Loss Data Analytics, Second Edition

An open text authored by the Actuarial Community

Contents

Preface

Date: 04 October 2024

Book Description

Loss Data Analytics is an interactive, online, freely available text.

- The online version contains many interactive objects (quizzes, computer demonstrations, interactive graphs, video, and the like) to promote *deeper learning.*
- A subset of the book is available for *offline reading* in pdf and EPUB formats.
- The online text will be available in multiple languages to promote access to a *worldwide audience.*

What will success look like?

The online text will be freely available to a worldwide audience. The online version will contain many interactive objects (quizzes, computer demonstrations, interactive graphs, video, and the like) to promote deeper learning. Moreover, a subset of the book will be available in pdf format for low-cost printing. The online text will be available in multiple languages to promote access to a worldwide audience.

How will the text be used?

This book will be useful in actuarial curricula worldwide. It will cover the loss data learning objectives of the major actuarial organizations. Thus, it will be suitable for classroom use at universities as well as for use by independent learners seeking to pass professional actuarial examinations. Moreover, the text will also be useful for the continuing professional development of actuaries and other professionals in insurance and related financial risk management industries.

Why is this good for the profession?

An online text is a type of open educational resource (OER). One important benefit of an OER is that it equalizes access to knowledge, thus permitting a broader community to learn about the actuarial profession. Moreover, it has the capacity to engage viewers through active learning that deepens the learning process, producing analysts more capable of solid actuarial work.

Why is this good for students and teachers and others involved in the learning process? Cost is often cited as an important factor for students and teachers in textbook selection (see a recent post on the $400 textbook). Students will also appreciate the ability to "carry the book around" on their mobile devices.

Why loss data analytics?

The intent is that this type of resource will eventually permeate throughout the actuarial curriculum. Given the dramatic changes in the way that actuaries treat data, loss data seems like a natural place to start. The idea behind the name *loss data analytics* is to integrate classical loss data models from applied probability with modern analytic tools. In particular, we recognize that big data (including social media and usage based insurance) are here to stay and that high speed computation is readily available.

Project Goal

The project goal is to have the actuarial community author our textbooks in a collaborative fashion. To get involved, please visit our Open Actuarial Textbooks Project Site.

Acknowledgements

Edward Frees acknowledges the John and Anne Oros Distinguished Chair for Inspired Learning in Business which provided seed money to support the project. Frees and his Wisconsin colleagues also acknowledge a Society of Actuaries Center of Excellence Grant that provided funding to support work in dependence modeling and health initiatives. Wisconsin also provided an education innovation grant that provided partial support for the many students who have worked on this project.

We acknowledge the Society of Actuaries for permission to use problems from their examinations.

We thank Rob Hyndman, Monash University, for allowing us to use his excellent style files to produce the online version of the book.

We thank Yihui Xie and his colleagues at Rstudio for the R bookdown package that allows us to produce this book.

We also wish to acknowledge the support and sponsorship of the Interna-

tional Association of Black Actuaries in our joint efforts to provide actuarial educational content to all.

Contributors

The project goal is to have the actuarial community author our textbooks in a collaborative fashion. The following contributors have taken a leadership role in developing *Loss Data Analytics*.

Zeinab Amin

- **Zeinab Amin** is a Professor at the Department of Mathematics and Actuarial Science and Associate Provost for Assessment and Accreditation at the American University in Cairo (AUC). Amin holds a PhD in Statistics and is an Associate of the Society of Actuaries. Amin is the recipient of the 2016 Excellence in Academic Service Award and the 2009 Excellence in Teaching Award from AUC. Amin has designed and taught a variety of statistics and actuarial science courses. Amin's current area of research includes quantitative risk assessment, reliability assessment, general statistical modelling, and Bayesian statistics.
- **Katrien Antonio**, KU Leuven

Jean-François Bégin

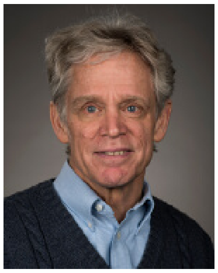
- **Jean-François Bégin** is an Assistant Professor in the Department of Statistics and Actuarial Science at Simon Fraser University in British Columbia, Canada. Bégin holds a PhD in Financial Engineering from HEC Montréal, Canada, and is a Fellow of the Society of Actuaries and of the Canadian Institute of Actuaries. His current research interests include financial modelling, financial econometrics, Bayesian statistics, filtering methods, credit risk, option pricing, and pension economics. Bégin has designed and taught a variety of actuarial finance and actuarial communication courses.
- **Jan Beirlant**, KU Leuven

Arthur Charpentier

- **Arthur Charpentier** is a professor in the Department of Mathematics at the Université du Québec á Montréal. Prior to that, he worked at a large general insurance company in Hong Kong, China, and the French Federation of Insurers in Paris, France. He received a MS on mathematical economics at Université Paris Dauphine and a MS in actuarial science at ENSAE (National School of Statistics) in Paris, and a PhD degree from KU Leuven, Belgium. His research interests include econometrics, applied probability and actuarial science. He has published several books (the most recent one on *Computational Actuarial Science with R*, CRC) and papers on a variety of topics. He is a Fellow of the French Institute of Actuaries, and was in charge of the ‘Data Science for Actuaries’ program from 2015 to 2018.

Curtis Gary Dean

- **Curtis Gary Dean** is the Lincoln Financial Distinguished Professor of Actuarial Science at Ball State University. He is a Fellow of the Casualty Actuarial Society and a CFA charterholder. He has extensive practical experience as an actuary at American States Insurance, SAFECO, and Travelers. He has served the CAS and actuarial profession as chair of the Examination Committee, first editor-in-chief for *Variance: Advancing the Science of Risk*, and as a member of the Board of Directors and the Executive Council. He contributed a chapter to *Predictive Modeling Applications in Actuarial Science* published by Cambridge University Press.



Edward (Jed) Frees

- **Edward (Jed) Frees** is an emeritus professor, formerly the Hickman-Larson Chair of Actuarial Science at the University of Wisconsin-Madison. He is a Fellow of both the Society of Actuaries and the American Statistical Association. He has published extensively (a four-time winner of the Halmstad and Prize for best paper published in the actuarial literature) and has written three books. He also is a co-editor of the two-volume series *Predictive Modeling Applications in Actuarial Science* published by Cambridge University Press.

Guojun Gan

- **Guojun Gan** is an associate professor in the Department of Mathematics at the University of Connecticut, where he has been since August 2014. Prior to that, he worked at a large life insurance company in Toronto, Canada for six years. He received a BS degree from Jilin University, Changchun, China, in 2001 and MS and PhD degrees from York University, Toronto, Canada, in 2003 and 2007, respectively. His research interests include data mining and actuarial science. He has published several books and papers on a variety of topics, including data clustering, variable annuity, mathematical finance, applied statistics, and VBA programming.

Lisa Gao

- **Lisa Gao** is a PhD candidate in the Risk and Insurance department at the University of Wisconsin-Madison. She holds a BMath in Actuarial Science and Statistics from the University of Waterloo and is an Associate of the Society of Actuaries.
- **José Garrido**, Concordia University

Lei (Larry) Hua

- **Lei (Larry) Hua** is an Associate Professor of Actuarial Science at Northern

Illinois University. He earned a PhD degree in Statistics from the University of British Columbia. He is an Associate of the Society of Actuaries. His research work focuses on multivariate dependence modeling for non-Gaussian phenomena and innovative applications for financial and insurance industries.

Noriszura Ismail

- **Noriszura Ismail** is a Professor and Head of Actuarial Science Program, Universiti Kebangsaan Malaysia (UKM). She specializes in Risk Modelling and Applied Statistics. She obtained her BSc and MSc (Actuarial Science) in 1991 and 1993 from University of Iowa, and her PhD (Statistics) in 2007 from UKM. She also passed several papers from Society of Actuaries in 1994. She has received several research grants from Ministry of Higher Education Malaysia (MOHE) and UKM, totaling about MYR1.8 million. She has successfully supervised and co-supervised several PhD students (13 completed and 11 on-going). She currently has about 180 publications, consisting of 88 journals and 95 proceedings.

Joseph H.T. Kim

- **Joseph H.T. Kim**, Ph.D., FSA, CERA, is Associate Professor of Applied Statistics at Yonsei University, Seoul, Korea. He holds a Ph.D. degree in Actuarial Science from the University of Waterloo, at which he taught as Assistant Professor. He also worked in the life insurance industry. He has published papers in *Insurance Mathematics and Economics*, *Journal of Risk and Insurance*, *Journal of Banking and Finance*, *ASTIN Bulletin*, and *North American Actuarial Journal*, among others.

Nii-Armah Okine

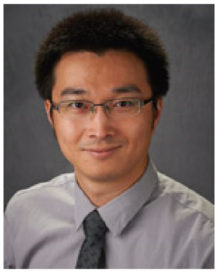
- **Nii-Armah Okine** is an assistant professor at the Mathematical Sciences Department at Appalachian State University. He holds a Ph.D. in Business (Actuarial Science) from the University of Wisconsin - Madison and obtained his master's degree in Actuarial science from Illinois State University. His research interest includes micro-level reserving, joint longitudinal-survival modeling, dependence modeling, micro-insurance, and machine learning.

Rajesh (Raj) Sahasrabuddhe

- **Rajesh (Raj) Sahasrabuddhe** is a Partner and Philadelphia Office Leader with Oliver Wyman Actuarial Consulting. Raj is a Fellow of the Casualty Actuarial Society (CAS), an Associate of the Canadian Institute of Actuaries, and a Member of the American Academy of Actuaries. Raj has been an active volunteer with CAS Admissions committees throughout his career, including a term as Chairperson of the Syllabus Committee from 2010 to 2013. He currently serves on the MAS-II Examination Committee. He has authored or co-authored papers that have appeared on syllabi for both the CAS and Society of Actuaries.

Emine Selin Sarıdaş

- **Emine Selin Sarıdaş** is a doctoral candidate in the Statistics department of Mimar Sinan University. She holds a bachelor degree in Actuarial Science with a minor in Economics and a master degree in Actuarial Science from Hacettepe University. Her research interest includes dependence modeling, regression, loss models and life contingencies.

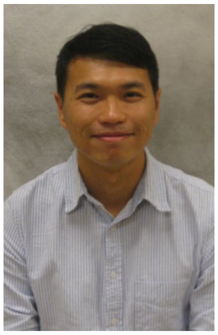


Peng Shi

- **Peng Shi** is an associate professor in the Risk and Insurance Department at the Wisconsin School of Business. He is also the Charles & Laura Albright Professor in Business and Finance. Professor Shi is an Associate of the Casualty Actuarial Society (ACAS) and a Fellow of the Society of Actuaries (FSA). He received a Ph.D. in actuarial science from the University of Wisconsin-Madison. His research interests are problems at the intersection of insurance and statistics. He has won several research awards, including the Charles A. Hachemeister Prize, the Ronald Bornhuetter Loss Reserve Prize, and the American Risk and Insurance Association Prize.

Nariankadu D. Shyamalkumar (Shyamal)

- **Nariankadu D. Shyamalkumar (Shyamal)** is an associate professor in the Department of Statistics and Actuarial Science at The University of Iowa. He is an Associate of the Society of Actuaries, and has volunteered in various elected and non-elected roles within the SoA. Having a broad theoretical interest as well as interest in computing, he has published in prominent actuarial, computer science, probability theory, and statistical journals. Moreover, he has worked in the financial industry, and since then served as an independent consultant to the insurance industry. He has experience educating actuaries in both Mexico and the US, serving in the roles of directing an undergraduate program, and as a graduate adviser for both masters and doctoral students.



Jianxi Su

- **Jianxi Su** is an Assistant Professor at the Department of Statistics at Purdue University. He is the Associate Director of Purdue's Actuarial Science. Prior to joining Purdue in 2016, he completed the PhD at York University (2012-2015). He obtained the Fellow of the Society of Actuaries (FSA) in 2017. His research expertise are in dependence modelling, risk management, and pricing. During the PhD candidature, Jianxi also worked as a research associate at the Model Validation and ORSA Implementation team of Sun Life Financial (Toronto office).

Chong It Tan

- **Chong It Tan** is a senior lecturer at Macquarie University in Australia, where he has served as the undergraduate actuarial program director since 2018. He obtained his PhD in 2015 from Nanyang Technological University in Singapore. He is a fully qualified actuary, holding the credentials from both the US Society of Actuaries and Australian Actuaries Institute. His major research interests are mortality modelling, longevity risk management and bonus-malus systems.

Tim Verdonck

- **Tim Verdonck** is associate professor at the University of Antwerp. He has a degree in Mathematics and a PhD in Science: Mathematics, obtained at the University of Antwerp. During his PhD he successfully took the Master in Insurance and the Master in Financial and Actuarial Engineering, both at KU Leuven. His research focuses on the adaptation and application of robust statistical methods for insurance and finance data.

Krupa Viswanathan

- **Krupa Viswanathan** is an Associate Professor in the Risk, Insurance and Healthcare Management Department in the Fox School of Business, Temple University. She is an Associate of the Society of Actuaries. She teaches courses in Actuarial Science and Risk Management at the undergraduate and graduate levels. Her research interests include corporate governance of insurance companies, capital management, and sentiment analysis. She received her Ph.D. from The Wharton School of the University of Pennsylvania.

Reviewers

Our goal is to have the actuarial community author our textbooks in a collaborative fashion. Part of the writing process involves many reviewers who generously donated their time to help make this book better. They are:

- Yair Babab
- David Back, Liberty Mutual
- Chunsheng Ban, Ohio State University
- Vytaras Brazauskas, University of Wisconsin - Milwaukee
- Yvonne Chueh, Central Washington University
- Chun Yong Chew, Universiti Tunku Abdul Rahman (UTAR)
- Benjamin Côté, Université Laval
- Eren Dodd, University of Southampton
- Gordon Enderle, University of Wisconsin - Madison
- Rob Erhardt, Wake Forest University
- Runhun Feng, University of Illinois
- Brian Hartman, Brigham Young University
- Liang (Jason) Hong, University of Texas at Dallas
- Fei Huang, Australian National University
- Hirokazu (Iwahiro) Iwasawa

- Himchan Jeong, University of Connecticut
- Min Ji, Towson University
- Paul Herbert Johnson, University of Wisconsin - Madison
- Dalia Khalil, Cairo University
- Samuel Kolins, Lebonan Valley College
- Andrew Kwon-Nakamura, Zurich North America
- Ambrose Lo, University of Iowa
- Mélina Mailhot, Concordia University
- Mark Maxwell, University of Texas at Austin
- Tatjana Miljkovic, Miami University
- Bell Ouelega, American University in Cairo
- Zhiyu (Frank) Quan, University of Connecticut
- Jiandong Ren, Western University
- Margie Rosenberg, University of Wisconsin - Madison
- Rajesh V. Sahasrabuddhe, Oliver Wyman
- Sherly Paola Alfonso Sanchez, Universidad Nacional de Colombia
- Ranee Thiagarajah, Illinois State University
- Ping Wang, Saint Johns University
- Chengguo Weng, University of Waterloo
- Toby White, Drake University
- Michelle Xia, Northern Illinois University
- Di (Cindy) Xu, University of Nebraska - Lincoln
- Lina Xu, Columbia University
- Lu Yang, University of Amsterdam
- Chun Yong
- Jorge Yslas, University of Copenhagen
- Jeffrey Zheng, Temple University
- Hongjuan Zhou, Arizona State University

Other Collaborators

- Alyaa Nuval Binti Othman, Aisha Nuval Binti Othman, and Khairina (Rina) Binti Ibraham were three of many students at the Univeristy of Wiscinson-Madison that helped with the text over the years.
- Maggie Lee, Macquarie University, and Anh Vu (then at University of New South Wales) contributed the end of the section quizzes.
- Jeffrey Zheng, Temple University, Lu Yang (University of Amsterdam), and Paul Johnson, University of Wisconsin-Madison, led the work on the glossary.

Version Number

- This is **Version 2.0**, October 2024. Edited by Hélène Cossette, Edward (Jed) Frees, Brian Hartman, and Tim Higgins.
- Version 1.1, August 2020. Edited by Edward (Jed) Frees and Paul Johnson.
- Version 1.0, January 2020, was edited by Edward (Jed) Frees.

You can also access pdf and epub (current and older) versions of the text in our Offline versions of the text.

For our Readers

We hope that you find this book worthwhile and even enjoyable. For your convenience, at our Github Landing site (https://openacttexts.github.io/), you will find links to the book that you can (freely) download for offline reading, including a pdf version (for Adobe Acrobat) and an EPUB version suitable for mobile devices. Data for running our examples are available at the same site.

In developing this book, we are emphasizing the online version that has lots of great features such as a glossary, code and solutions to examples that you can be revealed interactively. For example, you will find that the statistical code is hidden and can only be seen by clicking on terms such as

We hide the code because we don't want to insist that you use the `R` statistical software (although we like it). Still, we encourage you to try some statistical code as you read the book – we have opted to make it easy to learn `R` as you go. We have set up a separate R Code for Loss Data Analytics site to explain more of the details of the code.

Like any book, we have a set of notations and conventions. It will probably save you time if you regularly visit our Appendix Chapter 21 to get used to ours.

Freely available, interactive textbooks represent a new venture in actuarial education and we need your input. Although a lot of effort has gone into the development, we expect hiccoughs. Please let your instructor know about opportunities for improvement, write us through our project site, or contact chapter contributors directly with suggested improvements.

1

Loss Data and Insurance Activities

Chapter Preview. This book introduces readers to methods of analyzing insurance data. Section 1.1 begins with a discussion of why the use of data is important in the insurance industry. Section 1.2 gives a general overview of the purposes of analyzing insurance data which is reinforced in the Section 1.3 case study. Naturally, there is a huge gap between the broad goals summarized in the overview and a case study application; this gap is covered through the methods and techniques of data analysis covered in the rest of the text.

1.1 Data Driven Insurance Activities

In this section, you learn how to:

- Summarize the importance of insurance to consumers and the economy
- Describe the role that data plays in managing insurance activities
- Identify data generating events associated with the timeline of a typical insurance contract

1.1.1 Nature and Relevance of Insurance

This book introduces the process of using data to make decisions in an insurance context. It does not assume that readers are familiar with insurance but introduces insurance concepts as needed. Insurance is the exchange of a certain amount, known as a premium, for a promise to compensate another party upon the occurrence of an insured event.

If you are new to insurance, then it is probably easiest to think about an insurance policy that covers the contents of an apartment or house that you are renting (known as renters insurance) or the contents and property of a building that is owned by you or a friend (known as homeowners insurance). Another common example is automobile insurance. In the event of an accident, this policy may cover damage to your vehicle, damage to other vehicles in the accident, as well as medical expenses of those injured in the accident.

One way to think about the nature of insurance is who buys it. Renters, homeowners, and auto insurance are examples of personal insurance in that these are policies issued to people. Businesses also buy insurance, such as coverage on their properties, and this is known as commercial insurance. The seller, an insurance company, is also known as an insurer. Even insurance companies need insurance; this is known as reinsurance.

Another way to think about the nature of insurance is the type of risk being covered. In the U.S., policies such as renters and homeowners are known as property insurance whereas a policy such as auto that covers medical damages to people is known as casualty insurance. In the rest of the world, these are both known as non-life or general insurance, to distinguish them from life insurance.

Both life and non-life insurances are important components of the world economy. The The Organization for Economic Cooperation and Development (OECD) estimates that direct insurance premiums in the OECD (Organization for Economic Cooperation and Development) countries for 2020 was 2,520,220 for life and 2,704,799 for non-life; these figures are in *millions of U.S. dollars*. The total represents 9.447% of the OECD gross domestic product (GDP). As examples, premiums accounted for 30.9% of GDP in Luxembourg and 17.0% of GDP in Chinese Taipei (the two highest in the study) and represented 12.5% of GDP in the United States. Both life and non-life insurances represent important economic activities.

Insurance affects the financial livelihoods of many and, by almost any measure, insurance is a major economic activity. As noted earlier, on a global level insurance premiums comprised nearly 9.5% of GDP in 2020. On a personal level, almost everyone owning a home has insurance to protect themselves in the event of a fire, hailstorm, or some other calamitous event. Almost every country requires insurance for those driving a car. In sum, insurance plays an important role in the economies of nations and the lives of individuals.

1.1.2 Why Data Driven?

Insurance is a data-driven industry. Like all major corporations and organizations, insurers use data when trying to decide how much to pay employees, how many employees to retain, how to market their services and products, how to forecast financial trends, and so on. These represent general areas of activities that are not specific to the insurance industry. Although each industry has its own data nuances and needs, the collection, analysis and use of data is an activity shared by all, from the internet giants to a small business, by public and governmental organizations, and is not specific to the insurance industry. You will find that the data collection and analysis methods and tools introduced in this text are relevant for all.

In any data-driven industry, deriving and extracting information from data is critical. Making data-driven business decisions has been described as business analytics, business intelligence, and data science. These terms, among others, are sometimes used interchangeably and sometimes refer to distinct applications. *Business intelligence* may focus on processes of collecting data, often through databases and data warehouses, whereas *business analytics* utilizes tools and methods for statistical analyses of data. In contrast to these two terms that emphasize business applications, the term *data science* can encompass broader data related applications in many scientific domains. For our purposes, we use the term analytics to refer to the process of using data to make decisions. This process involves gathering data, understanding concepts and models of uncertainty, making general inferences, and communicating results. Chapter 2 describes data analytics in further detail.

When introducing methods in this text, we focus on **loss data** that arise from, or are related to, obligations in insurance contracts. This could be the amount of damage to one's apartment under a renter's insurance agreement, the amount needed to compensate someone that you hurt in a driving accident, and the like. We call such payments an insurance claim. With this focus, we are able to introduce and directly use generally applicable statistical tools and techniques.

1.1.3 Insurance Processes

Yet another way to think about the nature of insurance is by the duration of an insurance contract, known as the term. This text will focus on short-term insurance contracts. By short-term, we mean contracts where the insurance coverage is typically provided for a year or six months. Most non-life commercial and personal contracts are for a year so that is our default duration. An important exception is U.S. auto policies that are often six months in length.

In contrast, we typically think of life insurance as a long-term contract where the default is to have a multi-year contract. For example, if a person 25 years old purchases a whole life policy that pays upon death of the insured and that person does not die until age 100, then the contract is in force for 75 years.

There are other important differences between life and non-life products. In life insurance, the benefit amount is often stipulated in the contract provisions. In contrast, most non-life contracts provide for compensation of insured losses which are unknown before the accident. (There are usually limits placed on the compensation amounts.) In a life insurance contract that stretches over many years, the time value of money plays a prominent role. In a non-life contract, the random amount of compensation takes priority.

In both life and non-life insurances, the frequency of claims is very important. For many life insurance contracts, the insured event (such as death) happens only once. In contrast, for non-life insurances such as automobile, it is common for individuals (especially young male drivers) to get into more than one accident during a year. So, our models need to reflect this observation; we introduce different frequency models that you may also see when studying life insurance.

For short-term insurance, the framework of the probabilistic model is straightforward. We think of a one-period model (the period length, e.g., one year, will be specified in the situation).

- At the beginning of the period, the insured pays the insurer a known premium that is agreed upon by both parties to the contract.
- At the end of the period, the insurer reimburses the insured for a (possibly multivariate) random loss.

This framework will be developed as we proceed; but we first focus on integrating this framework with concerns about how the data may arise. From an insurer's viewpoint, contracts may be only for a year but they tend to be renewed. Moreover, payments arising from claims during the year may extend well beyond a single year. One way to describe the data arising from operations of an insurance company is to use a timeline granular approach. A **process** approach provides an overall view of the events occurring during the life of an insurance contract, and their nature – random or planned, loss events (claims) and contract changes events, and so forth. In this micro oriented view, we can think about what happens to a contract at various stages of its existence.

Figure 1.1 traces a timeline of a typical insurance contract. Throughout the life of the contract, the company regularly processes events such as premium collection and valuation, described in Section 1.2; these are marked with an **x** on the timeline. Non-regular and unanticipated events also occur. To illustrate, t_2 and t_4 mark the event of an insurance claim (some contracts, such as life insurance, can have only a single claim). Times t_3 and t_5 mark events when a policyholder wishes to alter certain contract features, such as the choice of a deductible or the amount of coverage. From a company perspective, one can even think about the contract initiation (arrival, time t_1) and contract termination (departure, time t_6) as uncertain events. (Alternatively, for some purposes, you may condition on these events and treat them as certain.)

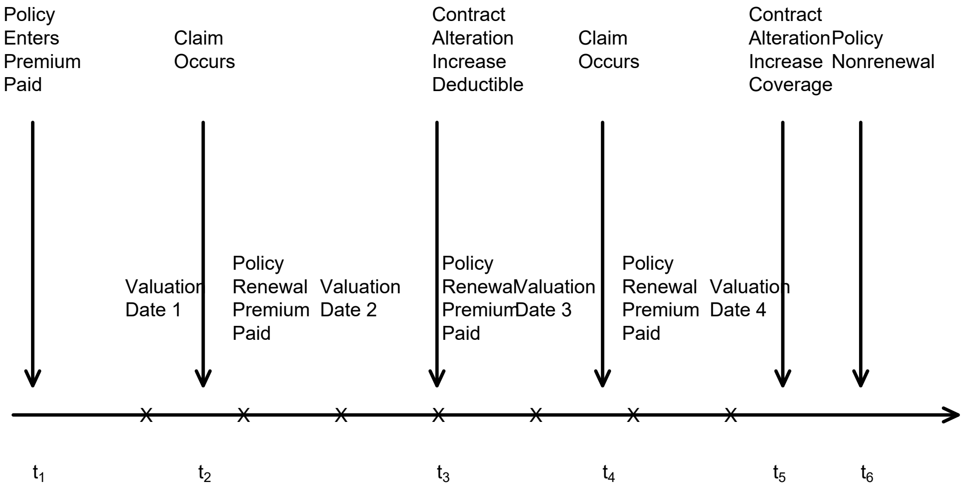

Policy Enters Premium Paid
Claim Occurs
Contract Alteration Increase Deductible
Claim Occurs
Contract Alteration Increase Coverage
Policy Nonrenewal
Valuation Date 1
Policy Renewal Premium Paid
Valuation Date 2
Policy Renewal Premium Paid
Valuation Date 3
Policy Renewal Premium Paid
Valuation Date 4
x
x
x
x
x
x
x
t_1
t_2
t_3
t_4
t_5
t_6

FIGURE 1.1: **Timeline of a Typical Insurance Policy.** Arrows mark the occurrences of random events. Each x marks the time of scheduled events that are typically non-random.

1.2 Insurance Company Operations

In this section, you learn how to:

- Describe five major operational areas of insurance companies.
- Identify the role of data and analytics opportunities within each operational area.

Armed with insurance data, the end goal is to use data to make decisions. We will learn more about methods of analyzing and extrapolating data in future chapters. To begin, let us think about why we want to do the analysis. We take the insurance company's viewpoint (not the insured person) and introduce ways of bringing money in, paying it out, managing costs, and making sure that we have enough money to meet obligations. The emphasis is on insurance-specific operations rather than on general business activities such as advertising, marketing, and human resources management.

Specifically, in many insurance companies, it is customary to aggregate detailed insurance processes into larger operational units; many companies use these functional areas to segregate employee activities and areas of responsibilities. Actuaries, other financial analysts, and insurance regulators work within these units and use data for the following activities:

1. **Initiating Insurance**. At this stage, the company makes a decision as to whether or not to take on a risk (the underwriting stage) and assign an appropriate premium (or rate). Insurance analytics has its actuarial roots in *ratemaking*, where analysts seek to determine the right price for the right risk.
2. **Renewing Insurance**. Many contracts, particularly in general insurance, have relatively short durations such as 6 months or a year. Although there is an implicit expectation that such contracts will be renewed, the insurer has the opportunity to decline coverage and to adjust the premium. Analytics is also used at this policy renewal stage where the goal is to retain profitable customers.
3. **Claims Management**. Analytics has long been used in (1) detecting and preventing claims fraud, (2) managing claim costs, including identifying the appropriate support for claims handling expenses, as well as (3) understanding excess layers for reinsurance and retention.
4. **Loss Reserving**. Analytic tools are used to provide management

with an appropriate estimate of future obligations and to quantify the uncertainty of those estimates.

5. **Solvency and Capital Allocation**. Deciding on the requisite amount of capital and on ways of allocating capital among alternative investments are also important analytics activities. Companies must understand how much capital is needed so that they have sufficient flow of cash available to meet their obligations at the times they are expected to materialize (solvency). This is an important question that concerns not only company managers but also customers, company shareholders, regulatory authorities, as well as the public at large. Related to issues of how much capital is the question of how to allocate capital to differing financial projects, typically to maximize an investor's return. Although this question can arise at several levels, insurance companies are typically concerned with how to allocate capital to different lines of business within a firm and to different subsidiaries of a parent firm.

Although data represent a critical component of solvency and capital allocation, other components including the local and global economic framework, the financial investments environment, and quite specific requirements according to the regulatory environment of the day, are also important. Because of the background needed to address these components, we do not address solvency, capital allocation, and regulation issues in this text.

Nonetheless, for all operating functions, we emphasize that analytics in the insurance industry is not an exercise that a small group of analysts can do by themselves. It requires an insurer to make significant investments in their information technology, marketing, underwriting, and actuarial functions. As these areas represent the primary end goals of the analysis of data, additional background on each operational unit is provided in the following subsections.

1.2.1 Initiating Insurance

Setting the price of an insurance product can be a perplexing problem. This is in contrast to other industries such as manufacturing where the cost of a product is (relatively) known and provides a benchmark for assessing a market demand price. Similarly, in other areas of financial services, market prices are available and provide the basis for a market-consistent pricing structure of products. However, for many lines of insurance, the cost of a product is uncertain and market prices are unavailable. Expectations of the random cost is a reasonable place to start for a price. (If you have studied finance, then you will recall that an expectation is the optimal price for a risk-neutral insurer.) It has been traditional in insurance pricing to begin with the expected cost. Insurers then add margins to this, to account for the product's riskiness, expenses incurred in servicing the product, and an allowance for profit/surplus of the company.

Use of expected costs as a foundation for pricing is prevalent in some lines of the insurance business. These include automobile and homeowners insurance. For these lines, analytics has served to sharpen the market by making the calculation of the product's expected cost more precise. The increasing availability of the internet to consumers has also promoted transparency in pricing; in today's marketplace, consumers have ready access to competing quotes from a host of insurers. Insurers seek to increase their market share by refining their risk classification systems, thus achieving a better approximation of the products' prices and enabling cream-skimming underwriting strategies ("cream-skimming" is a phrase used when the insurer underwrites only the best risks). Surveys (e.g., Earnix (2013)) indicate that pricing is the most common use of analytics among insurers.

Underwriting, the process of classifying risks into homogeneous categories and assigning policyholders to these categories, lies at the core of ratemaking. Policyholders within a class (category) have similar risk profiles and so are charged the same insurance price. This is the concept of an actuarially fair premium; it is fair to charge different rates to policyholders only if they can be separated by identifiable risk factors. An early article, *Two Studies in Automobile Insurance Ratemaking* (Bailey and LeRoy, 1960), provided a catalyst to the acceptance of analytic methods in the insurance industry. This paper addresses the problem of classification ratemaking. It describes an example of automobile insurance that has five use classes cross-classified with four merit rating classes. At that time, the contribution to premiums for use and merit rating classes were determined independently of each other. Thinking about the interacting effects of different classification variables is a more difficult problem.

When the risk is initially obtained, the insurer's obligations can be managed by imposing contract parameters that modify contract payouts. Chapter 4 describes common modifications including coinsurance, deductibles and policy upper limits.

1.2.2 Renewing Insurance

Insurance is a type of financial service and, like many service contracts, insurance coverage is often agreed upon for a limited time period at which time coverage commitments are complete. Particularly for general insurance, the need for coverage continues and so efforts are made to issue a new contract providing similar coverage when the existing contract comes to the end of its term. This is called *policy renewal.* Renewal issues can also arise in life insurance, e.g., term (temporary) life insurance. At the same time other contracts, such as life annuities, terminate upon the insured's death and so issues of renewability are irrelevant.

In the absence of legal restrictions, at renewal the insurer has the opportunity to:

- accept or decline to underwrite the risk; and
- determine a new premium, possibly in conjunction with a new classification of the risk.

Risk classification and rating at renewal is based on two types of information. First, at the initial stage, the insurer has available many rating variables upon which decisions can be made. Many variables are not likely to change, e.g., sex, whereas others are likely to change, e.g., age, and still others may or may not change, e.g., credit score. Second, unlike the initial stage, at renewal the insurer has available a history of policyholder's loss experience, and this history can provide insights into the policyholder that are not available from rating variables. Modifying premiums with claims history is known as *experience rating*, also sometimes referred to as *merit rating*.

Experience rating methods are either applied retrospectively or prospectively. With retrospective methods, a refund of a portion of the premium is provided to the policyholder in the event of favorable (to the insurer) experience. Retrospective premiums are common in life insurance arrangements (where policyholders earn dividends in the U.S., bonuses in the U.K., and profit sharing in Israeli term life coverage). In general insurance, prospective methods are more common, where favorable insured experience is rewarded through a lower renewal premium.

Claims history can provide information about a policyholder's risk appetite. For example, in personal lines it is common to use a variable to indicate whether or not a claim has occurred in the last three years. As another example, in a commercial line such as worker's compensation, one may look to a policyholder's average claim frequency or severity over the last three years. Claims history can reveal information that is otherwise hidden (to the insurer) about the policyholder.

1.2.3 Claims and Product Management

In some of types of insurance, the process of paying claims for insured events is relatively straightforward. For example, in life insurance, a simple death certificate is all that is needed to pay the benefit amount as provided in the contract. However, in non-life areas such as property and casualty insurance, the process can be much more complex. Think about a relatively simple insured event such as an automobile accident. Here, it is often required to determine which party is at fault and then one needs to assess damage to all of the vehicles and people involved in the incident, both insured and non-insured. Further, the expenses incurred in assessing the damages must be assessed, and so forth. The process of determining coverage, legal liability, and settling claims is known as claims adjustment.

Insurance managers sometimes use the phrase claims leakage to mean dollars lost through claims management inefficiencies. There are many ways in which analytics can help manage the claims process, c.f., Gorman and Swenson (2013). Historically, the most important has been fraud detection. The claim adjusting process involves reducing information asymmetry (the claimant knows what happened; the company knows some of what happened). Mitigating fraud is an important part of the claims management process.

Fraud detection is only one aspect of managing claims. More broadly, one can think about claims management as consisting of the following components:

- **Claims triaging**. Just as in the medical world, early identification and appropriate handling of high cost claims (patients, in the medical world), can lead to dramatic savings. For example, in workers compensation, insurers look to achieve early identification of those claims that run the risk of high medical costs and a long payout period. Early intervention into these cases could give insurers more control over the handling of the claim, the medical treatment, and the overall costs with an earlier return-to-work.
- **Claims processing**. The goal is to use analytics to identify routine situations that are anticipated to have small payouts. More complex situations may require more experienced adjusters and legal assistance to appropriately handle claims with high potential payouts.
- **Adjustment decisions**. Once a complex claim has been identified and assigned to an adjuster, analytic driven routines can be established to aid subsequent decision-making processes. Such processes can also be helpful for adjusters in developing case reserves, an estimate of the insurer's future liability. This is an important input to the insurer's loss reserves, described in Section 1.2.4.

In addition to the insured's reimbursement for losses, the insurer also needs to be concerned with another source of revenue outflow, expenses. Loss adjustment expenses are part of an insurer's cost of managing claims. Analytics can be used to reduce expenses directly related to claims handling (allocated) as well as general staff time for overseeing the claims processes (unallocated). The insurance industry has high operating costs relative to other portions of the financial services sectors.

In addition to claims payments, there are many other ways in which insurers use data to manage their products. We have already discussed the need for analytics in underwriting, that is, risk classification at the initial acquisition and renewal stages. Insurers are also interested in which policyholders elect to renew their contracts and, as with other products, monitor customer loyalty.

Analytics can also be used to manage the portfolio, or collection, of risks that an insurer has acquired. As described in Chapter 13, after the contract has been agreed upon with an insured, the insurer may still modify its net obligation by entering into a reinsurance agreement. This type of agreement is with a reinsurer, an insurer of an insurer. It is common for insurance companies to purchase insurance on its portfolio of risks to gain protection from unusual events, just as people and other companies do.

1.2.4 Loss Reserving

An important feature that distinguishes insurance from other sectors of the economy is the timing of the exchange of considerations. In manufacturing, payments for goods are typically made at the time of a transaction. In contrast, for insurance, money received from a customer occurs in advance of benefits or services; these are rendered at a later date if the insured event occurs. This leads to the need to hold a reservoir of wealth to meet future obligations in respect to obligations made, and to gain the trust of the insureds that the company will be able to fulfill its commitments. The size of this reservoir of wealth, and the importance of ensuring its adequacy, is a major concern for the insurance industry.

Setting aside money for unpaid claims is known as loss reserving; in some jurisdictions, reserves are also known as *technical provisions*. We saw in Figure 1.1 several times at which a company summarizes its financial position; these times are known as valuation dates. Claims that arise prior to valuation dates have either been paid, are in the process of being paid, or are about to be paid; claims in the future of these valuation dates are unknown. A company must estimate these outstanding liabilities when determining its financial strength. Accurately determining loss reserves is important to insurers for many reasons.

1. Loss reserves represent an anticipated claim that the insurer owes its customers. Under-reserving may result in a failure to meet claim liabilities. Conversely, an insurer with excessive reserves may present a conservative estimate of surplus and thus portray a weaker financial position than it truly has.
2. Reserves provide an estimate for the unpaid cost of insurance that can be used for pricing contracts.

3. Loss reserving is required by laws and regulations. The public has a strong interest in the financial strength and solvency of insurers.
4. In addition to regulators, other stakeholders such as insurance company management, investors, and customers make decisions that depend on company loss reserves. Whereas regulators and customers appreciate conservative estimates of unpaid claims, managers and investors seek more unbiased estimates to represent the true financial health of the company.

Loss reserving is a topic where there are substantive differences between life and general (also known as property and casualty, or non-life) insurance. In life insurance, the severity (amount of loss) is often not a source of uncertainty as payouts are specified in the contract. The frequency, driven by mortality of the insured, is a concern. However, because of the lengthy time for settlement of life insurance contracts, the time value of money uncertainty as measured from issue to date of payment can dominate frequency concerns. For example, for an insured who purchases a life contract at age 20, it would not be unusual for the contract to still be open in 60 years time, when the insured celebrates his or her 80th birthday. See, for example, Bowers et al. (1986) or Dickson et al. (2013) for introductions to reserving for life insurance. In contrast, for most lines of non-life business, severity is a major source of uncertainty and contract durations tend to be shorter.

1.3 Case Study: Wisconsin Property Fund

In this section, we use the Wisconsin Property Fund as a case study. You learn how to:

- Describe how data generating events can produce data of interest to insurance analysts.
- Produce relevant summary statistics for each variable.
- Describe how these summary statistics can be used in each of the major operational areas of an insurance company.

Let us illustrate the kind of data under consideration and the goals that we wish to achieve by examining the Local Government Property Insurance Fund (LGPIF), an insurance pool administered by the Wisconsin Office of the Insurance Commissioner. The LGPIF was established to provide property insurance for local government entities that include counties, cities, towns, villages, school districts, and library boards. The fund insures local government property such as government buildings, schools, libraries, and motor vehicles. It covers all property losses except those resulting from flood, earthquake, wear and tear, extremes in temperature, mold, war, nuclear reactions, and embezzlement or theft by an employee.

The fund covers over a thousand local government entities who pay approximately 25 million dollars in premiums each year and receive insurance coverage of about 75 billion. State government buildings are not covered; the LGPIF is for local government entities that have separate budgetary responsibilities and who need insurance to moderate the budget effects of uncertain insurable events. Coverage for local government property has been made available by the State of Wisconsin since 1911, thus providing a wealth of historical data.

In this illustration, we restrict consideration to claims from coverage of building and contents; we do not consider claims from motor vehicles and specialized equipment owned by local entities (such as snow plowing machines). We also consider only claims that are closed, with obligations fully met.

1.3.1 Fund Claims Variables: Frequency and Severity

At a fundamental level, insurance companies accept premiums in exchange for promises to compensate a policyholder upon the occurrence of an insured event. Indemnification is the compensation provided by the insurer for incurred hurt, loss, or damage that is covered by the policy. This compensation is also known as a *claim*. The extent of the payout, known as the *severity*, is a key financial expenditure for an insurer.

In terms of money outgo to customers, an insurer is indifferent to having ten claims of 100 when compared to one claim of 1,000. Nonetheless, it is common for insurers to study how often claims arise, known as the *frequency* of claims. The frequency is important for expenses, but it also influences contractual parameters (such as deductibles and policy limits that are described later) that are written to limit amounts paid for each occurrence of an insured event. Frequency is routinely monitored by insurance regulators and can be a key driver in the overall indemnification obligation of the insurer. We shall consider the frequency and severity as the two main claim variables that we wish to understand, model, and manage.

To illustrate, in 2010 there were 1,110 policyholders in the property fund who experienced a total of 1,377 claims. Table 1.1 shows the distribution. Almost two-thirds (0.637) of the policyholders did not have any claims and an additional 18.8% had only one claim. The remaining 17.5% (=1 - 0.637 -

TABLE 1.1: **2010 Claims Frequency Distribution**

Number	0	1	2	3	4	5	6	7	8	9 or more	Sum
Policies	707	209	86	40	18	12	9	4	6	19	1110
Claims	0	209	172	120	72	60	54	28	48	614	1377
Proportion	0.637	0.188	0.077	0.036	0.016	0.011	0.008	0.004	0.005	0.017	1

TABLE 1.2: **2010 Average Severity Distribution**

Minimum	First Quartile	Median	Mean	Third Quartile	Maximum
167	2,226	4,951	56,332	11,900	12,922,218

0.188) had more than one claim; the policyholder with the highest number recorded 239 claims. The average number of claims for this sample was 1.24 (=1377/1110).

For the severity distribution, a common approach is to examine the distribution of the sample of 1,377 claims. However, another common approach is to examine the distribution of the average claims of those policyholders with claims. In our 2010 sample, there were 403 (=1110-707) such policyholders. For 209 of these policyholders with one claim, the average claim equals the only claim they experienced. For the policyholder with highest frequency, the average claim is an average over 239 separately reported claim events.

Table 1.2 summarizes the sample distribution of average severities from the 403 policyholders who made a claim; it shows that the average claim amount was 56,330 (all amounts are in U.S. Dollars). However, the average gives only a limited look at the distribution. More information can be gleaned from the summary statistics which show a very large claim in the amount of 12,920,000. Figure 1.2 provides further information about the distribution of sample claims, showing a distribution that is dominated by this single large claim so that the histogram is not very helpful. Even when removing the large claim, you will find a distribution that is skewed to the right. A generally accepted technique is to work with claims in logarithmic units especially for graphical purposes; the corresponding figure in the right-hand panel is much easier to interpret.

1.3.2 Fund Rating Variables

Developing models to represent and manage the two outcome variables, frequency and severity, is the focus of the early chapters of this text. However, when actuaries and other financial analysts use those models, they do so in the context of external variables. In general statistical terminology, one might call these explanatory or predictor variables; there are many other names in

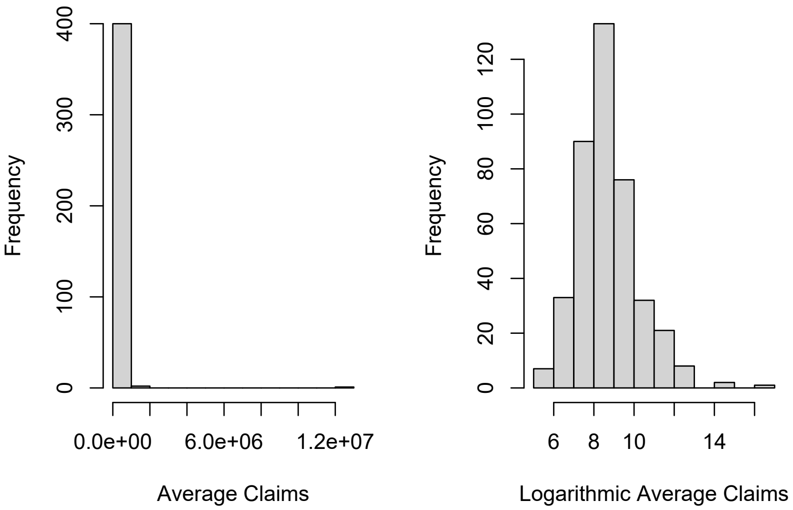


FIGURE 1.2: **Distribution of Positive Average Severities**

statistics, economics, psychology, and other disciplines. Because of our insurance focus, we call them rating variables as they are useful in setting insurance rates and premiums.

We earlier considered observations from a sample of 1,110 policyholders which may seem like a lot. However, as we will see in our forthcoming applications, because of the preponderance of zeros and the skewed nature of claims, actuaries typically yearn for more data. One common approach that we adopt here is to examine outcomes from multiple years, thus increasing the sample size. We will discuss the strengths and limitations of this strategy later but, at this juncture, we just wish to show the reader how it works.

Specifically, Table 1.3 shows that we now consider policies over five years of data, 2006, . . . , 2010, inclusive. The data begins in 2006 because there was a shift in claim coding in 2005 so that comparisons with earlier years are not helpful. To mitigate the effect of open claims, we consider policy years prior to 2011. An open claim means that not all of the obligations for the claim are known at the time of the analysis; for some claims, such an injury to a person in an auto accident or in the workplace, it can take years before costs are fully known.

Table 1.3 shows that the average claim varies over time, especially with the high 2010 value (that we saw was due to a single large claim)[1]. The total

[1] Note that the average severity in Table 1.3 differs from that reported in Table 1.2. This

TABLE 1.3: **Claims Summary by Policyholder**

Year	Average Frequency	Average Severity	Average	Number of Policy-holders
2006	0.951	9,695	32,498,186	1,154
2007	1.167	6,544	35,275,949	1,138
2008	0.974	5,311	37,267,485	1,125
2009	1.219	4,572	40,355,382	1,112
2010	1.241	20,452	41,242,070	1,110

TABLE 1.4: **Summary of Claim Frequency and Severity, Deductibles, and Coverages**

	Minimum	Median	Average	Maximum
Claim Frequency	0	0	1.109	263
Claim Severity	0	0	9,292	12,922,218
Deductible	500	1,000	3,365	100,000
Coverage (000's)	8.937	11,354	37,281	2,444,797

number of policyholders is steadily declining and, conversely, the coverage is steadily increasing. The coverage variable is the amount of coverage of the property and contents. Roughly, you can think of it as the maximum possible payout of the insurer. For our immediate purposes, the coverage is our first rating variable. Other things being equal, we would expect that policyholders with larger coverage have larger claims. We will make this vague idea much more precise as we proceed, and also justify this expectation with data.

For a different look at the 2006-2010 data, Table 1.4 summarizes the distribution of our two outcomes, frequency and claims amount. In each case, the average exceeds the median, suggesting that the two distributions are right-skewed. In addition, the table summarizes our continuous rating variables, coverage and deductible amount. The table also suggests that these variables also have right-skewed distributions.

Table 1.5 describes the rating variables considered in this chapter. Hopefully, these are variables that you think might naturally be related to claims outcomes. You can learn more about them in Frees et al. (2016b). To handle the skewness, we henceforth focus on logarithmic transformations of coverage and deductibles.

is because the former includes policyholders with zero claims where as the latter does not. This is an important distinction that we will address in later portions of the text.

Table 1.5. **Description of Rating Variables**

Variable	*Description*
EntityType	Categorical variable that is one of six types: (Village, City, County, Misc, School, or Town)
LnCoverage	Total building and content coverage, in logarithmic millions of dollars
LnDeduct	Deductible, in logarithmic dollars
AlarmCredit	Categorical variable that is one of four types: (0, 5, 10, or 15) for automatic smoke alarms in main rooms
NoClaimCredit	Binary variable to indicate no claims in the past two years
Fire5	Binary variable to indicate the fire class is below 5 (The range of fire class is 1 to 10)

For the *alarm credit* variable, a zero means that no automatic smoke alarms exist in any of the main rooms. In the same way, a 5 means they exist in some of the main rooms and a 10 means they exist in all of the main rooms. At the 15 level, facilities are monitored on a 24 hours per day, 7 days per week basis by a police, fire, or security company. A *fire rating* is a similar type of score. It reflects how prepared a community and area is for fires. While it mainly focuses on the local fire departments and water supply, there are other factors that contribute to an area's score. This rating is used to determine how likely it is for a fire to do severe damage before help arrives with 1 being the best and 10 the worst.

To get a sense of the relationship between the non-continuous rating variables and claims, Table 1.6 relates the claims outcomes to these categorical variables. Table 1.6 suggests substantial variation in the claim frequency and average severity of the claims by entity type. It also demonstrates higher frequency and severity for the `Fire5` variable and the reverse for the `NoClaimCredit` variable. The relationship for the `Fire5` variable is counter-intuitive in that one would expect lower claim amounts for those policyholders in areas with better public protection (when the protection code is five or less). Naturally, there are other variables that influence this relationship. We will see that these background variables are accounted for in the subsequent multivariate regression analysis, which yields an intuitive, appealing (negative) sign for the `Fire5` variable.

Tables 1.7 and 1.8 show the claims experience by alarm credit. It underscores the difficulty of examining variables individually. For example, when looking at the experience for all entities, we see that policyholders with no alarm credit have on average lower frequency and severity than policyholders with the highest (15%, with 24/7 monitoring by a fire station or security company) alarm credit. In particular, when we look at the entity type School, the frequency is 0.422 and the severity 25,523 for no alarm credit, whereas for the highest

TABLE 1.6: **Claims Summary by Entity Type, Fire Class, and No Claim Credit**

	Number of Policies	Claim Frequency	Average Severity
Village	1,341	0.452	10,645
City	793	1.941	16,924
County	328	4.899	15,453
Misc	609	0.186	43,036
School	1,597	1.434	64,346
Town	971	0.103	19,831
Fire5–No	2,508	0.502	13,935
Fire5–Yes	3,131	1.596	41,421
NoClaimCredit–No	3,786	1.501	31,365
NoClaimCredit–Yes	1,853	0.31	30,499
Total	5,639	1.109	31,206

alarm level it is 2.008 and 85,140, respectively. This may simply imply that entities with more claims are the ones that are likely to have an alarm system. Summary tables do not examine multivariate effects; for example, Table 1.6 ignores the effect of size (as we measure through coverage amounts) that affect claims.

We will learn more about modeling count data in the Chapter 3 and about severity data in Chapters 4 and 7.

1.3.3 Fund Operations

We have now seen distributions of the Fund's two outcome variables: a count variable for the number of claims, and a continuous variable for the claims amount. We have also introduced a continuous rating variable (logarithmic coverage); a discrete quantitative variable (logarithmic deductibles); two binary rating variables (no claims credit and fire class); and two categorical rating variables (entity type and alarm credit). Subsequent chapters will explain how to analyze and model the distribution of these variables and their relationships. Before getting into these technical details, let us first think about where we want to go. General insurance company functional areas are described in Section 1.2; we now consider how these areas might apply in the context of the property fund.

Initiating Insurance

Because this is a government sponsored fund, we do not have to worry about selecting good or avoiding poor risks; the fund is not allowed to deny a

TABLE 1.7: **Claims Summary by Entity Type and Alarm Credit (AC) Categories 0 and 5**

	AC0 Claim Frequency	AC0 Avg. Severity	AC0 Num. Policies	AC5 Claim Frequency	AC5 Avg. Severity	AC5 Num. Policies
Village	0.326	11,078	829	0.278	8,086	54
City	0.893	7,576	244	2.077	4,150	13
County	2.14	16,013	50	0	0	1
Misc	0.117	15,122	386	0.278	13,064	18
School	0.422	25,523	294	0.41	14,575	122
Town	0.083	25,257	808	0.194	3,937	31
Total	0.318	15,118	2611	0.431	10,762	239

TABLE 1.8: **Claims Summary by Entity Type and Alarm Credit (AC) Categories 10 and 15**

	AC10 Claim Frequency	AC10 Avg. Severity	AC10 Num. Policies	AC15 Claim Frequency	AC15 Avg. Severity	AC15 Num. Policies
Village	0.5	8,792	50	0.725	10,544	408
City	1.258	8,625	31	2.485	20,470	505
County	2.125	11,688	8	5.513	15,476	269
Misc	0.077	3,923	26	0.341	87,021	179
School	0.488	11,597	168	2.008	85,140	1013
Town	0.091	2,338	44	0.261	9,490	88
Total	0.517	10,194	327	2.093	41,458	2462

coverage application from a qualified local government entity. If we do not have to underwrite, what about how much to charge?

We might look at the most recent experience in 2010, where the total fund claims were approximately 28.16 million USD (= 1377 claims × 20452 average severity). Dividing that among 1,110 policyholders, that suggests a rate of 24,370 ($\approx$ 28,160,000/1110). However, 2010 was a bad year; using the same method, our premium would be much lower based on 2009 data. This swing in premiums would defeat the primary purpose of the fund, to allow for a steady charge that local property managers could utilize in their budgets.

Having a single price for all policyholders is nice but hardly seems fair. For example, Table 1.6 suggests that schools have higher aggregate claims than other entities and so should pay more. However, simply doing the calculation on an entity by entity basis is not right either. For example, we saw in Tables 1.7 and 1.8 that had we used this strategy, entities with a 15% alarm credit (for good behavior, having top alarm systems) would actually wind up paying more.

So, we have the data for thinking about the appropriate rates to charge but need to dig deeper into the analysis. We will explore this topic further in Chapter 10 on *premium calculation fundamentals*. Selecting appropriate risks is introduced in Chapter 11 on *risk classification*.

Renewing Insurance

Although property insurance is typically a one-year contract, Table 1.3 suggests that policyholders tend to renew; this is typical of general insurance. For renewing policyholders, in addition to their rating variables we have their claims history and this claims history can be a good predictor of future claims. For example, Table 1.6 shows that policyholders without a claim in the last two years had much lower claim frequencies than those with at least one accident (0.310 compared to 1.501); a lower predicted frequency typically results in a lower premium. This is why it is common for insurers to use variables such as `NoClaimCredit` in their rating. We will explore this topic further in Chapters 12 and 15 on *experience rating*.

Claims Management

Of course, the main story line of the 2010 experience was the large claim of over 12 million USD, nearly half the amount of claims for that year. Are there ways that this could have been prevented or mitigated? Are their ways for the fund to purchase protection against such large unusual events? Another unusual feature of the 2010 experience noted earlier was the very large frequency of claims (239) for one policyholder. Given that there were only 1,377 claims that year, this means that a single policyholder had 17.4 % of the claims. These extreme features of the data suggests opportunities for managing claims, the subject of Chapter 13.

Loss Reserving

In our case study, we look only at the one year outcomes of closed claims (the opposite of open). However, like many lines of insurance, obligations from insured events to buildings such as fire, hail, and the like, are not known immediately and may develop over time. Other lines of business, including those where there are injuries to people, take much longer to develop. Chapter 14 introduces this concern and *loss reserving*, the discipline of determining how much the insurance company should retain to meet its obligations.

1.4 Exercises

These exercises ask you to work with data using statistical software, such as `R` code. If you would like some practice with `R` code, please visit the first chapter of a *Short Course on Loss Data Analytics*. As another method of learning, you can also get practice executing ‘R’ code at our Online Version R Code Site.

Exercise 1.1. Corporate Travel. Universities purchase corporate travel policies to cover employees and students traveling on official university business for a wide variety of accidents and incidents while away from the campus or primary workplace. This broad coverage includes medical care and evacuation, loss of personal property, extraction for political and weather related reasons, and more. These data represent experience from the Australian National University (ANU) and additional details can be found in ANU’s corporate travel policy. You can also learn more about this line of business from ANU’s insurer, Chubb Travel. The data provided are maintained by the insurer, Chubb, and were accessed on 29 July 2022. You can retrieve the data by going to Appendix Section 22.2.

a. Claim Frequency. The travel data history is long and stable. This coverage began on 1 November 2006. Table 1.9 shows the count of claims for years 2015-2019, inclusive. Produce a comparable table of claims frequency for the entire period. Comment on the unusual frequency surrounding the COVID pandemic.

b. Adjust for Zero Claims. From this data set, there are 2107 incurred claims. Of these claims, there are 269 zeros and an additional 3 claims where the

TABLE 1.9: **2015-2019 Travel Claims Frequency**

2015	2016	2017	2018	2019
158	154	139	205	274

incurred claim is less than 10. We omit these claims in our analysis. Reproduce your part (a) analysis by omitting incurred claims less than 10.

c. Loss Distributions over Time. There are 1835 incurred losses in the dataset with all available years (yet omitting claims less than 10). Figure 1.3 shows that the distribution of incurred losses is stable over the period 2015-2019, inclusive. Produce a comparable figure for the entire period.

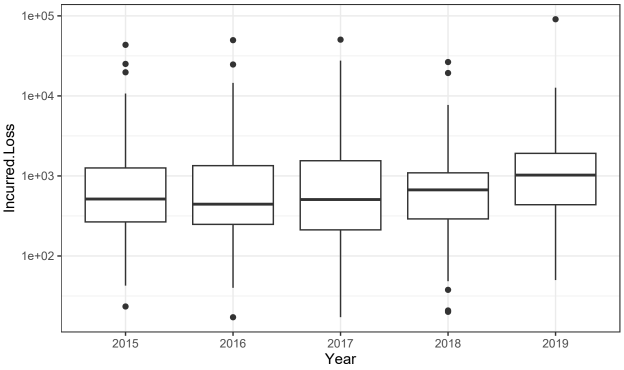


FIGURE 1.3: **Distribution of Travel Losses by Year**

d. Summary Statistics. In addition to graphs, it can be helpful to display several summary statistics. For the five year period 2015-2019, produce a set of summary statistics.

Exercise 1.2. Group Personal Accident. Group personal accident insurance offers financial protection in case of injury or death resulting from an incident that occurs on the job. Group personal accident offers insurance coverage and liability insurance protection against accidental death or injury. The insurance covers students and ANU's voluntary workers; ANU workers are covered through another system known as "workers' compensation."

Several limits apply including 1,000,000 for the period of insurance, 600,000 for non-scheduled flights, and others. These limits were not reached in the data we consider. For this coverage, there is a "7 day excess" for weekly benefits but none for general benefits. The database documentation provided to us, and the data we provide, do not indicate whether the excess has been triggered;

TABLE 1.10: **2015-2019 Group Personal Accident Claims Frequency**

2015	2016	2017	2018	2019
4	7	16	11	9

we have only paid claims. Because of the relatively small size of this class of insurance, we ignore the effects of deductibles for this line.

The data provided to us are maintained by the insurer, Chubb. These data began in underwriting year 2007 and were accessed on 29 July 2022. You can retrieve the data by going to Appendix Section 22.3.

a. Claim Frequency. From this data set, there are 148 incurred claims. Of these claims, there are 35 zeros and an additional 0 claims where the incurred claim is less than 10. We omit these claims in our analysis. Table 1.10 shows the count of claims for years 2015-2019, inclusive. Produce a comparable table of claims frequency for the entire period, omitting claims that are less than 10.

b. Skewness of Claims Severity Distribution. The left-hand panel of Figure 1.4 shows a histogram of incurred claims that reveals the right-skewed nature of this distribution. The right-hand panel shows the same claims but on the log (base 10) scale; this plot demonstrates that the log transform can symmetrize a distribution. These plots are for the 2015-2019 data. Replicate this work, using incurred claims for all available years (still omitting those less than 10).

c. Summary Statistics. Produce summary statistics for both claims and log claims using all available years (still omitting those less than 10). Comment on the relationship between the mean and the median for both claims and log claims, relating this to the symmetry of the distributions observed in part (b).

d. Loss Distributions over Time. There are 112 incurred losses. Figure 1.5 indicates that the incurred losses are stable over the period 2015-2019, inclusive. Produce a comparable figure for the entire period and comment on the stability of the distribution.

Exercise 1.3. Motor Vehicle. This policy covers ANU's vehicles including cars, vans, utilities, and motorcycles. There are two parts to this coverage, one for comprehensive damage to the insured vehicles and a second for legal liability. The comprehensive coverage for loss or damage is essentially limited by the market value of the insured vehicle. For legal liability, there is a $50 Million upper limit for all claims arising from the one accident or series of accidents resulting from the one original cause. There is also another upper

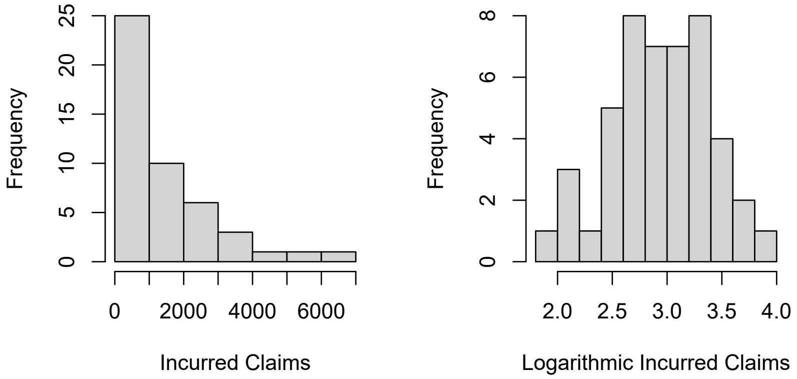


FIGURE 1.4: **Distribution of Incurred Claims 2015-2019**

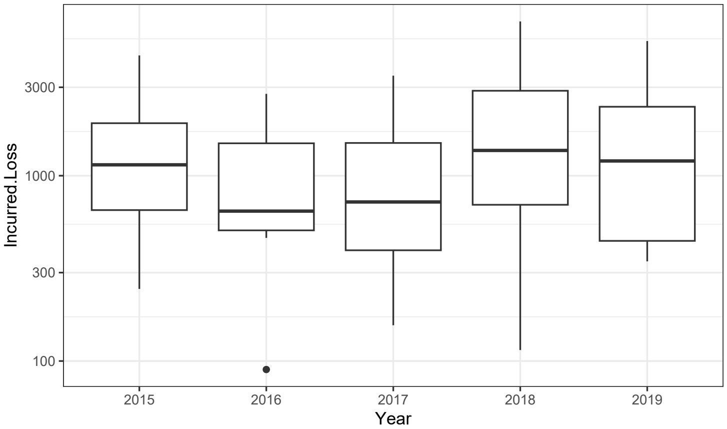


FIGURE 1.5: **Distribution of Group Personal Accident Losses by Year**

limit (that is lower than 50 million) when the vehicle is used for transportation of dangerous goods.

The data available contain the amount paid by the insurer (Vero Insurance Limited) which is the focus of our initial analysis. In addition, the data also contains a deductible (called an “excess” in the data file) that we explore in later parts.

The data provided to us are maintained by the insurer, Vero Insurance Limited. These data began in underwriting year 2012 and were accessed on 8 August 2022. You can retrieve the data by going to Appendix Section 22.4.

a. Adjust for Zeros. From this data set, check that:

- there are 318 incurred claims.
- Of these claims, there are 50 zeros and
- an additional 0 claims where the incurred claim is less than 10.

Remove these claims in your analysis, so that there are 268 incurred claims.

b. Claim Frequency. Produce a table that shows the count of claims for the entire period.

c. Loss Distributions over Time. Produce a figure that shows the distribution of motor vehicle paid amounts over time and comment on the stability of the distribution.

d. Year 2019. In your analysis from the prior steps, you may have noticed the unusual aspects of year 2019. In that year, ANU suffered extensive damage from a hailstorm that increased the frequency of claims as well as the severity. Produce a histogram of paid claims for that year.

e. Deductibles. For each event, or series of events arising from the one originating cause, ANU bears the amount of the excess in respect of each and every insured vehicle, unless stated otherwise. The standard deductible (or excess) in the dataset is 1000. However, a cursory examination of the dataset shows tremendous variation by vehicle and over time. Replicate Table 1.11 that shows, for each year, the number of claims with zero excess, positive excess less than 1000, an excess equal to 1000, and an excess greater than 1000.

(**Deductibles**. We recommend that motivated readers extend our analysis to account for this deductible in both the severity and frequency.)

TABLE 1.11: **Motor Vehicle Excess by Year**

UW.Year	Num 0	Num 0-1000	Num = 1000	Num >1000	Total
2011	1	1	7	0	9
2012	1	2	13	0	16
2013	4	1	22	0	27
2014	0	0	11	0	11
2015	1	1	14	0	16
2016	6	1	19	0	26
2017	16	0	4	1	21
2018	19	0	1	0	20
2019	99	0	6	0	105
2020	5	0	0	0	5
2021	10	0	0	0	10

1.5 Further Resources and Contributors

If you would like additional practice with R coding, please visit our companion LDA Short Course. In particular, see the Introduction to Loss Data Analytics Chapter.

Contributor

- **Edward (Jed) Frees**, University of Wisconsin-Madison and Australian National University, is the principal author of the initial version and second edition of this chapter. Email: jfrees@bus.wisc.edu for chapter comments and suggested improvements.
- Chapter reviewers include: Yair Babad, Chunsheng Ban, Aaron Bruhn, Gordon Enderle, Hirokazu (Iwahiro) Iwasawa, Dalia Khalil, Bell Ouelega, Michelle Xia.

This book introduces loss data analytic tools that are most relevant to actuaries and other financial risk analysts. We have also introduced you to many new insurance terms; more terms can be found at the NAIC Glossary (2018).

2

Introduction to Data Analytics

Chapter Preview. This introduction focuses on data analytics concepts relevant to insurance activities. As data analytics is used across various fields with different terminologies, we start in Section 2.1 by describing the basic ingredients or elements of data analytics. Then, Section 2.2 outlines a process an analyst can use to analyze insurance data. Many fields emphasize the development of data analytics with a focus on multiple variables, or "big" data. However, this often comes at the cost of excluding consideration of a single variable. So, Section 2.3 introduces an approach we call "single variable analytics," which includes a description of variable types, exploratory versus confirmatory analysis, and elements of model construction and selection, all of which can be done in the context of a single variable. Building on this, Section 2.4 explores the roles of supervised and unsupervised learning, which require the presence of many variables.

The final section of this chapter, Section 2.5, offers a broader introduction to data considerations beyond the scope of this book, intended for budding analysts who want to use this chapter to build a foundation for further studies in data analytics. Additionally, the technical supplements introduce other standard ingredients of data analytics, such as principal components, cluster analysis, and tree-based regression models. While these topics are not necessary for this book, they are important in a broader analytics context.

2.1 Elements of Data Analytics

In this section, you learn how to describe the essential ingredients of data analytics

- consisting of several key concepts, and
- two fundamental approaches, data and algorithmic modeling.

Data analysis involves inspecting, cleansing, transforming, and modeling data to discover useful information to suggest conclusions and make decisions. Data analysis has a long history. In 1962, statistician John Tukey defined data analysis as:

> procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.
>
> — (Tukey, 1962)

2.1.1 Key Data Analytic Concepts

Underpinning the elements of data analytics are the following key concepts:

- **Data Driven**. As described in Section 1.1.2, the conclusions and decisions made through a data analytic process depend heavily on data inputs. In comparison, econometricians have long recognized the difference between a data-driven model and a structural model, the latter being one that represents an explicit interplay between economic theory and stochastic models, Goldberger (1972) .
- **EDA** - exploratory data analysis - and **CDA** - confirmatory data analysis. Although some techniques overlap, e.g., taking the average of a dataset, these two approaches to analyzing data have different purposes. The purpose of EDA is to reveal aspects or patterns in the data without reference to any particular model. In contrast, CDA techniques use data to substantiate, or confirm, aspects or patterns in a model. See Section 2.3.2 for further discussions.
- **Estimation** and **Prediction**. Recall the traditional triad of statistical inference: hypothesis testing, parameter estimation, and prediction. Medical statisticians test the efficacy of a new drug and econometricians estimate parameters of an economic relationship. In insurance, one also uses hypothesis testing and parameter estimation. Moreover, predictions of yet to be realized random outcomes are critical for financial risk management (e.g., pricing) of existing risks in future periods, as well as not yet observed risks in a current period, cf. Frees (2015).
- **Model Complexity, Parsimony,** and **Interpretability**. A model is a mathematical representation of reality that, in statistics, is calibrated using a data set. One concern is the *complexity* of the model where the complexity may involve the number of parameters used to define the model, the number of variables upon which it relies, and the intricacies of relationships among the parameters and variables. As a rule of thumb, we will see that the more

complex is the model, the better it fares in fitting a set of data (and hence at estimation) but the worse it fares in predicting new outcomes. Other things being equal, a model with fewer parameters is said to be *parsimonious* and hence less complex. Moreover, a parsimonious model is typically easier to interpret than a comparable model that is more complex. Complexity hinders our ability to understand the inner workings of a model, its interpretability, and will be a key ingredient in our comparisons of data versus algorithmic models in Section 2.1.2.

- **Parametric** and **Nonparametric** models. Many models, including stochastic distributions, are known with the exception of a limited number of quantities known as parameters. For example, the mean and variance are parameters that determine a normal distribution. In contrast, other models may not rely on parameters; these are simply known as *nonparametric* models. Naturally, there is also a host of models that rely on parameters for some parts of the distribution and are distribution-free for other portions; these are referred to as *semi-parametric* models. Parametric and nonparametric approaches have different strengths and limitations; neither is strictly better than the other. We start the discussion in Section 2.3.3 to explain under what circumstances you might prefer one approach to another.
- **Robustness** means that a model, test, or procedure is resistant to unanticipated deviations in model assumptions or the data used to calibrate the model. When interpreting findings, it is natural to ask questions about how the results react to changes in assumptions or data, that is, the robustness of the results.
- **Computational Statistics**. Historically, statistical modeling relied extensively on summary statistics that were not only easy to interpret but also easy to compute. With modern-day computing power, definitions of "easy to compute" have altered drastically paving the way for measures that were once deemed far too computationally intensive to be of practical use. Moreover, ideas of subsampling and resampling data (e.g., through cross-validation and bootstrapping) have introduced new methods for understanding statistical sampling errors and a model's predictive capabilities.
- **Big Data**. This is about the process of using special methods and tools that can extract information rapidly from massive data. Examples of big data include text documents, videos, and audio files that are also known as *unstructured* data. Table 2.1 summarizes new types of data sources that lead to new data. As part of the analytics trends, different types of algorithms lead to new software for handling new types of data. See Section 2.5.4 for further discussions.

Table 2.1. **Analytic Trends** (from Frees and Gao (2019))

Data Sources	**Algorithms**
Mobile devices	Statistical learning
Auto telematics	Artificial intelligence
Home sensors (Internet of Things)	Structural models
Drones, micro satellites	
Data	**Software**
Big data (text, speech, image, video)	Text analysis, semantics
Behavioral data (including social media)	Voice recognition
Credit, trading, financial data	Image recognition
	Video recognition
Source : Stephen Mildenhall, Personal Communication	

2.1.2 Data versus Algorithmic Modeling

There are two cultures for the use of statistical modeling to reach conclusions from data: the data modeling culture and the algorithmic modeling culture. In the data modeling culture, data are assumed to be generated by a given stochastic model. In the algorithmic modeling culture, the data mechanism is treated as unknown and algorithmic models are used.

Data modeling allows statisticians to analyze data and acquire information about the data mechanisms. However, Breiman (2001) argued that the focus on data modeling in the statistical community has led to some side effects such as:

- It produced irrelevant theory and questionable scientific conclusions.
- It kept statisticians from using algorithmic models that might be more suitable.
- It restricted the ability of statisticians to deal with a wide range of problems.

Algorithmic modeling was used by industrial statisticians long time ago. Sadly, the development of algorithmic methods was taken up by communities outside statistics. The goal of algorithmic modeling is predictive accuracy. For some complex prediction problems, data models are not suitable. These prediction problems include voice recognition, image recognition, handwriting recognition, nonlinear time series prediction, and financial market prediction. The theory in algorithmic modeling focuses on the properties of algorithms, such as convergence and predictive accuracy.

2.2 Data Analysis Process

In this section, you learn how to describe the data analysis process as five steps:

- scoping phase,
- data splitting,
- model development,
- validation, and
- determining implications.

Table 2.2 outlines common steps used when analyzing data associated with insurance activities.

Table 2.2 **Data Analysis Process for Insurance Activities**

I. Scoping Phase	**II. Data Splitting**	**III. Model Development**	**IV. Validation**	**V. Determine Implications**
Use background knowledge and theory to define goals	Split the data into training and testing portions	Select a candidate model	Repeat Phase III to determine several candidate models	Use knowledge gained from exploring the data, fitting and predicting the models to make data-informed statements about the project goals
Prepare, collect, and revise data		Select variables to be used with the candidate model	Assess each model using the testing portion of the data to determine its predictive capabilities	
EDA Explore the data		Evaluate model fit using training data		
		Use deviations from model fit to improve suggested models		

I. Scoping Phase

Scoping, or problem formulation, can be divided into three components:

- **Use background knowledge and theory to define goals**. Insurance activity projects are commonly motivated by business pursuits that have been formulated to be consistent with background knowledge such as market conditions and theory such as a person's attitude towards risk-taking.

- **Prepare, collect, and revise data**. Getting the right data that gives insights into questions at hand is typically the most time-consuming aspect of most projects. Section 2.5 delves more into the devilish details of data structures, quality, cleaning, and so forth.
- **EDA** - Exploring the data, without reference to any particular model, can reveal unsuspected aspects or patterns in the data.

These three components can be performed *iteratively*. For example, a question may suggest collecting certain data types. Then, a preliminary analysis of the data raises additional questions of interest that can lead to seeking more data - this cycle can be repeated many times. Note that this iterative approach differs from the traditional "scientific method" whereby the analyst develops a hypothesis, collects data, and then employs the data to test the hypothesis.

II. Data Splitting

Although optional, splitting the data into training and testing portions has some important advantages. If the available dataset is sufficiently large, one can split the data into a portion used to calibrate one or more candidate models, the training portion, and another portion that can be used for testing, that is, evaluating the predictive capabilities of the model. The data splitting procedure guards against overfitting a model and emphasizes predictive aspects of a model. For many applications, the splitting is done randomly to mitigate unanticipated sources of bias. For some applications such as insurance, it is common to use data from an earlier time period to predict, or *forecast*, future behavior. For example, with the Section 1.3 Wisconsin Property Fund data, one might use 2006-2010 data for training and 2011 data for assessing predictions.

For large datasets, some analysts prefer to split the data into three portions, one for training (model estimation), one for validation (estimate prediction error for model selection), and one for testing (assessment of the generalization error of the final chosen model), c.f. Hastie et al. (2009) (Chapter 7). In contrast, for moderate and smaller datasets, it is common to use cross-validation techniques where one repeatedly splits the dataset into training and testing portions and then averages results over many applications. These techniques are described further in Chapter 8.

III. Model Development

The objective of the model development phase is to consider different types of model and provide the best fit for each "candidate" model. As with the scoping phase, developing a model is an iterative procedure.

- **Select a candidate model.** One starts with a model that, from the analyst's perspective, is a likely "candidate" to be the recommended model. Although analysts will focus on familiar models, such as through their past applications of a model or its acceptance in industry, in principle one remains open to all types of models.

- **Select variables to be used with the candidate model.** For simpler situations, only a single outcome, or variable, is of interest. However, many (if not most) situations deal with multivariate outcomes and, as will be seen in Section 2.4, analysts give a great deal of thought as to which variables are considered inputs to a system and which variables can be treated as outcomes.
- **Evaluate model fit on training data.** Given a candidate model based on one or more selected variables, the next step is to calibrate the model based on the training data and evaluate the model fit. Many measures of model fit are available - analysts should focus on those likely to be consistent with the project goals and intended audience of the data analysis process.
- **Use deviations from the model fit to suggest improvements to the candidate model.** When comparing the training data to model fits, it may be that certain patterns are revealed that suggest model improvements. In regression analysis, this tactic is known as *diagnostic checking.*

IV. Validation

- **Repeat Phase III to determine several candidate models.** There is a wealth of potential models from which an analyst can choose. Some are parametric, others non-parametric, and some a mixture between the two. Some focus on simplicity such as through linear relationships whereas others are much more complex. And so on. Through repeated applications of the Phase III process, it is customary to narrow the field of candidates down to a handful based on their fit to the training data.
- **Assess each model using the testing portion of the data to determine its predictive capabilities.** With the handful of models that perform the best in the model development phase, one assesses the predictive capabilities of each model. Specifically, each fitted model is used to make predictions with the predicted outcomes compared to the held-out test data. This comparison may also be done using cross-validation. Models are then compared based on their predictive capabilities.

V. Determine Implications

The scoping, model development, and validation phases all contribute to making data-informed statements about the project goals. Although most projects result in a single recommended model, each phase has the potential to lend powerful insights.

For data analytic projects associated with insurance activities, it is common to select the model with best predictive capabilities. However, analysts are also mindful of the intended audiences of their analyses, and it is also common to favor models that are simpler and easier to interpret. The relative importance of interpretability very much depends on the project goals. For example, a model devoted to enticing potential customers to view a webpage can be judged more on its predictive capabilities. In contrast, a model that provides the foundations for insurance prices typically undergoes scrutiny by regulators and consumer advocacy groups; here, interpretation plays an important role.

2.3 Single Variable Analytics

In this section, you learn how to describe analytics based on a single variable in terms of

- the type of variable,
- exploratory versus confirmatory analyses,
- model construction and
- model selection.

Rather than starting with multiple variables consisting of inputs and outputs as is common in analytics, in this section we restrict considerations to a single variable. Single variable analytics is motivated by statistical data modeling. Moreover, as will be seen in Chapters 3-8, single variable analytics plays a prominent role in fundamental insurance and risk management applications.

2.3.1 Variable Types

This section describes basic variable types traditionally encountered in statistical data analysis. Section 2.5 will provide a framework for more extensive types that include big data.

Qualitative Variables

A qualitative, or categorical variable is one for which the measurement denotes membership in a set of groups, or categories. For example, if you were coding in which area of the country an insured resides, you might use 1 for the northern part, 2 for southern, and 3 for everything else. Any analysis of categorical variables should not depend on the labeling of the categories. For example, instead of using a 1,2,3 for north, south, other, one should arrive at the same set of summary statistics if I used a 2,1,3 coding instead, interchanging north and south.

In contrast, an ordinal variable is a variation of categorical variable for which an ordering exists. For example, with a survey to see how satisfied customers are with our claims servicing department, we might use a five point scale that ranges from 1 meaning dissatisfied to a 5 meaning satisfied. Ordinal variables provide a clear ordering of levels of a variable although the amount of separation between levels is unknown.

A binary variable is a special type of categorical variable where there are only two categories commonly taken to be 0 and 1.

Earlier, in the Section 1.3 case study, we saw in Table 1.5 several examples of qualitative variables. These included the categorical `EntityType` and binary variables `NoClaimCredit` and `Fire5`. We also treated `AlarmCredit` as a categorical variable although some analysts may wish to explore its use as an ordinal variable.

Quantitative Variables

Unlike a qualitative variable, a quantitative variable is one in which each numerical level is a realization from some scale so that the distance between any two levels of the scale takes on meaning. A continuous variable is one that can take on any value within a finite interval. For example, one could represent a policyholder's age, weight, or income, as continuous variables. In contrast, a discrete variable is one that takes on only a finite number of values in any finite interval. For example, when examining a policyholder's choice of deductibles, it may be that values of 0, 250, 500, and 1000 are the only possible outcomes. Like an ordinal variable, these represent distinct categories that are ordered. Unlike an ordinal variable, the numerical difference between levels takes on economic meaning. A special type of discrete variable is a count variable, one with values on the nonnegative integers. For example, we will be particularly interested in the number of claims arising from a policy during a given period. Another interesting variation is an interval variable, one that gives a range of possible outcomes.

Earlier, in the Section 1.3 case study, we encountered several examples of quantitative variables. These included the deductible (in logarithmic dollars), total building and content coverage (in logarithmic dollars), claim severity and claim frequency.

Loss Data

This introduction to data analytics is motivated by features of **loss data** that arise from, or are related to, obligations in insurance contracts. Loss data rarely arise from a bell-shaped normal distribution that has motivated the development of much of classical statistics. As a consequence, the treatment of data analytics in this text differs from that typically encountered in other introductions to data analytics.

What features of loss data warrant special treatment?

- We have already seen in the Section 1.3 case study that we will be concerned with the frequency of losses, a type of count variable.
- Further, when a loss occurs, the interest is in the amount of the claim, a quantitative variable. This claim severity is commonly modeled using skewed and long-tailed distributions so that extremely large outcomes are associated with relatively large probabilities. Typically, the normal distribution is a poor choice for a loss distribution.
- When a loss does occur, often the analyst only observes a value that is modified by insurance contractual features such as deductibles, upper limits, and co-insurance parameters.
- Loss data are frequently a *combination of discrete and continuous* components. For example, when we analyze the insured loss of a policyholder, we will encounter a discrete outcome at zero, representing no insured loss, and a continuous amount for positive outcomes, representing the amount of the insured loss.

2.3.2 Exploratory versus Confirmatory

There are two phases of data analysis: exploratory data analysis (EDA) and confirmatory data analysis (CDA). Table 2.3 summarizes some differences between EDA and CDA. EDA is usually applied to observational data with the goal of looking for patterns and formulating hypotheses. In contrast, CDA is often applied to experimental data (i.e., data obtained by means of a formal design of experiments) with the goal of quantifying the extent to which discrepancies between the model and the data could be expected to occur by chance.

Table 2.3. **Comparison of Exploratory Data Analysis and Confirmatory Data Analysis**

	EDA	**CDA**
Data	Observational data	Experimental data
Goal	Pattern recognition, formulate hypotheses	Hypothesis testing, estimation, prediction
Techniques	Descriptive statistics, visualization, clustering	Traditional statistical tools of inference, significance, and confidence

As we have seen in the Section 1.3 case study, the techniques for single variable EDA include descriptive statistics (e.g., mean, median, standard deviation, quantiles) and summaries of distributions such as through histograms. In contrast, the techniques for CDA include the traditional statistical tools of inference, significance, and confidence.

2.3.3 Model Construction

As we learned in Section 2.1.2, models may have a stochastic basis from the statistical modeling paradigm or may simply be the result of an algorithm. When constructing a model, it is helpful to think about how it is parameterized and to identify the purpose of constructing the model.

Parametric versus Nonparametric

Data analysis models can be parametric or nonparametric. Parametric models are representations that are known up to a few terms known as *parameters*. These may be representations of a stochastic distribution or simply an algorithm used to predict data outcomes. Typically, data are used to determine the parameters and in this way calibrate the model. In contrast, nonparametric methods make no such assumption of a known functional form. For example, Section 4.4.1 will introduce nonparametric methods that do not assume distributions for the data and therefore are also called *distribution-free* methods.

Because a functional form is known with a parametric model, this approach works well when data size is relatively limited. This reasoning extends to the situation when one is considering many variables simultaneously so that the so-called "curse of dimensionality" effectively limits the sample size. For example if you are trying to determine the expected cost of automobile losses, you are likely to consider a driver's age, gender, driving location, type of vehicle, and dozens of other variables. Approaches that use some parametric relationships among these variables are common because a purely non-parametric approach would require data sets too large to be useful in practice.

Nonparametric methods are very valuable particularly at the exploratory stages of an analysis where one tries to understand the distribution of each variable. Because nonparametric methods make fewer assumptions, they can be more flexible, more robust, and more applicable to non-quantitative data. However, a drawback of nonparametric methods is that it is more difficult to extrapolate findings outside of the observed domain of the data, a key consideration in *predictive modeling*.

Explanation versus Prediction

There are two goals in data analysis: explanation and prediction. In some scientific areas such as economics, psychology, and environmental science, the focus of data analysis is to explain the causal relationships between the input variables and the response variable. In other scientific areas such as natural language processing, bioinformatics, and actuarial science, the focus of data analysis is to predict what the responses are going to be given the input variables.

Shmueli (2010) discussed in detail the distinction between explanatory modeling and predictive modeling. Explanatory modeling is commonly used for theory building and testing and is typically done as follows:

- State the prevailing theory.
- State causal hypotheses, which are given in terms of theoretical constructs rather than measurable variables. A causal diagram is usually included to illustrate the hypothesized causal relationship between the theoretical constructs.
- Operationalize constructs. In this step, previous literature and theoretical justification are used to build a bridge between theoretical constructs and observable measurements.
- Collect data and build models alongside the statistical hypotheses, which are operationalized from the research hypotheses.
- Reach research conclusions and recommend policy. The statistical conclusions are converted into research conclusions or policy recommendations.

In contrast, predictive modeling is the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations. Predictions include point predictions, interval predictions, regions, distributions, and rankings of new observations. A predictive model can be any method that produces predictions.

2.3.4 Model Selection

Although hypothesis testing is one approach to model selection that is viable in many fields, it does have its drawbacks. For example, the asymmetry between the null and alternative hypotheses raises issues; hypothesis testing is biased towards a null hypothesis unless there is strong evidence to the contrary.

For modeling insurance activities, it is typically preferable to estimate the predictive power of various models and select a model with the best predictive power. The motivation for this is that we want good model selection methods achieve a balance between goodness of fit and parsimony. This is a trade-off because on the one hand, better fits to the data can be achieved by adding more parameters, making the model more complex and less parsimonious. On the other hand, models with fewer parameters (parsimonious) are attractive because of their simplicity and interpretability; they are also less subject to estimation variability and so can yield more accurate predictions, Ruppert et al. (2003).

One way of measuring this balance is through information criteria such as Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These measures each contain a component that summarizes how well the model fits the data, a goodness of fit piece, plus a component to penalize the complexity of the model.

Although attractive due to their simplicity, there are drawbacks to these measures. In particular, both rely on knowledge of the underlying distribution of the outcomes (or at least good estimates). A more robust approach is to split a data set in a portion that can used to calibrate a model, the *training* portion, and another portion used to quantify the predictive power of the model, the *test* portion. It is more robust in the sense that it does not rely on any distributional assumptions and can be used to validate general models.

The data splitting approach is attractive because it directly aligns with the concept of assessing predictive power and can be used in general, and complex, situations. However, it does introduce additional variability into the process by introducing extra randomness of the uncertainty of which observations fall into the training and testing portions. To mitigate this problem, it is common to use an approach known as *cross-validation.* To illustrate, suppose that one randomly partitions a dataset into five subsets of roughly equivalent sizes

Train	Test	Train	Train	Train

Then, based on the first, third, fourth, and fifth subsets, estimate a model, use this fitted model to predict outcomes in the second, and compare the predictions to the held-out values in the test portion. Repeat this process by selecting each subset as the test portion, with the others being used for training, and take an average over the comparison which results in a cross-validation statistic. Cross-validation is used widely in modeling insurance activities and is described in more detail in Chapter 5.

Example 2.3.1. Under- and Over-Fitting. Suppose that we have a set of claims that potentially varies by a single categorical variable with six levels. For example, in the Section 1.3 case study there are six entity types. If each level is truly distinct, then in classical statistics one uses the level average to make predictions for future claims. Another option is to ignore information in the categorical variable and use the overall average to make predictions; this is known as a "community-rating" approach.

For illustrative purposes, we assume that two of the six levels are the same and are different from the others. For example, the Table 1.6 summary statistics suggest that Schools and the Miscellaneous levels can be viewed similarly yet warrant a higher predicted claims amount than the other four levels. For illustrative purposes, we generated 100 claims that follow this pattern (using simulation techniques that will be described in Chapter 8).

Results are summarized in Table 2.4 for three fitted models. These are the "Community Rating" corresponding to using the overall mean, the "Two Levels" corresponding to using two averages, and the "Six Levels" corresponding to using an average for each level of the categorical variable. The data set of size 100 was randomly split into five folds; for each fold, the other folds were used to train/estimate the model and then that fold was used to assess predictions. The first five rows of Table 2.4 give the results of the root mean square error for each fold. The sixth row provides the average over the five folds and the last row gives a similar result for another goodness of fit statistic, the AIC. This approach is known as "cross-validation" that will be described in greater detail in Chapters 6 and 8.

Table 2.4 shows that in each case the "Two Level" model has the lowest root mean square error and AIC, indicating that it is the preferred model. The overfit model with six levels came in second and the underfit model, community rating, was a distant third. This analysis demonstrates techniques for selecting the appropriate model. Unlike analysis of real data, in this demonstration we enjoyed the additional luxury of knowing that we got things correct because we in fact generated the data - an approach that analysts often use to develop analytic procedures prior to utilizing the procedures on real data.

TABLE 2.4: **Under- and Over-Fitting of Models**

	Community Rating	Two Levels	Six Levels
Rmse - Fold 1	1.318	1.192	1.239
Rmse - Fold 2	1.034	0.972	1.023
Rmse - Fold 3	0.816	0.660	0.759
Rmse - Fold 4	0.807	0.796	0.824
Rmse - Fold 5	0.886	0.539	0.671
Rmse - Average	0.972	0.832	0.903
AIC - Average	227.171	206.769	211.333

2.4 Analytics with Many Variables

In this section, you learn how to describe analytics based on many variables in terms of

- supervised and unsupervised learning,
- types of algorithmic models, including linear, ridge, and lasso regressions, as well as regularization, and
- types of data models, including Poisson regressions and generalized linear models.

Just as with a single variable in Section 2.3, with many variables analysts follow the same structure of identifying variables, exploring data, constructing and selecting models. However, the potential applications become much richer when considering many variables. With many potential applications, it is natural that techniques for data analysis have developed in different but overlapping fields; these fields include statistics, machine learning, pattern recognition, and data mining.

- Statistics is a field that addresses reliable ways of gathering data and making inferences.
- The term machine learning was coined by Samuel in 1959 (Samuel, 1959). Originally, machine learning referred to the field of study where computers have the ability to learn without being explicitly programmed. Nowadays, machine learning has evolved to a broad field of study where computational methods use experience (i.e., the past information available for analysis) to improve performance or to make accurate predictions.

- Originating in engineering, pattern recognition is a field that is closely related to machine learning, which grew out of computer science. In fact, pattern recognition and machine learning can be considered to be two facets of the same field (Bishop, 2007).
- Data mining is a field that concerns collecting, cleaning, processing, analyzing, and gaining useful insights from data (Aggarwal, 2015).

2.4.1 Supervised and Unsupervised Learning

With multiple variables, the essential tasks of identifying variable types, exploring data, and selecting models are similar in principle to that described for single variables in Section 2.3. When exploring data in multiple dimensions, additional considerations such as clustering like observations and reducing the dimension arise. As these considerations will not arise in the applications in this book, we provide only a brief introduction in Technical Supplement Section 2.6.1.

The construction of models differs dramatically when comparing single to multiple variable modeling. With many variables, we have the opportunity to think about some of them as “inputs” and others “outputs” of a system. Models based on input and output variables are known as supervised learning methods or as regression methods. Table 2.5 gives a list of common names for different types of variables (Frees, 2009). When the target variable is a categorical variable, supervised learning methods are called classification methods.

Table 2.5. **Common Names of Different Variables**

Target Variable	**Explanatory Variable**
Dependent variable	Independent variable
Response	Treatment
Output	Input
Endogenous variable	Exogenous variable
Predicted variable	Predictor variable
Regressand	Regressor

Methods for data analysis can be divided into two types (Abbott, 2014; James et al., 2013): supervised learning methods and unsupervised learning methods. Unsupervised learning methods work where our data are treated the same and there is no artificial divide between “inputs” and “outputs.” As a result, unsupervised learning methods are particularly useful at the exploratory stage of an analysis.

2.4.2 Algorithmic Modeling

Early data analysis traced the movements of orbits of bodies about the sun using astronomical observations in the 1750's by Boscovich and was continued in the early 1800's by Legendre and Gauss (the latter two in connection with their development of least squares). This work was done using algorithmic *fitting* approaches (such as least squares) without regard to distributions of random variables.

The idea underpinning algorithmic fitting is easy to interpret. One variable, Y, is determined to be a target variable. Other variables, $X_1, X_2, \ldots, X_p$, are used to understand or explain the target Y. The goal is to determine an appropriate function $f(\cdot)$ so that $f(X_1, X_2, \ldots, X_k)$ is a useful predictor of Y.

Linear Regression. To illustrate, consider the classic linear regression context. In this case, we have n observations of a target and explanatory variables, with the ith observation denoted as $(x_{i1}, \ldots, x_{ik}, y_i) = (\mathbf{x}_i, y_i)$. One would like to determine a single function f so that $f(\mathbf{x}_i)$ is a reasonable approximation for y_i, for each i. For the linear regression, one restricts considerations to functions of the form

$$f(x_{i1}, \ldots, x_{ik}) = \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = \mathbf{x}_i' \boldsymbol{\beta}.$$

Here, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ is a vector of *regression coefficients*. This function is *linear* in the explanatory variables that gives rise to the name linear regression.

The *ordinary least squares (OLS)* estimates are the solution of the following minimization problem,

$$\text{minimize}_{\boldsymbol{\beta}} \quad \tfrac{1}{n} \textstyle\sum_{i=1}^{n} (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2.$$

The *OLS* estimates are historically prominent in part because of their ease of computation and interpretation. Naturally, a squared difference such as $(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$ is not the only way to measure the deviation between a target y_i and an estimate $\mathbf{x}_i'\boldsymbol{\beta}$. In general, analysts use the term *loss function* $l(y_i, \mathbf{x}_i'\boldsymbol{\beta})$ to measure this deviation; as an alternative, it is not uncommon to use an absolute deviation.

Algorithmic Modeling Culture. As introduced in Section 2.1.2, a culture has developed across widespread communities that emphasizes algorithmic fitting particularly in complex problems such as voice, image, and handwriting recognition. Algorithmic methods are especially useful when the goal is prediction, as noted in Section 2.3.3. Many of these algorithms take an approach similar to linear regression. As examples, other widely used algorithmic fitting methods include ridge and lasso regression, as well as regularization methods.

Ridge Regression. One limitation of *OLS* is that it tends to overfit, particularly when the number of regression coefficients k becomes large. In fact, with $k = n$ one gets an exact match between the targets y_i and the predictor function. A modification introduced in 1970 by Hoerl and Kennard (1970) is known as *ridge regression* where one determines regression coefficients $\boldsymbol{\beta}$ as in equation (2.4.2) although subject to the constraint that $\sum_{j=1}^{p} |\beta_j|^2 \leq c_{ridge}$, where c_{ridge} is an appropriately chosen constant. Naturally, if c_{ridge} is very large, then the constraint has no effect and the ridge estimates equal the *OLS* solution. However, as c_{ridge} becomes small, it reduces the size of the regression coefficients. In this sense, the ridge regression estimator is said to be "shrunk towards zero."

Adding the constraint on the size of the coefficients can mean smaller and more stable coefficients when compared to *OLS*. As such, ridge regression is particularly useful when dealing with high-dimensional datasets, where the number of predictors is very large compared to the number of observations. In the actuarial applications, we might have a portfolio of only a few thousand risks that we wish to model. With ridge regression, we can utilize millions of variables as potential inputs to develop predictive models.

Lasso Regression. Similar to ridge regression, one can determine regression coefficients $\boldsymbol{\beta}$ as in equation (2.4.2) although subject to the constraint that $\sum_{j=1}^{p} |\beta_j| \leq c_{lasso}$, where c_{lasso} is an appropriately chosen constant. This procedure is known as *lasso regression.* Here, one uses absolute values in the constraint function (although still squared errors for the loss function).

The lasso overcomes an important limitation of ridge regression. With ridge regression, we might reduce the size of the constant c_{ridge} that forces the regression coefficient to become small but does not ensure that they become zero. In contrast, the lasso ensures that trivial regression coefficients become zero. In the linear regression approximation, a zero regression coefficient means that the variable drops from the function approximation, thus reducing model complexity.

Regularization. Both the ridge and lasso regressions are constrained minimization problems. It is not too hard to show that they can be written as

$$\text{minimize}_{\boldsymbol{\beta}} \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + LM \sum_{j=1}^{p} |\beta_j|^s \right),$$

where $s = 2$ is for ridge regression and $s = 1$ is for lasso regression. We can interpret the first part inside the minimization operation as the goodness of fit and the second part as a penalty for size of the regression coefficients. As we have discussed, reducing the coefficients can mean reducing modeling complexity. In this sense, this expression demonstrates a balance between goodness of fit and model complexity, controlled by the parameter LM (In this case, because it is a constrained optimization problem, the parameter is a Lagrange multiplier.). The choice of $LM = 0$ reduces to the *OLS* estimator that focuses on goodness of fit. As LM becomes large, the focus moves away from the data (and hence goodness of fit). This is an example of a *regularization* method in data analytics, where one expresses a prior belief concerning the smoothness of functions used for our predictions.

2.4.3 Data Modeling

One way to motivate an algorithmic development is through the use of a data model introduced in Section 2.1.2. Here, we can also think of this as a “probability” or “likelihood” based model, in that our main goal is to understand the target (Y) distribution, typically in terms of the explanatory variables. Thus, data models are particularly useful for the goal of explanation previously discussed in Section 2.3.3.

Data models were initially developed in the early twentieth century through the work of R.A. Fisher and E.P. George Box (among many, many others) whose work focused on data as the result of experiments with a small number of outcomes and even fewer explanatory (control) variables.

Linear Regression. The (algorithmic) linear regression with *OLS* estimates can be motivated using a probabilistic framework, as follows. We can think of the target variable y_i as having a normal distribution with unknown variance and a mean equal to $\mathbf{x}_i'\boldsymbol{\beta}$, a linear combination of the explanatory variables. Assuming independence among observations, it can be shown that the maximum likelihood estimates are equivalent to the *OLS* estimates determine in equation (2.4.2).

Maximum likelihood estimation is used extensively in this text, *you can get a quick overview in Chapter 18 Appendix C.* For additional background on *OLS* and maximum likelihood in the linear regression case see, for example, Frees (2009) for more details.

Poisson Regression. In the case where the target variable Y represents a count (such as the number of insurance losses), then it is common to use a Poisson distribution to represent the likelihood of potential outcomes. The Poisson has only one parameter, the mean, and if explanatory variables are available, then one can take the mean to equal $\exp(\mathbf{x}_i'\boldsymbol{\beta})$. One motivation for using the exponential ($\exp(\cdot)$) function is that it ensures that estimated means are non-negative (a necessary condition for the Poisson distribution). When maximum likelihood is used to estimate the regression coefficients, then this is known as *Poisson regression.*

Generalized Linear Model. The generalized linear model (*GLM*) consists of a wide family of regression models that include linear and Poisson regression models as special cases. In a *GLM*, the mean of the target variable is assumed to be a function of a linear combination of the explanatory variables. As with a Poisson regression, the mean can vary by observations by allowing some parameters to change yet the regression parameters $\boldsymbol{\beta}$ are assumed to be constant.

In a *GLM*, the target variable is assumed to follow a distribution from the *linear exponential family*, a collection of distributions that includes the normal, Poisson, Bernoulli, Weibull, and others. Thus, a *GLM* is one way of developing a broader class that includes linear and Poisson regression. Using a Bernoulli distribution, it also includes zero-one target variables resulting in what is known as *logistic regression.* Thus, the *GLM* provides a unifying framework to handle different types of target variables, including discrete and continuous variables. Extensions to other distributions that are not part of linear exponential family, such as a Pareto distribution, are also possible. But, *GLMs* have historically been found useful because their form permits efficient calculation of estimators (through what is known as *iterative reweighted least squares*). For more information about *GLM*s, readers are referred to De Jong and Heller (2008) and Frees (2009).

2.5 Data

In this section, you learn how to describe data considerations in terms of

- data types,
- data structure and storage,
- data cleaning,
- big data issues, and
- ethical issues.

Data constitute the backbone of “data analytics.” Without data containing useful information, no level of sophisticated analytic techniques can provide useful guidance for making good decisions.

The prior sections of this chapter provide the foundations of data considerations needed for the rest of this book. However, for readers who wish to specialize in data analytics, the following subsections provide a useful starting point for further study.

2.5.1 Data Types

In terms of how data are collected, data can be divided into two types (Hox and Boeije, 2005): primary and secondary data. Primary data are the original data that are collected for a specific research problem. Secondary data are data originally collected for a different purpose and reused for another research problem. A major advantage of using primary data is that the theoretical constructs, the research design, and the data collection strategy can be tailored to the underlying research question to ensure that data collected help to solve the problem. A disadvantage of using primary data is that data collection can be costly and time consuming. Using secondary data has the advantage of lower cost and faster access to relevant information. However, using secondary data may not be optimal for the research question under consideration.

In terms of the degree of organization, data can be also divided into two types: structured data and unstructured data. Structured data have a predictable and regularly occurring format. In contrast, unstructured data lack any regularly occurring format and have no structure that is recognizable to a computer. Structured data consist of records, attributes, keys, and indices and are typically managed by a database management system such as IBM DB2, Oracle, MySQL, and Microsoft SQL Server. As a result, most units of structured data can be located quickly and easily. Unstructured data have many different forms and variations. One common form of unstructured data is text. Accessing unstructured data can be awkward. To find a given unit of data in a long text, for example, a sequential search is usually performed.

2.5.2 Data Structures and Storage

As mentioned in the previous subsection, there are structured data as well as unstructured data. Structured data are highly organized data and usually have the following tabular format:

	V_1	V_2	$\cdots$	V_d
$\mathbf{x}_1$	x_{11}	x_{12}	$\cdots$	x_{1d}
$\mathbf{x}_2$	x_{21}	x_{22}	$\cdots$	x_{2d}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
$\mathbf{x}_n$	x_{n1}	x_{n2}	$\cdots$	x_{nd}

In other words, structured data can be organized into a table consisting of rows and columns. Typically, each row represents a record and each column represents an attribute. A table can be decomposed into several tables that can be stored in a relational database such as the Microsoft SQL Server. The

SQL (Structured Query Language) can be used to access and modify the data easily and efficiently.

Unstructured data do not follow a regular format. Examples of unstructured data include documents, videos, and audio files. Most of the data we encounter are unstructured data. In fact, the term "big data" was coined to reflect this fact. Traditional relational databases cannot meet the challenges on the varieties and scales brought by massive unstructured data nowadays. NoSQL databases have been used to store massive unstructured data.

There are three main NoSQL databases (Chen et al., 2014): key-value databases, column-oriented databases, and document-oriented databases. Key-value databases use a simple data model and store data according to key values. Modern key-value databases have higher expandability and smaller query response times than relational databases. Examples of key-value databases include Dynamo used by Amazon and Voldemort used by LinkedIn. Column-oriented databases store and process data according to columns rather than rows. The columns and rows are segmented in multiple nodes to achieve expandability. Examples of column-oriented databases include BigTable developed by Google and Cassandra developed by FaceBook. Document databases are designed to support more complex data forms than those stored in key-value databases. Examples of document databases include MongoDB, SimpleDB, and CouchDB. MongoDB is an open-source document-oriented database that stores documents as binary objects. SimpleDB is a distributed NoSQL database used by Amazon. CouchDB is another open-source document-oriented database.

2.5.3 Data Cleaning

Raw data usually need to be cleaned before useful analysis can be conducted. In particular, the following areas need attention when preparing data for analysis (Janert, 2010):

- **Missing values.** It is common to have missing values in raw data. Depending on the situation, we can discard the record, discard the variable, or impute the missing values.
- **Outliers.** Raw data may contain unusual data points such as outliers. We need to handle outliers carefully. We cannot just remove outliers without knowing the reason for their existence. Although sometimes outliers can be simple mistakes such as those caused by clerical errors, sometimes their unusual behavior can point to precisely the type of effect that we are looking for.
- **Junk.** Raw data may contain garbage, or junk, such as nonprintable characters. When it happens, junk is typically rare and not easily noticed. However, junk can cause serious problems in downstream applications.

- **Format.** Raw data may be formatted in a way that is inconvenient for subsequent analysis. For example, components of a record may be split into multiple lines in a text file. In such cases, lines corresponding to a single record should be merged before loading to a data analysis software such as `R`.
- **Duplicate records.** Raw data may contain duplicate records. Duplicate records should be recognized and removed. This task may not be trivial depending on what you consider "duplicate."
- **Merging datasets.** Raw data may come from different sources. In such cases, we need to merge data from different sources to ensure compatibility.

For more information about how to handle data in `R`, readers are referred to Forte (2015) and Buttrey and Whitaker (2017).

2.5.4 Big Data Analysis

Unlike traditional data analysis, big data analysis employs additional methods and tools that can extract information rapidly from massive data. In particular, big data analysis uses the following processing methods (Chen et al., 2014):

- A **bloom filter** is a space-efficient probabilistic data structure that is used to determine whether an element belongs to a set. It has the advantages of high space efficiency and high query speed. A drawback of using bloom filter is that there is a certain nonrecognition rate.
- **Hashing** is a method that transforms data into fixed-length numerical values through a hash function. It has the advantages of rapid reading and writing. However, sound hash functions are difficult to find.
- **Indexing** refers to a process of partitioning data in order to speed up reading. Hashing is a special case of indexing.
- A **trie,** also called digital tree, is a method to improve query efficiency by using common prefixes of character strings to reduce comparisons among character strings.
- **Parallel computing** uses multiple computing resources to complete a computation task. Parallel computing tools include Message Passing Interface (MPI), MapReduce, and Dryad.

Big data analysis can be conducted in the following levels (Chen et al., 2014): memory-level, business intelligence (BI) level, and massive level. Memory-level analysis is conducted when data can be loaded to the memory of a cluster of computers. Current hardware can handle hundreds of gigabytes (GB) of data in memory. BI level analysis can be conducted when data surpass the memory level. It is common for BI level analysis products to support data over terabytes (TB). Massive level analysis is conducted when data surpass the capabilities of products for BI level analysis. Usually Hadoop and MapReduce are used in massive level analysis.

2.5.5 Ethical Issues

Analysts may face ethical issues and dilemmas during the data analysis process. In some fields, ethical issues and dilemmas include participant consent, benefits, risk, confidentiality, and data ownership (Miles et al., 2014). For example, regarding privacy and confidentiality, one might confront the following questions: How do we make sure that the information is kept confidentially? How do we verify where raw data and analysis results are stored? How will we have access to them? These questions should be addressed and documented in explicit confidentiality agreements.

Within the insurance sector, discrimination, privacy, and confidentiality are major concerns. Discrimination in insurance is particularly difficult because the entire industry is based on "discriminating," or classifying, insureds into homogeneous categories for the purposes of risk sharing. Many variables that insurers use are seemingly innocuous (e.g., blindness for auto insurance), yet others can be viewed as "wrong" (e.g., religious affiliation), "unfair" (e.g., onset of cancer for health insurance), "sensitive" (e.g., marital status), or "mysterious" (e.g., Artificial Intelligence produced). Regulators and policymakers decide whether it is not permitted to use a variable for classification. In part because they depend on differing attitudes, perspectives can vary dramatically across jurisdictions. For example, gender-based pricing of auto insurance is permitted in all but a handful of U.S. states (the exceptions being Hawaii, Massachusetts, Montana, North Carolina, Pennsylvania, and, as of 2019, California) yet not permitted within the European Union. Moreover, for personal lines such as auto and homeowners, availability of big data may also lead to issues regarding *proxy discrimination*. Proxy discrimination occurs when a surrogate, or proxy, is used in place of a prohibited trait such as race or gender, see, for example, Frees and Huang (2021).

2.6 Further Resources and Contributors

Contributors

- **Guojun Gan**, University of Connecticut, was the principal author of the initial version of this chapter.
  - Chapter reviewers include: Runhuan Feng, Himchan Jeong, Lei Hua, Min Ji, and Toby White.
- **Hirokazu (Iwahiro) Iwasawa** and **Edward (Jed) Frees**, University of Wisconsin-Madison and Australian National University, are the authors of the second edition of this chapter. Email: iwahiro@bb.mbn.or.jp and/or jfrees@bus.wisc.edu for chapter comments and suggested improvements.

Further Readings and References

- Stigler (1986) gives a definitive account of the early contributions of Boscovich, Legendre and Gauss.
- Breiman (2001) compares the data modeling and the algorithmic modeling cultures.
- Good (1983) compares the two phases of data analysis, exploratory data analysis (EDA) and confirmatory data analysis (CDA)
- See, for example, Breiman (2001) and Shmueli (2010), for more discussions of the two goals in data analysis: explanation and prediction.
- Comparisons of structured data and unstructured data can be found in Inmon and Linstedt (2014), O'Leary (2013) ,Hashem et al. (2015), Abdullah and Ahmad (2013), and Pries and Dunnigan (2015), among others.

2.6.1 Technical Supplement: Multivariate Exploratory Analysis

Principal Component Analysis

Principal component analysis (PCA) is a statistical procedure that transforms a dataset described by possibly correlated variables into a dataset described by linearly uncorrelated variables, which are called principal components and are ordered according to their variances. PCA is a technique for dimension reduction. If the original variables are highly correlated, then the first few principal components can account for most of the variation of the original data.

The principal components of the variables are related to the eigenvalues and eigenvectors of the covariance matrix of the variables. For $i = 1, 2, \ldots, d$, let $(\lambda_i, \mathbf{e}_i)$ be the ith eigenvalue-eigenvector pair of the covariance matrix Σ of d variables $X_1, X_2, \ldots, X_d$ such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$ and the eigenvectors are normalized. Then the ith principal component is given by

$$Z_i = \mathbf{e}_i'\mathbf{X} = \sum_{j=1}^{d} e_{ij} X_j,$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_d)'$. It can be shown that Var $(Z_i) = \lambda_i$. As a result, the proportion of variance explained by the ith principal component is calculated as

$$\frac{\text{Var }(Z_i)}{\sum_{j=1}^{d} \text{Var }(Z_j)} = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_d}.$$

For more information about PCA, readers are referred to Mirkin (2011).

Cluster Analysis

Cluster analysis (aka data clustering) refers to the process of dividing a dataset into homogeneous groups or clusters such that points in the same cluster are similar and points from different clusters are quite distinct (Gan et al., 2007; Gan, 2011). Data clustering is one of the most popular tools for exploratory data analysis and has found its applications in many scientific areas.

During the past several decades, many clustering algorithms have been proposed. Among these clustering algorithms, the k-means algorithm is perhaps the most well-known algorithm due to its simplicity. To describe the k-means algorithm, let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a dataset containing n points, each of which is described by d numerical features. Given a desired number of clusters k, the k-means algorithm aims at minimizing the following objective function:

$$P(U, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{il} \|\mathbf{x}_i - \mathbf{z}_l\|^2,$$

where $U = (u_{il})_{n \times k}$ is an $n \times k$ partition matrix, $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k\}$ is a set of cluster centers, and $\| \cdot \|$ is the L^2 norm or Euclidean distance. The partition matrix U satisfies the following conditions:

$$u_{il} \in \{0, 1\}, \quad i = 1, 2, \ldots, n, \; l = 1, 2, \ldots, k,$$

$$\sum_{l=1}^{k} u_{il} = 1, \quad i = 1, 2, \ldots, n.$$

The k-means algorithm employs an iterative procedure to minimize the objective function. It repeatedly updates the partition matrix U and the cluster centers Z alternately until some stop criterion is met. For more information about k-means, readers are referred to Gan et al. (2007) and Mirkin (2011).

2.6.2 Tree-based Models

Decision trees, also known as tree-based models, involve dividing the predictor space (i.e., the space formed by independent variables) into a number of simple regions and using the mean or the mode of the region for prediction (Breiman et al., 1984). There are two types of tree-based models: classification trees and regression trees. When the dependent variable is categorical, the resulting tree models are called classification trees. When the dependent variable is continuous, the resulting tree models are called regression trees.

The process of building classification trees is similar to that of building regression trees. Here we only briefly describe how to build a regression tree. To do that, the predictor space is divided into non-overlapping regions such that the following objective function

$$f(R_1, R_2, \ldots, R_J) = \sum_{j=1}^{J} \sum_{i=1}^{n} I_{R_j}(\mathbf{x}_i)(y_i - \mu_j)^2$$

is minimized, where I is an indicator function, R_j denotes the set of indices of the observations that belong to the jth box, μ_j is the mean response of the observations in the jth box, $\mathbf{x}_i$ is the vector of predictor values for the ith observation, and y_i is the response value for the ith observation.

In terms of predictive accuracy, decision trees generally do not perform to the level of other regression and classification models. However, tree-based models may outperform linear models when the relationship between the response and the predictors is nonlinear. For more information about decision trees, readers are referred to Breiman et al. (1984) and Mitchell (1997).

2.6.3 Technical Supplement: Some R Functions

`R` is an open-source software for statistical computing and graphics. The `R` software can be downloaded from the `R` project website at https://www.r-project.org/. In this section, we give some `R` function for data analysis, especially the data analysis tasks mentioned in previous sections.

Table 2.6. **Some R Functions for Data Analysis**

Data Analysis Task	R Package	R Function
Descriptive Statistics	base	summary
Principal Component Analysis	stats	prcomp
Data Clustering	stats	kmeans, hclust
Fitting Distributions	MASS	fitdistr
Linear Regression Models	stats	lm
Generalized Linear Models	stats	glm
Regression Trees	rpart	rpart
Survival Analysis	survival	survfit

Table 2.6 lists a few R functions for different data analysis tasks. Readers can go to the R documentation to learn how to use these functions. There are also other R packages that do similar things. However, the functions listed in this table provide good starting points for readers to conduct data analysis in R. For analyzing large datasets in R in an efficient way, readers are referred to Daroczi (2015).

3

Frequency Modeling

Chapter Preview. A primary focus for insurers is estimating the magnitude of aggregate claims it must bear under its insurance contracts. Aggregate claims are affected by both the frequency and the severity of the insured event. Decomposing aggregate claims into these two components, each of which warrant significant attention, is essential for analysis and pricing. This chapter discusses frequency distributions, summary measures, and parameter estimation techniques.

In Section 3.1, we present terminology and discuss reasons why we study frequency and severity separately. The foundations of frequency distributions and measures are presented in Section 3.2 along with three principal distributions: the binomial, the Poisson, and the negative binomial. These three distributions are members of what is known as the $(a, b, 0)$ class of distributions, a distinguishing, identifying feature which allows for efficient calculation of probabilities, further discussed in Section 3.3. When fitting a dataset with a distribution, parameter values need to be estimated and in Section 3.4, the procedure for maximum likelihood estimation is explained.

For insurance datasets, the observation at zero denotes no occurrence of a particular event; this often deserves additional attention. As explained further in Section 3.5, for some datasets it may be impossible to have zero of the studied event or zero events may follow a different model than other event counts. In either case, direct fitting of typical count models could lead to improper estimates. Zero truncation or modification techniques allow for more appropriate distribution fit.

Noting that our insurance portfolio could consist of different sub-groups, each with its own set of individual characteristics, Section 3.6 introduces mixture distributions and methodology to allow for this heterogeneity within a portfolio. In Section 3.7 an example is given that demonstrates how standard frequency distributions can often provide a good fit to real data. Exercises are presented in Section 3.8 and Section 3.9.1 concludes the chapter with R Code for plots depicted in Section 3.4.

3.1 Frequency Distributions

In this section, you learn how to summarize the importance of frequency modeling in terms of

- contractual,
- behavioral,
- database, and
- regulatory/administrative motivations.

3.1.1 How Frequency Augments Severity Information

Basic Terminology

In this chapter, **loss**, also referred to as ground-up loss, denotes the amount of financial loss suffered by the insured. We use **claim** to denote the indemnification upon the occurrence of an insured event, thus the amount paid by the insurer. While some texts use **loss** and **claim** interchangeably, we wish to make a distinction here to recognize how insurance contractual provisions, such as deductibles and limits, affect the size of the claim stemming from a loss. Frequency represents how often an insured event occurs, typically within a policy contract. Here, we focus on count random variables that represent the number of claims, that is, how frequently an event occurs. Severity denotes the amount, or size, of each payment for an insured event. In Chapter 7, the aggregate model, which combines frequency models with severity models, is examined.

The Importance of Frequency

Recall from Section 1.2 that setting the price of an insurance good can be a complex problem. In manufacturing, the cost of a good is (relatively) known. In other financial service areas, market prices are available. In insurance, we can generalize the price setting as follows. Start with an expected cost, then add "margins" to account for the product's riskiness, expenses incurred in servicing the product, and a profit/surplus allowance for the insurer.

The expected cost for insurance can be determined as the expected number of claims times the amount per claim, that is, expected value of *frequency times severity*. The focus on claim count allows the insurer to consider those factors which directly affect the occurrence of a loss, thereby potentially generating a claim.

Why Examine Frequency Information?

Insurers and other stakeholders, including governmental organizations, have various motivations for gathering and maintaining frequency datasets.

- **Contractual.** In insurance contracts, it is common for particular deductibles and policy limits to be listed and invoked for each occurrence of an insured event. Correspondingly, the claim count data generated would indicate the number of claims which meet these criteria, offering a unique claim frequency measure. Extending this, models of total insured losses would need to account for deductibles and policy limits for each insured event.
- **Behavioral.** In considering factors that influence loss frequency, the risk-taking and risk-reducing behavior of individuals and companies should be considered. Explanatory (rating) variables can have different effects on models of how often an event occurs in contrast to the size of the event.
  - In healthcare, the decision to utilize healthcare by individuals, and minimize such healthcare utilization through preventive care and wellness measures, is related primarily to his or her personal characteristics. The cost per user is determined by the patient's medical condition, potential treatment measures, and decisions made by the healthcare provider (such as the physician) and the patient. While there is overlap in those factors and how they affect total healthcare costs, attention can be focused on those separate drivers of healthcare visit frequency and healthcare cost severity.
  - In personal lines, prior claims history is an important underwriting factor. It is common to use an indicator of whether or not the insured had a claim within a certain time period prior to the contract. Also, the number of claims incurred by the insured in previous periods has predictive power.
  - In homeowners insurance, in modeling potential loss frequency, the insurer could consider loss prevention measures that the homeowner has adopted, such as visible security systems. Separately, when modeling loss severity, the insurer would examine those factors that affect repair and replacement costs.
- **Databases**. Insurers may hold separate data files that suggest developing separate frequency and severity models. For example, a policyholder file is established when a policy is written. This file records much underwriting information about the insured(s), such as age, gender, and prior claims experience, policy information such as coverage, deductibles and limitations,

as well as any insurance claims event. A separate file, known as the "claims" file, records details of the claim against the insurer, including the amount. (There may also be a "payments" file that records the timing of the payments although we shall not deal with that here.) This recording process could then extend to insurers modeling the frequency and severity as separate processes.

- **Regulatory and Administrative.** Insurance is a highly regulated and monitored industry, given its importance in providing financial security to individuals and companies facing risk. As part of their duties, regulators routinely require the reporting of claims numbers as well as amounts. This may be due to the fact that there can be alternative definitions of an "amount," e.g., paid versus incurred, and there is less potential error when reporting claim numbers. This continual monitoring helps ensure financial stability of these insurance companies.

3.2 Basic Frequency Distributions

In this section, you learn how to:

- Determine quantities that summarize a distribution such as the distribution and survival function, as well as moments such as the mean and variance
- Define and compute the moment and probability generating functions
- Describe and understand relationships among three important frequency distributions: the binomial, Poisson, and negative binomial distributions

In this section, we introduce the distributions that are commonly used in actuarial practice to model count data. The claim count random variable is denoted by N; by its very nature it assumes only non-negative integer values. Hence the distributions below are all discrete distributions supported on the set of non-negative integers $\{0, 1, \ldots\}$.

3.2.1 Foundations

Since N is a discrete random variable taking values in $\{0, 1, \ldots\}$, the most natural full description of its distribution is through the specification of the probabilities with which it assumes each of the non-negative integer values. This leads us to the concept of the probability mass function (pmf) of N, denoted as $p_N(\cdot)$ and defined as follows:

$$p_N(k) = \Pr(N = k), \quad \text{for } k = 0, 1, \ldots$$

We note that there are alternate complete descriptions, or characterizations, of the distribution of N; for example, the distribution function of N defined by $F_N(x) = \Pr(N \leq x)$ and determined as:

$$F_N(x) = \begin{cases} \sum\limits_{k=0}^{\lfloor x \rfloor} \Pr(N = k), & x \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

In the above, $\lfloor \cdot \rfloor$ denotes the floor function; $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x. This expression also suggests the descriptor *cumulative distribution function,* a commonly used alternative way of expressing the distribution function. We also note that the survival function of N, denoted by $S_N(\cdot)$, is defined as the ones'-complement of $F_N(\cdot)$, *i.e.* $S_N(\cdot) = 1 - F_N(\cdot)$. Clearly, the latter is another characterization of the distribution of N.

Often one is interested in quantifying a certain aspect of the distribution and not in its complete description. This is particularly useful when comparing distributions. A *center of location* of the distribution is one such aspect, and there are many different measures that are commonly used to quantify it. Of these, the mean is the most popular; the mean of N, denoted by μ_N,[1] is defined as

Another basic aspect of a distribution is its dispersion, and of the various measures of dispersion studied in the literature, the standard deviation is the most popular. Towards defining it, we first define the variance of N, denoted

$$\mu_N = \sum_{k=0}^{\infty} k \; p_N(k).$$

We note that μ_N is the expected value of the random variable N, *i.e.* $\mu_N = \mathrm{E}[N]$. This leads to a general class of measures, the moments of the distribution; the r-th raw moment of N, for $r > 0$, is defined as $\mathrm{E}[N^r]$ and denoted by $\mu'_N(r)$. We remark that the prime $'$ here does *not* denote differentiation. Rather, it is commonly used notation to distinguish a raw from a central moment, as will be introduced in Section 4.1.1. For $r > 0$, we have

$$\mu'_N(r) = \mathrm{E}[N^r] = \sum_{k=0}^{\infty} k^r \; p_N(k).$$

We note that $\mu'_N(\cdot)$ is a well-defined non-decreasing function taking values in $[0, \infty]$, as $\Pr(N \in \{0, 1, \ldots\}) = 1$; also, note that $\mu_N = \mu'_N(1)$. In the following, when we refer to a moment it will be implicit that it is finite unless mentioned otherwise.

[1] For convenience, we have indexed μ_N with the random variable N instead of F_N or p_N, even though it is solely a function of the distribution of the random variable.

by Var$[N]$, as Var$[N] = \mathrm{E}[(N-\mu_N)^2]$ when μ_N is finite. By basic properties of the expected value of a random variable, we see that Var$[N] = \mathrm{E}[N^2] - [\mathrm{E}(N)]^2$. The standard deviation of N, denoted by σ_N, is defined as the square root of Var$[N]$. Note that the latter is well-defined as Var$[N]$, by its definition as the average squared deviation from the mean, is non-negative; Var$[N]$ is denoted by σ_N^2. Note that these two measures take values in $[0, \infty]$.

3.2.2 Moment and Probability Generating Functions

Now we introduce two generating functions that are found to be useful when working with count variables. For a discrete random variable, the moment generating function (mgf) of N, denoted as $M_N(\cdot)$, is defined as

$$M_N(t) = \mathrm{E}\left[e^{tN}\right] = \sum_{k=0}^{\infty} e^{tk}\, p_N(k), \quad t \in \mathbb{R}.$$

We note that while $M_N(\cdot)$ is well defined as it is the expectation of a non-negative random variable (e^{tN}), it can assume the value ∞. Note that for a count random variable, $M_N(\cdot)$ is finite valued on $(-\infty, 0]$ with $M_N(0) = 1$. The following theorem, whose proof can be found in Billingsley (2008) (pages 285-6), encapsulates the reason for its name.

Theorem 3.1.
Let N be a count random variable such that $\mathrm{E}\left[e^{t^*N}\right]$ is finite for some $t^* > 0$. We have the following:

a. All moments of N are finite, *i.e.*

$$\mathrm{E}[N^r] < \infty, \quad r > 0.$$

b. The *mgf* can be used to *generate* its moments as follows:

$$\left.\frac{\mathrm{d}^m}{\mathrm{d}t^m} M_N(t)\right|_{t=0} = \mathrm{E}[N^m], \quad m \geq 1.$$

c. The *mgf* $M_N(\cdot)$ characterizes the distribution; in other words it uniquely specifies the distribution.

Another reason that the *mgf* is very useful as a tool is that for two independent random variables X and Y, with their mgfs existing in a neighborhood of 0, the *mgf* of $X+Y$ is the product of their respective mgfs, that is, $M_{X+Y}(t) = M_X(t)M_Y(t)$, for small t.

A related generating function to the *mgf* is the probability generating function (pgf), and is a useful tool for random variables taking values in the non-negative integers. For a random variable N, by $P_N(\cdot)$ we denote its *pgf* and we define it as follows[2]:

$$P_N(s) = \mathrm{E}\ [s^N], \quad s \geq 0.$$

It is straightforward to see that if the *mgf* $M_N(\cdot)$ exists on $(-\infty, t^*)$ then

$$P_N(s) = M_N(\log(s)), \quad s < e^{t^*}.$$

Moreover, if the *pgf* exists on an interval $[0, s^*)$ with $s^* > 1$, then the *mgf* $M_N(\cdot)$ exists on $(-\infty, \log(s^*))$, and hence uniquely specifies the distribution of N by Theorem 3.1. (As a reminder, throughout this text we use *log* as the natural logarithm, not the base ten (common) logarithm or other version.) The following result for *pgf* is an analog of Theorem 3.1, and in particular justifies its name.

Theorem 3.2.
Let N be a count random variable such that $\mathrm{E}\ (s^*)^N$ is finite for some $s^* > 1$. We have the following:

a. All moments of N are finite, *i.e.*

$$\mathrm{E}\ N^r < \infty, \quad r \geq 0.$$

b. The pmf of N can be derived from the *pgf* as follows:

$$p_N(m) = \begin{cases} P_N(0), & m = 0; \\ \\ \left(\frac{1}{m!}\right) \frac{\mathrm{d}^m}{\mathrm{d}s^m} P_N(s)\Big|_{s=0}, & m \geq 1. \end{cases}$$

c. The factorial moments of N can be derived as follows:

$$\frac{\mathrm{d}^m}{\mathrm{d}s^m} P_N(s)\bigg|_{s=1} = \mathrm{E} \prod_{i=0}^{m-1} (N - i), \quad m \geq 1.$$

d. The *pgf* $P_N(\cdot)$ characterizes the distribution; in other words it uniquely specifies the distribution.

[2] $0^0 = 1$

3.2.3 Important Frequency Distributions

In this sub-section we study three important frequency distributions used in statistics, namely the binomial, the Poisson, and the negative binomial distributions. In the following, a risk denotes a unit covered by insurance. A risk could be an individual, a building, a company, or some other identifier for which insurance coverage is provided. For context, imagine an insurance data set containing the number of claims by risk or stratified in some other manner. The above mentioned distributions also happen to be the most commonly used in insurance practice for reasons, some of which we mention below.

- These distributions can be motivated by natural random experiments which are good approximations to real life processes from which many insurance data arise. Hence, not surprisingly, they together offer a reasonable fit to many insurance data sets of interest. The appropriateness of a particular distribution for the set of data can be determined using standard statistical methodologies, as we discuss later in this chapter.
- They provide a rich enough basis for generating other distributions that even better approximate or well cater to more real situations of interest to us.
  - The three distributions are either one-parameter or two-parameter distributions. In fitting to data, a parameter is assigned a particular value. The set of these distributions can be enlarged to their convex hulls by treating the parameter(s) as a random variable (or vector) with its own probability distribution, with this larger set of distributions offering greater flexibility. A simple example that is better addressed by such an enlargement is a portfolio of claims generated by insureds belonging to many different risk classes.
  - In insurance data, we may observe either a marginal or inordinate number of zeros, that is, zero claims by risk. When fitting to the data, a frequency distribution in its standard specification often fails to reasonably account for this occurrence. The natural modification of the above three distributions, however, accommodate this phenomenon well towards offering a better fit.
  - In insurance we are interested in total claims paid, whose distribution results from compounding the fitted frequency distribution with a severity distribution. These three distributions have properties that make it easy to work with the resulting aggregate severity distribution.

Binomial Distribution

We begin with the binomial distribution which arises from any finite sequence of identical and independent experiments with binary outcomes. The most canonical of such experiments is the (biased or unbiased) coin tossing experiment with the outcome being heads or tails. So if N denotes the number of heads in a sequence of m independent coin tossing experiments with an identical coin which turns heads up with probability q, then the distribution of N is called the binomial distribution with parameters (m, q), with m a positive integer and $q \in [0, 1]$. Note that when $q = 0$ (resp., $q = 1$) then the distribution is degenerate with $N = 0$ (resp., $N = m$) with probability 1. Clearly, its support when $q \in (0, 1)$ equals $\{0, 1, \dots, m\}$ with *pmf* given by [3]

$$p_k = \binom{m}{k} q^k (1-q)^{m-k}, \quad k = 0, \dots, m.$$

where

$$\binom{m}{k} = \frac{m!}{k!(m-k)!}$$

The reason for its name is that the *pmf* takes values among the terms that arise from the binomial expansion of $(q + (1-q))^m$. This realization then leads to the the following expression for the *pgf* of the binomial distribution:

$$\begin{aligned} P_N(z) &= \textstyle\sum_{k=0}^m z^k \binom{m}{k} q^k (1-q)^{m-k} \\ &= \textstyle\sum_{k=0}^m \binom{m}{k} (zq)^k (1-q)^{m-k} \\ &= (qz + (1-q))^m = (1 + q(z-1))^m. \end{aligned}$$

Note that the above expression for the *pgf* confirms the fact that the binomial distribution is the m-convolution of the Bernoulli distribution, which is the binomial distribution with $m = 1$ and *pgf* $(1 + q(z-1))$. By "m-convolution," we mean that we can write N as the sum of $N_1, \dots, N_m$. Here, N_i are iid Bernoulli variates. Also, note that the *mgf* of the binomial distribution is given by $(1 + q(e^t - 1))^m$.

The mean and variance of the binomial distribution can be found in a few different ways. To emphasize the key property that it is a m-convolution of the Bernoulli distribution, we derive below the moments using this property. We begin by observing that the Bernoulli distribution with parameter q assigns probability of q and $1 - q$ to 1 and 0, respectively. So its mean equals q $(= 0 \times (1-q) + 1 \times q)$; note that its raw second moment equals its mean as $N^2 = N$ with probability 1. Using these two facts we see that the variance equals $q(1-q)$. Moving on to the binomial distribution with parameters m and q, using the fact that it is the m-convolution of the Bernoulli distribution, we write N as the sum of $N_1, \dots, N_m$, where N_i are *iid* Bernoulli variates, as above. Now using the moments of Bernoulli and linearity of the expectation,

[3] In the following we suppress the reference to N and denote the *pmf* by the sequence $\{p_k\}_{k \geq 0}$, instead of the function $p_N(\cdot)$.

we see that

$$\mathrm{E}[N] = \mathrm{E}\left[\sum_{i=1}^{m} N_i\right] = \sum_{i=1}^{m} \mathrm{E}[N_i] = mq.$$

Also, using the fact that the variance of the sum of independent random variables is the sum of their variances, we see that

$$\mathrm{Var}[N] = \mathrm{Var}\left[\sum_{i=1}^{m} N_i\right] = \sum_{i=1}^{m} \mathrm{Var}[N_i] = mq(1-q).$$

Alternate derivations of the above moments are suggested in the exercises. One important observation, especially from the point of view of applications, is that the mean is greater than the variance unless $q = 0$.

Poisson Distribution

After the binomial distribution, the Poisson distribution (named after the French polymath Simeon Denis Poisson) is probably the most well known of discrete distributions. This is partly due to the fact that it arises naturally as the distribution of the count of the random occurrences of a type of event in a certain time period, if the rate of occurrences of such events is a constant. It also arises as the asymptotic limit of the binomial distribution with $m \to \infty$ and $mq \to \lambda$.

The Poisson distribution is parametrized by a single parameter usually denoted by λ which takes values in $(0, \infty)$. Its *pmf* is given by

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, \ldots$$

It is easy to check that the above specifies a *pmf* as the terms are clearly non-negative, and that they sum to one follows from the infinite Taylor series expansion of e^{λ}. More generally, we can derive its *pgf*, $P_N(\cdot)$, as follows:

$$P_N(z) = \sum_{k=0}^{\infty} p_k z^k = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k z^k}{k!} = e^{-\lambda}e^{\lambda z} = e^{\lambda(z-1)}, \forall z \in \mathbb{R}.$$

From the above, we derive its *mgf* as follows:

$$M_N(t) = P_N(e^t) = e^{\lambda(e^t-1)}, t \in \mathbb{R}.$$

Towards deriving its mean, we note that for the Poisson distribution

$$kp_k = \begin{cases} 0, & k = 0 \\ \lambda\ p_{k-1}, & k \geq 1. \end{cases}$$

This can be checked easily. In particular, this implies that

$$\mathrm{E}[N] = \sum_{k\geq 0} k\ p_k = \lambda \sum_{k\geq 1} p_{k-1} = \lambda \sum_{j\geq 0} p_j = \lambda.$$

In fact, more generally, using either a generalization of the above or using Theorem 3.1, we see that

$$\mathrm{E} \prod_{i=0}^{m-1} (N-i) = \frac{\mathrm{d}^m}{\mathrm{d}s^m} P_N(s)\bigg|_{s=1} = \lambda^m, \quad m \geq 1.$$

This, in particular, implies that

$$\mathrm{Var}[N] = \mathrm{E}[N^2] - [\mathrm{E}(N)]^2 = \mathrm{E}\ [N(N-1)] + \mathrm{E}[N] - (\mathrm{E}[N])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Note that interestingly for the Poisson distribution $\mathrm{Var}[N] = \mathrm{E}[N]$.

Negative Binomial Distribution

The third important count distribution is the negative binomial distribution. Recall that the binomial distribution arose as the distribution of the number of *successes* in m independent repetitions of an experiment with binary outcomes. If we instead consider the number of *successes* until we observe the r-th *failure* in independent repetitions of an experiment with binary outcomes, then its distribution is a negative binomial distribution. A particular case, when $r = 1$, is the geometric distribution. However when r in not an integer, the above random experiment would not be applicable. In the following, we allow the parameter r to be any positive real number to then motivate the distribution more generally. To explain its name, we recall the binomial series, *i.e.*

$$(1+x)^s = 1 + sx + \frac{s(s-1)}{2!}x^2 + \ldots..., \quad s \in \mathbb{R}; |x| < 1.$$

If we define $\binom{s}{k}$, the generalized binomial coefficient, by

$$\binom{s}{k} = \frac{s(s-1)\cdots(s-k+1)}{k!},$$

then we have

$$(1+x)^s = \sum_{k=0}^{\infty} \binom{s}{k} x^k, \quad s \in \mathbb{R}; |x| < 1.$$

If we let $s = -r$, then we see that the above yields

$$(1-x)^{-r} = 1 + rx + \frac{(r+1)r}{2!}x^2 + \ldots... = \sum_{k=0}^{\infty} \binom{r+k-1}{k} x^k, \quad r \in \mathbb{R}; |x| < 1.$$

This implies that if we define p_k as

$$p_k = \binom{r+k-1}{k}\left(\frac{1}{1+\beta}\right)^r\left(\frac{\beta}{1+\beta}\right)^k, \quad k = 0, 1, \ldots$$

for $r > 0$ and $\beta \geq 0$, then it defines a valid *pmf*. Such defined distribution is called the negative binomial distribution with parameters (r, β) with $r > 0$ and $\beta \geq 0$. Moreover, the binomial series also implies that the *pgf* of this distribution is given by

$$P_N(z) = (1-\beta(z-1))^{-r}, \quad |z| < 1 + \frac{1}{\beta}, \beta \geq 0.$$

The above implies that the *mgf* is given by

$$M_N(t) = (1-\beta(e^t-1))^{-r}, \quad t < \log\left(1+\frac{1}{\beta}\right), \beta \geq 0.$$

We derive its moments using Theorem 3.1 as follows:

$$\begin{aligned}
\mathrm{E}[N] &= M'(0) = r\beta e^t(1-\beta(e^t-1))^{-r-1}\Big|_{t=0} = r\beta; \\
\mathrm{E}[N^2] &= M''(0) = \Big[r\beta e^t(1-\beta(e^t-1))^{-r-1} + r(r+1)\beta^2 e^{2t}(1-\beta(e^t-1))^{-r-2}\Big]\Big|_{t=0} \\
&= r\beta(1+\beta) + r^2\beta^2; \\
\text{and } \mathrm{Var}[N] &= \mathrm{E}[N^2] - (\mathrm{E}[N])^2 = r\beta(1+\beta) + r^2\beta^2 - r^2\beta^2 = r\beta(1+\beta)
\end{aligned}$$

We note that when $\beta > 0$, we have $\mathrm{Var}[N] > \mathrm{E}[N]$. In other words, this distribution is overdispersed (relative to the Poisson); similarly, when $q > 0$ the binomial distribution is said to be underdispersed (relative to the Poisson).

Finally, we observe that the Poisson distribution also emerges as a limit of negative binomial distributions. Towards establishing this, let β_r be such that as r approaches infinity $r\beta_r$ approaches $\lambda > 0$. Then we see that the mgfs of negative binomial distributions with parameters (r, β_r) satisfies

$$\lim_{r\to 0}(1-\beta_r(e^t-1))^{-r} = \exp\{\lambda(e^t-1)\},$$

with the right hand side of the above equation being the *mgf* of the Poisson distribution with parameter λ.[4]

3.3 The $(a, b, 0)$ Class

[4] For the theoretical basis underlying the above argument, see Billingsley (2008).

In this section, you learn how to:

- Define the $(a,b,0)$ class of frequency distributions
- Discuss the importance of the recursive relationship underpinning this class of distributions
- Identify conditions under which this general class reduces to each of the binomial, Poisson, and negative binomial distributions

---

In the previous section we studied three distributions, namely the binomial, the Poisson and the negative binomial distributions. In the case of the Poisson, to derive its mean we used the the fact that

$$kp_k = \lambda p_{k-1}, \quad k \geq 1,$$

which can be expressed equivalently as

$$\frac{p_k}{p_{k-1}} = \frac{\lambda}{k}, \quad k \geq 1.$$

Interestingly, we can similarly show that for the binomial distribution

$$\frac{p_k}{p_{k-1}} = \frac{-q}{1-q} + \left(\frac{(m+1)q}{1-q}\right)\frac{1}{k}, \quad k = 1, \ldots, m,$$

and that for the negative binomial distribution

$$\frac{p_k}{p_{k-1}} = \frac{\beta}{1+\beta} + \left(\frac{(r-1)\beta}{1+\beta}\right)\frac{1}{k}, \quad k \geq 1.$$

The above relationships are all of the form

$$\frac{p_k}{p_{k-1}} = a + \frac{b}{k}, \quad k \geq 1; \tag{3.1}$$

this raises the question if there are any other distributions which satisfy this seemingly general recurrence relation. Note that the ratio on the left, the ratio of two probabilities, is non-negative.

Snippet of Theory. To begin with, let $a < 0$. In this case as $k \to \infty$, $(a + b/k) \to a < 0$. It follows that if $a < 0$ then b should satisfy $b = -ka$, for some $k \geq 1$. Any such pair (a, b) can be written as

$$\left(\frac{-q}{1-q}, \frac{(m+1)q}{1-q}\right), \quad q \in (0,1), m \geq 1;$$

note that the case $a < 0$ with $a + b = 0$ yields the degenerate at 0 distribution which is the binomial distribution with $q = 0$ and arbitrary $m \geq 1$.

In the case of $a = 0$, again by non-negativity of the ratio p_k/p_{k-1}, we have $b \geq 0$. If $b = 0$ the distribution is degenerate at 0, which is a binomial with $q = 0$ or a Poisson distribution with $\lambda = 0$ or a negative binomial distribution with $\beta = 0$. If $b > 0$, then clearly such a distribution is a Poisson distribution with mean (*i.e.* λ) equal to b, as presented at the beginning of this section.

In the case of $a > 0$, again by non-negativity of the ratio p_k/p_{k-1}, we have $a + b/k \geq 0$ for all $k \geq 1$. The most stringent of these is the inequality $a + b \geq 0$. Note that $a + b = 0$ again results in degeneracy at 0; excluding this case we have $a + b > 0$ or equivalently $b = (r - 1)a$ with $r > 0$. Some algebra easily yields the following expression for p_k:

$$p_k = \binom{r+k-1}{k} p_0 a^k, \quad k = 1, 2, \ldots .$$

The above series converges for $a < 1$ when $r > 0$, with the sum given by $p_0 \cdot ((1-a)^{(-r)} - 1)$. Hence, equating the latter to $1 - p_0$ we get $p_0 = (1-a)^{(r)}$. So in this case the pair (a, b) is of the form $(a, (r-1)a)$, for $r > 0$ and $0 < a < 1$; since an equivalent parametrization is $(\beta/(1+\beta), (r-1)\beta/(1+\beta))$, for $r > 0$ and $\beta > 0$, we see from above that such distributions are negative binomial distributions.

From the above development we see that not only does the recurrence (3.1) tie these three distributions together, but also it characterizes them. For this reason these three distributions are collectively referred to in the actuarial literature as (a,b,0) class of distributions, with 0 referring to the starting point of the recurrence. Note that the value of p_0 is implied by (a, b) since the probabilities have to sum to one. Of course, (3.1) as a recurrence relation for p_k makes the computation of the *pmf* efficient by removing redundancies. Later, we will see that it does so even in the case of compound distributions with the frequency distribution belonging to the $(a, b, 0)$ class - this fact is the more important motivating reason to study these three distributions from this viewpoint.

Example 3.3.1. A discrete probability distribution has the following properties

$$\begin{aligned} p_k &= c\left(1 + \frac{2}{k}\right) p_{k-1} \quad k = 1, 2, 3, \ldots \\ p_1 &= \frac{9}{256} \end{aligned}$$

Determine the expected value of this discrete random variable.

Example Solution. Since the pmf satisfies the $(a, b, 0)$ recurrence relation we know that the underlying distribution is one among the binomial, Poisson, and negative binomial distributions. Since the ratio of the parameters (*i.e.* b/a) equals 2, we know that it is negative binomial and that $r = 3$. Moreover, since for a negative binomial $p_1 = r(1+\beta)^{-(r+1)}\beta$, we have

$$\begin{aligned} \frac{9}{256} =& 3\frac{\beta}{(1+\beta)^4} \\ \implies \quad \frac{3}{(1+3)^4} =& \frac{\beta}{(1+\beta)^4} \\ \implies \quad \beta =& 3. \end{aligned}$$

Finally, since the mean of a negative binomial is $r\beta$ we have the mean of the given distribution equals 9.

3.4 Estimating Frequency Distributions

In this section, you learn how to:

- Define a likelihood for a sample of observations from a discrete distribution
- Define the maximum likelihood estimator for a random sample of observations from a discrete distribution
- Calculate the maximum likelihood estimator for the binomial, Poisson, and negative binomial distributions

3.4.1 Parameter Estimation

In Section 3.2 we introduced three distributions of importance in modeling various types of count data arising from insurance. Let us now suppose that we have a set of count data to which we wish to fit a distribution, and that we have determined that one of these $(a, b, 0)$ distributions is more appropriate than the others. Since each one of these forms a class of distributions if we allow its parameter(s) to take any permissible value, there remains the task of determining the **best** value of the parameter(s) for the data at hand. This is a statistical point estimation problem, and in parametric inference problems the statistical inference paradigm of *maximum likelihood* usually yields efficient estimators. In this section we describe this paradigm and derive the maximum likelihood estimators.

Let us suppose that we observe the independent and identically distributed, *iid*, random variables $X_1, X_2, \ldots, X_n$ from a distribution with *pmf* p_θ, where θ is a vector of parameters and an unknown value in the parameter space $\Theta \subseteq \mathbb{R}^d$. For example, in the case of the Poisson distribution, there is a single parameter so that $d = 1$ and

$$p_\theta(x) = e^{-\theta} \frac{\theta^x}{x!}, \quad x = 0, 1, \ldots,$$

with $\theta > 0$. In the case of the binomial distribution we have

$$p_\theta(x) = \binom{m}{x} q^x (1-q)^{m-x}, \quad x = 0, 1, \ldots, m.$$

For some applications, we can view m as a parameter and so take $d = 2$ so that $\theta = (m, q) \in \{0, 1, 2, \ldots\} \times [0, 1]$.

Let us suppose that the observations are $x_1, \ldots, x_n$, observed values of the random sample $X_1, X_2, \ldots, X_n$ presented earlier. In this case, the probability of observing this sample from p_θ equals

$$\prod_{i=1}^{n} p_\theta(x_i).$$

The above, denoted by $L(\theta)$, viewed as a function of θ, is called the *likelihood*. Note that we suppressed its dependence on the data, to emphasize that we are viewing it as a function of the parameter vector. For example, in the case of the Poisson distribution we have

$$L(\lambda) = e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i} \left(\prod_{i=1}^{n} x_i! \right)^{-1}.$$

In the case of the binomial distribution we have

$$L(m, q) = \left(\prod_{i=1}^{n} \binom{m}{x_i} \right) q^{\sum_{i=1}^n x_i} (1-q)^{nm - \sum_{i=1}^n x_i}.$$

The maximum likelihood estimator (mle) for θ is any maximizer of the likelihood; in a sense the *mle* chooses the set of parameter values that best explains the observed observations. Appendix Section 17.2.2 reviews the foundations of maximum likelihood estimation with more mathematical details in Appendix Chapter 19.

Special Case: Three Bernoulli Outcomes. To illustrate, consider a sample of size $n = 3$ from a Bernoulli distribution (binomial with $m = 1$) with values $0, 1, 0$. The likelihood in this case is easily checked to equal

$$L(q) = q(1-q)^2,$$

and the plot of the likelihood is given in Figure 3.1. As shown in the plot, the maximum value of the likelihood equals $4/27$ and is attained at $q = 1/3$, and hence the maximum likelihood estimate for q is $1/3$ for the given sample. In this case one can resort to algebra to show that

$$q(1-q)^2 = \left(q - \frac{1}{3}\right)^2 \left(q - \frac{4}{3}\right) + \frac{4}{27},$$

and conclude that the maximum equals $4/27$, and is attained at $q = 1/3$ (using the fact that the first term is non-positive in the interval $[0, 1]$).

But as is apparent, this way of deriving the *mle* using algebra does not generalize. In general, one resorts to calculus to derive the *mle* - note that for some likelihoods one may have to resort to other optimization methods, especially when the likelihood has many local extrema. It is customary to equivalently maximize the logarithm of the likelihood[5] $L(\cdot)$, denoted by $l(\cdot)$, and look at the set of zeros of its first derivative[6] $l'(\cdot)$. In the case of the above likelihood, $l(q) = \log(q) + 2\log(1-q)$, and

$$l'(q) = \frac{\mathrm{d}}{\mathrm{d}q} l(q) = \frac{1}{q} - \frac{2}{1-q}.$$

The unique zero of $l'(\cdot)$ equals $1/3$, and since $l''(\cdot)$ is negative, we have $1/3$ is the unique maximizer of the likelihood and hence its maximum likelihood estimate.

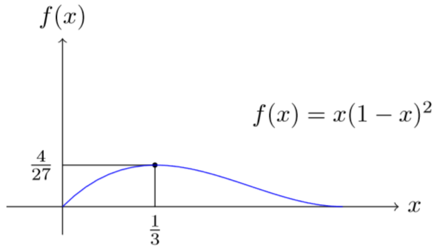


FIGURE 3.1: **Likelihood of a $(0, 1, 0)$ 3-sample from Bernoulli**

[5] The set of maximizers of $L(\cdot)$ are the same as the set of maximizers of any strictly increasing function of $L(\cdot)$, and hence the same as those for $l(\cdot)$.

[6] A slight benefit of working with $l(\cdot)$ is that constant terms in $L(\cdot)$ do not appear in $l'(\cdot)$ whereas they do in $L'(\cdot)$.

3.4.2 Frequency Distributions MLE

In the following, we derive the maximum likelihood estimator, *mle*, for the three members of the $(a, b, 0)$ class. We begin by summarizing the discussion above. In the setting of observing *iid*, independent and identically distributed, random variables $X_1, X_2, \ldots, X_n$ from a distribution with *pmf* p_θ, where θ takes an unknown value in $\Theta \subseteq \mathbb{R}^d$, the likelihood $L(\cdot)$, a function on Θ is defined as

$$L(\theta) = \prod_{i=1}^{n} p_\theta(x_i),$$

where $x_1, \ldots, x_n$ are the observed values. The *mle* of θ, denoted as $\hat{\theta}_{\mathrm{MLE}}$, is a function which maps the observations to an element of the set of maximizers of $L(\cdot)$, namely

$$\{\theta | L(\theta) = \max_{\eta \in \Theta} L(\eta)\}.$$

Note the above set is a function of the observations, even though this dependence is not made explicit. In the case of the three distributions that we study, and quite generally, the above set is a singleton with probability tending to one (with increasing sample size). In other words, for many commonly used distributions and when the sample size is large, the likelihood estimate is uniquely defined with high probability.

In the following, we assume that we have observed n *iid* random variables $X_1, X_2, \ldots, X_n$ from the distribution under consideration, even though the parametric value is unknown. Also, $x_1, x_2, \ldots, x_n$ will denote the observed values. We note that in the case of count data, and data from discrete distributions in general, the likelihood can alternately be represented as

$$L(\theta) = \prod_{k \geq 0} \left(p_\theta(k)\right)^{m_k},$$

where m_k is the number of observations equal to k. Mathematically, we have

$$m_k = \left|\{i | x_i = k, 1 \leq i \leq n\}\right| = \sum_{i=1}^{n} I(x_i = k), \quad k \geq 0.$$

Note that this transformation retains all of the data, compiling it in a streamlined manner. For large n it leads to compression of the data in the sense of *sufficiency*. Below, we present expressions for the *mle* in terms of $\{m_k\}_{k \geq 1}$ as well.

Special Case: Poisson Distribution. In this case, as noted above, the likelihood is given by

$$L(\lambda) = \left(\prod_{i=1}^{n} x_i!\right)^{-1} e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}.$$

Taking logarithms, the log-likelihood is

$$l(\lambda) = -\sum_{i=1}^{n} \log(x_i!) - n\lambda + \log(\lambda) \cdot \sum_{i=1}^{n} x_i.$$

Taking a derivative, we have

$$l'(\lambda) = -n + \frac{1}{\lambda} \sum_{i=1}^{n} x_i.$$

In evaluating $l''(\lambda)$, when $\sum_{i=1}^{n} x_i > 0$, $l'' < 0$. Consequently, the maximum is attained at the sample mean, $\overline{x}$, presented below. When $\sum_{i=1}^{n} x_i = 0$, the likelihood is a decreasing function and hence the maximum is attained at the least possible parameter value; this results in the maximum likelihood estimate being zero. Hence, we have

$$\overline{x} = \hat{\lambda}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Note that the sample mean can be computed also as

$$\overline{x} = \frac{1}{n} \sum_{k \geq 1} k \cdot m_k \ .$$

It is noteworthy that in the case of the Poisson, the exact distribution of $\hat{\lambda}_{\text{MLE}}$ is available in closed form - it is a scaled Poisson - when the underlying distribution is a Poisson. This is so as the sum of independent Poisson random variables is a Poisson as well. Of course, for large sample size one can use the ordinary Central Limit Theorem (CLT) to derive a normal approximation. Note that the latter approximation holds even if the underlying distribution is any distribution with a finite second moment.

Special Case: Binomial Distribution with known m. Unlike the case of the Poisson distribution, the parameter space in the case of the binomial is 2-dimensional. Hence the optimization problem is a bit more challenging. We first discuss the case where m is taken to be known - this is not a realistic assumption in insurance applications but is appropriate in circumstances where we are observing m iid binary outcomes with unknown probabilities.

We begin by observing that the likelihood is given by

$$L(m,q) = \left(\prod_{i=1}^{n} \binom{m}{x_i} \right) q^{\sum_{i=1}^{n} x_i} (1-q)^{nm - \sum_{i=1}^{n} x_i}.$$

Taking logarithms, the log-likelihood is

$$\begin{aligned} l(m,q) &= \textstyle\sum_{i=1}^{n} \log\left(\binom{m}{x_i}\right) + \left(\sum_{i=1}^{n} x_i\right) \log(q) \\ &\quad + \left(nm - \textstyle\sum_{i=1}^{n} x_i\right) \log(1-q) \\ &= \textstyle\sum_{i=1}^{n} \log\left(\binom{m}{x_i}\right) + n\overline{x} \log(q) + n\left(m - \overline{x}\right) \log(1-q), \end{aligned}$$

where $\overline{x} = n^{-1}\sum_{i=1}^{n} x_i$. Since we have assumed that m is known, maximizing $l(m,q)$ with respect to q involves taking the first differential and equating to zero:

$$\frac{\mathrm{d}l(m,q)}{\mathrm{d}q} = \frac{n\overline{x}}{q} - \frac{n\,(m-\overline{x})}{1-q} = 0,$$

which implies that

$$\hat{q}_{MLE} = \frac{\overline{x}}{m}.$$

Special Case: Binomial Distribution with unknown m. Note that since m takes only non-negative integer values, we cannot use multivariate calculus to find the optimal values. Nevertheless, we can use single variable calculus to show that

$$\hat{q}_{MLE} \times \hat{m}_{MLE} = \overline{x}. \tag{3.2}$$

Towards this we note that for a fixed value of m,

$$\frac{\delta}{\delta q} l(m,q) = \frac{n\overline{x}}{q} - \frac{n\,(m-\overline{x})}{1-q},$$

and that

$$\frac{\delta^2}{\delta q^2} l(m,q) = -\frac{n\overline{x}}{q^2} + \frac{n\,(m-\overline{x})}{(1-q)^2} \leq 0.$$

The above implies that for any fixed value of m, the maximizing value of q satisfies

$$mq = \overline{x},$$

and hence we establish equation (3.2).

With equation (3.2), the above reduces the task to the search for $\hat{m}_{\text{MLE}}$, which is a maximizer of

$$L\left(m, \frac{\overline{x}}{m}\right). \tag{3.3}$$

Note the likelihood would be zero for values of m smaller than $\max\limits_{1\leq i\leq n} x_i$, and hence $\hat{m}_{\text{MLE}} \geq \max_{1\leq i\leq n} x_i$.

Towards specifying an algorithm to compute $\hat{m}_{\text{MLE}}$, we first point out that for some data sets $\hat{m}_{\text{MLE}}$ could equal ∞, indicating that a Poisson distribution would render a better fit than any binomial distribution. This is so as the binomial distribution with parameters $(m, \overline{x}/m)$ approaches the Poisson distribution with parameter $\overline{x}$ with m approaching infinity. The fact that some data sets **prefer** a Poisson distribution should not be surprising since in the above sense the set of Poisson distribution is on the boundary of the set of binomial distributions. Interestingly, in Olkin et al. (1981) they show that if the sample mean is less than or equal to the sample variance then $\hat{m}_{\mathrm{MLE}} = \infty$; otherwise, there exists a finite m that maximizes equation (3.3).

In Figure 3.2 below we display the plot of $L\left(m, \overline{x}/m\right)$ for three different samples of size 5; they differ only in the value of the sample maximum. The first sample of $(2, 2, 2, 4, 5)$ has the ratio of sample mean to sample variance greater than 1 (1.875), the second sample of $(2, 2, 2, 4, 6)$ has the ratio equal to 1.25 which is closer to 1, and the third sample of $(2, 2, 2, 4, 7)$ has the ratio less than 1 (0.885). For these three samples, as shown in Figure 3.2, $\hat{m}_{\mathrm{MLE}}$ equals 7, 18 and ∞, respectively. Note that the limiting value of $L\left(m, \overline{x}/m\right)$ as m approaches infinity equals

$$\left(\prod_{i=1}^{n} x_i!\right)^{-1} \exp\left(-n\overline{x}\ \right)\left(\ \overline{x}\ \right)^{n\overline{x}}. \tag{3.4}$$

Also, note that Figure 3.2 shows that the *mle* of m is non-robust, *i.e.* changes in a small proportion of the data set can cause large changes in the estimator.

The above discussion suggests the following simple algorithm:

- *Step 1*. If the sample mean is less than or equal to the sample variance, then set $\hat{m}_{MLE} = \infty$. The *mle* suggested distribution is a Poisson distribution with $\hat{\lambda} = \overline{x}$.
- *Step 2*. If the sample mean is greater than the sample variance, then compute $L(m, \overline{x}/m)$ for m values greater than or equal to the sample maximum until $L(m, \overline{x}/m)$ is close to the value of the Poisson likelihood given in (3.4). The value of m that corresponds to the maximum value of $L(m, \overline{x}/m)$ among those computed equals $\hat{m}_{MLE}$.

We note that if the underlying distribution is the binomial distribution with parameters (m, q) (with $q > 0$) then $\hat{m}_{MLE}$ equals m for large sample sizes. Also, $\hat{q}_{MLE}$ will have an asymptotically normal distribution and converge with probability one to q.

Special Case: Negative Binomial Distribution. The case of the negative binomial distribution is similar to that of the binomial distribution in the sense that we have two parameters and the *mle*s are not available in closed form. A difference between them is that unlike the binomial parameter m which takes positive integer values, the parameter r of the negative binomial can assume any positive real value. This makes the optimization problem a tad

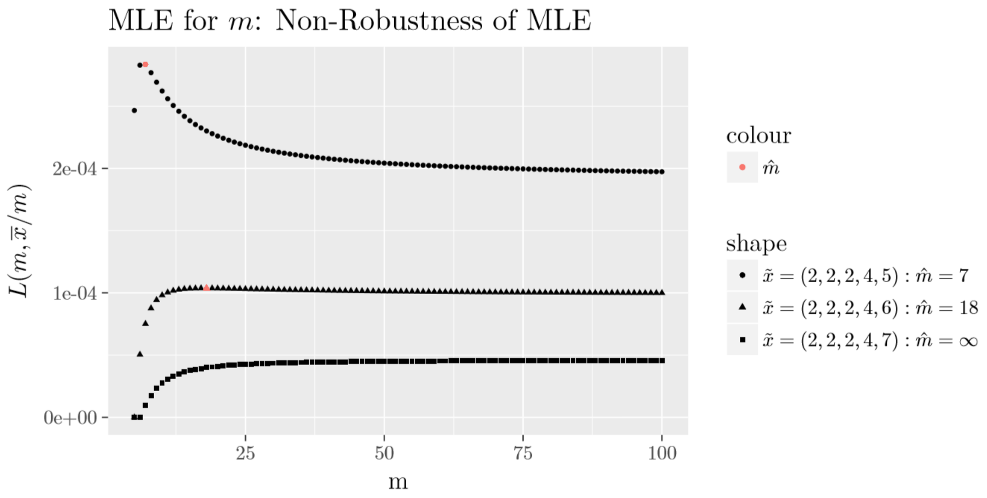


FIGURE 3.2: **Plot of $L(m, \bar{x}/m)$ for a Binomial Distribution**

more complex. We begin by observing that the likelihood can be expressed in the following form:

$$L(r, \beta) = \left(\prod_{i=1}^{n} \binom{r + x_i - 1}{x_i} \binom{r + x_i - 1}{x_i} \right) (1 + \beta)^{-n(r+\overline{x})} \beta^{n\overline{x}}.$$

The above implies that log-likelihood is given by

$$l(r, \beta) = \sum_{i=1}^{n} \log \binom{r + x_i - 1}{x_i} - n(r + \overline{x}) \log(1 + \beta) + n\overline{x} \log \beta,$$

and hence

$$\frac{\delta}{\delta \beta} l(r, \beta) = -\frac{n(r + \overline{x})}{1 + \beta} + \frac{n\overline{x}}{\beta}.$$

Equating the above to zero, we get

$$\hat{r}_{MLE} \times \hat{\beta}_{MLE} = \overline{x}.$$

The above reduces the two dimensional optimization problem to a one-dimensional problem - we need to maximize

$$l(r, \overline{x}/r) = \sum_{i=1}^{n} \log \binom{r + x_i - 1}{x_i} - n(r + \overline{x}) \log(1 + \overline{x}/r) + n\overline{x} \log(\overline{x}/r),$$

with respect to r, with the maximizing r being its *mle* and $\hat{\beta}_{MLE} = \overline{x}/\hat{r}_{MLE}$. In Levin et al. (1977) it is shown that if the sample variance is greater than the sample mean then there exists a unique $r > 0$ that maximizes $l(r, \overline{x}/r)$

and hence a unique *mle* for r and β. Also, they show that if $\hat{\sigma}^2 \leq \overline{x}$, then the negative binomial likelihood will be dominated by the Poisson likelihood with $\hat{\lambda} = \overline{x}$. In other words, a Poisson distribution offers a better fit to the data. The guarantee in the case of $\hat{\sigma}^2 > \hat{\mu}$ permits us to use some algorithm to maximize $l(r, \overline{x}/r)$. Towards an alternate method of computing the likelihood, we note that

$$\begin{aligned} l(r, \overline{x}/r) \quad &= \textstyle\sum_{i=1}^{n} \sum_{j=1}^{x_i} \log(r-1+j) - \sum_{i=1}^{n} \log(x_i!) \\ &\quad -n(r+\overline{x}) \log(r+\overline{x}) + nr \log(r) + n\overline{x} \log(\overline{x}), \end{aligned}$$

which yields

$$\left(\frac{1}{n}\right) \frac{\delta}{\delta r} l(r, \overline{x}/r) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{x_i} \frac{1}{r-1+j} - \log(r+\overline{x}) + \log(r).$$

We note that, in the above expressions for the terms involving a double summation, the inner sum equals zero if $x_i = 0$. The *maximum likelihood estimate* for r is a root of the last expression and we can use a root finding algorithm to compute it. Also, we have

$$\left(\frac{1}{n}\right) \frac{\delta^2}{\delta r^2} l(r, \overline{x}/r) = \frac{\overline{x}}{r(r+\overline{x})} - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{x_i} \frac{1}{(r-1+j)^2}.$$

A simple but quickly converging iterative root finding algorithm is the Newton's method, which incidentally the Babylonians are believed to have used for computing square roots. Under this method, an initial approximation is selected for the root and new approximations for the root are successively generated until convergence. Applying the Newton's method to our problem results in the following algorithm:
Step i. Choose an approximate solution, say r_0. Set k to 0.
Step ii. Define r_{k+1} as

$$r_{k+1} = r_k - \frac{\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{x_i} \frac{1}{r_k-1+j} - \log(r_k+\overline{x}) + \log(r_k)}{\frac{\overline{x}}{r_k(r_k+\overline{x})} - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{x_i} \frac{1}{(r_k-1+j)^2}}$$

Step iii. If $r_{k+1} \sim r_k$, then report r_{k+1} as *maximum likelihood estimate*; else increment k by 1 and repeat *Step ii.*

For example, we simulated a 5 observation sample of $41, 49, 40, 27, 23$ from the negative binomial with parameters $r = 10$ and $\beta = 5$. Choosing the starting value of r such that

$$r\beta = \hat{\mu} \quad \text{and} \quad r\beta(1+\beta) = \hat{\sigma}^2$$

where $\hat{\mu}$ represents the estimated mean and $\hat{\sigma}^2$ is the estimated variance. This leads to the starting value for r of 23.14286. The iterates of r from the Newton's method are

$$21.39627, 21.60287, 21.60647, 21.60647;$$

the rapid convergence seen above is typical of the Newton's method. Hence in this example, $\hat{r}_{MLE} \sim 21.60647$ and $\hat{\beta}_{MLE} = 1.66616$.

```
Newton <- function(x, abserr) {
    mu <- mean(x)
    sigma2 <- mean(x^2) - mu^2
    r <- mu^2/(sigma2 - mu)
    b <- TRUE
    iter <- 0
    while (b) {
        tr <- r
        m1 <- mean(c(x[x == 0], sapply(x[x > 0], function(z) {
            sum(1/(tr:(tr - 1 + z)))
        })))
        m2 <- mean(c(x[x == 0], sapply(x[x > 0], function(z) {
            sum(1/(tr:(tr - 1 + z))^2)
        })))
        r <- tr - (m1 - log(1 + mu/tr))/(mu/(tr * (tr + mu)) - m2)
        b <- !(abs(tr - r) < abserr)
        iter <- iter + 1
    }
    c(r, iter)
}
```

To summarize our discussion of MLE for the $(a, b, 0)$ class of distributions, in Figure 3.3 below we plot the maximum value of the Poisson likelihood, $L(m, \overline{x}/m)$ for the binomial, and $L(r, \overline{x}/r)$ for the negative binomial, for the three samples of size 5 given in Table 3.1. The data was constructed to cover the three orderings of the sample mean and variance. As shown in the Figure 3.3, and supported by theory, if $\hat{\mu} < \hat{\sigma}^2$ then the negative binomial results in a higher maximum likelihood value; if $\hat{\mu} = \hat{\sigma}^2$ the Poisson has the highest likelihood value; and finally in the case that $\hat{\mu} > \hat{\sigma}^2$ the binomial gives a better fit than the others. So before fitting a frequency data with an $(a, b, 0)$ distribution, it is best to start with examining the ordering of $\hat{\mu}$ and $\hat{\sigma}^2$. We again emphasize that the Poisson is on the **boundary** of the negative binomial and binomial distributions. So in the case that $\hat{\mu} \geq \hat{\sigma}^2$ ($\hat{\mu} \leq \hat{\sigma}^2$, resp.) the Poisson yields a better fit than the negative binomial (binomial, resp.), which is indicated by $\hat{r} = \infty$ ($\hat{m} = \infty$, respectively).

Table 3.1. **Three Samples of Size 5**

Data	Mean ($\hat{\mu}$)	Variance ($\hat{\sigma}^2$)
$(2, 3, 6, 8, 9)$	5.60	7.44
$(2, 5, 6, 8, 9)$	6	6
$(4, 7, 8, 10, 11)$	8	6

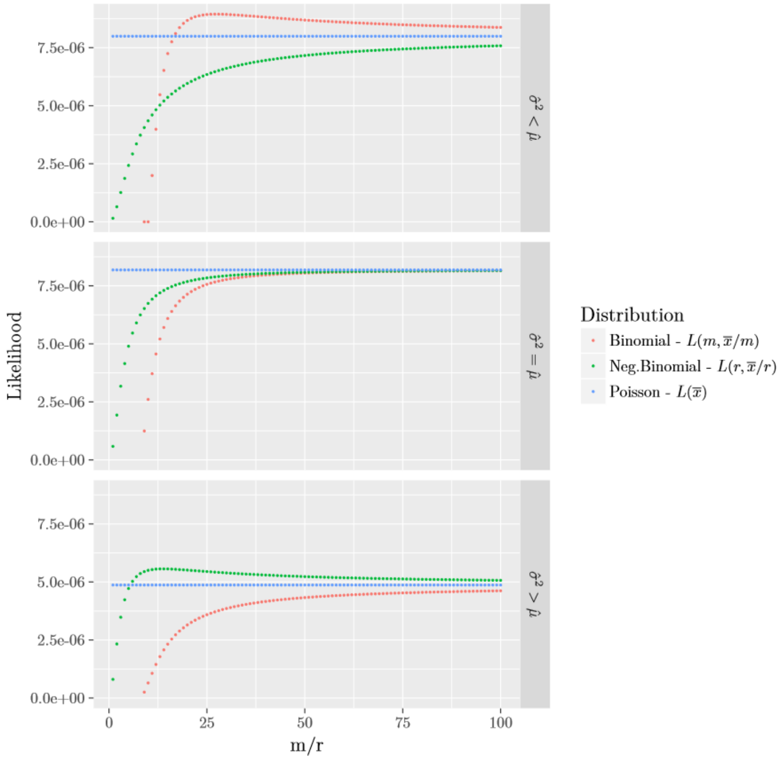


FIGURE 3.3: **Plot of $(a, b, 0)$ Partially Maximized Likelihoods**

3.5 Other Frequency Distributions

In this section, you learn how to:

- Define the (a,b,1) class of frequency distributions and discuss the importance of the recursive relationship underpinning this class of distributions
- Interpret zero truncated and modified versions of the binomial, Poisson, and negative binomial distributions

- Compute probabilities using the recursive relationship

In the previous sections, we discussed three distributions with supports contained in the set of non-negative integers, which well cater to many insurance applications. Moreover, typically by allowing the parameters to be a function of known (to the insurer) explanatory variables such as age, sex, geographic location (territory), and so forth, these distributions allow us to explain claim probabilities in terms of these variables. The field of statistical study that studies such models is known as regression analysis - it is an important topic of actuarial interest that we will not pursue in this book; see Frees (2009).

There are clearly infinitely many other count distributions, and more importantly the above distributions by themselves do not cater to all practical needs. In particular, one feature of some insurance data is that the proportion of zero counts can be out of place with the proportion of other counts to be explainable by the above distributions. In the following we modify the above distributions to allow for arbitrary probability for zero count irrespective of the assignment of relative probabilities for the other counts. Another feature of a data set which is naturally comprised of homogeneous subsets is that while the above distributions may provide good fits to each subset, they may fail to do so to the whole data set. Later we naturally extend the $(a, b, 0)$ distributions to be able to cater to, in particular, such data sets.

3.5.1 Zero Truncation or Modification

Let us suppose that we are looking at auto insurance policies which appear in a database of auto claims made in a certain period. If one is to study the number of claims that these policies have made during this period, then clearly the distribution has to assign a probability of zero to the count variable assuming the value zero. In other words, by restricting attention to count data from policies in the database of claims, we have in a sense zero-truncated the count data of all policies. In personal lines (like auto), policyholders may not want to report that first claim because of fear that it may increase future insurance rates - this behavior inflates the proportion of zero counts. Examples such as the latter modify the proportion of zero counts. Interestingly, natural modifications of the three distributions considered above are able to provide good fits to zero-modified/truncated data sets arising in insurance.

As presented below, we modify the probability assigned to zero count by the $(a, b, 0)$ class while maintaining the relative probabilities assigned to non-zero counts - zero modification. Note that since the $(a, b, 0)$ class of distributions satisfies the recurrence (3.1), maintaining relative probabilities of non-zero counts implies that recurrence (3.1) is satisfied for $k \geq 2$. This leads to the definition of the following class of distributions.

Definition. A count distribution is a member of the $(a, b, 1)$ class if for some constants a and b the probabilities p_k satisfy

$$\frac{p_k}{p_{k-1}} = a + \frac{b}{k}, \quad k \geq 2. \tag{3.5}$$

Note that since the recursion starts with p_1, and not p_0, we refer to this super-class of $(a, b, 0)$ distributions by (a,b,1). To understand this class, recall that each valid pair of values for a and b of the $(a, b, 0)$ class corresponds to a unique vector of probabilities $\{p_k\}_{k\geq 0}$. If we now look at the probability vector $\{\tilde{p}_k\}_{k\geq 0}$ given by

$$\tilde{p}_k = \frac{1-\tilde{p}_0}{1-p_0} \cdot p_k, \quad k \geq 1,$$

where $\tilde{p}_0 \in [0, 1)$ is arbitrarily chosen, then since the relative probabilities for positive values according to $\{p_k\}_{k\geq 0}$ and $\{\tilde{p}_k\}_{k\geq 0}$ are the same, we have $\{\tilde{p}_k\}_{k\geq 0}$ satisfies recurrence (3.5). This, in particular, shows that the class of $(a, b, 1)$ distributions is strictly wider than that of $(a, b, 0)$.

In the above, we started with a pair of values for a and b that led to a valid $(a, b, 0)$ distribution, and then looked at the $(a, b, 1)$ distributions that corresponded to this $(a, b, 0)$ distribution. We now argue that the $(a, b, 1)$ class allows for a larger set of permissible distributions for a and b than the $(a, b, 0)$ class. Recall from Section 3.3 that in the case of $a < 0$ we did not use the fact that the recurrence (3.1) started at $k = 1$, and hence the set of pairs (a, b) with $a < 0$ that are permissible for the $(a, b, 0)$ class is identical to those that are permissible for the $(a, b, 1)$ class. The same conclusion is easily drawn for pairs with $a = 0$. In the case that $a > 0$, instead of the constraint $a + b > 0$ for the $(a, b, 0)$ class we now have the weaker constraint of $a + b/2 > 0$ for the $(a, b, 1)$ class. With the parametrization $b = (r - 1)a$ as used in Section 3.3, instead of $r > 0$ we now have the weaker constraint of $r > -1$. In particular, we see that while zero modifying a $(a, b, 0)$ distribution leads to a distribution in the $(a, b, 1)$ class, the conclusion does not hold in the other direction.

Zero modification of a count distribution F such that it assigns zero probability to zero count is called a zero truncation of F. Hence, the zero truncated version of probabilities $\{p_k\}_{k\geq 0}$ is given by

$$\tilde{p}_k = \begin{cases} 0, & k = 0; \\ \frac{p_k}{1-p_0}, & k \geq 1. \end{cases}$$

In particular, we have that a zero modification of a count distribution $\{p_k^T\}_{k\geq 0}$, denoted by $\{p_k^M\}_{k\geq 0}$, can be written as a convex combination of the degenerate distribution at 0 and the zero truncation of $\{p_k\}_{k\geq 0}$, denoted by $\{p_k^T\}_{k\geq 0}$. That is we have

$$p_k^M = p_0^M \cdot \delta_0(k) + (1-p_0^M) \cdot p_k^T, \quad k \geq 0.$$

Example 3.5.1. Zero Truncated/Modified Poisson. Consider a Poisson distribution with parameter $\lambda = 2$. Calculate $p_k, k = 0, 1, 2, 3$, for the usual (unmodified), truncated and a modified version with $(p_0^M = 0.6)$.

Example Solution. For the Poisson distribution as a member of the $(a, b, 0)$ class, we have $a = 0$ and $b = \lambda = 2$. Thus, we may use the recursion $p_k = \lambda p_{k-1}/k = 2p_{k-1}/k$ for each type, after determining starting probabilities. The calculation of probabilities for $k \leq 3$ is shown in the following table.

Table. **Calculation of Probabilities for** $k \leq 3$

k	p_k	p_k^T	p_k^M
0	$p_0 = e^{-\lambda} = 0.135335$	0	0.6
1	$p_1 = p_0(0 + \frac{\lambda}{1}) = 0.27067$	$\frac{p_1}{1-p_0} = 0.313035$	$\frac{1-p_0^M}{1-p_0}\ p_1 = 0.125214$
2	$p_2 = p_1 \left(\frac{\lambda}{2}\right) = 0.27067$	$p_2^T = p_1^T \left(\frac{\lambda}{2}\right) = 0.313035$	$p_2^M = p_1^M \left(\frac{\lambda}{2}\right) = 0.125214$
3	$p_3 = p_2 \left(\frac{\lambda}{3}\right) = 0.180447$	$p_3^T = p_2^T \left(\frac{\lambda}{3}\right) = 0.208690$	$p_3^M = p_2^M \left(\frac{\lambda}{3}\right) = 0.083476$

3.6 Mixture Distributions

In this section, you learn how to:

- Define a mixture distribution when the mixing component is based on a finite number of sub-groups
- Compute mixture distribution probabilities from mixing proportions and knowledge of the distribution of each subgroup
- Define a mixture distribution when the mixing component is continuous

In many applications, the underlying population consists of naturally defined sub-groups with some homogeneity within each sub-group. In such cases it is convenient to model the individual sub-groups, and in a ground-up manner model the whole population. As we shall see below, beyond the aesthetic appeal of the approach, it also extends the range of applications that can be catered to by standard parametric distributions.

Let k denote the number of defined sub-groups in a population, and let F_i denote the distribution of an observation drawn from the i-th subgroup. If we let α_i denote the proportion of the population in the i-th subgroup, with $\sum_{i=1}^{k} \alpha_i = 1$, then the distribution of a randomly chosen observation from the population, denoted by F, is given by

$$F(x) = \sum_{i=1}^{k} \alpha_i \cdot F_i(x). \tag{3.6}$$

The above expression can be seen as a direct application of the Law of Total Probability. As an example, consider a population of drivers split broadly into two sub-groups, those with at most five years of driving experience and those with more than five years experience. Let α denote the proportion of drivers with less than 5 years experience, and $F_{\leq 5}$ and $F_{>5}$ denote the distribution of the count of claims in a year for a driver in each group, respectively. Then the distribution of claim count of a randomly selected driver is given by

$$\alpha \cdot F_{\leq 5}(x) + (1-\alpha) F_{>5}(x).$$

An alternate definition of a mixture distribution is as follows. Let N_i be a random variable with distribution F_i, $i = 1, \ldots, k$. Let I be a random variable taking values $1, 2, \ldots, k$ with probabilities $\alpha_1, \ldots, \alpha_k$, respectively. Then the random variable N_I has a distribution given by equation (3.6)[7].

In (3.6) we see that the distribution function is a convex combination of the component distribution functions. This result easily extends to the probability mass function, the survival function, the raw moments, and the expectation as these are all linear mappings of the distribution function. We note that this is not true for central moments like the variance, and conditional measures like the hazard rate function. In the case of variance it is easily seen as

$$\mathrm{Var}[N_I] = \mathrm{E}[\mathrm{Var}[N_I|I]] + \mathrm{Var}[\mathrm{E}[N_I|I]] = \sum_{i=1}^{k} \alpha_i \mathrm{Var}[N_i] + \mathrm{Var}[\mathrm{E}[N_I|I]]. \tag{3.7}$$

Appendix Chapter 18 provides additional background about this important expression.

Example 3.6.1. Actuarial Exam Question. In a certain town the number

[7] This in particular lays out a way to simulate from a mixture distribution that makes use of efficient simulation schemes that may exist for the component distributions.

of common colds an individual will get in a year follows a Poisson distribution that depends on the individual's age and smoking status. The distribution of the population and the mean number of colds are as follows:

Table 3.2. **The Distribution of the Population and the Mean Number of Colds**

	Proportion of population	Mean number of colds
Children	0.3	3
Adult Non-Smokers	0.6	1
Adult Smokers	0.1	4

1. Calculate the probability that a randomly drawn person has 3 common colds in a year.
2. Calculate the conditional probability that a person with exactly 3 common colds in a year is an adult smoker.

Example Solution.

1. Using Law of Total Probability, we can write the required probability as $\Pr(N_I = 3)$, with I denoting the group of the randomly selected individual with $1, 2$ and 3 signifying the groups *Children*, *Adult Non-Smoker*, and *Adult Smoker*, respectively. Now by conditioning we get

$$\Pr(N_I = 3) = 0.3 \cdot \Pr(N_1 = 3) + 0.6 \cdot \Pr(N_2 = 3) + 0.1 \cdot \Pr(N_3 = 3),$$

with N_1, N_2 and N_3 following Poisson distributions with means $3, 1$, and 4, respectively. Using the above, we get $\Pr(N_I = 3) \sim 0.1235$ 2. The conditional probability of event A given event B, $\Pr(A|B) = \frac{\Pr(A,B)}{\Pr(B)}$. The required conditional probability in this problem can then be written as $\Pr(I = 3|N_I = 3)$, which equals

$$\Pr(I = 3|N_I = 3) = \frac{\Pr(I = 3, N_3 = 3)}{\Pr(N_I = 3)} \sim \frac{0.1 \times 0.1954}{0.1235} \sim 0.1581.$$

In the above example, the number of subgroups k was equal to three. In general, k can be any natural number, but when k is large it is parsimonious from a modeling point of view to take the following *infinitely many subgroup* approach. To motivate this approach, let the i-th subgroup be such that its component distribution F_i is given by $G_{\tilde{\theta}_i}$, where $G_{\cdot}$ is a parametric family of distributions with parameter space $\Theta \subseteq \mathbb{R}^d$. With this assumption, the distribution function

F of a randomly drawn observation from the population is given by

$$F(x) = \sum_{i=1}^{k} \alpha_i G_{\tilde{\theta}_i}(x), \quad \forall x \in \mathbb{R},$$

similar to equation (3.6). Alternately, it can be written as

$$F(x) = \mathrm{E}[G_{\tilde{\vartheta}}(x)], \quad \forall x \in \mathbb{R},$$

where $\tilde{\vartheta}$ takes values $\tilde{\theta}_i$ with probability α_i, for $i = 1, \ldots, k$. The above makes it clear that when k is large, one could model the above by treating $\tilde{\vartheta}$ as continuous random variable.

To illustrate this approach, suppose we have a population of drivers with the distribution of claims for an individual driver being distributed as a Poisson. Each person has their own (personal) expected number of claims λ - smaller values for good drivers, and larger values for others. There is a distribution of λ in the population; a popular and convenient choice for modeling this distribution is a gamma distribution with parameters (α, θ) (the gamma distribution will be introduced formally in Section 4.2.1). With these specifications it turns out that the resulting distribution of N, the claims of a randomly chosen driver, is a negative binomial with parameters $(r = \alpha, \beta = \theta)$. This can be shown in many ways, but a straightforward argument is as follows:

$$\begin{aligned} \Pr(N=k) &= \int_0^\infty \tfrac{e^{-\lambda}\lambda^k}{k!} \tfrac{\lambda^{\alpha-1}e^{-\lambda/\theta}}{\Gamma(\alpha)\theta^\alpha} d\lambda = \tfrac{1}{k!\Gamma(\alpha)\theta^\alpha} \int_0^\infty \lambda^{\alpha+k-1} e^{-\lambda(1+1/\theta)}\, d\lambda \\ &= \tfrac{\Gamma(\alpha+k)}{k!\Gamma(\alpha)\theta^\alpha(1+1/\theta)^{\alpha+k}} \\ &= \binom{\alpha+k-1}{k} \left(\tfrac{1}{1+\theta}\right)^\alpha \left(\tfrac{\theta}{1+\theta}\right)^k, \quad k = 0, 1, \ldots \end{aligned}$$

Note that the above derivation implicitly uses the following:

$$f_{N|\Lambda=\lambda}(N=k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k \geq 0; \quad \text{and} \quad f_\Lambda(\lambda) = \frac{\lambda^{\alpha-1}e^{-\lambda/\theta}}{\Gamma(\alpha)\theta^\alpha}, \quad \lambda > 0.$$

By considering mixtures of a parametric class of distributions, we increase the richness of the class. This expansion of distributions results in the mixture class being able to cater well to more applications that the parametric class we started with. Mixture modeling is an important modeling technique in insurance applications and later chapters will cover more aspects of this modeling technique.

Example 3.6.2. Suppose that $N|\Lambda \sim$ Poisson(Λ) and that $\Lambda \sim$ gamma with mean of 1 and variance of 2. Determine the probability that $N = 1$.

Example Solution. For a gamma distribution with parameters (α, θ), we have that the mean is $\alpha\theta$ and the variance is $\alpha\theta^2$. Using these expressions we have

$$\alpha = \frac{1}{2} \text{ and } \theta = 2.$$

Now, one can directly use the above result to conclude that N is distributed as a negative binomial with $r = \alpha = \frac{1}{2}$ and $\beta = \theta = 2$. Thus

$$\begin{aligned}\Pr(N = 1) &= \binom{1 + r - 1}{1} (\frac{1}{(1+\beta)^r}) \left(\frac{\beta}{1+\beta}\right)^1 \\ &= \binom{1 + \frac{1}{2} - 1}{1} \frac{1}{(1+2)^{1/2}} \left(\frac{2}{1+2}\right)^1 \\ &= \frac{1}{3^{3/2}} = 0.19245.\end{aligned}$$

3.7 Real Data Example

In this section, you learn how to:

- Compare a fitted distribution to empirical data to assess the adequacy of the fit

In the above, we have discussed three basic frequency distributions, along with their extensions through zero modification/truncation, and by looking at mixtures of these distributions. Nevertheless, these classes remain parametric and hence a small subset of the class of all possible frequency distributions (that is, the set of distributions on non-negative integers). Hence, even though we have discussed methods for estimating the unknown parameters, the *fitted* distribution need not be a good representation of the underlying distribution if the latter is **far** from the class of distribution used for modeling.

While the class of distributions considered above is relatively narrow, via a real example, we present some evidence that they serve insurance purposes quite well.

In 1993, a portfolio of $n = 7,483$ automobile insurance policies from a major

Singaporean insurance company had the distribution of auto accidents per policyholder as given in Table 3.3.

Table 3.3. **Singaporean Automobile Accident Data**

Count (k)	0	1	2	3	4	Total
No. of Policies with k accidents (m_k)	$6,996$	455	28	4	0	$7,483$

If we a fit a Poisson distribution, then the *mle* for λ, the Poisson mean, is the sample mean which is given by

$$\overline{N} = \frac{0 \cdot 6996 + 1 \cdot 455 + 2 \cdot 28 + 3 \cdot 4 + 4 \cdot 0}{7483} = 0.06989.$$

Now if we use Poisson ($\hat{\lambda}_{MLE}$) as the fitted distribution, then a tabular comparison of the fitted counts and observed counts is given by Table 3.4 below, where $\hat{p}_k$ represents the estimated probabilities under the fitted Poisson distribution.

Table 3.4. **Comparison of Observed to Fitted Counts: Singaporean Auto Data**

Count (k)	Observed (m_k)	Fitted Counts Using Poisson ($n\hat{p}_k$)
0	$6,996$	$6,977.86$
1	455	487.70
2	28	17.04
3	4	0.40
≥ 4	0	0.01
Total	$7,483$	$7,483.00$

Notice that the fit seems *quite reasonable* from the above tabular comparison, suggesting that the Poisson distribution is a good model of the underlying distribution. Nevertheless, it is worth pointing out that such a tabular comparison falls short of a statistical test of the hypothesis that the underlying distribution is indeed Poisson. In Section 6.1.2, we present Pearson's chi-square statistic as a goodness-of-fit statistical measure for this purpose.

3.8 Exercises

Theoretical Exercises

Exercise 3.1. Derive an expression for $p_N(\cdot)$ in terms of $F_N(\cdot)$ and $S_N(\cdot)$.

Exercise 3.2. A measure of center of location must be **equi-variant** with respect to shifts, or location transformations. In other words, if N_1 and N_2 are two random variables such that $N_1 + c$ has the same distribution as N_2, for some constant c, then the difference between the measures of the center of location of N_2 and N_1 must equal c. Show that the mean satisfies this property.

Exercise 3.3. Measures of dispersion should be invariant with respect to shifts and scale equi-variant. Show that standard deviation satisfies these properties by doing the following:

- Show that for a random variable N, its standard deviation equals that of $N + c$, for any constant c.
- Show that for a random variable N, its standard deviation equals $1/c$ times that of cN, for any positive constant c.

Exercise 3.4. Let N be a random variable with probability mass function given by

$$p_N(k) = \begin{cases} \left(\frac{6}{\pi^2}\right)\left(\frac{1}{k^2}\right), & k \geq 1; \\ 0, & \text{otherwise.} \end{cases}$$

Show that the mean of N is ∞.

Exercise 3.5. Let N be a random variable with a finite second moment. Show that the function $\psi(\cdot)$ defined by

$$\psi(x) = \mathrm{E}(N - x)^2. \quad x \in \mathbb{R}$$

is minimized at μ_N without using calculus. Also, give a proof of this fact using derivatives. Conclude that the minimum value equals the variance of N.

Exercise 3.6. Derive the first two central moments of the $(a, b, 0)$ distributions using the methods mentioned below:

- For the binomial distribution, derive the moments using only its *pmf*, then its *mgf*, and then its *pgf*.
- For the Poisson distribution, derive the moments using only its *mgf*.
- For the negative binomial distribution, derive the moments using only its *pmf*, and then its *pgf*.

Exercise 3.7. Let N_1 and N_2 be two independent Poisson random variables with means λ_1 and λ_2, respectively. Identify the conditional distribution of N_1 given $N_1 + N_2$.

Exercise 3.8. (**Non-Uniqueness of the MLE**) Consider the following parametric family of densities indexed by the parameter p taking values in $[0, 1]$:

$$f_p(x) = p \cdot \phi(x + 2) + (1 - p) \cdot \phi(x - 2), \quad x \in \mathbb{R},$$

where $\phi(\cdot)$ represents the standard normal density.

- Show that for all $p \in [0,1]$, $f_p(\cdot)$ above is a valid density function.
- Find an expression in p for the mean and the variance of $f_p(\cdot)$.
- Let us consider a sample of size one consisting of x. Show that when x equals 0, the set of *maximum likelihood estimates* for p equals $[0,1]$; also show that the *mle* is unique otherwise.

Exercise 3.9. Graph the region of the plane corresponding to values of (a,b) that give rise to valid $(a,b,0)$ distributions. Do the same for $(a,b,1)$ distributions.

Exercise 3.10. (**Computational Complexity**) For the $(a,b,0)$ class of distributions, count the number of basic mathematical operations (addition, subtraction, multiplication, division) needed to compute the n probabilities $p_0 \ldots p_{n-1}$ using the recurrence relationship. For the negative binomial distribution with non-integer r, count the number of such operations. What do you observe?

Exercise 3.11. Using the development of Section 3.3 rigorously show that not only does the recurrence (3.1) tie the binomial, the Poisson and the negative binomial distributions together, but that it also characterizes them.

Exercise 3.12. Actuarial Exam Question. You are given:

1. p_k denotes the probability that the number of claims equals k for $k = 0, 1, 2, \ldots$
2. $\frac{p_n}{p_m} = \frac{m!}{n!}, m \geq 0, n \geq 0$

Using the corresponding zero-modified claim count distribution with $p_0^M = 0.1$, calculate p_1^M.

Exercises with a Practical Focus

Exercise 3.13. Singaporean Automobile Accident. In this exercise, we replicate and extend the real-data example introduced in Section 3.7 using R.

- **a**. From the package `CASdatasets`, retrieve the data `sgautonb` in order to work with the variable `Clm_Count` which is a count of claims. Refer to Section 22.6 for a description of this package. Verify that the mean claim count is 0.
- **b**. Compute the fitted Poisson distribution and reproduce Table 3.5.
- **c**. Compute the maximum likelihood estimates for the negative binomial distribution. One way to do this is to create a negative logarithmic likelihood function and use the R function `optim` for minimization. Use the resulting

TABLE 3.5: **Singaporean Automobile Comparison of Empirical to Poisson Fitted Percentiles**

Claim Count	Empirical Percentile	Poisson Perc
6996	0.9349	0.9325
455	0.9957	0.9977
28	0.9995	0.9999
4	1.0000	1.0000

maximum likelihood estimates to create a fitted distribution and augment the Table in part (b) with this alternative distribution.

In part (c), you learn that the more complex negative binomial distribution produces roughly the same fits as the simpler Poisson distribution. As a result of this analysis, an analyst would typically prefer the simpler Poisson distribution.

Exercise 3.14. Corporate Travel. This exercise is based on the data set introduced in Exercise 1.1 where now the focus is on frequency modeling. For corporate travel, the number of claims are sufficient that a separate frequency model could be considered. For the frequency of claims, there are 2107 claims over the 2006-2021 period that amounts to 131.69 per year. One might assume that annual claims can be fit using a single distribution to the entire period, such as a Poisson or a negative binomial. Another option is to fit a distribution starting in years 2009, where this is an increase in the amount of claims from prior years. A third option is to omit experience from underwriting year 2019 and on where the number of claims fluctuated dramatically, in part due to the Covid epidemic. In this exercise, we pursue the first option.

- **a**. Fit a Poisson distribution and a negative binomial distribution to all claims.
- **b**. Fit a negative binomial distribution to all claims using the strategy introduced in part (c) of Exercise 3.13.
- **c**. To check your work, use the `fitdist` function from the package `fitdistrplus`, with the negative binomial (`nbinom`) option.
- **d**. Use the `ecdf` function in `R` to produce empirical cumulative probabilities. Produce a table that compares the empirical percentiles to those under the Poisson and negative binomial.

From part (d), you learn that both fitted distributions did well and neither outperformed the other.

3.9 Further Resources and Contributors

Appendix Chapter 17 gives a general introduction to maximum likelihood theory regarding estimation of parameters from a parametric family. Appendix Chapter 19 gives more specific examples and expands some of the concepts.

If you would like additional practice with R coding, please visit our companion LDA Short Course. In particular, see the Frequency Modeling Chapter.

Contributors

- **N.D. Shyamalkumar**, The University of Iowa, and **Krupa Viswanathan**, Temple University, are the principal authors of the initial version and also the second edition of this chapter. Email: shyamal-kumar@uiowa.edu for chapter comments and suggested improvements.
- Chapter reviewers include: Chunsheng Ban, Paul Johnson, Hirokazu (Iwahiro) Iwasawa, Dalia Khalil, Tatjana Miljkovic, Rajesh Sahasrabuddhe, and Michelle Xia.

3.9.1 TS 3.A. R Code for Plots

Code for Figure 3.2:

```
likm<-function(m){
  prod((dbinom(x,m,mean(x)/m)))
}
x<-c(2,2,2,4,5);
n<-(5:100);
# Computing the Likelihood
ll<-sapply(n,likm);
# Computing the MLE
n[ll==max(ll)]
# Storing the Likelihood Curve
y<-cbind(n,ll);

# Second Dataset
x<-c(2,2,2,4,6);
ll<-sapply(n,likm);
n[ll==max(ll)]
y<-cbind(y,ll);

# Third Dataset
x<-c(2,2,2,4,7);
ll<-sapply(n,likm);
n[ll==max(ll)]
y<-cbind(y,ll);

colnames(y)<-c("m","$\\tilde{x}=(2,2,2,4,5)$",
               "$\\tilde{x}=(2,2,2,4,6)$",
               "$\\tilde{x}=(2,2,2,4,7)$");
dy<-data.frame(y);
library(tikzDevice);
library(ggplot2);
options(tikzMetricPackages =
          c("\\usepackage[utf8]{inputenc}","\\usepackage[T1]{fontenc}",
            "\\usetikzlibrary{calc}", "\\usepackage{amssymb}",
            "\\usepackage{amsmath}","\\usepackage[active]{preview}"))
tikz(file = "plot_test.tex", width = 6.25, height = 3.125);
ggplot(dy) +
  geom_point(aes(x=m, y=(X..tilde.x...2.2.2.4.5..),
                 shape="$\\tilde{x}=(2,2,2,4,5):\\hat{m}=7$"), size=0.75) +
  geom_point(aes(x=m, y=(X..tilde.x...2.2.2.4.6..),
                 shape="$\\tilde{x}=(2,2,2,4,6):\\hat{m}=18$"),size=0.75) +
  geom_point(aes(x=m, y=(X..tilde.x...2.2.2.4.7..),
                 shape="$\\tilde{x}=(2,2,2,4,7):\\hat{m}=\\infty$"),size=0.75) +
  geom_point(aes(x=c(7),y=dy$X..tilde.x...2.2.2.4.5..[3],colour="$\\hat{m}$",
                 shape="$\\tilde{x}=(2,2,2,4,5):\\hat{m}=7$"),size=0.75) +
  geom_point(aes(x=c(18),y=dy$X..tilde.x...2.2.2.4.6..[14],colour="$\\hat{m}$",
                 shape="$\\tilde{x}=(2,2,2,4,6):\\hat{m}=18$"),size=0.75)+
  labs(x="m",y="$L(m,\\overline{x}/m)$",
       title="MLE for $m$: Non-Robustness of MLE ");
dev.off();
```

Code for Figure 3.3:

```
likbinm<-function(m){
  # binomial likelihood maximized w.r.t. p
  prod((dbinom(x,m,mean(x)/m)))
  }

liknbinm<-function(r){
  # negative binomial likelihood maximized w.r.t. beta
  prod(dnbinom(x,r,1-mean(x)/(mean(x)+r)))
  }

# Data Matrix; Three Samples, one in each Column;
# First Sample has Var<Mean
# Second Sample has Var=Mean
# Third Sample has Var>Mean

  X<-cbind(c(2,5,6,8,9)+2,c(2,5,6,8,9),c(2,3,6,8,9))

# Used for creating the labels in the z matrix
  ord_char<-c("<","=",">")

# Empty matrices;
  Y<-matrix(1,ncol=2,nrow=0)
  Z<-matrix(1,ncol=2,nrow=0)

for (i in (1:3)) {
  # Work with data in the i-th sample
    x<-X[,i]

  # Binomial Likelihood
      # Interval of n values covering the MLE
        n<-(9:100)
      # Evaluating the Likelihood at various values of n
        ll<-sapply(n,likbinm)
      # Finding the MLE of n
        n[ll==max(ll[!is.na(ll)])]
      # Storing the data and the labels
        Y<-rbind(Y,cbind(n,ll))
        Z<-rbind(Z,cbind(rep(paste("$\\hat{\\sigma}^2",ord_char[i],
                "\\hat{\\mu}$"),length(n)),rep("Binomial - L(m,\\overline{x}/m)$",
                length(n))))

  # Negative Binomial Likelihood
    # Interval of r values
      r<-(1:100)
    # Evaluating the Likelihood at various values of r
      ll<-sapply(r,liknbinm)
    # Finding the MLE of r
      ll[is.na(ll)]=0
      r[ll==max(ll[!is.na(ll)])]
    # Storing the data and the labels
      Y<-rbind(Y,cbind(r,ll))
      Z<-rbind(Z,cbind(rep(paste("$\\hat{\\sigma}^2",ord_char[i],
              "\\hat{\\mu}$"),length(r)),rep("Neg.Binomial -$L(r,\\overline{x}/r)$",
                                             length(r))))

  # Poisson Likelihood
    # Storing the data and the labels
    # In the Poisson case MLE is the sample mean
      Y<-rbind(Y,cbind(r,rep(prod(dpois(x,mean(x))),length(r))))
      Z<-rbind(Z,cbind(rep(paste("$\\hat{\\sigma}^2",ord_char[i],"\\hat{\\mu}$"),
          length(r)),rep("Poisson - $L(\\overline{x})$",length(r))))
  }

  # Assigning Column Names
    colnames(Y)<-c("x","lik")
    colnames(Z)<-c("dataset","Distribution")
  # Creating a Dataframe for using ggplot
    dy<-cbind(data.frame(Y),data.frame(Z))

library(tikzDevice)
library(ggplot2)
options(tikzMetricPackages = c("\\usepackage[utf8]{inputenc}",
                               "\\usepackage[T1]{fontenc}",
                               "\\usetikzlibrary{calc}",
                               "\\usepackage{amssymb}",
                               "\\usepackage{amsmath}",
                               "\\usepackage[active]{preview}") )
tikz(file = "plot_test_2.tex", width = 6.25, height = 6.25)
ggplot(data=dy,aes(x=x,y=lik,col=Distribution)) +
  geom_point(size=0.25) + facet_grid(dataset~.) +
  labs(x="m/r",y="Likelihood",title="")
dev.off()
```

4

Modeling Loss Severity

Chapter Preview. The traditional loss distribution approach to modeling aggregate losses starts by separately fitting a frequency distribution to the number of losses and a severity distribution to the size of losses. The estimated aggregate loss distribution combines the loss frequency distribution and the loss severity distribution by convolution. Discrete distributions often referred to as counting or frequency distributions were used in Chapter 3 to describe the number of events such as number of accidents to the driver or number of claims to the insurer. Lifetimes, asset values, losses and claim sizes are usually modeled as continuous random variables and as such are modeled using continuous distributions, often referred to as loss or severity distributions. A mixture distribution is a weighted combination of simpler distributions that is used to model phenomenon investigated in a heterogeneous population, such as modeling more than one type of claims in liability insurance (small frequent claims and large relatively rare claims). In this chapter we explore the use of continuous as well as mixture distributions to model the random size of loss.

Sections 4.1 to 4.3 present key attributes that characterize continuous models and means of creating new distributions from existing ones. Section 4.4.1 describes some principal non-parametric methods for estimating loss distributions: moment and percentile based, empirical, and density estimation methods. Section 4.4.2 covers parametric estimation methods including method of moments and percentile matching, and deepens our understanding of maximum likelihood methods. The frequency distributions from Chapter 3 will be combined with the ideas from this chapter to describe the aggregate losses over the whole portfolio in Chapter 7.

4.1 Basic Distributional Quantities

In this section, you learn how to define some basic distributional quantities:

- moments,

- moment generating functions, and
- percentiles

4.1.1 Moments and Moment Generating Functions

Let X be a continuous random variable with probability density function (*pdf*) $f_X(x)$ and distribution function $F_X(x)$. The k-th raw moment of X, denoted by μ_k', is the expected value of the k-th power of X, provided it exists. The first raw moment μ_1' is the mean of X usually denoted by μ. The formula for μ_k' is given as

$$\mu_k' = \mathrm{E}\left(X^k\right) = \int_0^\infty x^k f_X(x)\, dx.$$

Note that the notation used here for moments differs from the notation used in Section 3.2.1. The support of the random variable X is assumed to be nonnegative since actuarial phenomena are rarely negative. For example, an easy integration by parts shows that the raw moments for nonnegative variables can also be computed using

$$\mu_k' = \int_0^\infty k\ x^{k-1}\left[1 - F_X(x)\right] dx,$$

that is based on the survival function, denoted as $S_X(x) = 1 - F_X(x)$. This formula is particularly useful when $k = 1$. Section 5.1.2 discusses this approach in more detail.

The k-th central moment of X, denoted by μ_k, is the expected value of the k-th power of the deviation of X from its mean μ. The formula for μ_k is given as

$$\mu_k = \mathrm{E}\left[(X-\mu)^k\right] = \int_0^\infty (x-\mu)^k f_X(x)\, dx.$$

The second central moment μ_2 defines the variance of X, denoted by σ^2. The square root of the variance is the standard deviation σ.

From a classical perspective, further characterization of the shape of the distribution includes its degree of symmetry as well as its flatness compared to the normal distribution. The ratio of the third central moment to the cube of the standard deviation $\left(\mu_3/\sigma^3\right)$ defines the coefficient of skewness which is a measure of symmetry. A positive coefficient of skewness indicates that the distribution is skewed to the right (positively skewed). The ratio of the fourth central moment to the fourth power of the standard deviation $\left(\mu_4/\sigma^4\right)$ defines the coefficient of kurtosis. The normal distribution has a coefficient of kurtosis of 3. Distributions with a coefficient of kurtosis greater than 3 have heavier tails than the normal, whereas distributions with a coefficient of kurtosis less than 3 have lighter tails and are flatter. Section 13.2 describes the tails of distributions from an insurance and actuarial perspective.

Example 4.1.1. Actuarial Exam Question. Assume that the rv X has a gamma distribution with mean 8 and skewness 1. Find the variance of X. (*Hint*: The gamma distribution is reviewed in Section 4.2.1.)

Example Solution. The *pdf* of X is given by

$$f_X\left(x\right) = \frac{\left(x/\theta\right)^{\alpha}}{x\ \Gamma\left(\alpha\right)} e^{-x/\theta}$$

for $x > 0$. For $\alpha > 0$, the k-th raw moment is

$$\mu_k' = \mathrm{E}\left(X^k\right) = \int_0^\infty \frac{1}{\Gamma\left(\alpha\right)\theta^{\alpha}} x^{k+\alpha-1} e^{-x/\theta} dx = \frac{\Gamma\left(k+\alpha\right)}{\Gamma\left(\alpha\right)} \theta^k$$

Given $\Gamma\left(r+1\right) = r\Gamma\left(r\right)$ and $\Gamma\left(1\right) = 1$, then $\mu_1' = \mathrm{E}\left(X\right) = \alpha\theta$, $\mu_2' = \mathrm{E}\left(X^2\right) = \left(\alpha+1\right)\alpha\theta^2$, $\mu_3' = \mathrm{E}\left(X^3\right) = \left(\alpha+2\right)\left(\alpha+1\right)\alpha\theta^3$, and $\mathrm{Var}\left(X\right) = (\alpha+1)\alpha\theta^2 - (\alpha\theta)^2 = \alpha\theta^2$.

$$\begin{aligned} \text{Skewness} &= \frac{\mathrm{E}\left[(X-\mu_1')^3\right]}{(\mathrm{Var}X)^{3/2}} = \frac{\mu_3' - 3\mu_2'\mu_1' + 2{\mu_1'}^3}{(\mathrm{Var}X)^{3/2}} \\ &= \frac{(\alpha+2)(\alpha+1)\alpha\theta^3 - 3(\alpha+1)\alpha^2\theta^3 + 2\alpha^3\theta^3}{(\alpha\theta^2)^{3/2}} \\ &= \frac{2}{\alpha^{1/2}} = 1. \end{aligned}$$

Hence, $\alpha = 4$. Since, $\mathrm{E}\left(X\right) = \alpha\theta = 8$, then $\theta = 2$ and finally, $\mathrm{Var}\left(X\right) = \alpha\theta^2 = 16$.

The moment generating function (mgf), denoted by $M_X(t)$ uniquely characterizes the distribution of X. While it is possible for two different distributions to have the same moments and yet still differ, this is not the case with the moment generating function. That is, if two random variables have the same moment generating function, then they have the same distribution. The moment generating function is given by

$$M_X(t) = \mathrm{E}\left(e^{tX}\right) = \int_0^\infty e^{\mathrm{tx}} f_X\left(x\right) dx$$

for all t for which the expected value exists. The *mgf* is a real function whose k-th derivative at zero is equal to the k-th raw moment of X. In symbols, this is

$$\left.\frac{d^k}{dt^k} M_X(t)\right|_{t=0} = \mathrm{E}\left(X^k\right).$$

Example 4.1.2. Actuarial Exam Question. The random variable X has an exponential distribution with mean $\frac{1}{b}$. It is found that $M_X\left(-b^2\right) = 0.2$. Find b. (*Hint*: The exponential is a special case of the gamma distribution which is reviewed in Section 4.2.1.)

Example Solution. With X having an exponential distribution with mean $\frac{1}{b}$, we have that

$$M_X(t) = \mathrm{E}\left(e^{tX}\right) = \int_0^\infty e^{\mathrm{tx}} be^{-bx} dx = \int_0^\infty be^{-x(b-t)} dx = \frac{b}{(b-t)}.$$

Then,

$$M_X\left(-b^2\right) = \frac{b}{(b+b^2)} = \frac{1}{(1+b)} = 0.2.$$

Thus, $b = 4$.

Example 4.1.3. Actuarial Exam Question. Let $X_1, \ldots, X_n$ be independent random variables, where X_i has a gamma distribution with parameters α_i and θ. Find the distribution of $S = \sum_{i=1}^n X_i$, the mean $\mathrm{E}(S)$, and the variance $\mathrm{Var}(S)$.

Example Solution. The *mgf* of S is

$$M_S(t) = \mathrm{E}\left(e^{\mathrm{tS}}\right) = \mathrm{E}\left(e^{t\sum_{i=1}^n X_i}\right) = \mathrm{E}\left(\prod_{i=1}^n e^{tX_i}\right).$$

Using independence, we get

$$M_S(t) = \prod_{i=1}^n \mathrm{E}\left(e^{tX_i}\right) = \prod_{i=1}^n M_{X_i}(t).$$

The moment generating function of the gamma distribution X_i is $M_{X_i}(t) = (1-\theta t)^{\alpha_i}$. Then,

$$M_S(t) = \prod_{i=1}^n (1-\theta t)^{-\alpha_i} = (1-\theta t)^{-\sum_{i=1}^n \alpha_i}.$$

This indicates that the distribution of S is gamma with parameters $\sum_{i=1}^n \alpha_i$ and θ.

This is a demonstration of how we can use the uniqueness property of the moment generating function to determine the probability distribution of a function of random variables.

We can find the mean and variance from the properties of the gamma distribution. Alternatively, by finding the first and second derivatives of $M_S(t)$ at zero, we can show that $\mathrm{E}\left(S\right) = \left.\frac{\partial M_S(t)}{\partial t}\right|_{t=0} = \alpha\theta$ where $\alpha = \sum_{i=1}^{n} \alpha_i$, and

$$\mathrm{E}\left(S^2\right) = \left.\frac{\partial^2 M_S(t)}{\partial t^2}\right|_{t=0} = \left(\alpha + 1\right)\alpha\theta^2.$$

Hence, $\mathrm{Var}\left(S\right) = \alpha\theta^2$.

One can also use the moment generating function to compute the probability generating function

$$P_X(z) = \mathrm{E}\left(z^X\right) = M_X\left(\log z\right).$$

As introduced in Section 3.2.2, the probability generating function is more useful for discrete random variables.

4.1.2 Quantiles

Quantiles can also be used to describe the characteristics of the distribution of X. When the distribution of X is continuous, for a given fraction $0 \le p \le 1$ the corresponding quantile is the solution of the equation

$$F_X\left(\pi_p\right) = p.$$

For example, the middle point of the distribution, $\pi_{0.5}$, is the median. A percentile is a type of quantile; a $100p$ percentile is the number such that $100 \times p$ percent of the data is below it.

Example 4.1.4. Actuarial Exam Question. Let X be a continuous random variable with density function $f_X\left(x\right) = \theta e^{-\theta x}$, for $x > 0$ and 0 elsewhere. If the median of this distribution is $\frac{1}{3}$, find θ.

Example Solution. The distribution function is $F_X\left(x\right) = 1 - e^{-\theta x}$. So, $F_X\left(\pi_{0.5}\right) = 1 - e^{-\theta\pi_{0.5}} = 0.5$. As, $\pi_{0.5} = \frac{1}{3}$, we have $F_X\left(\frac{1}{3}\right) = 1 - e^{-\theta/3} = 0.5$ and $\theta = 3\log 2$.

Section 4.4.1 extends the definition of quantiles to include distributions that are discrete, continuous, or a hybrid combination.

4.2 Continuous Distributions for Modeling Loss Severity

In this section, you learn how to define and apply four fundamental severity distributions:

- gamma,
- Pareto,
- Weibull, and
- generalized beta distribution of the second kind.

4.2.1 Gamma Distribution

Recall that the traditional approach in modeling losses is to fit separate models for frequency and claim severity. When frequency and severity are modeled separately it is common for actuaries to use the Poisson distribution (introduced in Section 3.2.3) for claim count and the gamma distribution to model severity. An alternative approach for modeling losses that has recently gained popularity is to create a single model for pure premium (average claim cost).

The continuous variable X is said to have the gamma distribution with shape parameter α and scale parameter θ if its probability density function is given by

$$f_X\left(x\right) = \frac{\left(x/\theta\right)^{\alpha}}{x\ \Gamma\left(\alpha\right)} \exp\left(-x/\theta\right) \quad \text{for } x > 0.$$

Note that $\alpha > 0,\ \theta > 0$.

The two panels in Figure 4.1 demonstrate the effect of the scale and shape parameters on the gamma density function.

When $\alpha = 1$ the gamma reduces to an exponential distribution and when $\alpha = \frac{n}{2}$ and $\theta = 2$ the gamma reduces to a chi-square distribution with n degrees of freedom. As we will see in Section 17.4.1, the chi-square distribution is used extensively in statistical hypothesis testing.

The distribution function of the gamma model is the *incomplete gamma function*, denoted by $\Gamma\left(\alpha; \frac{x}{\theta}\right)$, and defined as

$$F_X\left(x\right) = \Gamma\left(\alpha; \frac{x}{\theta}\right) = \frac{1}{\Gamma\left(\alpha\right)} \int_0^{x/\theta} t^{\alpha-1} e^{-t}\, dt,$$

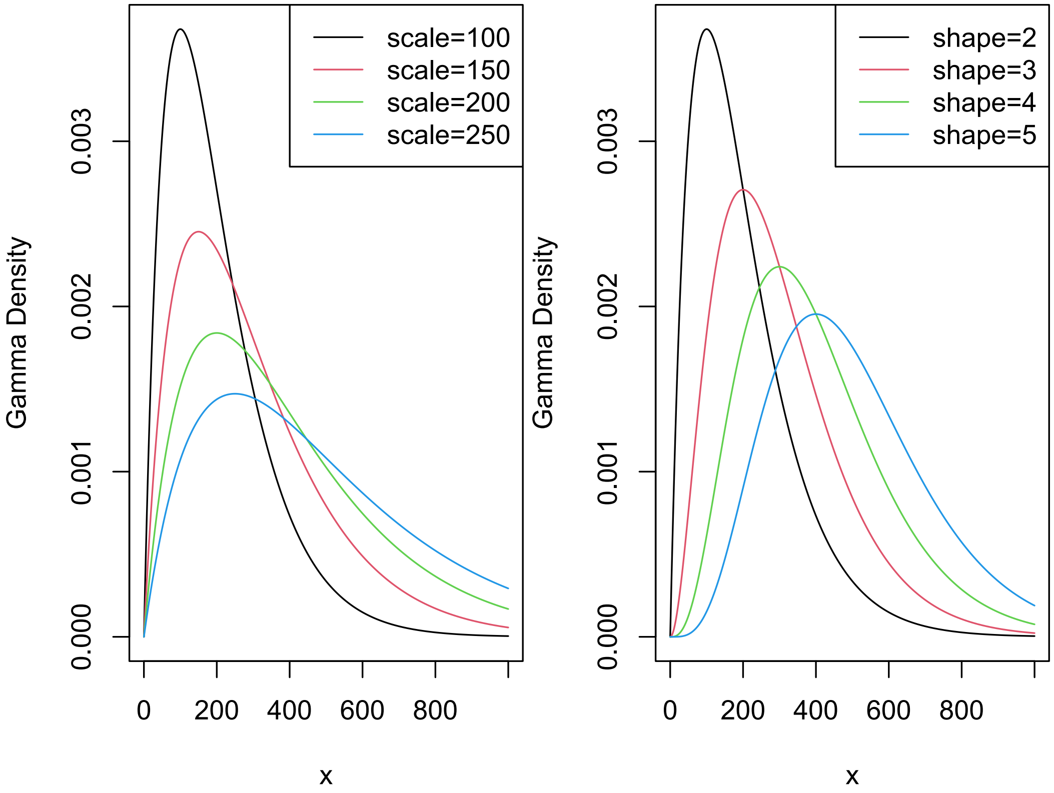


FIGURE 4.1: **Gamma Densities**. The left-hand panel is with shape=2 and varying scale. The right-hand panel is with scale=100 and varying shape.

with $\alpha > 0,\ \theta > 0$. For an integer α, it can be written as $\Gamma\left(\alpha; \frac{x}{\theta}\right) = 1 - e^{-x/\theta} \sum_{k=0}^{\alpha-1} \frac{(x/\theta)^k}{k!}$.

The k-th raw moment of the gamma distributed random variable for any positive k is given by

$$\mathrm{E}\left(X^k\right) = \theta^k \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)}.$$

The mean and variance are given by $\mathrm{E}(X) = \alpha\theta$ and $\mathrm{Var}(X) = \alpha\theta^2$, respectively.

Since all moments exist for any positive k, the gamma distribution is considered a light tailed distribution, which may not be suitable for modeling risky assets as it will not provide a realistic assessment of the likelihood of severe losses.

4.2.2 Pareto Distribution

The Pareto distribution, named after the Italian economist Vilfredo Pareto (1843-1923), has many economic and financial applications. It is a positively skewed and heavy-tailed distribution which makes it suitable for modeling income, high-risk insurance claims and severity of large casualty losses. The survival function of the Pareto distribution which decays slowly to zero was first used to describe the distribution of income where a small percentage of the population holds a large proportion of the total wealth. For extreme insurance claims, the tail of the severity distribution (losses in excess of a threshold) can be modeled using a Generalized Pareto distribution.

The continuous variable X is said to have the (two parameter) Pareto distribution with shape parameter α and scale parameter θ if its pdf is given by

$$f_X(x) = \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}} \quad x > 0,\ \alpha > 0,\ \theta > 0. \tag{4.1}$$

The two panels in Figure 4.2 demonstrate the effect of the scale and shape parameters on the Pareto density function. There are other formulations of the Pareto distribution including a one parameter version given in Appendix Section 20.2. Henceforth, when we refer the Pareto distribution, we mean the version given through the *pdf* in equation (4.1).

The distribution function of the Pareto distribution is given by

$$F_X(x) = 1 - \left(\frac{\theta}{x+\theta}\right)^\alpha \quad x > 0,\ \alpha > 0,\ \theta > 0.$$

It can be easily seen that the hazard function of the Pareto distribution is a decreasing function in x, another indication that the distribution is heavy

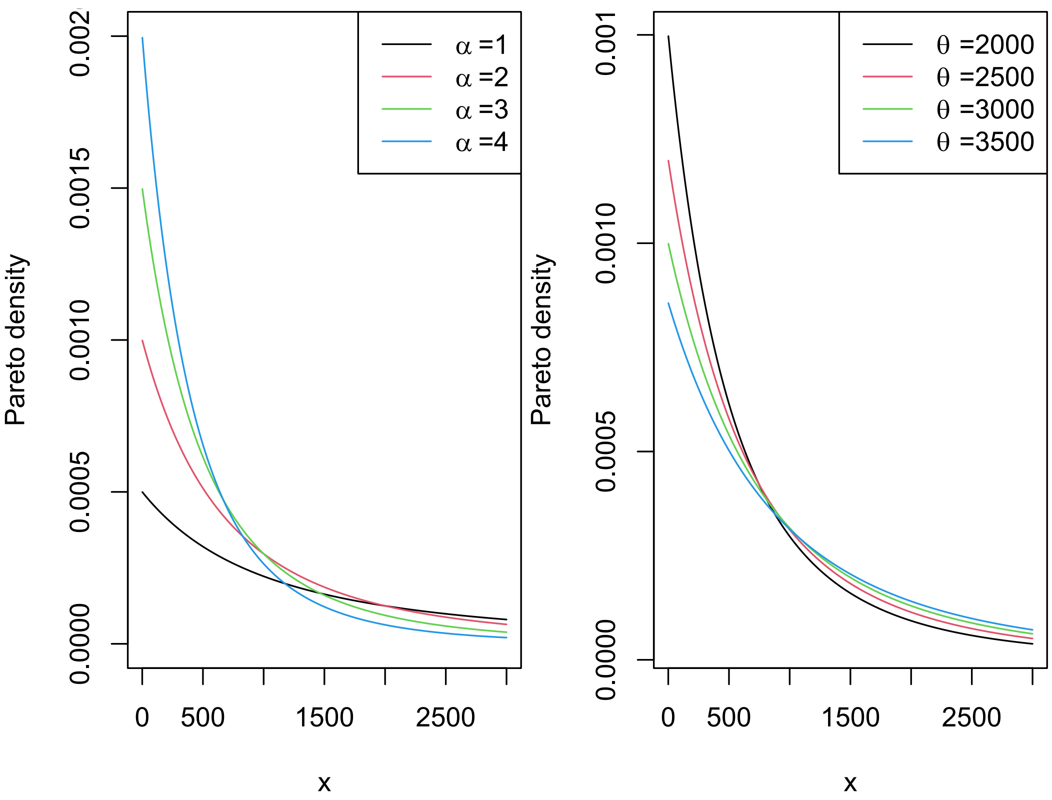


FIGURE 4.2: **Pareto Densities**. The left-hand panel is with scale=2000 and varying shape. The right-hand panel is with shape=3 and varying scale.

tailed. Again using the analogy of the income of a population, when the hazard function decreases over time the population dies off at a decreasing rate resulting in a heavier tail for the distribution. The hazard function reveals information about the tail distribution and is often used to model data distributions in survival analysis. The hazard function is defined as the instantaneous potential that the event of interest occurs within a very narrow time frame.

The k-th raw moment of the Pareto distributed random variable exists, if and only if, $\alpha > k$. If k is a positive integer then

$$\mathrm{E}\left(X^k\right) = \frac{\theta^k\ k!}{(\alpha-1)\cdots(\alpha-k)} \quad \alpha > k.$$

The mean and variance are given by

$$\mathrm{E}\left(X\right) = \frac{\theta}{\alpha-1} \quad \text{for } \alpha > 1$$

and

$$\mathrm{Var}\left(X\right) = \frac{\alpha\theta^2}{\left(\alpha-1\right)^2\left(\alpha-2\right)} \quad \text{for } \alpha > 2,$$

respectively.

Example 4.2.1. The claim size of an insurance portfolio follows the Pareto distribution with mean and variance of 40 and 1800, respectively. Find

a. The shape and scale parameters.
b. The 95-th percentile of this distribution.

Example Solution.

a. As, $X \sim Pa(\alpha, \theta)$, we have $\mathrm{E}\left(X\right) = \frac{\theta}{\alpha-1} = 40$ and $\mathrm{Var}\left(X\right) = \frac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)} = 1800$. By dividing the square of the first equation by the second we get $\frac{\alpha-2}{\alpha} = \frac{40^2}{1800}$. Thus, $\alpha = 18.02$ and $\theta = 680.72$.

b. The 95-th percentile, $\pi_{0.95}$, satisfies the equation

$$F_X\left(\pi_{0.95}\right) = 1 - \left(\frac{680.72}{\pi_{0.95} + 680.72}\right)^{18.02} = 0.95.$$

Thus, $\pi_{0.95} = 122.96$.

4.2.3 Weibull Distribution

The Weibull distribution, named after the Swedish physicist Waloddi Weibull (1887-1979) is widely used in reliability, life data analysis, weather forecasts and general insurance claims. Truncated data arise frequently in insurance studies. The Weibull distribution has been used to model excess of loss treaty over automobile insurance as well as earthquake inter-arrival times.

The continuous variable X is said to have the Weibull distribution with shape parameter α and scale parameter θ if its *pdf* is given by

$$f_X(x) = \frac{\alpha}{\theta}\left(\frac{x}{\theta}\right)^{\alpha-1}\exp\left(-\left(\frac{x}{\theta}\right)^{\alpha}\right) \quad x>0,\ \alpha>0,\ \theta>0.$$

The two panels in Figure 4.3 demonstrate the effects of the scale and shape parameters on the Weibull density function.

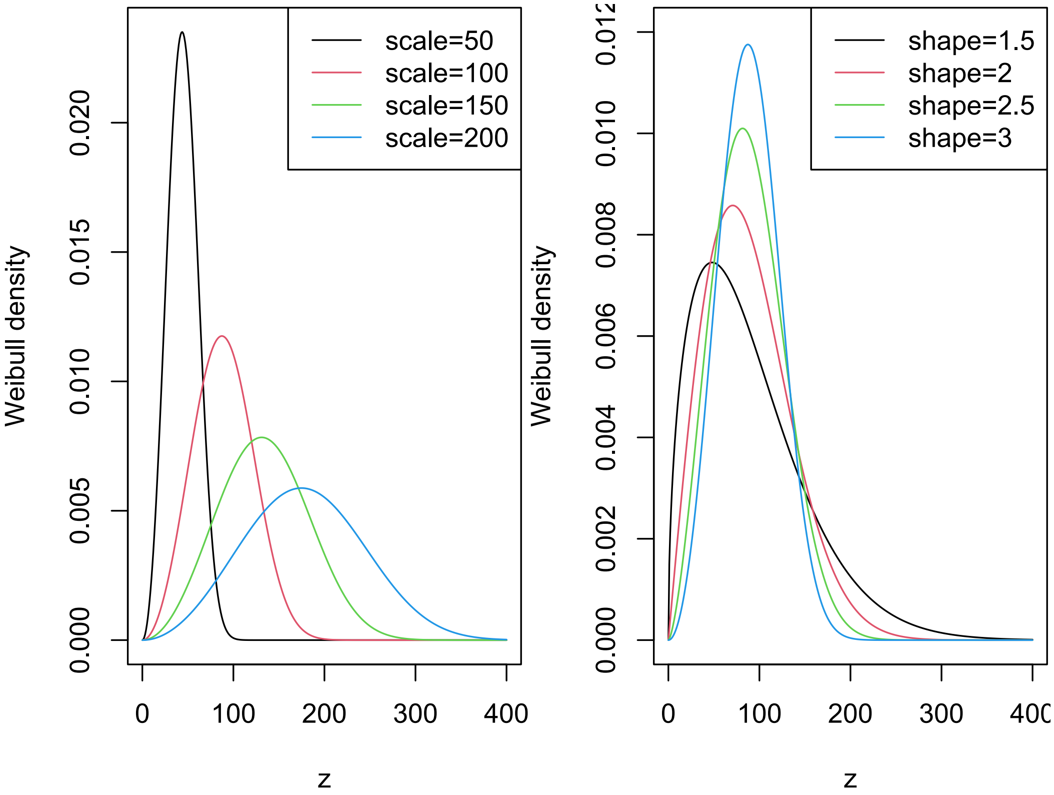


FIGURE 4.3: **Weibull Densities**. The left-hand panel is with shape=3 and varying scale. The right-hand panel is with scale=100 and varying shape.

The distribution function of the Weibull distribution is given by

$$F_X(x) = 1-\exp\left(-\left(\frac{x}{\theta}\right)^{\alpha}\right) \quad x>0,\ \alpha>0,\ \theta>0.$$

It can be easily seen that the shape parameter α describes the shape of the hazard function of the Weibull distribution. The hazard function is a decreasing function when $\alpha < 1$ (heavy tailed distribution), constant when $\alpha = 1$ and increasing when $\alpha > 1$ (light tailed distribution). This behavior of the hazard function makes the Weibull distribution a suitable model for a wide variety of phenomena such as weather forecasting, electrical and industrial engineering, insurance modeling, and financial risk analysis.

The k-th raw moment of the Weibull distributed random variable is given by

$$\mathrm{E}\left(X^k\right) = \theta^k \ \Gamma\left(1 + \frac{k}{\alpha}\right).$$

The mean and variance are given by

$$\mathrm{E}\left(X\right) = \theta \ \Gamma\left(1 + \frac{1}{\alpha}\right)$$

and

$$\mathrm{Var}(X) = \theta^2 \left(\Gamma\left(1 + \frac{2}{\alpha}\right) - \left[\Gamma\left(1 + \frac{1}{\alpha}\right)\right]^2\right),$$

respectively.

Example 4.2.2. Suppose that the probability distribution of the lifetime of AIDS patients (in months) from the time of diagnosis is described by the Weibull distribution with shape parameter 1.2 and scale parameter 33.33.

a. Find the probability that a randomly selected person from this population survives at least 12 months.
b. A random sample of 10 patients will be selected from this population. What is the probability that at most two will die within one year of diagnosis.
c. Find the 99-th percentile of the distribution of lifetimes.

Example Solution.

a. Let X be the lifetime of AIDS patients (in months) having a Weibull distribution with parameters $(1.2, 33.33)$. We have,

$$\Pr\left(X \geq 12\right) = S_X\left(12\right) = e^{-\left(\frac{12}{33.33}\right)^{1.2}} = 0.746.$$

b. Let Y be the number of patients who die within one year of diagnosis. Then, $Y \sim Bin\left(10,\ 0.254\right)$ and $\Pr\left(Y \leq 2\right) = 0.514$.

c. Let $\pi_{0.99}$ denote the 99-th percentile of this distribution. Then,

$$S_X\left(\pi_{0.99}\right) = \exp\left\{-\left(\frac{\pi_{0.99}}{33.33}\right)^{1.2}\right\} = 0.01.$$

Solving for $\pi_{0.99}$, we get $\pi_{0.99} = 118.99$.

4.2.4 The Generalized Beta Distribution of the Second Kind

The Generalized Beta Distribution of the Second Kind (*GB2*) was introduced by Venter (1983) in the context of insurance loss modeling and by McDonald (1984) as an income and wealth distribution. It is a four-parameter, very flexible, distribution that can model positively as well as negatively skewed distributions.

The continuous variable X is said to have the *GB2* distribution with parameters σ, θ, α_1 and α_2 if its *pdf* is given by

$$f_X\left(x\right) = \frac{(x/\theta)^{\alpha_2/\sigma}}{x\sigma\ \mathrm{B}\left(\alpha_1,\alpha_2\right)\left[1+(x/\theta)^{1/\sigma}\right]^{\alpha_1+\alpha_2}} \quad \text{for } x>0, \tag{4.2}$$

$\sigma, \theta, \alpha_1, \alpha_2 > 0$, and where the beta function $\mathrm{B}\left(\alpha_1,\alpha_2\right)$ is defined as

$$\mathrm{B}\left(\alpha_1,\alpha_2\right) = \int_0^1 t^{\alpha_1-1}\left(1-t\right)^{\alpha_2-1} dt.$$

The *GB2* provides a model for heavy as well as light tailed data. It includes the exponential, gamma, Weibull, Burr, Lomax, F, chi-square, Rayleigh, lognormal and log-logistic as special or limiting cases. For example, by setting the parameters $\sigma = \alpha_1 = \alpha_2 = 1$, the *GB2* reduces to the log-logistic distribution. When $\sigma = 1$ and $\alpha_2 \to \infty$, it reduces to the gamma distribution, and when $\alpha = 1$ and $\alpha_2 \to \infty$, it reduces to the Weibull distribution.

A *GB2* random variable can be constructed as follows. Suppose that G_1 and G_2 are independent random variables where G_i has a gamma distribution with shape parameter α_i and scale parameter 1. Then, one can show that the random variable $X = \theta\left(\frac{G_1}{G_2}\right)^{\sigma}$ has a *GB2* distribution with *pdf* summarized in equation (4.2). This theoretical result has several implications. For example, when the moments exist, one can show that the k-th raw moment of the *GB2* distributed random variable is given by

$$\mathrm{E}\left(X^k\right) = \frac{\theta^k\ \mathrm{B}\left(\alpha_1+k\sigma,\alpha_2-k\sigma\right)}{\mathrm{B}\left(\alpha_1,\alpha_2\right)}, \quad k>0.$$

As will be described in Section 4.3.1, the *GB2* is also related to an F-distribution, a result that can be useful in simulation and residual analysis.

Earlier applications of the *GB2* were on income data and more recently have been used to model long-tailed claims data (Section 13.2 describes different interpretations of the descriptor "long-tail"). The *GB2* has been used to model different types of automobile insurance claims, severity of fire losses, as well as medical insurance claim data.

4.3 Methods of Creating New Distributions

In this section, you learn how to:

- Understand connections among the distributions
- Give insights into when a distribution is preferred when compared to alternatives
- Provide foundations for creating new distributions

4.3.1 Functions of Random Variables and their Distributions

In Section 4.2 we discussed some elementary known distributions. In this section we discuss means of creating new parametric probability distributions from existing ones. Specifically, let X be a continuous random variable with a known *pdf* $f_X(x)$ and distribution function $F_X(x)$. We are interested in the distribution of $Y = g\left(X\right)$, where $g(X)$ is a one-to-one transformation defining a new random variable Y. In this section we apply the following techniques for creating new families of distributions: (a) multiplication by a constant (b) raising to a power, (c) exponentiation and (d) mixing.

Multiplication by a Constant

If claim data show change over time then such transformation can be useful to adjust for inflation. If the level of inflation is positive then claim costs are rising, and if it is negative then costs are falling. To adjust for inflation we multiply the cost X by 1+ inflation rate (negative inflation is deflation). To account for currency impact on claim costs we also use a transformation to apply currency conversion from a base to a counter currency.

Consider the transformation $Y = cX$, where $c > 0$, then the distribution function of Y is given by

$$F_Y(y) = \Pr(Y \le y) = \Pr(cX \le y) = \Pr\left(X \le \frac{y}{c}\right) = F_X\left(\frac{y}{c}\right).$$

Using the chain rule for differentiation, the *pdf* of interest $f_Y(y)$ can be written as

$$f_Y(y) = \frac{1}{c} f_X\left(\frac{y}{c}\right).$$

Suppose that X belongs to a certain set of parametric distributions and define a rescaled version $Y = cX$, $c > 0$. If Y is in the same set of distributions then the distribution is said to be a scale distribution. When a member of a scale distribution is multiplied by a constant c ($c > 0$), the scale parameter for this scale distribution meets two conditions:

- The parameter is changed by multiplying by c;
- All other parameters remain unchanged.

Example 4.3.1. Actuarial Exam Question. Losses of Eiffel Auto Insurance are denoted in Euro currency and follow a lognormal distribution with $\mu = 8$ and $\sigma = 2$. Given that 1 euro = 1.3 dollars, find the set of lognormal parameters which describe the distribution of Eiffel's losses in dollars.

Example Solution. Let X and Y denote the aggregate losses of Eiffel Auto Insurance in euro currency and dollars respectively. As $Y = 1.3X$, we have,

$$F_Y(y) = \Pr(Y \le y) = \Pr(1.3X \le y) = \Pr\left(X \le \frac{y}{1.3}\right) = F_X\left(\frac{y}{1.3}\right).$$

X follows a lognormal distribution with parameters $\mu = 8$ and $\sigma = 2$. The *pdf* of X is given by

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2\right\} \quad \text{for } x > 0.$$

As $\left|\frac{dx}{dy}\right| = \frac{1}{1.3}$, the *pdf* of interest $f_Y(y)$ is

$$\begin{aligned} f_Y(y) &= \tfrac{1}{1.3} f_X\left(\tfrac{y}{1.3}\right) \\ &= \tfrac{1}{1.3}\tfrac{1.3}{y\sigma\sqrt{2\pi}} \exp\left\{-\tfrac{1}{2}\left(\tfrac{\log(y/1.3)-\mu}{\sigma}\right)^2\right\} \\ &= \tfrac{1}{y\sigma\sqrt{2\pi}} \exp\left\{-\tfrac{1}{2}\left(\tfrac{\log y-(\log 1.3+\mu)}{\sigma}\right)^2\right\}. \end{aligned}$$

Then Y follows a lognormal distribution with parameters $\log 1.3 + \mu = 8.26$ and

$\sigma = 2.00$. If we let $\mu = \log(m)$ then it can be easily seen that $m = e^{\mu}$ is the scale parameter which was multiplied by 1.3 while σ is the shape parameter that remained unchanged.

Example 4.3.2. Actuarial Exam Question. Demonstrate that the gamma distribution is a scale distribution.

Example Solution. Let $X \sim Ga(\alpha, \theta)$ and $Y = cX$. As $\left|\frac{dx}{dy}\right| = \frac{1}{c}$, then

$$f_Y\left(y\right) = \frac{1}{c} f_X\left(\frac{y}{c}\right) = \frac{\left(\frac{y}{c\theta}\right)^{\alpha}}{y\,\Gamma\left(\alpha\right)} \exp\left(-\frac{y}{c\theta}\right).$$

We can see that $Y \sim Ga(\alpha, c\theta)$ indicating that gamma is a scale distribution and θ is a scale parameter.

Using the same approach as in the example, you can demonstrate that other distributions introduced in Section 4.2 are also scale distributions. In actuarial modeling, working with a scale distribution is very convenient because it allows to incorporate the effect of inflation and to accommodate changes in the currency unit.

Raising to a Power

In Section 4.2.3 we talked about the flexibility of the Weibull distribution in fitting reliability data. Looking to the origins of the Weibull distribution, we recognize that the Weibull is a power transformation of the exponential distribution. This is an application of another type of transformation which involves raising the random variable to a power.

Consider the transformation $Y = X^{\tau}$, where $\tau > 0$, then the distribution function of Y is given by

$$F_Y\left(y\right) = \Pr\left(Y \leq y\right) = \Pr\left(X^{\tau} \leq y\right) = \Pr\left(X \leq y^{1/\tau}\right) = F_X\left(y^{1/\tau}\right).$$

Hence, the *pdf* of interest $f_Y(y)$ can be written as

$$f_Y(y) = \frac{1}{\tau} y^{(1/\tau)-1} f_X\left(y^{1/\tau}\right).$$

On the other hand, if $\tau < 0$, then the distribution function of Y is given by

$$F_Y\left(y\right) = \Pr\left(Y \leq y\right) = \Pr\left(X^{\tau} \leq y\right) = \Pr\left(X \geq y^{1/\tau}\right) = 1 - F_X\left(y^{1/\tau}\right),$$

and

$$f_Y(y) = \left|\frac{1}{\tau}\right| y^{(1/\tau)-1} f_X\left(y^{1/\tau}\right).$$

Example 4.3.3. We assume that X follows the exponential distribution with mean θ and consider the transformed variable $Y = X^{\tau}$. Show that Y follows the Weibull distribution when τ is positive and determine the parameters of the Weibull distribution.

Example Solution. As X follows the exponential distribution with mean θ, we have

$$f_X(x) = \frac{1}{\theta} e^{-x/\theta} \qquad x > 0.$$

Solving for *x* yields $x = y^{1/\tau}$. Taking the derivative, we have

$$\left| \frac{dx}{dy} \right| = \frac{1}{\tau} y^{\frac{1}{\tau}-1}.$$

Thus,

$$f_Y(y) = \frac{1}{\tau} y^{\frac{1}{\tau}-1} f_X\left(y^{\frac{1}{\tau}}\right) = \frac{1}{\tau\theta} y^{\frac{1}{\tau}-1} e^{-\frac{y^{\frac{1}{\tau}}}{\theta}} = \frac{\alpha}{\beta}\left(\frac{y}{\beta}\right)^{\alpha-1} e^{-(y/\beta)^{\alpha}}.$$

where $\alpha = \frac{1}{\tau}$ and $\beta = \theta^{\tau}$. Then, Y follows the Weibull distribution with shape parameter α and scale parameter β.

Special Case. Relating a *GB2* to an F- Distribution. We can use tranforms such as multiplication by a constant and raising to a power to verify that the *GB2* distribution is related to an F-distribution, a distribution widely used in applied statistics.

To see this relationship, we first note that $\frac{1}{2}G_1$ has a gamma distribution with shape parameter α_1 and scale parameter 0.5. Readers with some background in applied statistics may also recognize this to be a *chi-square* distribution with degrees of freedom $2\alpha_1$. The ratio of independent chi-squares has an F-distribution. That is

$$\frac{G_1}{G_2} = \frac{0.5G_1}{0.5G_2} = F$$

has an F-distribution with numerator degrees of freedom $2\alpha_1$ and denominator degrees of freedom $2\alpha_2$. Thus, a random variable X with a *GB2* distribution can be expressed as $X = \theta\left(\frac{G_1}{G_2}\right)^{\sigma} = \theta\ F^{\sigma}$. With this, you can think of a *GB2* as a "power F" or a "generalized F", as it is sometimes known in the literature.

Simulation, discussed in Chapter 8, provides a direct application of this result. Suppose we know how to simulate an outcome with an $F - distribution$ (that is easy to do using, for example, the `R` function `rf(n,df1,df2)`), say F. Then we raise it to the power σ and multiply it by θ so that $\theta\ F^{\sigma}$ is an outcome that has a *GB2* distribution.

Residual analysis provides another direct application. Suppose we have an outcome, say X, that we think comes from a *GB2* distribution. Then we can examine the transformed version $X^* = (X/\theta)^{1/\sigma}$. If the original specification is correct, then X^* has an $F-$ distribution and there are many well-known techniques, some described in Chapter 6, for verifying this assertion.

Exponentiation

The normal distribution is a very popular model for a wide number of applications and when the sample size is large, it can serve as an approximate distribution for other models. If the random variable X has a normal distribution with mean μ and variance σ^2, then $Y = e^X$ has a lognormal distribution with parameters μ and σ^2. The lognormal random variable has a lower bound of zero, is positively skewed and has a long right tail. A lognormal distribution is commonly used to describe distributions of financial assets such as stock prices. It is also used in fitting claim amounts for automobile as well as health insurance. This is an example of another type of transformation which involves exponentiation.

In general, consider the transformation $Y = e^X$. Then, the distribution function of Y is given by

$$F_Y\left(y\right) = \Pr\left(Y \le y\right) = \Pr\left(e^X \le y\right) = \Pr\left(X \le \log y\right) = F_X\left(\log y\right).$$

Taking derivatives, we see that the *pdf* of interest $f_Y(y)$ can be written as

$$f_Y(y) = \frac{1}{y} f_X\left(\log y\right).$$

As an important special case, suppose that X is normally distributed with mean μ and variance σ^2. Then, the distribution of $Y = e^X$ is

$$f_Y(y) = \frac{1}{y} f_X\left(\log y\right) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\log y - \mu}{\sigma}\right)^2\right\}.$$

This is known as a *lognormal* distribution.

Example 4.3.4. Actuarial Exam Question. Assume that X has a uniform distribution on the interval $(0,\ c)$ and define $Y = e^X$. Find the distribution of Y.

Example Solution. We begin with the cdf of Y,

$$F_Y\left(y\right) = \Pr\left(Y \le y\right) = \Pr\left(e^X \le y\right) = \Pr\left(X \le \log y\right) = F_X\left(\log y\right).$$

Taking the derivative, we have,

$$f_Y\left(y\right) = \frac{1}{y} f_X\left(\log y\right) = \frac{1}{cy}.$$

Since $0 < x < c$, then $1 < y < e^c$.

4.3.2 Mixture Distributions for Severity

Mixture distributions represent a useful way of modeling data that are drawn from a heterogeneous population. This parent population can be thought to be divided into multiple subpopulations with distinct distributions.

Finite Mixtures

Two-point Mixture

If the underlying phenomenon is diverse and can actually be described as two phenomena representing two subpopulations with different modes, we can construct the two-point mixture random variable X. Given random variables X_1 and X_2, with *pdf*s $f_{X_1}\left(x\right)$ and $f_{X_2}\left(x\right)$ respectively, the *pdf* of X is the weighted average of the component *pdf* $f_{X_1}\left(x\right)$ and $f_{X_2}\left(x\right)$. The *pdf* and distribution function of X are given by

$$f_X\left(x\right) = a f_{X_1}\left(x\right) + \left(1-a\right) f_{X_2}\left(x\right),$$

and

$$F_X\left(x\right) = a F_{X_1}\left(x\right) + \left(1-a\right) F_{X_2}\left(x\right),$$

for $0 < a < 1$, where the mixing parameters a and $(1-a)$ represent the proportions of data points that fall under each of the two subpopulations respectively. This weighted average can be applied to a number of other distribution related quantities. The k-th raw moment and moment generating function of X are given by $\mathrm{E}\left(X^k\right) = a\mathrm{E}\left(X_1^K\right) + \left(1-a\right)\mathrm{E}\left(X_2^k\right)$, and

$$M_X(t) = a M_{X_1}(t) + \left(1-a\right) M_{X_2}(t),$$

respectively.

Example 4.3.5. Actuarial Exam Question. A collection of insurance policies consists of two types. 25% of policies are Type 1 and 75% of policies are Type 2. For a policy of Type 1, the loss amount per year follows an exponential distribution with mean 200, and for a policy of Type 2, the loss amount per year follows a Pareto distribution with parameters $\alpha = 3$ and $\theta = 200$. For a policy chosen at random from the entire collection of both types of policies, find the probability that the annual loss will be less than 100, and find the average loss.

Example Solution. The two types of losses are the random variables X_1 and X_2. X_1 has an exponential distribution with mean 100, so $F_{X_1}(100) = 1 - e^{-\frac{100}{200}} = 0.393$. X_2 has a Pareto distribution with parameters $\alpha = 3$ and $\theta = 200$, so $F_{X_1}(100) = 1 - \left(\frac{200}{100+200}\right)^3 = 0.704$. Hence, $F_X(100) = (0.25 \times 0.393) + (0.75 \times 0.704) = 0.626$.

The average loss is given by

$$\mathrm{E}(X) = 0.25\mathrm{E}(X_1) + 0.75\mathrm{E}(X_2) = (0.25 \times 200) + (0.75 \times 100) = 125.$$

k-*point Mixture*

In case of finite mixture distributions, the random variable of interest X has a probability p_i of being drawn from homogeneous subpopulation i, where $i = 1, 2, \ldots, k$ and k is the initially specified number of subpopulations in our mixture. The mixing parameter p_i represents the proportion of observations from subpopulation i. Consider the random variable X generated from k distinct subpopulations, where subpopulation i is modeled by the continuous distribution $f_{X_i}(x)$. The probability distribution of X is given by

$$f_X(x) = \sum_{i=1}^{k} p_i f_{X_i}(x),$$

where $0 < p_i < 1$ and $\sum_{i=1}^{k} p_i = 1$.

This model is often referred to as a finite mixture or a k-point mixture. The distribution function, r-th raw moment and moment generating functions of the k point mixture are given as

$$F_X(x) = \sum_{i=1}^{k} p_i F_{X_i}(x),$$

$$\mathrm{E}(X^r) = \sum_{i=1}^{k} p_i \mathrm{E}(X_i^r), \quad \text{and}$$

$$M_X(t) = \sum_{i=1}^{k} p_i M_{X_i}(t),$$

respectively.

Example 4.3.6. Actuarial Exam Question. Y_1 is a mixture of X_1 and X_2 with mixing weights a and $(1-a)$. Y_2 is a mixture of X_3 and X_4 with mixing weights b and $(1-b)$. Z is a mixture of Y_1 and Y_2 with mixing weights c and $(1-c)$.

Show that Z is a mixture of X_1, X_2, X_3 and X_4, and find the mixing weights.

Example Solution. Applying the formula for a mixed distribution, we get

$$f_{Y_1}(x) = af_{X_1}(x) + (1-a)\,f_{X_2}(x)$$

$$\begin{aligned} f_{Y_2}(x) &= bf_{X_3}(x) + (1-b)\,f_{X_4}(x) \\ f_Z(x) &= cf_{Y_1}(x) + (1-c)\,f_{Y_2}(x)\,. \end{aligned}$$

Substituting the first two equations into the third, we get

$$f_Z(x) = c\left[af_{X_1}(x) + (1-a)\,f_{X_2}(x)\right] + (1-c)\left[bf_{X_3}(x) + (1-b)\,f_{X_4}(x)\right]$$

$$= caf_{X_1}(x) + c\,(1-a)\,f_{X_2}(x) + (1-c)\,bf_{X_3}(x) + (1-c)\,(1-b)\,f_{X_4}(x)\,.$$

Then, Z is a mixture of X_1, X_2, X_3 and X_4, with mixing weights ca, $c\,(1-a)$, $(1-c)\,b$ and $(1-c)\,(1-b)$, respectively. It can be easily seen that the mixing weights sum to one.

Continuous Mixtures

A mixture with a very large number of subpopulations (k goes to infinity) is often referred to as a continuous mixture. In a continuous mixture, subpopulations are not distinguished by a discrete mixing parameter but by a continuous variable Θ, where Θ plays the role of p_i in the finite mixture. Consider the random variable X with a distribution depending on a parameter Θ, where Θ itself is a continuous random variable. This description yields the following model for X

$$f_X(x) = \int_{-\infty}^{\infty} f_X(x\,|\theta)\,g_\Theta(\theta)d\theta,$$

where $f_X(x|\theta)$ is the conditional distribution of X at a particular value of $\Theta = \theta$ and $g_\Theta(\theta)$ is the probability statement made about the unknown parameter θ. In a Bayesian context (to be described in Chapter 9), this is known as the prior distribution of Θ (the prior information or expert opinion to be used in the analysis).

The distribution function, k-th raw moment and moment generating functions of the continuous mixture are given as

$$F_X(x) = \int_{-\infty}^{\infty} F_X(x\,|\theta)\, g_\Theta(\theta) d\theta,$$

$$\mathrm{E}\left(X^k\right) = \int_{-\infty}^{\infty} \mathrm{E}\left(X^k\,|\theta\right) g_\Theta(\theta) d\theta,$$

$$M_X(t) = \mathrm{E}\left(e^{tX}\right) = \int_{-\infty}^{\infty} \mathrm{E}\left(e^{tx}\,|\theta\right) g_\Theta(\theta) d\theta,$$

respectively.

The k-th raw moment of the mixture distribution can be rewritten as

$$\mathrm{E}\left(X^k\right) = \int_{-\infty}^{\infty} \mathrm{E}\left(X^k\,|\theta\right) g_\Theta(\theta) d\theta \;=\; \mathrm{E}\left[\mathrm{E}\left(X^k\,|\Theta\right)\right].$$

Using the law of iterated expectations (see Appendix Chapter 18), we can define the mean and variance of X as

$$\mathrm{E}(X) = \mathrm{E}\left[\mathrm{E}(X\,|\Theta)\right]$$

and

$$\mathrm{Var}(X) = \mathrm{E}\left[\mathrm{Var}(X\,|\Theta)\right] + \mathrm{Var}\left[\mathrm{E}(X\,|\Theta)\right].$$

Example 4.3.7. Actuarial Exam Question. X has a normal distribution with a mean of Λ and variance of 1. Λ has a normal distribution with a mean of 1 and variance of 1. Find the mean and variance of X.

Example Solution. X is a continuous mixture with mean

$$\mathrm{E}(X) = \mathrm{E}\left[\mathrm{E}(X|\Lambda)\right] = \mathrm{E}(\Lambda) = 1$$

and

$$\mathrm{V}(X) = \mathrm{V}\left[\mathrm{E}(X|\Lambda)\right] + \mathrm{E}\left[\mathrm{V}(X|\Lambda)\right] = \mathrm{V}(\Lambda) + \mathrm{E}(1) = 1 + 1 = 2.$$

Example 4.3.8. Actuarial Exam Question. Claim sizes, X, are uniform on the interval $(\Theta, \Theta + 10)$ for each policyholder. Θ varies by policyholder according to an exponential distribution with mean 5. Find the unconditional distribution, mean and variance of X.

Example Solution. The conditional distribution of X is $f_X(x|\theta) = \frac{1}{10}$ for $\theta < x < \theta + 10$. The prior distribution of θ is $g_\Theta(\theta) = \frac{1}{5} e^{-\frac{\theta}{5}}$ for $0 < \theta < \infty$.

Multiplying and integrating yields the unconditional distribution of X

$$f_X(x) = \int f_X(x|\theta)\ g_\Theta(\theta) d\theta.$$

For this example, this is

$$f_X(x) = \begin{cases} \int_0^x \frac{1}{50} e^{-\frac{\theta}{5}} d\theta = \frac{1}{10}\left(1 - e^{-\frac{x}{5}}\right) & 0 \le x \le 10, \\ \int_{x-10}^x \frac{1}{50} e^{-\frac{\theta}{5}} d\theta = \frac{1}{10}\left(e^{-\frac{(x-10)}{5}} - e^{-\frac{x}{5}}\right) & 10 < x < \infty. \end{cases}$$

One can use this to derive the mean and variance of the unconditional distribution. Alternatively, start with the conditional mean and variance of X, given by

$$\mathrm{E}(X|\theta) = \frac{\theta + \theta + 10}{2} = \theta + 5$$

and

$$\mathrm{Var}(X|\theta) = \frac{[(\theta + 10) - \theta]^2}{12} = \frac{100}{12},$$

respectively. With these, the unconditional mean and variance of X are given by

$$\mathrm{E}(X) = \mathrm{E}[\mathrm{E}(X|\Theta)] = \mathrm{E}(\Theta + 5) = \mathrm{E}(\Theta) + 5 = 5 + 5 = 10,$$

and

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}[V(X|\Theta)] + \mathrm{Var}[\mathrm{E}(X|\Theta)] \\ &= \mathrm{E}\left(\tfrac{100}{12}\right) + \mathrm{Var}(\Theta + 5) = 8.33 + \mathrm{Var}(\Theta) = 33.33. \end{aligned}$$

4.4 Estimating Loss Distributions

In this section, you learn how to:

- Estimate moments, quantiles, and distributions without reference to a parametric distribution
- Summarize the data graphically without reference to a parametric distribution
- Use method of moments, percentile matching, and maximum likelihood estimation to estimate parameters for different distributions.

4.4.1 Nonparametric Estimation

In Section 3.4 for frequency and Section 4.1 for severity, we learned how to summarize a distribution by computing means, variances, quantiles/percentiles, and so on. To approximate these summary measures using a dataset, one strategy is to:

i. assume a parametric form for a distribution, such as a negative binomial for frequency or a gamma distribution for severity,
ii. estimate the parameters of that distribution, and then
iii. use the distribution with the estimated parameters to calculate the desired summary measure.

This is the parametric approach. Another strategy is to estimate the desired summary measure directly from the observations *without* reference to a parametric model. Not surprisingly, this is known as the nonparametric approach.

Let us start by considering the most basic type of sampling scheme and assume that observations are realizations from a set of random variables $X_1, \ldots, X_n$ that are iid draws from an unknown population distribution $F(\cdot)$. An equivalent way of saying this is that $X_1, \ldots, X_n$, is a *random sample* (with replacement) from $F(\cdot)$. We now describe nonparametric estimators of many important measures that summarize a distribution.

Moment Estimators

We learned how to define moments in Section 3.2.2 for frequency and Section 4.1.1 for severity. In particular, the k-th moment, $\mathrm{E}\left[X^k\right] = \mu_k'$, summarizes many aspects of the distribution for different choices of k. Here, μ_k' is sometimes called the kth *population* moment to distinguish it from the kth sample moment,

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k,$$

which is the corresponding nonparametric estimator. In typical applications, k is a positive integer, although it need not be in theory. The sample estimator for the population mean μ is called the *sample mean*, denoted with a bar on top of the random variable:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

A nonparametric, or sample, estimator of the *k-th central moment*, μ_k is

$$\frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^k .$$

Properties of the sample moment estimator of the variance such as $n^{-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$ have been studied extensively but is not the only possible estimator. The most widely used version is one where the effective sample size is reduced by one, and so we define

$$s^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2 .$$

Dividing by $n-1$ instead of n matters little when you have a large sample size n as is common in insurance applications. The *sample variance* estimator s^2 is unbiased in the sense that $\mathrm{E}\left[s^2\right] = \sigma^2$, a desirable property particularly when interpreting results of an analysis.

Empirical Distribution Function

We have seen how to compute nonparametric estimators of the kth moment $\mathrm{E}\left[X^k\right]$. In the same way, for any known function $\mathrm{g}(\cdot)$, we can estimate $\mathrm{E}\left[\mathrm{g}(X)\right]$ using $n^{-1}\sum_{i=1}^n \mathrm{g}(X_i)$.

Now consider the function $\mathrm{g}(X) = I(X \le x)$ for a fixed x. Here, the notation $I(\cdot)$ is the indicator function; it returns 1 if the event $(\cdot)$ is true and 0 otherwise. Note that now the random variable $\mathrm{g}(X)$ has Bernoulli distribution (a binomial distribution with $n = 1$). We can use this distribution to readily calculate quantities such as the mean and the variance. For example, for this choice of $\mathrm{g}(\cdot)$, the expected value is $\mathrm{E}\left[I(X \le x)\right] = \Pr(X \le x) = F(x)$, the distribution function evaluated at x. We define the nonparametric estimator of the distribution function

$$\begin{aligned} F_n(x) &= \frac{1}{n}\sum_{i=1}^n I\left(X_i \le x\right) \\ &= \frac{\text{number of observations less than or equal to } x}{n}. \end{aligned}$$

As $F_n(\cdot)$ is based on only observations and does not assume a parametric family for the distribution, it is nonparametric and also known as the empirical distribution function. It is also known as the *empirical cumulative distribution function* and, in `R`, one can use the `ecdf(.)` function to compute it.

Example 4.4.1. Toy Data Set. To illustrate, consider a fictitious, or "toy," data set of $n = 10$ observations. Determine the empirical distribution function.

i	1	2	3	4	5	6	7	8	9	10
X_i	10	15	15	15	20	23	23	23	23	30

You should check that the sample mean is $\overline{X} = 19.7$ and that the sample variance is $s^2 = 34.45556$. The corresponding empirical distribution function is

$$F_n(x) = \begin{cases} 0 & \text{for} \quad x < 10 \\ 0.1 & \text{for} \quad 10 \leq x < 15 \\ 0.4 & \text{for} \quad 15 \leq x < 20 \\ 0.5 & \text{for} \quad 20 \leq x < 23 \\ 0.9 & \text{for} \quad 23 \leq x < 30 \\ 1 & \text{for} \quad x \geq 30, \end{cases}$$

as shown in Figure 4.4. The empirical distribution is generally discrete and continuous from the right.

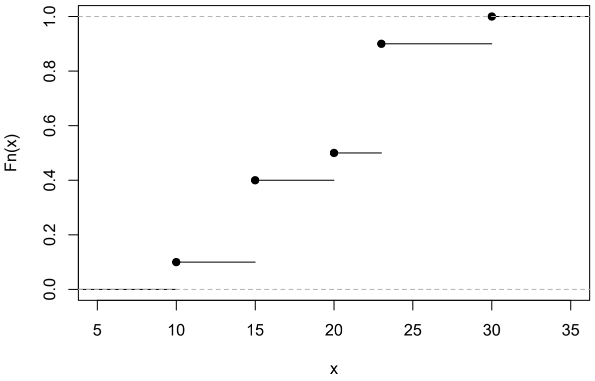


FIGURE 4.4: **Empirical Distribution Function of a Toy Example**

Quartiles, Percentiles and Quantiles

We have already seen in Section 4.1.2 the median, which is the number such that approximately half of a data set is below (or above) it. The first quartile is the number such that approximately 25% of the data is below it and the third quartile is the number such that approximately 75% of the data is below it. A $100p$ percentile is the number such that $100 \times p$ percent of the data is below it.

To generalize this concept, consider a distribution function $F(\cdot)$, which may or may not be continuous, and let q be a fraction so that $0 < q < 1$. We want to define a quantile, say q_F, to be a number such that $F(q_F) \approx q$. Notice that when $q = 0.5$, q_F is the median; when $q = 0.25$, q_F is the first quartile, and so on. In the same way, when $q = 0, 0.01, 0.02, \ldots, 0.99, 1.00$, the resulting q_F is a percentile. So, a quantile generalizes the concepts of median, quartiles, and percentiles.

To be precise, for a given $0 < q < 1$, define the q**th quantile** q_F to be *any* number that satisfies

$$F(q_F-) \leq q \leq F(q_F) \tag{4.3}$$

Here, the notation $F(x-)$ means to evaluate the function $F(\cdot)$ as a left-hand limit.

To get a better understanding of this definition, let us look at a few special cases. First, consider the case where X is a continuous random variable so that the distribution function $F(\cdot)$ has no jump points, as illustrated in Figure 4.5. In this figure, a few fractions, q_1, q_2, and q_3 are shown with their corresponding quantiles $q_{F,1}$, $q_{F,2}$, and $q_{F,3}$. In each case, it can be seen that $F(q_F-) = F(q_F)$ so that there is a unique quantile. Because we can find a unique inverse of the distribution function at any $0 < q < 1$, we can write $q_F = F^{-1}(q)$.

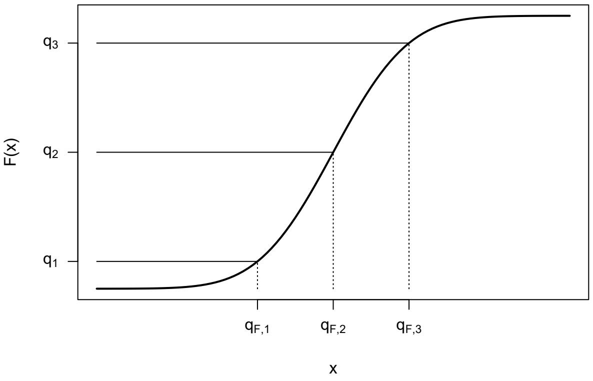


FIGURE 4.5: **Continuous Quantile Case**

Figure 4.6 shows three cases for distribution functions. The left panel corresponds to the continuous case just discussed. The middle panel displays a jump point similar to those we already saw in the empirical distribution function of Figure 4.4. For the value of q shown in this panel, we still have a unique value of the quantile q_F. Even though there are many values of q such that $F(q_F-) \leq q \leq F(q_F)$, for a particular value of q, there is only one solution to equation (4.3). The right panel depicts a situation in which the quantile cannot be uniquely determined for the q shown as there is a range of q_F's satisfying equation (4.3).

Example 4.4.2. Toy Data Set: Continued. Determine quantiles corresponding to the 20th, 50th, and 95th percentiles.

Solution. Consider Figure 4.4. The case of $q = 0.20$ corresponds to the middle

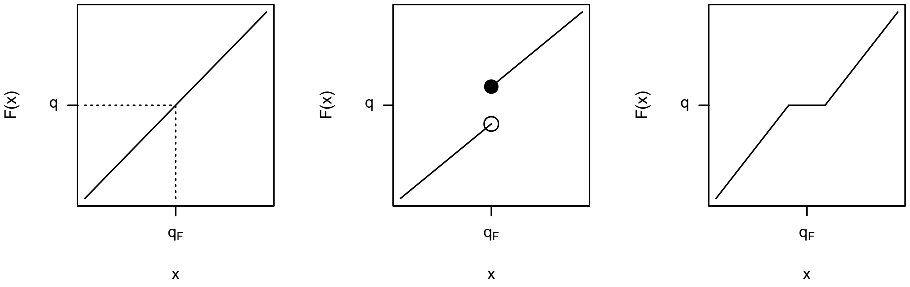


FIGURE 4.6: **Three Quantile Cases**

panel of Figure Figure 4.6, so the 20th percentile is 15. The case of $q = 0.50$ corresponds to the right panel, so the median is any number between 20 and 23 inclusive. Many software packages use the average 21.5 (e.g. `R`, as seen below). For the 95th percentile, the solution is 30. We can see from Figure 4.4 that 30 also corresponds to the 99th and the 99.99th percentiles.

```
xExample <- c(10, rep(15, 3), 20, rep(23, 4), 30)
quantile(xExample, probs = c(0.2, 0.5, 0.95), type = 6)
```

```
 20%  50%  95%
15.0 21.5 30.0
```

By taking a weighted average between data observations, smoothed empirical quantiles can handle cases such as the right panel in Figure 4.6. The qth smoothed empirical quantile is defined as

$$\hat{\pi}_q = (1-h)X_{(j)} + hX_{(j+1)}$$

where $j = \lfloor (n+1)q \rfloor$, $h = (n+1)q - j$, and $X_{(1)}, \ldots, X_{(n)}$ are the ordered values (known as the *order statistics*) corresponding to $X_1, \ldots, X_n$. (Recall that the brackets $\lfloor \cdot \rfloor$ are the floor function denoting the greatest integer value.) Note that $\hat{\pi}_q$ is simply a linear interpolation between $X_{(j)}$ and $X_{(j+1)}$.

Example 4.4.3. Toy Data Set: Continued. Determine the 50th and 20th smoothed percentiles.

Example Solution. Take $n = 10$ and $q = 0.5$. Then, $j = \lfloor (11)(0.5) \rfloor = \lfloor 5.5 \rfloor = 5$ and $h = (11)(0.5) - 5 = 0.5$. Then the 0.5-th smoothed empirical quantile is

$$\hat{\pi}_{0.5} = (1-0.5)X_{(5)} + (0.5)X_{(6)} = 0.5(20) + (0.5)(23) = 21.5.$$

Now take $n = 10$ and $q = 0.2$. In this case, $j = \lfloor (11)(0.2) \rfloor = \lfloor 2.2 \rfloor = 2$ and $h = (11)(0.2) - 2 = 0.2$. Then the 0.2-th smoothed empirical quantile is

$$\hat{\pi}_{0.2} = (1 - 0.2)X_{(2)} + (0.2)X_{(3)} = 0.8(15) + (0.2)(15) = 15.$$

Density Estimators

Discrete Variable. When the random variable is discrete, estimating the probability mass function $f(x) = \Pr(X = x)$ is straightforward. We simply use the sample average, defined to be

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i = x),$$

which is the proportion of the sample equal to x.

Continuous Variable within a Group. For a continuous random variable, consider a discretized formulation in which the domain of $F(\cdot)$ is partitioned by constants $\{c_0 < c_1 < \cdots < c_k\}$ into intervals of the form $[c_{j-1}, c_j)$, for $j = 1, \ldots, k$. The data observations are thus "grouped" by the intervals into which they fall. Then, we might use the basic definition of the empirical mass function, or a variation such as

$$f_n(x) = \frac{n_j}{n \times (c_j - c_{j-1})} \qquad c_{j-1} \leq x < c_j,$$

where n_j is the number of observations (X_i) that fall into the interval $[c_{j-1}, c_j)$.

Continuous Variable (not grouped). Extending this notion to instances where we observe individual data, note that we can always create arbitrary groupings and use this formula. More formally, let $b > 0$ be a small positive constant, known as a bandwidth, and define a density estimator to be

$$f_n(x) = \frac{1}{2nb} \sum_{i=1}^{n} I(x - b < X_i \leq x + b) \tag{4.4}$$

Snippet of Theory. The idea is that the estimator $f_n(x)$ in equation (4.4) is the average over n *iid* realizations of a random variable with mean

$$\begin{aligned} \mathrm{E}\left[\frac{1}{2b} I(x - b < X \leq x + b)\right] &= \frac{1}{2b}\left(F(x + b) - F(x - b)\right) \\ &\to F'(x) = f(x), \end{aligned}$$

as $b \to 0$. That is, $f_n(x)$ is an asymptotically unbiased estimator of $f(x)$ (its expectation approaches the true value as sample size increases to infinity). This development assumes some smoothness of $F(\cdot)$, in particular, twice differentiability at x, but makes no assumptions on the form of the distribution function F. Because of this, the density estimator f_n is said to be *nonparametric*.

More generally, define the kernel density estimator of the pdf at x as

$$f_n(x) = \frac{1}{nb} \sum_{i=1}^{n} w\left(\frac{x - X_i}{b}\right), \tag{4.5}$$

where w is a probability density function centered about 0. Note that equation (4.4) is a special case of the kernel density estimator where $w(x) = \frac{1}{2}I(-1 < x \leq 1)$, also known as the *uniform kernel*. Other popular choices are shown in Table 4.1.

Table 4.1. Popular Kernel Choices

Kernel	$w(x)$
Uniform	$\frac{1}{2}I(-1 < x \leq 1)$
Triangle	$(1 - \lvert x \rvert) \times I(\lvert x \rvert \leq 1)$
Epanechnikov	$\frac{3}{4}(1 - x^2) \times I(\lvert x \rvert \leq 1)$
Gaussian	$\phi(x)$

Here, $\phi(\cdot)$ is the standard normal density function. As we will see in the following example, the choice of bandwidth b comes with a bias-variance tradeoff between matching local distributional features and reducing the volatility.

Example 4.4.4. Property Fund. Figure 4.7 shows a histogram (with shaded gray rectangles) of logarithmic property claims from 2010. The (blue) thick curve represents a Gaussian kernel density where the bandwidth was selected automatically using an ad hoc rule based on the sample size and volatility of these data. For this dataset, the bandwidth turned out to be $b = 0.3255$. For comparison, the (red) dashed curve represents the density estimator with a bandwidth equal to 0.1 and the green smooth curve uses a bandwidth of 1. As anticipated, the smaller bandwidth (0.1) indicates taking local averages over less data so that we get a better idea of the local average, but at the price of higher volatility. In contrast, the larger bandwidth (1) smooths out local fluctuations, yielding a smoother curve that may miss perturbations in the local average. For actuarial applications, we mainly use the kernel density estimator to get a quick visual impression of the data. From this perspective, you can simply use the default ad hoc rule for bandwidth selection, knowing that you have the ability to change it depending on the situation at hand.

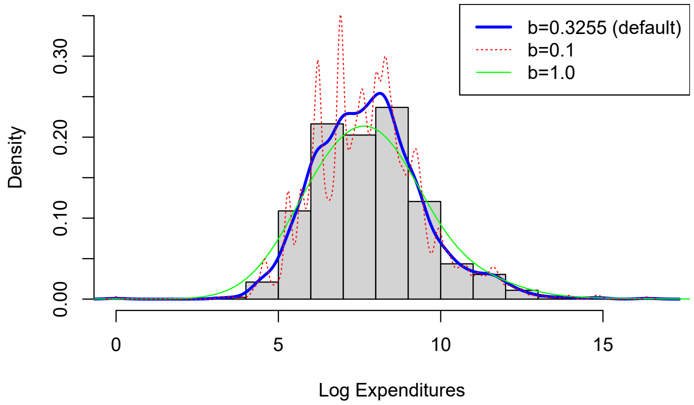


FIGURE 4.7: **Histogram of Logarithmic Property Claims with Superimposed Kernel Density Estimators**

Nonparametric density estimators, such as the kernel estimator, are regularly used in practice. The concept can also be extended to give smooth versions of an empirical distribution function. Given the definition of the kernel density estimator, the *kernel estimator of the distribution function* can be found as

$$\tilde{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x - X_i}{b}\right).$$

where W is the distribution function associated with the kernel density w. To illustrate, for the uniform kernel, we have $w(y) = \frac{1}{2}I(-1 < y \leq 1)$, so

$$W(y) = \begin{cases} 0 & y < -1 \\ \frac{y+1}{2} & -1 \leq y < 1 \\ 1 & y \geq 1 \end{cases}.$$

Example 4.4.5. Actuarial Exam Question. You study five lives to estimate the time from the onset of a disease to death. The times to death are:

$$2 \quad 3 \quad 3 \quad 3 \quad 7$$

Using a triangular kernel with bandwidth 2, calculate the density function estimate at 2.5.

Example Solution. For the kernel density estimate, we have

$$f_n(x) = \frac{1}{nb}\sum_{i=1}^{n} w\left(\frac{x-X_i}{b}\right),$$

where $n=5$, $b=2$, and $x=2.5$. For the triangular kernel, $w(x) = (1-|x|)\times I(|x|\leq 1)$. Thus,

X_i	$\frac{x-X_i}{b}$	$w\left(\frac{x-X_i}{b}\right)$
2	$\frac{2.5-2}{2} = \frac{1}{4}$	$(1-\frac{1}{4})(1) = \frac{3}{4}$
3	$\frac{2.5-3}{2} = \frac{-1}{4}$	$\left(1-\left\lvert\frac{-1}{4}\right\rvert\right)(1) = \frac{3}{4}$
3		
3		
7	$\frac{2.5-7}{2} = -2.25$	$(1-\lvert-2.25\rvert)(0) = 0$

Then the kernel density estimate at $x=2.5$ is

$$f_n(2.5) = \frac{1}{5(2)}\left(\frac{3}{4} + (3)\frac{3}{4} + 0\right) = \frac{3}{10}.$$

4.4.2 Parametric Estimation

Section 4.2 has focused on parametric distributions that are commonly used in insurance applications. However, to be useful in applied work, these distributions must use "realistic" values for the parameters. In this section we cover three methods for estimating parameters: Method of moments, Percentile matching, and Maximum likelihood estimation.

Method of Moments

Under the method of moments, we approximate the moments of the parametric distribution using the empirical moments described in Section 4.4.1. We can then algebraically solve for the parameter estimates.

Example 4.4.6. Property Fund. For the 2010 property fund, there are $n = 1,377$ individual claims (in thousands of dollars) with

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = 26.62259 \quad \text{and} \quad m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = 136154.6.$$

Fit the parameters of the gamma and Pareto distributions using the method of moments.

Solution. To fit a gamma distribution, we have $\mu_1 = \alpha\theta$ and $\mu_2' = \alpha(\alpha+1)\theta^2$. Equating the two yields the method of moments estimators, easy algebra shows that

$$\alpha = \frac{\mu_1^2}{\mu_2' - \mu_1^2} \quad \text{and} \quad \theta = \frac{\mu_2' - \mu_1^2}{\mu_1}.$$

Thus, the method of moment estimators are

$$\begin{aligned}\hat{\alpha} &= \frac{26.62259^2}{136154.6 - 26.62259^2} = 0.005232809 \\ \hat{\theta} &= \frac{136154.6 - 26.62259^2}{26.62259} = 5,087.629.\end{aligned}$$

For comparison, the maximum likelihood values (see Section 4.4.2) turn out to be $\hat{\alpha}_{MLE} = 0.2905959$ and $\hat{\theta}_{MLE} = 91.61378$, so there are big discrepancies between the two estimation procedures. This is one indication, as we have seen before, that the gamma model fits poorly.

In contrast, now assume a Pareto distribution so that $\mu_1 = \theta/(\alpha-1)$ and $\mu_2' = 2\theta^2/((\alpha-1)(\alpha-2))$. Note that this expression for μ_2' is only valid for $\alpha > 2$. Easy algebra shows

$$\alpha = 1 + \frac{\mu_2'}{\mu_2' - \mu_1^2} \qquad \text{and} \qquad \theta = (\alpha-1)\mu_1.$$

Thus, the method of moment estimators are

$$\begin{aligned}\hat{\alpha} &= 1 + \frac{136154.6}{136154.6 - 26,62259^2} = 2.005233 \\ \hat{\theta} &= (2.005233 - 1) \cdot 26.62259 = 26.7619.\end{aligned}$$

The maximum likelihood values turn out to be $\hat{\alpha}_{MLE} = 0.9990936$ and $\hat{\theta}_{MLE} = 2.2821147$. It is interesting that $\hat{\alpha}_{MLE} < 1$; for the Pareto distribution, recall that $\alpha < 1$ means that the mean is infinite. This is another indication that the property claims data set is a long tail distribution.

As the above example suggests, there is flexibility with the method of moments. For example, we could have matched the second and third moments instead of the first and second, yielding different estimators. Furthermore, there is no guarantee that a solution will exist for each problem. For data that are censored or truncated, matching moments is possible for a few problems but, in general, this is a more difficult scenario. Finally, for distributions where the moments do not exist or are infinite, method of moments is not available. As an alternative, one can use the percentile matching technique.

Percentile Matching

Under percentile matching, we approximate the quantiles or percentiles of the parametric distribution using the empirical quantiles or percentiles described in Section 4.4.1.

Example 4.4.7. Property Fund. For the 2010 property fund, we illustrate matching on quantiles. In particular, the Pareto distribution is intuitively pleasing because of the closed-form solution for the quantiles. Recall that the distribution function for the Pareto distribution is

$$F(x) = 1 - \left(\frac{\theta}{x+\theta} \right)^{\alpha} .$$

Easy algebra shows that we can express the quantile as

$$F^{-1}(q) = \theta \left((1-q)^{-1/\alpha} - 1 \right),$$

for a fraction q, $0 < q < 1$.

Determine estimates of the Pareto distribution parameters using the 25th and 95th empirical quantiles.

Example Solution. The 25th percentile (the first quartile) turns out to be 0.78853 and the 95th percentile is 50.98293 (both in thousands of dollars). With two equations

$$0.78853 = \theta \left(1 - (1-.25)^{-1/\alpha} \right) \quad \text{and} \quad 50.98293 = \theta \left(1 - (1-.75)^{-1/\alpha} \right)$$

and two unknowns, the solution is $\hat{\alpha} = 0.9412076$ and $\hat{\theta} = 2.205617$.

We remark here that a numerical routine is required for these solutions as no analytic solution is available. Furthermore, recall that the maximum likelihood estimates are $\hat{\alpha}_{MLE} = 0.9990936$ and $\hat{\theta}_{MLE} = 2.2821147$, so the percentile matching provides a better approximation for the Pareto distribution than the method of moments.

Example 4.4.8. Actuarial Exam Question. You are given:

(i) Losses follow a loglogistic distribution with cumulative distribution function:

$$F(x) = \frac{(x/\theta)^{\gamma}}{1+(x/\theta)^{\gamma}}$$

(ii) The sample of losses is:

10 35 80 86 90 120 158 180 200 210 1500

Calculate the estimate of θ by percentile matching, using the 40th and 80th empirically smoothed percentile estimates.

Example Solution. With 11 observations, we have $j = \lfloor (n+1)q \rfloor = \lfloor 12(0.4) \rfloor = \lfloor 4.8 \rfloor = 4$ and $h = (n+1)q - j = 12(0.4) - 4 = 0.8$. By interpolation, the 40th empirically smoothed percentile estimate is $\hat{\pi}_{0.4} = (1-h)X_{(j)} + hX_{(j+1)} = 0.2(86) + 0.8(90) = 89.2$.

Similarly, for the 80th empirically smoothed percentile estimate, we have $12(0.8) = 9.6$ so the estimate is $\hat{\pi}_{0.8} = 0.4(200) + 0.6(210) = 206$.

Using the loglogistic cumulative distribution, we need to solve the following two equations for parameters $\hat{\theta}$ and $\hat{\gamma}$:

$$0.4 = \frac{(89.2/\hat{\theta})^{\hat{\gamma}}}{1 + (89.2/\hat{\theta})^{\hat{\gamma}}} \quad \text{and} \quad 0.8 = \frac{(206/\hat{\theta})^{\hat{\gamma}}}{1 + (206/\hat{\theta})^{\hat{\gamma}}}.$$

Solving for each parenthetical expression gives $\frac{2}{3} = (89.2/\theta)^{\hat{\gamma}}$ and $4 = (206/\hat{\theta})^{\hat{\gamma}}$. Taking the ratio of the second equation to the first gives $6 = (206/89.2)^{\hat{\gamma}} \Rightarrow \hat{\gamma} = \frac{\log(6)}{\log(206/89.2)} = 2.1407$. Then $4^{1/2.1407} = 206/\hat{\theta} \Rightarrow \hat{\theta} = 107.8$.

Like the method of moments, percentile matching is almost too flexible in the sense that estimators can vary depending on different percentiles chosen. For example, one actuary may use estimation on the 25th and 95th percentiles whereas another uses the 20th and 80th percentiles. In general estimated parameters will differ and there is no compelling reason to prefer one over the other. Also as with the method of moments, percentile matching is appealing because it provides a technique that can be readily applied in selected situations and has an intuitive basis. Although most actuarial applications use maximum likelihood estimators, it can be convenient to have alternative approaches such as method of moments and percentile matching available.

Maximum Likelihood Estimators for Complete Data

At a foundational level, we assume that the analyst has available a random sample $X_1, \ldots, X_n$ from a distribution with distribution function F_X (for brevity, we sometimes drop the subscript X). As is common, we use the vector $\boldsymbol{\theta}$ to denote the set of parameters for F. This basic sample scheme is reviewed in Appendix Section 17.1.1. Although basic, this sampling scheme provides the foundations for understanding more complex schemes that are regularly used in practice, and so it is important to master the basics.

Before drawing from a distribution, we consider potential outcomes summarized by the random variable X_i (here, i is 1, 2, ..., n). After the draw, we observe x_i. Notationally, we use uppercase roman letters for random variables and lower case ones for realizations. We have seen this set-up already in Section 3.4, where we used $\Pr(X_1 = x_1, \ldots, X_n = x_n)$ to quantify the "likelihood" of drawing a sample $\{x_1, \ldots, x_n\}$. With continuous data, we use the joint probability density function instead of joint probabilities. With the independence assumption, the joint *pdf* may be written as the product of pdfs. Thus, we define the **likelihood** to be

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i). \tag{4.6}$$

From the notation, note that we consider this to be a function of the parameters in $\boldsymbol{\theta}$, with the data $\{x_1, \ldots, x_n\}$ held fixed. The maximum likelihood estimator is that value of the parameters in $\boldsymbol{\theta}$ that maximize $L(\boldsymbol{\theta})$.

From calculus, we know that maximizing a function produces the same results as maximizing the logarithm of a function (this is because the logarithm is a monotone function). Because we get the same results, to ease computational considerations, it is common to consider the **logarithmic likelihood**, denoted as

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i). \tag{4.7}$$

Appendix Section 17.2.2 reviews the foundations of maximum likelihood estimation with more mathematical details in Appendix Chapter 19.

Example 4.4.9. Actuarial Exam Question. You are given the following five observations: 521, 658, 702, 819, 1217. You use the single-parameter Pareto with distribution function:

$$F(x) = 1 - \left(\frac{500}{x}\right)^{\alpha}, \qquad x > 500.$$

With $n = 5$, the log-likelihood function is

$$l(\alpha) = \sum_{i=1}^{5} \log f(x_i; \alpha) = 5\alpha \log 500 + 5 \log \alpha - (\alpha + 1) \sum_{i=1}^{5} \log x_i.$$

Figure 4.8 shows the logarithmic likelihood as a function of the parameter α.

We can determine the maximum value of the logarithmic likelihood by taking

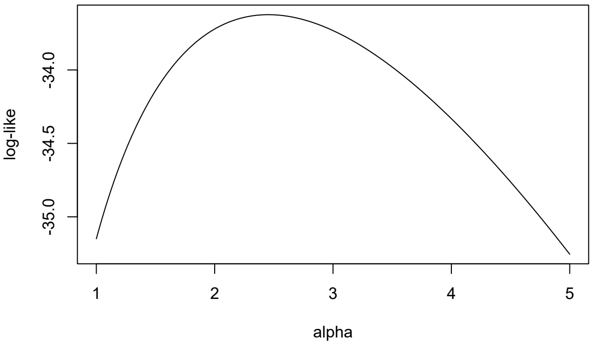


FIGURE 4.8: **Logarithmic Likelihood for a One-Parameter Pareto**

derivatives and setting it equal to zero. This yields

$$\begin{array}{ll} \frac{\partial}{\partial \alpha} l(\alpha) & = 5 \log 500 + 5/\alpha - \sum_{i=1}^5 \log x_i =_{set} 0 \Rightarrow \\ \hat{\alpha}_{MLE} & = \frac{5}{\sum_{i=1}^5 \log x_i - 5 \log 500} = 2.453. \end{array}$$

Naturally, there are many problems where it is not practical to use hand calculations for optimization. Fortunately there are many statistical routines available such as the `R` function `optim`.

This code confirms our hand calculation result where the maximum likelihood estimator is $\alpha_{MLE} = 2.453125$.

We present a few additional examples to illustrate how actuaries fit a parametric distribution model to a set of claim data using maximum likelihood.

Example 4.4.10. Actuarial Exam Question. Consider a random sample of claim amounts: 8000 10000 12000 15000. You assume that claim amounts follow an inverse exponential distribution, with parameter θ. Calculate the maximum likelihood estimator for θ.

Example Solution. The *pdf* is

$$f_X(x) = \frac{\theta e^{-\frac{\theta}{x}}}{x^2},$$

where $x > 0$.

The likelihood function, $L(\theta)$, can be viewed as the probability of the observed data, written as a function of the model's parameter θ

$$L(\theta) = \prod_{i=1}^{4} f_{X_i}(x_i) = \frac{\theta^4 e^{-\theta \sum_{i=1}^{4} \frac{1}{x_i}}}{\prod_{i=1}^{4} x_i^2}.$$

The log-likelihood function, $\log L(\theta)$, is the sum of the individual logarithms

$$\log L(\theta) = 4 \log \theta - \theta \sum_{i=1}^{4} \frac{1}{x_i} - 2 \sum_{i=1}^{4} \log x_i.$$

Taking a derivative, we have

$$\frac{d \log L(\theta)}{d\theta} = \frac{4}{\theta} - \sum_{i=1}^{4} \frac{1}{x_i}.$$

The maximum likelihood estimator of θ, denoted by $\hat{\theta}$, is the solution to the equation

$$\frac{4}{\hat{\theta}} - \sum_{i=1}^{4} \frac{1}{x_i} = 0.$$

Thus, $\hat{\theta} = \frac{4}{\sum_{i=1}^{4} \frac{1}{x_i}} = 10,667$.

The second derivative of $\log L(\theta)$ is given by

$$\frac{d^2 \log L(\theta)}{d\theta^2} = \frac{-4}{\theta^2}.$$

Evaluating the second derivative of the loglikelihood function at $\hat{\theta} = 10,667$ gives a negative value, indicating $\hat{\theta}$ as the value that maximizes the loglikelihood function.

Example 4.4.11. Actuarial Exam Question. A random sample of size 6 is from a lognormal distribution with parameters μ and σ. The sample values are

$$200 \quad 3000 \quad 8000 \quad 60000 \quad 60000 \quad 160000.$$

Calculate the maximum likelihood estimator for μ and σ.

Example Solution. The *pdf* is

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2\right),$$

where $x > 0$.

The likelihood function, $L(\mu,\sigma)$, is the product of the *pdf* for each data point.

$$L(\mu,\sigma)=\prod_{i=1}^{6} f_{X_i}(x_i)=\frac{1}{\sigma^6(2\pi)^3\prod_{i=1}^{6}x_i}\exp\left(-\frac{1}{2}\sum_{i=1}^{6}\left(\frac{\log x_i-\mu}{\sigma}\right)^2\right).$$

Taking a logarithm yields the loglikelihood function, $\log L(\mu,\sigma)$, which is the sum of the individual logarithms.

$$\log L(\mu,\sigma)=-6\log\sigma-3\log(2\pi)-\sum_{i=1}^{6}\log x_i-\frac{1}{2}\sum_{i=1}^{6}\left(\frac{\log x_i-\mu}{\sigma}\right)^2.$$

The first partial derivatives are

$$\begin{array}{ll}\frac{\partial\log L(\mu,\sigma)}{\partial\mu} & =\frac{1}{\sigma^2}\sum_{i=1}^{6}(\log x_i-\mu)\\ \frac{\partial\log L(\mu,\sigma)}{\partial\sigma} & =\frac{-6}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{6}(\log x_i-\mu)^2.\end{array}$$

The maximum likelihood estimators of μ and σ, denoted by $\hat{\mu}$ and $\hat{\sigma}$, are the solutions to the equations

$$\begin{array}{ll}\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{6}(\log x_i-\hat{\mu}) & =0\\ \frac{-6}{\hat{\sigma}}+\frac{1}{\hat{\sigma}^3}\sum_{i=1}^{6}(\log x_i-\hat{\mu})^2 & =0.\end{array}$$

These yield the estimates

$$\hat{\mu}=\frac{\sum_{i=1}^{6}\log x_i}{6}=9.38\quad\text{and}\quad\hat{\sigma}^2=\frac{\sum_{i=1}^{6}(\log x_i-\hat{\mu})^2}{6}=5.12.$$

To check that these estimates maximize, and do not minimize, the likelihood, you may also wish to compute the second partial derivatives. These are

$$\frac{\partial^2\log L(\mu,\sigma)}{\partial\mu^2}=\frac{-6}{\sigma^2},\qquad\frac{\partial^2\log L(\mu,\sigma)}{\partial\mu\partial\sigma}=\frac{-2}{\sigma^3}\sum_{i=1}^{6}(\log x_i-\mu)$$

and

$$\frac{\partial^2\log L(\mu,\sigma)}{\partial\sigma^2}=\frac{6}{\sigma^2}-\frac{3}{\sigma^4}\sum_{i=1}^{6}(\log x_i-\mu)^2.$$

Two follow-up questions rely on large sample properties that you may have seen in an earlier course. Appendix Chapter 19 reviews the definition of the likelihood function, introduces its properties, reviews the maximum likelihood estimators, extends their large-sample properties to the case where there are multiple parameters in the model, and reviews statistical inference based on maximum likelihood estimators. In the solutions of these examples we derive the asymptotic variance of maximum-likelihood estimators of the model parameters. We use the delta method to derive the asymptotic variances of functions of these parameters.

Example 4.4.10 - Follow - Up. Refer to **Example 4.4.10.**

a. Approximate the variance of the maximum likelihood estimator.
b. Determine an approximate 95% confidence interval for θ.
c. Determine an approximate 95% confidence interval for $\Pr(X \leq 9,000)$.

Example Solution.

a. Taking reciprocal of negative expectation of the second derivative of $\log L(\theta)$, we obtain an estimate of the variance of $\hat{\theta}$, $\widehat{Var}\left(\hat{\theta}\right) = \left.\left[E\left(\frac{d^2 \log L(\theta)}{d\theta^2}\right)\right]^{-1}\right|_{\theta=\hat{\theta}} = \frac{\hat{\theta}^2}{4} = 28,446,222.$

It should be noted that as the sample size $n \to \infty$, the distribution of the maximum likelihood estimator $\hat{\theta}$ converges to a normal distribution with mean θ and variance $\hat{V}\left(\hat{\theta}\right)$. The approximate confidence interval in this example is based on the assumption of normality, despite the small sample size, only for the purpose of illustration.

b. The 95

$$10,667 \pm 1.96\sqrt{28,446,222} = (213.34,\ 21120.66).$$

c. The distribution function of X is $F(x) = 1 - e^{-\frac{x}{\theta}}$. Then, the maximum likelihood estimate of $g_\Theta(\theta) = F(9,000)$ is

$$g\left(\hat{\theta}\right) = 1 - e^{-\frac{9,000}{10,667}} = 0.57.$$

We use the delta method to approximate the variance of $g\left(\hat{\theta}\right)$.

$$\frac{\mathrm{dg}(\theta)}{d\theta} = -\frac{9000}{\theta^2} e^{-\frac{9000}{\theta}}.$$

$\widehat{Var}\left[g\left(\hat{\theta}\right)\right] = \left(-\frac{9000}{\hat{\theta}^2} e^{-\frac{9000}{\hat{\theta}}}\right)^2 \hat{V}\left(\hat{\theta}\right) = 0.0329.$

The 95

$$0.57 \pm 1.96\sqrt{0.0329} = (0.214,\ 0.926).$$

Example 4.4.11 - Follow - Up. Refer to **Example 4.4.11.**

a. Estimate the covariance matrix of the maximum likelihood estimator.
b. Determine approximate 95% confidence intervals for μ and σ.
c. Determine an approximate 95% confidence interval for the mean of the lognormal distribution.

a. To derive the covariance matrix of the mle we need to find the expectations of the second derivatives. Since the random variable X is from a lognormal distribution with parameters μ and σ, then $\log X$ is normally distributed with mean μ and variance σ^2.

$$\mathrm{E}\left(\frac{\partial^2 \log \mathrm{L}\left(\mu,\sigma\right)}{\partial\mu^2}\right) = \mathrm{E}\left(\frac{-6}{\sigma^2}\right) = \frac{-6}{\sigma^2},$$

$$\begin{aligned}\mathrm{E}\left(\tfrac{\partial^2 \log \mathrm{L}(\mu,\sigma)}{\partial\mu\partial\sigma}\right) &= \tfrac{-2}{\sigma^3}\textstyle\sum_{i=1}^{6}\mathrm{E}\left(\log x_i-\mu\right)\\ &= \tfrac{-2}{\sigma^3}\textstyle\sum_{i=1}^{6}\left[\mathrm{E}\left(\log x_i\right)-\mu\right] = \tfrac{-2}{\sigma^3}\textstyle\sum_{i=1}^{6}\left(\mu-\mu\right)=0,\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}\left(\tfrac{\partial^2 \log \mathrm{L}(\mu,\sigma)}{\partial\sigma^2}\right) &= \tfrac{6}{\sigma^2}-\tfrac{3}{\sigma^4}\textstyle\sum_{i=1}^{6}\mathrm{E}\left(\log x_i-\mu\right)^2\\ &= \tfrac{6}{\sigma^2}-\tfrac{3}{\sigma^4}\textstyle\sum_{i=1}^{6}\mathrm{Var}\left(\log x_i\right) = \tfrac{6}{\sigma^2}-\tfrac{3}{\sigma^4}\textstyle\sum_{i=1}^{6}\sigma^2\\ &= \tfrac{-12}{\sigma^2}.\end{aligned}$$

Using the negatives of these expectations we obtain the Fisher information matrix

$$\begin{bmatrix}\frac{6}{\sigma^2} & 0\\ 0 & \frac{12}{\sigma^2}\end{bmatrix}.$$

The covariance matrix, Σ, is the inverse of the Fisher information matrix

$$\Sigma = \begin{bmatrix}\frac{\sigma^2}{6} & 0\\ 0 & \frac{\sigma^2}{12}\end{bmatrix}.$$

The estimated matrix is given by

$$\hat{\Sigma} = \begin{bmatrix}0.8533 & 0\\ 0 & 0.4267\end{bmatrix}.$$

b. The 95% confidence interval for μ is given by $9.38 \pm 1.96\sqrt{0.8533} = (7.57,\ 11.19)$.

The 95% confidence interval for σ^2 is given by $5.12 \pm 1.96\sqrt{0.4267} = (3.84,\ 6.40)$.

c. The mean of X is $\exp\left(\mu+\frac{\sigma^2}{2}\right)$. Then, the maximum likelihood estimate of

$$g\left(\mu,\sigma\right) = \exp\left(\mu+\frac{\sigma^2}{2}\right)$$

is

$$g\left(\hat{\mu},\hat{\sigma}\right) = \exp\left(\hat{\mu}+\frac{\hat{\sigma}^2}{2}\right) = 153,277.$$

We use the delta method to approximate the variance of the mle $g\left(\hat{\mu},\hat{\sigma}\right)$.

$\frac{\partial g(\mu,\sigma)}{\partial\mu} = \exp\left(\mu+\frac{\sigma^2}{2}\right)$ and $\frac{\partial g(\mu,\sigma)}{\partial\sigma} = \sigma\exp\left(\mu+\frac{\sigma^2}{2}\right)$.

Using the delta method, the approximate variance of $g\left(\hat{\mu},\hat{\sigma}\right)$ is given by

$$\begin{aligned}
\widehat{Var}\left(g\left(\hat{\mu},\hat{\sigma}\right)\right) &= \left.\begin{bmatrix}\frac{\partial g(\mu,\sigma)}{\partial\mu} & \frac{\partial g(\mu,\sigma)}{\partial\sigma}\end{bmatrix}\Sigma\begin{bmatrix}\frac{\partial g(\mu,\sigma)}{\partial\mu}\\ \frac{\partial g(\mu,\sigma)}{\partial\sigma}\end{bmatrix}\right|_{\mu=\hat{\mu},\sigma=\hat{\sigma}}\\
&= \begin{bmatrix}153,277 & 346,826\end{bmatrix}\begin{bmatrix}0.8533 & 0\\ 0 & 0.4267\end{bmatrix}\begin{bmatrix}153,277\\ 346,826\end{bmatrix}\\
&= 71,374,380,000.
\end{aligned}$$

The 95% confidence interval for $\exp\left(\mu+\frac{\sigma^2}{2}\right)$ is given by

$$153277 \pm 1.96\sqrt{71,374,380,000} = \left(-370356,\ 676910\right).$$

Since the mean of the lognormal distribution cannot be negative, we should replace the negative lower limit in the previous interval by a zero.

Example 4.4.12. Wisconsin Property Fund. To see how maximum likelihood estimators work with real data, we return to the 2010 claims data introduced in Section 1.3.

The following snippet of code shows how to fit the exponential, gamma, Pareto, lognormal, and $GB2$ models. For consistency, the code employs the `R` package `VGAM`. The acronym stands for *Vector Generalized Linear and Additive Models*; as suggested by the name, this package can do far more than fit these models although it suffices for our purposes. The one exception is the $GB2$ density which is not widely used outside of insurance applications; however, we can code this density and compute maximum likelihood estimators using the `optim` general purpose optimizer.

Results from the fitting exercise are summarized in Figure 4.9. Here, the black "longdash" curve is a density estimator of the actual data (introduced in Section 4.4.1); the other curves are parametric curves where the parameters are computed via maximum likelihood. We see poor fits in the red dashed line from the exponential distribution fit and the blue dotted line from the gamma distribution fit. Fits of the other curves, Pareto, lognormal, and GB2, all seem to provide reasonably good fits to the actual data. Chapter 6 describes in more detail the principles of model selection.

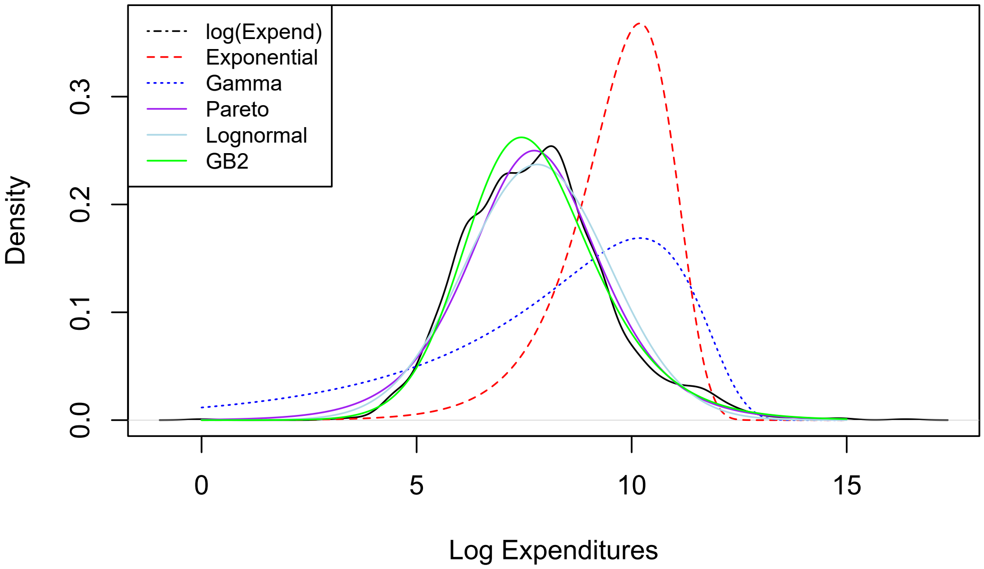


FIGURE 4.9: **Density Comparisons for the Wisconsin Property Fund**

Starting Values

Generally, maximum likelihood is the preferred technique for parameter estimation because it employs data more efficiently. (See Appendix Chapter 19 for precise definitions of efficiency.) However, methods of moments and percentile matching are useful because they are easier to interpret and therefore allow the actuary or analyst to explain procedures to others. Additionally, the numerical estimation procedure (e.g. if performed in `R`) for the maximum likelihood is iterative and requires starting values to begin the recursive process. Although many problems are robust to the choice of the starting values, for some complex situations it can be important to have a starting value that is close to the (unknown) optimal value. Method of moments and percentile matching can produce desirable estimates without a serious computational investment and can thus be used as a *starting value* for computing maximum likelihood.

4.5 Exercises with a Practical Focus

Exercise 4.1. Corporate Travel This exercise is based on the data set introduced in Exercise 1.1 where now the focus is on severity modeling. As in Exercise 3.14, we fit data for the period 2006-2021 but restrict claims to be greater than or equal to 10 (Australian dollars).

- **a**. Using the R function density, provide a nonparametric density estimate of the claims on both the original and logarithmic scale over the range of the data. Use this display to verify that the display is more interpretable on the logarithmic scale.
- **b**. Fit a normal distribution to logarithmic claims and compare the fitted distribution to the nonparametric (empirical) distribution. Interpret this comparison to mean that the lognormal distribution is an excellent candidate to represent these data.
- **c**. As an alternative, fit a Pareto distribution to the claims data using maximum likelihood. To check your work, do this in two ways. A basic approach is to create a log likelihood function and minimize it (using the function optim). A second approach is to the the vglm function from the VGAM package.
- **d**. We have fit X to be a Pareto distribution but wish to plot $Y = \ln(X)$. From Section 4.3.1.3, we saw that $F_Y(y) = F_X(e^y)$ and $f_Y(y) = e^y f_X(e^y)$. Use this transformation to augment the plot in part (b) to include the Pareto distribution.

From this analysis, you learn that the lognormal and Pareto distribution fit the data approximately the same with the lognormal as a slight favorite.

Exercise 4.2. Wisconsin Property Fund. Replicate the real-data example introduced in Example 4.4.12 using the techniques demonstrated in Exercise 4.1.

Exercise 4.3. Group Personal Accident. This exercise is based on the data set introduced in Exercise 1.2. We use incurred claims for all available years, still omitting those less than 10.

One can fit a distribution to the losses. An analysis, summarized in Figure 4.10, shows the results from fitting via maximum likelihood the gamma, Pareto, and lognormal distributions to incurred losses. This figure suggests that the lognormal distribution appears to be the best fit.

Following the outlines in Exercises 4.1 and 4.2, fit these data via maximum likelihood and reproduce the figure that summarizes the results.

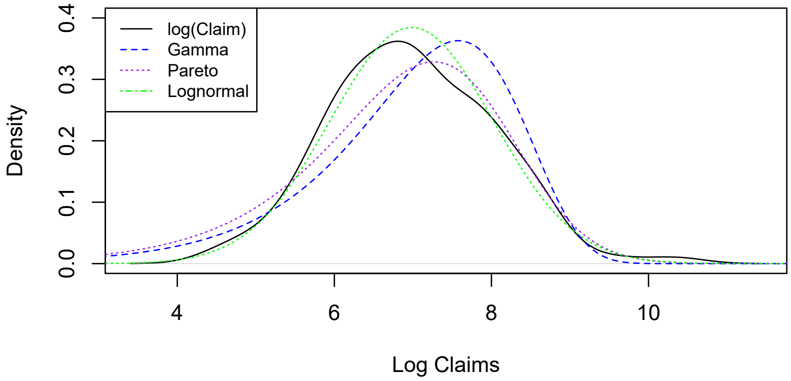


FIGURE 4.10: **Distribution of Group Personal Accident Losses with Superimposed Fitted Distributions**

4.6 Further Resources and Contributors

Contributors

- **Zeinab Amin**, The American University in Cairo, is the principal author of the initial version and also the second edition of this chapter. **Edward (Jed) Frees** and **Lisa Gao**, University of Wisconsin-Madison, are the principal authors of the sections on nonparametric estimation which appeared in chapter 4 of the first edition of the text. Email: zeinabha@aucegypt.edu for chapter comments and suggested improvements.
- Many helpful comments have been provided by Hirokazu (Iwahiro) Iwasawa, iwahiro@bb.mbn.or.jp .
- Other chapter reviewers include: Rob Erhardt, Samuel Kolins, Tatjana Miljkovic, Michelle Xia, and Jorge Yslas.

Further Readings and References

Notable contributions include: Cummins and Derrig (2012), Frees and Valdez (2008), Klugman et al. (2012), Kreer et al. (2015), McDonald (1984), McDonald and Xu (1995), Tevet (2016), and Venter (1983).

If you would like additional practice with R coding, please visit our companion LDA Short Course. In particular, see the Modeling Loss Severity Chapter.

5

Modeling Claim Severity

Chapter Preview. In Chapter 4 we explored the use of continuous as well as mixture distributions to model the random size of loss. Often the risk of loss is shared between the policyholder (the insured) and the insurer. Sharing risk can take the form of a deductible that is paid out-of-pocket of the insured before the insurer contributes to the loss, the form of a limit that caps the insured's liability for loss to a certain amount, or the form of a portion of the loss the insurer is responsible for covering after the insured covers his/her share of the cost, among other forms of cost-sharing. In Sections 5.1.1 to 5.1.3 we introduce the policy deductible feature of the insurance contract, the limited policy, and the co-insurance cost-sharing arrangement. In Section 5.1.4 we explore how insurance companies transfer part of the underlying insured risk by securing coverage from a reinsurer. Section 5.2 covers parametric estimation methods for modified data including grouped, censored and truncated data. In Section 5.3 we apply some non-parametric estimation tools like the ogive estimator, the plug-in principle, the Kaplan-Meier product-limit estimator, and the Nelson Aalon estimator on the modified data.

5.1 Coverage Modifications

In this section, you learn how to:

- Describe the policy deductible feature of the insurance contract, the limited policy, and the coinsurance factor.
- Describe the distinction between the loss incurred to the insured and the amount of paid claim by the insurer under different policy modifications.
- Derive the distribution functions and raw moments for the amount of paid claim by the insurer for the different insurance contracts.
- Calculate the percentage decrease in the expected payment of the insurer as a result of imposing the deductible.
- Describe the insurance mechanism for insurance companies (reinsurance).

- Calculate the raw moments of the amount retained by the primary insurer in the reinsurance agreement.

In this section we evaluate the impacts of coverage modifications: a) deductibles, b) policy limit, c) coinsurance and d) inflation on insurer's costs.

5.1.1 Policy Deductibles

Under an ordinary deductible policy, the insured (policyholder) agrees to cover a fixed amount of an insurance claim before the insurer starts to pay. This fixed expense paid out of pocket is called the deductible and often denoted by d. If the loss exceeds d then the insurer is responsible for covering the loss X less the deductible d. Depending on the agreement, the deductible may apply to each covered loss or to the total losses during a defined benefit period (such as a month, year, etc.)

Deductibles reduce premiums for the policyholders by eliminating a large number of small claims, the costs associated with handling these claims, and the potential moral hazard arising from having insurance. Moral hazard occurs when the insured takes more risks, increasing the chances of loss due to perils insured against, knowing that the insurer will incur the cost (e.g. a policyholder with collision insurance may be encouraged to drive recklessly). The larger the deductible, the less the insured pays in premiums for an insurance policy.

Let X denote the loss incurred to the insured and Y denote the amount of paid claim by the insurer. Speaking of the benefit paid to the policyholder, we differentiate between two variables: The payment per loss and the payment per payment. The payment per loss variable, denoted by Y^L or $(X-d)_+$ is left censored because values of X that are less than d are set equal to zero. This variable is defined as

$$Y^L = (X-d)_+ = \begin{cases} 0 & X \le d, \\ X-d & X > d \end{cases}.$$

Y^L is often referred to as left censored and shifted variable because the values below d are not ignored and all losses are shifted by a value d.

On the other hand, the payment per payment variable, denoted by Y^P, is defined only when there is a payment. Specifically, Y^P equals $X-d$ on the event $\{X > d\}$, denoted as $Y^P = X-d|X > d$. Another way of expressing this that is commonly used is

$$Y^P = \begin{cases} \text{Undefined} & X \le d \\ X-d & X > d. \end{cases}$$

Here, Y^P is often referred to as left truncated and shifted variable or excess loss variable because the claims smaller than d are not reported and values above d are shifted by d.

Even when the distribution of X is continuous, the distribution of Y^L is a hybrid combination of discrete and continuous components. The discrete part of the distribution is concentrated at $Y = 0$ (when $X \leq d$) and the continuous part is spread over the interval $Y > 0$ (when $X > d$). For the discrete part, the probability that no payment is made is the probability that losses fall below the deductible; that is,

$$\Pr\left(Y^L = 0\right) = \Pr\left(X \leq d\right) = F_X\left(d\right).$$

Using the transformation $Y^L = X - d$ for the continuous part of the distribution, we can find the *pdf* of Y^L given by

$$f_{Y^L}\left(y\right) = \begin{cases} F_X\left(d\right) & y = 0 \\ f_X\left(y + d\right) & y > 0. \end{cases}$$

We can see that the payment per payment variable is the payment per loss variable conditional on the loss exceeding the deductible ($X > d$); that is, $Y^P = \left. Y^L \right| X > d$. Alternatively, it can be expressed as $Y^P = (X - d)|X > d$, that is, Y^P is the loss in excess of the deductible given that the loss exceeds the deductible. Hence, the *pdf* of Y^P is given by

$$f_{Y^P}\left(y\right) = \frac{f_X\left(y + d\right)}{1 - F_X\left(d\right)},$$

for $y > 0$. Accordingly, the distribution functions of $Y^L$and Y^P are given by

$$F_{Y^L}\left(y\right) = \begin{cases} F_X\left(d\right) & y = 0 \\ F_X\left(y + d\right) & y > 0, \end{cases}$$

and

$$F_{Y^P}\left(y\right) = \frac{F_X\left(y + d\right) - F_X\left(d\right)}{1 - F_X\left(d\right)},$$

for $y > 0$, respectively.

The raw moments of Y^L and Y^P can be found directly using the *pdf* of X as follows

$$\mathrm{E}\left[\left(Y^L\right)^k\right] = \int_d^\infty \left(x - d\right)^k f_X\left(x\right) dx,$$

and

$$\mathrm{E}\left[\left(Y^P\right)^k\right] = \frac{\int_d^\infty \left(x - d\right)^k f_X\left(x\right) dx}{1 - F_X\left(d\right)} = \frac{\mathrm{E}\left[\left(Y^L\right)^k\right]}{1 - F_X\left(d\right)},$$

respectively. For $k = 1$, we can use the survival function to calculate $\mathrm{E}(Y^L)$ as

$$\mathrm{E}(Y^L) = \int_d^\infty [1 - F_X(x)]\, dx.$$

This could be easily proved if we start with the initial definition of $\mathrm{E}(Y^L)$ and use integration by parts.

We have seen that the deductible d imposed on an insurance policy is the amount of loss that has to be paid out of pocket before the insurer makes any payment. The deductible d imposed on an insurance policy reduces the insurer's payment. The loss elimination ratio (LER) is the percentage decrease in the expected payment of the insurer as a result of imposing the deductible. It is defined as

$$LER = \frac{\mathrm{E}\,(X) - \mathrm{E}\left(Y^L\right)}{\mathrm{E}\,(X)}.$$

A little less common type of policy deductible is the franchise deductible. The franchise deductible will apply to the policy in the same way as ordinary deductible except that when the loss exceeds the deductible d, the full loss is covered by the insurer. The payment per loss and payment per payment variables are defined as

$$Y^L = \begin{cases} 0 & X \le d, \\ X & X > d, \end{cases}$$

and

$$Y^P = \begin{cases} \text{Undefined} & X \le d, \\ X & X > d, \end{cases}$$

respectively.

Example 5.1.1. Actuarial Exam Question. A claim severity distribution is exponential with mean 1000. An insurance company will pay the amount of each claim in excess of a deductible of 100. Calculate the variance of the amount paid by the insurance company for one claim, including the possibility that the amount paid is 0.

Example Solution. Let Y^L denote the amount paid by the insurance company for one claim.

$$Y^L = (X - 100)_+ = \begin{cases} 0 & X \le 100, \\ X - 100 & X > 100. \end{cases}$$

The first and second moments of Y^L are

$$E\left(Y^L\right) = \int_{100}^\infty (x - 100)\, f_X(x)\, dx = \int_{100}^\infty S_X(x) dx = 1000e^{-\frac{100}{1000}},$$

and

$$E\left[\left(Y^L\right)^2\right] = \int_{100}^{\infty} (x-100)^2 f_X(x)\, dx = 2 \times 1000^2 e^{-\frac{100}{1000}}.$$

So,

$$\operatorname{Var}\left(Y^L\right) = \left(2 \times 1000^2 e^{-\frac{100}{1000}}\right) - \left(1000 e^{-\frac{100}{1000}}\right)^2 = 990,944.$$

An arguably simpler path to the solution is to make use of the relationship between X and Y^P. If X is exponentially distributed with mean 1000, then Y^P is also exponentially distributed with the same mean, because of the memoryless property of the exponential distribution. Hence, $E\left(Y^P\right)$=1000 and

$$E\left[\left(Y^P\right)^2\right] = 2 \times 1000^2.$$

Using the relationship between Y^L and Y^P we find

$$E\left(Y^L\right) = E\left(Y^P\right) S_X(100) = 1000 e^{-\frac{100}{1000}}$$

$$E\left[\left(Y^L\right)^2\right] = E\left[\left(Y^P\right)^2\right] S_X(100) = 2 \times 1000^2 e^{-\frac{100}{1000}}.$$

The relationship between X and Y^P can also be used when dealing with the uniform or the Pareto distributions. You can easily show that if X is uniform over the interval $(0,\theta)$ then Y^P is uniform over the interval $(0,\theta-d)$ and if X is Pareto with parameters α and θ then Y^P is Pareto with parameters α and $\theta+d$.

Example 5.1.2. Actuarial Exam Question. For an insurance:

- Losses have a density function

$$f_X(x) = \begin{cases} 0.02x & 0 < x < 10, \\ 0 & \text{elsewhere.} \end{cases}$$

- The insurance has an ordinary deductible of 4 per loss.
- Y^P is the claim payment per payment random variable.

Calculate E $\left(Y^P\right)$.

Example Solution. We define Y^P as follows

$$Y^P = \begin{cases} \text{Undefined} & X \le 4, \\ X-4 & X > 4. \end{cases}$$

So, $E\left(Y^P\right) = \int_4^{10} (x-4)\, 0.02x\ dx / 1 - F_X(4) = \frac{2.88}{0.84} = 3.43.$

Note that we divide by $S_X(4) = 1 - F_X(4)$, as this is the probability where the variable Y^P is defined.

Example 5.1.3. Actuarial Exam Question. You are given:

- Losses follow an exponential distribution with the same mean in all years.
- The loss elimination ratio this year is 70%.
- The ordinary deductible for the coming year is 4/3 of the current deductible.

Compute the loss elimination ratio for the coming year.

Example Solution. Let the losses $X \sim Exp(\theta)$ and the deductible for the coming year $d' = \frac{4}{3}d$, the deductible of the current year. The LER for the current year is

$$\frac{E\left(X\right) - E\left(Y^L\right)}{E\left(X\right)} = \frac{\theta - \theta e^{-d/\theta}}{\theta} = 1 - e^{-d/\theta} = 0.7.$$

Then, $e^{-d/\theta} = 0.3$.

The LER for the coming year is

$$\begin{aligned} &\tfrac{\theta - \theta\exp(-\frac{d'}{\theta})}{\theta} = \tfrac{\theta - \theta\exp\left(-\frac{\frac{4}{3}d}{\theta}\right)}{\theta} \\ &= 1 - \exp\left(-\frac{\frac{4}{3}d}{\theta}\right) = 1 - \left(e^{-d/\theta}\right)^{4/3} = 1 - 0.3^{4/3} = 0.8. \end{aligned}$$

5.1.2 Policy Limits

Under a limited policy, the insurer is responsible for covering the actual loss X up to the limit of its coverage. This fixed limit of coverage is called the policy limit and often denoted by u. If the loss exceeds the policy limit, the difference $X - u$ has to be paid by the policyholder. While a higher policy limit means a higher payout to the insured, it is associated with a higher premium.

Let X denote the loss incurred to the insured and Y denote the amount of paid claim by the insurer. The variable Y is known as the *limited loss variable* and is denoted by $X \wedge u$. It is a right censored variable because values above u are set equal to u. The limited loss random variable Y is defined as

$$Y = X \wedge u = \begin{cases} X & X \leq u \\ u & X > u. \end{cases}$$

It can be seen that the distinction between Y^L and Y^P is not needed under limited policy as the insurer will always make a payment.

Using the definitions of $(X-u)_+$ and $(X\wedge u)$, it can be easily seen that the expected payment without any coverage modification, X, is equal to the sum of the expected payments with deductible u and limit u. That is, $X=(X-u)_+ + (X\wedge u)$.

Even when the distribution of X is continuous, the distribution of Y is a hybrid combination of discrete and continuous components. The discrete part of the distribution is concentrated at $Y=u$ (when $X>u$), while the continuous part is spread over the interval $Y<u$ (when $X\le u$). For the discrete part, the probability that the benefit paid is u, is the probability that the loss exceeds the policy limit u; that is,

$$\Pr(Y=u)=\Pr(X>u)=1-F_X(u).$$

For the continuous part of the distribution $Y=X$, hence the *pdf* of Y is given by

$$f_Y(y)=\begin{cases} f_X(y) & 0<y<u\\ 1-F_X(u) & y=u.\end{cases}$$

Accordingly, the distribution function of Y is given by

$$F_Y(y)=\begin{cases} F_X(x) & 0<y<u\\ 1 & y\ge u.\end{cases}$$

The raw moments of Y can be found directly using the *pdf* of X as follows

$$\mathrm{E}\left(Y^k\right)=\mathrm{E}\left[(X\wedge u)^k\right]=\int_0^u x^k f_X(x)\,dx+\int_u^\infty u^k f_X(x)dx=\int_0^u x^k f_X(x)\,dx+u^k\left[1-F_X(u)\right].$$

An alternative expression using the survival function is

$$\mathrm{E}\left[(X\wedge u)^k\right]=\int_0^u kx^{k-1}\left[1-F_X(x)\right]dx.$$

In particular, for $k=1$, this is

$$\mathrm{E}(Y)=\mathrm{E}(X\wedge u)=\int_0^u[1-F_X(x)]dx.$$

This could be easily proved if we start with the initial definition of $\mathrm{E}(Y)$ and use integration by parts. Alternatively, see the following justification of this limited expectation result.

$$\begin{aligned}\mathrm{E}\left[(X\wedge u)^k\right] &= \mathrm{E}\left[\int_0^{X\wedge u} kx^{k-1}dx\right]\\ &= \mathrm{E}\left[\int_0^u kx^{k-1}I(X>x)dx\right]\\ &= \int_0^u kx^{k-1}\mathrm{E}I(X>x)dx\\ &= \int_0^u kx^{k-1}\left[1-F_X(x)\right]dx.\end{aligned}$$

This approach uses the Fubini-Tonelli theorem to exchange the expectation and integration. Note that it does not make any continuity assumptions about the distribution of X.

When a loss is subject to a deductible d and a limit u, the per-loss variable Y^L is defined as

$$Y^L = \begin{cases} 0 & X \le d \\ X-d & d < X \le u \\ u-d & X > u. \end{cases}$$

Hence, Y^L can be expressed as $Y^L = (X \wedge u) - (X \wedge d)$.

Example 5.1.4. Actuarial Exam Question. Under a group insurance policy, an insurer agrees to pay 100% of the medical bills incurred during the year by employees of a small company, up to a maximum total of one million dollars. The total amount of bills incurred, X, has *pdf*

$$f_X(x) = \begin{cases} \frac{x(4-x)}{9} & 0 < x < 3 \\ 0 & \text{elsewhere.} \end{cases}$$

where x is measured in millions. Calculate the total amount, in millions of dollars, the insurer would expect to pay under this policy.

Example Solution. Define the total amount of bills paid by the insurer as

$$Y = X \wedge 1 = \begin{cases} X & X \le 1 \\ 1 & X > 1. \end{cases}$$

So $\mathrm{E}(Y) = \mathrm{E}(X \wedge 1) = \int_0^1 (x^2(4-x))/9 \; dx + 1 \cdot \int_1^3 (x(4-x))/9 \; dx = 0.935$.

5.1.3 Coinsurance and Inflation

As we have seen in Section 5.1.1, the amount of loss retained by the policyholder can be losses up to the deductible d. The retained loss can also be a percentage of the claim. The percentage α, often referred to as the coinsurance factor, is the percentage of claim the insurance company is required to cover. If the policy is subject to an ordinary deductible and policy limit, coinsurance refers to the percentage of claim the insurer is required to cover, after imposing the ordinary deductible and policy limit. The payment per loss variable, Y^L, is defined as

$$Y^L = \begin{cases} 0 & X \le d, \\ \alpha(X-d) & d < X \le u, \\ \alpha(u-d) & X > u. \end{cases}$$

The maximum amount paid by the insurer in this case is $\alpha(u-d)$, while u is the maximum covered loss.

We have seen in Section 5.1.2 that when a loss is subject to both a deductible d and a limit u the per-loss variable Y^L can be expressed as $Y^L = (X \wedge u) - (X \wedge d)$. With coinsurance, this becomes Y^L can be expressed as $Y^L = \alpha[(X \wedge u) - (X \wedge d)]$.

The k-th raw moment of Y^L is given by

$$\mathrm{E}\left[\left(Y^L\right)^k\right] = \int_d^u [\alpha(x-d)]^k f_X(x)\,dx + [\alpha(u-d)]^k [1 - F_X(u)].$$

A growth factor $(1+r)$ may be applied to X resulting in an inflated loss random variable $(1+r)X$ (the prespecified d and u remain unchanged). The resulting per loss variable can be written as

$$Y^L = \begin{cases} 0 & X \le \frac{d}{1+r} \\ \alpha[(1+r)X - d] & \frac{d}{1+r} < X \le \frac{u}{1+r} \\ \alpha(u-d) & X > \frac{u}{1+r}. \end{cases}$$

The first and second moments of Y^L can be expressed as

$$\mathrm{E}\left(Y^L\right) = \alpha(1+r)\left[\mathrm{E}\left(X \wedge \frac{u}{1+r}\right) - \mathrm{E}\left(X \wedge \frac{d}{1+r}\right)\right],$$

and

$$\begin{aligned}\mathrm{E}\left[\left(Y^L\right)^2\right] &= \alpha^2(1+r)^2\left\{\mathrm{E}\left[\left(X \wedge \tfrac{u}{1+r}\right)^2\right] - \mathrm{E}\left[\left(X \wedge \tfrac{d}{1+r}\right)^2\right]\right. \\ &\quad \left. -2\left(\tfrac{d}{1+r}\right)\left[\mathrm{E}\left(X \wedge \tfrac{u}{1+r}\right) - \mathrm{E}\left(X \wedge \tfrac{d}{1+r}\right)\right]\right\},\end{aligned}$$

respectively.

The formulas given for the first and second moments of Y^L are general. Under full coverage, $\alpha = 1$, $r = 0$, $u = \infty$, $d = 0$ and $\mathrm{E}\left(Y^L\right)$ reduces to $\mathrm{E}(X)$. If only an ordinary deductible is imposed, $\alpha = 1$, $r = 0$, $u = \infty$ and $\mathrm{E}\left(Y^L\right)$ reduces to $\mathrm{E}(X) - \mathrm{E}(X \wedge d)$. If only a policy limit is imposed $\alpha = 1$, $r = 0$, $d = 0$ and $\mathrm{E}\left(Y^L\right)$ reduces to $\mathrm{E}(X \wedge u)$.

Example 5.1.5. Actuarial Exam Question. The ground up loss random variable for a health insurance policy in 2006 is modeled with X, a random variable with an exponential distribution having mean 1000. An insurance policy pays the loss above an ordinary deductible of 100, with a maximum annual payment of 500. The ground up loss random variable is expected to be 5% larger in 2007, but the insurance in 2007 has the same deductible and maximum payment as in 2006. Find the percentage increase in the expected cost per payment from 2006 to 2007.

Example Solution. We define the amount per loss Y^L in both years as

$$Y^L_{2006} = \begin{cases} 0 & X \leq 100, \\ X - 100 & 100 < X \leq 600, \\ 500 & X > 600. \end{cases}$$

$$Y^L_{2007} = \begin{cases} 0 & X \leq 95.24, \\ 1.05X - 100 & 95.24 < X \leq 571.43, \\ 500 & X > 571.43. \end{cases}$$

So,

$$\begin{aligned} E\left(Y^L_{2006}\right) &= E\left(X \wedge 600\right) - E\left(X \wedge 100\right) \\ &= 1000\left(1 - e^{-\frac{600}{1000}}\right) - 1000\left(1 - e^{-\frac{100}{1000}}\right) \\ &= 356.026. \end{aligned}$$

Further,

$$\begin{aligned} E\left(Y^L_{2007}\right) &= 1.05\left[E\left(X \wedge 571.43\right) - E\left(X \wedge 95.24\right)\right] \\ &= 1.05\left[1000\left(1 - e^{-\frac{571.43}{1000}}\right) - 1000\left(1 - e^{-\frac{95.24}{1000}}\right)\right] \\ &= 361.659. \end{aligned}$$

$$\begin{aligned} E\left(Y^P_{2006}\right) &= \frac{356.026}{e^{-(100/1000)}} = 393.469. \\ E\left(Y^P_{2007}\right) &= \frac{361.659}{e^{-(95.24/1000)}} = 397.797. \end{aligned}$$

Because $\frac{E\left(Y^P_{2007}\right)}{E\left(Y^P_{2006}\right)} - 1 = 0.011$, there is an increase of 1.1 percent from 2006 to 2007. Due to the policy limit, the cost per payment event grew by only 1.1 percent between 2006 and 2007 even though the ground up losses increased by 5 percent between the two years.

5.1.4 Reinsurance

In Section 5.1.1 we introduced the policy deductible feature of the insurance contract. In this feature, there is a contractual arrangement under which an insured transfers part of the risk by securing coverage from an insurer in return for an insurance premium. Under that policy, the insured must pay all losses up to the deductible, and the insurer only pays the amount (if any) above the deductible. We now introduce reinsurance, a mechanism of insurance for insurance companies. Reinsurance is a contractual arrangement under which an insurer transfers part of the underlying insured risk by securing coverage from another insurer (referred to as a reinsurer) in return for a reinsurance premium. Although reinsurance involves a relationship between three parties: the original insured, the insurer (often referred to as cedent or cedant) and the reinsurer, the parties of the reinsurance agreement are only the primary insurer and the reinsurer. There is no contractual agreement between the original insured and the reinsurer. Though many different types of reinsurance contracts exist, a common form is excess of loss coverage. In such contracts, the primary insurer must make all required payments to the insured until the primary insurer's total payments reach a fixed reinsurance deductible. The reinsurer is then only responsible for paying losses above the reinsurance deductible. The maximum amount retained by the primary insurer in the reinsurance agreement (the reinsurance deductible) is called retention.

Reinsurance arrangements allow insurers with limited financial resources to increase the capacity to write insurance and meet client requests for larger insurance coverage while reducing the impact of potential losses and protecting the insurance company against catastrophic losses. Reinsurance also allows the primary insurer to benefit from underwriting skills, expertise and proficient complex claim file handling of the larger reinsurance companies.

Example 5.1.6. Actuarial Exam Question. Losses arising in a certain portfolio have a two-parameter Pareto distribution with $\alpha = 5$ and $\theta = 3,600$. A reinsurance arrangement has been made, under which (a) the reinsurer accepts 15% of losses up to $u = 5,000$ and all amounts in excess of 5,000 and (b) the insurer pays for the remaining losses.

a) Express the random variables for the reinsurer's and the insurer's payments as a function of X, the portfolio losses.
b) Calculate the mean amount paid on a single claim by the insurer.
c) By assuming that the upper limit is $u = \infty$, calculate an upper bound on the standard deviation of the amount paid on a single claim by the insurer (retaining the 15% copayment).

Example Solution.

a). The reinsurer's portion is

$$Y_{reinsurer} = \left\{ \begin{array}{cc} 0.15X & X < 5000, \\ 0.15(5000) + X - 5000 & X \geq 5000 \end{array} \right. .$$

and the insurer's portion is

$$Y_{insurer} = \left\{ \begin{array}{cc} 0.85X & X < 5000, \\ 0.85(5000) & X \geq 5000 \end{array} \right. = 0.85(X \wedge 5000).$$

b) Using the limited expected value tables for the Pareto distribution, we have

$$\begin{aligned}\mathrm{E}\left(Y_{insurer}\right) &= 0.85\ \mathrm{E}\left(X \wedge 5000\right) = 0.85\ \tfrac{\theta}{\alpha-1}\left[1-\left(\tfrac{\theta}{5000+\theta}\right)^{\alpha-1}\right] \\ &= 0.85\ \tfrac{3600}{5-1}\left[1-\left(\tfrac{3600}{5000+3600}\right)^{5-1}\right] = 741.5103.\end{aligned}$$

c) The unlimited variable is $0.85X$. For the first moment, we have

$$0.85\ \mathrm{E}\ X = 0.85\ \frac{\theta}{\alpha-1} = 0.85\ \frac{3600}{5-1} = 765.$$

For the second moment of the unlimited variable, we use the table of distributions to get

$$0.85^2\ \mathrm{E}\ X^2 = 0.85^2\ \frac{\theta^2\Gamma(2+1)\Gamma(\alpha-2)}{\Gamma(\alpha)} = 0.85^2\ \frac{3600^2\cdot 2\cdot 2}{24} = 1560600.$$

Thus, the variance is $1560600 - 765^2 = 975375$. Alternatively, you can use the formula

$$0.85^2\ \mathrm{Var}\ X = 0.85^2\ \frac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)} = 0.85^2\ \frac{5(3600^2)}{(5-1)^2(5-2)} = 975375.$$

Taking square roots, the standard deviation is $\sqrt{975375} \approx 987.6108$.

Further discussions of reinsurance will be provided in Section 13.4.

5.2 Parametric Estimation using Modified Data

In this section, you learn how to:

- Describe grouped, censored, and truncated data
- Estimate parametric distributions based on grouped, censored, and truncated data

Basic theory and many applications are based on *individual* observations that are "*complete*" and "*unmodified*," as we have seen in the Chapter 4. Section 5.1.1 introduced the concept of observations that are "*modified*" due to two common types of limitations: **censoring** and **truncation**. For example, it is common to think about an insurance deductible as producing data that are truncated (from the left) or policy limits as yielding data that are censored (from the right). This viewpoint is from the primary insurer (the seller of the insurance). Another viewpoint is that of a reinsurer (an insurer of an insurance company) that will be discussed more in Chapter 13. A reinsurer may not observe a claim smaller than an amount, only that a claim exists; this is an example of censoring from the left. So, in this section, we cover the full gamut of alternatives. Specifically, this section will address parametric estimation methods for three alternatives to individual, complete, and unmodified data: **interval-censored** data available only in groups, data that are limited or **censored**, and data that may not be observed due to **truncation**.

5.2.1 Parametric Estimation using Grouped Data

Consider a sample of size n observed from the distribution $F(\cdot)$, but in groups so that we only know the group into which each observation fell, not the exact value. This is referred to as **grouped** or **interval-censored** data. For example, we may be looking at two successive years of annual employee records. People employed in the first year but not the second have left sometime during the year. With an exact departure date (individual data), we could compute the amount of time that they were with the firm. Without the departure date (grouped data), we only know that they departed sometime during a year-long interval.

Formalizing this idea, suppose there are k groups or intervals delimited by boundaries $c_0 < c_1 < \cdots < c_k$. For each observation, we only observe the interval into which it fell (e.g. (c_{j-1}, c_j)), not the exact value. Thus, we only know the number of observations in each interval. The constants $\{c_0 < c_1 < \cdots < c_k\}$ form some partition of the domain of $F(\cdot)$. Then the probability of an observation X_i falling in the jth interval is

$$\Pr\left(X_i \in (c_{j-1}, c_j]\right) = F(c_j) - F(c_{j-1}).$$

The corresponding probability mass function for an observation is

$$\begin{aligned} f(x) &= \begin{cases} F(c_1) - F(c_0) & \text{if } \ x \in (c_0, c_1] \\ \vdots & \vdots \\ F(c_k) - F(c_{k-1}) & \text{if } \ x \in (c_{k-1}, c_k] \end{cases} \\ &= \prod_{j=1}^{k} \left\{F(c_j) - F(c_{j-1})\right\}^{I(x \in (c_{j-1}, c_j])} \end{aligned}$$

Now, define n_j to be the number of observations that fall in the jth interval,

$(c_{j-1}, c_j]$. Thus, the likelihood function (with respect to the parameter(s) θ) is

$$L(\theta) = \prod_{j=1}^{n} f(x_i) = \prod_{j=1}^{k} \{F(c_j) - F(c_{j-1})\}^{n_j}$$

And the log-likelihood function is

$$l(\theta) = \log L(\theta) = \log \prod_{j=1}^{n} f(x_i) = \sum_{j=1}^{k} n_j \log \{F(c_j) - F(c_{j-1})\}$$

Maximizing the likelihood function (or equivalently, maximizing the log-likelihood function) would then produce the maximum likelihood estimates for grouped data.

Example 5.2.1. Actuarial Exam Question. You are given:

(i) Losses follow an exponential distribution with mean θ.
(ii) A random sample of 20 losses is distributed as follows:

Loss Range	Frequency
$[0, 1000]$	7
$(1000, 2000]$	6
$(2000, \infty)$	7

Calculate the maximum likelihood estimate of θ.

Example Solution.

$$\begin{aligned}
L(\theta) &= F(1000)^7[F(2000) - F(1000)]^6[1 - F(2000)]^7 \\
&= (1 - e^{-1000/\theta})^7(e^{-1000/\theta} - e^{-2000/\theta})^6(e^{-2000/\theta})^7 \\
&= (1 - p)^7(p - p^2)^6(p^2)^7 \\
&= p^{20}(1 - p)^{13}
\end{aligned}$$

where $p = e^{-1000/\theta}$. Maximizing this expression with respect to p is equivalent to maximizing the likelihood with respect to θ. The maximum occurs at $p = \frac{20}{33}$ and so $\hat{\theta} = \frac{-1000}{\log(20/33)} = 1996.90$.

5.2.2 Censored Data

Censoring occurs when we record only a limited value of an observation. The most common form is **right-censoring**, in which we record the *smaller* of the "true" dependent variable and a censoring value. Using notation, let X

represent an outcome of interest, such as the loss due to an insured event or time until an event. Let C_U denote the censoring amount. With right-censored observations, we record $X_U^* = \min(X, C_U) = X \wedge C_U$. We also record whether or not censoring has occurred. Let $\delta_U = I(X \leq C_U)$ be a binary variable that is 0 if censoring occurs and 1 if it does not, that is, δ_U indicates whether or not X is uncensored.

For an example that we saw in Section 5.1.2, C_U may represent the upper limit of coverage of an insurance policy (we used u for the upper limit in that section). The loss may exceed the amount C_U, but the insurer only has C_U in its records as the amount paid out and does not have the amount of the actual loss X in its records.

Similarly, with **left-censoring**, we record the *larger* of a variable of interest and a censoring variable. If C_L is used to represent the censoring amount, we record $X_L^* = \max(X, C_L)$ along with the censoring indicator $\delta_L = I(X > C_L)$.

As an example, we gave a brief introduction to reinsurance (insurance for insurers) in Section 5.1.4 and more is given in Chapter 13. Suppose a reinsurer will cover insurer losses greater than C_L; this means that the reinsurer is responsible for the excess of X_L^* over C_L. Using notation, the loss of the reinsurer is $Y = X_L^* - C_L$. To see this, first consider the case where the policyholder loss $X < C_L$. Then, the insurer will pay the entire claim and $Y = C_L - C_L = 0$, no loss for the reinsurer. For contrast, if the loss $X \geq C_L$, then $Y = X - C_L$ represents the reinsurer's retained claims. Put another way, if a loss occurs, the reinsurer records the actual amount if it exceeds the limit C_L and otherwise it only records that it had a loss of 0.

5.2.3 Truncated Data

Censored observations are recorded for study, although in a limited form. In contrast, **truncated** outcomes are a type of missing data. An outcome is potentially truncated when the availability of an observation depends on the outcome.

In insurance, it is common for observations to be **left-truncated** at C_L when the amount is

$$Y = \begin{cases} \text{we do not observe } X & X \leq C_L \\ X & X > C_L \end{cases} .$$

In other words, if X is less than the threshold C_L, then it is not observed.

For an example we saw in Section 5.1.1, C_L may represent the deductible of an insurance policy (we used d for the deductible in that section). If the insured loss is less than the deductible, then the insurer may not observe or record the loss at all. If the loss exceeds the deductible, then the excess $X - C_L$ is the claim that the insurer covers. In Section 5.1.1, we defined the per payment loss to be

$$Y^P = \begin{cases} \text{Undefined} & X \le d \\ X - d & X > d \end{cases},$$

so that if a loss exceeds a deductible, we record the excess amount $X - d$. This is very important when considering amounts that the insurer will pay. However, for estimation purposes of this section, it matters little if we subtract a known constant such as $C_L = d$. So, for our truncated variable Y, we use the simpler convention and do not subtract d.

Similarly for **right-truncated** data, if X exceeds a threshold C_U, then it is not observed. In this case, the amount is

$$Y = \begin{cases} X & X \le C_U \\ \text{we do not observe } X & X > C_U. \end{cases}$$

Classic examples of truncation from the right include X as a measure of distance to a star. When the distance exceeds a certain level C_U, the star is no longer observable.

Figure 5.1 compares truncated and censored observations.

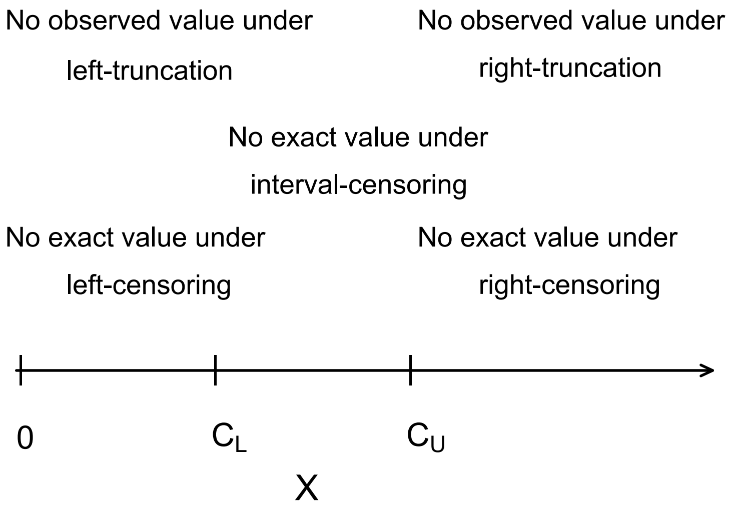


FIGURE 5.1: **Censoring and Truncation**

Example – Mortality Study. Suppose that you are conducting a two-year study of mortality of high-risk subjects, beginning January 1, 2010 and finishing January 1, 2012. Figure 5.2 graphically portrays the six types of subjects recruited. For each subject, the beginning of the arrow represents that the subject was recruited and the arrow end represents the event time where in this example the event represents death. The arrow represents exposure time.

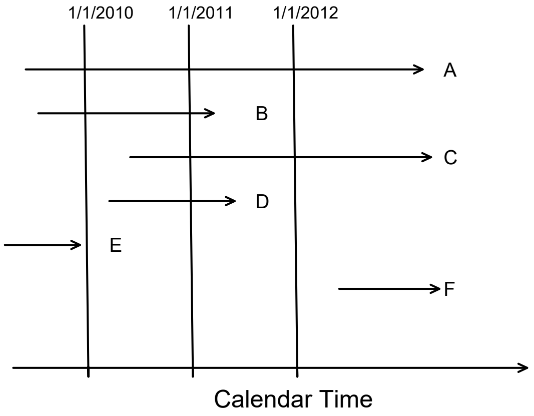


FIGURE 5.2: **Timeline for Several Subjects on Test in a Mortality Study**

- **Type A - Right-censored.** This subject is alive at the beginning and the end of the study. Because the time of death is not known by the end of the study, it is right-censored. Most subjects are Type A.
- **Type B - Complete** information is available for a type B subject. The subject is alive at the beginning of the study and the death occurs within the observation period.
- **Type C - Right-censored and left-truncated.** A type C subject is right-censored, in that death occurs after the observation period. However, the subject entered after the start of the study and is said to have a *delayed entry time*. Because the subject would not have been observed had death occurred before entry, it is left-truncated.
- **Type D - Left-truncated.** A type D subject also has delayed entry. Because death occurs within the observation period, this subject is not right censored.
- **Type E - Left-truncated.** A type E subject is not included in the study because death occurs prior to the observation period.
- **Type F - Right-truncated.** Similarly, a type F subject is not included because the entry time occurs after the observation period.

To summarize, for outcome X and constants C_L and C_U,

Limitation Type	Limited Variable	Recording Information
right censoring	$X_U^* = \min(X, C_U)$	$\delta_U = I(X \le C_U)$
left censoring	$X_L^* = \max(X, C_L)$	$\delta_L = I(X > C_L)$
interval censoring		
right truncation	X	observe X if $X \le C_U$
left truncation	X	observe X if $X > C_L$

5.2.4 Parametric Estimation using Censored and Truncated Data

For simplicity, we assume non-random censoring amounts and a continuous outcome X. To begin, consider the case of right-censored data where we record $X_U^* = \min(X, C_U)$ and censoring indicator $\delta = I(X \le C_U)$. If censoring occurs so that $\delta = 0$, then $X > C_U$ and the likelihood is $\Pr(X > C_U) = 1 - F(C_U)$. If censoring does not occur so that $\delta = 1$, then $X \le C_U$ and the likelihood is $f(x)$. Summarizing, we have the likelihood of a single observation as

$$\begin{cases} 1 - F(C_U) & \text{if } \delta = 0 \\ f(x) & \text{if } \delta = 1 \end{cases} = \{f(x)\}^{\delta} \{1 - F(C_U)\}^{1-\delta}.$$

The right-hand expression allows us to present the likelihood more compactly. Now, for an *iid* sample of size n, the likelihood is

$$L(\theta) = \prod_{i=1}^{n} \{f(x_i)\}^{\delta_i} \{1 - F(C_{Ui})\}^{1-\delta_i} = \prod_{\delta_i=1} f(x_i) \prod_{\delta_i=0} \{1 - F(C_{Ui})\},$$

with potential censoring times $\{C_{U1}, \ldots, C_{Un}\}$. Here, the notation "$\prod_{\delta_i=1}$" means to take the product over uncensored observations, and similarly for "$\prod_{\delta_i=0}$."

On the other hand, truncated data are handled in likelihood inference via conditional probabilities. Specifically, we adjust the likelihood contribution by dividing by the probability that the variable was observed. To summarize, we have the following contributions to the likelihood function for six types of outcomes:

Outcome	Likelihood Contribution
exact value	$f(x)$
right-censoring	$1 - F(C_U)$
left-censoring	$F(C_L)$
right-truncation	$f(x)/F(C_U)$
left-truncation	$f(x)/(1 - F(C_L))$
interval-censoring	$F(C_U) - F(C_L)$

For known outcomes and censored data, the likelihood is

$$L(\theta) = \prod_{E} f(x_i) \prod_{R} \{1 - F(C_{Ui})\} \prod_{L} F(C_{Li}) \prod_{I} (F(C_{Ui}) - F(C_{Li})),$$

where "$\prod_E$" is the product over observations with *E*xact values, and similarly for *R*ight-, *L*eft- and *I*nterval-censoring.

For right-censored and left-truncated data, the likelihood is

$$L(\theta) = \prod_E \frac{f(x_i)}{1 - F(C_{Li})} \prod_R \frac{1 - F(C_{Ui})}{1 - F(C_{Li})},$$

and similarly for other combinations. To get further insights, consider the following.

Special Case: Exponential Distribution. Consider data that are right-censored and left-truncated, with random variables X_i that are exponentially distributed with mean θ. With these specifications, recall that $f(x) = \theta^{-1}\exp(-x/\theta)$ and $F(x) = 1 - \exp(-x/\theta)$.

For this special case, the log-likelihood is

$$\begin{aligned} l(\theta) &= \sum_E \left\{ \log f(x_i) - \log(1 - F(C_{Li})) \right\} + \sum_R \left\{ \log(1 - F(C_{Ui})) - \log(1 - \mathrm{F}(C_{Li})) \right\} \\ &= \sum_E (-\log\theta - (x_i - C_{Li})/\theta) - \sum_R (C_{Ui} - C_{Li})/\theta. \end{aligned}$$

To simplify the notation, define $\delta_i = I(X_i < C_{Ui})$ to be a binary variable that indicates right-censoring. Let $X_i^{**} = \min(X_i, C_{Ui}) - C_{Li}$ be the amount that the observed variable exceeds the lower truncation limit. With this, the log-likelihood is

$$l(\theta) = -\sum_{i=1}^{n} \left((1 - \delta_i)\log\theta + \frac{x_i^{**}}{\theta} \right) \tag{5.1}$$

Taking derivatives with respect to the parameter θ and setting it equal to zero yields the maximum likelihood estimator

$$\widehat{\theta} = \frac{1}{n_u} \sum_{i=1}^{n} x_i^{**},$$

where $n_u = \sum_i (1 - \delta_i)$ is the number of uncensored observations.

Example 5.2.2. Actuarial Exam Question. You are given:

(i) A sample of losses is: 600 700 900
(ii) No information is available about losses of 500 or less.
(iii) Losses are assumed to follow an exponential distribution with mean θ.

Calculate the maximum likelihood estimate of θ.

Example Solution. These observations are truncated at 500. The contribution of each observation to the likelihood function is

$$\frac{f(x)}{1-F(500)} = \frac{\theta^{-1}e^{-x/\theta}}{e^{-500/\theta}}$$

Then the likelihood function is

$$L(\theta) = \frac{\theta^{-1}e^{-600/\theta}\theta^{-1}e^{-700/\theta}\theta^{-1}e^{-900/\theta}}{(e^{-500/\theta})^3} = \theta^{-3}e^{-700/\theta}$$

The log-likelihood is

$$l(\theta) = \log L(\theta) = -3\log\theta - 700\theta^{-1}$$

Maximizing this expression by setting the derivative with respect to θ equal to 0, we have

$$L'(\theta) = -3\theta^{-1} + 700\theta^{-2} = 0 \ \Rightarrow\ \hat{\theta} = \frac{700}{3} = 233.33.$$

Example 5.2.3. Actuarial Exam Question. You are given the following information about a random sample:

(i) The sample size equals five.
(ii) The sample is from a Weibull distribution with $\tau = 2$.
(iii) Two of the sample observations are known to exceed 50, and the remaining three observations are 20, 30, and 45.

Calculate the maximum likelihood estimate of θ.

Example Solution. The likelihood function is

$$\begin{aligned}
L(\theta) &= f(20)f(30)f(45)[1-F(50)]^2 \\
&= \frac{2(20/\theta)^2e^{-(20/\theta)^2}}{20}\frac{2(30/\theta)^2e^{-(30/\theta)^2}}{30}\frac{2(45/\theta)^2e^{-(45/\theta)^2}}{45}(e^{-(50/\theta)^2})^2 \\
&\propto \frac{1}{\theta^6}e^{-8325/\theta^2}
\end{aligned}$$

The natural logarithm of the above expression is $-6\log\theta - \frac{8325}{\theta^2}$. Maximizing this expression by setting its derivative to 0, we get

$$\frac{-6}{\theta} + \frac{16650}{\theta^3} = 0 \Rightarrow \hat{\theta} = \left(\frac{16650}{6}\right)^{\frac{1}{2}} = 52.6783.$$

5.3 Nonparametric Estimation using Modified Data

In this section, you learn how to:

- Estimate the distribution function for grouped data using the ogive.
- Create a nonparametric estimator of the loss elimination ratio using the plug-in principle.
- Apply the Kaplan-Meier product-limit and the Nelson Aalon estimators to estimate the distribution function in the presence of censoring.
- Apply Greenwood's formula to estimate the variance of the product-limit estimator.

Nonparametric estimators provide useful benchmarks, so it is helpful to understand the estimation procedures for grouped, censored, and truncated data.

5.3.1 Grouped Data

As we have seen in Section 5.2.1, observations may be grouped (also referred to as interval censored) in the sense that we only observe them as belonging in one of k intervals of the form $(c_{j-1}, c_j]$, for $j = 1, \ldots, k$. At the boundaries, the empirical distribution function is defined in the usual way:

$$F_n(c_j) = \frac{\text{number of observations } \leq c_j}{n}.$$

Ogive Estimator. For other values of $x \in (c_{j-1}, c_j)$, we can estimate the distribution function with the ogive estimator, which linearly interpolates between $F_n(c_{j-1})$ and $F_n(c_j)$, i.e. the values of the boundaries $F_n(c_{j-1})$ and $F_n(c_j)$ are connected with a straight line. This can formally be expressed as

$$F_n(x) = \frac{c_j - x}{c_j - c_{j-1}} F_n(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}} F_n(c_j) \quad \text{for } c_{j-1} \leq x < c_j$$

The corresponding density is

$$f_n(x) = F_n'(x) = \frac{F_n(c_j) - F_n(c_{j-1})}{c_j - c_{j-1}} \quad \text{for } c_{j-1} < x < c_j.$$

Example 5.3.1. Actuarial Exam Question. You are given the following information regarding claim sizes for 100 claims:

Claim Size	Number of Claims
$0 - 1,000$	16
$1,000 - 3,000$	22
$3,000 - 5,000$	25
$5,000 - 10,000$	18
$10,000 - 25,000$	10
$25,000 - 50,000$	5
$50,000 - 100,000$	3
over $100,000$	1

Using the ogive, calculate the estimate of the probability that a randomly chosen claim is between 2000 and 6000.

Example Solution. At the boundaries, the empirical distribution function is defined in the usual way, so we have

$$F_{100}(1000) = 0.16,\ F_{100}(3000) = 0.38,\ F_{100}(5000) = 0.63,\ F_{100}(10000) = 0.81.$$

For other claim sizes, the ogive estimator linearly interpolates between these values:

$$\begin{aligned} F_{100}(2000) &= 0.5F_{100}(1000) + 0.5F_{100}(3000) = 0.5(0.16) + 0.5(0.38) = 0.27 \\ F_{100}(6000) &= 0.8F_{100}(5000) + 0.2F_{100}(10000) = 0.8(0.63) + 0.2(0.81) = 0.666 \end{aligned}$$

Thus, the probability that a claim is between 2000 and 6000 is $F_{100}(6000) - F_{100}(2000) = 0.666 - 0.27 = 0.396$.

5.3.2 Plug-in Principle

One way to create a nonparametric estimator of some quantity is to use the *analog* or plug-in principle where one replaces the unknown cdf F with a known estimate such as the empirical *cdf* F_n. So, if we are trying to estimate $\mathrm{E}\left[\mathrm{g}(X)\right] = \mathrm{E}_F\left[\mathrm{g}(X)\right]$ for a generic function g, then we define a nonparametric estimator to be $\mathrm{E}_{F_n}\left[\mathrm{g}(X)\right] = n^{-1}\sum_{i=1}^{n} \mathrm{g}(X_i)$.

To see how this works, as a special case of g we consider the loss per payment random variable is $Y = (X-d)_+$ and the *loss elimination ratio* introduced in Section 4.4.1. We can express this as

$$LER(d) = \frac{\mathrm{E}\left[X-(X-d)_+\right]}{\mathrm{E}\left[X\right]} = \frac{\mathrm{E}\left[\min(X,d)\right]}{\mathrm{E}\left[X\right]},$$

for a fixed deductible d.

Example. 5.3.2. Bodily Injury Claims and Loss Elimination Ratios

We use a sample of 432 closed auto claims from Boston from Derrig et al. (2001). Losses are recorded for payments due to bodily injuries in auto accidents. Losses are not subject to deductibles but are limited by various maximum coverage amounts that are also available in the data. It turns out that only 17 out of 432 ($\approx$ 4%) were subject to these policy limits and so we ignore these data for this illustration.

The average loss paid is 6906 in U.S. dollars. Figure 5.3 shows other aspects of the distribution. Specifically, the left-hand panel shows the empirical distribution function, the right-hand panel gives a nonparametric density plot.

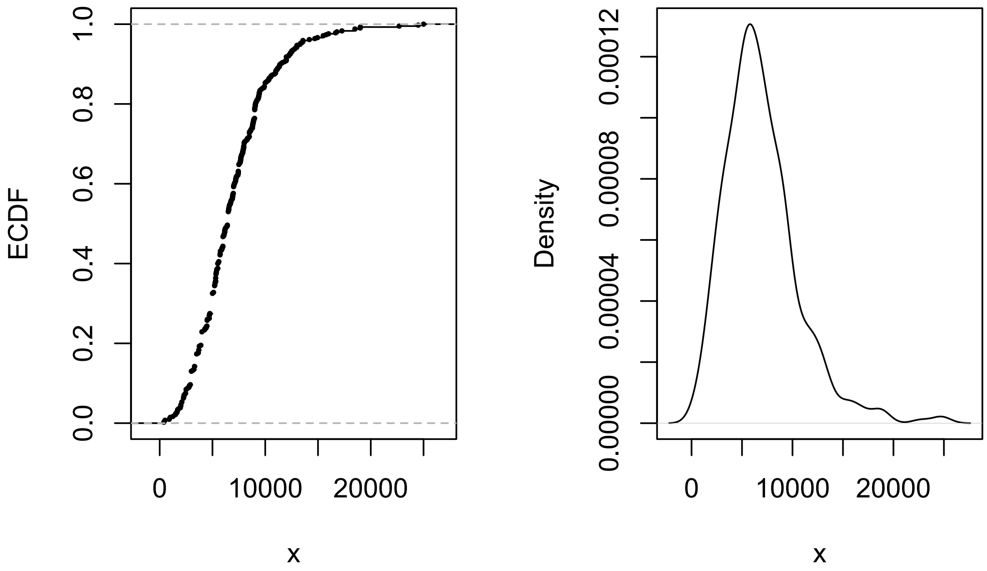


FIGURE 5.3: **Bodily Injury Claims.** The left-hand panel gives the empirical distribution function. The right-hand panel presents a nonparametric density plot.

The impact of bodily injury losses can be mitigated by the imposition of limits or purchasing reinsurance policies (see Section 10.3). To quantify the impact of these risk mitigation tools, it is common to compute the *loss elimination ratio (LER)* as introduced in Section 4.4.1. The distribution function is not available and so must be estimated in some way. Using the plug-in principle, a nonparametric estimator can be defined as

$$LER_n(d) = \frac{n^{-1}\sum_{i=1}^n \min(X_i, d)}{n^{-1}\sum_{i=1}^n X_i} = \frac{\sum_{i=1}^n \min(X_i, d)}{\sum_{i=1}^n X_i}.$$

Figure 5.4 shows the estimator $LER_n(d)$ for various choices of d. For example, at $d = 1,000$, we have $LER_n(1000) \approx 0.1442$. Thus, imposing a limit of 1,000 means that expected retained claims are 14.42 percent lower when compared to expected claims with a zero deductible.

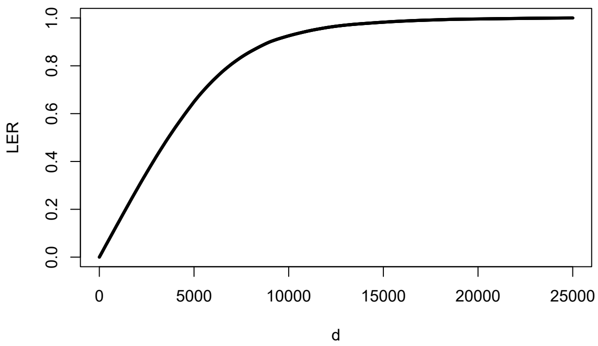


FIGURE 5.4: **LER for Bodily Injury Claims.** The figure presents the loss elimination ratio (LER) as a function of deductible d.

5.3.3 Right-Censored Empirical Distribution Function

It can be useful to calibrate parametric estimators with nonparametric methods that do not rely on a parametric form of the distribution. The product-limit estimator due to (Kaplan and Meier, 1958) is a well-known estimator of the distribution function in the presence of censoring.

Motivation for the Kaplan-Meier Product Limit Estimator. To explain why the product-limit works so well with censored observations, let us first return to the "usual" case without censoring. Here, the empirical distribution function $F_n(x)$ is an *unbiased* estimator of the distribution function $F(x)$. This is because $F_n(x)$ is the average of indicator variables each of which are unbiased, that is, $\mathrm{E}\,[I(X_i \le x)] = \Pr(X_i \le x) = F(x)$.

Now suppose the random outcome is censored on the right by a limiting amount, say, C_U, so that we record the smaller of the two, $X^* = \min(X, C_U)$. For values of x that are smaller than C_U, the indicator variable still provides an unbiased estimator of the distribution function before we reach the censoring limit. That is, $\mathrm{E}\left[I(X^* \leq x)\right] = F(x)$ because $I(X^* \leq x) = I(X \leq x)$ for $x < C_U$. In the same way, $\mathrm{E}\left[I(X^* > x)\right] = 1 - F(x) = S(x)$. But, for $x > C_U$, $I(X^* \leq x)$ is in general not an unbiased estimator of $F(x)$.

As an alternative, consider *two* random variables that have different censoring limits. For illustration, suppose that we observe $X_1^* = \min(X_1, 5)$ and $X_2^* = \min(X_2, 10)$ where X_1 and X_2 are independent draws from the same distribution. For $x \leq 5$, the empirical distribution function $F_2(x)$ is an unbiased estimator of $F(x)$. However, for $5 < x \leq 10$, the first observation cannot be used for the distribution function because of the censoring limitation. Instead, the strategy developed by (Kaplan and Meier, 1958) is to use $S_2(5)$ as an estimator of $S(5)$ and then to use the second observation to estimate the survival function conditional on survival to time 5, $\Pr(X > x | X > 5) = \frac{S(x)}{S(5)}$. Specifically, for $5 < x \leq 10$, the estimator of the survival function is

$$\hat{S}(x) = S_2(5) \times I(X_2^* > x).$$

Kaplan-Meier Product Limit Estimator. Extending this idea, for each observation i, let u_i be the upper censoring limit ($= \infty$ if no censoring). Thus, the recorded value is x_i in the case of no censoring and u_i if there is censoring. Let $t_1 < \cdots < t_k$ be k distinct points at which an uncensored loss occurs, and let s_j be the number of uncensored losses x_i's at t_j. The corresponding risk set is the number of observations that are active (not censored) at a value *less than* t_j, denoted as $R_j = \sum_{i=1}^n I(x_i \geq t_j) + \sum_{i=1}^n I(u_i \geq t_j)$.

With this notation, the **product-limit estimator** of the distribution function is

$$\hat{F}(x) = \begin{cases} 0 & x < t_1 \\ 1 - \prod_{j:t_j \leq x} \left(1 - \frac{s_j}{R_j}\right) & x \geq t_1 \end{cases} . \tag{5.2}$$

For example, if x is smaller than the smallest uncensored loss, then $x < t_1$ and $\hat{F}(x) = 0$. As another example, if x falls between then second and third smallest uncensored losses, then $x \in (t_2, t_3]$ and $\hat{F}(x) = 1 - \left(1 - \frac{s_1}{R_1}\right)\left(1 - \frac{s_2}{R_2}\right)$.

As usual, the corresponding estimate of the survival function is $\hat{S}(x) = 1 - \hat{F}(x)$.

Example 5.3.3. Actuarial Exam Question. The following is a sample of 10 payments:

4 4 5+ 5+ 5+ 8 10+ 10+ 12 15

where + indicates that a loss has exceeded the policy limit.

Using the Kaplan-Meier product-limit estimator, calculate the probability that the loss on a policy exceeds 11, $\hat{S}(11)$.

Example Solution. There are four event times (non-censored observations). For each time t_j, we can calculate the number of events s_j and the risk set R_j as the following:

j	t_j	s_j	R_j
1	4	2	10
2	8	1	5
3	12	1	2
4	15	1	1

Thus, the Kaplan-Meier estimate of $S(11)$ is

$$\begin{aligned}\hat{S}(11) &= \prod_{j:t_j \leq 11} \left(1 - \frac{s_j}{R_j}\right) = \prod_{j=1}^{2} \left(1 - \frac{s_j}{R_j}\right) \\ &= \left(1 - \frac{2}{10}\right)\left(1 - \frac{1}{5}\right) = (0.8)(0.8) = 0.64.\end{aligned}$$

Example. 5.3.4. Bodily Injury Claims. We consider again the Boston auto bodily injury claims data from Derrig et al. (2001) that was introduced in Example 5.1.11. In that example, we omitted the 17 claims that were censored by policy limits. Now, we include the full dataset and use the Kaplan-Meier product limit to estimate the survival function. This is given in Figure 5.5.

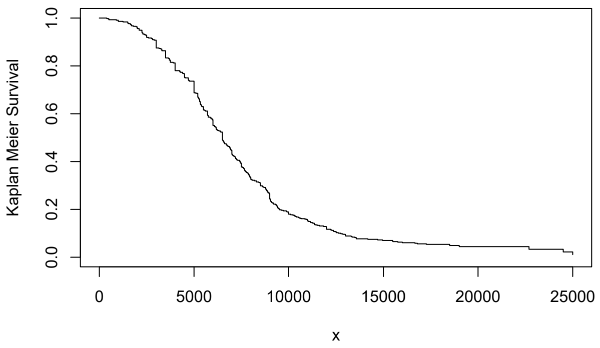


FIGURE 5.5: **Kaplan-Meier Estimate of the Survival Function for Bodily Injury Claims**

Right-Censored, Left-Truncated Empirical Distribution Function. In addition to right-censoring, we now extend the framework to allow for left-truncated data. As before, for each observation i, let u_i be the upper censoring limit ($= \infty$ if no censoring). Further, let d_i be the lower truncation limit (0 if no truncation). Thus, the recorded value (if it is greater than d_i) is x_i in the case of no censoring and u_i if there is censoring. Let $t_1 < \cdots < t_k$ be k distinct points at which an event of interest occurs, and let s_j be the number of recorded events x_i's at time point t_j. The corresponding risk set is

$$R_j = \sum_{i=1}^{n} I(x_i \geq t_j) + \sum_{i=1}^{n} I(u_i \geq t_j) - \sum_{i=1}^{n} I(d_i \geq t_j).$$

With this new definition of the risk set, the product-limit estimator of the distribution function is as in equation (5.2).

Greenwood's Formula. (Greenwood, 1926) derived the formula for the estimated variance of the product-limit estimator to be

$$\widehat{Var}(\hat{F}(x)) = (1-\hat{F}(x))^2 \sum_{j:t_j \leq x} \frac{s_j}{R_j(R_j - s_j)}.$$

As usual, we refer to the square root of the estimated variance as a *standard error*, a quantity that is routinely used in confidence intervals and for hypothesis testing. To compute this, `R`'s `survfit` method takes a survival data object and creates a new object containing the Kaplan-Meier estimate of the survival function along with confidence intervals. The Kaplan-Meier method (`type='kaplan-meier'`) is used by default to construct an estimate of the survival curve. The resulting discrete survival function has point masses at the observed event times (discharge dates) t_j, where the probability of an event given survival to that duration is estimated as the number of observed events at the duration s_j divided by the number of subjects exposed or 'at-risk' just prior to the event duration R_j.

Alternative Estimators. Two alternate types of estimation are also available for the `survfit` method. The alternative (`type='fh2'`) handles ties, in essence, by assuming that multiple events at the same duration occur in some arbitrary order. Another alternative (`type='fleming-harrington'`) uses the Nelson-Aalen (see (Aalen, 1978)) estimate of the **cumulative hazard function** to obtain an estimate of the survival function. The estimated cumulative hazard $\hat{H}(x)$ starts at zero and is incremented at each observed event duration t_j by the number of events s_j divided by the number at risk R_j. With the same notation as above, the **Nelson-Äalen** estimator of the distribution function is

$$\hat{F}_{NA}(x) = \begin{cases} 0 & x < t_1 \\ 1 - \exp\left(-\sum_{j:t_j \leq x} \frac{s_j}{R_j}\right) & x \geq t_1 \end{cases}.$$

Note that the above expression is a result of the Nelson-Äalen estimator of the cumulative hazard function

$$\hat{H}(x) = \sum_{j:t_j \leq x} \frac{s_j}{R_j}$$

and the relationship between the survival function and cumulative hazard function, $\hat{S}_{NA}(x) = e^{-\hat{H}(x)}$.

Example 5.3.5. Actuarial Exam Question. For observation i of a survival study:

- d_i is the left truncation point
- x_i is the observed value if not right censored
- u_i is the observed value if right censored

You are given:

Observation (i)	1	2	3	4	5	6	7	8	9	10
d_i	0	0	0	0	0	0	0	1.3	1.5	1.6
x_i	0.9	–	1.5	–	–	1.7	–	2.1	2.1	–
u_i	–	1.2	–	1.5	1.6	–	1.7	–	–	2.3

Calculate the Kaplan-Meier product-limit estimate, $\hat{S}(1.6)$

Example Solution. Recall the risk set $R_j = \sum_{i=1}^{n} \{I(x_i \geq t_j) + I(u_i \geq t_j) - I(d_i \geq t_j)\}$. Then

j	t_j	s_j	R_j	$\hat{S}(t_j)$
1	0.9	1	$10-3=7$	$1-\frac{1}{7}=\frac{6}{7}$
2	1.5	1	$8-2=6$	$\frac{6}{7}\left(1-\frac{1}{6}\right)=\frac{5}{7}$
3	1.7	1	$5-0=5$	$\frac{5}{7}\left(1-\frac{1}{5}\right)=\frac{4}{7}$
4	2.1	2	3	$\frac{4}{7}\left(1-\frac{2}{3}\right)=\frac{4}{21}$

The Kaplan-Meier estimate is therefore $\hat{S}(1.6) = \frac{5}{7}$.

Example 5.3.6. Actuarial Exam Question. - Continued.

a) Using the Nelson-Äalen estimator, calculate the probability that the loss on a policy exceeds 11, $\hat{S}_{NA}(11)$.
b) Calculate Greenwood's approximation to the variance of the product-limit estimate $\hat{S}(11)$.

Example Solution. As before, there are four event times (non-censored observations). For each time t_j, we can calculate the number of events s_j and the risk set R_j as the following:

j	t_j	s_j	R_j
1	4	2	10
2	8	1	5
3	12	1	2
4	15	1	1

The Nelson-Aalen estimate of $S(11)$ is $\hat{S}_{NA}(11) = e^{-\hat{H}(11)} = e^{-0.4} = 0.67$, since

$$\begin{aligned}\hat{H}(11) &= \sum_{j:t_j \leq 11} \frac{s_j}{R_j} = \sum_{j=1}^{2} \frac{s_j}{R_j} \\ &= \frac{2}{10} + \frac{1}{5} = 0.2 + 0.2 = 0.4.\end{aligned}$$

From earlier work, the Kaplan-Meier estimate of $S(11)$ is $\hat{S}(11) = 0.64$. Then Greenwood's estimate of the variance of the product-limit estimate of $S(11)$ is

$$\widehat{Var}(\hat{S}(11)) = (\hat{S}(11))^2 \sum_{j:t_j \leq 11} \frac{s_j}{R_j(R_j - s_j)} \quad = (0.64)^2 \left(\frac{2}{10(8)} + \frac{1}{5(4)} \right) = 0.0307.$$

5.4 Further Resources and Contributors

Exercises

Contributors

- **Zeinab Amin**, The American University in Cairo, is the principal author of this chapter. Email: zeinabha@aucegypt.edu for chapter comments and suggested improvements.
- **Edward W. (Jed) Frees** and **Lisa Gao**, University of Wisconsin-Madison, are the principal authors of the sections on estimation using modified data which appeared in chapter 4 of the first edition of the text.

- Chapter reviewers include: Vytaras Brazauskas, Yvonne Chueh, Eren Dodd, Hirokazu (Iwahiro) Iwasawa, Joseph Kim, Andrew Kwon-Nakamura, Jiandong Ren, and Di (Cindy) Xu.

Further Readings and References

If you would like additional practice with `R` coding, please visit our companion LDA Short Course. In particular, see the Model Selection and Estimation Chapter.

6

Model Selection

Chapter Preview. Model selection is a fundamental aspect of statistical modeling. In this chapter, the process of model selection is summarized, including tools for model comparisons and diagnostics. In addition to nonparametric tools for model selection based on marginal distributions of outcomes ignoring explanatory variables, this chapter underscores the idea that model selection is an iterative process in which models are cyclically (re)formulated and tested for appropriateness before using them for inference. After an overview, we describe the model selection process based on:

- an in-sample or training dataset,
- an out-of-sample or test dataset, and
- a method that combines these approaches known as cross-validation.

Although our focus is predominantly on data from continuous distributions, the same process can be used for discrete versions or data that come from a hybrid combination of discrete and continuous distributions.

In this chapter, you learn how to:

- Determine measures that summarize deviations of a parametric from a nonparametric fit
- Describe the iterative model selection specification process
- Outline steps needed to select a parametric model
- Describe pitfalls of model selection based purely on in-sample data when compared to the advantages of out-of-sample model validation

6.1 Tools for Model Selection and Diagnostics

Section 4.1.1 introduced nonparametric estimators in which there was no parametric form assumed about the underlying distributions. However, in many actuarial applications, analysts seek to employ a parametric fit of a distribution for ease of explanation and the ability to readily extend it to more complex situations such as including explanatory variables in a regression setting. When fitting a parametric distribution, one analyst might try to use a gamma distribution to represent a set of loss data. However, another analyst may prefer to use a Pareto distribution. How does one determine which model to select?

Nonparametric tools can be used to corroborate the selection of parametric models. Essentially, the approach is to compute selected summary measures under a fitted parametric model and to compare it to the corresponding quantity under the nonparametric model. As the nonparametric model does not assume a specific distribution and is merely a function of the data, it is used as a benchmark to assess how well the parametric distribution/model represents the data. Also, as the sample size increases, the empirical distribution converges almost surely to the underlying population distribution (by the strong law of large numbers). Thus the empirical distribution is a good proxy for the population. The comparison of parametric to nonparametric estimators may alert the analyst to deficiencies in the parametric model and sometimes point ways to improving the parametric specification. Procedures geared towards assessing the validity of a model are known as model diagnostics.

6.1.1 Graphical Comparison of Distributions

We have already seen the technique of overlaying graphs for comparison purposes. To reinforce the application of this technique, Figure 6.1 compares the empirical distribution to two parametric fitted distributions for log claims from the Property Fund data introduced in Section 1.3. The left panel shows the distribution functions of claims distributions. The dots forming an "S-shaped" curve represent the empirical distribution function at each observation. The thick blue curve gives corresponding values for the fitted gamma distribution and the light purple is for the fitted Pareto distribution. Because the Pareto is much closer to the empirical distribution function than the gamma, this provides evidence that the Pareto is the better model for this dataset. The right panel gives similar information for the density function and provides a consistent message. Based (only) on these figures, the Pareto distribution is the clear choice for the analyst.

For another way to compare the appropriateness of two fitted models, consider the probability-probability (pp) plot. A *pp* plot compares cumulative probabilities under two models. For our purposes, these two models are the nonparametric empirical distribution function and the parametric fitted model. Figure 6.2 shows *pp* plots for the Property Fund data. The fitted gamma is on the left and the fitted Pareto is on the right, compared to the same empirical

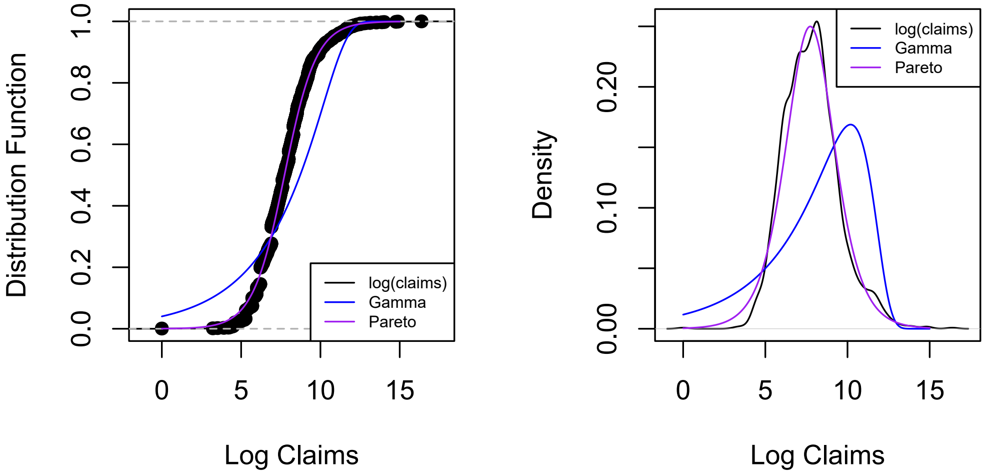


FIGURE 6.1: **Nonparametric Versus Fitted Parametric Distribution and Density Functions.** The left-hand panel compares distribution functions, with the dots corresponding to the empirical distribution, the thick blue curve corresponding to the fitted gamma and the light purple curve corresponding to the fitted Pareto. The right hand panel compares these three distributions summarized using probability density functions.

distribution function of the data. The straight line represents equality between the two distributions being compared, so points close to the line are desirable. As seen in earlier demonstrations, the Pareto is much closer to the empirical distribution than the gamma, providing additional evidence that the Pareto is the better model.

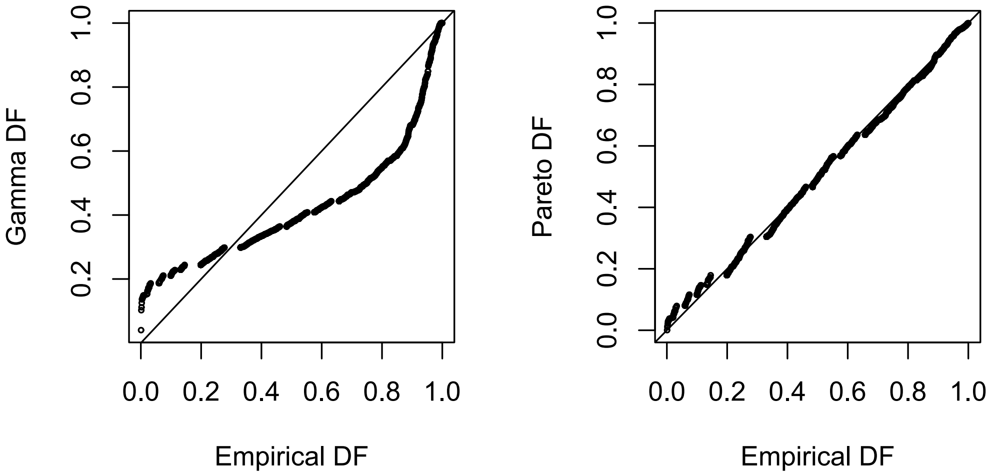


FIGURE 6.2: **Probability-Probability (*pp*) Plots.** The horizontal axis gives the empirical distribution function at each observation. In the left-hand panel, the corresponding distribution function for the gamma is shown in the vertical axis. The right-hand panel shows the fitted Pareto distribution. Lines of $y = x$ are superimposed.

A *pp* plot is useful in part because no artificial scaling is required, such as with the overlaying of densities in Figure 6.1, in which we switched to the log scale to better visualize the data. Note further that *pp* plots are available in multivariate settings where more than one outcome variable is available. However, a limitation of the *pp* plot is that, because it plots *cumulative* distribution functions, it can sometimes be difficult to detect *where* a fitted parametric distribution is deficient. As an alternative, it is common to use a quantile-quantile (qq) plot, as demonstrated in Figure 6.3.

A *qq* plot compares two fitted models through their quantiles. As with *pp* plots, we compare the nonparametric to a parametric fitted model. Quantiles may be evaluated at each point of the dataset, or on a grid (e.g., at $0, 0.001, 0.002, \ldots, 0.999, 1.000$), depending on the application. In Figure 6.3, for each point on the aforementioned grid, the horizontal axis displays the empirical quantile and the vertical axis displays the corresponding fitted parametric quantile (gamma for the upper two panels, Pareto for the lower two). Quantiles are plotted on the original scale in the left panels and on the log scale in the right panels to allow us to see where a fitted distribution is deficient. The straight line represents equality between the empirical distribution and fitted distribution. From these plots, we again see that the Pareto is an overall better fit than the gamma. Furthermore, the lower-right panel suggests that the Pareto distribution does a good job with large claims, but provides a poorer fit for small claims.

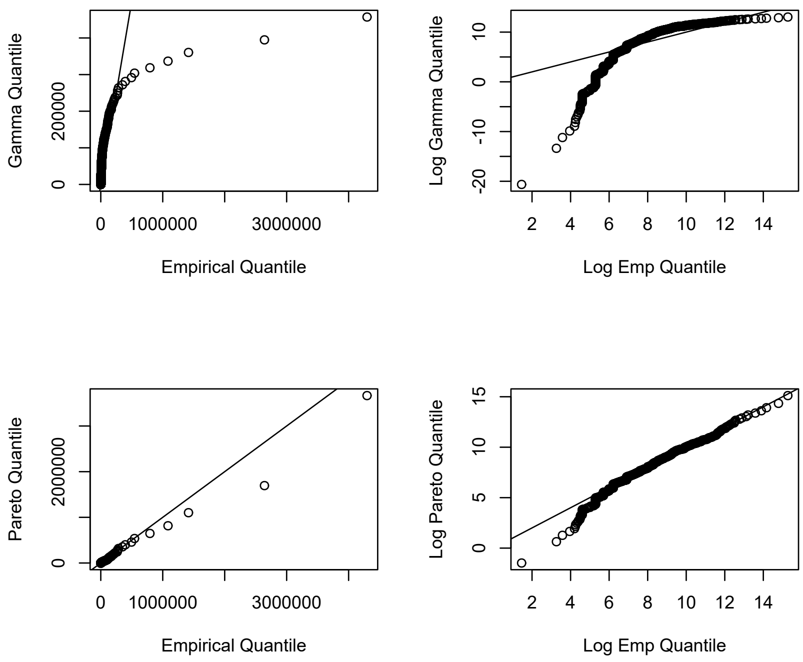


FIGURE 6.3: **Quantile-Quantile (*qq*) Plots.** The horizontal axis gives the empirical quantiles at each observation. The right-hand panels they are graphed on a logarithmic basis. The vertical axis gives the quantiles from the fitted distributions; gamma quantiles are in the upper panels, Pareto quantiles are in the lower panels.

Example 6.1.1. Actuarial Exam Question. Figure 6.4 shows a *pp* plot of a fitted distribution compared to a sample.

Comment on the two distributions with respect to left tail, right tail, and median probabilities.

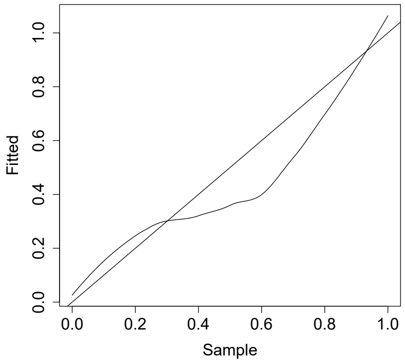


FIGURE 6.4: **Example 6.1.1 Plot**

Example Solution. The tail of the fitted distribution is too thick on the left, too thin on the right, and the fitted distribution has less probability around the median than the sample. To see this, recall that the *pp* plot graphs the cumulative distribution of two distributions on its axes (empirical on the x-axis and fitted on the y-axis in this case). For small values of x, the fitted model assigns greater probability to being below that value than occurred in the sample (i.e. $F(x) > F_n(x)$). This indicates that the model has a heavier left tail than the data. For large values of x, the model again assigns greater probability to being below that value and thus less probability to being above that value (i.e. $S(x) < S_n(x)$). This indicates that the model has a lighter right tail than the data. In addition, as we go from 0.4 to 0.6 on the horizontal axis (thus looking at the middle 20data), the *pp* plot increases from about 0.3 to 0.4. This indicates that the model puts only about 10

6.1.2 Statistical Comparison of Distributions

When selecting a model, it is helpful to make the graphical displays presented. However, for reporting results, it can be effective to supplement the graphical displays with selected statistics that summarize model goodness of fit. Table 6.1 provides three commonly used goodness of fit statistics. In this table, F_n is the empirical distribution, F is the fitted or hypothesized distribution, and $F_i^* = F(x_i)$.

Table 6.1. **Three Goodness of Fit Statistics**

Statistic	Definition	Computational Expression
Kolmogorov-	$\max_x \lvert F_n(x) - F(x)\rvert$	$\max(D^+, D^-)$ where
Smirnov		$D^+ = \max_{i=1,\ldots,n} \left\lvert \frac{i}{n} - F_i^* \right\rvert$
		$D^- = \max_{i=1,\ldots,n} \left\lvert F_i^* - \frac{i-1}{n} \right\rvert$
Cramer-von Mises	$n \int (F_n(x) - F(x))^2 f(x)dx$	$\frac{1}{12n} + \sum_{i=1}^n \left(F_i^* - (2i-1)/n\right)^2$
Anderson-Darling	$n \int \frac{(F_n(x)-F(x))^2}{F(x)(1-F(x))} f(x)dx$	$-n - \frac{1}{n}\sum_{i=1}^n (2i-1)\log\left(F_i^*(1-F_{n+1-i})\right)^2$

The *Kolmogorov-Smirnov statistic* is the maximum absolute difference between the fitted distribution function and the empirical distribution function. Instead of comparing differences between single points, the *Cramer-von Mises statistic* integrates the difference between the empirical and fitted distribution functions over the entire range of values. The *Anderson-Darling statistic* also integrates this difference over the range of values, although weighted by the inverse of the variance. It therefore places greater emphasis on the tails of the distribution (i.e when $F(x)$ or $1 - F(x) = S(x)$ is small).

Example 6.1.2. Actuarial Exam Question (modified). A sample of claim payments is:

$$29 \quad 64 \quad 90 \quad 135 \quad 182$$

Compare the empirical claims distribution to an exponential distribution with mean 100 by calculating the value of the Kolmogorov-Smirnov test statistic.

r SolnBegin()‘ For an exponential distribution with mean 100, the cumulative distribution function is $F(x) = 1 - e^{-x/100}$. Thus,

x	$F(x)$	$F_n(x)$	$F_n(x-)$	$\max(\lvert F(x) - F_n(x)\rvert, \lvert F(x) - F_n(x-)\rvert)$
29	0.2517	0.2	0	$\max(0.0517, 0.2517) = 0.2517$
64	0.4727	0.4	0.2	$\max(0.0727, 0.2727) = 0.2727$
90	0.5934	0.6	0.4	$\max(0.0066, 0.1934) = 0.1934$
135	0.7408	0.8	0.6	$\max(0.0592, 0.1408) = 0.1408$
182	0.8380	1	0.8	$\max(0.1620, 0.0380) = 0.1620$

The Kolmogorov-Smirnov test statistic is therefore

$$KS = \max(0.2517, 0.2727, 0.1934, 0.1408, 0.1620) = 0.2727.$$

r SolnEnd()‘

Pearson's chi-square test

In this section we introduce another goodness of fit test - Pearson's chi-square test - which can be used for testing whether a discrete distribution provides a good fit to discrete data. For more details on the Pearson's chi-square test, at an introductory mathematical statistics level, we refer the reader to Section 9.1 of Hogg et al. (2015).

To illustrate application of the Pearson's chi-square test, we use the example introduced in Section 3.7: In 1993, a portfolio of $n = 7,483$ automobile insurance policies from a major Singaporean insurance company had the distribution of auto accidents per policyholder as given in Table 6.2.

Table 6.2. **Singaporean Automobile Accident Data**

Count (k)	0	1	2	3	4	Total
No. of Policies with k accidents (m_k)	$6,996$	455	28	4	0	$7,483$

If we a fit a Poisson distribution, then the *mle* for λ, the Poisson mean, is the sample mean which is given by

$$\overline{N} = \frac{0 \cdot 6996 + 1 \cdot 455 + 2 \cdot 28 + 3 \cdot 4 + 4 \cdot 0}{7483} = 0.06989.$$

Now if we use Poisson $(\hat{\lambda}_{MLE})$ as the fitted distribution, then a tabular comparison of the fitted counts and observed counts is given by Table 6.3, where $\hat{p}_k$ represents the estimated probabilities under the fitted Poisson distribution.

Table 6.3. **Comparison of Observed to Fitted Counts: Singaporean Auto Data**

Count (k)	Observed (m_k)	Fitted Counts Using Poisson ($n\hat{p}_k$)
0	$6,996$	$6,977.86$
1	455	487.70
2	28	17.04
3	4	0.40
≥ 4	0	0.01
Total	$7,483$	$7,483.00$

While the fit seems reasonable, the Pearson's chi-square statistic is a goodness of fit measure that can be used to test the hypothesis that the underlying distribution is Poisson. To explain this statistic let us suppose that a dataset of size n is grouped into k cells with m_k/n and $\hat{p}_k$, for $k = 1\ldots, K$ being the observed and estimated probabilities of an observation belonging to the k-th cell, respectively. The Pearson's chi-square test statistic is then given by

$$\sum_{k=1}^{K} \frac{(m_k - n\widehat{p}_k)^2}{n\widehat{p}_k}.$$

The motivation for the above statistic derives from the fact that

$$\sum_{k=1}^{K} \frac{(m_k - np_k)^2}{np_k}$$

has a limiting chi-square distribution with $K-1$ degrees of freedom if p_k, $k = 1, \ldots, K$ are the true cell probabilities. Now suppose that only the summarized data represented by m_k, $k = 1, \ldots, K$ is available. Further, if p_k's are functions of s parameters, replacing p_k's by any *efficiently* estimated probabilities $\widehat{p}_k$'s results in the statistic continuing to have a limiting chi-square distribution but with degrees of freedom given by $K-1-s$. Such efficient estimates can be derived for example by using the *mle* method (with a multinomial likelihood) or by estimating the s parameters which minimizes the Pearson's chi-square statistic above. For example, the `R` code below does calculate an estimate for λ doing the latter and results in the estimate 0.06623153, close but different from the *mle* of λ using the full data:

```
m  <- c(6996,455,28,4,0);
op <- m/sum(m);
g  <- function(lam){ sum( (op-c(dpois(0:3,lam),1-ppois(3,lam)) )^2) };
optim( sum(op*(0:4)), g, method="Brent", lower=0, upper=10)$par
```

When one uses the full data to estimate the probabilities, the asymptotic distribution is *in between* chi-square distributions with parameters $K-1$ and $K-1-s$. In practice it is common to ignore this subtlety and assume the limiting chi-square has $K-1-s$ degrees of freedom. Interestingly, this practical shortcut works quite well in the case of the Poisson distribution.

For the Singaporean auto data the Pearson's chi-square statistic equals 41.98 using the full data *mle* for λ. Using the limiting distribution of chi-square with $5-1-1 = 3$ degrees of freedom, we see that the value of 41.98 is way out in the tail (99-th percentile is below 12). Hence we can conclude that the Poisson distribution provides an inadequate fit for the data.

In the above, we started with the cells as given in the above tabular summary. In practice, a relevant question is how to define the cells so that the chi-square distribution is a good approximation to the finite sample distribution of the statistic. A rule of thumb is to define the cells in such a way to have at least 80%, if not all, of the cells having expected counts greater than 5. Also, it is clear that a larger number of cells results in a higher power of the test, and hence a simple rule of thumb is to maximize the number of cells such that each cell has at least 5 observations.

6.2 Iterative Model Selection

In our model development, we examine the data graphically, hypothesize a model structure, and compare the data to a candidate model in order to formulate an improved model. Box (1980) describes this as an *iterative process* which is shown in Figure 6.5.

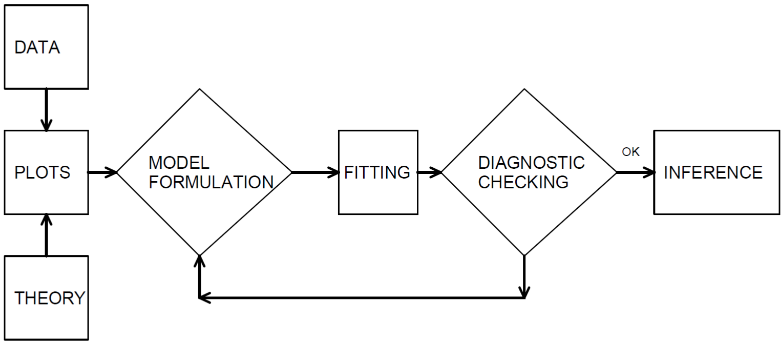


FIGURE 6.5: **Iterative Model Specification Process**

This iterative process provides a useful recipe for structuring the task of specifying a model to represent a set of data.

1. The first step, the model formulation stage, is accomplished by examining the data graphically and using prior knowledge of relationships, such as from economic theory or industry practice.
2. The second step in the iteration is fitting based on the assumptions of the specified model. These assumptions must be consistent with the data to make valid use of the model.
3. The third step is *diagnostic checking*; the data and model must be consistent with one another before additional inferences can be made. Diagnostic checking is an important part of the model formulation; it can reveal mistakes made in previous steps and provide ways to correct these mistakes.

The iterative process also emphasizes the skills you need to make data analytics work. First, you need a willingness to summarize information numerically and portray this information graphically. Second, it is important to develop an understanding of model properties. You should understand how a probabilistic model behaves in order to match a set of data to it. Third, theoretical properties of the model are also important for inferring general relationships based on the behavior of the data.

6.3 Model Selection Based on a Training Dataset

As introduced in Section 2.2, it is common to refer to a dataset used for fitting the model as a *training* or an *in-sample* dataset. Techniques available for selecting a model depend upon whether the outcomes X are discrete, continuous, or a hybrid of the two, although the principles are the same.

Graphical and other Basic Summary Measures. Begin by summarizing the data graphically and with statistics that do not rely on a specific parametric form, as summarized in Section 4.4.1. Specifically, you will want to graph both the empirical distribution and density functions. Particularly for loss data that contain many zeros and that can be skewed, deciding on the appropriate scale (e.g., logarithmic) may present some difficulties. For discrete data, tables are often preferred. Determine sample moments, such as the mean and variance, as well as selected quantiles, including the minimum, maximum, and the median. For discrete data, the mode (or most frequently occurring value) is usually helpful.

These summaries, as well as your familiarity of industry practice, will suggest one or more candidate parametric models. Generally, start with the simpler parametric models (for example, one parameter exponential before a two parameter gamma), gradually introducing more complexity into the modeling process.

Critique the candidate parametric model numerically and graphically. For the graphs, utilize the tools introduced in Section 6.1 such as *pp* and *qq* plots. For the numerical assessments, examine the statistical significance of parameters and try to eliminate parameters that do not provide additional information. In addition to statistical significance of parameters, you may use the following model comparison tools.

Likelihood Ratio Tests. For comparing model fits, if one model is a subset of another, then a likelihood ratio test may be employed; the general approach to likelihood ratio testing is described in Appendix Sections 17.4.3 and 19.1.

Goodness of Fit Statistics. Generally, models are not proper subsets of one another in which case overall goodness of fit statistics are helpful for comparing models. *Information criteria* are one type of goodness of statistic. The most widely used examples are Akaike's Information Criterion (AIC) and the (Schwarz) Bayesian Information Criterion (BIC); they are widely cited because they can be readily generalized to multivariate settings. Appendix Section 17.4.4 provides a summary of these statistics.

For selecting the appropriate distribution, statistics that compare a parametric fit to a nonparametric alternative, summarized in Section 6.1.2, are useful for model comparison. For discrete data, a *goodness of fit* statistic is generally preferred as it is more intuitive and simpler to explain.

6.4 Model Selection Based on a Test Dataset

Model validation introduced in Section 2.2 is the process of confirming that the proposed model is appropriate based on a *test* or an *out-of-sample* dataset, especially in light of the purposes of the investigation. Model validation is important since the model selection process based only on training or in-sample data can be susceptible to data-snooping, that is, fitting a great number of models to a single set of data. By looking at a large number of models, we may overfit the data and understate the natural variation in our representation.

Selecting a model based only on in-sample data also does not support the goal of predictive inference. Particularly in actuarial applications, our goal is to make statements about *new* experience rather than a dataset at hand. For example, we use claims experience from one year to develop a model that can be used to price insurance contracts for the following year. As an analogy, we can think about the training dataset as experience from one year that is used to predict the behavior of the next year's test dataset.

We can respond to these criticisms by using a technique known as *out-of-sample validation*. The ideal situation is to have available two sets of data, one for training, or model development, and the other for testing, or model validation. We initially develop one or several models on the first dataset that we call *candidate* models. Then, the relative performance of the candidate models can be measured on the second set of data. In this way, the data used to validate the model are unaffected by the procedures used to formulate the model.

Random Split of the Data. Unfortunately, rarely will two sets of data be available to the investigator. As mentioned in Section 2.2, we can implement the validation process by splitting the dataset into *training* and *test* subsamples, respectively. Figure 6.6 illustrates this splitting of the data.

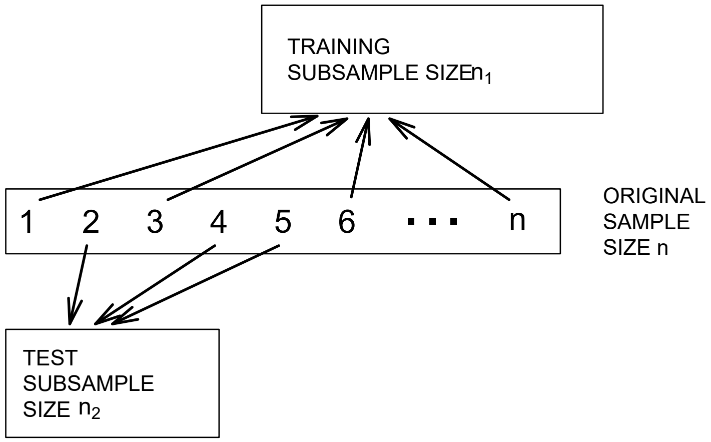


FIGURE 6.6: **Model Validation.** A dataset is randomly split into two subsamples.

Various researchers recommend different proportions for the allocation. Snee (1977) suggests that data-splitting not be done unless the sample size is moderately large. The guidelines of Picard and Berk (1990) show that the greater the number of parameters to be estimated, the greater the proportion of observations is needed for the training subsample for model development.

Selecting a Distribution. Still, our focus so far has been to select a distribution for a dataset that can be used for actuarial modeling without additional explanatory or input variables $x_1, \ldots, x_k$. Even in this more fundamental problem, the model validation approach is valuable. If we base all inference on only in-sample data, then there is a tendency to select more complicated models than needed. For example, we might select a four parameter GB2, generalized beta of the second kind, distribution when only a two parameter Pareto is needed. Information criteria such as AIC and BIC introduced in Appendix Section 17.4.4 include penalties for model complexity and thus provide protection against over-fitting, but using a test sample may also help achieve parsimonious models. From a quote often attributed to Albert Einstein, we want to "use the simplest model as possible but no simpler."

Example 6.4.1. Wisconsin Property Fund. For the 2010 property fund data from Section 1.3, we may try to select a severity distribution based on out-of-sample prediction. In particular, we may randomly select 1,000 observations as our training data, and use the remaining 377 claims to validate the two models based respectively on gamma and Pareto distributions. For illustration purposes, We compare the Kolmogorov-Smirnov statistics respectively for the training and test datasets using the models fitted from training data.

Based on in-sample prediction, the Kolmogorov-Smirnov goodness of fit statistic for the gamma distribution turns out to be 0.2771 and for the Pareto distribution is 0.046. Based on out-of-sample prediction, the Kolmogorov-Smirnov goodness of fit statistic for the gamma distribution turns out to be 0.2693 and for the Pareto distribution is 0.0746. Based on both in-sample and out-of-sample prediction, the Pareto model seems to give considerably better goodness of fit under the random seed used in the code for splitting the training and test data.

Model Validation Statistics. In addition to the nonparametric tools introduced earlier for comparing marginal distributions of the outcome or output variables ignoring potential explanatory or input variables, much of the literature supporting the establishment of a model validation process is based on regression and classification models that you can think of as an *input-output* problem (James et al. (2013)). That is, we have several inputs or predictor variables $x_1, \ldots, x_k$ that are related to an output or outcome y through a function such as

$$y = \mathrm{g}\left(x_1, \ldots, x_k\right).$$

For model selection, one uses the training sample to develop an estimate of g, say, $\hat{\mathrm{g}}$, and then calibrate the average distance from the observed outcomes to the predictions using a criterion of the form

$$\frac{1}{n} \sum_i \mathrm{d}(y_i, \hat{\mathrm{g}}\left(x_{i1}, \ldots, x_{ik}\right)). \tag{6.1}$$

Here, "d" is some measure of distance and the sum i is over the test data. The function g may not have an analytical form and can be estimated for each observation using the different different types of algorithms and models introduced earlier in Section 2.4. In many regression applications, it is common to use the squared Euclidean distance of the form $\mathrm{d}(y_i, \mathrm{g}) = (y_i - \mathrm{g})^2$ under which the criterion in equation (6.1) is called the *mean squared error (MSE)*. Using data simulated from linear models, Example 2.3.1 uses the *root mean squared error (Rmse)* which is the squared root of the MSE. From equation (6.1), the MSE criteria works the best for linear models under normal distributions with constant variance, as minimizing MSE is equivalent to the maximum likelihood and least squares criterion in training data. In data analytics and linear regression, one may consider transformations of the outcome variable in order for the MSE criteria to work more effectively. In actuarial applications, the *mean absolute error (MAE)* under the Euclidean distance $\mathrm{d}(y_i, \mathrm{g}) = |y_i - \mathrm{g}|$ may be preferred because of the skewed nature of loss data. For right-skewed outcomes, it may require a larger sample size for the validation statistics to pickup the correct model when large outlying values of y can have a large effect on the measures.

Following Example 2.3.1, we use simulated data in Examples 6.4.2 through 6.4.4 to compare the AIC information criteria from Appendix Chapter 17.4.4 with out-of-sample MSE and MAE criterion for selecting the distribution and input variables for outcomes that are respectively from normal and right-skewed distributions including lognormal and gamma distributions. For right skewed distributions, we find that the AIC information criteria seems to work consistently for selecting the correct distributional form and mean structure (input variables), whereas out-of-sample MSE and MAE may not work for right-skewed outcomes like those from gamma distributions, even with relatively large sample sizes. Therefore, model validation statistics commonly used in data analytics may only work for minimizing specific cost functions, such as the MAE that represents the average absolute error for out-of-sample prediction, and do not necessarily guarantee correct selection of the underlying data generating mechanism.

Example 6.4.2. In-sample AIC and out-of-sample MSE for normal outcomes. Example 2.3.1 assumes that there is a set of claims that potentially varies by a single categorical variable with six levels. To illustrating in-sample over-fitting, it also assumes that two of the six levels share a common mean that differs from rest of levels. For Example 2.3.1, the claim amounts were generated from a linear model with constant variance, for which in-sample AIC and out-of-sample Rmse provide consistent results from the cross-validation procedure to be introduced in the next section. Here, we may use the same data generation mechanism to compare the performance of in-sample AIC with the in-sample and out-of-sample Rmse criteria. In particular, we generate a total of 200 samples and split them equally into the training and test datasets. From Table 6.4, we observe the two-level model was correctly selected by both in-sample AIC and out-of-sample MSE criteria, whereas in-sample MSE prefers an over-fitted model with six levels. Thus, due to concerns of model over-fitting, we do not use in-sample distance measures such as the MSE and MAE criterion that favors more complicated models.

TABLE 6.4: **Model Selection based on MSE and AIC for normal outputs**

	Community Rating	Two Levels	Six Levels
Rmse - Train	1.186	1.016	0.990
Rmse - Test	1.081	0.958	1.012
AIC - Train	321.935	293.028	295.694

Example 6.4.3. MSE and MAE for right-skewed outcomes - lognormal claims. For claims modeling, one may wonder how the MSE and MAE types of criterion may perform for right-skewed data. Using the same data generating procedure, we may generate lognormal claim amounts by exponentiating the normal outcomes from the previous example. We fit the lognormal claim amounts with lognormal and gamma regression commonly used for ratemaking and claims analytics. Results are summarized in Tables 6.5 and 6.6, respectively. For the specific data generating mechanism, we observe that it requires a larger sample size for out-of-sample Rmse and MAE to select the correct distributional form and mean structure, when compared with in-sample AIC criteria. The AIC criteria is able to pick out the correct model with a sample size of 200, while out-of-sample MSE and MAE fail to. Thus, for right skewed output, precautions need to be taken when using model validation statistics that may be sensitive to large claim values, particularly when the sample size is relatively small.

TABLE 6.5: **Model Selection based on in-sample AIC and out-of-sample MSE and MAE from lognormal model**

	Community Rating	Two Levels	Six Levels
Rmse - Train	4.365	4.185	4.192
Rmse - Test	3.881	3.686	3.679
MAE - Train	2.077	1.821	1.807
MAE - Test	2.166	2.056	2.073
AIC - Train	1800.716	1681.550	1686.142

TABLE 6.6: **Model Selection based on in-sample AIC and out-of-sample MSE and MAE from gamma model**

	Community Rating	Two Levels	Six Levels
Rmse - Train	4.634	4.572	4.572
Rmse - Test	4.298	4.232	4.235
MAE - Train	1.862	1.815	1.817
MAE - Test	2.127	2.123	2.128
AIC - Train	1906.398	1789.312	1795.662

Example 6.4.4. MSE and MAE for right-skewed outcomes - gamma claims. For right-skewed outcomes, we may be interested in studying how the MSE and MAE types of measures work for another loss severity distribution, the gamma distribution, that is widely used in ratemaking and claims analytics. Here, we use a similar mean structure for generating claims amounts based on a gamma regression with the log link function. We fit the data using lognormal and gamma regression. Results are summarized in Tables 6.7 and 6.8, respectively. For gamma outcomes, Table 6.8 shows that out-of-sample MSE and MAE criterion fail to select the correct distributional form or the mean structure even with a total of 1000 samples. By changing the gamma shape parameter, you may see that the out-of-sample MSE and MAE criterion work in certain settings for correctly selecting the distributional form or the mean structure, but the performance of such model validation statistics does not seem to be consistent across different parameter values and sample sizes for right-skewed gamma outcomes. Again, the AIC criteria seems to be working consistently in selecting the correct distribution and mean structure for the data generated from gamma distributions, even with a smaller sample size of 200.

TABLE 6.7: **Model Selection based on in-sample AIC and out-of-sample MSE and MAE from lognormal model**

	Community Rating	Two Levels	Six Levels
Rmse - Train	1.083	0.763	0.760
Rmse - Test	1.128	0.815	0.812
MAE - Train	0.800	0.535	0.529
MAE - Test	0.830	0.565	0.566
AIC - Train	1212.218	864.776	868.794

TABLE 6.8: **Model Selection based on in-sample AIC and out-of-sample MSE and MAE from gamma model**

	Community Rating	Two Levels	Six Levels
Rmse - Train	1.553	1.476	1.475
Rmse - Test	1.594	1.523	1.522
MAE - Train	1.121	1.226	1.227
MAE - Test	1.138	1.253	1.253
AIC - Train	1249.211	852.292	856.850

6.5 Model Selection Based on Cross-Validation

Although out-of-sample validation is the gold standard in predictive modeling, it is not always practical to do so. The main reason is that we have limited sample sizes and the out-of-sample model selection criterion in equation (6.1) depends on a *random* split of the data. This means that different analysts, even when working the same dataset and same approach to modeling, may select different models. This is likely in actuarial applications because we work with skewed datasets where there is a large chance of getting some very large outcomes and large outcomes may have a great influence on the parameter estimates.

Cross-Validation Procedure. Alternatively, one may use *cross-validation*, as follows.

- The procedure begins by using a random mechanism to split the data into K subsets of roughly equal size known as *folds*, where analysts typically use 5 to 10.
- Next, one uses the first K-1 subsamples to estimate model parameters. Then,

"predict" the outcomes for the Kth subsample and use a measure such as in equation (6.1) to summarize the fit.
- Now, repeat this by holding out each of the K subsamples, summarizing with an out-of-sample statistic. Thus, summarize these K statistics, typically by averaging, to give a single overall statistic for comparison purposes.

Repeat these steps for several candidate models and choose the model with the lowest overall cross-validation statistic.

In Example 2.3.1, you have seen that the MSE criteria seems to work with k-fold cross-validation in selecting the correct mean structure for claims outcome data generated from linear models with constant variance. From Examples 6.4.3 and 6.4.4, however, the out-of-sample MSE and MAE criterion does not seem to provide consistent performance for selecting the distributional form and the mean structure under right-skewed claims distributions. Thus, we may use the k-folder cross-validation instead of out-of-sample prediction to see whether the MSE and MAE types of criterion work for right-skewed distributions based on lognormal and gamma regression with a log link function.

Example 6.5.1. Cross-validation in right-skewed outcomes - lognormal claims For lognormal claims, we use the data generating mechanism from Example 6.4.3 to generate a total of 100 samples, and use the k-fold cross validation procedure in Example 2.3.1 to select the distributional form and mean structure. Using cross-validation, we note that both AIC and out-of-sample MSE and MAE seem to be working for selecting the model with the correct distribution and mean structure, even with a total of 100 samples.

Example 6.5.2. Cross-validation in right-skewed outcomes - gamma claims For gamma claims, we use the data generating mechanism from Example 6.4.4 to generate a total of 100 samples, and use the k-fold cross validation procedure to select the distributional form and mean structure. Using cross-validation, we note that in-sample AIC seems to be working for selecting the model with the correct distribution and mean structure, while out-of-sample MSE and MAE seem to fail in selecting the distributional form or the mean structure correctly even after we increase the sample size to 1000.

Cross-validation is widely used because it retains the predictive flavor of the out-of-sample model validation process but, due to the re-use of the data, is more stable over random samples. In addition, Example 8.4.1 in Chapter 8 uses the Wisconsin Property Fund to perform k-fold cross-validation of the gamma and Pareto models based on the Kolmogorov-Smirnov goodness

TABLE 6.9: **Cross-validation based on in-sample AIC, and out-of-sample MSE and MAE from lognormal model**

	Community Rating	Two Levels	Six Levels
Rmse - Fold 1	1.808	1.750	1.891
Rmse - Fold 2	2.145	1.773	1.813
Rmse - Fold 3	3.461	3.335	3.333
Rmse - Fold 4	1.425	1.723	1.865
Rmse - Fold 5	4.848	4.450	4.454
Rmse - Average	2.738	2.606	2.671
MAE - Fold 1	1.341	1.408	1.502
MAE - Fold 2	1.881	1.264	1.255
MAE - Fold 3	2.037	2.142	2.146
MAE - Fold 4	1.225	1.345	1.476
MAE - Fold 5	2.421	2.022	2.051
MAE - Average	1.781	1.636	1.686
AIC - Average	286.257	266.223	271.200

TABLE 6.10: **Cross-validation based on in-sample AIC, and out-of-sample MSE and MAE from gamma model**

	Community Rating	Two Levels	Six Levels
Rmse - Fold 1	2.557	2.642	2.677
Rmse - Fold 2	1.930	1.999	2.005
Rmse - Fold 3	4.088	4.155	4.187
Rmse - Fold 4	1.181	1.273	1.318
Rmse - Fold 5	5.232	5.262	5.286
Rmse - Average	2.998	3.066	3.095
MAE - Fold 1	1.929	2.069	2.114
MAE - Fold 2	1.060	1.116	1.124
MAE - Fold 3	2.488	2.660	2.725
MAE - Fold 4	0.887	0.949	0.999
MAE - Fold 5	2.251	2.312	2.345
MAE - Average	1.723	1.821	1.861
AIC - Average	299.063	281.455	282.816

TABLE 6.11: **Cross-validation based on in-sample AIC, and out-of-sample MSE and MAE from lognormal model**

	Community Rating	Two Levels	Six Levels
Rmse - Fold 1	1.080	0.794	0.799
Rmse - Fold 2	0.953	0.639	0.639
Rmse - Fold 3	1.354	0.914	0.916
Rmse - Fold 4	1.097	0.725	0.727
Rmse - Fold 5	1.171	0.695	0.695
Rmse - Average	1.131	0.753	0.755
MAE - Fold 1	0.837	0.579	0.583
MAE - Fold 2	0.755	0.473	0.474
MAE - Fold 3	0.952	0.600	0.602
MAE - Fold 4	0.852	0.523	0.525
MAE - Fold 5	0.897	0.503	0.507
MAE - Average	0.859	0.536	0.538
AIC - Average	1980.018	1381.321	1388.351

TABLE 6.12: **Cross-validation based on in-sample AIC, and out-of-sample MSE and MAE from gamma model**

	Community Rating	Two Levels	Six Levels
Rmse - Fold 1	1.455	1.620	1.620
Rmse - Fold 2	1.347	1.543	1.543
Rmse - Fold 3	1.865	2.006	2.005
Rmse - Fold 4	1.558	1.738	1.738
Rmse - Fold 5	1.690	1.838	1.838
Rmse - Average	1.583	1.749	1.749
MAE - Fold 1	1.003	1.223	1.223
MAE - Fold 2	0.975	1.195	1.195
MAE - Fold 3	1.301	1.478	1.479
MAE - Fold 4	1.118	1.342	1.342
MAE - Fold 5	1.228	1.420	1.420
MAE - Average	1.125	1.332	1.332
AIC - Average	2047.108	1349.855	1357.246

of fit statistic. Additional information and examples regarding re-sampling procedures including leave-one-out cross-validation and bootstrap can also be found in Chapter 8.

6.6 Model Selection for Modified Data

So far we have discussed model selection using unmodified data. For modified data including grouped, censored and truncated data, you learned parametric and nonparametric estimation of distribution functions in Chapter 5. For model selection, the tools from Section 6.1 can be extended to cases of modified data.

For selection of distributions, the nonparametric tools introduced in Section 6.1 are based on estimated parametric and nonparametric distribution functions, and thus can be extended to modified data for which both types of estimators exist.

For graphical comparisons, the *pp* and *qq* plots introduced earlier can be created for modified data by plotting the parametric estimates from Section 5.2 against nonparametric estimates of the probability or distribution functions from Section 5.3. For example, the `qqPlotCensored` and `qqtrunc` functions in `R` generate *qq* plots respectively for censored (left or right) and truncated data, whereas the `probPlot` function creates both *pp* and *qq* plots with a larger selection of distributions for right-censored and unmodified data. Additional graphical tools such as cumulative hazard plots are available in the `R` package `GofCens`.

Example 6.6.1. Bodily Injury Claims and *qq*-Plots. For the Boston auto bodily injury claims data from Example 5.3.2, we include the full dataset with right-censoring, and use the *qq*-plot to compare the estimated quantiles from lognormal, normal and exponential distributions with those from the nonparametric Kaplan-Meier method. From the *qq*-plots in Figure 6.7, the lognormal distribution seems to fit the censored data much better those based on the normal and exponential distributions.

In addition to graphical tools, you may use tools from Section 6.1.2 for statistical comparisons of models fitted from modified data based on parametric and nonparametric estimates of distribution functions. For example, the `R` package `GofCens` provides functions calculating the three goodness of fit statistics from Section 6.1.2 for both right-censored and unmodified data. The `R` package

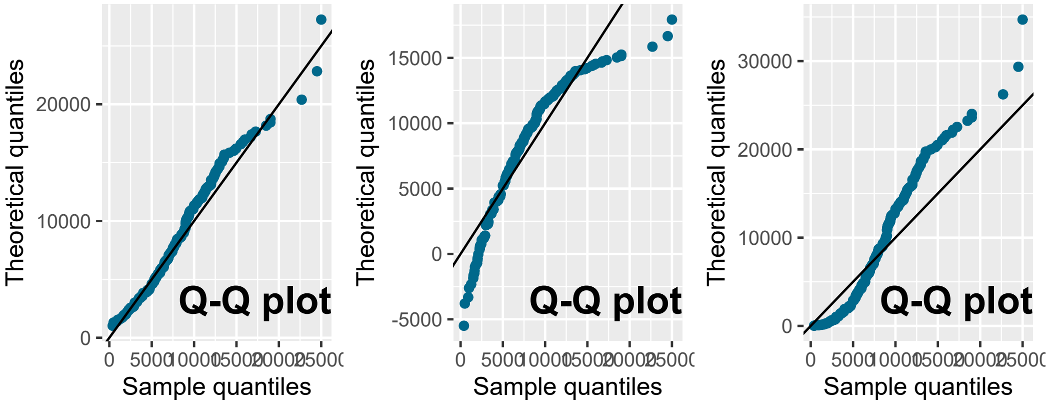


FIGURE 6.7: **Quantile-Quantile (*qq*) Plots for Bodily Injury Claims.** The horizontal axis gives the empirical quantiles at each observation. The vertical axis gives the quantiles from the fitted distributions; lognormal quantiles are in the left-hand panel, normal quantiles are in the middle, and exponential in the right-hand panel.

TABLE 6.13: **Nonparametric goodness of fit statistics for right-censored Bodily Injury Claims**

	Kolmogorov-Smirnov	Cramer-von Mises	Anderson-Darling
Lognormal	1.994	0.305	1.770
Normal	3.096	1.335	9.437
Exponential	4.811	4.065	21.659

`truncgof`, on the other hand, provides functions for calculating the three goodness of fit statistics for left-truncated data.

Example 6.6.2. Bodily Injury Claims and Goodness of Fit Stastistics. For the Boston auto bodily injury claims with right-censoring, we may use the goodness of fit statistics to evaluate the fitted lognormal, normal and exponential distributions. For the Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling statistics, the lognormal distribution gives values that are much lower than those from normal and exponential distributions. The conclusion from the goodness of fit statistics is consistent to that revealed by the *qq* plots.

Other than selecting the distributional form, model comparison measures such as the likelihood ratio test and information criterion including the AIC from Section 6.3 can be obtained for models fitted based on likelihood criteria based on the likelihood functions introduced earlier for modified data. For modified data, the `survreg` and `flexsurvreg` functions in `R` fit parametric regression models on censored and/or truncated outcomes based on maximum likelihood estimation which allows use of likelihood ratio tests and information criterion such as AIC for in-sample model comparisons. For censored and truncated data, the functions also provide output of residuals that allow calculation of model validation statistics such as the MSE and MAE for the iterative model selection procedure introduced in Section 6.2.

6.7 Further Resources and Contributors

Contributors

- **Lei (Larry) Hua** and **Michelle Xia**, Northern Illinois University, are the principal authors of the second edition of this chapter.
- **Edward (Jed) Frees** and **Lisa Gao**, University of Wisconsin-Madison, are the principal authors of the initial version of this chapter.
- Chapter reviewers include: Vytaras Brazauskas, Yvonne Chueh, Eren Dodd, Hirokazu (Iwahiro) Iwasawa, Joseph Kim, Andrew Kwon-Nakamura, Jiandong Ren, and Di (Cindy) Xu.

Further Readings and References

If you would like additional practice with `R` coding, please visit our companion LDA Short Course. In particular, see the Model Selection and Estimation Chapter.

7

Aggregate Loss Models

Chapter Preview. This chapter introduces probability models for describing the aggregate (total) claims that arise from a portfolio of insurance contracts. We present two standard modeling approaches, the individual risk model and the collective risk model. Further, we discuss strategies for computing the distribution of the aggregate claims, including exact methods for special cases, recursion, and simulation. Finally, we examine the effects of individual policy modifications such as deductibles, coinsurance, and inflation, on the frequency and severity distributions, and thus on the aggregate loss distribution.

7.1 Introduction

In this section, you learn:

- the concept of aggregate claims for an insurance system
- alternative methods to describe the aggregate losses
- the interpretation of different models for aggregate claims

The objective of this chapter is to build a probability model to describe the aggregate claims by an insurance system occurring in a fixed time period. The insurance system could be a single policy, a group insurance contract, a business line, or an entire book of an insurer's business. In this chapter, aggregate claims refer to either the number or the amount of claims from a portfolio of insurance contracts. However, the modeling framework can be readily applied in a general setup.

Consider an insurance portfolio of n individual contracts, and let S denote the aggregate losses of the portfolio in a given time period. There are two approaches to modeling the aggregate losses S, the individual risk model and the collective risk model. The individual risk model emphasizes the loss from each individual contract and represents the aggregate losses as:

$$S_n = X_1 + X_2 + \cdots + X_n,$$

where X_i $(i = 1, \ldots, n)$ is interpreted as the loss amount from the ith contract. It is worth stressing that n denotes the number of contracts in the portfolio and thus is a fixed number rather than a random variable. For the individual risk model, one usually assumes the X_i's are independent. Because of different contract features such as coverage and exposure, the X_i's are not necessarily identically distributed. A notable feature of the distribution of each X_i is the probability mass at zero corresponding to the event of no claims.

The collective risk model represents the aggregate losses in terms of a frequency distribution and a severity distribution:

$$S_N = X_1 + X_2 + \cdots + X_N.$$

Here, one thinks of a random number of claims N that may represent either the number of losses or the number of payments. In contrast, in the individual risk model, we use a fixed number of contracts n. We think of $X_1, X_2, \ldots, X_N$ as representing the amount of each loss. Each loss may or may not correspond to a unique contract. For instance, there may be multiple claims arising from a single contract. It is natural to think about $X_i > 0$ because if $X_i = 0$ then no claim has occurred. Typically we assume that conditional on $N = n$, $X_1, X_2, \ldots, X_n$ are iid random variables. The distribution of N is known as the frequency distribution, and the common distribution of X is known as the severity distribution. We further assume N and X are independent. With the collective risk model, we may decompose the aggregate losses into the frequency (N) process and the severity (X) model. This flexibility allows the analyst to comment on these two separate components. For example, sales growth due to lower underwriting standards could lead to higher frequency of losses but might not affect severity. Similarly, inflation or other economic forces could have an impact on severity but not on frequency.

The rest of the chapter is structured as follows: Section 7.2 and Section 7.3 provide details on the individual risk model and collective risk model respectively. Section 7.4 presents methods for computing the distribution of aggregate claims. Section 7.5 discusses the effect of coverage modifications on the aggregate losses. Technical materials are summarized in Section 7.6.

7.2 Individual Risk Model

In this section, you learn:

- mathematical representation of the individual risk model
- applications of individual risk model to life and non-life insurance
- how to evaluate moments, generating functions, and the distribution function of the individual risk model

7.2.1 Moments and Distribution

As noted earlier, for the *individual risk model*, we think of X_i as the loss from ith contract and interpret

$$S_n = X_1 + X_2 + \cdots + X_n,$$

to be the aggregate loss from all contracts in a portfolio or group of contracts. Here, the X_i's are not necessarily identically distributed and we have

$$\mathrm{E}(S_n) = \sum_{i=1}^{n} \mathrm{E}(X_i) \ .$$

Under the independence assumption on X_i's (which implies $\mathrm{Cov}\left(X_i, X_j\right) = 0$ for all $i \neq j$), it can further be shown that

$$\begin{aligned} \mathrm{Var}(S_n) &= \sum_{i=1}^{n} \mathrm{Var}(X_i) \\ P_{S_n}(z) &= \prod_{i=1}^{n} P_{X_i}(z) \\ M_{S_n}(t) &= \prod_{i=1}^{n} M_{X_i}(t), \end{aligned}$$

where $P_{S_n}(\cdot)$ and $M_{S_n}(\cdot)$ are the probability generating function (*pgf*) and the moment generating function (*mgf*) of S_n, respectively. The distribution of each X_i contains a probability mass at zero, corresponding to the event of no claims from the ith contract. One strategy to incorporate the zero mass in the distribution is to use the two-part framework:

$$X_i = I_i \times B_i = \begin{cases} 0 \ , & \text{if } \ I_i = 0 \\ B_i \ , & \text{if } \ I_i = 1. \end{cases}$$

Here, I_i is a Bernoulli variable indicating whether or not a loss occurs for the ith contract, and B_i is a random variable with nonnegative support representing the amount of losses of the contract given loss occurrence. Assume that $I_1, \ldots, I_n, B_1, \ldots, B_n$ are mutually independent. Denote $\Pr(I_i = 1) = q_i$, $\mu_i = \mathrm{E}(B_i)$, and $\sigma_i^2 = \mathrm{Var}(B_i)$. It can be shown (see *Technical Supplement 7.A.1* in Section 7.6 for details) that

$$\begin{aligned}
\mathrm{E}(S_n) &= \sum_{i=1}^{n} q_i \ \mu_i \\
\mathrm{Var}(S_n) &= \sum_{i=1}^{n} \left(q_i \sigma_i^2 + q_i(1-q_i)\mu_i^2 \right) \\
P_{S_n}(z) &= \prod_{i=1}^{n} \left(1 - q_i + q_i P_{B_i}(z)\right) \\
M_{S_n}(t) &= \prod_{i=1}^{n} \left(1 - q_i + q_i M_{B_i}(t)\right).
\end{aligned}$$

A special case of the above model is when B_i follows a degenerate distribution with $\mu_i = b_i$ and $\sigma_i^2 = 0$. One example is term life insurance or a pure endowment insurance where b_i represents the insurance benefit amount of the ith contract.

Another strategy to accommodate the zero mass in the loss from each contract is to consider them in aggregate at the portfolio level, as in the *collective risk model*. Here, the aggregate loss is $S_N = X_1 + \cdots + X_N$, where N is a random variable representing the number of non-zero claims that occurred out of the entire group of contracts. Thus, not every contract in the portfolio may be represented in this sum, and $S_N = 0$ when $N = 0$. The collective risk model will be discussed in detail in the next section.

Example 7.2.1. Actuarial Exam Question. An insurance company sold 300 fire insurance policies as follows:

Number of Policies	Policy Maximum (M_i)	Probability of Claim Per Policy (q_i)
100	400	0.05
200	300	0.06

You are given:
(i) The claim amount for each policy, X_i, is uniformly distributed between 0 and the policy maximum M_i.

(ii) The probability of more than one claim per policy is 0.
(iii) Claim occurrences are independent.

Calculate the mean, E (S_{300}), and variance, Var (S_{300}), of the aggregate claims. How would these results change if every claim is equal to the policy maximum?

Example Solution. The aggregate claims are $S_{300} = X_1 + \cdots + X_{300}$, where $X_1, \ldots, X_{300}$ are independent but not identically distributed. Policy claims amounts are uniformly distributed on $(0, M_i)$, so the mean claim amount is $M_i/2$ and the variance is $M_i^2/12$. Thus, for policy $i = 1, \ldots, 300$, we have

Number of Policies	Policy Maximum (M_i)	Probability of Claim Per Policy (q_i)	Mean Amount (μ_i)	Variance Amount (σ_i^2)
100	400	0.05	200	$400^2/12$
200	300	0.06	150	$300^2/12$

The mean of the aggregate claims is

$$\mathrm{E}\ (S_{300}) = \sum_{i=1}^{300} q_i \mu_i = 100\left\{0.05(200)\right\} + 200\left\{0.06(150)\right\} = 2,800$$

The variance of the aggregate claims is

$$\begin{aligned} \mathrm{Var}\ (S_{300}) &= \textstyle\sum_{i=1}^{300} \left(q_i \sigma_i^2 + q_i(1-q_i)\mu_i^2\right) \qquad \text{since } X_i\text{'s are independent} \\ &= 100\left\{0.05\left(\tfrac{400^2}{12}\right) + 0.05(1-0.05)200^2\right\} \\ &\quad +200\left\{0.06\left(\tfrac{300^2}{12}\right) + 0.06(1-0.06)150^2\right\} \\ &= 600,467. \end{aligned}$$

Example 7.2.1. Continued.

Now suppose everybody receives the policy maximum M_i if a claim occurs. What is the expected aggregate loss E $(\tilde{S})$ and variance of the aggregate loss Var $(\tilde{S})$?

Example Solution. Each policy claim amount X_i is now deterministic and fixed at M_i instead of a randomly distributed amount, so $\sigma_i^2 = \mathrm{Var}\ (X_i) = 0$ and $\mu_i = M_i$. Again, the probability of a claim occurring for each policy is q_i. Under these circumstances, the expected aggregate loss is

$$\mathrm{E}\ (\tilde{S}) = \sum_{i=1}^{300} q_i \mu_i = 100\left\{0.05(400)\right\} + 200\left\{0.06(300)\right\} = 5,600.$$

The variance of the aggregate loss is

$$\begin{aligned}\text{Var}\ (\tilde{S}) &= \sum_{i=1}^{300} (q_i\sigma_i^2 + q_i(1-q_i)\mu_i^2) = \sum_{i=1}^{300} (q_i(1-q_i)\mu_i^2) \\ &= 100\left\{(0.05)(1-0.05)400^2\right\} + 200\left\{(0.06)(1-0.06)300^2\right\} \\ &= 1,775,200.\end{aligned}$$

The individual risk model can also be used for claim frequency. If X_i denotes the number of claims from the ith contract, then S_n is interpreted as the total number of claims from the portfolio. In this case, the above two-part framework still applies since there is a probability mass at zero for contracts that do not experience any claims. Assume X_i belongs to the $(a,b,0)$ class with pmf denoted by $p_{ik} = \Pr(X_i = k)$ for $k = 0, 1, \ldots$ (see Section 3.3). Let X_i^T denote the associated zero-truncated distribution in the $(a,b,1)$ class with *pmf* $p_{ik}^T = p_{ik}/(1-p_{i0})$ for $k = 1, 2, \ldots$ (see Section 3.5.1). Using the relationship between their probability generating functions (see *Technical Supplement 7.A.2* in Section 7.6 for details):

$$P_{X_i}(z) = p_{i0} + (1-p_{i0})P_{X_i^T}(z),$$

we can write $X_i = I_i \times B_i$ with $q_i = \Pr(I_i = 1) = \Pr(X_i > 0) = 1 - p_{i0}$ and $B_i = X_i^T$. Notice that in this case, we have a zero-modified distribution since the I_i variable covers the modified probability mass at zero with $q_i = \Pr(I_i = 1)$, while the $B_i = X_i^T$ covers the discrete non-zero frequency portion. See Section 3.5.1 for the relationship between zero-truncated and zero-modified distributions.

Example 7.2.2. An insurance company sold a portfolio of 100 independent homeowners insurance policies, each of which has claim frequency following a zero-modified Poisson distribution, as follows:

Type of	Number of	Probability of	λ
Policy	Policies	At Least 1 Claim	
Low-risk	40	0.03	1
High-risk	60	0.05	2

Find the expected value and variance of the claim frequency for the entire portfolio.

Example Solution. For each policy, we can write the zero-modified Poisson claim frequency N_i as $N_i = I_i \times B_i$, where

$$q_i = \Pr(I_i = 1) = \Pr(N_i > 0) = 1 - p_{i0}.$$

For the low-risk policies, we have $q_i = 0.03$ and for the high-risk policies, we have $q_i = 0.05$. Further, $B_i = N_i^T$, the zero-truncated version of N_i. Thus, we have

$$\begin{aligned}
\mu_i &= \mathrm{E}(B_i) = \mathrm{E}(N_i^T) = \frac{\lambda}{1-e^{-\lambda}} \\
\sigma_i^2 &= \mathrm{Var}(B_i) = \mathrm{Var}(N_i^T) = \frac{\lambda[1-(\lambda+1)e^{-\lambda}]}{(1-e^{-\lambda})^2}.
\end{aligned}$$

Using $n = 100$, let the portfolio claim frequency be $S_{100} = \sum_{i=1}^{100} N_i$. Using the formulas above, the expected claim frequency of the portfolio is

$$\begin{aligned}
\mathrm{E}\ (S_{100}) &= \sum_{i=1}^{100} q_i \mu_i \\
&= 40 \left[0.03 \left(\frac{1}{1-e^{-1}}\right)\right] + 60 \left[0.05 \left(\frac{2}{1-e^{-2}}\right)\right] \\
&= 40(0.03)(1.5820) + 60(0.05)(2.3130) = 8.8375.
\end{aligned}$$

The variance of the claim frequency of the portfolio is

$$\begin{aligned}
\mathrm{Var}\ (S_{100}) &= \sum_{i=1}^{100} \left(q_i \sigma_i^2 + q_i(1-q_i)\mu_i^2\right) \\
&= 40 \left[0.03 \left(\frac{1-2e^{-1}}{(1-e^{-1})^2}\right) + 0.03(0.97)(1.5820^2)\right] \\
&\quad + 60 \left[0.05 \left(\frac{2[1-3e^{-2}]}{(1-e^{-2})^2}\right) + 0.05(0.95)(2.3130^2)\right] \\
&= 23.7214.
\end{aligned}$$

Note that equivalently, we could have calculated the mean and variance of an individual policy directly using the relationship between the zero-modified and zero-truncated Poisson distributions (see Section 3.3).

7.2.2 Aggregate Loss Distribution

To understand the distribution of the aggregate loss, one could use the central limit theorem to approximate the distribution of S_n for large n. Denote $\mu_{S_n} = \mathrm{E}(S_n)$ and $\sigma_{S_n}^2 = \mathrm{Var}(S_n)$ and let $Z \sim N(0,1)$, a standard normal random variable with cdf Φ. Then the *cdf* of S_n can be approximated as follows:

$$\begin{aligned} F_{S_n}(s) &= \Pr(S_n \leq s) = \Pr\left(\frac{S_n - \mu_{S_n}}{\sigma_{S_n}} \leq \frac{s - \mu_{S_n}}{\sigma_{S_n}}\right) \\ &\approx \Pr\left(Z \leq \frac{s - \mu_{S_n}}{\sigma_{S_n}}\right) = \Phi\left(\frac{s - \mu_{S_n}}{\sigma_{S_n}}\right). \end{aligned}$$

Example 7.2.3. Actuarial Exam Question - Follow-Up. As in the Example 7.2.1 earlier, an insurance company sold 300 fire insurance policies, with claim amounts X_i uniformly distributed between 0 and the policy maximum M_i. Using the normal approximation, calculate the probability that the aggregate claim amount S_{300} exceeds \$3,500.

Example Solution. We have seen earlier that $\mathrm{E}(S_{300}) = 2,800$ and $\mathrm{Var}(S_{300}) = 600,467$. Then

$$\begin{aligned} \Pr(S_{300} > 3,500) &= 1 - \Pr(S_{300} \leq 3,500) \\ &\approx 1 - \Phi\left(\frac{3,500 - 2,800}{\sqrt{600,467}}\right) = 1 - \Phi\left(0.90334\right) \\ &= 1 - 0.8168 = 0.1832. \end{aligned}$$

For small n, the distribution of S_n is likely skewed, and the normal approximation would be a poor choice. To examine the aggregate loss distribution, we go back to first principles. Specifically, the distribution can be derived recursively. Define $S_k = X_1 + \cdots + X_k, k = 1, \ldots, n$.

For $k = 1$:

$$F_{S_1}(s) = \Pr(S_1 \leq s) = \Pr(X_1 \leq s) = F_{X_1}(s).$$

For $k = 2, \ldots, n$:

$$\begin{aligned} F_{S_k}(s) &= \Pr(X_1 + \cdots + X_k \leq s) = \Pr(S_{k-1} + X_k \leq s) \\ &= \mathrm{E}_{X_k}\left[\Pr(S_{k-1} \leq s - X_k | X_k)\right] = \mathrm{E}_{X_k}\left[F_{S_{k-1}}(s - X_k)\right]. \end{aligned}$$

A special case is when X_i's are identically distributed. Let $F_X(x) = \Pr(X \leq x)$ be the common distribution of $X_i,\ i = 1, \ldots, n$. We define

$$F_X^{*n}(x) = \Pr(X_1 + \cdots + X_n \leq x),$$

the n-fold convolution of F_X. More generally, we can compute F_X^{*n} recursively.

Begin the recursion at $k = 1$ using $F_X^{*1}(x) = F_X(x)$. Next, for $k = 2$, we have

$$\begin{aligned}
F_X^{*2}(x) &= \Pr(X_1 + X_2 \le x) = \mathrm{E}_{X_2}\left[\Pr(X_1 \le x - X_2 | X_2)\right] \\
&= \mathrm{E}_{X_2}\left[F(x - X_2)\right] \\
&= \begin{cases} \int_0^x F(x-y)f(y)dy & \text{for continuous } X_i\text{'s} \\ \sum_{y \le x} F(x-y)f(y) & \text{for discrete } X_i\text{'s} \end{cases}
\end{aligned}$$

Recall $F(0) = 0$.

Similarly for $k = n$, we have $S_n = X_1 + X_2 + \cdots + X_n$ and

$$\begin{aligned}
F^{*n}(x) &= \Pr(S_n \le x) = \Pr(S_{n-1} + X_n \le x) \\
&= \mathrm{E}_{X_n}\left[\Pr(S_{n-1} \le x - X_n | X_n)\right] \\
&= \mathrm{E}_X\left[F^{*(n-1)}(x - X)\right] \\
&= \begin{cases} \int_0^x F^{*(n-1)}(x-y)f(y)dy & \text{for continuous } X_i\text{'s} \\ \sum_{y \le x} F^{*(n-1)}(x-y)f(y) & \text{for discrete } X_i\text{'s} \end{cases}
\end{aligned}$$

When the X_i's are independent and belong to the same family of distributions, there are some simple cases where S_n has a closed form. This makes it easy to compute $\Pr(S_n \le x)$. This property is known as *closed under convolution*, meaning the distribution of the sum of independent random variables belongs to the same family of distributions as that of the component variables, just with different parameters. Table 7.1 provides a few examples.

Table 7.1. **Closed Form Partial Sum Distributions**

Distribution of X_i	Abbreviation	Distribution of S_n
Normal with mean μ_i and variance σ_i^2	$N(\mu_i, \sigma_i^2)$	$N\left(\sum_{i=1}^n \mu_i,\ \sum_{i=1}^n \sigma_i^2\right)$
Exponential with mean θ	$Exp(\theta)$	$Gam(n, \theta)$
Gamma with shape α_i and scale θ	$Gam(\alpha_i, \theta)$	$Gam\left(\sum_{i=1}^n \alpha_i, \theta\right)$
Poisson with mean (and variance) λ_i	$Poi(\lambda_i)$	$Poi\left(\sum_{i=1}^n \lambda_i\right)$
Binomial with m_i trials and q success probability	$Bin(m_i, q)$	$Bin\left(\sum_{i=1}^n m_i, q\right)$
Geometric with mean β	$Geo(\beta)$	$NB(\beta, n)$
Negative binomial with mean $r_i\beta$ and variance $r_i\beta(1+\beta)$	$NB(\beta, r_i)$	$NB\left(\beta, \sum_{i=1}^n r_i\right)$

Example 7.2.4. Gamma Distribution. Assume that $X_1, \ldots, X_n$ are independent random variables with $X_i \sim Gam(\alpha_i, \theta)$. The *mgf* of X_i is

$M_{X_i}(t) = (1-\theta t)^{-\alpha_i}$. Thus, the *mgf* of the sum $S_n = X_1 + \cdots + X_n$ is

$$\begin{aligned} M_{S_n}(t) &= \prod_{i=1}^{n} M_{X_i}(t) \quad \text{from the independence of } X_i\text{'s} \\ &= \prod_{i=1}^{n} (1-\theta t)^{-\alpha_i} = (1-\theta t)^{-\sum_{i=1}^{n} \alpha_i} \ , \end{aligned}$$

which is the *mgf* of a gamma random variable with parameters $(\sum_{i=1}^{n} \alpha_i, \theta)$. Thus, $S_n \sim Gam(\sum_{i=1}^{n} \alpha_i, \theta)$.

Example 7.2.5. Negative Binomial Distribution. Assume that $X_1, \ldots, X_n$ are independent random variables with $X_i \sim NB(\beta, r_i)$. The *pgf* of X_i is $P_{X_i}(z) = [1-\beta(z-1)]^{-r_i}$. Thus, the *pgf* of the sum $S_n = X_1 + \cdots + X_n$ is

$$\begin{aligned} P_{S_n}(z) &= \mathrm{E}\left[z^{S_n}\right] \\ &= \mathrm{E}\left[z^{X_1}\right] \cdots \mathrm{E}\left[z^{X_n}\right] \quad \text{from the independence of } X_i\text{'s} \\ &= \prod_{i=1}^{n} P_{X_i}(z) = \prod_{i=1}^{n} [1-\beta(z-1)]^{-r_i} = [1-\beta(z-1)]^{-\sum_{i=1}^{n} r_i} \ , \end{aligned}$$

which is the *pgf* of a negative binomial random variable with parameters $(\beta, \sum_{i=1}^{n} r_i)$. Thus, $S_n \sim NB(\beta, \sum_{i=1}^{n} r_i)$.

Example 7.2.6. Actuarial Exam Question (modified). The annual number of doctor visits for each individual in a family of 4 has geometric distribution with mean 1.5. The annual numbers of visits for the family members are mutually independent. An insurance pays 100 per doctor visit beginning with the 4th visit per family. Calculate the probability that the family will not receive an insurance payment this year.

Example Solution. Let $X_i \sim Geo(\beta = 1.5)$ be the number of doctor visits for one individual in the family and $S_4 = X_1 + X_2 + X_3 + X_4$ be the number of doctor visits for the family. The sum of 4 independent geometric random variables each with mean $\beta = 1.5$ follows a negative binomial distribution, i.e. $S_4 \sim NB(\beta = 1.5, r = 4)$.

If the insurance pays 100 per visit beginning with the 4th visit for the family, then the family will not receive an insurance payment if they have less than 4 claims. This probability is

$$\begin{aligned}\Pr(S_4 < 4) &= \Pr(S_4 = 0) + \Pr(S_4 = 1) + \Pr(S_4 = 2) + \Pr(S_4 = 3) \\ &= (1+1.5)^{-4} + \frac{4(1.5)}{(1+1.5)^5} + \frac{4(5)(1.5^2)}{2(1+1.5)^6} + \frac{4(5)(6)(1.5^3)}{3!(1+1.5)^7} \\ &= 0.0256 + 0.0614 + 0.0922 + 0.1106 = 0.2898.\end{aligned}$$

7.3 Collective Risk Model

In this section, you learn:

- mathematical representation of the collective risk model
- how to evaluate moments, generating functions, and the distribution function of the collective risk model
- applications of collective risk model in stop-loss insurance
- Tweedie compound Poisson distribution as a special case of the collective risk model

7.3.1 Moments and Distribution

Under the collective risk model $S_N = X_1 + \cdots + X_N$, $\{X_i\}$ are *iid*, and independent of N. Let $\mu = \mathrm{E}\,(X_i)$ and $\sigma^2 = \mathrm{Var}\,(X_i)$ for all i. Thus, conditional on N, we have that the expectation of the sum is the sum of expectations and that the variance of the sum is the sum of variances,

$$\begin{aligned}\mathrm{E}(S|N) &= \mathrm{E}(X_1 + \cdots + X_N|N) = \mu N \\ \mathrm{Var}(S|N) &= \mathrm{Var}(X_1 + \cdots + X_N|N) = \sigma^2 N.\end{aligned}$$

Using the law of iterated expectations from Appendix Section 18.2, the mean of the aggregate loss is

$$\mathrm{E}(S_N) = \mathrm{E}_N[\mathrm{E}_S(S|N)] = \mathrm{E}_N(N\mu) = \mu\ \mathrm{E}(N).$$

Using the law of total variance from Appendix Section 18.2, the variance of the aggregate loss is

$$\begin{aligned}\mathrm{Var}(S_N) &= \mathrm{E}_N[\mathrm{Var}(S_N|N)] + \mathrm{Var}_N[\mathrm{E}(S_N|N)] \\ &= \mathrm{E}_N\left[\sigma^2 N\right] + \mathrm{Var}_N\left[\mu N\right] \\ &= \sigma^2\ \mathrm{E}[N] + \mu^2\ \mathrm{Var}[N].\end{aligned}$$

Special Case: Poisson Distributed Frequency. If $N \sim Poi(\lambda)$, then

$$\begin{aligned}
\mathrm{E}(N) &= \mathrm{Var}(N) = \lambda \\
\mathrm{E}(S_N) &= \lambda \ \mathrm{E}(X) \\
\mathrm{Var}(S_N) &= \lambda(\sigma^2 + \mu^2) = \lambda \ \mathrm{E}(X^2).
\end{aligned}$$

Example 7.3.1. Actuarial Exam Question. The number of accidents follows a Poisson distribution with mean 12. Each accident generates 1, 2, or 3 claimants with probabilities 1/2, 1/3, and 1/6 respectively.

Calculate the variance in the total number of claimants.

Example Solution.

$$\begin{aligned}
\mathrm{E}(X^2) &= 1^2\left(\frac{1}{2}\right) + 2^2\left(\frac{1}{3}\right) + 3^2\left(\frac{1}{6}\right) = \frac{10}{3} \\
\Rightarrow \mathrm{Var}(S_N) &= \lambda \ \mathrm{E}(X^2) = 12\left(\frac{10}{3}\right) = 40.
\end{aligned}$$

Alternatively, using the general approach, $\mathrm{Var}(S_N) = \sigma^2 \mathrm{E}(N) + \mu^2 \mathrm{Var}(N)$, where

$$\begin{aligned}
\mathrm{E}(N) &= \mathrm{Var}(N) = 12 \\
\mu &= \mathrm{E}(X) = 1\left(\frac{1}{2}\right) + 2\left(\frac{1}{3}\right) + 3\left(\frac{1}{6}\right) = \frac{5}{3} \\
\sigma^2 &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = \frac{10}{3} - \frac{25}{9} = \frac{5}{9} \\
\Rightarrow \ \mathrm{Var}(S_N) &= \left(\frac{5}{9}\right)(12) + \left(\frac{5}{3}\right)^2(12) = 40.
\end{aligned}$$

In general, the moments of S_N can be derived from its moment generating function (*mgf*). Because X_i's are *iid*, we denote the *mgf* of X as $M_X(t) = \mathrm{E}\ (e^{tX})$. Using the law of iterated expectations, the *mgf* of S_N is

$$\begin{aligned}
M_{S_N}(t) &= \mathrm{E}(e^{tS_N}) = \mathrm{E}_N[\ \mathrm{E}(e^{tS_N}|N)\ ] \\
&= \mathrm{E}_N\left[\ \mathrm{E}\left(e^{t(X_1+\cdots+X_N)}\right)\ \right] = \mathrm{E}_N\left[\mathrm{E}(e^{tX_1})\cdots\mathrm{E}(e^{tX_N})\right] \qquad \text{since } X_i\text{'s are independent} \\
&= \mathrm{E}_N[\ (M_X(t))^N\ ].
\end{aligned}$$

Now, recall that the probability generating function (*pgf*) of N is $P_N(z) = \mathrm{E}(z^N)$. Denote $M_X(t) = z$. Substituting into the expression for the *mgf* of S_N above, it is shown

$$M_{S_N}(t) = \mathrm{E}\ (z^N) = P_N(z) = P_N[M_X(t)].$$

Similarly, if S_N is discrete, one can show the *pgf* of S_N is:

$$P_{S_N}(z) = P_N[P_X(z)].$$

To get $\mathrm{E}(S_N) = M'_{S_N}(0)$, we use the chain rule

$$M'_{S_N}(t) = \frac{\partial}{\partial t} P_N(M_X(t)) = P'_N(M_X(t))M'_X(t)$$

and recall $M_X(0) = 1, M'_X(0) = \mathrm{E}(X) = \mu, P'_N(1) = \mathrm{E}(N)$. So,

$$\mathrm{E}(S_N) = M'_{S_N}(0) = P'_N(M_X(0))M'_X(0) = \mu\mathrm{E}(N).$$

Similarly, one could use relation $\mathrm{E}(S_N^2) = M''_{S_N}(0)$ to get

$$\mathrm{Var}(S_N) = \sigma^2\mathrm{E}(N) + \mu^2\mathrm{Var}(N).$$

Special Case. Poisson Frequency. Let $N \sim Poi(\lambda)$. Thus, the *pgf* of N is $P_N(z) = e^{\lambda(z-1)}$ and the *mgf* of S_N is

$$M_{S_N}(t) = P_N[M_X(t)] = e^{\lambda(M_X(t)-1)}.$$

Taking derivatives yields

$$\begin{aligned} M'_{S_N}(t) &= e^{\lambda(M_X(t)-1)}\ \lambda\ M'_X(t) = M_{S_N}(t)\ \lambda\ M'_X(t) \\ M''_{S_N}(t) &= M_{S_N}(t)\ \lambda\ M''_X(t) + [\ M_{S_N}(t)\ \lambda\ M'_X(t)\ ]\ \lambda\ M'_X(t). \end{aligned}$$

Evaluating these at $t = 0$ yields

$$\mathrm{E}(S_N) = M'_{S_N}(0) = \lambda\mathrm{E}(X) = \lambda\mu$$

and

$$\begin{aligned} M''_{S_N}(0) &= \lambda\mathrm{E}(X^2) + \lambda^2\mu^2 \\ \Rightarrow \mathrm{Var}(S_N) &= \lambda\mathrm{E}(X^2) + \lambda^2\mu^2 - (\lambda\mu)^2 = \lambda\ \mathrm{E}(X^2). \end{aligned}$$

Example 7.3.2. Actuarial Exam Question. You are the producer of a television quiz show that gives cash prizes. The number of prizes, N, and prize amount, X, have the following distributions:

n	$\Pr(N = n)$	x	$\Pr(X = x)$
1	0.8	0	0.2
2	0.2	100	0.7
		1000	0.1

Your budget for prizes equals the expected aggregate cash prizes plus the standard deviation of aggregate cash prizes. Calculate your budget.

Example Solution. We need to calculate the mean and standard deviation of the aggregate (sum) of cash prizes. The moments of the frequency distribution N are

$$\begin{aligned} \mathrm{E}(N) &= 1(0.8) + 2(0.2) = 1.2 \\ \mathrm{E}(N^2) &= 1^2(0.8) + 2^2(0.2) = 1.6 \\ \mathrm{Var}(N) &= \mathrm{E}(N^2) - [\mathrm{E}(N)]^2 = 0.16. \end{aligned}$$

The moments of the severity distribution X are

$$\begin{aligned} \mathrm{E}(X) &= 0(0.2) + 100(0.7) + 1000(0.1) = 170 = \mu \\ \mathrm{E}(X^2) &= 0^2(0.2) + 100^2(0.7) + 1000^2(0.1) = 107,000 \\ \mathrm{Var}(X) &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = 78,100 = \sigma^2. \end{aligned}$$

Thus, the mean and variance of the aggregate cash prize are

$$\begin{aligned} \mathrm{E}(S_N) &= \mu \mathrm{E}(N) = 170(1.2) = 204 \\ \mathrm{Var}(S_N) &= \sigma^2 \mathrm{E}(N) + \mu^2 \mathrm{Var}(N) \\ &= 78,100(1.2) + 170^2(0.16) = 98,344. \end{aligned}$$

This gives the following required budget

$$\begin{aligned} Budget &= \mathrm{E}(S_N) + \sqrt{\mathrm{Var}(S_N)} \\ &= 204 + \sqrt{98,344} = 517.60. \end{aligned}$$

The distribution of S_N is called a compound distribution, and it can be derived based on the convolution of F_X as follows:

$$\begin{aligned} F_{S_N}(s) &= \Pr\left(X_1 + \cdots + X_N \le s\right) \\ &= \mathrm{E}\left[\Pr\left(X_1 + \cdots + X_N \le s | N = n\right)\right] \\ &= \mathrm{E}\left[F_X^{*N}(s)\right] \\ &= p_0 + \sum_{n=1}^{\infty} p_n F_X^{*n}(s). \end{aligned}$$

Example 7.3.3. Actuarial Exam Question. The number of claims in a period has a geometric distribution with mean 4. The amount of each claim X follows $\Pr(X = x) = 0.25,\ x = 1, 2, 3, 4$, i.e. a discrete uniform distribution on $\{1, 2, 3, 4\}$. The number of claims and the claim amounts are independent. Let S_N denote the aggregate claim amount in the period. Calculate $F_{S_N}(3)$.

Example Solution. By definition, we have

$$\begin{aligned}
F_{S_N}(3) &= \Pr\left(\sum_{i=1}^{N} X_i \le 3\right) = \sum_{n=0}^{\infty} \Pr\left(\sum_{i=1}^{n} X_i \le 3|N=n\right) \Pr(N=n) \\
&= \sum_{n} F^{*n}(3)\, p_n = \sum_{n=0}^{3} F^{*n}(3) p_n \\
&= p_0 + F^{*1}(3)\ p_1 + F^{*2}(3)\ p_2 + F^{*3}(3)\ p_3.
\end{aligned}$$

Because $N \sim Geo(\beta = 4)$, we know that

$$p_n = \frac{1}{1+\beta}\left(\frac{\beta}{1+\beta}\right)^n = \frac{1}{5}\left(\frac{4}{5}\right)^n.$$

For the claim severity distribution, recursively, we have

$$\begin{aligned}
F^{*1}(3) &= \Pr(X \le 3) = \frac{3}{4} \\
F^{*2}(3) &= \sum_{y \le 3} F^{*1}(3-y) f(y) = F^{*1}(2) f(1) + F^{*1}(1) f(2) \\
&= \frac{1}{4}\left[F^{*1}(2) + F^{*1}(1)\right] = \frac{1}{4}\left[\Pr(X \le 2) + \Pr(X \le 1)\right] \\
&= \frac{1}{4}\left(\frac{2}{4} + \frac{1}{4}\right) = \frac{3}{16} \\
F^{*3}(3) &= \Pr(X_1 + X_2 + X_3 \le 3) = \Pr(X_1 = X_2 = X_3 = 1) = \left(\frac{1}{4}\right)^3.
\end{aligned}$$

Notice that we did not need to recursively calculate $F^{*3}(3)$ by recognizing that each $X \in \{1,2,3,4\}$, so the only way of obtaining $X_1 + X_2 + X_3 \le 3$ is to have $X_1 = X_2 = X_3 = 1$. Additionally, for $n \ge 4$, $F^{*n}(3) = 0$ since it is impossible for the sum of 4 or more X's to be less than 3. For $n = 0$, $F^{*0}(3) = 1$ since the sum of 0 X's is 0, which is always less than 3. Laying out the probabilities systematically,

x	$F^{*1}(x)$	$F^{*2}(x)$	$F^{*3}(x)$
0			
1	$\frac{1}{4}$	0	
2	$\frac{2}{4}$	$\left(\frac{1}{4}\right)^2$	
3	$\frac{3}{4}$	$\frac{3}{16}$	$\left(\frac{1}{4}\right)^3$

Finally,

$$\begin{aligned}
F_{S_N}(3) &= p_0 + F^{*1}(3)\ p_1 + F^{*2}(3)\ p_2 + F^{*3}(3)\ p_3 \\
&= \frac{1}{5} + \frac{3}{4}\left(\frac{4}{25}\right) + \frac{3}{16}\left(\frac{16}{125}\right) + \frac{1}{64}\left(\frac{64}{625}\right) = 0.3456.
\end{aligned}$$

Example 7.3.4. Convolution Method to Compute the Aggregate Loss Distribution. Consider the Wisconsin Property Fund data that was introduced in Section 1.3 and is available in Appendix Section 22.1. Specifically, we examine building and content claims with frequence of claims given by the variable `Freq` and amount of claims given by `BCClaim`. Assume a Poisson distribution for the frequency and a gamma distribution for the severity. The following block of `R` code illustrates how to retrieve the data and reviews parameter estimation from prior chapters.

```
datraw <- read.csv("Data/WiscPropFund.csv")
# remove extreme observations to speed up the evaluation of distribution of
# aggregate losses
index <- which(datraw$Freq < 100 & datraw$BCClaim < 250000)
dat <- datraw[index, ]
# head(dat,n=3) tail(dat,n=3)

# Assume a Poisson for claim frequency
lambda <- mean(dat$Freq)
# print(lambda) Assume a gamma for claim severity
index <- which(dat$BCClaim > 0)
n <- dat$Freq[index]
xbar <- dat$BCClaim[index]/dat$Freq[index]
fit <- glm(xbar ~ 1, family = Gamma(link = "log"), weight = 1/n)
mu <- unname(exp(fit$coefficients))
phi <- summary(fit)$dispersion
a = 1/phi
s = mu * phi
# print(c(a,s))
```

With the parameter estimates in place, we are now in a position to calculate distribution of $S = X_1 + X_2 + \cdots + X_N$ using convolution method. Figure 7.1 summarizes the aggregate loss distribution. The following block of code demonstrates its calculation.

```
Nmax <- 1000
# CDF
FAgg <- function(y) {
    re <- dpois(0, lambda)
    for (i in 1:Nmax) {
        re <- re + dpois(i, lambda) * pgamma(y, shape = i * a, scale = s)
    }
    re <- ifelse(y < 0, NA, re)
    return(re)
}
# PDF
fAgg <- function(y) {
    re <- dpois(0, lambda)
    for (i in 1:Nmax) {
        re <- re + dpois(i, lambda) * dgamma(y * (y > 0) - 1 * (y <= 0), shape = i *
            a, scale = s)
    }
    re <- ifelse(y < 0, NA, re)
    return(re)
}
# Numerical examples
obs <- c(-1, 0, 1, 10, 100, 1000, 10000, 100000, 1000000)
# FAgg(obs) fAgg(obs)
```

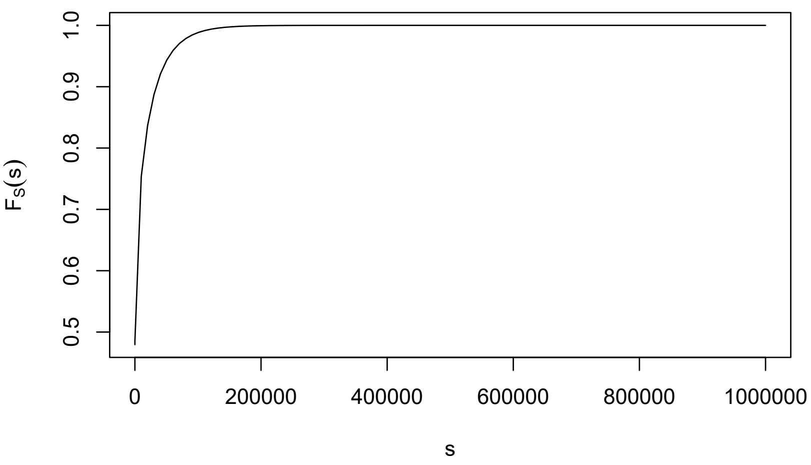


FIGURE 7.1: **Aggregate Loss Distribution for Wisconsin Property Fund Building and Loss Claims**

When $\mathrm{E}(N)$ and $\mathrm{Var}(N)$ are known, one may also use a type of central limit theorem to approximate the distribution of S_N as in the individual risk model. That is, $\frac{S_N - \mathrm{E}(S_N)}{\sqrt{\mathrm{Var}(S_N)}}$ approximately follows the standard normal distribution $N(0,1)$. From this type of central limit theorem, the approximation works well if $\mathrm{E}[N]$ is sufficiently large.

Example 7.3.5. Actuarial Exam Question. You are given:

	Mean	Standard Deviation
Number of Claims	8	3
Individual Losses	$10,000$	$3,937$

As a benchmark, use the normal approximation to determine the probability that the aggregate loss will exceed 150% of the expected loss.

Example Solution. To use the normal approximation, we must first find the mean and variance of the aggregate loss S

$$\begin{aligned}
\mathrm{E}(S_N) &= \mu\ \mathrm{E}(N) = 10,000(8) = 80,000 \\
\mathrm{Var}(S_N) &= \sigma^2\ \mathrm{E}(N) + \mu^2\ \mathrm{Var}(N) \\
&= 3937^2(8) + 10000^2(3^2) = 1,023,999,752 \\
\sqrt{\mathrm{Var}(S_N)} &= 31,999.996 \approx 32,000.
\end{aligned}$$

Then under the normal approximation, aggregate loss S_N is approximately normal with mean 80,000 and standard deviation 32,000. The probability that S_N will exceed 150% of the expected aggregate loss is therefore

$$\begin{aligned}
\Pr(S_N > 1.5\mathrm{E}(S_N)\ ) &= \Pr\left(\frac{S_N - \mathrm{E}(S_N)}{\sqrt{\mathrm{Var}(S_N)}} > \frac{1.5\ \mathrm{E}(S_N) - \mathrm{E}(S_N)}{\sqrt{\mathrm{Var}(S_N)}}\right) \\
&\approx \Pr\left(Z > \frac{0.5\ \mathrm{E}(S_N)}{\sqrt{\mathrm{Var}(S_N)}}\right), \quad \text{where } Z \sim N(0,1) \\
&= \Pr\left(Z > \frac{0.5(80,000)}{32,000}\right) = \Pr(Z > 1.25) \\
&= 1 - \Phi(1.25) = 0.1056.
\end{aligned}$$

Example 7.3.6. Actuarial Exam Question. For an individual over 65:

(i) The number of pharmacy claims is a Poisson random variable with mean 27.

(ii) The amount of each pharmacy claim is uniformly distributed between 5 and 97.

(iii) The amounts of the claims and the number of claims are mutually independent.

Estimate the probability that aggregate claims for this individual will exceed 2000 using the normal approximation.

Example Solution. We have claim frequency $N \sim Poi(\lambda = 25)$ and claim severity $X \sim U\,(5, 95)$. To use the normal approximation, we need to find the mean and variance of the aggregate claims S_N. Note

$$\begin{array}{ll} \mathrm{E}(N) = 25 & \mathrm{Var}(N) = 25 \\ \mathrm{E}(X) = \frac{5+95}{2} = 50 = \mu & \mathrm{Var}(X) = \frac{(95-5)^2}{12} = 675 = \sigma^2. \end{array}$$

Then for S_N,

$$\begin{aligned} \mathrm{E}(S_N) &= \mu \ \mathrm{E}(N) = 50(25) = 1,250 \\ \mathrm{Var}(S_N) &= \sigma^2 \ \mathrm{E}(N) + \mu^2 \ \mathrm{Var}(N) \\ &= 675(25) + 50^2(25) = 79,375. \end{aligned}$$

Using the normal approximation, S_N is approximately normal with mean 1,250 and variance 79,375. The probability that S_N exceeds 2,000 is

$$\begin{aligned} \Pr(S_N > 2,000) &= \Pr\left(\frac{S_N - \mathrm{E}(S_N)}{\sqrt{\mathrm{Var}(S_N)}} > \frac{2,000 - \mathrm{E}(S_N)}{\sqrt{\mathrm{Var}(S_N)}}\right) \\ &\approx \Pr\left(Z > \frac{2,000 - 1,250}{\sqrt{79,375}}\right), \quad \text{where } Z \sim N(0,1) \\ &= \Pr(Z > 2.662) = 1 - \Phi(2.662) = 0.003884. \end{aligned}$$

7.3.2 Stop-loss Insurance

Recall the coverage modifications on the individual policy level in Section 5.1. Insurance on the aggregate loss S_N, subject to a deductible d, is called *net stop-loss insurance*. The expected value of the amount of the aggregate loss in excess of the deductible,

$$\mathrm{E}[(S_N - d)_+]$$

is known as the *net stop-loss premium*.

To calculate the net stop-loss premium, we have

$$\begin{aligned} \mathrm{E}(S_N - d)_+ &= \begin{cases} \int_d^\infty (s-d) f_{S_N}(s) ds & \text{for continuous } S_N \\ \sum_{s>d} (s-d) f_{S_N}(s) & \text{for discrete } S_N \end{cases} \\ &= \mathrm{E}(S_N) - \mathrm{E}(S_N \wedge d) \end{aligned}$$

Example 7.3.7. Actuarial Exam Question. In a given week, the number of projects that require you to work overtime has a geometric distribution with $\beta = 2$. For each project, the distribution of the number of overtime hours in the week, X, is as follows:

x	$f(x)$
5	0.2
10	0.3
20	0.5

The number of projects and the number of overtime hours are independent. You will get paid for overtime hours in excess of 15 hours in the week. Calculate the expected number of overtime hours for which you will get paid in the week.

Example Solution. The number of projects in a week requiring overtime work has distribution $N \sim Geo(\beta = 2)$, while the number of overtime hours worked per project has distribution X as described above. The aggregate number of overtime hours in a week is S_N and we are therefore looking for

$$\mathrm{E}(S_N - 15)_+ = \mathrm{E}(S_N) - \mathrm{E}(S_N \wedge 15).$$

To find $\mathrm{E}(S_N) = \mathrm{E}(X)\ \mathrm{E}(N)$, we have

$$\begin{aligned} &\mathrm{E}(X) = 5(0.2) + 10(0.3) + 20(0.5) = 14 \\ &\mathrm{E}(N) = 2 \\ \Rightarrow\ &\mathrm{E}(S) = \mathrm{E}(X)\ \mathrm{E}(N) = 14(2) = 28. \end{aligned}$$

To find $\mathrm{E}(S_N \wedge 15) = 0 \Pr(S_N = 0) + 5 \Pr(S_N = 5) + 10 \Pr(S_N = 10) + 15 \Pr(S_N \geq 15)$, we have

$$\begin{aligned} \Pr(S_N = 0) &= \Pr(N = 0) = \frac{1}{1+\beta} = \frac{1}{3} \\ \Pr(S_N = 5) &= \Pr(X = 5,\ N = 1) = 0.2\left(\frac{2}{9}\right) = \frac{0.4}{9} \\ \Pr(S_N = 10) &= \Pr(X = 10,\ N = 1) + \Pr(X_1 = X_2 = 5, N = 2) \\ &= 0.3\left(\frac{2}{9}\right) + (0.2)(0.2)\left(\frac{4}{27}\right) = 0.0726 \\ \Pr(S_N \geq 15) &= 1 - \left(\frac{1}{3} + \frac{0.4}{9} + 0.0726\right) = 0.5496 \\ \Rightarrow \mathrm{E}(S_N \wedge 15) &= 0 \Pr(S_N = 0) + 5 \Pr(S_N = 5) + 10 \Pr(S_N = 10) + 15 \Pr(S_N \geq 15) \\ &= 0\left(\frac{1}{3}\right) + 5\left(\frac{0.4}{9}\right) + 10(0.0726) + 15(0.5496) = 9.193. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{E}(S_N - 15)_+ &= \mathrm{E}(S_N) - \mathrm{E}(S_N \wedge 15) \\ &= 28 - 9.193 = 18.807. \end{aligned}$$

Recursive Net Stop-Loss Premium Calculation. For the discrete case, this can be computed recursively as

$$\mathrm{E}\left[(S_N-(j+1)h)_+\right]=\mathrm{E}\left[(S_N-jh)_+\right]-h\left[1-F_{S_N}(jh)\right].$$

This assumes that the support of S_N is equally spaced over units of h.

To establish this, we assume that $h=1$. We have

$$\begin{aligned}\mathrm{E}\left[(S_N-(j+1))_+\right]&=\mathrm{E}(S_N)-\mathrm{E}[S_N\wedge(j+1)]\ , \quad \text{and}\\ \mathrm{E}\left[(S_N-j)_+\right]&=\mathrm{E}(S_N)-\mathrm{E}[S_N\wedge j]\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{E}\left[(S_N-(j+1))_+\right]-\mathrm{E}\left[(S_N-j)_+\right]&=\{\mathrm{E}(S_N)-\mathrm{E}(S_N\wedge(j+1))\}-\{\mathrm{E}(S_N)-\mathrm{E}(S_N\wedge j)\}\\ &=\mathrm{E}\left(S_N\wedge j\right)-\mathrm{E}\left[S\wedge(j+1)\right]\end{aligned}$$

We can write

$$\begin{aligned}\mathrm{E}\left[S_N\wedge(j+1)\right]&=\sum_{x=0}^{j}xf_{S_N}(x)+(j+1)\ \Pr(S_N\geq j+1)\\ &=\sum_{x=0}^{j-1}xf_{S_N}(x)+j\ \Pr(S_N=j)+(j+1)\ \Pr(S_N\geq j+1)\end{aligned}$$

Similarly,

$$\mathrm{E}(S_N\wedge j)=\sum_{x=0}^{j-1}xf_{S_N}(x)+j\ \Pr(S_N\geq j)$$

With these expressions, we have

$$\begin{aligned}&\mathrm{E}\left[(S_N-(j+1))_+\right]-\mathrm{E}\ \left[(S_N-j)_+\right]\\ &=\mathrm{E}\left(S_N\wedge j\right)-\mathrm{E}\left[S\wedge(j+1)\right]\\ &=\left\{\sum_{x=0}^{j-1}xf_{S_N}(x)+j\ \Pr(S_N\geq j)\right\}-\left\{\sum_{x=0}^{j-1}xf_{S_N}(x)+j\ \Pr(S_N=j)+(j+1)\ \Pr(S_N\geq j+1)\right\}\\ &=j\ \left[\Pr(S_N\geq j)-\Pr(S_N=j)\right]-(j+1)\ \Pr(S_N\geq j+1)\\ &=j\ \Pr(S_N>j)-(j+1)\ \Pr(S_N\geq j+1)\qquad(\text{note }\Pr(S_N>j)=\Pr(S_N\geq j+1))\\ &=-\Pr(S_N\geq j+1)=-\left[1-F_{S_N}(j)\right],\end{aligned}$$

as required.

Example 7.3.8. Actuarial Exam Question - Continued. Recall that the goal of this question was to calculate E $(S_N - 15)_+$. Note that the support of S_N is equally spaced over units of 5, so this question can also be done recursively, using the expression above with steps of $h = 5$:

- Step 1:

$$\begin{aligned} \mathrm{E}\ (S_N - 5)_+ &= \mathrm{E}(S_N) - 5[1 - \Pr(S_N \le 0)] \\ &= 28 - 5\left(1 - \frac{1}{3}\right) = \frac{74}{3} = 24.6667. \end{aligned}$$

- Step 2:

$$\begin{aligned} \mathrm{E}\ (S_N - 10)_+ &= \mathrm{E}\ (S_N - 5)_+ - 5[1 - \Pr(S_N \le 5)] \\ &= \frac{74}{3} - 5\left(1 - \frac{1}{3} - \frac{0.4}{9}\right) = 21.555. \end{aligned}$$

- Step 3:

$$\begin{aligned} \mathrm{E}\ (S_N - 15)_+ &= \mathrm{E}\ (S_N - 10)_+ - 5[1 - \Pr(S_N \le 10)] \\ &= \mathrm{E}\ (S_N - 10)_+ - 5\Pr(S_N \ge 15) \\ &= 21.555 - 5(0.5496) = 18.807. \end{aligned}$$

7.3.3 Closed-form Distributions

There are a few combinations of claim frequency and severity distributions that result in an easy-to-compute distribution for aggregate losses. This section provides some simple examples. Although these examples are computationally convenient, they are generally too simple to be used in practice.

Example 7.3.9. Geometric Frequency, Exponential Severity. One has a closed-form expression for the aggregate loss distribution by assuming a geometric frequency distribution and an exponential severity distribution.

Assume that claim count N is geometric with mean $\mathrm{E}(N) = \beta$, and that claim amount X is exponential with $\mathrm{E}(X) = \theta$. Recall that the *pgf* of N and the *mgf* of X are:

$$\begin{aligned} P_N(z) &= \frac{1}{1 - \beta(z-1)} \\ M_X(t) &= \frac{1}{1 - \theta t}. \end{aligned}$$

Thus, the *mgf* of aggregate loss S_N can be expressed two ways (for details, see *Technical Supplement 7.A.3*)

$$\begin{aligned} M_{S_N}(t) &= P_N[M_X(t)] = \frac{1}{1-\beta\left(\frac{1}{1-\theta t}-1\right)} \\ &= 1+\frac{\beta}{1+\beta}\left([1-\theta(1+\beta)t]^{-1}-1\right) \qquad (7.1) \\ &= \frac{1}{1+\beta}(1)+\frac{\beta}{1+\beta}\left(\frac{1}{1-\theta(1+\beta)t}\right). \qquad (7.2) \end{aligned}$$

From (7.1), we note that S_N is equivalent to the compound distribution of $S_N = X_1^* + \cdots + X_{N^*}^*$, where N^* is a Bernoulli with mean $\beta/(1+\beta)$ and X^* is an exponential with mean $\theta(1+\beta)$. To see this, we examine the *mgf* of S:

$$M_{S_N}(t) = P_N[M_X(t)] = P_{N^*}[M_{X^*}(t)],$$

where

$$\begin{aligned} P_{N^*}(z) &= 1+\frac{\beta}{1+\beta}(z-1), \\ M_{X^*}(t) &= \frac{1}{1-\theta(1+\beta)t}. \end{aligned}$$

From (7.2), we note that S_N is also equivalent to a two-point mixture of 0 and X^*. Specifically,

$$S_N = \begin{cases} 0 & \text{with probability } \Pr(N^*=0) = 1/(1+\beta) \\ Y^* & \text{with probability } \Pr(N^*=1) = \beta/(1+\beta). \end{cases}$$

The distribution function of S_N is:

$$\begin{aligned} \Pr(S_N = 0) &= \frac{1}{1+\beta} \\ \Pr(S_N > s) &= \Pr(X^* > s) = \frac{\beta}{1+\beta}\exp\left(-\frac{s}{\theta(1+\beta)}\right) \end{aligned}$$

with pdf for $s > 0$,

$$f_{S_N}(s) = \frac{\beta}{\theta(1+\beta)^2}\exp\left(-\frac{s}{\theta(1+\beta)}\right).$$

Example 7.3.10. Exponential Severity. Consider a collective risk model with an exponential severity and an arbitrary frequency distribution. Recall that if $X_i \sim Exp(\theta)$, then the sum of *iid* exponential random variables,

$S_n = X_1 + \cdots + X_n$, has a gamma distribution, i.e. $S_n \sim Gam(n, \theta)$. This has cdf:

$$\begin{aligned} F_X^{*n}(s) &= \Pr(S_n \le s) = \int_0^s \frac{1}{\Gamma(n)\theta^n} s^{n-1} \exp\left(-\frac{s}{\theta}\right) ds \\ &= 1 - \sum_{j=0}^{n-1} \frac{1}{j!} \left(\frac{s}{\theta}\right)^j e^{-s/\theta}. \end{aligned}$$

The last equality is derived by applying integration by parts $n-1$ times.

For the aggregate loss distribution, we can interchange the order of summations in the second line below to get

$$\begin{aligned} F_S(s) &= p_0 + \textstyle\sum_{n=1}^{\infty} p_n F_X^{*n}(s) \\ &= 1 - \textstyle\sum_{n=1}^{\infty} p_n \sum_{j=0}^{n-1} \frac{1}{j!} \left(\frac{s}{\theta}\right)^j e^{-s/\theta} \\ &= 1 - e^{-s/\theta} \textstyle\sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{s}{\theta}\right)^j \overline{P}_j \end{aligned}$$

where $\overline{P}_j = p_{j+1} + p_{j+2} + \cdots = \Pr(N > j)$ is the "survival function" of the claims count distribution.

7.3.4 Tweedie Distribution

In this section, we examine a particular compound distribution where the number of claims has a Poisson distribution and the amount of claims has a gamma distribution. This specification leads to what is known as a Tweedie distribution. The Tweedie distribution has a mass probability at zero and a continuous component for positive values. Because of this feature, it is widely used in insurance claims modeling, where the zero mass is interpreted as no claims and the positive component as the amount of claims.

Specifically, consider the collective risk model $S_N = X_1 + \cdots + X_N$. Suppose that N has a Poisson distribution with mean λ, and each X_i has a gamma distribution with shape parameter α and scale parameter γ. The Tweedie distribution is derived as the Poisson sum of gamma variables. To understand the distribution of S_N, we first examine the mass probability at zero. The aggregate loss is zero when no claims occurred, i.e.

$$\Pr(S_N = 0) = \Pr(N = 0) = e^{-\lambda}.$$

In addition, note that S_N conditional on $N = n$, denoted by $S_n = X_1 + \cdots + X_n$, follows a gamma distribution with shape $n\alpha$ and scale γ. Thus, for $s > 0$, the density of a Tweedie distribution can be calculated as

$$\begin{aligned} f_{S_N}(s) &= \sum_{n=1}^{\infty} p_n f_{S_n}(s) \\ &= \sum_{n=1}^{\infty} e^{-\lambda} \frac{(\lambda)^n}{n!} \frac{\gamma^{na}}{\Gamma(n\alpha)} s^{n\alpha-1} e^{-s\gamma}. \end{aligned}$$

Thus, the Tweedie distribution can be thought of a mixture of zero and a positive valued distribution, which makes it a convenient tool for modeling insurance claims and for calculating pure premiums. The mean and variance of the Tweedie compound Poisson model are:

$$\mathrm{E}(S_N) = \lambda \frac{\alpha}{\gamma} \quad \text{and} \quad \mathrm{Var}(S_N) = \lambda \frac{\alpha(1+\alpha)}{\gamma^2}.$$

As another important feature, the Tweedie distribution is a special case of exponential dispersion models, a class of models used to describe the random component in generalized linear models. To see this, we consider the following reparameterization:

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}, \qquad \alpha = \frac{2-p}{p-1}, \qquad \frac{1}{\gamma} = \phi(p-1)\mu^{p-1}.$$

With the above relationships, one can show that the distribution of S_N is

$$f_{S_N}(s) = \exp\left[\frac{1}{\phi}\left(\frac{-s}{(p-1)\mu^{p-1}} - \frac{\mu^{2-p}}{2-p}\right) + C(s;\phi)\right]$$

where

$$C(s;\phi) = \begin{cases} 0 & \text{if } s = 0 \\ \log \sum_{n\geq 1} \left\{ \frac{(1/\phi)^{1/(p-1)} s^{(2-p)/(p-1)}}{(2-p)(p-1)^{(2-p)/(p-1)}} \right\}^n \frac{1}{n!\ \Gamma[n(2-p)/(p-1)]s} & \text{if } s > 0. \end{cases}$$

Hence, the distribution of S_N belongs to the exponential family with parameters μ, ϕ, and $1 < p < 2$, and we have

$$\mathrm{E}(S_N) = \mu \quad \text{and} \quad \mathrm{Var}(S_N) = \phi\mu^p.$$

This allows us to use the Tweedie distribution with generalized linear models to model claims. It is also worth mentioning the two limiting cases of the Tweedie model: $p \to 1$ results in the Poisson distribution and $p \to 2$ results in the gamma distribution. Thus, the Tweedie model accommodates the situations in between the gamma and Poisson distributions, which makes intuitive sense as it is the Poisson sum of gamma random variables.

7.4 Computing the Aggregate Claims Distribution

In this section, you learn:

- the recursive method to compute the aggregate claims distribution
- the simulation approach to compute the aggregate claims distribution

Computing the distribution of aggregate losses is a difficult, yet important, problem. As we have seen, for both individual risk model and collective risk model, computing the distribution frequently involves the evaluation of a n-fold convolution. To make the problem tractable, one strategy is to use a distribution that is easy to evaluate to approximate the aggregate loss distribution. For instance, normal distribution is a natural choice based on central limit theorem where parameters of the normal distribution can be estimated by matching the moments. This approach has its strength and limitations. Its main advantage is the ease of computation. The disadvantages are: first, the size and direction of approximation error are unknown; second, the approximation may fail to capture some special features of the aggregate loss such as mass point at zero.

This section discusses two practical approaches to computing the distribution of aggregate loss, the recursive method and simulation.

7.4.1 Recursive Method

The recursive method applies to compound models where the frequency component N belongs to either $(a,b,0)$ or $(a,b,1)$ class (see Sections 3.3 and 3.5.1) and the severity component X has a discrete distribution. For continuous X, a common practice is to first discretize the severity distribution, after which the recursive method is ready to apply.

Assume that N is in the $(a,b,1)$ class so that $p_k = \left(a+\frac{b}{k}\right)p_{k-1}, k = 2,3,\ldots$. Further assume that the support of X is $\{0,1,\ldots,m\}$, discrete and finite. Then, the probability function of S_N is:

$$\begin{aligned} f_{S_N}(s) &= \Pr(S_N = s) \\ &= \frac{1}{1-af_X(0)}\left\{[p_1-(a+b)p_0]\,f_X(s) + \sum_{x=1}^{s\wedge m}\left(a+\frac{bx}{s}\right)f_X(x)f_{S_N}(s-x)\right\}. \end{aligned}$$

If N is in the $(a,b,0)$ class, then $p_1=(a+b)p_0$ and so

$$f_{S_N}(s)=\frac{1}{1-af_X(0)}\left\{\sum_{x=1}^{s\wedge m}\left(a+\frac{bx}{s}\right)f_X(x)f_{S_N}(s-x)\right\}.$$

Special Case: Poisson Frequency. If $N\sim Poi(\lambda)$, then $a=0$ and $b=\lambda$, and thus

$$f_{S_N}(s)=\frac{\lambda}{s}\left\{\sum_{x=1}^{s\wedge m}xf_X(x)f_{S_N}(s-x)\right\}.$$

Example 7.4.1. Actuarial Exam Question. The number of claims in a period N has a geometric distribution with mean 4. The amount of each claim X follows $\Pr(X=x)=0.25$, for $x=1,2,3,4$. The number of claims and the claim amount are independent. S_N is the aggregate claim amount in the period. Calculate $F_{S_N}(3)$.

Example Solution. The severity distribution X follows

$$f_X(x)=\frac{1}{4},\quad x=1,2,3,4.$$

The frequency distribution N is geometric with mean 4, which is a member of the $(a,b,0)$ class with $b=0$, $a=\frac{\beta}{1+\beta}=\frac{4}{5}$, and $p_0=\frac{1}{1+\beta}=\frac{1}{5}$. The support of severity component X is $\{1,\ldots,m=4\}$, discrete and finite. Thus, we can use the recursive method

$$\begin{aligned}f_{S_N}(x)&=1\sum_{y=1}^{x\wedge m}(a+0)f_X(y)f_{S_N}(x-y)\\&=\frac{4}{5}\sum_{y=1}^{x\wedge m}f_X(y)f_{S_N}(x-y).\end{aligned}$$

Specifically, we have

$$\begin{aligned}
f_{S_N}(0) &= \Pr(N=0) = p_0 = \frac{1}{5} \\
f_{S_N}(1) &= \frac{4}{5}\sum_{y=1}^{1} f_X(y) f_{S_N}(1-y) = \frac{4}{5} f_X(1) f_{S_N}(0) \\
&= \frac{4}{5}\left(\frac{1}{4}\right)\left(\frac{1}{5}\right) = \frac{1}{25} \\
f_{S_N}(2) &= \frac{4}{5}\sum_{y=1}^{2} f_X(y) f_{S_N}(2-y) = \frac{4}{5}\left[f_X(1) f_{S_N}(1) + f_X(2) f_{S_N}(0)\right] \\
&= \frac{4}{5}\left[\frac{1}{4}\left(\frac{1}{25}+\frac{1}{5}\right)\right] = \frac{4}{5}\left(\frac{6}{100}\right) = \frac{6}{125} \\
f_{S_N}(3) &= \frac{4}{5}\left[f_X(1) f_{S_N}(2) + f_X(2) f_{S_N}(1) + f_X(3) f_{S_N}(0)\right] \\
&= \frac{4}{5}\left[\frac{1}{4}\left(\frac{1}{25}+\frac{1}{5}+\frac{6}{125}\right)\right] = \frac{1}{5}\left(\frac{5+25+6}{125}\right) = 0.0576 \\
\Rightarrow \quad F_{S_N}(3) &= f_{S_N}(0) + f_{S_N}(1) + f_{S_N}(2) + f_{S_N}(3) = 0.3456.
\end{aligned}$$

Example 7.4.2. Convolution Method to Compute the Aggregate Loss Distribution - Continued. This is a continuation of Example 7.3.4 where we now compute the aggregate loss distribution using the recursive method. This requires discretization of the severity amounts and this illustration rounds claims to the nearest thousand. The following block of code illustrates the calculation.

```
# Discretized severity distribution
round_any = function(y, accuracy, f = round) {
    f(y/accuracy) * accuracy
}
# round to $1000
acc <- 1000
xbar_disc <- round_any(xbar, acc)/acc
dSev <- function(y) {
    re <- ecdf(xbar_disc)(y) - ecdf(xbar_disc)(y - 1)
    re
}

Fs0 <- function(y) {
    if (y < 0)
        return(NA)
    y_scale <- round_any(y, acc)/acc
    y_scale <- ifelse(y_scale > max(xbar_disc), max(xbar_disc), y_scale)
    s.out <- rep(NA, y_scale + 1)
    s.out[1] <- exp(-lambda)
    if (y_scale > 0) {
        for (i in 1:y_scale) {
            re <- 0
            for (j in 1:i) {
                re <- re + j * dSev(j) * s.out[i + 1 - j]
            }
            s.out[i + 1] <- re * lambda/i
        }
    }
    return(sum(s.out))
}
Fs <- function(y) sapply(y, Fs0)
obs <- c(-1, 0, 1, 10, 100, 1000, 10000, 100000, 1000000)
# Fs(obs)
```

7.4.2 Simulation

The distribution of aggregate loss can be evaluated using Monte Carlo simulation. You can get a broad introduction to simulation procedures in Chapter 8. For aggregate losses, the idea is that one can calculate the empirical distribution of S_N using a random sample. The expected value and variance of the aggregate loss can also be estimated using the sample mean and sample variance of the simulated values.

We now summarize simulation procedures for aggregate loss models. Let R be the size of the generated random sample of aggregate losses.

1. Individual Risk Model: $S_n = X_1 + \cdots + X_n$

   - Let $j = 1, \ldots, R$ be a counter. Start by setting $j = 1$.
   - Generate each individual loss realization x_{ij} for $i = 1, \ldots, n$. For example, this can be done using the inverse transformation method (Section 8.1.2).
   - Calculate the aggregate loss $s_j = x_{1j} + \cdots + x_{nj}$.
   - Repeat the above two steps for $j = 2, \ldots, R$ to obtain a size-R sample of S_n, i.e. $\{s_1, \ldots, s_R\}$.

2. Collective Risk Model: $S_N = X_1 + \cdots + X_N$

   - Let $j = 1, \ldots, R$ be a counter. Start by setting $j = 1$.
   - Generate the number of claims n_j from the frequency distribution N.
   - Given n_j, generate the amount of each claim independently from severity distribution X, denoted by $x_{1j}, \ldots, x_{n_j j}$.
   - Calculate the aggregate loss $s_j = x_{1j} + \cdots + x_{n_j j}$.
   - Repeat the above three steps for $j = 2, \ldots, R$ to obtain a size-R sample of S_N, i.e. $\{s_1, \ldots, s_R\}$.

Given the random sample of S, the empirical distribution can be calculated as

$$\hat{F}_S(s) = \frac{1}{R}\sum_{j=1}^{R} I(s_j \leq s),$$

where $I(\cdot)$ is an indicator function. The empirical distribution $\hat{F}_S(s)$ will converge to $F_S(s)$ almost surely as the sample size $R \to \infty$.

The above procedure assumes that the probability distributions, including the parameter values, of the frequency and severity distributions are known. In practice, one would need to first assume these distributions, estimate their parameters from data, and then assess the quality of model fit using various model validation tools (see Chapter 6). For instance, the assumptions in the collective risk model suggest a two-stage estimation where one model is developed for the number of claims N from data on claim counts, and another model is developed for the severity of claims X from data on the amount of claims.

Example 7.4.3. Recall Example 7.3.6 with an individual's claim frequency N has a Poisson distribution with mean $\lambda = 25$ and claim severity X is uniformly distributed on the interval $(5, 95)$. Using a simulated sample of 10,000 observations, estimate the mean and variance of the aggregate loss S_N. In addition, use the simulated sample to estimate the probability that aggregate claims for this individual will exceed 2,000 and compare with the normal approximation estimates from Example 7.3.6.

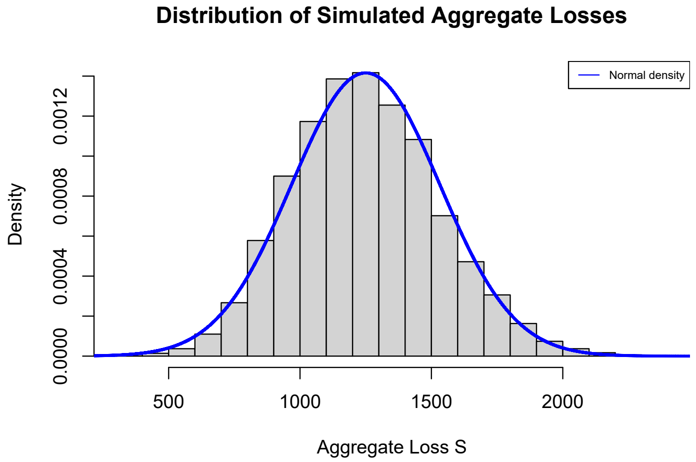
Solution. We follow the algorithm for the collective risk model, where we first simulate frequencies $n_1, \ldots, n_{10000}$, and conditional on $n_j,\ j = 1, \ldots, 10000$, simulate each individual loss $x_{ij},\ i = 1, \ldots n_j$.

```
set.seed(4321)  # For reproducibility of results
m <- 10000  # Number of observations to simulate
lambda <- 25  # Parameter for frequency distribution N
a <- 5
b <- 95  # Parameters for severity distribution X
S <- rep(NA, m)  # Initialize an empty vector to store S observations

n <- rpois(m, lambda)  # Generate m=10000 observations of N from Poisson
for (j in 1:m) {
    n_j <- n[j]  # Given each n_j (j=1,...,m), generate n_j observations of X from uniform
    x_j <- runif(n_j, min = a, max = b)
    s_j <- sum(x_j)  # Calculate the aggregate loss s_j
    S[j] <- s_j  # Store s_j in the vector of observations
}
# mean(S) # Compare to theoretical value of 1,250 var(S) # Compare to
# theoretical value of 79,375 mean(S>2000) # Proportion of simulated
# observations s_j that are > 2000 Compare to normal approximation method of
# 0.003884
```

Using simulation, we estimate the mean and variance of the aggregate claims to be approximately 1248 and 77441 respectively, compared to the theoretical values of 1,250 and 79,375. In addition, we estimate the probability that aggregate losses exceed 2000 to be 0.0062, compared to the normal approximation estimate of 0.003884.

We can assess the appropriateness of the normal approximation by comparing the empirical distribution of the simulated aggregate losses to the density of the normal distribution used for the normal approximation, $N(\mu = 1,250\ ,\ \sigma^2 = 79,375)$:



The simulated losses are slightly more right-skewed than the normal distribution, with a longer right tail. This explains why the normal approximation estimate of $\Pr(S_N > 2000)$ is lower than the simulated estimate.

7.5 Effects of Coverage Modifications

In this section, you learn to evaluate:

- the effect of exposure change on the aggregate claim count
- the effect of per-occurrence deductible on the claim frequency

- the effect of coverage modifications on the aggregate losses

---

7.5.1 Impact of Exposure on Frequency

This section focuses on an individual risk model for claim counts. Recall the individual risk model involves a fixed n number of contracts and independent loss random variables X_i. Consider the number of claims from a group of n policies:

$$S = X_1 + \cdots + X_n,$$

where we assume X_i are *iid* representing the number of claims from policy i. In this case, the exposure for the portfolio is n, using policy as exposure base. In Section 10.4 we will introduce other exposure bases. The *pgf* of S is

$$\begin{aligned} P_S(z) &= \mathrm{E}(z^S) = \mathrm{E}\left(z^{\sum_{i=1}^n X_i}\right) \\ &= \prod_{i=1}^n \mathrm{E}(z^{X_i}) = [P_X(z)]^n. \end{aligned}$$

Special Case: Poisson. If $X_i \sim Poi(\lambda)$, its *pgf* is $P_X(z) = e^{\lambda(z-1)}$. Then the *pgf* of S is

$$P_S(z) = [e^{\lambda(z-1)}]^n = e^{n\lambda(z-1)}.$$

So $S \sim Poi(n\lambda)$. That is, the sum of n independent Poisson random variables each with mean λ has a Poisson distribution with mean $n\lambda$.

Special Case: Negative Binomial. If $X_i \sim NB(\beta, r)$, its *pgf* is $P_X(z) = [1-\beta(z-1)]^{-r}$. Then the *pgf* of S is

$$P_S(z) = [[1-\beta(z-1)]^{-r}]^n = [1-\beta(z-1)]^{-nr}.$$

So $S \sim NB(\beta, nr)$.

Example 7.5.1. Assume that the number of claims for each vehicle is Poisson with mean λ. Given the following data on the observed number of claims for each household, calculate the MLE of λ.

Household ID	Number of vehicles	Number of claims
1	2	0
2	1	2
3	3	2
4	1	0
5	1	1

Example Solution. Each of the 5 households has number of exposures n_j (number of vehicles) and number of claims S_j, $j = 1, ..., 5$. Note for each household, the number of claims $S_j \sim Poi(n_j\lambda)$. The likelihood function is

$$\begin{aligned} L(\lambda) &= \prod_{j=1}^{5} \Pr(S_j = s_j) = \prod_{j=1}^{5} \frac{e^{-n_j\lambda}(n_j\lambda)^{s_j}}{s_j!} \\ &= \left(\frac{e^{-2\lambda}(2\lambda)^0}{0!}\right)\left(\frac{e^{-1\lambda}(1\lambda)^2}{2!}\right)\left(\frac{e^{-3\lambda}(3\lambda)^2}{2!}\right)\left(\frac{e^{-1\lambda}(1\lambda)^0}{0!}\right)\left(\frac{e^{-1\lambda}(1\lambda)^1}{1!}\right) \\ &\propto e^{-8\lambda}\lambda^5 \end{aligned}$$

Taking the logarithm, we have

$$l(\lambda) = \log L(\lambda) = -8\lambda + 5\log(\lambda).$$

Setting the first derivative of the log-likelihood to 0, we get $\hat{\lambda} = \frac{5}{8}$.

If the exposure of the portfolio changes from n_1 to n_2, we can establish the following relation between the aggregate claim counts:

$$P_{S_{n_2}}(z) = [P_X(z)]^{n_2} = [P_X(z)^{n_1}]^{n_2/n_1} = P_{S_{n_1}}(z)^{n_2/n_1}.$$

7.5.2 Impact of Deductibles on Claim Frequency

This section examines the effect of deductibles on claim frequency. Intuitively, there will be fewer claims filed when a policy deductible is imposed because a loss below the deductible level may not result in a claim. Even if an insured does file a claim, this may not result in a payment by the policy, since the claim may be denied or the loss amount may ultimately be determined to be below deductible. Let N^L denote the number of losses (i.e. the number of claims with no deductible), and N^P denote the number of payments when a deductible d is imposed. Our goal is to identify the distribution of N^P given the distribution of N^L. We show below that the relationship between N^L and N^P can be established within an aggregate risk model framework.

Note that sometimes changes in deductibles will affect policyholder claim behavior. We assume that this is not the case, i.e. the underlying distributions of losses for both frequency and severity remain unchanged when the deductible changes.

Given there are N^L losses, let $X_1, X_2 \ldots, X_{N^L}$ be the associated amount of losses. For $j = 1, \ldots, N^L$, define

$$I_j = \begin{cases} 1 & \text{if } X_j > d \\ 0 & \text{otherwise} \end{cases}.$$

Then we establish

$$N^P = I_1 + I_2 + \cdots + I_{N^L},$$

that is, the total number of payments is equal to the number of losses above the deductible level. Given that I_j's are independent Bernoulli random variables with probability of success $v = \Pr(X > d)$, the sum of a *fixed number* of such variables is then a binomial random variable. Thus, conditioning on N^L, N^P has a binomial distribution, i.e. $N^P|N^L \sim Bin(N^L, v)$, where $v = \Pr(X > d)$. This implies that

$$\mathrm{E}\left(z^{N^P}|N^L\right) = [1 + v(z-1)]^{N^L}$$

So the *pgf* of N^P is

$$\begin{aligned} P_{N^P}(z) &= \mathrm{E}_{N^P}\left(z^{N^P}\right) = \mathrm{E}_{N^L}\left[\mathrm{E}_{N^P}\left(z^{N^P}|N^L\right)\right] \\ &= \mathrm{E}_{N^L}\left[(1 + v(z-1))^{N^L}\right] \\ &= P_{N^L}\left(1 + v(z-1)\right). \end{aligned}$$

Thus, we can write the *pgf* of N^P as the *pgf* of N^L, evaluated at a new argument $z^* = 1 + v(z-1)$. That is, $P_{N^P}(z) = P_{N^L}(z^*)$.

Special Cases:

- $N^L \sim Poi(\lambda)$. The *pgf* of N^L is $P_{N^L} = e^{\lambda(z-1)}$. Thus the *pgf* of N^P is

$$\begin{aligned} P_{N^P}(z) &= e^{\lambda(1+v(z-1)-1)} \\ &= e^{\lambda v(z-1)}, \end{aligned}$$

  So $N^P \sim Poi(\lambda v)$. This means the number of payments has the same distribution as the number of losses, but with the expected number of payments equal to $\lambda v = \lambda \Pr(X > d)$.

- $N^L \sim NB(\beta, r)$. The *pgf* of N^L is $P_{N^L}(z) = [1 - \beta(z-1)]^{-r}$. Thus the *pgf* of N^P is

$$\begin{aligned} P_{N^P}(z) &= (1 - \beta(1 + v(z-1) - 1))^{-r} \\ &= (1 - \beta v(z-1))^{-r}, \end{aligned}$$

  So $N^P \sim NB(\beta v, r)$. This means the number of payments has the same distribution as the number of losses, but with parameters βv and r.

Example 7.5.2. Suppose that loss amounts $X_i \sim Pareto(\alpha = 4,\ \theta = 150)$. You are given that the loss frequency is $N^L \sim Poi(\lambda)$ and the payment frequency distribution is $N_1^P \sim Poi(0.4)$ at deductible level $d_1 = 30$. Find the distribution of the payment frequency N_2^P when the deductible level is $d_2 = 100$.

Example Solution. Because the loss frequency N^L is Poisson, we can relate the means of the loss distribution N^L and the first payment distribution N_1^P (under deductible $d_1 = 30$) through $0.4 = \lambda v_1$, where

$$v_1 = \Pr(X > 30) = \left(\frac{150}{30 + 150}\right)^4 = \left(\frac{5}{6}\right)^4$$
$$\Rightarrow \lambda = 0.4 \left(\frac{6}{5}\right)^4.$$

With this, we can assess the second payment distribution N_2^P (under deductible $d_2 = 100$) as being Poisson with mean $\lambda_2 = \lambda v_2$, where

$$v_2 = \Pr(X > 100) = \left(\frac{150}{100 + 150}\right)^4 = \left(\frac{3}{5}\right)^4$$
$$\Rightarrow \lambda_2 = \lambda v_2 = 0.4 \left(\frac{6}{5}\right)^4 \left(\frac{3}{5}\right)^4 = 0.1075.$$

Example 7.5.3. Follow-Up. Now suppose instead that the loss frequency is $N^L \sim NB(\beta,\ r)$ and for deductible $d_1 = 30$, the payment frequency N_1^P is negative binomial with mean 0.4. Find the mean of the payment frequency N_2^P for deductible $d_2 = 100$.

Example Solution. Because the loss frequency N^L is negative binomial, we can relate the parameter β of the N^L distribution and the parameter β_1 of the first payment distribution N_1^P using $\beta_1 = \beta v_1$, where

$$v_1 = \Pr(X > 30) = \left(\frac{5}{6}\right)^4$$

Thus, the mean of N_1^P and the mean of N^L are related via

$$0.4 = r\beta_1 = r\,(\beta v_1)$$
$$\Rightarrow r\beta = \frac{0.4}{v_1} = 0.4 \left(\frac{6}{5}\right)^4.$$

Note that $v_2 = \Pr(X > 100) = \left(\frac{3}{5}\right)^4$ as in the original example. Then the second payment frequency distribution under deductible $d_2 = 100$ is $N_2^P \sim NB(\beta v_2,\ r)$ with mean

$$r(\beta v_2) = (r\beta) v_2 = 0.4 \left(\frac{6}{5}\right)^4 \left(\frac{3}{5}\right)^4 = 0.1075.$$

Next, we examine the more general case where N^L is a zero-modified distribution. Recall that a zero-modified distribution can be defined in terms of an unmodified one (as was shown in Section 3.5.1). That is,

$$p_k^M = c\ p_k^0, \text{ for } k = 1, 2, 3, \ldots, \text{ with } c = \frac{1 - p_0^M}{1 - p_0^0},$$

where p_k^0 is the *pmf* of the unmodified distribution. In the case that $p_0^M = 0$, we call this a *zero-truncated* distribution, or ZT. For other arbitrary values of p_0^M, this is a zero-modified, or ZM, distribution. The *pgf* for the modified distribution is shown as

$$P^M(z) = 1 - c + c\ P^0(z),$$

expressed in terms of the *pgf* of the unmodified distribution, $P^0(z)$. When N^L follows a zero-modified distribution, the distribution of N^P is established using the same relation from earlier, $P_{N^P}(z) = P_{N^L}\left(1 + v(z-1)\right)$.

Special Cases:

- N^L is a ZM-Poisson random variable with parameters λ and p_0^M. The *pgf* of N^L is

  $$P_{N^L}(z) = 1 - \frac{1 - p_0^M}{1 - e^{-\lambda}} + \frac{1 - p_0^M}{1 - e^{-\lambda}} \left(e^{\lambda(z-1)}\right).$$

  Thus the *pgf* of N^P is

  $$P_{N^P}(z) = 1 - \frac{1 - p_0^M}{1 - e^{-\lambda}} + \frac{1 - p_0^M}{1 - e^{-\lambda}} \left(e^{\lambda v(z-1)}\right).$$

  So the number of payments is also a ZM-Poisson distribution with parameters λv and p_0^M. The probability at zero can be evaluated using $\Pr(N^P = 0) = P_{N^P}(0)$.

- N^L is a ZM-negative binomial random variable with parameters β, r, and p_0^M. The *pgf* of N^L is

  $$P_{N^L}(z) = 1 - \frac{1 - p_0^M}{1 - (1+\beta)^{-r}} + \frac{1 - p_0^M}{1 - (1+\beta)^{-r}} \left[1 - \beta\left(z - 1\right)\right]^{-r}.$$

  Thus the *pgf* of N^P is

  $$P_{N^P}(z) = 1 - \frac{1 - p_0^M}{1 - (1+\beta)^{-r}} + \frac{1 - p_0^M}{1 - (1+\beta)^{-r}} \left[1 - \beta v\left(z - 1\right)\right]^{-r}.$$

  So the number of payments is also a ZM-negative binomial distribution with parameters βv, r, and p_0^M. Similarly, the probability at zero can be evaluated using $\Pr(N^P = 0) = P_{N^P}(0)$.

Example 7.5.4. Aggregate losses are modeled as follows:
(i) The number of losses follows a zero-modified Poisson distribution with $\lambda = 3$ and $p_0^M = 0.5$.
(ii) The amount of each loss has a Burr distribution with $\alpha = 3, \theta = 50, \gamma = 1$.
(iii) There is a deductible of $d = 30$ on each loss.
(iv) The number of losses and the amounts of the losses are mutually independent.

Calculate $\mathrm{E}(N^P)$ and $\mathrm{Var}(N^P)$.

Example Solution. Since N^L follows a ZM-Poisson distribution with parameters λ and p_0^M, we know that N^P also follows a ZM-Poisson distribution, but with parameters λv and p_0^M, where

$$v = \Pr(X > 30) = \left(\frac{1}{1 + (30/50)}\right)^3 = 0.2441.$$

Thus, N^P follows a ZM-Poisson distribution with parameters $\lambda^* = \lambda v = 0.7324$ and $p_0^M = 0.5$. Finally,

$$\begin{aligned}
\mathrm{E}(N^P) &= (1 - p_0^M)\frac{\lambda^*}{1 - e^{-\lambda^*}} = 0.5\left(\frac{0.7324}{1 - e^{-0.7324}}\right)\\
&= 0.7053\\
\mathrm{Var}(N^P) &= (1 - p_0^M)\left(\frac{\lambda^*[1 - (\lambda^* + 1)e^{-\lambda^*}]}{(1 - e^{-\lambda^*})^2}\right) + p_0^M(1 - p_0^M)\left(\frac{\lambda^*}{1 - e^{-\lambda^*}}\right)^2\\
&= 0.5\left(\frac{0.7324(1 - 1.7324e^{-0.7324})}{(1 - e^{-0.7324})^2}\right) + 0.5^2\left(\frac{0.7324}{1 - e^{-0.7324}}\right)^2\\
&= 0.7244.
\end{aligned}$$

7.5.3 Impact of Policy Modifications on Aggregate Claims

In this section, we examine how a change in the deductible affects the aggregate payments from an insurance portfolio. We assume that the presence of policy limits (u), coinsurance (α), and inflation (r) have no effect on the underlying distribution of frequency of payments made by an insurer. As in the previous section, we further assume that deductible changes do not impact the underlying distributions of losses for both frequency and severity.

Recall the notation N^L for the number of losses. With ground-up loss amount X and policy deductible d, we use N^P for the number of payments (as defined in the previous section 7.5.2). Also, define the amount of payment on a per-loss basis as

$$Y^L = \begin{cases} 0 \,, & \text{if } X < \dfrac{d}{1+r} \\ \alpha[(1+r)X - d] \,, & \text{if } \dfrac{d}{1+r} \leq X < \dfrac{u}{1+r} \\ \alpha(u-d) \,, & \text{if } X \geq \dfrac{u}{1+r} \end{cases} ,$$

and the amount of payment on a per-payment basis as

$$Y^P = \begin{cases} \text{undefined} \,, & \text{if } X < \dfrac{d}{1+r} \\ \alpha[(1+r)X - d] \,, & \text{if } \dfrac{d}{1+r} \leq X < \dfrac{u}{1+r} \\ \alpha(u-d) \,, & \text{if } X \geq \dfrac{u}{1+r}. \end{cases} .$$

In the above, r, u, and α represent the inflation rate, policy limit, and coinsurance, respectively. Hence, aggregate costs (payment amounts) can be expressed either on a per loss or per payment basis:

$$\begin{aligned} S &= Y_1^L + \cdots + Y_{N^L}^L \\ &= Y_1^P + \cdots + Y_{N^P}^P \,. \end{aligned}$$

(Recall that when we introduced the per-loss and per-payment bases in Section 5.1, we used another letter Y to distinguish losses from insurance payments, or claims. At this point in our development, we use the letter X to reduce notation complexity.)

The fundamentals regarding collective risk models are ready to apply. For instance, we have:

$$\begin{aligned} \mathrm{E}(S) &= \mathrm{E}\left(N^L\right)\mathrm{E}\left(X^L\right) = \mathrm{E}\left(N^P\right)\mathrm{E}\left(Y^P\right) \\ \mathrm{Var}(S) &= \mathrm{E}\left(N^L\right)\mathrm{Var}\left(Y^L\right) + \left[\mathrm{E}\left(Y^L\right)\right]^2 \mathrm{Var}(N^L) \\ &= \mathrm{E}\left(N^P\right)\mathrm{Var}\left(Y^P\right) + \left[\mathrm{E}\left(Y^P\right)\right]^2 \mathrm{Var}(N^P) \\ M_S(z) &= P_{N^L}\left[M_{Y^L}(z)\right] = P_{N^P}\left[M_{Y^P}(z)\right]. \end{aligned}$$

Example 7.5.5. Actuarial Exam Question. A group dental policy has a negative binomial claim count distribution with mean 300 and variance 800.

Ground-up severity is given by the following table:

Severity	Probability
40	0.25
80	0.25
120	0.25
200	0.25

You expect severity to increase 50% with no change in frequency. You decide to impose a per claim deductible of 100. Calculate the expected total claim payment S after these changes.

Example Solution. The cost per loss with a 50% increase in severity and a 100 deductible per claim is

$$X^L = \begin{cases} 0 & 1.5x < 100 \\ 1.5x - 100 & 1.5x \geq 100 \end{cases}$$

This has expectation

$$\begin{aligned} \mathrm{E}(X^L) &= \frac{1}{4}\left[(1.5(40)-100)_+ + (1.5(80)-100)_+ + (1.5(120)-100)_+ + (1.5(200)-100)_+\right] \\ &= \frac{1}{4}\left[(60-100)_+ + (120-100)_+ + (180-100)_+ + (300-100)_+\right] \\ &= \frac{1}{4}\left[0+20+80+200\right] = 75. \end{aligned}$$

Thus, the expected aggregate loss is

$$\mathrm{E}(S) = \mathrm{E}(N)\ \mathrm{E}\left(X^L\right) = 300(75) = 22,500..$$

Example 7.5.6. Follow-Up. What is the variance of the total claim payment, Var (S)?

Example Solution. On a per loss basis, we have

$$\mathrm{Var}(S) = \mathrm{E}(N)\ \mathrm{Var}\left(X^L\right) + \left[\mathrm{E}\left(X^L\right)\right]^2\ \mathrm{Var}(N)$$

where $\mathrm{E}(N) = 300$ and $\mathrm{Var}(N) = 800$. We find

$$\begin{aligned} \mathrm{E}\left[(X^L)^2\right] &= \frac{1}{4}\left[0^2+20^2+80^2+200^2\right] = 11,700 \\ \Rightarrow \mathrm{Var}(X^L) = \mathrm{E}\left[(X^L)^2\right] - \left[\mathrm{E}(X^L)\right]^2 &= 11,700 - 75^2 = 6,075 \end{aligned}$$

Thus, the variance of the aggregate claim payment is

$$\mathrm{Var}(S) = 300(6,075) + 75^2(800) = 6,322,500.$$

Alternative Method: Using the Per Payment Basis. Previously, we calculated the expected total claim payment by multiplying the expected number of losses by the expected payment *per loss*. Recall that we can also multiply the expected number of payments by the expected payment *per payment*. In this case, we have

$$S = X_1^P + \cdots + X_{N_P}^P$$

The probability of a payment is

$$\Pr(1.5X \geq 100) = \Pr(X \geq 66.\bar{6}) = \frac{3}{4}.$$

Thus, the number of payments, N^P has a negative binomial distribution (see negative binomial special case in Section 7.5.2) with mean

$$\mathrm{E}(N^P) = \mathrm{E}(N^L)\ \Pr(1.5X \geq 100) = 300\left(\frac{3}{4}\right) = 225.$$

The cost per payment is

$$X^P \quad = \quad \begin{cases} \text{undefined}\ , & \text{if}\ \ 1.5x < 100 \\ 1.5x - 100\ , & \text{if}\ \ 1.5x \geq 100 \end{cases}$$

This has expectation

$$\mathrm{E}(X^P) = \frac{\mathrm{E}(X^L)}{\Pr(1.5X > 100)} = \frac{75}{(3/4)} = 100.$$

Thus, as before, the expected aggregate loss is

$$\mathrm{E}(S) = \mathrm{E}(X^P)\ \mathrm{E}(N^P) = 100(225) = 22,500.$$

Example 7.5.7. Actuarial Exam Question. A company insures a fleet of vehicles. Aggregate losses have a compound Poisson distribution. The expected number of losses is 20. Loss amounts, regardless of vehicle type, have exponential distribution with $\theta = 200$. To reduce the cost of the insurance, two modifications are to be made:
(i) A certain type of vehicle will not be insured. It is estimated that this will reduce loss frequency by 20%.
(ii) A deductible of 100 per loss will be imposed.

Calculate the expected aggregate amount paid by the insurer after the modifications.

Example Solution. On a per loss basis, we have a 100 deductible. Thus, the expectation per loss is

$$\begin{aligned}\mathrm{E}(X^L) &= E[(X-100)_+] = E(X) - E(X \wedge 100)\\ &= 200 - 200(1-e^{-100/200}) = 121.31.\end{aligned}$$

Loss frequency has been reduced by 20%, resulting in an expected number of losses

$$\mathrm{E}(N^L) = 0.8(20) = 16.$$

Thus, the expected aggregate amount paid after the modifications is

$$\mathrm{E}(S) = \mathrm{E}(X^L)\ \mathrm{E}(N^L) = 121.31(16) = 1,941.$$

Alternative Method: Using the Per Payment Basis. We can also use the per payment basis to find the expected aggregate amount paid after the modifications. With the deductible of 100, the probability that a payment occurs is $\Pr(X > 100) = e^{-100/200}$. For the per payment severity, plugging in the expression for $\mathrm{E}(X^L)$ from the original example, we have

$$\mathrm{E}(X^P) = \frac{\mathrm{E}(X^L)}{\Pr(X>100)} = \frac{200 - 200(1-e^{-100/200})}{e^{-100/200}} = 200$$

This is not surprising – recall that the exponential distribution is memoryless, so the expected claim amount paid in excess of 100 is still exponential with mean 200.

Now we look at the payment frequency

$$\mathrm{E}(N^P) = \mathrm{E}(N^L)\ \Pr(X>100) = 16\ e^{-100/200} = 9.7.$$

Putting this together, we produce the same answer using the per payment basis as the per loss basis from earlier

$$\mathrm{E}(S) = \mathrm{E}(X^P)\ \mathrm{E}(N^P) = 200(9.7) = 1,941.$$

7.6 Further Resources and Contributors

Contributors

- **Peng Shi** and **Lisa Gao**, University of Wisconsin-Madison, are the principal authors of the initial version of this chapter.

- **Peng Shi**, University of Wisconsin-Madison, is the author of the second edition of this chapter. Email: pshi@bus.wisc.edu for chapter comments and suggested improvements.
- Chapter reviewers include: Himchan Jeong, Vytaras Brazauskas, Mark Maxwell, Jiadong Ren, Sherly Paola Alfonso Sanchez, and Di (Cindy) Xu.

Further Readings and References

If you would like additional practice with `R` coding, please visit our companion LDA Short Course. In particular, see the Aggregate Loss Models Chapter.

TS 7.A.1. Individual Risk Model Properties

For the expected value of the aggregate loss under the individual risk model,

$$\begin{aligned}
\mathrm{E}(S_n) &= \sum_{i=1}^{n}\ \mathrm{E}(X_i) = \sum_{i=1}^{n}\ \mathrm{E}(I_i \times B_i) = \sum_{i=1}^{n}\ \mathrm{E}(I_i)\ \ \mathrm{E}(B_i) \quad \text{from independence of } I_i\text{'s and } B_i\text{'s} \\
&= \sum_{i=1}^{n} \Pr(I_i = 1)\ \mu_i \quad \text{since } E(I_i) \text{ is the probability } I_i \text{ is 1} \\
&= \sum_{i=1}^{n}\ q_i\ \mu_i .
\end{aligned}$$

For the variance of the aggregate loss under the individual risk model,

$$\begin{aligned}
\mathrm{Var}(S_n) &= \sum_{i=1}^{n} \mathrm{Var}(X_i) \quad \text{from independence of } X_i\text{'s} \\
&= \sum_{i=1}^{n}\ (\ \mathrm{E}\left[\mathrm{Var}(X_i|I_i)\right] + \mathrm{Var}\left[\mathrm{E}(X_i|I_i)\right]\ ) \quad \text{from conditional variance formulas} \\
&= \sum_{i=1}^{n} \left(q_i\ \sigma_i^2\ +\ q_i\ (1-q_i)\ \mu_i^2\right) .
\end{aligned}$$

To see this, note that

$$\begin{aligned}
\mathrm{E}\left[\mathrm{Var}(X_i|I_i)\right] &= \mathrm{Var}(X_i|I_i = 0)\ \ \Pr(I_i = 0) + \mathrm{Var}(X_i|I_i = 1)\ \ \Pr(I_i = 1) \\
&= q_i\ \sigma_i^2 + (1-q_i)\ (0) = q_i\ \sigma_i^2,
\end{aligned}$$

and

$$\mathrm{Var}\left[\mathrm{E}(X_i|I_i)\right] = q_i\ (1-q_i)\ \mu_i^2\ ,$$

using the Bernoulli variance shortcut since $\mathrm{E}(X_i|I_i) = 0$ when $I_i = 0$ (probability $\Pr(I_i = 0) = 1 - q_i$) and $\mathrm{E}(X_i|I_i) = \mu_i$ when $I_i = 1$ (probability $\Pr(I_i = 1) = q_i$).

For the probability generating function of the aggregate loss under the individual risk model,

$$
\begin{aligned}
P_{S_n}(z) &= \prod_{i=1}^{n} \ P_{X_i}(z) \quad \text{ from the independence of } X_i\text{'s} \\
&= \prod_{i=1}^{n} \ \mathrm{E}(z^{\ X_i}) = \prod_{i=1}^{n} \ \mathrm{E}(z^{\ I_i \times B_i}) = \prod_{i=1}^{n} \ \mathrm{E}\left[\mathrm{E}(z^{\ I_i \times B_i} | I_i)\right] \quad \text{ from law of iterated expectations} \\
&= \prod_{i=1}^{n} \left[\ E\left(z^{\ I_i \times B_i} | I_i = 0\right) \ \Pr(I_i = 0) + E\left(z^{\ I_i \times B_i} | I_i = 1\right) \ \Pr(I_i = 1) \ \right] \\
&= \prod_{i=1}^{n} \ [\ (1) \ (1 - q_i) + P_{B_i}(z) \ q_i \ ] = \prod_{i=1}^{n} (\ 1 - q_i + q_i \ P_{B_i}(z) \ )
\end{aligned}
$$

Lastly, for the moment generating function of the aggregate loss under the individual risk model,

$$
\begin{aligned}
M_{S_n}(t) &= \prod_{i=1}^{n} \ M_{X_i}(t) \quad \text{ from the independence of } X_i\text{'s} \\
&= \prod_{i=1}^{n} \ \mathrm{E}(e^{t \ X_i}) = \prod_{i=1}^{n} \ \mathrm{E}\left(e^{\ t \ (I_i \times B_i)}\right) \\
&= \prod_{i=1}^{n} \ \mathrm{E}\left[\mathrm{E}\left(e^{\ t \ (I_i \times B_i)} | I_i\right)\right] \quad \text{ from law of iterated expectations} \\
&= \prod_{i=1}^{n} \ \left[\ \mathrm{E}\left(e^{\ t \ (I_i \times B_i)} | I_i = 0\right) \ \Pr(I_i = 0) + \mathrm{E}\left(e^{\ t \ (I_i \times B_i)} | I_i = 1\right) \ \Pr(I_i = 1) \ \right] \\
&= \prod_{i=1}^{n} \ [\ (1) \ (1 - q_i) + M_{B_i}(t) \ q_i \ ] = \prod_{i=1}^{n} (\ 1 - q_i + q_i \ M_{B_i}(t) \ ) \ .
\end{aligned}
$$

TS 7.A.2. Relationship Between Probability Generating Functions of X_i and X_i^T

Let X_i belong to the $(a, b, 0)$ class with *pmf* $p_{ik} = \Pr(X_i = k)$ for $k = 0, 1, \ldots$ and X_i^T be the associated zero-truncated distribution in the $(a, b, 1)$ class with *pmf* $p_{ik}^T = p_{ik}/(1 - p_{i0})$ for $k = 1, 2, \ldots$. Then the relationship between the *pgf* of X_i and the *pgf* of X_i^T is shown by

$$
\begin{aligned}
P_{X_i}(z) &= \mathrm{E}\ (z^{X_i}) = \mathrm{E}\left[\mathrm{E}\left(z^{X_i} | X_i\right)\right] \quad \text{ from law of iterated expectations} \\
&= \mathrm{E}\left(z^{X_i} | X_i = 0\right) \ \Pr(X_i = 0) + \mathrm{E}\left(z^{X_i} | X_i > 0\right) \ \Pr(X_i > 0) \\
&= (1) \ p_{i0} + \mathrm{E}(z^{X_i^T}) \ (1 - p_{i0}) \quad \text{ since } (X_i | X_i > 0) \text{ is zero-truncated r.v. } X_i^T \\
&= p_{i0} + (1 - p_{i0}) P_{X_i^T}(z).
\end{aligned}
$$

TS 7.A.3. Moment Generating Function of Aggregate Loss S_N in Example 7.3.9

For $N \sim Geo(\beta)$ and $X \sim Exp(\theta)$, we have

$$\begin{aligned} P_N(z) &= \frac{1}{1-\beta(z-1)} \\ M_X(t) &= \frac{1}{1-\theta t}. \end{aligned}$$

Thus, the *mgf* of aggregate loss S_N is

$$\begin{aligned} M_{S_N}(t) = P_N[M_X(t)] &= \frac{1}{1-\beta\left(\frac{1}{1-\theta t}-1\right)} \\ &= \frac{1}{1-\beta\left(\frac{\theta t}{1-\theta t}\right)} + 1 - 1 = 1 + \frac{\beta\left(\frac{\theta t}{1-\theta t}\right)}{1-\beta\left(\frac{\theta t}{1-\theta t}\right)} \\ &= 1 + \frac{\beta\theta t}{(1-\theta t)-\beta\theta t} = 1 + \frac{\beta\theta t}{1-\theta t(1+\beta)} \cdot \frac{1+\beta}{1+\beta} \\ &= 1 + \frac{\beta}{1+\beta}\left[\frac{\theta(1+\beta)t}{1-\theta(1+\beta)t}\right] \\ &= 1 + \frac{\beta}{1+\beta}\left[\frac{1}{1-\theta(1+\beta)t} - 1\right], \end{aligned}$$

which gives the expression (7.1). For the alternate expression of the *mgf* (7.2), we continue from where we just left off:

$$\begin{aligned} M_{S_N}(t) &= 1 + \frac{\beta}{1+\beta}\left[\frac{\theta(1+\beta)t}{1-\theta(1+\beta)t}\right] \\ &= \frac{1+\beta}{1+\beta} + \frac{\beta}{1+\beta}\left[\frac{\theta(1+\beta)t}{1-\theta(1+\beta)t}\right] \\ &= \frac{1}{1+\beta} + \frac{\beta}{1+\beta} + \frac{\beta}{1+\beta}\left[\frac{\theta(1+\beta)t}{1-\theta(1+\beta)t}\right] \\ &= \frac{1}{1+\beta} + \frac{\beta}{1+\beta}\left[1 + \frac{\theta(1+\beta)t}{1-\theta(1+\beta)t}\right] \\ &= \frac{1}{1+\beta} + \frac{\beta}{1+\beta}\left[\frac{1}{1-\theta(1+\beta)t}\right]. \end{aligned}$$

8

Simulation and Resampling

Chapter Preview. Simulation is a computationally intensive method used to solve difficult problems. Instead of creating physical processes and experimenting with them in order to understand their operational characteristics, a simulation study is based on a computer representation - it considers various hypothetical conditions as inputs and summarizes the results. Through simulation, a vast number of hypothetical conditions can be quickly and inexpensively examined. Section 8.1 introduces simulation as a valuable computational tool, particularly effective in complex, multivariate settings.

Analysts find simulation especially useful for computing measures that summarize intricate distributions, as discussed in Section 8.2. This encompasses all the examples mentioned in the book thus far, such as measures that summarize the frequency and severity of losses, along with many additional cases. Simulation can also be used to compute complex distributions necessary for hypothesis testing. In addition, we can also use simulation to draw from an empirical distribution - this process is known as *resampling*. Resampling allows us to assess the uncertainty of estimates in complex models. Section 8.3 introduces resampling in the context of bootstrapping to determine the precision of estimators. Section 8.4 on cross-validation shows how to use it for model selection and validation.

8.1 Random Number Generation

In this section, you learn how to:

- Generate approximately independent realizations that are uniformly distributed
- Transform the uniformly distributed realizations to observations from a probability distribution of interest
- Generate simulated values directly from common distributions using ready-made random number generators

- Generate simulated values from complex distributions by combining simulated values from common distributions
- Generate simulated values from distributions whose domain is restricted to specific regions of interest, such as with deductible and long-tailed actuarial applications.

8.1.1 Generating Independent Uniform Observations

The simulations that we consider are generated by computers. A major strength of this approach is that they can be replicated, allowing us to check and improve our work. Naturally, this also means that they are not really random. Nonetheless, algorithms have been produced so that results appear to be random for all practical purposes. Specifically, they pass sophisticated tests of independence and can be designed so that they come from a single distribution - our iid assumption, identically and independently distributed.

To get a sense as to what these algorithms do, we consider a historically prominent method.

Linear Congruential Generator. To generate a sequence of random numbers, start with B_0, a starting value that is known as a *seed*. This value is updated using the recursive relationship

$$B_{n+1} = (aB_n + c) \text{ modulo } m, \quad n = 0, 1, 2, \ldots .$$

This algorithm is called a linear congruential generator. The case of $c = 0$ is called a *multiplicative* congruential generator; it is particularly useful for really fast computations.

For illustrative values of a and m, Microsoft's Visual Basic uses $m = 2^{24}$, $a = 1,140,671,485$, and $c = 12,820,163$ (see https://en.wikipedia.org/wiki/Linear_congruential_generator). This is the engine underlying the random number generation in Microsoft's Excel program.

The sequence used by the analyst is defined as $U_n = B_n/m$. The analyst may interpret the sequence $\{U_i\}$ to be (approximately) identically and independently uniformly distributed on the interval (0,1). To illustrate the algorithm, consider the following.

Example 8.1.1. Illustrative Sequence. Take $m = 15$, $a = 3$, $c = 2$ and $B_0 = 1$. Then we have:

step n	B_n		U_n
0	$B_0 = 1$		
1	$B_1 =$	$\mod (3 \times 1 + 2) = 5$	$U_1 = \frac{5}{15}$
2	$B_2 =$	$\mod (3 \times 5 + 2) = 2$	$U_2 = \frac{2}{15}$
3	$B_3 =$	$\mod (3 \times 2 + 2) = 8$	$U_3 = \frac{8}{15}$
4	$B_4 =$	$\mod (3 \times 8 + 2) = 11$	$U_4 = \frac{11}{15}$

The linear congruential generator is just one method of producing pseudo-random outcomes. It is easy to understand and is widely used. The linear congruential generator does have limitations, including the fact that it is possible to detect long-run patterns over time in the sequences generated (recall that we can interpret *independence* to mean a total lack of functional patterns). Not surprisingly, advanced techniques have been developed that address some of this method's drawbacks. The random number generated by `R` utilizes such advanced techniques.

Sometimes computer generated random results are known as pseudo-random numbers to reflect the fact that they are machine generated and can be replicated. That is, despite the fact that $\{U_i\}$ appears to be i.i.d, it can be reproduced by using the same seed number (and the same algorithm).

Example 8.1.2. Generating Uniform Random Numbers in `R`. The following code shows how to generate three uniform (0,1) numbers in `R` using the `runif` command. The `set.seed()` function sets the initial seed. In many computer packages, the initial seed is set using the system clock unless specified otherwise.

Three Uniform Random Variates

```
set.seed(2017)
U <- runif(3)
knitr::kable(U, digits = 5, align = "c", col.names = "Uniform") %>%
    kableExtra::kable_classic(full_width = F) %>%
    kable_styling(latex_options = "hold_position", font_size = 10)
```

Uniform
0.92424
0.53718
0.46920

8.1.2 Inverse Transform Method

With the sequence of uniform random numbers, we next transform them to a distribution of interest, say F. A prominent technique is the inverse transform method, defined as

$$X_i = F^{-1}\left(U_i\right).$$

Here, recall from Section 4.1.1 that we introduced the inverse of the distribution function, F^{-1}, and referred to it also as the quantile function. Specifically, it is defined to be

$$F^{-1}(y) = \inf_x ~ \{F(x) \geq y\}.$$

Recall that inf stands for *infimum* or the greatest lower bound. It is essentially the smallest value of x that satisfies the inequality $\{F(x) \geq y\}$. The result is that the sequence $\{X_i\}$ is *iid* with distribution function F if the $\{U_i\}$ are *iid* with uniform on $(0,1)$ distribution function.

The inverse transform result is available when the underlying random variable is continuous, discrete or a hybrid combination of the two. We now present a series of examples to illustrate its scope of applications.

Example 8.1.3. Generating Exponential Random Numbers. Suppose that we would like to generate observations from an exponential distribution with scale parameter θ so that $F(x) = 1 - e^{-x/\theta}$. To compute the inverse transform, we can use the following steps:

$$\begin{aligned} y = F(x) &\Leftrightarrow y = 1 - e^{-x/\theta} \\ &\Leftrightarrow -\theta \ln(1-y) = x = F^{-1}(y). \end{aligned}$$

Thus, if U has a uniform (0,1) distribution, then $X = -\theta \ln(1-U)$ has an exponential distribution with parameter θ.

The following `R` code shows how we can start with the same three uniform random numbers as in Example 8.1.2 and transform them to independent exponentially distributed random variables with a mean of 10. Alternatively, you can directly use the `rexp` function in `R` to generate random numbers from the exponential distribution. The algorithm built into this routine is different so even with the same starting seed number, individual realizations will differ.

```
set.seed(2017)
U <- runif(3)
X1 <- -10 * log(1 - U)
set.seed(2017)
X2 <- rexp(3, rate = 1/10)
```

Three Uniform Random Variates

Example 8.1.4. Generating Pareto Random Numbers. Suppose that we would like to generate observations from a Pareto distribution with parameters

Uniform	Exponential 1	Exponential 2
0.92424	25.80219	3.25222
0.53718	7.70409	8.47652
0.46920	6.33362	5.40176

α and θ so that $F(x) = 1 - \left(\frac{\theta}{x+\theta}\right)^{\alpha}$. To compute the inverse transform, we can use the following steps:

$$\begin{aligned} y = F(x) \Leftrightarrow 1 - y &= \left(\frac{\theta}{x+\theta}\right)^{\alpha} \\ \Leftrightarrow (1-y)^{-1/\alpha} &= \frac{x+\theta}{\theta} = \frac{x}{\theta} + 1 \\ \Leftrightarrow \theta\left((1-y)^{-1/\alpha} - 1\right) &= x = F^{-1}(y). \end{aligned}$$

Thus, $X = \theta\left((1-U)^{-1/\alpha} - 1\right)$ has a Pareto distribution with parameters α and θ.

Inverse Transform Justification. Why does the random variable $X = F^{-1}(U)$ have a distribution function F?

This is easy to establish when F is strictly increasing, where the distribution is continuous. Because U is a uniform random variable on (0,1), we know that $\Pr(U \le y) = y$, for $0 \le y \le 1$. Thus,

$$\begin{aligned} \Pr[X \le x] &= \Pr[F^{-1}(U) \le x] \\ &= \Pr[F(F^{-1}(U)) \le F(x)] \\ &= \Pr[U \le F(x)] = F(x), \end{aligned}$$

as required. The key step is that $F[F^{-1}(u)] = u$ for each u, which is true when F is strictly increasing.

We now consider some discrete examples.

Example 8.1.5. Generating Bernoulli Random Numbers. Suppose that we wish to simulate random variables from a Bernoulli distribution with parameter $q = 0.85$.

A graph of the cumulative distribution function in Figure 8.1 shows that the quantile function can be written as

$$F^{-1}(y) = \begin{cases} 0 & 0 < y \le 0.85 \\ 1 & 0.85 < y \le 1.0. \end{cases}$$

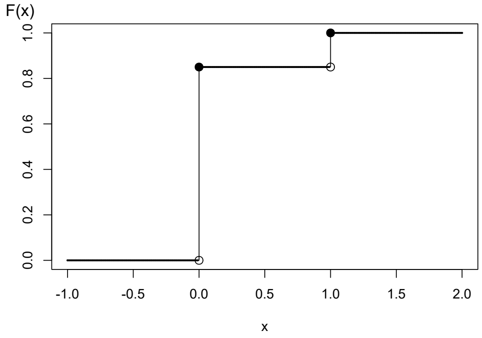


FIGURE 8.1: **Distribution Function of a Binary Random Variable**

Thus, with the inverse transform we may define

$$X = \begin{cases} 0 & 0 < U \le 0.85 \\ 1 & 0.85 < U \le 1.0 \end{cases}$$

For illustration, we generate three random numbers to get

```
set.seed(2017)
U <- runif(3)
X <- 1 * (U > 0.85)
```

Three Random Variates

Uniform	Binary X
0.92424	1
0.53718	0
0.46920	0

Example 8.1.6. Generating Random Numbers from a Discrete Distribution. Consider the time of a machine failure in the first five years. The distribution of failure times is given as:

Discrete Distribution

Time	1.0	2.0	3.0	4.0	5.0
Probability	0.1	0.2	0.1	0.4	0.2
Distribution Function	0.1	0.3	0.4	0.8	1.0

Using the graph of the distribution function in Figure 8.2, with the inverse

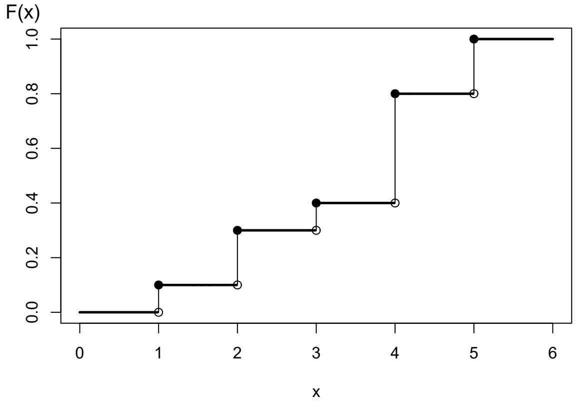


FIGURE 8.2: **Distribution Function of a Discrete Random Variable**

transform we may define

$$X = \begin{cases} 1 & 0 < U \leq 0.1 \\ 2 & 0.1 < U \leq 0.3 \\ 3 & 0.3 < U \leq 0.4 \\ 4 & 0.4 < U \leq 0.8 \\ 5 & 0.8 < U \leq 1.0. \end{cases}$$

For general discrete random variables there may not be an ordering of outcomes. For example, a person could own one of five types of life insurance products and we might use the following algorithm to generate random outcomes:

$$X = \begin{cases} \text{whole life} & 0 < U \leq 0.1 \\ \text{endowment} & 0.1 < U \leq 0.3 \\ \text{term life} & 0.3 < U \leq 0.4 \\ \text{universal life} & 0.4 < U \leq 0.8 \\ \text{variable life} & 0.8 < U \leq 1.0. \end{cases}$$

Another analyst may use an alternative procedure such as:

$$X = \begin{cases} \text{whole life} & 0.9 < U < 1.0 \\ \text{endowment} & 0.7 \leq U < 0.9 \\ \text{term life} & 0.6 \leq U < 0.7 \\ \text{universal life} & 0.2 \leq U < 0.6 \\ \text{variable life} & 0 \leq U < 0.2. \end{cases}$$

Both algorithms produce (in the long-run) the same probabilities, e.g., Pr(whole life) = 0.1, and so forth. So, neither is incorrect. You should be aware that there is more than one way to accomplish a goal. Similarly, you could use an alternative algorithm for ordered outcomes (such as failure times 1, 2, 3, 4, or 5, above).

8.1.3 Ready-made Random Number Generators

Sections 8.1.1 and 8.1.2 showed how one can generate simulated values from the foundations. This approach is important so that analyst can appreciate why the simulation works so well. However, because simulation is used so widely, it is not surprising that packages have been developed as time-saving devices.

For example, we have already seen in Example 8.1.3 that one can generate exponentially distributed random variates through the `rexp` function. This function means that analyst need not generate uniform random variates and then transform them using the inverse exponential distribution function. Instead, this is done in a single step using the `rexp` function.

Table 8.2 summarizes a few of the standard random number generators in `R`; the **r** at the beginning of each function refers to a **r**andom number generator. Additional documentation for these functions are in Appendix Chapter 20. Note that the Pareto distribution requires the package `actuar`.

Table 8.2. **Random Number Generators (RNGs)**

Discrete Distributions		**Continuous Distributions**	
Distribution	*RNG Function*	*Distribution*	*RNG Function*
Binomial	rbinom	Exponential	rexp
Poisson	rpoisson	Gamma	rgamma
Negative Binomial	rnbinom	Pareto	actuar::rpareto
		Normal	rnorm
		Weibull	rweibull

8.1.4 Simulating from Complex Distributions

In statistical software programs such as `R`, analysts will find several ready-made random number generators. However, for many complex actuarial applications, it is likely that ready-made generators will not be available and so one must return to the foundations.

To illustrate, consider the aggregate claims distributions introduced in Chapter 7. There, in Section 7.4.2, we have already seen how to simulate aggregate loss distributions. As we saw in [Example 7.4.2], the process is to first simulate the number of losses and then simulate individual losses.

As another example of a complex distribution, consider the following example.

Example 8.1.7. Generating Random Numbers from a Hybrid Distribution. Consider a random variable that is 0 with probability 70% and is exponentially distributed with parameter $\theta = 10,000$ with probability 30%. In an insurance application, this might correspond to a 70% chance of having no insurance claims and a 30% chance of a claim - if a claim occurs, then it is exponentially distributed. The distribution function, depicted in Figure 8.3, is given as

$$F(y) = \begin{cases} 0 & x < 0 \\ 1 - 0.3\exp(-x/10000) & x \geq 0. \end{cases}$$

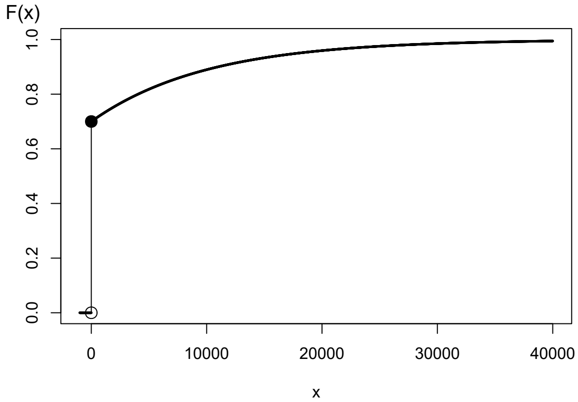


FIGURE 8.3: **Distribution Function of a Hybrid Random Variable**

From Figure 8.3, we can see that the inverse transform for generating random variables with this distribution function is

$$X = F^{-1}(U) = \begin{cases} 0 & 0 < U \leq 0.7 \\ -1000\ln(\frac{1-U}{0.3}) & 0.7 < U < 1. \end{cases}$$

For discrete and hybrid random variables, the key is to draw a graph of the distribution function that allows you to visualize potential values of the inverse function.

You can think of this hybrid distribution as a special case of a mixture model that was introduced in Sections 3.6 and 4.3.2. Mixture models are straightforward to evaluate using simulation. In the first stage, one simulates a variable indicating the subpopulation. In the second stage, one simulates from that subpopulation. The resulting variate is a realization from the mixture model. To illustrate, let's revisit Example 4.3.5.

Example 4.3.5. Continued. In this problem, we can label draws from the Type 1 subpopulation as X_1 from an exponential distribution with mean 200, and those from the Type 2 subpopulation as X_2 from a Pareto distribution with parameters $\alpha = 3$ and $\theta = 200$. Here, 25% of policies are Type 1 and 75% of policies are Type 2.

We can use simulation to find the probability that the annual loss will be less than 100, and find the average loss. The illustrative code uses the ready-made random number generator functions `rbinom`, `rexp`, and `actuar::pareto`.

```r
nsim <- 100000
Z <- rbinom(nsim, prob = 0.75, size = 1)
X1 <- rexp(nsim, rate = 1/200)
X2 <- actuar::rpareto(nsim, shape = 3, scale = 200)
X <- (1 - Z) * X1 + Z * X2
# sum(X<100)/nsim mean(X)
```

8.1.5 Importance Sampling

Another class of important problems utilize distributions that are from a limited region. For example, when a loss has a deductible, the resulting claim represents the payment by an insurer that is not observed for amounts less than the deductible. This type of problem was considered extensively in Chapter 5. As another example, for claims that are extremely large, one may wish to restrict an analysis to only extremely large outcomes - discussions of *tails* of distributions will be taken up in Section 13.2. To address both types of problems, we now suppose that we wish to draw according to X, conditional on $X \in [a, b]$.

To this end, one can use an accept-reject mechanism : draw x from distribution F

- if $x \in [a, b]$: keep it ("*accept*")
- if $x \notin [a, b]$: draw another one ("*reject*")

Observe that from n values initially generated, we keep here only $[F(b) - F(a)] \cdot n$ draws, on average.

Example 8.1.8. Draws from a Normal Distribution. Suppose that we draw from a normal distribution with mean 2.5 and variance 1, $N(2.5, 1)$, but are only interested in draws greater that $a = 2$ and less than $b = 4$. That is, we can only use $F(4) - F(2) = \Phi(4 - 2.5) - \Phi(2 - 2.5) = 0.9332$ - $0.3085 =$ 0.6247 proportion of the draws. Figure 8.4 demonstrates that some draws lie with the interval $(2, 4)$ and some are outside.

Instead, one can draw according to the conditional distribution $F^{\star}$ defined as

$$F^{\star}(x) = \Pr(X \le x | a < X \le b) = \frac{F(x) - F(a)}{F(b) - F(a)}, \quad \text{for } a < x \le b.$$

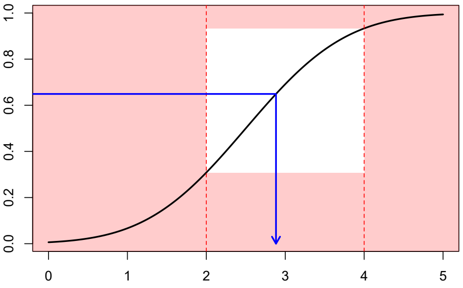


FIGURE 8.4: **Demonstration of Draws In and Outside of (2,4)**

Using the inverse transform method in Section 8.1.2, we have that the draw

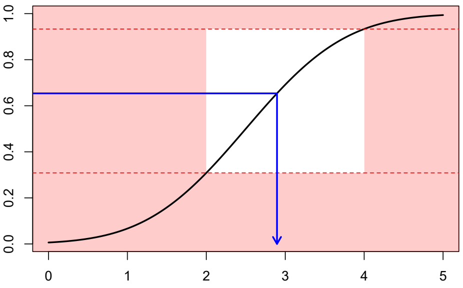
$$X^\star = F^{\star-1}(U) = F^{-1}(F(a) + U \cdot [F(b) - F(a)])$$

has distribution $F^\star$. Expressed another way, define

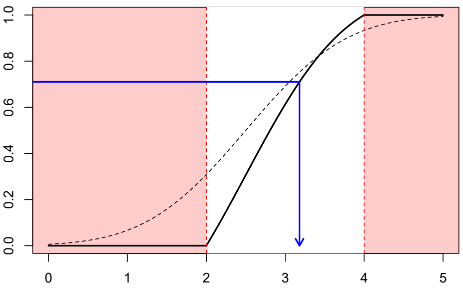
$$\tilde{U} = (1 - U) \cdot F(a) + U \cdot F(b)$$

and then use $F^{-1}(\tilde{U})$. With this approach, each draw counts.

This can be related to the importance sampling mechanism : we draw more frequently in regions where we expect to have quantities that have some interest. This transform can be considered as a "a change of measure."



In Example 8.1.8, the inverse of the normal distribution is readily available (in R, the function is `qnorm`). However, for other applications, this is not always the case. Then, one simply uses numerical methods to determine $X^\star$ as the solution of the equation $F(X^\star) = \tilde{U}$ where $\tilde{U} = (1 - U) \cdot F(a) + U \cdot F(b)$.



8.2 Computing Distribution Parameters

In this section, you learn how to:

- Calculate quantities of interest and determine the precision of the calculated quantities
- Determine the appropriate number of replications for a simulation study
- Calculate complex distributions needed for hypothesis testing.

8.2.1 Simulating Parameters

One use of the term **parameter** is as a quantity that serves as an index for a known parametric family. For example, one usually thinks of a mean μ and standard deviation σ as parameters of a normal distribution. Statisticians also use the term *parameter* to mean any quantity that summarizes a distribution. In this sense, a parameter can be written as $\theta(F)$, that is, if one knows the distribution function $F(\cdot)$, then one can compute the summary measure θ.

In the previous subsection, we learned how to generate independent simulated realizations from a distribution of interest. With these realizations, we can construct an empirical distribution and approximate the underlying distribution as precisely as needed. As we introduce more actuarial applications in this book, you will see that simulation can be applied in a wide variety of contexts.

Many of these applications can be reduced to the problem of approximating a parameter $\mathrm{E}\left[h(X)\right]$, where $h(\cdot)$ is some known function. Based on R simulations (replications), we get $X_1, \ldots, X_R$. From this simulated sample, we calculate an average

$$\overline{h}_R = \frac{1}{R}\sum_{i=1}^{R} h(X_i)$$

that we use as our simulated approximate (estimate) of $\mathrm{E}\left[h(X)\right]$. To estimate the precision of this approximation, we use the simulation variance

$$s^2_{h,R} = \frac{1}{R-1}\sum_{i=1}^{R}\left(h(X_i) - \overline{h}_R\right)^2 .$$

From the independence, the standard error of the estimate is $s_{h,R}/\sqrt{R}$. This can be made as small as we like by increasing the number of replications R.

Example 8.2.1. Portfolio Management. In Section 5.1, we learned how to calculate the expected value of policies with deductibles. For an example of something that cannot be done with closed form expressions, we now consider two risks. This is a variation of a more complex example that will be covered as Example 13.4.6.

We consider two property risks of a telecommunications firm:

- X_1 - buildings, modeled using a gamma distribution with mean 200 and scale parameter 100.
- X_2 - motor vehicles, modeled using a gamma distribution with mean 400 and scale parameter 200.

Denote the total risk as $X = X_1 + X_2$. For simplicity, you assume that these risks are independent.

To manage the risk, you seek some insurance protection. You are willing to retain internally small building and motor vehicles amounts, up to M, say. Random amounts in excess of M will have an unpredictable affect on your budget and so for these amounts you seek insurance protection. Stated mathematically, your retained risk is $Y_{retained} = \min(X_1 + X_2, M)$ and the insurer's portion is $Y_{insurer} = X - Y_{retained}$.

To be specific, we use $M = 400$ as well as $R = 1000000$ simulations.

a. With these settings, we wish to determine the expected claim amount and the associated standard deviation of (i) that retained by you, (ii) that accepted by the insurer, and (iii) the total overall amount. **b.** For insured claims, the standard error of the simulation approximation is $s_{h,R}/\sqrt{1000000} = 280.86$ $/\sqrt{1000000} = 0.281$. For this example, simulation is quick and so a large value such as 1000000 is an easy choice. However, for more complex problems, the simulation size may be an issue.

Example Solution. For part (a), the results of these calculations are:

```
                   Retained Insurer  Total
Mean                 365.17  235.01 600.18
Standard Deviation    69.51  280.86 316.36
```

For part (b), Figure 8.5 allows us to visualize the development of the approximation as the number of simulations increases.

You can learn more about the R code for this example at the online version of this book, Actuarial Community (2025).

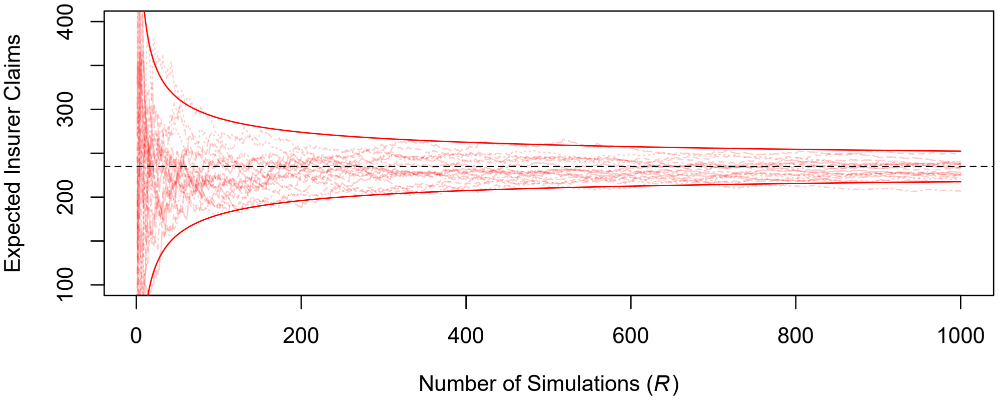


FIGURE 8.5: **Estimated Expected Insurer Claims versus Number of Simulations**

8.2.2 Determining the Number of Simulations

How many simulated values are recommended? 100? 1,000,000? We can use the central limit theorem to respond to this question.

As one criterion for your confidence in the result, suppose that you wish to be within 1% of the mean with 95% certainty. That is, you want $\Pr\left(|\overline{h}_R - \mathrm{E}\left[h(X)\right]| \le 0.01\mathrm{E}\left[h(X)\right]\right) \ge 0.95$. According to the central limit theorem, your estimate should be approximately normally distributed and so we want to have R large enough to satisfy $0.01\mathrm{E}\left[h(X)\right]/\sqrt{\mathrm{Var}\left[h(X)\right]/R}) \ge 1.96$. (Recall that 1.96 is the 97.5th percentile from the standard normal distribution.) Replacing $\mathrm{E}\left[h(X)\right]$ and $\mathrm{Var}\left[h(X)\right]$ with estimates, you continue your simulation until

$$\frac{0.01~\overline{h}_R}{s_{h,R}/\sqrt{R}} \ge 1.96$$

or equivalently

$$R \geq 38,416 \frac{s_{h,R}^2}{\overline{h}_R^2}. \tag{8.1}$$

This criterion is a direct application of the approximate normality. Note that $\overline{h}_R$ and $s_{h,R}$ are not known in advance, so you will have to come up with estimates, either by doing a small pilot study in advance or by interrupting your procedure intermittently to see if the criterion is satisfied.

Example 8.2.1. Portfolio Management - continued. For this example, the average insurance claim is 235.011 and the corresponding standard deviation is 280.862. Using equation (8.1), to be within 1% of the mean, we would only require at least 54.87 thousand simulations. In addition, to be within 0.1% we would want at least 5.49 million simulations.

Example 8.2.2. Approximation Choices. An important application of simulation is the approximation of $\mathrm{E}\left[h(X)\right]$. In this example, we show that the choice of the $h(\cdot)$ function and the distribution of X can play a role.

Consider the following question : what is $\Pr[X > 2]$ when X has a Cauchy distribution, with density $f(x) = \left[\pi(1+x^2)\right]^{-1}$, on the real line? The true value is

$$\Pr\left[X > 2\right] = \int_2^\infty \frac{dx}{\pi(1+x^2)}.$$

One can use an `R` numerical integration function (which usually works well on improper integrals)

which is equal to 0.14758.

Approximation 1. Alternatively, one can use simulation techniques to approximate that quantity. From calculus, you can check that the quantile function of the Cauchy distribution is $F^{-1}(y) = \tan\left[\pi(y-0.5)\right]$. Then, with simulated uniform (0,1) variates, $U_1, \ldots, U_R$, we can construct the estimator

$$p_1 = \frac{1}{R}\sum_{i=1}^R \mathrm{I}(F^{-1}(U_i) > 2) = \frac{1}{R}\sum_{i=1}^R \mathrm{I}(\tan\left[\pi(U_i - 0.5)\right] > 2).$$

With one million simulations, we obtain an estimate of 0.14744 with standard error 0.355 (divided by 1000). The estimated variance of p_1 can be written as $0.127/R$.

Approximation 2. With other choices of $h(\cdot)$ and $F(\cdot)$ it is possible to reduce uncertainty even using the same number of simulations R. To begin, one can use the symmetry of the Cauchy distribution to write $\Pr[X > 2] = 0.5 \cdot \Pr[|X| > 2]$. With this, can construct a new estimator,

$$p_2 = \frac{1}{2R} \sum_{i=1}^{R} \mathrm{I}(|F^{-1}(U_i)| > 2).$$

With one million simulations, we obtain an estimate of 0.14748 with standard error 0.228 (divided by 1000). The estimated variance of p_2 can be written as $0.052/R$.

Approximation 3. But one can go one step further. The improper integral can be written as a proper one by a simple symmetry property (since the function is symmetric and the integral on the real line is equal to 1)

$$\int_2^\infty \frac{dx}{\pi(1+x^2)} = \frac{1}{2} - \int_0^2 \frac{dx}{\pi(1+x^2)}.$$

From this expression, a natural approximation would be

$$p_3 = \frac{1}{2} - \frac{1}{R} \sum_{i=1}^{R} h_3(2U_i), \qquad \text{where } h_3(x) = \frac{2}{\pi(1+x^2)}.$$

With one million simulations, we obtain an estimate of 0.14756 with standard error 0.169 (divided by 1000). The estimated variance of p_3 can be written as $0.0285/R$.

Approximation 4. Finally, one can also consider some change of variable in the integral

$$\int_2^\infty \frac{dx}{\pi(1+x^2)} = \int_0^{1/2} \frac{y^{-2}dy}{\pi(1-y^{-2})}.$$

From this expression, a natural approximation would be

$$p_4 = \frac{1}{R} \sum_{i=1}^{R} h_4(U_i/2), \qquad \text{where } h_4(x) = \frac{1}{2\pi(1+x^2)}.$$

The expression seems rather similar to the previous one.

With one million simulations, we obtain an estimate of 0.14759 with standard error 0.01 (divided by 1000). The estimated variance of p_4 can be written as $0.00009/R$, which is much smaller than what we had so far!

Table 8.1 summarizes the four choices of $h(\cdot)$ and $F(\cdot)$ to approximate $\Pr[X > 2] = 0.14758$. The standard error varies dramatically. Thus, if we have a desired degree of accuracy, then the *number of simulations* depends strongly on how we write the integrals we try to approximate.

Table 8.1. **Summary of Four Choices to Approximate** $\Pr[X > 2]$

Estimator	Definition	Support Function	Estimate	Standard Error
p_1	$\frac{1}{R}\sum_{i=1}^{R} \mathrm{I}(F^{-1}(U_i) > 2)$	$F^{-1}(u) = \tan\left(\pi(u-0.5)\right)$	0.147439	0.000355
p_2	$\frac{1}{2R}\sum_{i=1}^{R} \mathrm{I}(\lvert F^{-1}(U_i)\rvert > 2)$	$F^{-1}(u) = \tan\left(\pi(u-0.5)\right)$	0.147477	0.000228
p_3	$\frac{1}{2} - \frac{1}{R}\sum_{i=1}^{R} h_3(2U_i)$	$h_3(x) = \frac{2}{\pi(1+x^2)}$	0.147558	0.000169
p_4	$\frac{1}{R}\sum_{i=1}^{R} h_4(U_i/2)$	$h_4(x) = \frac{1}{2\pi(1+x^2)}$	0.147587	0.000010

8.2.3 Simulation and Statistical Inference

Simulations not only help us approximate expected values but are also useful in calculating other aspects of distribution functions. As described in Section 8.2.1, the logic is that one wishes to calculate a parameter $\theta(F)$, use the same rule for calculating the parameter but replace the distribution function $F(\cdot)$ with an empirical one from a simulated sample. For example, in addition to expected values, analysts can use simulation to compute quantiles from complex distributions.

In addition, simulation is very useful when distributions of test statistics are too complicated to derive; in this case, one can use simulations to approximate the reference distribution. We now illustrate this with the Kolmogorov-Smirnov test which we learned about in Section 6.1.2.

Example 8.2.3. Kolmogorov-Smirnov Test of Distribution. Suppose that we have available $n = 100$ observations $\{x_1, \cdots, x_n\}$ that, unknown to the analyst, were generated from a gamma distribution with parameters $\alpha = 6$ and $\theta = 2$. The analyst believes that the data come from a lognormal distribution with parameters 1 and 0.4 and would like to test this assumption.

The first step is to visualize the data.

With this set-up, Figure 8.6 provides a graph of a histogram and empirical distribution. For reference, superimposed are red dashed lines from the lognormal distribution.

Recall that the Kolmogorov-Smirnov statistic equals the largest discrepancy between the empirical and the hypothesized distribution. This is $\max_x |F_n(x) - F_0(x)|$, where F_0 is the hypothesized lognormal distribution. We can calculate this directly.

Fortunately, for the lognormal distribution, R has built-in tests that allow us to determine this without complex programming:

```
ks.test(x, plnorm, mean = 1, sd = 0.4)
```

```

	Asymptotic one-sample Kolmogorov-Smirnov test

data:  x
```

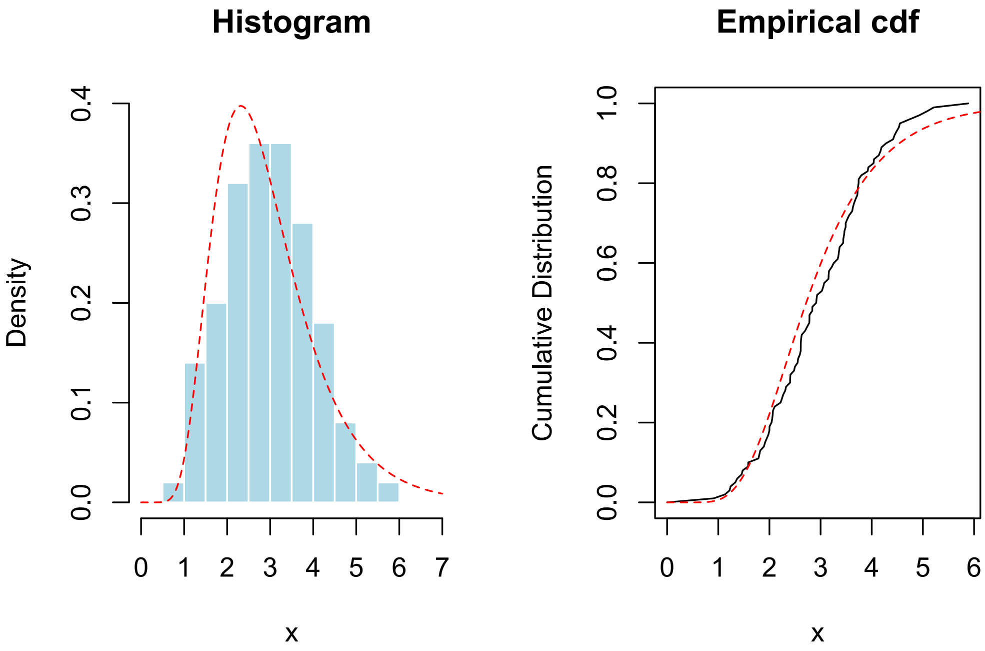


FIGURE 8.6: **Histogram and Empirical Distribution Function of Data used in Kolmogorov-Smirnov Test**. The red dashed lines are fits based on (incorrectly) hypothesized lognormal distribution.

```
D = 0.09703666, p-value = 0.303148
alternative hypothesis: two-sided
```

However, for many distributions of actuarial interest, pre-built programs are not available. We can use simulation to test the relevance of the test statistic. Specifically, to compute the p-value, let us generate thousands of random samples from a $LN(1, 0.4)$ distribution (with the same size), and compute empirically the distribution of the statistic,

```
ns <- 10000
d_KS <- rep(NA, ns)
# compute the test statistics for a large (ns) number of simulated samples
for (s in 1:ns) d_KS[s] <- D(rlnorm(n, 1, 0.4), function(x) plnorm(x, 1, 0.4))
mean(d_KS > D(x, function(x) plnorm(x, 1, 0.4)))
```

```
[1] 0.2843
```

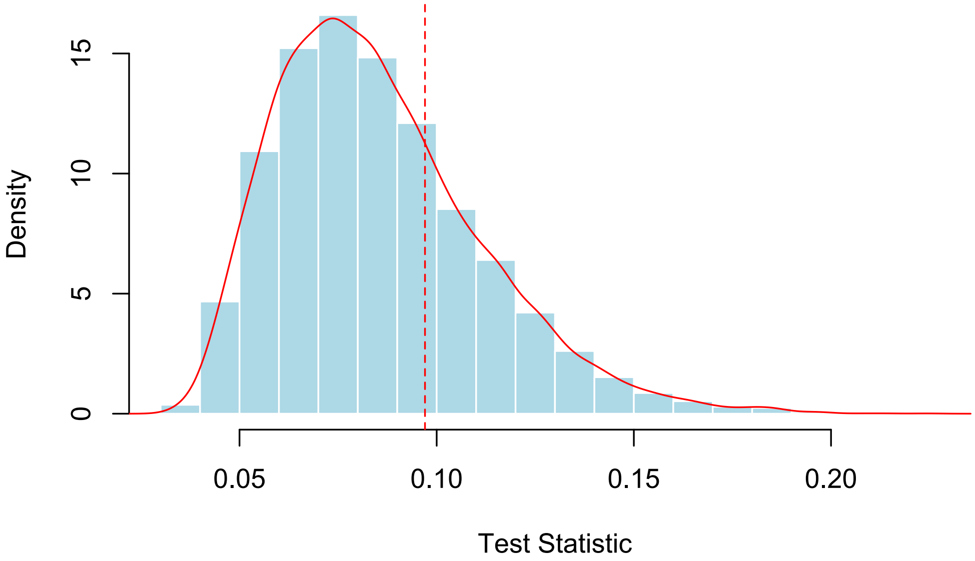


FIGURE 8.7: **Simulated Distribution of the Kolmogorov-Smirnov Test Statistic**. The vertical red dashed line marks the test statistic for the sample of 100.

The simulated distribution based on 10,000 random samples is summarized in Figure 8.7. Here, the statistic exceeded the empirical value (0.09704) in 28.43% of the scenarios, while the *theoretical* p-value is 0.3031. For both the simulation and the theoretical p-values, the conclusions are the same; the data do not provide sufficient evidence to reject the hypothesis of a lognormal distribution.

Although only an approximation, the simulation approach works in a variety of distributions and test statistics without needing to develop the nuances of the underpinning theory for each situation. We summarize the procedure for developing simulated distributions and p-values as follows:

1. Draw a sample of size n, say, $X_1, \ldots, X_n$, from a known distribution function F. Compute a statistic of interest, denoted as $\hat{\theta}(X_1, \ldots, X_n)$. Call this $\hat{\theta}^r$ for the rth replication.
2. Repeat this $r = 1, \ldots, R$ times to get a sample of statistics, $\hat{\theta}^1, \ldots, \hat{\theta}^R$.
3. From the sample of statistics in Step 2, $\{\hat{\theta}^1, \ldots, \hat{\theta}^R\}$, compute a summary measure of interest, such as a p-value.

8.3 Bootstrapping and Resampling

In this section, you learn how to:

- Generate a nonparametric bootstrap distribution for a statistic of interest
- Use the bootstrap distribution to generate estimates of precision for the statistic of interest, including bias, standard deviations, and confidence intervals
- Perform bootstrap analyses for parametric distributions

8.3.1 Bootstrap Foundations

Simulation presented up to now is based on sampling from a **known** distribution. Section 8.1 showed how to use simulation techniques to sample and compute quantities from known distributions. However, statistical science is dedicated to providing inferences about distributions that are *unknown*. We gather summary statistics based on this unknown population distribution. But how do we sample from an unknown distribution?

Naturally, we cannot simulate draws from an unknown distribution but we can draw from a sample of observations. If the sample is a good representation from the population, then our simulated draws from the sample should well approximate the simulated draws from a population. The process of sampling from a sample is called *resampling* or *bootstrapping*. The term bootstrap comes from the phrase "pulling oneself up by one's bootstraps" Efron (1979). With resampling, the original sample plays the role of the population and estimates from the sample play the role of true population parameters.

The resampling algorithm is the same as introduced in Section 8.2.3 except that now we use simulated draws from a sample. It is common to use $\{X_1, \ldots, X_n\}$ to denote the original sample and let $\{X_1^*, \ldots, X_n^*\}$ denote the simulated draws. We draw them with replacement so that the simulated draws will be independent from one another, the same assumption as with the original sample. For each sample, we also use n simulated draws, the same number as the original sample size. To distinguish this procedure from the simulation, it is common to use B (for bootstrap) to be the number of simulated samples. We could also write $\{X_1^{(b)}, \ldots, X_n^{(b)}\}$, $b = 1, \ldots, B$ to clarify this.

There are two basic resampling methods, *model-free* and *model-based*, which are, respectively, as *nonparametric* and *parametric*. In the nonparametric approach, no assumption is made about the distribution of the parent population. The simulated draws come from the empirical distribution function $F_n(\cdot)$, so each draw comes from $\{X_1, \ldots, X_n\}$ with probability $1/n$.

In contrast, for the parametric approach, we assume that we have knowledge of the distribution family F. The original sample $X_1, \ldots, X_n$ is used to estimate parameters of that family, say, $\hat{\theta}$. Then, simulated draws are taken from the $F(\hat{\theta})$. Section 8.3.4 discusses this approach in further detail.

Nonparametric Bootstrap

The idea of the nonparametric bootstrap is to use the inverse transform method on F_n, the empirical cumulative distribution function, depicted in Figure 8.8.

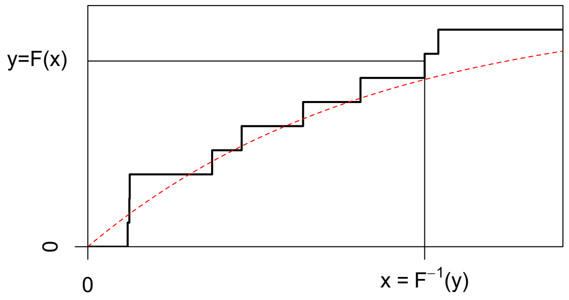


FIGURE 8.8: **Inverse of an Empirical Distribution Function**

Because F_n is a step-function, F_n^{-1} takes values in $\{x_1, \cdots, x_n\}$. More precisely, as illustrated in Figure 8.9.

- if $y \in (0, 1/n)$ (with probability $1/n$) we draw the smallest value ($\min\{x_i\}$)
- if $y \in (1/n, 2/n)$ (with probability $1/n$) we draw the second smallest value,
- $\vdots \; \vdots \; \vdots$
- if $y \in ((n-1)/n, 1)$ (with probability $1/n$) we draw the largest value ($\max\{x_i\}$).

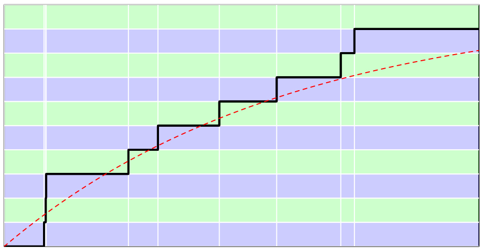


FIGURE 8.9: **Inverse of an Empirical Distribution Function**

Using the inverse transform method with F_n means sampling from $\{x_1, \cdots, x_n\}$, with probability $1/n$. Generating a bootstrap sample of size B means sampling from $\{x_1, \cdots, x_n\}$, with probability $1/n$, with replacement. See the following illustrative R code.

```
set.seed(1)
n <- 10
x <- rexp(n, 1/6)
m <- 10
bootvalues <- sample(x, size = m, replace = TRUE)
```

```
 [1] 2.6164 5.7394 5.7394 2.6164 2.6164 7.0899 0.8823 5.7394 4.5311 0.8388
```

Observe that value 5.7394 was obtained three times.

8.3.2 Bootstrap Precision: Bias, Standard Deviation, and Mean Square Error

We summarize the nonparametric bootstrap procedure as follows:

1. From the sample $\{X_1, \ldots, X_n\}$, draw a sample of size n (with re-

placement), say, $X_1^*, \ldots, X_n^*$. From the simulated draws compute a statistic of interest, denoted as $\hat{\theta}(X_1^*, \ldots, X_n^*)$. Call this $\hat{\theta}_b^*$ for the bth replicate.

2. Repeat this $b = 1, \ldots, B$ times to get a sample of statistics, $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.
3. From the sample of statistics in Step 2, $\{\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\}$, compute a summary measure of interest.

In this section, we focus on three summary measures, the bias, the standard deviation, and the mean square error (MSE). Table 8.3 summarizes these three measures. Here, $\overline{\hat{\theta}^*}$ is the average of $\{\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\}$.

Table 8.3. **Bootstrap Summary Measures**

Population Measure	*Population Definition*	*Bootstrap Approximation*	*Bootstrap Symbol*
Bias	$\mathrm{E}(\hat{\theta}) - \theta$	$\overline{\hat{\theta}^*} - \hat{\theta}$	$Bias_{boot}(\hat{\theta})$
Standard Deviation	$\sqrt{\mathrm{Var}(\hat{\theta})}$	$\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \overline{\hat{\theta}^*}\right)^2}$	$s_{boot}(\hat{\theta})$
Mean Square Error	$\mathrm{E}(\hat{\theta} - \theta)^2$	$\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \hat{\theta}\right)^2$	$MSE_{boot}(\hat{\theta})$

Example 8.3.1. Bodily Injury Claims and Loss Elimination Ratios. To show how the bootstrap can be used to quantify the precision of estimators, we return to the Example 5.3.2 bodily injury claims data where we introduced a nonparametric estimator of the loss elimination ratio.

Table 8.4 summarizes the results of the bootstrap estimation. For example, at $d = 14000$, the nonparametric estimate of LER is 0.97678. This has an estimated bias of 0.00016 with a standard deviation of 0.00687. For some applications, you may wish to apply the estimated bias to the original estimate to give a bias-corrected estimator. This is the focus of the next example. For this illustration, the bias is small and so such a correction is not relevant.

Table 8.4. **Bootstrap Estimates of LER at Selected Deductibles**

d	NP	Bootstrap	Bootstrap	Lower Normal	Upper Normal
	Estimate	Bias	SD	95% CI	95% CI
4000	0.54113	0.00011	0.01237	0.51678	0.56527
5000	0.64960	0.00027	0.01412	0.62166	0.67700
10500	0.93563	0.00004	0.01017	0.91567	0.95553
11500	0.95281	-0.00003	0.00941	0.93439	0.97128
14000	0.97678	0.00016	0.00687	0.96316	0.99008
18500	0.99382	0.00014	0.00331	0.98719	1.00017

The bootstrap standard deviation gives a measure of precision. For one application of standard deviations, we can use the normal approximation to create a confidence interval. For example, the R function `boot.ci` produces the normal confidence intervals at 95%. These are produced by creating an interval of twice the length of 1.95994 bootstrap standard deviations, centered about the bias-corrected estimator (1.95994 is the 97.5th quantile of the standard normal distribution). For example, the lower normal 95% CI at $d = 14000$ is $(0.97678 - 0.00016) - 1.95994 \times 0.00687 = 0.96316$. We further discuss bootstrap confidence intervals in the next section.

Example 8.3.2. Estimating $\log(\mu)$. The bootstrap can be used to quantify the bias of an estimator, for instance. Consider here a sample $\mathbf{x} = \{x_1, \cdots, x_n\}$ that is iid with mean μ.

```
sample_x <- c(2.46, 2.8, 3.28, 3.86, 2.85, 3.67, 3.37, 3.4, 5.22, 2.55, 2.79, 4.5,
    3.37, 2.88, 1.44, 2.56, 2, 2.07, 2.19, 1.77)
```

Suppose that the quantity of interest is $\theta = \log(\mu)$. A natural estimator would be $\widehat{\theta}_1 = \log(\overline{x})$. This estimator is biased (due to the Jensen inequality) but is asymptotically unbiased. For our sample, the estimate is as follows.

```
(theta_1 <- log(mean(sample_x)))
```

```
[1] 1.08231352
```

One can use a bootstrap strategy to get a correction: given a bootstrap sample, $\mathbf{x}_b^*$, let $\overline{x}_b^*$ denote its mean, and set

$$\widehat{\theta}_2 = \frac{1}{B}\sum_{b=1}^{B} \log(\overline{x}_b^*).$$

To implement this, we have the following code where we now use the function `boot` from the R package `boot`.

```
library(boot)
results <- boot(data = sample_x, statistic = function(y, indices) {
    log(mean(y[indices]))
}, R = 1000)
theta_2 <- 2 * theta_1 - mean(results$t)
```

Then, you can `plot(results)` and `print(results)` to see the following.

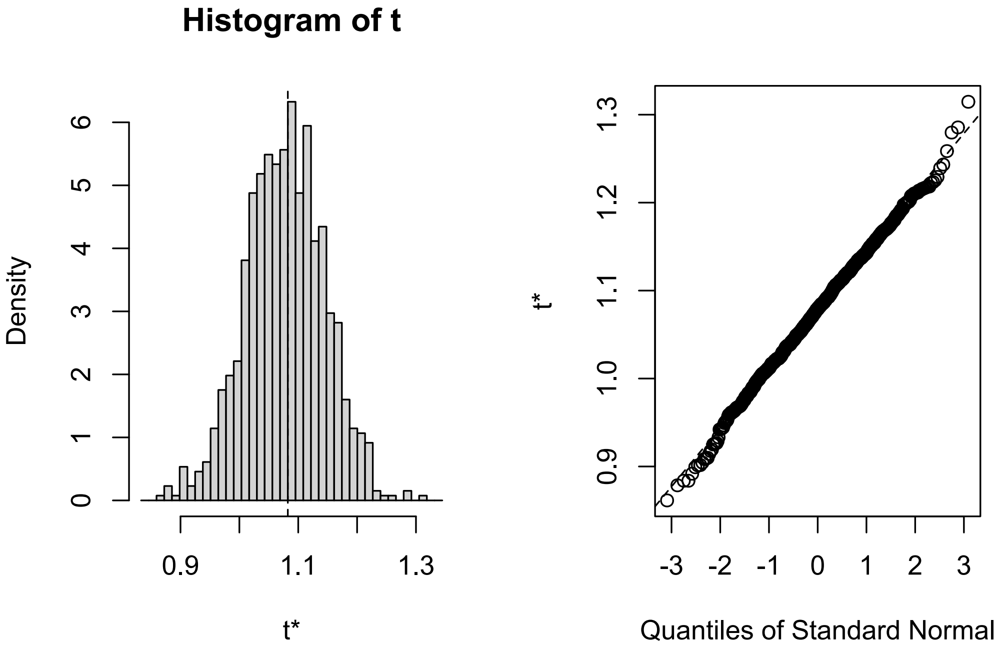


FIGURE 8.10: **Distribution of Bootstrap Replicates**. The left-hand panel is a histogram of replicates. The right-hand panel is a quantile-quantile plot, comparing the bootstrap distribution to the standard normal distribution.

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = sample_x, statistic = function(y, indices) {
    log(mean(y[indices]))
}, R = 1000)


Bootstrap Statistics :
      original         bias     std. error
t1* 1.08231352 -0.00438957075 0.0669312212
```

This results in two estimators, the raw estimator $\widehat{\theta}_1 = 1.082$ and the bootstrap estimator $\widehat{\theta}_2 = 1.087$.

How does this work with differing sample sizes? We now suppose that the x_i's are generated from a gamma distribution with shape parameter $\alpha = 0.25$ and scale parameter $\theta = 12$. We use simulation to draw the sample sizes but then act as if they were a realized set of observations. See the following illustrative code.

```
param <- function(x) {
    n <- length(x)
    theta_1 <- log(mean(x))
    results <- boot(data = x, statistic = function(y, indices) {
        log(mean(y[indices]))
    }, R = 999)
    theta_2 <- 2 * theta_1 - mean(results$t)
    return(c(theta_1, theta_2))
}
set.seed(2074)
ns <- 200
est <- function(n) {
    call_param <- function(i) {
        param(rgamma(n, shape = 0.25, scale = 12))
    }
    V <- Vectorize(call_param)(1:ns)
    apply(V, 1, median)
}
VN <- seq(15, 100, by = 5)
Est <- Vectorize(est)(VN)

save(VN, Est, file = "../IntermediateCalcs/SimulationChapter/Section832Bootstrap.Rdata")
```

The results of the comparison are summarized in Figure 8.11. This figure shows that the bootstrap estimator is closer to the true parameter value for many of the sample sizes. The bias of both estimators decreases as the sample size increases.

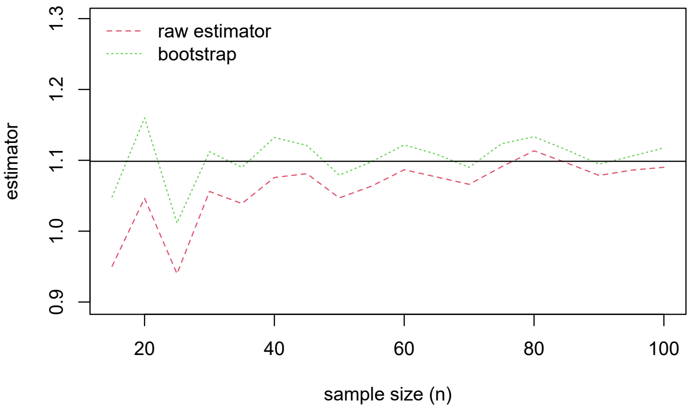


FIGURE 8.11: **Comparison of Estimates.** True value of the parameter is given by the solid horizontal line at $\log(3) \approx 1.099$.

Although successful in this example, we remark that the bootstrap bias adjusted estimator is generally not used in practice because the bias adjustment introduces extra variability into the estimator. Instead, the bias estimate provides information as to whether or not the estimate contains bias; this information gives additional information about the reliability of the estimate.

8.3.3 Confidence Intervals

The bootstrap procedure generates B replicates $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ of the estimator $\hat{\theta}$. In Example 8.3.1, we saw how to use standard normal approximations to create a confidence interval for parameters of interest. However, given that a major point is to use bootstrapping to avoid relying on assumptions of approximate normality, it is not surprising that there are alternative confidence intervals available.

For an estimator $\hat{\theta}$, the *basic* bootstrap confidence interval is

$$\left(2\hat{\theta} - q_U, 2\hat{\theta} - q_L\right), \tag{8.2}$$

where q_L and q_U are lower and upper 2.5% quantiles from the bootstrap sample $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.

To see where this comes from, start with the idea that (q_L, q_U) provides a 95% interval for $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$. So, for a random $\hat{\theta}_b^*$, there is a 95% chance that $q_L \le \hat{\theta}_b^* \le q_U$. Reversing the inequalities and adding $\hat{\theta}$ to each side gives a 95% interval

$$\hat{\theta} - q_U \le \hat{\theta} - \hat{\theta}_b^* \le \hat{\theta} - q_L.$$

So, $\left(\hat{\theta} - q_U, \hat{\theta} - q_L\right)$ is an 95% interval for $\hat{\theta} - \hat{\theta}_b^*$. The bootstrap approximation idea says that this is also a 95% interval for $\theta - \hat{\theta}$. Adding $\hat{\theta}$ to each side gives the 95% interval in equation (8.2).

Many alternative bootstrap intervals are available. The easiest to explain is the percentile bootstrap interval which is defined as (q_L, q_U). However, this has the drawback of potentially poor behavior in the tails which can be of concern in some actuarial problems of interest.

Example 8.3.3. Bodily Injury Claims and Risk Measures. To see how the bootstrap confidence intervals work, we return to the bodily injury auto claims considered in Example 8.3.1. Instead of the loss elimination ratio, suppose we wish to estimate the 95th percentile $F^{-1}(0.95)$ and a measure defined as

$$ES_{0.95}[X] = \mathrm{E}[X|X > F^{-1}(0.95)].$$

This measure is called the expected shortfall. In this formulation, it is the expected value of X conditional on X exceeding the 95th percentile which is also sometimes known as the *conditional value at risk*. Section 13.2 explains how quantiles and the expected shortfall are the two most important examples of so-called *risk measures*. For now, we will simply think of these as measures that we wish to estimate. For the percentile, we use the nonparametric estimator $F_n^{-1}(0.95)$ defined in Section 4.4.1. For the expected shortfall, we use the plug-in principle to define the nonparametric estimator

$$ES_{n,0.95}[X] = \frac{\sum_{i=1}^n X_i I[X_i > F_n^{-1}(0.95)]}{\sum_{i=1}^n I[X_i > F_n^{-1}(0.95)]} .$$

In this expression, the denominator counts the number of observations that exceed the 95th percentile $F_n^{-1}(0.95)$. The numerator adds up losses for those observations that exceed $F_n^{-1}(0.95)$. Table 8.5 summarizes the estimator for selected fractions.

Table 8.5. **Bootstrap Estimates of Quantiles at Selected Fractions**

Fraction	NP	Bootstrap	Bootstrap	Lower Normal	Upper Normal	Lower Basic	Upper Basic	Lower Percentile	Upper Percentile
	Estimate	Bias	SD	95% CI	95% CI	95% CI	95% CI	95% CI	95% CI
0.50	6500.00	-128.02	200.36	6235.32	7020.72	6300.00	7000.00	6000.00	6700.00
0.80	9078.40	89.51	200.27	8596.38	9381.41	8533.20	9230.40	8926.40	9623.60
0.90	11454.00	55.95	480.66	10455.96	12340.13	10530.49	12415.00	10493.00	12377.51
0.95	13313.40	13.59	667.74	11991.07	14608.55	11509.70	14321.00	12305.80	15117.10
0.98	16758.72	101.46	1273.45	14161.34	19153.19	14517.44	19326.95	14190.49	19000.00

For example, when the fraction is 0.50, we see that lower and upper 2.5th quantiles of the bootstrap simulations are $q_L = 6000$ and $q_u = 6700$, respectively. These form the percentile bootstrap confidence interval. With the nonparametric estimator 6500, these yield the lower and upper bounds of the basic confidence interval 6300 and 7000, respectively. Table 8.5 also shows bootstrap estimates of the bias, standard deviation, and a normal confidence interval, concepts introduced in Section 8.3.2.

Table 8.6 shows similar calculations for the expected shortfall. In each case, we see that the bootstrap standard deviation increases as the fraction increases. This is because there are fewer observations to estimate quantiles as the fraction increases, leading to greater imprecision. Confidence intervals also become wider. Interestingly, there does not seem to be the same pattern in the estimates of the bias.

Table 8.6. **Bootstrap Estimates of ES at Selected Risk Levels**

Fraction	NP	Bootstrap	Bootstrap	Lower Normal	Upper Normal	Lower Basic	Upper Basic	Lower Percentile	Upper Percentile
	Estimate	Bias	SD	95% CI	95% CI	95% CI	95% CI	95% CI	95% CI
0.50	9794.69	-120.82	273.35	9379.74	10451.27	9355.14	10448.87	9140.51	10234.24
0.80	12454.18	30.68	481.88	11479.03	13367.96	11490.62	13378.52	11529.84	13417.74
0.90	14720.05	17.51	718.23	13294.82	16110.25	13255.45	16040.72	13399.38	16184.65
0.95	17072.43	5.99	1103.14	14904.31	19228.56	14924.50	19100.88	15043.97	19220.36
0.98	20140.56	73.43	1587.64	16955.40	23178.85	16942.36	22984.40	17296.71	23338.75

8.3.4 Parametric Bootstrap

The idea of the nonparametric bootstrap is to resample by drawing independent variables from the empirical cumulative distribution function F_n. In contrast, with parametric bootstrap, we draw independent variables from $F_{\widehat{\theta}}$ where the underlying distribution is assumed to be in a parametric family such as a gamma or lognormal distribution. Typically, parameters from this distribution are estimated based on a sample and denoted as $\hat{\theta}$.

Example 8.3.4. Lognormal distribution. Consider again the dataset

```
sample_x <- c(2.46, 2.8, 3.28, 3.86, 2.85, 3.67, 3.37, 3.4, 5.22, 2.55, 2.79, 4.5,
    3.37, 2.88, 1.44, 2.56, 2, 2.07, 2.19, 1.77)
```

The classical (nonparametric) bootstrap was based on the following samples.

```
x <- sample(sample_x, replace = TRUE)
```

Instead, for the parametric bootstrap, we have to assume that the distribution of x_i's is from a specific family. As an example, the following code utilizes a lognormal distribution.

```
library(MASS)
fit <- fitdistr(sample_x, dlnorm, list(meanlog = 1, sdlog = 1))
fit
```

```
     meanlog          sdlog
  1.0363069735   0.3059343996
 (0.0684090114) (0.0483702729)
```

Then we draw from that distribution.

```
x <- rlnorm(length(sample_x), meanlog = fit$estimate[1], sdlog = fit$estimate[2])
```

Figure 8.12 compares the bootstrap distributions for the coefficient of variation, one based on the nonparametric approach and the other based on a parametric approach, assuming a lognormal distribution.

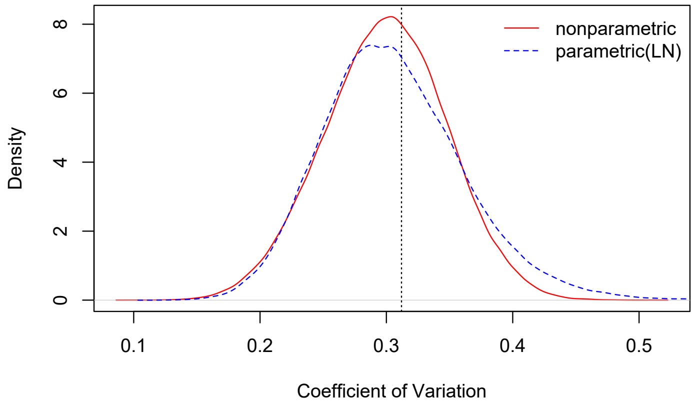


FIGURE 8.12: **Comparison of Nonparametric and Parametric Bootstrap Distributions for the Coefficient of Variation**

Example 8.3.5. Bootstrapping Censored Observations. The parametric bootstrap draws simulated realizations from a parametric estimate of the distribution function. In the same way, we can draw simulated realizations from estimates of a distribution function. As one example, we might draw from smoothed estimates of a distribution function introduced in Section 4.4.1. Another special case, considered here, is to draw an estimate from the Kaplan-Meier estimator introduced in Section 5.3.3. In this way, we can handle observations that are censored.

Specifically, return to the bodily injury data in Examples 8.2.1 and 8.2.3 but now we include the 17 claims that were censored by policy limits. In Example 4.3.6, we used this full dataset to estimate the Kaplan-Meier estimator of the survival function introduced in Section 5.3.3. Table 8.7 presents bootstrap estimates of the quantiles from the Kaplan-Meier survival function estimator. These include the bootstrap precision estimates, bias and standard deviation, as well as the basic 95% confidence interval.

Table 8.7. **Bootstrap Kaplan-Meier Estimates of Quantiles at Selected Fractions**

Fraction	KM NP	Bootstrap	Bootstrap	Lower Basic	Upper Basic
	Estimate	Bias	SD	95% CI	95% CI
0.50	6500	18.77	177.38	6067	6869
0.80	9500	167.08	429.59	8355	9949
0.90	12756	37.73	675.21	10812	13677
0.95	18500	Inf	NaN	12500	22300
0.98	25000	Inf	NaN	-Inf	27308

Results in Table 8.7 are consistent with the results for the uncensored subsample in Table 8.5. In Table 8.7, we note the difficulty in estimating quantiles at large fractions due to the censoring. However, for moderate size fractions (0.50, 0.80, and 0.90), the Kaplan-Meier nonparametric (KM NP) estimates of the quantile are consistent with those Table 8.5. The bootstrap standard deviation is smaller at the 0.50 (corresponding to the median) but larger at the 0.80 and 0.90 levels. The censored data analysis summarized in Table 8.7 uses more data than the uncensored subsample analysis in Table 8.5 but also has difficulty extracting information for large quantiles.

8.4 Model Selection and Cross-Validation

In this section, you learn how to:

- Compare and contrast cross-validation to simulation techniques and bootstrap methods.
- Use cross-validation techniques for model selection
- Explain the jackknife method as a special case of cross-validation and calculate jackknife estimates of bias and standard errors

Cross-validation, briefly introduced in Chapter 2 and Section 6.5, is a technique based on simulated outcomes that is especially useful for selecting an appropriate model. We now compare and contrast cross-validation to other simulation techniques already introduced in this chapter.

- Simulation, or Monte-Carlo, introduced in Section 8.1, allows us to compute expected values and other summaries of statistical distributions, such as p-values, readily.

- Bootstrap, and other resampling methods introduced in Section 8.3, provides estimators of the precision, or variability, of statistics.
- Cross-validation is important when assessing how accurately a predictive model will perform in practice.

Overlap exists but nonetheless it is helpful to think about the broad goals associated with each statistical method.

To discuss cross-validation, let us recall from Chapter 2 some of the key ideas of model validation. When assessing, or validating, a model, we look to performance measured on *new* data, or at least not those that were used to fit the model. A classical approach is to split the sample in two: a subpart (the *training* dataset) is used to fit the model and the other one (the *testing* dataset) is used to validate. However, a limitation of this approach is that results depend on the split; even though the overall sample is fixed, the split between training and test subsamples varies randomly. A different training sample means that model estimated parameters will differ. Different model parameters and a different test sample means that validation statistics will differ. Two analysts may use the same data and same models yet reach different conclusions about the viability of a model (based on different random splits), a frustrating situation.

8.4.1 k-Fold Cross-Validation

To mitigate this difficulty, it is common to use a cross-validation approach as introduced in Section 4.2.4. The key idea is to emulate the basic test/training approach to model validation by repeating it many times through averaging over different splits of the data. A key advantage is that the validation statistic is not tied to a specific parametric (or nonparametric) model - one can use a nonparametric statistic or a statistic that has economic interpretations - and so this can be used to compare models that are not nested (unlike likelihood ratio procedures).

Example 8.4.1. Wisconsin Property Fund. For the 2010 property fund data introduced in Section 1.3, we fit gamma and Pareto distributions to the 1,377 claims data. For details of the related goodness of fit, see Appendix Section 15.4.4. We now consider the Kolmogorov-Smirnov statistic introduced in Section 6.1.2. When the entire dataset was fit, the Kolmogorov-Smirnov goodness of fit statistic for the gamma distribution turns out to be 0.2639 and for the Pareto distribution is 0.0478. The lower value for the Pareto distribution indicates that this distribution is a better fit than the gamma.

To see how k-fold cross-validation works, we randomly split the data into $k = 8$ groups, or folds, each having about $1377/8 \approx 172$ observations. Then, we fit gamma and Pareto models to a data set with the first seven folds (about $172 \times 7 = 1,204$ observations), determine estimated parameters, and then used these fitted models with the held-out data to determine the Kolmogorov-Smirnov statistic.

The results appear in Figure 8.13 where horizontal axis is Fold=1. This process was repeated for the other seven folds. The results summarized in Figure 8.13 show that the Pareto consistently provides a more reliable predictive distribution than the gamma.

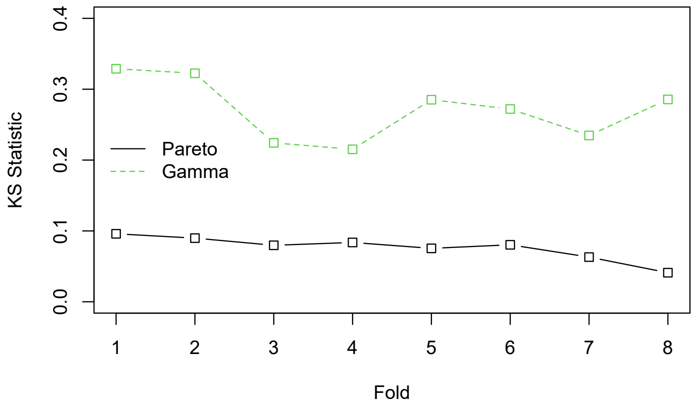


FIGURE 8.13: **Cross Validated Kolmogorov-Smirnov (KS) Statistics for the Property Fund Claims Data.** The solid black line is for the Pareto distribution, the green dashed line is for the gamma distribution. The KS statistic measures the largest deviation between the fitted distribution and the empirical distribution for each of 8 groups, or folds, of randomly selected data.

8.4.2 Leave-One-Out Cross-Validation

A special case where $k = n$ is known as leave-one-out cross validation. This case is historically prominent and is closely related to jackknife statistics, a precursor of the bootstrap technique.

Even though we present it as a special case of cross-validation, it is helpful to given an explicit definition. Consider a generic statistic $\widehat{\theta} = t(\boldsymbol{x})$ that is an estimator for a parameter of interest θ. The idea of the jackknife is to compute n values $\widehat{\theta}_{-i} = t(\boldsymbol{x}_{-i})$, where $\boldsymbol{x}_{-i}$ is the subsample of $\boldsymbol{x}$ with the i-th value removed. The average of these values is denoted as

$$\overline{\widehat{\theta}}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\theta}_{-i}.$$

These values can be used to create estimates of the bias of the statistic $\widehat{\theta}$

$$Bias_{jack} = (n-1)\left(\overline{\widehat{\theta}}_{(\cdot)} - \widehat{\theta}\right) \tag{8.3}$$

as well as a standard deviation estimate

$$s_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}_{-i} - \overline{\widehat{\theta}}_{(\cdot)}\right)^2}\ . \tag{8.4}$$

Example 8.4.2. Coefficient of Variation. To illustrate, consider a small fictitious sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ with realizations

```
sample_x <- c(2.46, 2.8, 3.28, 3.86, 2.85, 3.67, 3.37, 3.4, 5.22, 2.55, 2.79, 4.5,
    3.37, 2.88, 1.44, 2.56, 2, 2.07, 2.19, 1.77)
```

Suppose that we are interested in the coefficient of variation $\theta = CV = \sqrt{\mathrm{Var}\,[X]}/\mathrm{E}\,[X]$.

With this dataset, the estimator of the coefficient of variation turns out to be 0.31196. But how reliable is it? To answer this question, we can compute the jackknife estimates of bias and its standard deviation. The following code shows that the jackknife estimator of the bias is $Bias_{jack} =$ -0.00627 and the jackknife standard deviation is $s_{jack} = 0.01293$.

```
# Sample Code for Example 8.4.2
CVar <- function(x) sqrt(var(x))/mean(x)
JackCVar <- function(i) sqrt(var(sample_x[-i]))/mean(sample_x[-i])
JackTheta <- Vectorize(JackCVar)(1:length(sample_x))
BiasJack <- (length(sample_x) - 1) * (mean(JackTheta) - CVar(sample_x))
sdJack <- sd(JackTheta)
```

Example 8.4.3. Bodily Injury Claims and Loss Elimination Ratios. In Example 8.3.1, we showed how to compute bootstrap estimates of the bias and standard deviation for the loss elimination ratio using the Example 4.1.11 bodily injury claims data. We follow up now by providing comparable quantities using jackknife statistics.

Table 8.8 summarizes the results of the jackknife estimation. It shows that jackknife estimates of the bias and standard deviation of the loss elimination ratio E $[\min(X, d)]/\mathrm{E}\ [X]$ are largely consistent with the bootstrap methodology. Moreover, one can use the standard deviations to construct normal based confidence intervals, centered around a bias-corrected estimator. For example, at $d = 14000$, we saw in Example 4.1.11 that the nonparametric estimate of LER is 0.97678. This has an estimated bias of 0.00010, resulting in the (jackknife) *bias-corrected* estimator 0.97688. The 95% confidence intervals are produced by creating an interval of twice the length of 1.96 jackknife standard deviations, centered about the bias-corrected estimator (1.96 is the approximate 97.5th quantile of the standard normal distribution).

Table 8.8. **Jackknife Estimates of LER at Selected Deductibles**

d	NP	Bootstrap	Bootstrap	Jackknife	Jackknife	Lower Jackknife	Upper Jackknife
	Estimate	Bias	SD	Bias	SD	95% CI	95% CI
4000	0.54113	0.00011	0.01237	0.00031	0.00061	0.53993	0.54233
5000	0.64960	0.00027	0.01412	0.00033	0.00068	0.64825	0.65094
10500	0.93563	0.00004	0.01017	0.00019	0.00053	0.93460	0.93667
11500	0.95281	-0.00003	0.00941	0.00016	0.00047	0.95189	0.95373
14000	0.97678	0.00016	0.00687	0.00010	0.00034	0.97612	0.97745
18500	0.99382	0.00014	0.00331	0.00003	0.00017	0.99350	0.99415

You can learn more about the R code for this example at the online version of this book, Actuarial Community (2025).

Discussion. One of the many interesting things about the leave-one-out special case is the ability to replicate estimates exactly. That is, when the size of the fold is only one, then there is no additional uncertainty induced by the cross-validation. This means that analysts can exactly replicate work of one another, an important consideration.

Jackknife statistics were developed to understand precision of estimators, producing estimators of bias and standard deviation in equations (8.3) and (8.4). This crosses into goals that we have associated with bootstrap techniques, not cross-validation methods. This demonstrates how statistical techniques can be used to achieve different goals.

8.4.3 Cross-Validation and Bootstrap

The bootstrap is useful in providing estimators of the precision, or variability, of statistics. It can also be useful for model validation. The bootstrap approach to model validation is similar to the leave-one-out and k-fold validation procedures:

- Create a bootstrap sample by re-sampling (with replacement) n indices in

$\{1, \cdots, n\}$. That will be our *training sample*. Estimate the model under consideration based on this sample.

- The *test*, or *validation sample*, consists of those observations not selected for training. Evaluate the fitted model (based on the training data) using the test data.

Repeat this process many (say B) times. Take an average over the results and choose the model based on the average evaluation statistic.

Example 8.4.4. Wisconsin Property Fund. Return to Example 8.3.1 where we investigate the fit of the gamma and Pareto distributions on the property fund data. We again compare the predictive performance using the Kolmogorov-Smirnov (KS) statistic but this time using the bootstrap procedure to split the data between training and testing samples. The following provides illustrative code.

We did the sampling using $B = 100$ replications. The average KS statistic for the Pareto distribution was 0.058 compared to the average for the gamma distribution, 0.262. This is consistent with earlier results and provides another piece of evidence that the Pareto is a better model for these data than the gamma.

You can learn more about the R code for this example at the online version of this book, Actuarial Community (2025).

8.5 Further Resources and Contributors

Section 8.4.2 presented the jackknife statistic as an application of (leave one out) cross-validation methods. Another way to present this material is to consider the historical development. Efron (1982) attributes the jackknife idea to Quenouille (1949). Even in this simpler time before modern computing power became widely available, the jackknife provided a handy tool to estimate the bias and standard deviation for virtually any statistic. In addition, this provided motivation for the 1979 introduction of the bootstrap in Efron (1979) (see also Efron (1992)). The bootstrap provided a tool to understand the uncertainty of a statistic, including the standard deviation.

The presentation in this book, outlined in Chapter 2, follows strategies adopted by analysts. We think of the jackknife and the bootstrap as tools that helps one understand qualities of a statistic of interest. In addition, cross-validation is a resampling strategy primarily devoted to model validation. As noted in Efron (1982), the historical development of cross-validation is a bit murkier. It is a method borne from the very simple strategy of splitting a sample in half, then using a model trained on one half to predict performance in the other half. Comparing cross-validation methods to the jackknifing and bootstrapping techniques, all are based on resampling. In addition, questions of statistical inference naturally overlap with model validation issues, so there is a natural overlap among these methods.

- For further reading, a classic, and still very readable, introduction to the jackknife and bootstrap is provided by Efron (1982).
- Here are some links to learn more about reproducibility and randomness and how to go from a random generator to a sample function.

Contributors

- **Arthur Charpentier**, Université du Quebec á Montreal, and **Edward (Jed) Frees**, University of Wisconsin-Madison, are the principal authors of the initial version of this chapter.
  - Chapter reviewers include Yvonne Chueh and Brian Hartman.
- **Edward (Jed) Frees**, University of Wisconsin-Madison and Australian National University, is the author of the second edition of this chapter. Email: jfrees@bus.wisc.edu for chapter comments and suggested improvements.
  - This chapter has benefited significantly from suggestions by Hirokazu (Iwahiro) Iwasawa.

9

Bayesian Statistics and Modeling

Chapter Preview. Up to this point in the book, we have focused almost exclusively on the frequentist approach to estimate our various loss distribution parameters. In this chapter, we switch gears and discuss a different paradigm: Bayesianism. These approaches are different as Bayesian and frequentist statisticians disagree on the source of the uncertainty: Bayesian statistics assumes that the observed data sample is fixed and that model parameters are random, whereas frequentism considers the opposite (i.e., the sample data are random, and the model parameters are fixed but unknown).

In this chapter, we introduce Bayesian statistics and modeling with a particular focus on loss data analytics. We begin in Section 9.1 by explaining the basics of Bayesian statistics: we compare it to frequentism and provide some historical context for the paradigm. We also introduce the seminal Bayes' rule that serves as a key component in Bayesian statistics. Then, building on this, we present the main ingredients of Bayesian statistics in Section 9.2: the posterior distribution, the likelihood function, and the prior distribution. Section 9.3 provides some examples of simple cases where the prior distribution is chosen for algebraic convenience, giving rise to a closed-form expression for the posterior; these are called conjugate families in the literature. Section 9.4 is dedicated to cases where we cannot get closed-form expressions and for which numerical integration is needed. Specifically, we discuss two influential Markov chain Monte Carlo samplers: the Gibbs sampler and the Metropolis–Hastings algorithm. We also discuss how to interpret the chains obtained by these methods (i.e., Markov chain diagnostics). Finally, the last section of this chapter, Section 9.5, explains the main computing resources available and gives an illustration in the context of loss data.

9.1 A Gentle Introduction to Bayesian Statistics

In Section 9.1, you learn how to:

- Describe qualitatively Bayesianism as an alternative to the frequentist approach.
- Give the historical context for Bayesian statistics.
- Use Bayes' rule to find conditional probabilities.
- Understand the basics of Bayesian statistics.

9.1.1 Bayesian versus Frequentist Statistics

Classic frequentist statistics rely on frequentist probability—an interpretation of probability in which an event's probability is defined as the limit of its relative frequency (or propensity) in many, repeatable trials. It draws conclusion from a sample that is one of many hypothetical datasets that could have been collected; the uncertainty is therefore due to the sampling error associated with the sample, while model parameters and various quantities of interest are fixed (but unknown to the experimenter).

Example 9.1.1. Coin Toss. Considering the simple case of coin tossing, if we flip a fair coin many times, we expect to see heads about 50% of the time. If we flip the coin only a few times, however, we could see a different sample just by chance. Indeed, there is a non-zero probability of observing all heads (and this even if the sample is very large). Figure 9.1 illustrates this the number of heads observed in 100 samples of five iid tosses; in this specific example, we observe six samples for which all tosses are heads.[1]

Yet, as the sample size increases, the relative frequency of heads should get closer to 50% if the coin is fair. Figure 9.2 reports that, if the number of tosses increases, then relative frequency of heads gets closer to 0.5—the probability of seeing heads on a given coin toss. In other words, increasing the sample size makes the resulting parameter estimate less uncertain, and the experimenter should be reaching a probability of 0.5 in the limit, assuming they can reproduce the experiment an infinite number of times.

[1] Each coin toss can be seen as a Bernoulli random variable, meaning that their sum is a binomial with parameters $q = 0.5$ and $m = 5$. See Chapter 20.1 for more details.

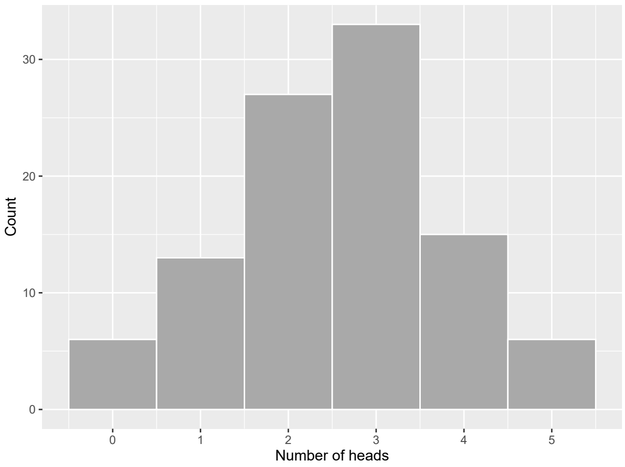


FIGURE 9.1: **Frequency histogram of the number of heads in a sample of five data points**

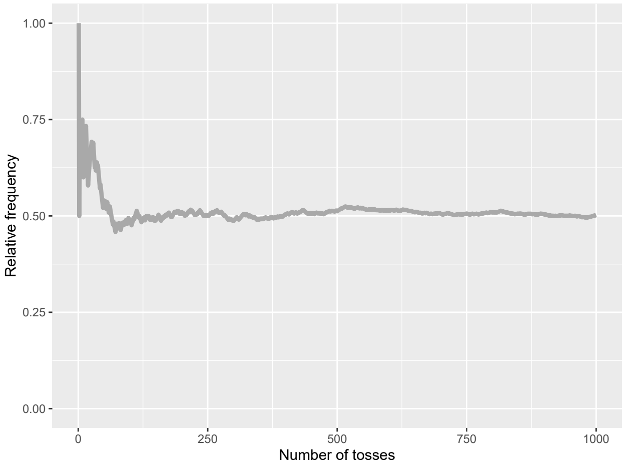


FIGURE 9.2: **Cumulative relative frequencies of heads for an increasing sample size**

Bayesians see things differently: they interpret probabilities as degrees of certainty about some quantity of interest. To find such probabilities, they draw on prior knowledge about those quantities, expressing one's beliefs before some data are taken into account. Then, as data are collected, knowledge about the world is updated, allowing us to incorporate such new information in a consistent manner; the resulting distribution is referred to as the posterior, which summarizes the information in both the prior and the data.

In the context of Bayesian statistics and modeling, this interpretation of probability implies that model parameters are assumed to be random variables—unlike the frequentist approach that considers them fixed. Starting from the prior distribution, the data—summarized via the likelihood function—are used to update the prior distribution and create a posterior distribution of the parameters (see Section 9.2 for more details on the posterior distribution, the likelihood function, and the prior distribution). The influence of the prior distribution on the posterior distribution becomes weaker as the size of the observed data sample increases: the prior information is less and less relevant as new information comes in.

Example 9.1.1. Coin Toss, continued. We now reconsider the coin tossing experiment above through a Bayesian lens. Let us first assume that we have a (potentially unfair) coin, and we wish to understand the probability of obtaining heads, denoted by q in this example. Consistent with the Bayesian paradigm, this parameter is random; let us assume that the random variable associated with the probability of observing heads is denoted by Q. For simplicity, we assume that we do not have prior information on the specific coin under investigation.[2] Assuming again that our sample contains only five iid tosses, we know that the probability of observing x heads is given by the binomial distribution with $m = 5$ such that

$$p_{X|Q=q}(x) = \Pr(X = x \,|\, Q = q) = \binom{5}{x} q^x (1-q)^{5-x}, \quad x \in \{0, 1, ..., 5\},$$

where $0 \leq q \leq 1$, which emphasizes the fact that this probability depends on parameter q by explicitly conditioning on it (unlike the notation used so far in this book, note that we append subscripts to the various pdf and pmf in this chapter to denote the random variables under study; this additional notation allows us to consider the pdf and pmf of different random variables in the same problem).

[2] Specifically, we use a uniform over $[0, 1]$ for our prior distribution. As explained in Section 9.2.3, this type of prior is said to be noninformative.

Let us generate a sample of these five tosses:

```r
set.seed(1)
nbheads <- c(1)
num_flips <- 5
coin <- c("heads", "tails")
flips <- sample(coin, size = 5, replace = TRUE)
nbheads <- sum(flips == "heads")
cat("Number of heads:", nbheads)
```

```
Number of heads: 3
```

Based on this simulation, we obtain a data sample that contains three heads and two tails. Therefore, using Bayesian statistics, we can show that

$$f_{Q|X=3}(q) \propto q^3(1-q)^2,$$

where $\propto$ means proportional to (note that obtaining this equation requires some tools that will be introduced in Section 9.2).[3] Figure 9.3 illustrates this pdf and reports the uncertainty about parameter q based on this sample of five data points. In this example, one can see that the uncertainty is quite large; this is a by-product of using only five data points. Indeed, based on these five observations, one could argue that the probability should be close to $\frac{3}{5} = 0.6$. This Bayesian analysis shows that 0.6 is likely, but that it is also very uncertain—a conclusion that is not direct in the frequentist approach.

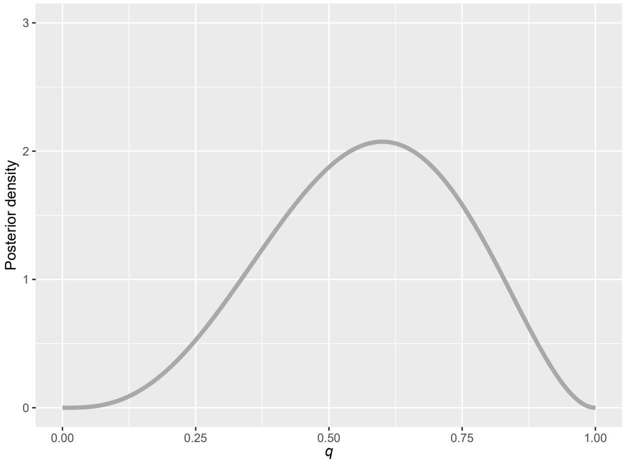


FIGURE 9.3: **Posterior probability density function of parameter q for a sample of five data points**

Figure 9.4 reports the analog of Figure 9.2 through a Bayesian lens: we see the

[3] This is also an application of the beta–binomial conjugate family that will be explained in Section 9.3.1

evolution of the posterior density of parameter q as a function of the sample size for the same sample used in Figure 9.2. As we obtain more evidence, the posterior density becomes more concentrated around 0.5—a consequence of using a fair coin in the simulations above. Yet, even if the sample size if 1,000, we still see some parameter uncertainty.

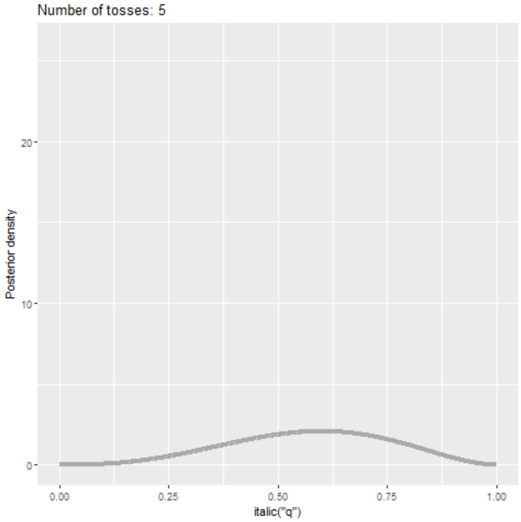


FIGURE 9.4: **Posterior probability density function of parameter q as a function of the sample size**

But why be Bayesian? There are indeed several advantages to the Bayesian approach. First, this approach allows us to describe the entire distribution of parameters conditional on the data and to provide probability statements regarding the parameters that could be interpreted as such. Second, it provides a unified approach for estimating parameters. Some non-Bayesian methods, such as least squares, require a separate approach to estimate variance components. In contrast, in Bayesian methods, all parameters can be treated in a similar fashion. Third, it allows experimenters to blend prior information from other sources with the data in a coherent manner.[4]

Are there any disadvantages to being Bayesian? Well, of course: while the Bayesian approach has many advantages, it is not without its disadvan-

[4] There is also a rich history blending prior information with data in loss modeling and in actuarial science, generally speaking; it is known as credibility. In technical terms, credibility theory's main challenge lies in identifying the optimal linear approximation to the mean of the Bayesian predictive density. This is the reason credibility theory shares numerous outcomes with both linear filtering and the broader field of Bayesian statistics. For more details on experience rating using credibility theory, see Chapter 12.

tages. First, it tends to be very computationally demanding (i.e., Bayesian methods often require complex computations, especially when dealing with high-dimensional problems or large datasets). For instance, complex models may not have closed-form solutions and require specialized computational techniques, which can be time-consuming. Second, there is some subjectivity in selecting priors—our initial beliefs and knowledge about the parameters—and this can lead to different results in the end. Third, Bayesian analysis often produces results that can be challenging to communicate effectively to non-experts.

Despite these disadvantages, the Bayesian approach remains powerful and flexible for many actuarial problems.

Do I need to be a Bayesian to embrace Bayesian statistics? No, this can be decided on a case-by-case basis. Consider a Bayesian study when you have prior knowledge or beliefs about the parameters, need to explicitly quantify uncertainty in your estimates, have limited data, require a flexible framework for complex models, or when decision-making under uncertainty is a key aspect of your analysis.

Even if one does not want to be a Bayesian truly, they can still recognize the usefulness of some of the methods. Indeed, some modern statistical tools in artificial intelligence and machine learning rely heavily on Bayesian techniques (e.g., Bayesian neural networks, Gaussian processes, and Bayesian classifiers, to name a few).

9.1.2 A Brief History Lesson

Interestingly, some have argued that the birth of Bayesian statistics is intimately related to insurance; see, for instance, Cowles (2013). Specifically, the Great Fire of London in 1666—destroying more than 10,000 homes and about 100 churches—led to the rise of insurance as we know it today. Shortly after, the first full-fledged fire insurance company came into existence in England during the 1680s. By the turn of the century, the idea of insurance was well ingrained and its use was booming in England; see, for instance, Haueter (2017). Yet, the lack of statistical models and methods—much needed to understand risk—drove some insurers to bankruptcy.

Thomas Bayes, an English statistician, philosopher and Presbyterian minister, applied his mind to some of these important statistical questions raised by insurers. This culminated into Bayes' theory of probability in his seminal essay entitled *Essay towards solving a problem in the doctrine of chances*, published posthumously in 1763. This essay laid out the foundation of what we now know as Bayesian statistics.

FIGURE 9.5: **Portrait of an unknown Presbyterian clergyman identified as Thomas Bayes in O'Donnell (1936)**

Thomas Bayes' work also helped Pierre-Simon Laplace, a famous French scholar and polymath, to develop and popularize the Bayesian interpretation of probability in the late 1700s and early 1800s. He also moved beyond Bayes' essay and generalized his framework. Laplace's efforts were followed by many, and Bayesian thinking continued to progress throughout the years with the help of statisticians like Bruno de Finetti, Harold Jeffreys, Dennis Lindley, and Leonard Jimmie Savage.

Nowadays, Bayesian statistics and modeling is widely used in science, thanks to the increase in computational power over the past 30 years. Actuarial science and loss modeling, more specifically, have also been breeding grounds for Bayesian methodology. So, Bayesian statistics circles back to insurance, in a sense, where it all started.

9.1.3 Bayes' Rule

This subsection introduces how the Bayes' rule is applied to calculating conditional probabilities for events.

Conditional Probability. The concept of conditional probability considers the relationship between probabilities of two (or more) events happening. In its most simple form, being interested in conditional probability boils down to answering this question: *given that event B happened, how does this affect the probability that A happens?* To answer this question, we can define formally the concept of conditional probability:

$$\Pr\left(A \,|\, B\right) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

To be properly defined, we must assume that $\Pr(B)$ is larger than zero; that is, event B is not impossible. Simply put, a conditional probability turns B

into the new probability space, and then cares only about the part of A that is inside B (i.e., $A \cap B$).

Example 9.1.2. Actuarial Exam Question. An insurance company estimates that 40% of policyholders who have an extended health policy but no long-term disability policy will renew next year, and 70% of policyholders who have a long-term disability policy but no extended health policy will renew next year. The company also estimates that 50% of their clients who have both policies will renew at least one next year. The company records report that 65% of clients have an extended health policy, 40% have a long-term disability policy, and 10% have both. Using the data above, calculate the percentage of policyholders that will renew at least one policy next year.[5]

Example Solution. Let E be the event that a policyholder has an extended health policy, D be the event that a policyholder has a long-term disability policy, and R be the event that a policyholder renews a policy. We are given:

- $\Pr(E) = 0.65$, - $\Pr(D) = 0.40$, - $\Pr(E \cap D) = 0.10$, - $\Pr(R \mid E \cap D^c) = 0.40$, - $\Pr(R \mid E^c \cap D) = 0.70$, - $\Pr(R \mid E \cap D) = 0.50$.

We are looking for $\Pr(R)$.

Note that

$$\Pr(E \cap D^c) = \Pr(E) - \Pr(E \cap D) = 0.65 - 0.10 = 0.55,$$

and

$$\Pr(E^c \cap D) = \Pr(D) - \Pr(E \cap D) = 0.40 - 0.10 = 0.30.$$

Moreover, note that $E \cap D^c$, $E^c \cap D$, and $E \cap D$ are mutually disjoint, and that=

$$\begin{aligned}\Pr(R) &= \Pr(R \cap (E \cap D^c)) + \Pr(R \cap (E^c \cap D)) + \Pr(R \cap (E \cap D)) \\ &= \Pr(R \mid (E \cap D^c)) \Pr(E \cap D^c) + \Pr(R \mid (E^c \cap D)) \Pr(E^c \cap D) \\ &\quad + \Pr(R \mid (E \cap D)) \Pr(E \cap D) \\ &= 0.40 \times 0.55 + 0.70 \times 0.30 + 0.50 \times 0.10 \\ &= 0.48.\end{aligned}$$

Independence. If two events are unrelated to one another, we say that they are independent. Specifically, A and B are independent if

$$\Pr(A \cap B) = \Pr(A)\ \Pr(B).$$

[5] This question was adapted from the Be An Actuary website. See here for more details.

For positive probability events, independence between A and B is also equivalent to

$$\Pr(A \mid B) = \Pr(A) \quad \text{and} \quad \Pr(B \mid A) = \Pr(B),$$

which means that the occurrence of event B does not have an impact on the occurrence of A, and vice versa.

Bayes' Rule. Intuitively speaking, Bayes' rule provides a mechanism to put our Bayesian thinking into practice. It allows us to update our information by combining the data—from the likelihood—and the prior together to obtain a posterior probability.

Proposition 9.1.1. Bayes' Rule for Events. For events A and B, the posterior probability of event A given B follows

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)},$$

where the law of total probability allows us to find

$$\Pr(B) = \Pr(A)\Pr(B \mid A) + \Pr(A^{\mathrm{c}})\Pr(B \mid A^{\mathrm{c}}).$$

Note, again, that this works as long as event B is possible (i.e., $\Pr(B) > 0$).[6]

Proof. Bayes' rule may be derived from the definition of conditional probability shown above:

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

if $\Pr(B) > 0$. Similarly,

$$\Pr(B \mid A) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

if $\Pr(A) > 0$. Solving for $\Pr(A \cap B)$ in the last equation and substituting into the first one yields Bayes' rule:

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}.$$

[6] The law of total probability states that the total probability of an event B is equal to the sum of the probabilities of B occurring under different conditions, weighted by the probabilities of those conditions. In the case where there are only two different conditions (let us say A and A^{c}), we simply need to consider these two conditions. In all generality, however, we would need to consider more possibilities if the sample space cannot be divided into only two events.

Simply put, the posterior probability of event A given B is obtained by combining the likelihood of B given a fixed A—proxied by $\Pr(B \mid A)$—with the prior probability of observing A, and then dividing it by the marginal probability of event B to make sure that the probabilities sum up to one.

Example 9.1.3. Actuarial Exam Question. An automobile insurance company insures drivers of all ages. An actuary compiled the probability of having an accident for some age bands as well as an estimate of the portion of the company's insured drivers in each age band:

Age of Driver	Probability of Accident	Portion of Company's Insured Drivers
16-20	0.06	0.08
21-30	0.03	0.15
31-65	0.02	0.49
66-99	0.04	0.28

A randomly selected driver that the company insures has an accident. Calculate the probability that the driver was age 16-20.[7]

Example Solution. Let B be the event of an insured driver having an accident, and let

- A_1 be the event related to the driver's age being in the range 16-20,
- A_2 be the event related to the driver's age being in the range 21-30,
- A_3 be the event related to the driver's age being in the range 31-65,
- A_4 be the event related to the driver's age being in the range 66-99.

Then,

$$\begin{aligned}\Pr(A_1 \mid B) &= \frac{\Pr(B \mid A_1)\Pr(A_1)}{\Pr(B \mid A_1)\Pr(A_1) + \Pr(B \mid A_2)\Pr(A_2) + \Pr(B \mid A_3)\Pr(A_3) + \Pr(B \mid A_4)\Pr(A_4)} \\ &= \frac{0.06 \times 0.08}{0.06 \times 0.08 + 0.03 \times 0.15 + 0.02 \times 0.49 + 0.04 \times 0.28} \\ &= 0.1584.\end{aligned}$$

9.1.4 An Introductory Example of Bayes' Rule

The example above illustrates how to use Bayes' rule in an academic context; the focus of this book is, nonetheless, data analytics. We therefore also wish to illustrate Bayes' rule by using *real* data. In this introductory example, we use

[7] This question was taken from the Society of Actuaries Sample Questions for Exam P. See here for more details.

the Singapore auto data `sgautonb` of the R package `CASdatasets` that was already used in Chapter 3.

```
library("CASdatasets")
data(sgautonb)
```

This dataset contains information about the number of car accidents and some risk factors (i.e., the type of the vehicle insured, the age of the vehicle, the sex of the policyholder, and the age of the policyholder grouped into seven categories).[8]

Example 9.1.4. Singapore Insurance Data. A new insurance company—targeting an older segment of the population—estimates that 20% of their policyholders will be 65 years old and older. The actuaries working at the insurance company believes that the Singapore insurance dataset is credible to understand the accident occurrence of the new company. Based on this information, find the probability that a randomly selected driver who has (at least) one accident, is 65 years or older.

Example Solution. Let O denote the event related to the policyholder being 65 years old and older (i.e., Age Category 6 in the dataset), and A the event of a policyholder having at least an accident. Using Bayes' rule, we have that

$$\Pr(O \mid A) = \frac{\Pr(A \mid O) \Pr(O)}{\Pr(A)},$$

where the prior probability $\Pr(O)$ is given by the problem statement: $\Pr(O) = 0.20$. This implies that $\Pr(O^c) = 1 - 0.20 = 0.80$. From the Singapore insurance data, we know that $\Pr(A \mid O) = 0.1082803$ and $\Pr(A \mid O^c) = 0.06415506$, which allow us to use the law of total probability to obtain:

$$\Pr(A) = \Pr(A \mid O) \Pr(O) + \Pr(A \mid O^c) \Pr(O^c).$$

```
# Example 9.1.4 Illustrative Code
n <- length(sgautonb$AgeCat)
nO <- sum(sgautonb$AgeCat == 6)
nOc <- sum(sgautonb$AgeCat != 6)
nAandO <- sum(sgautonb$AgeCat == 6 & sgautonb$Clm_Count > 0)
```

[8] The data are from the General Insurance Association of Singapore, an organization consisting of non-life insurers in Singapore. These data contain the number of car accidents for $n = 7{,}483$ auto insurance policies with several categorical explanatory variables and the exposure for each policy.

```
nAandOc <- sum(sgautonb$AgeCat != 6 & sgautonb$Clm_Count > 0)

PAO <- nAandO/nO
PAOc <- nAandOc/nOc

POA <- PAO * 0.2/(PAO * 0.2 + PAOc * 0.8)
cat("The probability that policyholder having accident \n is 65 years old and older is",
    POA)
```

```
The probability that policyholder having accident
 is 65 years old and older is 0.296739115
```

The probability that a randomly selected driver who has (at least) one accident, is 65 years or older is therefore about 29.7

In the next section, we expand on the idea of Bayes' rule and apply it to slightly more general cases involving random variables instead of events.

9.2 Building Blocks of Bayesian Statistics

In Section 9.2, you learn how to:

- Describe the main components of Bayesian statistics; that is, the posterior distribution, the likelihood function, and the prior distribution.
- Summarize the different classes of priors used in practice.

Proposition 9.1.1 above deals with the elementary case of Bayes' rule for events. Although this version of Bayes' rule is useful to understand the foundation of Bayesian statistics, we will need slightly more general versions of it to achieve our aim. Specifically, Proposition 9.1.1 needs to be generalized to the case of random variables.

Let us first consider the case of discrete random variables. Assume X and Y are both discrete random variables that allow for the following joint pmf of

$$p_{X,Y}(x,y) = \Pr(X = x \text{ and } Y = y)$$

as well as the following marginal distributions for X and Y:

$$p_X(x) = \Pr(X = x) = \sum_k p_{X,Y}(x,k) \quad \text{and} \quad p_Y(y) = \Pr(Y = y) = \sum_k p_{X,Y}(k,y),$$

respectively. Using the result of Proposition 9.1.1 and setting event A as $\{Y = y\}$ and B as $\{X = x\}$ yields

$$p_{Y|X=x}(y) = \frac{p_{X|Y=y}(x)\, p_Y(y)}{p_X(x)},$$

where $p_{Y|X=x}(y) = \Pr(Y = y \,|\, X = x)$ is the conditional pmf of Y conditional on X being equal to x. Using the law of total probability,

$$p_X(x) = \sum_k p_{X,Y}(x, k) = \sum_k p_{X|Y=k}(x)\, p_Y(k),$$

we can rewrite the denominator above to get the following version of Bayes' rule:

$$p_{Y|X=x}(y) = \frac{p_{X|Y=y}(x)\, p_Y(y)}{\sum_k p_{X|Y=k}(x)\, p_Y(k)}.$$

We can also obtain a similar Bayes' rule for continuous random variables by replacing probability mass functions by probability density functions, and sums by integrals.

Proposition 9.2.1. Bayes' Rule for Continuous Random Variables. For two continuous random variables X and Y, the conditional probability density function of Y given $X = x$ follows

$$f_{Y|X=x}(y) = \frac{f_{X|Y=y}(x)\, f_Y(y)}{f_X(x)},$$

where the marginal distributions of X and Y are given as follows:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, u)\, du \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(u, y)\, du,$$

respectively. Similar to the discrete random variable case, we can swap the denominator of the equation above for

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, u)\, du = \int_{-\infty}^{\infty} f_{X|Y=u}(x)\, f_Y(u)\, du$$

by using the law of total probability.

Proof. Bayes' rule for continuous random variables may be derived from the definition of conditional probability density functions:

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x, y)}{f_X(x)},$$

if $f_X(x) > 0$. Similarly,

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

if $f_Y(y) > 0$. Solving for $f_{X,Y}(x,y)$ in the last equation and substituting into the first one yields Bayes' rule for continuous random variables:

$$f_{Y|X=x}(y) = \frac{f_{X|Y=y}(x)\, f_Y(y)}{f_X(x)}.$$

Note that one can mix the discrete and continuous definitions of Bayes' rule to accommodate for cases where the parameters have continuous random variables and the observations are expressed via discrete random variables, or vice versa.

9.2.1 Posterior Distribution

Model parameters are assumed to be random variables under the Bayesian paradigm, meaning that Bayes' rule for (discrete or continuous) random variables can be applied to update the prior knowledge about parameters by using new data. This is indeed similar to the process used in Section 9.1.1.

Let us consider only one unknown model parameter θ associated with random variable Θ for now.[9] Further, consider n observations

$$\mathbf{x} = (x_1, x_2, ..., x_n),$$

which are realizations of the collection of random variables

$$\mathbf{X} = (X_1, X_2, ..., X_n).$$

If Y in Proposition 9.2.1 is replaced by Θ and X by $\mathbf{X}$, we obtain

$$f_{\Theta|\mathbf{X}=\mathbf{x}}(\theta) = \frac{f_{\mathbf{X}|\Theta=\theta}(\mathbf{x})\, f_\Theta(\theta)}{f_{\mathbf{X}}(\mathbf{x})},$$

which represents the posterior distribution of the model parameter after updating the distribution based on the new observations $\mathbf{x}$, and where

- $f_{\mathbf{X}|\Theta=\theta}(\mathbf{x})$ is the likelihood function, also known as the conditional joint pdf of the observations assuming a given value of parameter θ,
- $f_\Theta(\theta)$ is the unconditional pdf of the parameter that represents the prior information, and

[9] For the sake of simplicity, we only consider one parameter in our derivation here. Note that, later, we will consider cases with more than one parameter and that this extension does not change the bulk of our results and derivations.

- $f_{\mathbf{X}}(\mathbf{x})$ is the marginal likelihood, which is a constant term with respect to θ, making the posterior density integrate to one.

In other words, Bayes' rule provides a way to update the prior distribution of the parameter into a posterior distribution—by considering the observations $\mathbf{x}$.

Note that the marginal likelihood is constant once we have the observations. It does not depend on θ and does not impact the overall shape of the pdf: it only provides the adequate scaling to ensure that the density integrates to one. For this reason, it is common to write down the posterior distribution using a proportional relationship instead:

$$f_{\Theta|\mathbf{X}=\mathbf{x}}(\theta) \propto \underbrace{f_{\mathbf{X}|\Theta=\theta}(\mathbf{x})}_{\text{Likelihood}} \; \underbrace{f_{\Theta}(\theta)}_{\text{Prior}}.$$

Example 9.2.1. A Problem Inspired from Meyers (1994). A car insurance pays the following (independent) claim amounts on an automobile insurance policy:

$$1050, \qquad 1250, \qquad 1550, \qquad 2600, \qquad 5350, \qquad 10200.$$

The amount of a single payment is distributed as a single-parameter Pareto distribution with $\theta = 1000$ and α unknown, such that

$$f_{X_i|A=\alpha}(x_i) = \frac{\alpha\, 1000^{\alpha}}{x_i^{\alpha+1}}, \quad x_i \in \mathbb{R}_+.$$

We assume that the prior distribution of α is given by a gamma distribution with shape parameter 2 and scale parameter 1, and its pdf is given by

$$f_A(\alpha) = \alpha\, e^{-\alpha}, \quad \alpha \in \mathbb{R}_+.$$

Find the posterior distribution of parameter α.

Example Solution. The likelihood function is constructed by multiplying the pdf of the single payment amounts because they are independent; that is,

$$f_{\mathbf{X}|A=\alpha}(\mathbf{x}) = \prod_{i=1}^{6} f_{X_i|A=\alpha}(x) = \frac{\alpha^6\, 1000^{6\alpha}}{\prod_{i=1}^{6} x_i^{\alpha+1}} = \alpha^6\, e^{-5.66518\alpha-41.44653}.$$

The posterior distribution is given by

$$f_{A|\mathbf{X}=\mathbf{x}}(\alpha) = \frac{\alpha^7\, e^{-6.66518\alpha-41.44653}}{\int_0^\infty \alpha^7\, e^{-6.66518\alpha-41.44653}\, d\alpha} = \frac{\alpha^7\, e^{-6.66518\alpha}}{\int_0^\infty \alpha^7\, e^{-6.66518\alpha}\, d\alpha}.$$

Interestingly, we do not need to solve the integral in the denominator to find this distribution. As we know that the results should be a proper pdf and that the numerator looks like a gamma distribution, we can deduce that

$$f_{A|\mathbf{X}=\mathbf{x}}(\alpha) = \frac{6.66518^8}{\Gamma(8)}\,\alpha^7\,e^{-6.66518\,\alpha},$$

which is a gamma distribution with shape parameter 8 and scale parameter $\frac{1}{6.66518}$. Figure 9.6 reports the posterior distribution of α.

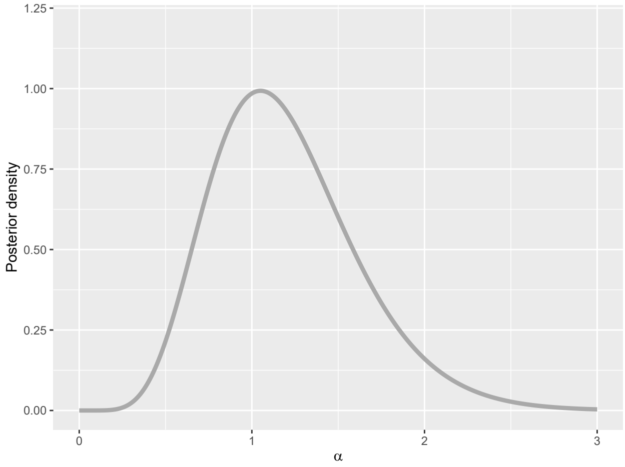


FIGURE 9.6: **Posterior densities of parameter α**

The discussion above considered continuous random variables, but the same logic can be applied to discrete random variables by replacing probability density functions by probability mass functions.

Example 9.2.2. Coin Toss Revisited. Assume that you observe three heads out of five (independent) tosses. Each toss has a probability of q of observing heads and $1-q$ of observing tails. Find the posterior distribution of q assuming a uniform prior distribution over the interval $[0,1]$.

Example Solution. The prior distribution of q is given by

$$f_Q(q) = 1, \quad q \in [0,1].$$

Assuming the likelihood function conditional on $Q = q$ is given by a binomial distribution with $m = 5$ and $x = 3$,

$$p_{X|Q=q}(x) = \binom{5}{3} q^3 (1-q)^2,$$

we have that the posterior distribution of q is given by

$$f_{Q|X=3}(q) \propto p_{X|Q=q}(x)\, f_Q(q) = q^3\,(1-q)^2,$$

which is a beta distribution with $a = 4$, $b = 3$, and $\theta = 1$; that is, we can easily deduce that

$$f_{Q|X=3}(q) = \frac{\Gamma(7)}{\Gamma(4)\Gamma(3)}\, q^3\,(1-q)^2.$$

In the following subsections, we will discuss at greater length the two main building blocks used to build the posterior distribution: the likelihood function and the prior distribution.

9.2.2 Likelihood Function

The likelihood function is a fundamental concept in statistics. It is used to estimate the parameters of a statistical model based on observed data. As mentioned in previous chapters, the likelihood function can be used to find the maximum likelihood estimator. In Bayesian statistics, the likelihood function is used to update the prior based on the evidence (or data).

As explained above and in Chapter 17, the likelihood function is defined as the conditional joint pdf or pmf of the observed data, given the model parameters. In other words, it is the probability of observing the data given a specific parameter values.

Mathematically, the likelihood function is written as $f_{\mathbf{X}|\Theta=\theta}(x)$ (for continuous random variables) or $p_{\mathbf{X}|\Theta=\theta}(x)$ (for discrete random variables). Note that, throughout the book, the notation $L(\theta|\mathbf{x})$ has also been used for the likelihood function, and we will use both interchangeably in this chapter.

Special Case: Independent and Identically Distributed Observations. Oftentimes, in many problems and real-world applications, the observations are assumed to be iid. If they are, then we can easily write the likelihood function as:

$$f_{\mathbf{X}|\Theta=\theta}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i|\Theta=\theta}(x_i) \quad \text{or} \quad p_{\mathbf{X}|\Theta=\theta}(\mathbf{x}) = \prod_{i=1}^{n} p_{X_i|\Theta=\theta}(x_i).$$

9.2.3 Prior Distribution

In the Bayesian paradigm, the prior distribution represents our knowledge or beliefs about the unknown parameters before we observe any data. It is a probability distribution that expresses the uncertainty about the values of the parameters. The prior distribution is typically specified by choosing a family of probability distributions and selecting specific values for its parameters.

The choice of prior distribution is subjective and often based on external information or previous studies. In some cases, noninformative priors can be used, which represent minimal prior knowledge or assumptions about the parameters. In other cases, informative and weakly informative priors can be used, which incorporate prior knowledge or assumptions based on external sources. The selection of the prior distribution should be carefully considered, and sensitivity analysis can be performed to assess the robustness of the results to different prior assumptions.

Why Does It Matter? The choice of prior distribution can have a significant impact on the results of a Bayesian analysis. Different prior distributions can lead to different posterior distributions, which are the updated probability distributions for the parameters after we observe the data. Therefore, it is important to choose a prior distribution that reflects our prior knowledge or beliefs about the parameters.

Informative and Weakly Informative Priors

Informative and weakly informative priors are terms used to describe the amount of prior knowledge or beliefs that is incorporated into a statistical model. Informative priors contain substantial prior knowledge about the parameters of a model, while weakly informative priors contain moderate prior knowledge.

Informative priors are useful when there is strong, potentially subjective prior information available about the model parameters, which can help to constrain the posterior distribution and improve inference. For example, in an insurance claims analysis study, an informative prior may be used to incorporate previous knowledge, such as the results of a previous claims study.

On the other hand, weakly informative priors are used when there is some—yet little—prior knowledge available or when the goal is to allow the data to drive the analysis. Weakly informative priors are designed to mildly impact the posterior distribution and are often chosen based on principles such as symmetry or scale invariance.

Overall, the choice of prior depends on the specific problem at hand and the available prior knowledge or beliefs. Informative priors can be useful when prior information is available and can improve the precision of the posterior distribution. In contrast, weakly informative priors can be useful when the goal is to allow the data to drive the analysis and avoid imposing strong prior assumptions.

Example 9.2.3. Actuarial Exam Question. You are given:

- Annual claim frequencies follow a Poisson distribution with mean λ.
- The prior distribution of λ has the following pdf:

$$f_\Lambda(\lambda) = (0.3)\frac{1}{6}e^{-\frac{\lambda}{6}} + (0.7)\frac{1}{12}e^{-\frac{\lambda}{12}}, \quad \text{where } \lambda > 0.$$

Ten claims are observed for an insured in Year 1. Calculate the expected value of the posterior distribution of λ.[10]

Example Solution. The posterior distribution can be found from:

$$\begin{aligned} f_{\Lambda|X=10}(\lambda) &= \frac{p_{X|\Lambda=\lambda}(10) f_\Lambda(\lambda)}{p_X(10)} \\ &= \frac{\frac{e^{-\lambda}\lambda^{10}}{10!}\left((0.3)\frac{1}{6}e^{-\frac{\lambda}{6}} + (0.7)\frac{1}{12}e^{-\frac{\lambda}{12}}\right)}{\int_0^\infty \frac{e^{-\lambda}\lambda^{10}}{10!}\left((0.3)\frac{1}{6}e^{-\frac{\lambda}{6}} + (0.7)\frac{1}{12}e^{-\frac{\lambda}{12}}\right) d\lambda} \\ &= \frac{\lambda^{10}\left(\frac{0.3}{6}e^{-\frac{7\lambda}{6}} + \frac{0.7l}{12}e^{-\frac{13\lambda}{12}}\right)}{121050}. \end{aligned}$$

The posterior mean is therefore given by

$$\begin{aligned} \mathrm{E}\left[\Lambda \mid X = 10\right] &= \frac{1}{121050}\int_0^\infty \lambda^{11}\left(\frac{0.3}{6}e^{-\frac{7\lambda}{6}} + \frac{0.7}{12}e^{-\frac{13\lambda}{12}}\right) d\lambda \\ &= \frac{1}{118170}\left(\frac{0.3}{6}(11!)(6/7)^{12} + \frac{0.7}{12}(11!)(12/13)^{12}\right) \\ &= 9.95442. \end{aligned}$$

Noninformative Priors

It is possible to take the idea of weakly informative priors to the extreme by using noninformative priors. A noninformative prior is a prior distribution that is intentionally chosen to allow the data to have a more decisive influence on the posterior distribution rather than being overly influenced by prior beliefs or assumptions.

[10] This question is a modified version of Sample Question 184 of the Society of Actuaries Exam C sample questions.

Noninformative priors can take different forms, such as flat priors, for instance. A flat prior assigns equal probability to all possible parameter values without additional information or assumptions.

Example 9.2.4. Informative Versus Noninformative Priors. You wish to investigate the impact of having informative and noninformative priors on a claim frequency analysis. Assume that the claim frequency for each policy follows a Bernoulli random variable with a probability of q such that

$$q_{X_i|Q=q}(x_i) = q^{x_i}(1-q)^{1-x_i}, \quad x_i \in \{0,1\},$$

where $q \in [0,1]$, and consider two different prior distributions:

- Informative: Based on past experience, you know that the claim probability is typically less than 5%, thus justifying the use of a uniform distribution over $[0, 0.05]$.
- Noninformative: You do not wish your posterior distribution to be impacted by your prior assumption and simply select a uniform distribution over the domain of q, which is $[0,1]$.

Using the first 100 lines of the Singapore insurance dataset (see Example 9.1.4 for more details on this dataset), find the two posterior distributions as well as the posterior expected value of the probability q under both prior assumptions.

Example Solution. Let us start with the informative prior, where

$$f_Q(q) = \frac{1}{0.05-0} = 20, \quad \text{if } q \in [0, 0.05],$$

and zero otherwise. In this case, assuming $x = \sum_{i=1}^{100} x_i$, the posterior density is given by

$$\begin{aligned} f_{Q|\mathbf{X}=\mathbf{x}}(q) &\propto f_{\mathbf{X}|Q=q}(\mathbf{x}) f_Q(q) \\ &\propto \prod_{i=1}^{100} q^{x_i}(1-q)^{1-x_i} \\ &= q^x(1-q)^{100-x}, \quad \text{if } 0 \le q \le 0.05, \end{aligned}$$

and zero otherwise. We can numerically obtain the shape of this posterior distribution by dividing $q^x(1-q)^{100-x}$ by

$$\int_0^{0.05} q^x(1-q)^{100-x}\, dq.$$

Note that this prior makes it impossible for the estimated frequency to be greater than 0.05.

The second prior is still uniform, but over $[0, 1]$ this time, which is given mathematically by

$$f_Q(q) = \frac{1}{1-0} = 1, \quad \text{if } q \in [0, 1],$$

and zero otherwise, leading to the following posterior distribution:

$$\begin{aligned} f_{Q|\mathbf{X}=\mathbf{x}}(q) &\propto f_{\mathbf{X}|Q=q}(\mathbf{x}) f_Q(q) \\ &\propto \prod_{i=1}^{100} q^{x_i}(1-q)^{1-x_i} \\ &= q^x(1-q)^{100-x}, \quad \text{if } 0 \le q \le 1, \end{aligned}$$

and zero otherwise.

```
qs <- seq(from = 0, to = 0.12, by = 0.0001)
x <- sum(sgautonb$Clm_Count[1:100])

integrandposterior1 <- function(q) {
    q^x * (1 - q)^(100 - x) * ifelse(q >= 0 & q <= 0.05, 1, 0)
}
marglikelihood1 <- integrate(integrandposterior1, 0, 1, abs.tol = .Machine$double.eps^2)$value
posterior1 <- integrandposterior1(qs)/marglikelihood1

integrandposterior2 <- function(q) {
    q^x * (1 - q)^(100 - x) * ifelse(q >= 0 & q <= 1, 1, 0)
}
marglikelihood2 <- integrate(integrandposterior2, 0, 1, abs.tol = .Machine$double.eps^2)$value
posterior2 <- integrandposterior2(qs)/marglikelihood2
```

We also wish to obtain the expected value of q for both posterior distribution. This can be obtained by numerically integrating the following equation:

$$\mathrm{E}[Q|\mathbf{X} = \mathbf{x}] = \int_0^1 q\, f_{Q|\mathbf{X}=\mathbf{x}}(q)\, dq.$$

```
integrandexpvalue1 <- function(q) {
    integrandposterior1(q)/marglikelihood1 * q
}
expectedvalue1 <- integrate(integrandexpvalue1, 0, 1, abs.tol = .Machine$double.eps^2)$value
cat("The posterior expected value of the parameter \n
    when using the informative prior is",
    expectedvalue1)
```

```
The posterior expected value of the parameter

    when using the informative prior is 0.0304525117
```

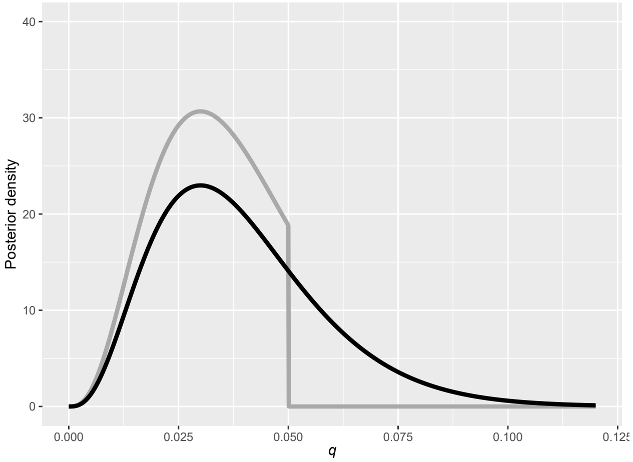


FIGURE 9.7: **Posterior densities based on informative (gray) and noninformative priors (black)**

```
integrandexpvalue2 <- function(q) {
    integrandposterior2(q)/marglikelihood2 * q
}
expectedvalue2 <- integrate(integrandexpvalue2, 0, 1, abs.tol = .Machine$double.eps^2)$value
cat("The posterior expected value of the parameter \n
    when using the noninformative prior is",
    expectedvalue2)
```

```
The posterior expected value of the parameter

    when using the noninformative prior is 0.0392156863
```

As one can see, these values are different, meaning that the prior distribution can have a material impact on the posterior distribution. One should therefore be careful when selecting a prior distribution.

Improper Priors

An improper prior is a prior distribution that is not a proper probability distribution, meaning that it does not integrate (or sum) to one over the entire parameter space. Improper priors can be used in Bayesian analyses, but they require careful handling because they can lead to improper posterior distributions.

Improper priors are typically used when there is little or no prior information about the parameter of interest—some noninformative priors are indeed improper—and they can be thought of as representing a very diffuse or noncommittal prior belief. For instance, the uniform distribution on an infinite interval is a common choice of improper prior.

Example 9.2.5. Improper Prior, Proper Posterior. Let us assume a random sample $\mathbf{x}$ of size n, which is a realization of the collection of random variables $\mathbf{X} = (X_1, X_2, ..., X_n)$. Further, assume that each random variable X_i is independent and normally distributed with mean of μ and variance of 1:

$$f_{X_i|M=\mu}(x_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2\right), \quad x_i \in \mathbb{R},$$

where μ is a (random) parameter. Obtain the posterior distribution of μ assuming that its prior distribution is improper and given by $f_M(\mu) \propto 1$, where $\mu \in \mathbb{R}$.

Example Solution. According to Bayes' rule, we have that

$$f_{M|\mathbf{X}=\mathbf{x}}(\mu) = \frac{f_{\mathbf{X}|M=\mu}(\mathbf{x})\, f_M(\mu)}{f_{\mathbf{X}}(\mathbf{x})} \propto \prod_{i=1}^{n} f_{X_i|M=\mu}(x_i)$$

because $f_M(\mu) \propto 1$ and $f_{\mathbf{X}}(\mathbf{x})$ does not depend on μ. Using the equation above, we can obtain the posterior distribution by simplifying the following equation:

$$\begin{aligned}
f_{M|\mathbf{X}=\mathbf{x}}(\mu) &\propto \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2\right)\right) \\
&\propto \exp\left(-\frac{n}{2}\left(\frac{\sum_{i=1}^{n} x_i^2}{n} - \frac{2\mu\sum_{i=1}^{n} x_i}{n} + \mu^2\right)\right) \\
&\propto \exp\left(-\frac{n}{2}\left(-\frac{2\mu\sum_{i=1}^{n} x_i}{n} + \mu^2\right)\right) \\
&\propto \exp\left(-\frac{n}{2}\left(\mu - \frac{\sum_{i=1}^{n} x_i}{n}\right)^2\right) \\
&\propto \frac{1}{\sqrt{2\pi\frac{1}{n}}} \exp\left(-\frac{1}{2}\frac{\left(\mu - \frac{\sum_{i=1}^{n} x_i}{n}\right)^2}{\frac{1}{n}}\right),
\end{aligned}$$

which is a normal distribution with mean $\frac{\sum_{i=1}^{n} x_i}{n}$ and variance $\frac{1}{n}$. Interestingly, this posterior distribution is proper even though the prior distribution was improper.

Special care is needed when dealing with improper priors. Indeed, if one can derive the posterior distribution in closed form and show that it is proper—like in Example 9.2.5—it should not be a concern. On the other hand, in cases where the posterior distribution cannot be obtained in closed form, there is no assurance that the posterior will be proper and extra attention is required.

Choice of the Prior Distribution

The selection of a prior in Bayesian statistics is a crucial step that reflects the experimenter's prior beliefs, knowledge, or assumptions about the parameters of interest. There are different approaches to selecting priors.

1. Informative priors are generally based on the experimenter's subjective beliefs, knowledge, or experience. For instance, one might have a subjective belief that a parameter is likely to fall within a certain range, and this belief is formalized as an informative prior distribution.
2. Noninformative priors are chosen to be minimally informative, expressing little or no prior information about the parameters. For example, uniform priors are commonly used as noninformative priors, expressing a lack of prior preference for any particular parameter value.
3. Empirical Bayes priors rely on the data itself, combining empirical information with Bayesian methodology. This can be done by estimating a prior distribution hyperparameter by using the observed data to inform the prior distribution.
4. Priors that rely on expert elicitation involves seeking input from domain experts to inform the prior. For instance, the experimenter might have additional knowledge about the problem at hand and use a prior distribution that represents their beliefs about the parameters.

Prior Sensitivity Analysis

Prior sensitivity analysis is an important step in Bayesian modeling processes. It refers to the examination and evaluation of the impact of different prior assumptions on the results of a statistical analysis. In other words, such analyses aim to verify the robustness of the conclusions drawn from Bayesian inference to the choice of the prior distribution. By exploring a range of plausible prior distributions, experimenters can gain insights into how much the choice of prior influences the final results and whether those conclusions remain consistent under different prior assumptions.

For instance, prior distributions may significantly influence the posterior estimates, leading to different conclusions. Some of these might be subjective (i.e., informative priors) or based on expert knowledge, and assessing the impact of such assumptions promotes transparency and objectivity in the analysis.

9.3 Conjugate Families

In Section 9.3, you learn how to:

- Describe three specific classes of conjugate families.
- Use conjugate distributions to determine posterior distributions of parameters.
- Understand the pros and cons of conjugate family models.

In Bayesian statistics, if a posterior distribution comes from the same distribution as the prior distribution, the prior and posterior are called conjugate distributions. Note that both posterior and prior have similar shapes but will have different parameters, generally speaking.

But Why? Two main reasons explain why conjugate families have been so popular historically:

1. They are easy to use from a computational standpoint: posterior distributions in most conjugate families can be obtained in closed form, making this class of models easy to use even if we do not have access to computing power.
2. They tend to be easy to interpret: posterior distributions are compromises between data and prior distributions. Having both prior and posterior distributions in the same family—but with different parameters—allows us to understand and quantify how the data changed our initial assumptions.

9.3.1 The Beta–Binomial Conjugate Family

The first conjugate family that we investigate in this book is the beta–binomial family. Let $\mathbf{X} = (X_1, X_2, ...X_m)$ represent a sample of iid Bernoulli random variables such that

$$X_i = \begin{cases} 1 & \text{if success} \\ 0 & \text{if failure} \end{cases},$$

with probabilities q and $1-q$, respectively. Let us further define $x = \sum_{i=1}^{m} x_i$ the sum of the realized successes.

We know from elementary probability that $X = \sum_{i=1}^{m} X_i$ follows a binomial distribution (i.e., the number of successes x in m Bernoulli trials) with unknown probability of success q in $[0,1]$, similar to the coin tossing case of Example 9.1.1, such that the likelihood function is given by

$$p_{\mathbf{X}|Q=q}(\mathbf{x}) = \binom{m}{x} q^x (1-q)^{m-x}, \quad x \in \{0, 1, ..., m\},$$

where $x = \sum_{i=1}^{m} x_i$. The latter represents our evidence. Then, we combine it with its usual conjugate prior—the beta distribution with parameters a and b. The pdf of the beta distribution is given as follows:

$$f_Q(q) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} q^{a-1}(1-q)^{b-1}, \quad q \in [0,1],$$

where a and b are shape parameters of the beta distribution.[11]

We can now combine the prior distribution—beta—with the likelihood function—binomial—to obtain the posterior distribution.

Proposition 9.3.1. Beta–Binomial Conjugate Family. Consider a sample of m iid Bernoulli experiments $(X_1, X_2, ..., X_m)$ each with success probability q. Further assume that the random variable associated with the success probability, Q, has a prior that is beta with shape parameters a and b. The posterior distribution of Q is therefore given by

$$f_{Q|\mathbf{X}=\mathbf{x}}(q) = \frac{\Gamma(a+b+m)}{\Gamma(a+x)\Gamma(b+m-x)} q^{a+x-1}(1-q)^{b+m-x-1},$$

where $x = \sum_{i=1}^{m} x_i$, which is a beta distribution with shape parameters $a+x$ and $b+m-x$.

[11] Here, we assume that the domain of the beta is $[0,1]$, meaning that $\theta = 1$. For more details, see Chapter 20.

Proof. From Section 9.2.1, we know that

$$f_{Q|\mathbf{X}=\mathbf{x}}(q) = \frac{p_{\mathbf{X}|Q=q}(\mathbf{x})\, f_Q(q)}{p_{\mathbf{X}}(\mathbf{x})} \propto \binom{m}{x} q^x (1-q)^{m-x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} q^{a-1}(1-q)^{b-1}$$
$$\propto q^{a+x-1}(1-q)^{b+m-x-1}.$$

We therefore only need to find the normalizing constant that ensures that the right-hand of the equation above is a density. Interestingly, the right-hand side looks like a beta distribution; specifically,

$$\begin{aligned}
&\int_0^1 q^{a+x-1}(1-q)^{b+m-x-1}\, dq \\
&= \frac{\Gamma(a+x)\Gamma(b+m-x)}{\Gamma(a+b+m)} \int_0^1 \frac{\Gamma(a+b+m)}{\Gamma(a+x)\Gamma(b+m-x)} q^{a+x-1}(1-q)^{b+m-x-1}\, dq \\
&= \frac{\Gamma(a+x)\Gamma(b+m-x)}{\Gamma(a+b+m)},
\end{aligned}$$

and

$$f_{Q|\mathbf{X}=\mathbf{x}}(q) = \frac{\Gamma(a+b+m)}{\Gamma(a+x)\Gamma(b+m-x)} q^{a+x-1}(1-q)^{b+m-x-1}.$$

Parameters Versus Hyperparameters. In this context, a and b are called hyperparameters—parameters of the prior. These are different from parameters of the underlying model (i.e., q in the beta–binomial family). Hyperparameters are typically assumed and determined by the experimenter, whereas the underlying model parameters are random in the Bayesian context.

Example 9.3.1. Actuarial Exam Question. You are given:

- The annual number of claims in Year i for a policyholder has a binomial distribution with pmf

$$p_{X_i|Q=q}(x_i) = \binom{2}{x} q^{x_i}(1-q)^{2-x_i}, \quad x_i \in \{0,1,2\}.$$

- The prior distribution is

$$f_Q(q) = 4q^3, \quad q \in [0,1].$$

The policyholder had one claim in each of Years 1 and 2. Calculate the Bayesian estimate of the expected number of claims in Year 3.[12]

[12] This question is Sample Question 5 of the Society of Actuaries Exam C sample questions.

Example Solution. The likelihood function based on this policyholder's number of claims in Years 1 and 2 is given by:

$$p_{\mathbf{X}|Q=q}(\mathbf{x}) = p_{X_1|Q=q}(1)\, p_{X_2|Q=q}(1) = \binom{2}{1} q^1(1-q)^1 \binom{2}{1} q^1(1-q)^1 \propto q^2(1-q)^2,$$

which is proportional to a binomial pmf with $m = 4$, two successes, and a success probability of q. Because the prior distribution is beta distributed with $a = 4$ and $b = 1$, we know that the posterior distribution of parameter q is given by

$$\begin{aligned} f_{Q|\mathbf{X}=\mathbf{x}}(q) &= \frac{\Gamma(4+1+4)}{\Gamma(4+2)\Gamma(1+4-2)} q^{4+2-1}(1-q)^{1+4-2-1} \\ &= \frac{\Gamma(9)}{\Gamma(6)\Gamma(3)} q^5 (1-q)^2 \\ &= 168 q^5 (1-q)^2, \end{aligned}$$

which is also a beta distribution with shape parameters 6 and 3, respectively.

The expected number of claim in Year 3 is

$$\mathrm{E}\left[\mathrm{E}\left[X_3 \,|\, Q = q\right] | \, X_1, X_2\right] = \mathrm{E}\left[2q \,|\, X_1, X_2\right] = 2\,\mathrm{E}\left[q \,|\, X_1, X_2\right],$$

and $\mathrm{E}\left[q \,|\, X_1, X_2\right]$ is the expected value of the beta distribution, which is given by

$$\mathrm{E}\left[q \,|\, X_1, X_2\right] = \frac{6}{6+3} = \frac{2}{3}.$$

Ultimately, this leads to an expected number of claim in Year 3 of $2\left(\frac{2}{3}\right) = \frac{4}{3}$.

Example 9.3.2. Impact of Beta Prior on Posterior. You wish to investigate the impact of having different beta hyperparameters on the posterior distribution. Assume that the claim frequency for each policy follows a Bernoulli random variable with a probability of q such that

$$p_{X_i|Q=q}(x_i) = q^{x_i}(1-q)^{1-x_i}, \quad x_i \in \{0, 1\},$$

where $q \in [0, 1]$, and consider two different sets of hyperparameters:

- Set 1: $a = 1$ and $b = 10$.
- Set 2: $a = 2$ and $b = 2$.

Figure 9.8 shows the pdf of these two prior distributions. The first prior assumes a small prior mean frequency of $\frac{1}{11}$, whereas the second prior distribution has a mean of $\frac{1}{2}$.

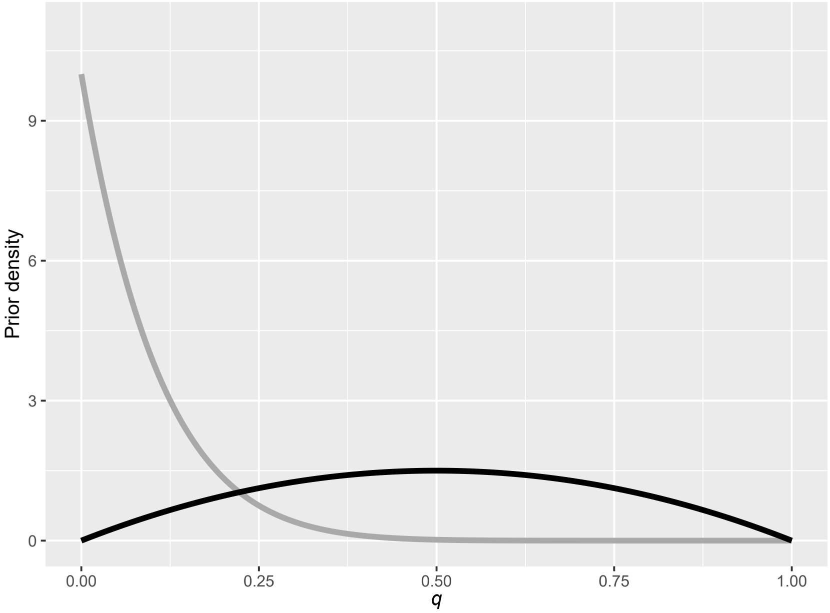


FIGURE 9.8: **Beta prior densities:** $a = 1$ **and** $b = 10$ **(gray), and** $a = 2$ **and** $b = 2$ **(black)**

Using again the first 100 lines of the Singapore insurance dataset (see Example 9.1.4 for more details on this dataset), find the two posterior distributions.

Example Solution. The likelihood function associated with the observations is given by

$$p_{\mathbf{X}|Q=q}(\mathbf{x}) = \binom{100}{x} q^x (1-q)^{100-x}, \quad \text{where } x = \sum_{i=1}^{100} x_i,$$

as mentioned already in Example 9.2.4. Combining this likelihood with a beta prior gives a beta posterior:

$$f_{Q|\mathbf{X}=\mathbf{x}}(q) = \frac{\Gamma(a+b+100)}{\Gamma\left(a+x\right)\,\Gamma\left(b+100-x\right)}\, q^{a+x-1}(1-q)^{b+100-x-1},$$

that can be evaluated for various values of a and b. Figure 9.9 reports the two posterior distributions associated with the priors mentioned above.

```
x <- sum(sgautonb$Clm_Count[1:100])

posterior1 <- dbeta(qs, shape1 = 1 + x, shape2 = 10 + 100 - x)
posterior2 <- dbeta(qs, shape1 = 2 + x, shape2 = 2 + 100 - x)
```

```
dataposterior <- data.frame(x = qs, y1 = posterior1, y2 = posterior2)

ggplot(dataposterior, aes(x = x, y = y1)) + geom_line(color = "darkgray", lwd = 1.5) +
    geom_line(aes(y = y2), color = "black", lwd = 1.5) + xlim(0, 1) + ylim(0, 35) +
    xlab(expression(italic("q"))) + ylab("Posterior density")
```

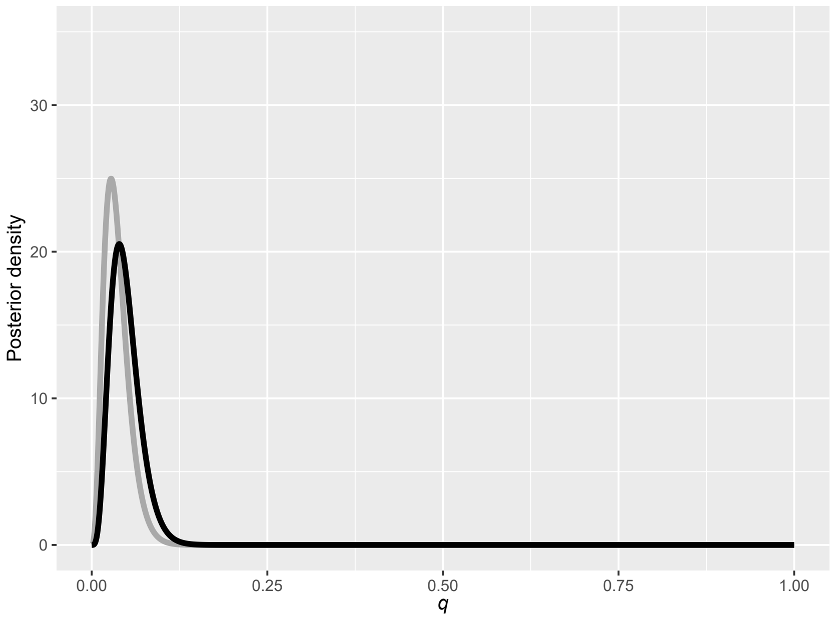


FIGURE 9.9: **Posterior densities based on two different priors:** $a = 1$ **and** $b = 10$ **(gray), and** $a = 2$ **and** $b = 2$ **(black)**

The prior distribution (and its hyperparameters) clearly have an impact on the posterior distribution. As a general rule of thumb for the beta prior, a higher a puts more weight on higher values of q and a higher b puts more weight on lower values of q.

9.3.2 The Gamma–Poisson Conjugate Family

We now present a second conjugate family: the gamma–Poisson family. Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a sample of iid Poisson random variables such that

$$p_{X_i|\Lambda=\lambda}(x_i) = \frac{\lambda^{x_i}\, e^{-\lambda}}{x_i!}, \quad x_i \in \mathbb{R}_+.$$

The likelihood function associated with this sample would therefore be given by

$$f_{\mathbf{X}|\Lambda=\lambda}(\mathbf{x}) = \prod_{i=1}^{n} p_{X_i|\Lambda=\lambda}(x_i) = \prod_{i=1}^{n} \frac{\lambda^{x_i}\, e^{-\lambda}}{x_i!} = \frac{\lambda^{x}\, e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \propto \lambda^{x}\, e^{-n\lambda},$$

where $x = \sum_{i=1}^{n} x_i$. The shape of this likelihood function, as a function of λ, is reminiscent of a gamma distribution, hinting to the fact that this distribution would be a good contender for a conjugate prior. Indeed, if we let the prior distribution be gamma with shape hyperparameter α and scale hyperparameter θ,

$$f_{\Lambda}(\lambda) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} \lambda^{\alpha-1}\, e^{-\frac{\lambda}{\theta}}, \quad \lambda \in \mathbb{R}_{+},$$

we can show that the posterior distribution of λ is also gamma.

Proposition 9.3.2. Gamma–Poisson Conjugate Family. Consider a sample of n iid Poisson experiments $(X_1, X_2, ..., X_n)$, each with rate parameter λ. Further assume that the random variable associated with the rate, Λ, has a prior that is gamma distributed with shape hyperparameter α and scale hyperparameter θ. The posterior distribution of Λ is therefore given by

$$f_{\Lambda|\mathbf{X}=\mathbf{x}}(\lambda) = \frac{1}{\Gamma(\alpha+x)\left(\frac{\theta}{n\theta+1}\right)^{\alpha+x}} \lambda^{\alpha+x-1}\, e^{-\frac{\lambda\,(n\theta+1)}{\theta}},$$

where $x = \sum_{i=1}^{n} x_i$, which is a gamma distribution with shape parameter $\alpha + x$ and scale parameter $\frac{\theta}{n\theta+1}$.

Proof. From Section 9.2.1, we know that

$$\begin{aligned} f_{\Lambda|\mathbf{X}=\mathbf{x}}(\lambda) = \frac{p_{\mathbf{X}|\Lambda=\lambda}(\mathbf{x})\, f_{\Lambda}(\lambda)}{p_{\mathbf{X}}(\mathbf{x})} &\propto \lambda^{x}\, e^{-n\lambda}\, \lambda^{\alpha-1}\, e^{-\frac{\lambda}{\theta}} \\ &\propto \lambda^{\alpha+x-1}\, e^{-\frac{\lambda(n\theta+1)}{\theta}}, \end{aligned}$$

where $x = \sum_{i=1}^{n} x_i$. We therefore only need to find the normalizing constant that ensures that the right-hand of the equation above is a density. Interestingly, the right-hand side looks like a gamma distribution; specifically,

$$\begin{aligned}
&\int_0^\infty \lambda^{\alpha+x-1}\, e^{-\frac{\lambda(n\theta+1)}{\theta}}\, d\lambda \\
=&\Gamma(\alpha+x)\left(\frac{\theta}{n\theta+1}\right)^{\alpha+x} \int_0^\infty \frac{1}{\Gamma(\alpha+x)\left(\frac{\theta}{n\theta+1}\right)^{\alpha+x}}\, \lambda^{\alpha+x-1}\, e^{-\frac{\lambda(n\theta+1)}{\theta}}\, d\lambda \\
=&\Gamma(\alpha+x)\left(\frac{\theta}{n\theta+1}\right)^{\alpha+x},
\end{aligned}$$

and

$$f_{\Lambda|\mathbf{X}=\mathbf{x}}(\lambda) = \frac{1}{\Gamma(\alpha+x)\left(\frac{\theta}{n\theta+1}\right)^{\alpha+x}} \lambda^{\alpha+x-1}\, e^{-\frac{\lambda\,(n\theta+1)}{\theta}}.$$

Example 9.3.3. Actuarial Exam Question. You are given:

- The number of claims incurred in a month by any insured has a Poisson distribution with mean λ.
- The claim frequencies of different insured are iid.
- The prior distribution is gamma with pdf

$$f_\Lambda(\lambda) = \frac{(100\lambda)^6}{120\lambda} e^{-100\lambda}, \quad \lambda \in \mathbb{R}_+.$$

- The number of claims every month is distributed as follows:

Month	Number of Insured	Number of Claims
1	100	6
2	150	8
3	200	11
4	300	?

Calculate the expected number of claims in Month 4.

Example Solution. The likelihood function based on this policyholder's number of claims in Months 1, 2, and 3 is given by:

$$p_{\mathbf{X}|\Lambda=\lambda}(\mathbf{x}) = p_{X_1|\Lambda=\lambda}(6)\, p_{X_2|\Lambda=\lambda}(8)\, p_{X_3|\Lambda=\lambda}(11) \propto \lambda^{6+8+11}\, e^{-\lambda(100+150+200)}.$$

Because the prior distribution is gamma distributed with $\alpha = 6$ and $\theta = \frac{1}{100}$, we know that the posterior distribution of parameter λ is also gamma distributed with shape parameter

$$\alpha + x = 6 + 6 + 8 + 11 = 31$$

and scale parameter

$$\frac{\theta}{n\theta+1} = \frac{\frac{1}{100}}{(100+150+200)\frac{1}{100}+1} = \frac{1}{550}.$$

The expected number of claim in Month 4 conditional on the information of Months 1, 2, and 3 is

$$\mathrm{E}\left[\mathrm{E}\left[X_4 \,|\, \Lambda = \lambda\right] \,|\, X_1, X_2, X_3\right] = \mathrm{E}\left[300\lambda \,|\, X_1, X_2, X_3\right] = 300\,\mathrm{E}\left[\lambda \,|\, X_1, X_2, X_3\right],$$

and $\mathrm{E}\left[\lambda \,|\, X_1, X_2, X_3\right]$ is the expected value of the posterior distribution, which is given by

$$\mathrm{E}\left[\lambda \,|\, X_1, X_2, X_3\right] = \frac{31}{550}.$$

Ultimately, this leads to an expected number of claim in Month 4 of $300\left(\frac{31}{550}\right) = \frac{930}{55} \approx 16.91$.

9.3.3 The Normal–Normal Conjugate Family

The last conjugate family is the normal–normal family. Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a sample of iid normal random variables such that

$$f_{X_i|M=\mu}(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right), \quad x_i \in \mathbb{R}.$$

Further, to keep our focus on μ, we will assume throughout our analysis that the variance parameter σ^2 is known.[13] The likelihood function associated with this sample would therefore be given by

$$\begin{aligned} f_{\mathbf{X}|M=\mu}(\mathbf{x}) &= \prod_{i=1}^{n} f_{X_i|M=\mu}(x_i) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\right) \\ &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\right). \end{aligned}$$

A very natural prior distribution that matches the likelihood structure is unsurprisingly the normal distribution. Let us assume that the prior distribution

[13] Conjugate families for the normal distribution with unknown σ^2 can also be derived. For the sake of simplicity, we will only focus on the case with known variance parameter in this book.

for μ is given by

$$f_M(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2}\frac{(\mu-\theta)^2}{\tau^2}\right),$$

where θ is the mean parameter and τ^2 is the variance parameter. We can then easily show that the posterior distribution of μ is also given by a normal distribution.

Proposition 9.3.3. Normal–Normal Conjugate Family. Consider a sample of n iid normals $(X_1, X_2, ..., X_n)$, each with mean parameter μ and variance parameter σ^2 that is known. Further assume that the random variable associated with the mean, M, has a prior that is normally distributed with mean hyperparameter θ and variance hyperparameter τ^2. The posterior distribution of M is therefore given by

$$f_{M|\mathbf{X}=\mathbf{x}}(\mu) = \frac{1}{\sqrt{2\pi\left(\frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}\right)}} \exp\left(-\frac{1}{2}\frac{\left(\mu - \left(\frac{x}{n}\frac{\tau^2}{n\tau^2+\sigma^2} + \theta\frac{\sigma^2}{n\tau^2+\sigma^2}\right)\right)^2}{\frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}}\right),$$

where $x = \sum_{i=1}^{n} x_i$, which is a normal distribution with mean parameter

$$\frac{x}{n}\frac{n\tau^2}{n\tau^2+\sigma^2} + \theta\frac{\sigma^2}{n\tau^2+\sigma^2}$$

and variance parameter

$$\frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}.$$

Proof. From Section 9.2.1, we know that

$$\begin{aligned}
f_{M|\mathbf{X}=\mathbf{x}}(\mu) &= \frac{f_{\mathbf{X}|M=\mu}(\mathbf{x})\, f_M(\mu)}{f_{\mathbf{X}}(\mathbf{x})} \\
&\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\right)\exp\left(-\frac{1}{2}\frac{(\mu-\theta)^2}{\tau^2}\right) \\
&\propto \exp\left(-\frac{1}{2}\frac{\sum_{i=1}^{n} x_i^2 - 2\mu x + n\mu^2}{\sigma^2} - \frac{1}{2}\frac{\mu^2-2\mu\theta+\theta^2}{\tau^2}\right) \\
&\propto \exp\left(-\frac{1}{2}\frac{n\mu^2-2\mu x}{\sigma^2} - \frac{1}{2}\frac{\mu^2-2\mu\theta}{\tau^2}\right) \\
&\propto \exp\left(-\frac{1}{2}\frac{\mu^2\left(n\tau^2+\sigma^2\right)-2\mu\tau^2 x-2\mu\sigma^2\theta}{\tau^2\sigma^2}\right) \\
&\propto \exp\left(-\frac{1}{2}\frac{\mu^2-2\mu\left(x\frac{\tau^2}{n\tau^2+\sigma^2}+\theta\frac{\sigma^2}{n\tau^2+\sigma^2}\right)}{\frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}}\right) \\
&\propto \frac{1}{\sqrt{2\pi\left(\frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}\right)}}\exp\left(-\frac{1}{2}\frac{\left(\mu-\left(\frac{x}{n}\frac{n\tau^2}{n\tau^2+\sigma^2}+\theta\frac{\sigma^2}{n\tau^2+\sigma^2}\right)\right)^2}{\frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}}\right),
\end{aligned}$$

where $x = \sum_{i=1}^{n} x_i$.

The prior distribution hyperparameters and posterior distribution parameters can be interpreted in the normal–normal conjugate family:

- For the prior, θ represents the *a priori* value of the mean parameter, and τ^2 is related to the precision of that prior mean (i.e., the larger the value, the less precise the prior mean is, and vice versa).
- For the posterior, the new mean parameter is a weighted average between the prior mean parameter θ and the sample mean $\frac{x}{n}$. The new variance parameter is informed by the prior variability τ^2 and the variability of the data σ^2.

Example 9.3.4. Impact of Normal Prior on Posterior. Assume the following observed automobile claims for a small portfolio of policies:

$$1050, \qquad 1250, \qquad 1550, \qquad 2600, \qquad 5350, \qquad 10200.$$

Further assume that the logarithm of the claim amount follows a normal distribution with parameters μ and $\sigma^2 = 1$. Find the posterior distribution of the mean parameter μ for a normal prior distribution where $\theta = 7$. Consider different values of τ^2; that is, $\tau^2 = 0.1$, $\tau^2 = 1$, and $\tau^2 = 10$. Figure 9.10 shows the pdf of these three prior distributions.

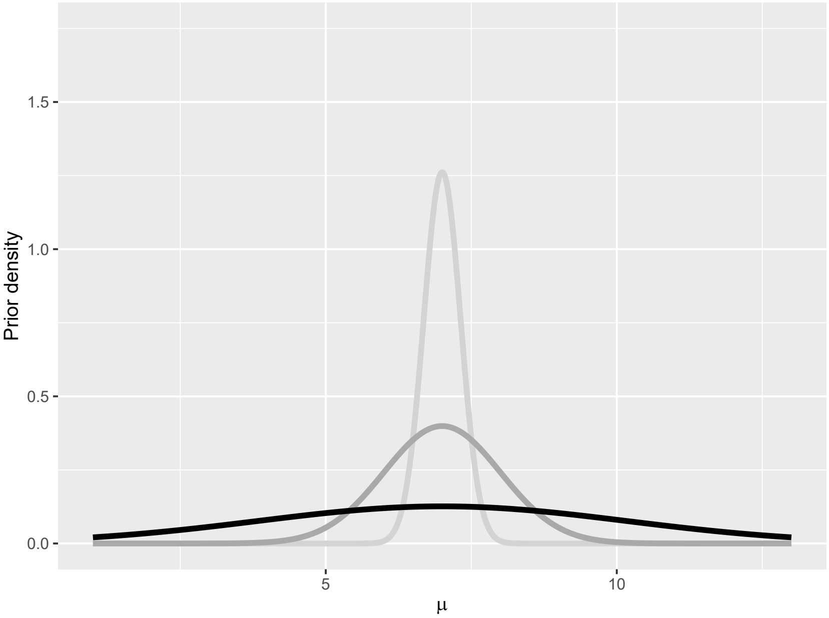


FIGURE 9.10: **Normal prior densities:** $\tau^2 = 0.1$ **(light gray),** $\tau^2 = 1$ **(gray), and** $\tau^2 = 10$ **(black)**

> **Example Solution.** Using the results of Proposition 9.3.3, we can obtain the following posterior distributions:

```
xi <- c(1050, 1250, 1550, 2600, 5350, 10200)
x <- sum(log(xi))
n <- length(xi)
sigma2 <- 1

mean1 <- theta * (sigma2/(n * tau21 + sigma2)) + x/n * ((n * tau21)/(n * tau21 +
    sigma2))
mean2 <- theta * (sigma2/(n * tau22 + sigma2)) + x/n * ((n * tau22)/(n * tau22 +
    sigma2))
mean3 <- theta * (sigma2/(n * tau23 + sigma2)) + x/n * ((n * tau23)/(n * tau23 +
    sigma2))

var1 <- (tau21 * sigma2)/(n * tau21 + sigma2)
var2 <- (tau22 * sigma2)/(n * tau22 + sigma2)
var3 <- (tau23 * sigma2)/(n * tau23 + sigma2)

posterior1 <- dnorm(xs, mean = mean1, sd = sqrt(var1))
posterior2 <- dnorm(xs, mean = mean2, sd = sqrt(var2))
posterior3 <- dnorm(xs, mean = mean3, sd = sqrt(var3))
```

```
dataposterior <- data.frame(x = xs, y1 = posterior1, y2 = posterior2, y3 = posterior3)

ggplot(dataposterior, aes(x = x, y = y1)) + geom_line(color = "lightgray", lwd = 1.5) +
    geom_line(aes(y = y2), color = "darkgray", lwd = 1.5) + geom_line(aes(y = y3),
    color = "black", lwd = 1.5) + xlim(1, 13) + ylim(0, 1.75) + xlab(expression(italic(mu))) +
    ylab("Posterior density")
```

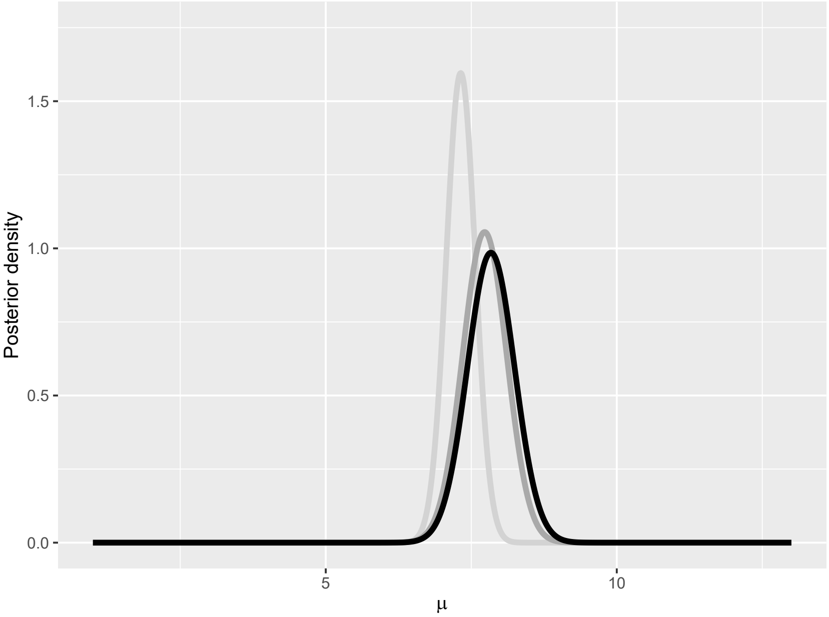


FIGURE 9.11: **Posterior densities based on three different priors:** $\tau^2 = 0.1$ **(light gray),** $\tau^2 = 1$ **(gray), and** $\tau^2 = 10$ **(black)**

Interestingly, as shown in Example 9.3.4, the prior distribution can have some impact on the final posterior distribution. When the prior assumption about the mean is very precise, having a few data points do not create a huge gap between the prior and the posterior (see the light gray curves in Figures 9.10 and 9.11). When the prior is very imprecise, on the other hand, then the data are allow to speak, and the posterior can be quite different from the prior distribution.

9.3.4 Criticism of Conjugate Family Models

While conjugate family models have some advantages, such as ease of interpretation and computational simplicity, they also have some limitations:

1. Conjugate families are oftentimes chosen for their mathematical con-

venience rather than their ability to accurately model the data under study. This can lead to models that are too simplistic and lack the flexibility needed to model real-world phenomena.

2. Conjugate family models rely on the choice of prior distribution, and different choices of possibly non-conjugate priors can lead to very different posterior distributions.
3. Conjugate family models are only applicable to a narrow range of problems, which limit their usefulness in practical applications.

It is important to note that while conjugate family models have their limitations, they can still be useful in certain situations, especially when the assumptions of the model are well understood and the data are relatively simple.

9.4 Posterior Simulation

In Section 9.4, you learn how to:

- Use the standard computational tools for Bayesian statistics.
- Diagnose Markov chain convergence.

9.4.1 Introduction to Markov Chain Monte Carlo Methods

Sometimes, using conjugate family models is ill-suited for the problem at hand, and more complicated priors need to be selected. Under other circumstances, complex models involve many parameters making the posterior distribution intractable. In these cases, the posterior distribution of the parameters will not have a closed-form solution, generally speaking, and will need to be estimated via numerical methods.

A common way to generate draws of the parameter posterior distribution is to create Markov chains for which their stationary distributions—the probability distribution that remains unchanged when the Markov chain has reached a state where the transition probabilities no longer evolve over time—correspond to the posterior of interest. These Markov chain-based methods are known as Markov chain Monte Carlo (MCMC) methods in the literature. This section provides a brief overview of these methods and of their uses. We do not intend to give much of the theory behind these methods, which would require a deep understanding of Markov chains and their theory.[14] Instead, we focus on their applications in insurance and loss modeling. Specifically, in the next two subsections, we introduce the two most common MCMC methods; that is, the Gibbs sampler of Gelfand and Smith (1990) and the Metropolis–Hastings algorithm of Hastings (1970) and Metropolis et al. (1953).

9.4.2 The Gibbs Sampler

As mentioned above, sometimes, we cannot use conjugate families. In other cases where the parameter space is large, it can be very hard to find the marginal likelihood $f_{\mathbf{X}}(\mathbf{x})$ (also known as the normalizing constant); that is, assuming that the model parameters are given by $\boldsymbol{\theta} = [\,\theta_1 \quad ...\theta_2 \quad ... \quad \theta_k\,]$ and contains k parameters, the marginal likelihood given by

$$f_{\mathbf{X}}(\mathbf{x}) = \int\int ... \int f_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathbf{x}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\, d\theta_1\, d\theta_2 ... d\theta_k$$

is hard to compute even when using typical quadrature-based rules, especially if k is large.

Fortunately, under very mild regularity conditions, samples of the joint estimates of parameters can be obtained by sequentially sampling each parameter individually and by keeping all the other parameters constant. To do so, the distribution of any given parameter conditional on all the other parameters (and the data) needs to be known. These distributions are known as full conditional distributions; that is,

$$f_{\Theta_i\,|\,\mathbf{X}=\mathbf{x},\,\boldsymbol{\Theta}_{\setminus i}=\boldsymbol{\theta}_{\setminus i}}(\theta_i),$$

for parameter θ_i, where $\boldsymbol{\theta}_{\setminus i}$ represents all parameters except for the i^{th} one, and $\boldsymbol{\Theta}_{\setminus i}$ is the random variable associated with this set of parameters.

The full conditional distribution is an important building block in Gibbs sampling. Indeed, if one can obtain each parameter's distribution conditional on having the value of all the other parameters in closed form, then it is possible to generate samples for each parameter. Specifically, starting from an arbitrary set of starting values $\boldsymbol{\theta}^{(0)} = [\,\theta_1^{(0)} \quad \theta_2^{(0)} \quad ... \quad \theta_k^{(0)}\,]$, samples for each parameter can be generated by performing the following steps for $m = 1, 2, ..., M$:

1. Draw $\theta_1^{(m)}$ from $f_{\Theta_1\,|\,\mathbf{X}=\mathbf{x},\,\Theta_2=\theta_2^{(m-1)},\,...,\,\Theta_k=\theta_k^{(m-1)}}(\theta_1)$.
2. Draw $\theta_2^{(m)}$ from $f_{\Theta_2\,|\,\mathbf{X}=\mathbf{x},\,\Theta_1=\theta_1^{(m)},\,\Theta_3=\theta_3^{(m-1)},\,...,\,\Theta_k=\theta_k^{(m-1)}}(\theta_2)$.

[14] For an overview of the theory behind MCMC methods, see Robert and Casella (1999).

3. Draw $\theta_3^{(m)}$ from $f_{\Theta_3\,|\,\mathbf{X}=\mathbf{x},\,\Theta_1=\theta_1^{(m)},\,\Theta_2=\theta_2^{(m)},\,\Theta_4=\theta_4^{(m-1)},\,...,\,\Theta_k=\theta_k^{(m-1)}}(\theta_3)$.

$\vdots$

k. Draw $\theta_k^{(m)}$ from $f_{\Theta_k\,|\,\mathbf{X}=\mathbf{x},\,\Theta_1=\theta_1^{(m)},\,...,\,\Theta_{k-1}=\theta_{k-1}^{(m)}}(\theta_k)$.

The sample, especially at first, will depend on the initial values, $\boldsymbol{\theta}^{(0)}$, and it might take some time until the sampler can get to the stationary distribution. For this reason, in practice, experimenters discard the first M^* iterations to make sure their analysis is not impacted by the choice of initial parameter; this initial period of discarded sample is known as the burn-in period.

The rest of the sample—the remaining $M - M^*$ iterations—is kept to estimate the posterior distribution and any quantities of interest.

Application to Bayesian Linear Regression

In statistics and in its most simple form, a linear regression is an approach for modeling the relationship between a scalar response and an explanatory variable. The former quantity is denoted by x_i for $i \in \{1, ..., n\}$, and the latter quantity is denoted by z_i for $i \in \{1, ..., n\}$ in this chapter. Mathematically, we can write this relationship as

$$x_i = \alpha + \beta z_i + \varepsilon_i,$$

where ε_i is a disturbance term that captures the potential for errors in the linear relationship. This error term is typically assumed to be normally distributed with mean zero and variance σ^2.

In general, the coefficients α and β are unknown and need to be estimated. The experimenter can rely on Bayesian statistics to find out the posterior distribution of the parameters α and β along with that of σ^2. For the rest of the subsection, we investigate a specific application of Gibbs sampling to the context of linear regression.

We begin by computing the likelihood function conditional on the parameter values:

$$\begin{aligned} f_{\mathbf{X}\,|\,A=\alpha,\,B=\beta,\,\Sigma^2=\sigma^2}(\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \alpha - \beta z_i)^2}{2\sigma^2}\right) \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \alpha - \beta z_i)^2}{2\sigma^2}\right), \end{aligned}$$

which is the first building block to construct our posterior distribution.

Then, we need a prior distribution, which could be informative, weakly informative, or noninformative. In this application, we select a prior that allows us to obtain each parameter's full conditional distribution in closed form. Specifically, we use a normal distribution for α and β, and an inverse gamma distribution for σ^2 with shape parameter $\frac{n_\sigma}{2}$ and scale parameter $\frac{\theta_\sigma}{2}$, where

$$\begin{aligned} f_A(\alpha) &= \frac{1}{\sqrt{2\pi\tau_\alpha^2}} \exp\left(-\frac{1}{2}\frac{(\alpha-\theta_\alpha)^2}{\tau_\alpha^2}\right), \\ f_B(\beta) &= \frac{1}{\sqrt{2\pi\tau_\beta^2}} \exp\left(-\frac{1}{2}\frac{(\beta-\theta_\beta)^2}{\tau_\beta^2}\right), \\ f_{\Sigma^2}(\sigma^2) &= \frac{(\theta_\sigma/2)^{n_\sigma/2}}{\Gamma(n_\sigma/2)} \left(\frac{1}{\sigma^2}\right)^{n_\sigma/2+1} \exp\left(-\frac{\theta_\sigma/2}{\sigma^2}\right). \end{aligned}$$

Proposition 9.4.1. Full Conditional Distributions of Bayesian Linear Regression Parameters. Consider a sample of n observations $\mathbf{x} = (x_1, ..., x_n)$ for which

$$x_i = \alpha + \beta z_i + \varepsilon_i,$$

where ε_i is normally distributed with mean zero and variance σ^2. The full conditional distributions of parameters α, β, and σ^2 are given by the following expressions:

$$\begin{aligned} &A \sim \text{Normal}\left(\frac{1}{n}\left(\sum_{i=1}^{n} x_i - \beta z_i\right)\frac{n\tau_\alpha^2}{n\tau_\alpha^2+\sigma^2} + \theta_\alpha\frac{\sigma^2}{n\tau_\alpha^2+\sigma^2}, \frac{\tau_\alpha^2\sigma^2}{n\tau_\alpha^2+\sigma^2}\right), \\ &B \sim \text{Normal}\left(\frac{1}{n}\left(\sum_{i=1}^{n} z_i\left(x_i-\alpha\right)\right)\frac{n\tau_\beta^2}{\tau_\beta^2\sum_{i=1}^n z_i^2+\sigma^2} + \theta_\beta\frac{\sigma^2}{\tau_\beta^2\sum_{i=1}^n z_i^2+\sigma^2}, \frac{\tau_\beta^2\sigma^2}{\tau_\beta^2\sum_{i=1}^n z_i^2+\sigma^2}\right), \\ &\Sigma^2 \sim \text{Inverse Gamma}\left(\frac{n_\sigma+n}{2}, \frac{\theta_\sigma+\sum_{i=1}^n\left(y_i-\alpha-\beta z_i\right)^2}{2}\right), \end{aligned}$$

respectively, assuming the prior distributions mentioned above.

Proof. From Section
refS:Sec92, we know that

$$f_{A,B,\Sigma^2|\mathbf{X}=\mathbf{x}}(\alpha,\beta,\sigma^2) \propto f_{\mathbf{X}|A=\alpha,\,B=\beta,\,\Sigma^2=\sigma^2}(\mathbf{x})\, f_A(\alpha)\, f_B(\beta)\, f_{\Sigma^2}(\sigma^2),$$

which is useful to derive the full conditional distributions of α, β, and σ^2.

Let us begin with α:

$$\begin{aligned}
&f_{A|\mathbf{X}=\mathbf{x},\, B=\beta,\, \Sigma^2=\sigma^2}(\alpha)\\
&\propto \exp\left(-\frac{1}{2}\frac{\sum_{i=1}^n (x_i-\alpha-\beta z_i)^2}{\sigma^2}\right)\exp\left(-\frac{1}{2}\frac{(\alpha-\theta_\alpha)^2}{\tau_\alpha^2}\right)\\
&\propto \exp\left(-\frac{1}{2}\left(\frac{n\alpha^2-2\alpha\sum_{i=1}^n(x_i-\beta z_i)}{\sigma^2}+\frac{\alpha^2-2\alpha\theta_\alpha}{\tau_\alpha^2}\right)\right)\\
&\propto \exp\left(-\frac{1}{2}\left(\frac{\alpha^2-2\alpha\left(\frac{1}{n}\left(\sum_{i=1}^n x_i-\beta z_i\right)\frac{n\tau_\alpha^2}{n\tau_\alpha^2+\sigma^2}+\theta_\alpha\frac{\sigma^2}{n\tau_\alpha^2+\sigma^2}\right)}{\frac{\tau_\alpha^2\sigma^2}{n\tau_\alpha^2+\sigma^2}}\right)\right)\\
&\propto \frac{1}{\sqrt{2\pi\left(\frac{\tau_\alpha^2\sigma^2}{n\tau_\alpha^2+\sigma^2}\right)}}\exp\left(-\frac{1}{2}\left(\frac{\left(\alpha-\left(\frac{1}{n}\left(\sum_{i=1}^n x_i-\beta z_i\right)\frac{n\tau_\alpha^2}{n\tau_\alpha^2+\sigma^2}+\theta_\alpha\frac{\sigma^2}{n\tau_\alpha^2+\sigma^2}\right)\right)^2}{\frac{\tau_\alpha^2\sigma^2}{n\tau_\alpha^2+\sigma^2}}\right)\right)
\end{aligned}$$

which is a normal distribution with mean parameter

$$\frac{1}{n}\left(\sum_{i=1}^n x_i-\beta z_i\right)\frac{n\tau_\alpha^2}{n\tau_\alpha^2+\sigma^2}+\theta_\alpha\frac{\sigma^2}{n\tau_\alpha^2+\sigma^2}$$

and variance parameter

$$\frac{\tau_\alpha^2\sigma^2}{n\tau_\alpha^2+\sigma^2}.$$

The derivation to obtain the full conditional distribution of β is similar to that of α:

$$
\begin{aligned}
&f_{B|\mathbf{X}=\mathbf{x},\,A=\alpha,\,\Sigma^2=\sigma^2}(\beta)\\
&\propto \exp\left(-\frac{1}{2}\frac{\sum_{i=1}^{n}\left(x_i-\alpha-\beta z_i\right)^2}{\sigma^2}\right)\exp\left(-\frac{1}{2}\frac{\left(\beta-\theta_\beta\right)^2}{\tau_\beta^2}\right)\\
&\propto \exp\left(-\frac{1}{2}\left(\frac{\beta^2\sum_{i=1}^{n}z_i^2-2\beta\sum_{i=1}^{n}z_i(x_i-\alpha)}{\sigma^2}+\frac{\beta^2-2\beta\theta_\beta}{\tau_\beta^2}\right)\right)\\
&\propto \exp\left(-\frac{1}{2}\left(\frac{\beta^2-2\beta\left(\frac{1}{n}\left(\sum_{i=1}^{n}z_i\left(x_i-\alpha\right)\right)\frac{n\tau_\beta^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}+\theta_\beta\frac{\sigma^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}\right)}{\frac{\sigma_\beta^2\sigma^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}}\right)\right)\\
&\propto \frac{1}{\sqrt{2\pi\left(\frac{\tau_\beta^2\sigma^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}\right)}}\\
&\quad\times\exp\left(-\frac{1}{2}\left(\frac{\left(\beta-\left(\frac{1}{n}\left(\sum_{i=1}^{n}z_i\left(x_i-\alpha\right)\right)\frac{n\tau_\beta^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}+\theta_\beta\frac{\sigma^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}\right)\right)^2}{\frac{\tau_\alpha^2\sigma^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}}\right)\right)
\end{aligned}
$$

which is a normal distribution with mean parameter

$$
\frac{1}{n}\left(\sum_{i=1}^{n}z_i\left(x_i-\alpha\right)\right)\frac{n\tau_\beta^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}+\theta_\beta\frac{\sigma^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}
$$

and variance parameter

$$
\frac{\tau_\beta^2\sigma^2}{\tau_\beta^2\sum_{i=1}^{n}z_i^2+\sigma^2}.
$$

Finally, we apply the same logic to the variance parameter, σ^2:

$$
\begin{aligned}
&f_{\Sigma^2|\mathbf{X}=\mathbf{x},\,A=\alpha,\,B=\beta}(\sigma^2)\\
&\propto \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2}\frac{\sum_{i=1}^{n}\left(x_i-\alpha-\beta z_i\right)^2}{\sigma^2}\right)\left(\frac{1}{\sigma^2}\right)^{n_\sigma/2+1}\exp\left(-\frac{\theta_\sigma/2}{\sigma^2}\right)\\
&\propto \exp\left(-\frac{1}{2}\frac{\theta_\sigma+\sum_{i=1}^{n}\left(x_i-\alpha-\beta z_i\right)^2}{\sigma^2}\right)\left(\frac{1}{\sigma^2}\right)^{(n_\sigma+n)/2+1}\\
&\propto \frac{\left(\frac{\theta_\sigma+\sum_{i=1}^{n}(x_i-\alpha-\beta z_i)^2}{2}\right)^{(n_\sigma+n)/2}}{\Gamma((n_\sigma+n)/2)}\left(\frac{1}{\sigma^2}\right)^{(n_\sigma+n)/2+1}\exp\left(-\frac{\frac{\theta_\sigma+\sum_{i=1}^{n}(x_i-\alpha-\beta z_i)^2}{2}}{\sigma^2}\right),
\end{aligned}
$$

which is an inverse gamma distribution with shape parameter $\frac{n_\sigma + n}{2}$ and scale parameter

$$\frac{\theta_\sigma + \sum_{i=1}^{n} (x_i - \alpha - \beta z_i)^2}{2}.$$

We now apply the Gibbs sampler on *real* data. The example will use motorcycle insurance data from Wasa, a Swedish insurance company, taken from `dataOhlsson` of the R package `insuranceData`; see Wolny-Dominiak and Trzesiok (2014) for more details.

```
library("insuranceData")
data(dataOhlsson)
```

This dataset contains information about the number of motorcycle accidents, their claim cost, and some risk factors (e.g., the age of the driver, the age of the vehicle, the geographic zone).

Example 9.4.1. Bayesian Linear Regression. You wish to understand the relationship between the age of the driver and the (logarithm of the) claim cost. Let x_i be the logarithm of the i^{th} claim cost and z_i be the age associated with the i^{th} claim. Further assume the following linear relationship between the two quantities:

$$x_i = \alpha + \beta z_i + \varepsilon_i,$$

where ε_i is normally distributed with mean zero and variance σ^2. Find the posterior density of the three parameters α, β, and σ^2 using the Gibbs sampler.

Example Solution. Let us begin by visualizing the data. Figure 9.12 reports the logarithm of the claim cost as a function of the driver's age. At first sight, it seems that the relationship between the claim cost and age is negative, so we should expect a negative β, generally speaking.

Let us now turn to Bayesian computation via Gibbs sampling to find the posterior distribution of the three parameters of interest. We will use 10,000 iterations and discard the first 5,000 iterations (i.e., burn-in period). For our prior distributions, we use weakly informative priors by setting $\theta_\alpha = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$, $\theta_\beta = 0$, $\tau_\alpha^2 = \tau_\beta^2 = 10$, $n_\sigma = 1$, and $\theta_\sigma = 0.1$. The initial values of the parameters are set to: $\alpha^{(0)} = \overline{x}$, $\beta^{(0)} = 0$, and $\sigma^{2\,(0)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

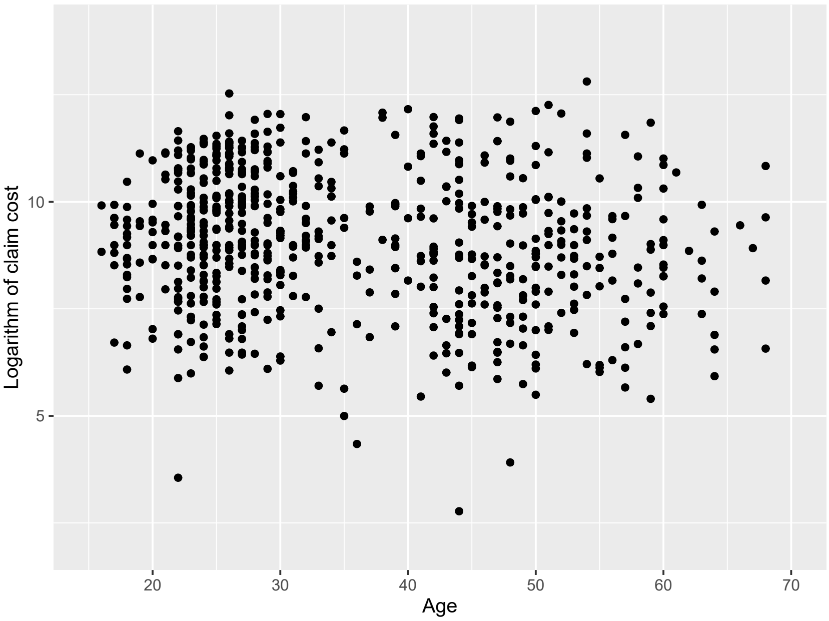


FIGURE 9.12: **Logarithm of the claim cost as a function of the driver's age**

```
set.seed(1)
library("nimble")
dataOhlsson <- dataOhlsson[dataOhlsson$skadkost > 0, ]
dataOhlsson$logskadkost <- log(dataOhlsson$skadkost)

x <- dataOhlsson$logskadkost
z <- dataOhlsson$agarald

n <- length(x)
M <- 10000
Mstar <- 5000
thetaa <- mean(x)
tau2a <- 10
thetab <- 0
tau2b <- 10
nsigma <- 1
thetasigma <- 0.1

alphas <- rep(NA, M + 1)
betas <- rep(NA, M + 1)
sigma2s <- rep(NA, M + 1)

alphas[1] <- mean(x)
betas[1] <- 0
sigma2s[1] <- var(x)

for (m in 2:(M + 1)) {
    # Generate alpha
    den_alpha <- n * tau2a + sigma2s[m - 1]
    mean_alpha <- (1/n) * (sum(x - betas[m - 1] * z)) * (n * tau2a)/den_alpha + thetaa *
        sigma2s[m - 1]/den_alpha
    var_alpha <- tau2a * sigma2s[m - 1]/den_alpha

    alphas[m] <- rnorm(1, mean = mean_alpha, sd = sqrt(var_alpha))

    # Generate beta
    den_beta <- tau2b * sum(z^2) + sigma2s[m - 1]
    mean_beta <- (1/n) * (sum(z * (x - alphas[m]))) * (n * tau2b)/den_beta + thetab *
        sigma2s[m - 1]/den_beta
    var_beta <- tau2b * sigma2s[m - 1]/den_beta

    betas[m] <- rnorm(1, mean = mean_beta, sd = sqrt(var_beta))

    # Generate sigma^2
    shape_sigma <- (nsigma + n)/2
    scale_sigma <- (thetasigma + sum((x - alphas[m] - betas[m] * z)^2))/2

    sigma2s[m] <- rinvgamma(1, shape = shape_sigma, scale = scale_sigma)
}
```

Once we have the posterior parameter samples, we can get multiple quantities of interest. For instance, the posterior mean of parameters α, β, and σ^2 are 9.843, -0.0208, and 2.551, respectively. These posterior means are obtained by simply taking the sample means of the respective posterior draws; that is, these are Monte Carlo estimates of the posterior means.

```
The posterior mean for coefficient alpha is 9.84259435

The posterior mean for coefficient beta is -0.0207847772

The posterior mean for the variance parameter is 2.55090494
```

We can also get histograms of the posterior distribution for α, β, and σ^2; Figure 9.13 reports histograms for the three parameters. The uncertainty around each parameter is very small.

The top panel of Figure 9.14 reports a plot of the post-burn-in values of α as a function of the iteration number; this type of plot is known as a trace plot in the literature. These samples are not impacted by the initial parameter value that was selected. Indeed, after about 20–30 iterations, the posterior parameter values obtained by the Gibbs sampler are very close to their posterior means. For instance, the bottom panel of Figure

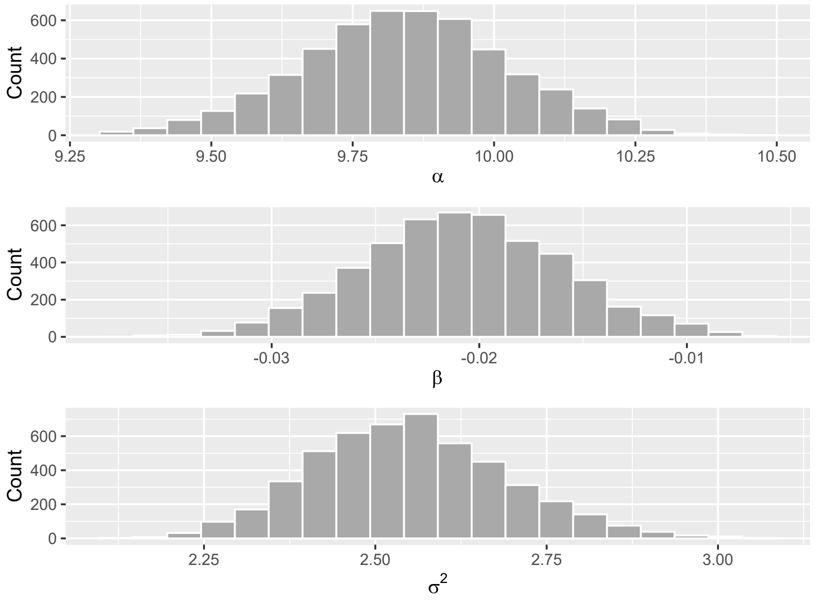


FIGURE 9.13: **Histogram of the posterior distribution for parameters α (top panel), β (middle panel), and σ^2 (bottom panel)**

reffig:Fig914 shows a plot of the first 50 values of α as a function of the iteration number.

9.4.3 The Metropolis–Hastings Algorithm

Gibbs sampling works well when the full conditional distribution for each parameter in the model can be found and is of a common form. This, unfortunately, is not always possible, meaning that we need to rely on other computational tools to find the posterior distribution of the parameters. One very popular method that copes with the shortcomings of Gibbs' method is the Metropolis–Hastings sampler.

Let us assume that the current value of the first model parameter is $\theta_1^{(0)}$. From this current value, we now wish to find a new value for this parameter. To do so, we propose a new value for this parameter, θ_1^*, from a candidate (or proposal) density $q\left(\theta_1^* \,\middle|\, \theta_1^{(0)}\right)$. Since this proposal has nothing to do with the posterior distribution of the parameter, we should not keep all candidates in our final sample—we only accept those samples that are representative of the posterior distribution of interest. To determine whether we accept or reject

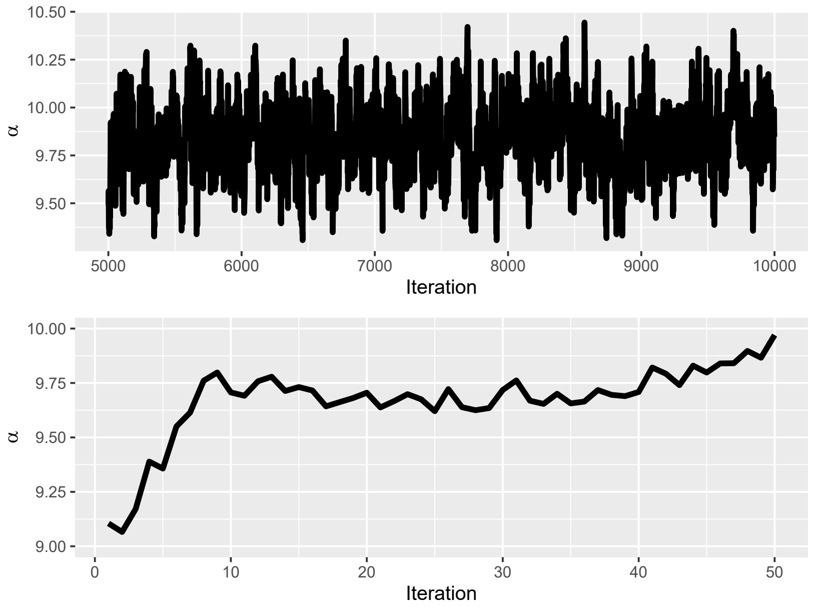


FIGURE 9.14: **Trace plot of α for the post-burn-in iterations (top panel) and for the first 50 iterations (bottom panel)**

the candidate, we compute a so-called acceptance ratio $\alpha\left(\theta_1^{(0)}, \theta_1^*\right)$ using

$$\alpha\left(\theta_1^{(0)}, \theta_1^*\right) = \frac{h\left(\theta_1^*\right) q\left(\theta_1^{(0)} \,\middle|\, \theta_1^*\right)}{h\left(\theta_1^{(1)}\right) q\left(\theta_1^* \,\middle|\, \theta_1^{(0)}\right)}$$

where

$$h(\theta_1) = f_{\mathbf{X} \,|\, \Theta_1 = \theta_1, \boldsymbol{\Theta}_{\setminus 1} = \boldsymbol{\theta}_{\setminus 1}}(\mathbf{x})\, f_{\Theta_1, \boldsymbol{\Theta}_{\setminus 1}}\left(\theta_1, \boldsymbol{\theta}_{\setminus 1}\right)$$

and $\boldsymbol{\theta}_{\setminus 1}$ represents all parameters except for the first one. Then, we accept the proposed value θ_1^* with probability $\alpha\left(\theta_1^{(0)}, \theta_1^*\right)$ and reject it with probability $1 - \alpha\left(\theta_1^{(0)}, \theta_1^*\right)$. Specifically,

$$\theta_1^{(1)} = \begin{cases} \theta_1^* & \text{with probability } \alpha\left(\theta_1^{(0)}, \theta_1^*\right) \\ \theta_1^{(0)} & \text{with probability } 1 - \alpha\left(\theta_1^{(0)}, \theta_1^*\right) \end{cases}$$

We can repeat the same process for all other parameters to obtain $\theta_2^{(1)}$ to $\theta_k^{(1)}$, while replacing the parameters $\boldsymbol{\theta}_{\setminus i}$ by their most current values in the chain. Once we have updated all values, we can repeat this process for all m in $\{2, 3, ..., M\}$, similar to the iterative process used in the Gibbs sampler.[15]

[15] The Gibbs sampler can be seen as a special case of the more general Metropolis–Hastings

Special Case: Symmetric Proposal Distribution. If a proposal distribution is symmetric, then

$$q\left(\theta_i^{(m)} \,\middle|\, \theta_i^*\right) = q\left(\theta_i^* \,\middle|\, \theta_i^{(m)}\right),$$

and those terms cancel out, leaving

$$\alpha\left(\theta_i^{(m)}, \theta_1^*\right) = \frac{h\left(\theta_i^*\right)}{h\left(\theta_i^{(m)}\right)}.$$

This special case is called the *Metropolis algorithm.*

The Metropolis–Hastings sampler requires a lot of fine-tuning, generally speaking, because the experimenter needs to select a proposal distribution for each parameter. A common approach is to assume a normal proposal distribution centered at the previous value; that is,

$$\Theta_i^* \sim \text{Normal}\left(\theta_i^{(m-1)}, \delta_i^2\right),$$

at step m, where δ_i^2 is the variance of the i^{th} parameter's proposal distribution.

Example 9.4.2. Impact of Proposal Density on the Acceptance Rate. Assume that each policyholder's claim count (frequency) is distributed as a Poisson random variable such that

$$p_{N_i \,|\, \Lambda=\lambda}(n_i) = \frac{\lambda^{n_i} e^{-\lambda}}{n_i!},$$

where n_i is the number of claims associated with the i^{th} policyholder. Further assume a noninformative, flat prior over $[0, \infty]$; that is,

$$f_\Lambda(\lambda) \propto 1, \quad \lambda \in [0, \infty].$$

Find the posterior distribution of the parameter using 1,000 iterations of the Metropolis–Hastings sampler assuming the claim count data of the Singapore Insurance Data (see Example 9.1.4 for more details). Use a normal proposal with small (1×10^{-7}), moderate (1×10^{-4}), and large (1×10^{-1}) values as the proposal variance δ in your tests and comment on the differences.

algorithm. Specifically, with Gibbs' method, all proposals are automatically accepted; that is, $\alpha\left(\theta_1^{(0)}, \theta_1^*\right) = 1$.

Example Solution. Starting from the the likelihood function and the prior distribution, we have that

$$h(\lambda) \propto \prod_{i=1}^{N} \frac{\lambda^{n_i} e^{-\lambda}}{n_i!}.$$

You can learn more about the R code for this example at the online version of this book, Actuarial Community (2025).

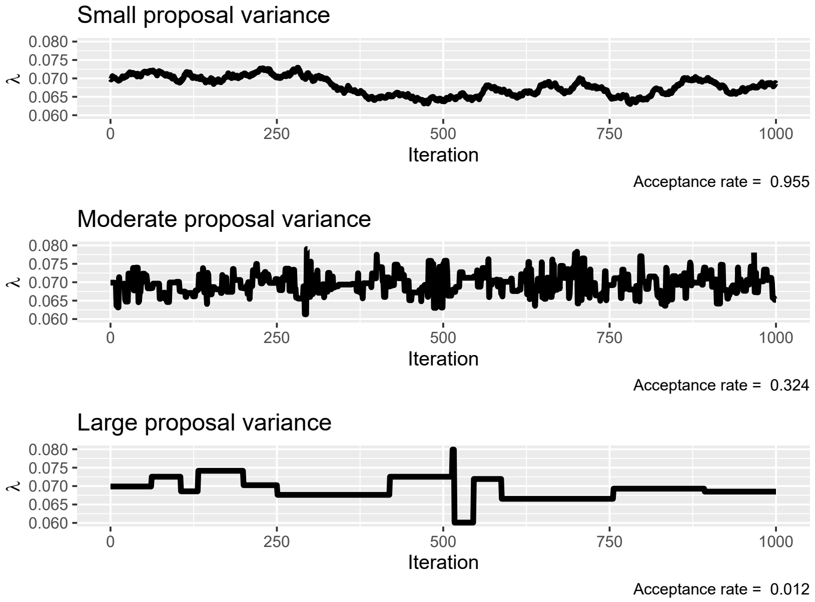


FIGURE 9.15: **Trace plots based on three different proposals: $\sigma^2 = 1 \times 10^{-7}$ (top panel), $\sigma^2 = 1 \times 10^{-4}$ (middle panel), and $\sigma^2 = 1 \times 10^{-1}$ (bottom panel)**

Different variance parameters lead to different results. In this example, if δ^2 is too small, then the experimenter tends to draw samples that are very similar from one iteration to the other. This increases the acceptance rate (i.e., the rate at which we accept the proposal), but also means that the chain is travelling slowly around the posterior distribution. This ultimately imply that it will take longer chains to visit the whole posterior distribution. One way to see this issue in practice is by computing autocorrelation coefficients for the sample of parameter (more details on this in Section 9.4.4). The top panel of Figure 9.1.5 indeed shows this strong autocorrelation and slow travelling around the posterior distribution.

On the other hand, if δ^2 is too large, then the proposal are seldom accepted, and the chain will tend to stick—exhibiting long period for which the chain stays constant. For instance, the case with large proposal variance above leads to an acceptance rate of 1.2 percent, which is very low. The bottom panel of Figure 9.1.5 reports this issue.

The moderate proposal variance case reports an acceptance rate of 32.4 percent, which is not too high nor too low. The general behavior of this chain resembles that of a hairy caterpillar—a good sign—meaning that the mixing seems adequate and that we accept a decent amount of proposed values.

Finding the right proposal variance values for problems of interest requires some fine-tuning. As a general guideline, experimenters should target acceptance rates between 20 and 50 percent.

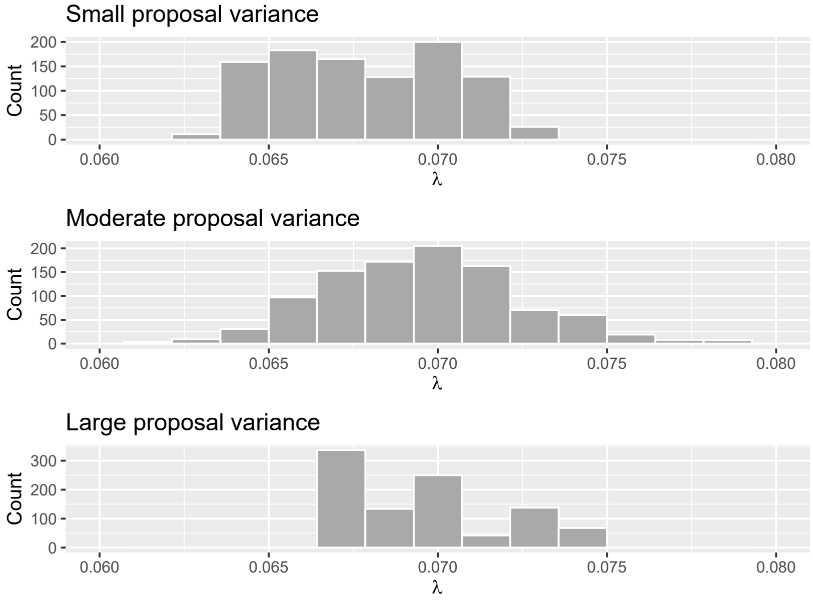


FIGURE 9.16: **Posterior densities based on three different proposals: $\sigma^2 = 1 \times 10^{-7}$ (top panel), $\sigma^2 = 1 \times 10^{-4}$ (middle panel), and $\sigma^2 = 1 \times 10^{-1}$ (bottom panel)**

Using the wrong proposal distribution can have an impact on the computational efficiency of the Metropolis–Hastings algorithm, as shown in Figure 9.16. A small variance takes a long time to travel throughout the posterior distribution, whereas a large variance tends to stick.

Example 9.4.3. Impact of Initial Parameters. Consider the motorcycle insurance data from Wasa used in Example 9.4.1. We wish to model the claim amount from motorcycle losses with a gamma distribution; that is,

$$f_{X_i \,|\, \Theta=\theta,\, A=\alpha}(x_i) = \frac{1}{\theta^{\alpha}\Gamma(\alpha)} x_i^{\alpha-1} e^{-\frac{x_i}{\theta}},$$

where x_i is the i^{th} claim amount. We assume that the prior distributions for both θ and α are noninformative and flat; that is,

$$f_{\Theta,A}(\theta,\alpha) \propto 1, \quad \theta \in [0,\infty], \quad \alpha \in [0,\infty].$$

Find the posterior distribution of the parameter using 1,000 iterations of the Metropolis–Hastings sampler. Use a normal proposal with a proposal variance 5×10^7 for θ and 1×10^{-2} for α, and rely on $\theta^{(0)} = 50,000$ and $\alpha^{(0)} = 0.5$ to start the Metropolis–Hastings sampler. Redo the experiment with $\theta^{(0)} = 10,000$ and $\alpha^{(0)} = 2.5$.

Example Solution. Starting from the the likelihood function and the prior distribution, we have that

$$h(\theta,\alpha) \propto \prod_{i=1}^{N} \frac{1}{\theta^{\alpha}\Gamma(\alpha)} x_i^{\alpha-1} e^{-\frac{x_i}{\theta}}.$$

You can learn more about the R code for this example at the online version of this book, Actuarial Community (2025).

Clearly, from Figure 9.17, the initial parameter value matters: for the first set, the starting value is close to the posterior mode, meaning that the final sample does not depend much on the starting value. For the second set, on the other hand, it takes about 200 iterations to get closer to where most of the density resides. Having a burn-in in the case of Metropolis–Hastings sampler is therefore a good idea to reduce the impact of initial guesses on the final posterior distribution.

In the next subsection, we learn a few methods and metrics to diagnose the convergence of the Markov chains generated via MCMC methods.

9.4.4 Markov Chain Diagnostics

There are many different tuning parameters in MCMC schemes, and they all have an impact on the convergence of the Markov chains generated by these methods. To understand the impact of these choices on the chains (e.g., number of iterations, length of burn-in, proposal distribution), we introduce a few methods to analyze their convergence.

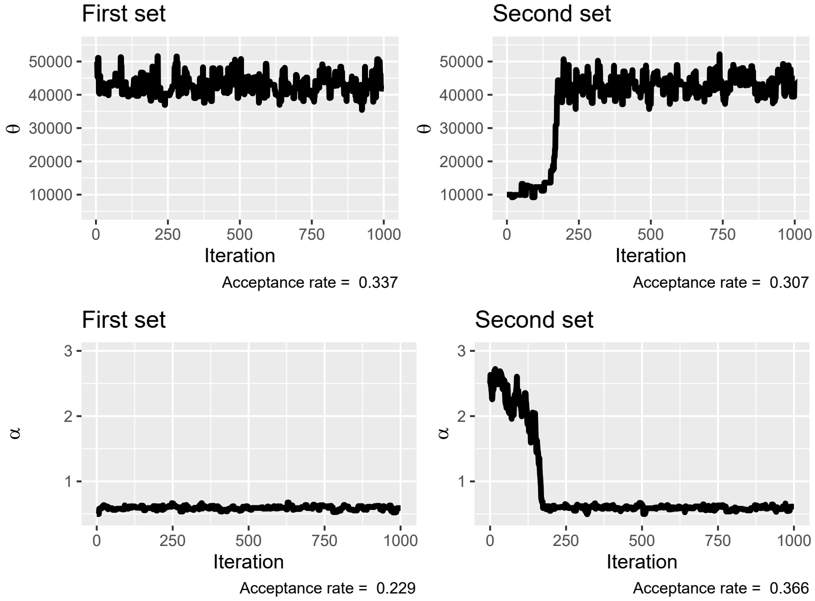


FIGURE 9.17: **Trace plots based on two different starting parameter sets:** $\theta^{(0)} = 50,000$ **and** $\alpha^{(0)} = 0.5$ **(left panels), and** $\theta^{(0)} = 10,000$ **and** $\alpha^{(0)} = 2.5$ **(right panels)**

Examining Trace Plots and Autocorrelation

Trace Plot. The most elementary tool to assess whether MCMC chains have converged to the posterior distribution is the trace plot. As mentioned above, a trace plot displays the sequence of samples as a function of the iteration number, with the sample value on the y-axis and the iteration number on the x-axis. If the chain has converged, the trace plot should show a stable sequence of samples around the true posterior distribution that looks like a hairy caterpillar. However, if the chain has not yet converged, the trace plot may show a sequence of samples that still appear to be changing or have not yet settled into a stable pattern.

In addition to assessing convergence, trace plots can also be used to diagnose potential problems with MCMC algorithms, such as poor mixing or autocorrelation. For example, if the trace plot shows long periods of no change followed by abrupt jumps, this may indicate poor mixing and suggest that the MCMC algorithm needs to be adjusted or a different method should be used.

Lag-1 Autocorrelation. Another quantity that might be helpful is the lag-1 autocorrelation—the correlation between consecutive samples in a given chain:

$$\mathrm{Cov}\left[\theta_i^{(m)}, \theta_i^{(m-1)}\right].$$

Note that if the autocorrelation is too high, it can indicate that the chain is not mixing well and is not sampling the posterior distribution effectively. This can result in poor convergence, longer run times, and decreased precision of the estimates obtained from the MCMC algorithm.

In addition to examining trace plots and computing autocorrelation coefficients, we can use other, more formal tools to evaluate whether the chains obtained are reliable and have converged.

Comparing Parallel Chains

Gelman–Rubin Statistic Another way to assess convergence is to run multiple chains in parallel from different starting points and check if their behavior is similar. In addition to comparing their trace plots, the chains can be compared by using a statistical test—the Gelman–Rubin test of Gelman and Rubin (1992). The latter test compares the within-chain variance to the between-chain variance; to calculate the statistic, we need to generate a small number of chains (say, R), each for $M - M^*$ post-burn-in iterations.

If the chains have converged, the within-chain variance should be similar to the between-chain variance. Assuming the parameter of interest is θ_i, the within-chain variance is

$$W = \frac{1}{R(M - M^* - 1)} \sum_{r=1}^{R} \sum_{m=M^*+1}^{M} \left(\theta_{i,r}^{(m)} - \overline{\theta}_{i,r}\right)^2 ,$$

where $\theta_{i,r}^{(m)}$ is the m^{th} draw of θ_i in the r^{th} chain and $\overline{\theta}_{i,r}$ is the sample mean of θ_i for the r^{th} chain. The between-chain variance is given by

$$B = \frac{M - M^*}{R - 1} \sum_{r=1}^{R} \left(\overline{\theta}_{i,r} - \overline{\theta}_i\right) ,$$

where $\overline{\theta}_i$ is the overall sample mean of θ_i from all chains. The Gelman–Rubin statistic is

$$\sqrt{\left(\frac{M - M^* - 1}{M - M^*} + \frac{R + 1}{R(M - M^*)} \frac{B}{W}\right) \frac{\text{df}}{\text{df} - 2}},$$

where df is the degrees of freedom from Student's t-distribution that approximates the posterior distribution. The statistic should produce a value close to 1 if the chain has converged. On the other hand, if the statistic value is greater than 1.1 or 1.2, this indicates that the chains may not have converged, and further analysis may be needed to determine why the chains are not mixing well.

Calculating Effective Sample Sizes

Effective Sample Size. The effective sample size (ESS) is a measure of the number of independent samples obtained from an MCMC chain. Recall that in an MCMC chain, each sample is correlated with the previous sample; as a result, the effective number of independent samples is usually much smaller than the total number of samples generated by the MCMC algorithm. The ESS takes this correlation into account and provides an estimate of the number of independent samples that are equivalent to the correlated samples in the chain.

In general, a higher effective sample size indicates that the MCMC algorithm has produced more independent samples and is more likely to have accurately sampled the posterior distribution. A lower effective sample size, on the other hand, suggests that the MCMC algorithm may require further tuning or optimization to produce reliable posterior estimates.

The function `multiESS`of the R package `mcmcse` contains a function that gives the ESS of a multivariate Markov chain as described in Vats et al. (2019). The package also includes an estimate of the minimum ESS required for a specified relative tolerance level (see function `minESS`).

We now apply these various diagnostics to an example.

Example 9.4.4. Markov Chain Diagnostics. Consider the setup of Example 9.4.2. Using chains of 51,000 iterations and a burn-in of 1,000 iterations, calculate the various Markov chain diagnostics mentioned above.

You can learn more about the R code for this example at the online version of this book, Actuarial Community (2025).

Example Solution. Let us begin by generating five chains.

Figure 9.18 reports the trace plot for the first chain: it indeed looks like a hairy caterpillar, which is a good sign.

```
The lag-1 autocorrelation coefficient is 0.651510865
```

The autocorrelation is also mild at 65%, again pointing towards good convergence behavior.

```
The Gelman-Rubin statistic is 1.00012696
```

The Gelman–Rubin statistic is very close to 1 in this case, meaning that the chains converged.

```
The ESS is 9927.29934 and the minimum ESS is 6146
```

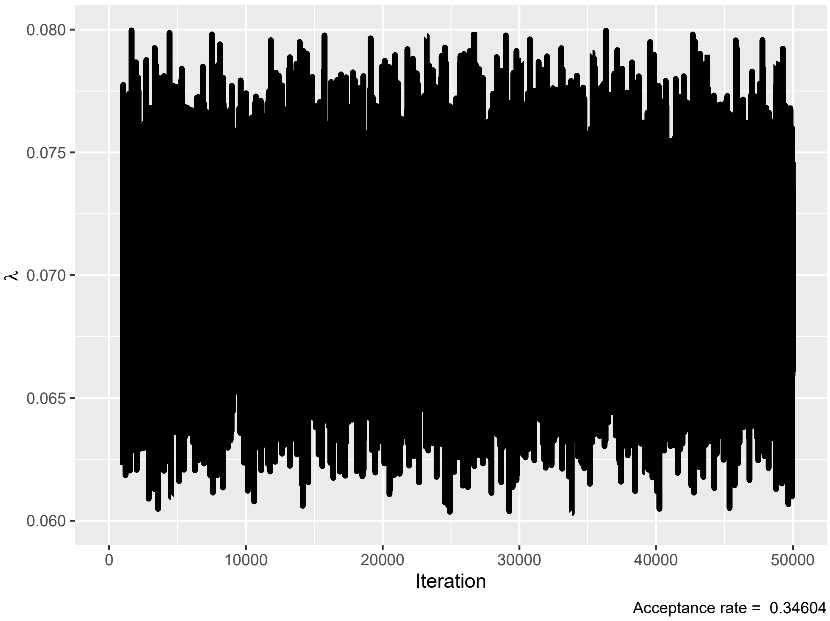


FIGURE 9.18: **Trace plot for parameter λ**

The last diagnostic refers to the ESS, and its comparison to the minimum ESS. In our case, the ESS is about 9,927, and the minimum ESS is 6,146. Since our ESS is above the minimum, we know we have a large enough sample to adequately capture the posterior distribution of λ.

9.5 Bayesian Statistics in Practice

In Section 9.5, you learn how to:

- Describe the main computing resources available for Bayesian statistics and modeling.
- Apply one of them to loss data.

Fortunately for end users, some of these methods are readily available in R, meaning that they are quite accessible. Some popular computing resources used in Bayesian statistics are listed below:

- `RSTAN`, named in honor of Stanislaw Ulam, is an R implementation of the widely used STAN probabilistic programming language for Bayesian statistical modeling and inference. It is highly flexible and allows users to define complex statistical models.
- `nimble` stands for Numerical Inference for Bayesian and Likelihood Estimation and is an R package designed for statistical computing and hierarchical modeling. `nimble` provides a high-level programming language that allows users to define complex statistical models with ease.
- `R2OpenBUGS` allows R users to use OpenBUGS, a classic and widely-used software package for Bayesian data analysis. It uses MCMC techniques like Gibbs sampling to obtain samples from the posterior distribution.
- `rjags` is an R implementation of the JAGS (Just Another Gibbs Sampler) program. It is an open-source software that was developed as an extension of BUGS. It provides a platform-independent engine for the BUGS language, allowing for the use of BUGS models in various environments. Like BUGS, `JAGS` is also used for Bayesian analysis through MCMC sampling techniques.

In what follows, we will use the `nimble` package in the context of loss data.

Example 9.5.1. The `nimble` package. Similar to the setup of Example 9.4.2, consider that each policyholder's claim count (frequency) is distributed as a Poisson random variable such that

$$p_{N_i\,|\,\Lambda=\lambda}(n_i) = \frac{\lambda^{n_i} e^{-\lambda}}{n_i!},$$

where n_i is the number of claims associated with the i^{th} policyholder. Unlike the previous example, however, let us assume an inverse gamma prior with a shape parameter of 2 and a scale parameter of 5.[16]

Find the posterior distribution of the parameter by creating a chain of 51,000 iterations and a burn-in of 1,000 iterations using the `nimble` package.

Example Solution. First, we need to define the model using the 'nimble' language. Simply put, the model is comprised of a likelihood density and a prior density. The former links the observations to a Poisson distribution with parameter λ, and the latter states the prior distribution, which is inverse gamma with shape and scale parameters of 2 and 5, respectively.

[16] Note that the inverse gamma prior combined with a Poisson distribution does not generally lead to closed-form posterior densities and thus requires us to use MCMC methods.

```
claimmodel <- nimbleCode({
    for (i in 1:N) {
        # Likelihood
        count[i] ~ dpois(lambda)
    }
    # Prior distribution
    lambda ~ dinvgamma(shape = 2, scale = 5)
})
```

Then, we define the data, the constant (i.e., the number of observations in this case), the parameter list (i.e., only λ here), and the initial value set to 0.05 in this illustration.

```
claimdata <- list(count = sgautonb$Clm_Count)
claimconstant <- list(N = length(sgautonb$Clm_Count))
claimparameters <- c("lambda")
claiminitial <- list(lambda = 0.05)
```

The MCMC chain is then run using for 51,000 iterations and a burn-in of 1,000 iterations.

```
mcmcoutput <- nimbleMCMC(code = claimmodel, data = claimdata, constants = claimconstant,
    inits = claiminitial, monitors = claimparameters, niter = 51000, nburnin = 1000,
    nchains = 1)
save(mcmcoutput, file = "../IntermediateCalcs/BayesChap/Example951.Rdata")
```

Finally, we display the trace plot, obtain the histogram of the posterior distribution of λ, and compute some descriptive statistics of the parameter.

```
The posterior mean of the parameter is 0.0781913941
```

```
The posterior standard deviation of the parameter is 0.00309462947
```

This simple illustration demonstrates the simplicity of utilizing R packages to generate MCMC chains, all without the need for writing extensive code. For more details on the `nimble` package, see Valpine et al. (2017).

9.6 Further Resources and Contributors

Many great books exist on Bayesian statistics and MCMC schemes. We refer the interested reader to Bernardo and Smith (2009) and Robert and Casella (1999) for an advanced treatment of these topics.

A number of academic articles in actuarial science relied on Bayesian statistics and MCMC schemes over the past 40 years; see, for instance, Heckman and

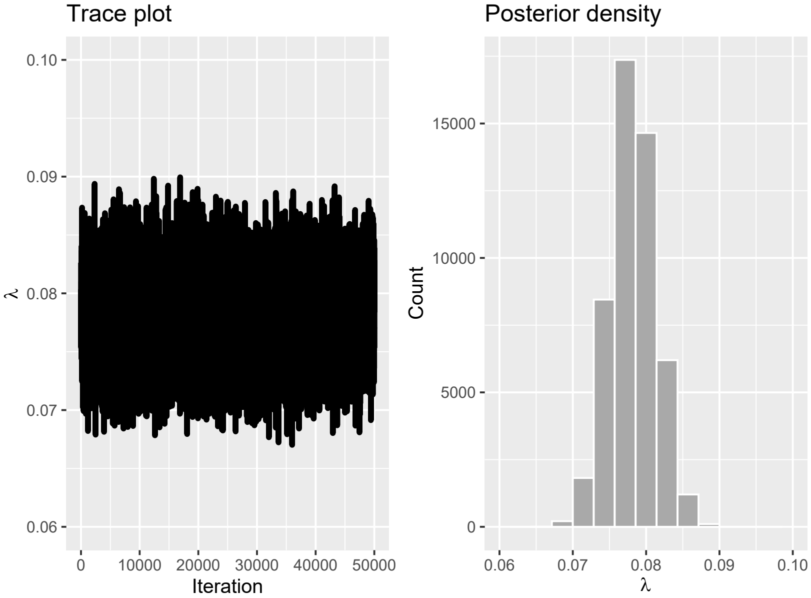


FIGURE 9.19: **Trace plot and posterior density for parameter λ**

Meyers (1983), Meyers and Schenker (1983), Cairns (2000), Cairns et al. (2006), Hartman and Heaton (2011), Bermúdez and Karlis (2011), Hartman and Groendyke (2013), Fellingham et al. (2015), Bignozzi and Tsanakas (2016), Bégin (2019), Huang and Meng (2020), Bégin (2021), Cheung et al. (2021), and Bégin (2023).

Contributors

- **Jean-François Bégin**, Simon Fraser University, is the principal author of the initial version of this chapter. Email: jbegin@sfu.ca for chapter comments and suggested improvements.
- Chapter reviewers include: Brian Hartman, Chun Yong, Margie Rosenberg, and Gary Dean.

10

Premium Foundations

Chapter Preview. Setting prices for insurance products, i.e., premiums, is an important task for actuaries and other data analysts. This chapter introduces the foundations for pricing non-life products.

The presentation of this chapter follows the premium equation.

- In Section 10.2, we first present the sources of information that support premium development.
- We discuss this development of the pure premiums in Section 10.5.
- In Section 10.6, we discuss fixed and variable non-claim expenses.
- In Section 10.7.3, we discuss the provision for profit.
- Section 10.9 summarizes alternative premium principles that incorporate uncertainty into our pricing.

10.1 Introduction to Ratemaking

In this section, you will learn how to:

- Describe relationship between between exposures, rates, and premiums
- Describe the components of the rate

This chapter explains how you can determine the appropriate price for an insurance product. As described in Section 1.2, one of the core actuarial functions is ratemaking, where the analyst seeks to determine the right price for a risk.

A price is the consideration exchanged for a good or service. In insurance, we refer to this consideration as the premium and the service provided by the insurer is protection against contingent events.

The amount of protection will vary by risk being insured. For example, in homeowners insurance, the amount of insurance protection depends on the home value. In life insurance, the amount of protection may depend on a policyholder's financial status (e.g., income and wealth) and their perceived need for financial security. So, it is common to express insurance prices as a unit of the protection being purchased, for example, a price per thousand dollars of coverage on a home or benefit in the event of death. We refer to the unit of protection as the exposure. These prices/premiums are known as rates because they are expressed in standardized units.

Unlike other products, the costs of insurance protection are not known at the sale of the contract. If the insured contingent event, such as an automobile accident of the loss of life, does not occur, then the contract costs are only administrative (e.g., to set up the contract) and are relatively minor. If an insured event occurs, then the cost includes not only administrative costs but also claim payment(s) and expenses to settle claims. So, the cost is random when the contract is written, and protection from that randomness is the basis of insurance.

Because costs are unknown at the time of sale, insurance pricing differs from common economic approaches. This chapter introduces traditional actuarial approaches to determine prices as a function of insurance costs. Insurance involves a promise of the insurer to pay a claim when presented by the insured. For this reason, insurance is a regulated business, particularly for personal lines insurance. The role of the regulator is to ensure that the insurer is able to satisfy its promise to its policyholders. In executing this mandate, the regulator often requires the insurer to file support for its rates. The regulator will review that filing to determine whether those rates are reasonable, not excessive, not inadequate, and not unfairly discriminatory.

The actuarial pricing approach we present is sufficient for some insurance markets, such as personal automobile or homeowners, where the insurer has a portfolio of many similar independent risks. However, there are other insurance markets where actuarial prices only provide an input to general market prices. To reinforce this distinction, actuarial cost-based premiums are sometimes known as technical prices.

To develop technical prices, it is helpful to think of a premium as revenue source that provides for

- Pure Premium - Claim payments are amounts due to the insured under the terms of the insurance contract. Pure premiums include claim payments costs to administer and investigate such claims.
- Insurer expenses - *Non-claim Expenses* include insurer costs that vary by premium (such as sales commissions), and those that do not (such as building

costs and employee salaries). We include those costs through the Fixed Expenses and Variable Expense Rate of the (10.2).
- Profit - An insurer requires capital to support operations. The capital provider will reasonably expect to earn a profit from insuring risk. Insurers have two sources of profit: underwriting income and investment income.

We formalize this relationship in our simplified premium equation.

$$\text{Premium} = \frac{\text{Pure Premiums} + \text{Fixed Expenses}}{1 - \text{Variable Expense Rate} - \text{Profit}} \tag{10.1}$$

where

$$\text{Pure Premiums} = \frac{\text{Estimated Claims and Claims Adjustment Expense}}{\text{Exposures}}$$

or

$$\text{Pure Premiums} = \frac{\text{Estimated Claim Counts}}{\text{Exposures}} \times \frac{\text{Estimated Claims and Claims Adjustment Expense}}{\text{Estimated Claim Counts}}$$

This simplified premium equation promotes a general understanding of the Relationship of insurance costs and revenue. We refer to this equation as the simplified premium equation because (i) it does not include explicit consideration for investment income, and (ii) we combine consideration of claims and claims adjustment expenses. We will refine the simplified premium equation to consider these items later in this chapter.

We observe that the *pure premium* in equation (10.1) is a ratio of claims and exposures. We discuss the development of the claims provision in Section 10.3 and the development of exposures in Section 10.4.

10.2 Data Sources

In this section, you will learn how to:

- Describe the types of data used to develop rates

Insurers consider aggregate information for ratemaking such as exposures, premiums, expenses, claims, and payments. This aggregate information is also useful for managing an insurer's activities. The information is typically summarized in financial reports which are commonly compiled at least annually and often quarterly. At any given financial reporting date, information about recent policies and claims will be ongoing and necessarily incomplete; this section introduces concepts for projecting risk information so that it is useful for ratemaking purposes.

Insurers generally store information about insured risks, such as exposures, premiums, claim counts, losses, and rating factors, in a relational database that will include:

- *policy database* - contains information about the risk being insured, the policyholder, and the contract provisions
- *claims database* - contains information about each claim. The claims database is linked to the policy database.
- *payment database* - contains information on each claims transaction, typically payments but may also include changes to *case reserves*. The payment database is linked to the claims database.

Insurers will aggregate the information in these detailed databases to develop the information needed for financial reports. As described in this chapter, insurers' actuaries will also use this information to develop the premiums.

10.3 Claims

In this section, you will learn how to:

- Describe the basis for the provision for claims, the numerator of the pure premium
- Adjust claims to the level of the prospective period

The terms loss and claim refer to the amount of compensation paid or payable to the claimant under the terms of the insurance policy. Definitions can vary:

- Sometimes, *claim* is used interchangeably with the term *loss*.
- In some insurance and actuarial sources, *loss* is the amount of damage sustained in an insured event. The *claim* is the amount paid by the insurer. Differences between *loss* and *claim* amounts are typically due to the coverage terms such as deductibles and policy limits.
- In economics, a *claim* is a demand for payment by an insured or by an injured third party under the terms and conditions of the insurance contract, and the *loss* is the amount paid by the insurer.

This text will follow the convention of the second bullet.

Also, there are two categories of claim adjustment expenses.

- Allocated claim adjustment expenses are attributed to a specific claim and are generally comprised of investigation and legal expenses to defend or settle the claim.

  Claims and allocated claim adjustment expenses are sometimes inversely correlated, as additional defense expenses may result in lower claim payments. In this section, references to claims also include allocated claims adjustment expenses.

- Unallocated claim adjustment expenses cannot be assigned to individual claims (e.g., claim adjuster salaries.) Actuaries often review claims and allocated claim adjustment expenses in aggregate, as the latter is often a function of the former.

  Insurers will include a provision for unallocated claims adjustment expenses either as a percentage of claims and allocated claims expenses or premiums, or a combination thereof. We discuss unallocated claims adjustment expenses separately in Section 10.3.2.

10.3.1 Estimated Ultimate Claims

Recall that a claim is the amount paid or payable to claimants under the terms of insurance policies. In more detail, one can consider *paid claims*, those losses for a particular period that have actually been paid to claimants. When there is an expectation that payment will be made in the future, a claim will have an associated *case reserve* representing the estimated amount of that payment. Case adjusters establish case reserves separately for each open claim based on the information available. In addition, *reported claims*, also known as *case incurred claims* or *incurred claims*, are the sum of paid claims and case reserves. The *ultimate claim* is the amount required to close and settle all claims for a defined group of policies. We describe the estimation of ultimate claims and claims adjustment expenses in Section 14.

Alternatively, we can estimate projected claims and claims expenses as the product of projected claim frequency and claims severity:

$$\text{Claims and Claims Expenses}_i^{(t)} = E[X] \times E[N]$$

where:

- $E[X]$ is the projected average ultimate severity per claim, and

- $E[N]$ is the projected ultimate number of claims.

We note the frequency-severity alternatives to support our discussion of trends in the following section.

10.3.2 Adjustments to Claims and Allocated Claims Adjustment Expenses

In this section, we review adjustments to experience period ultimate claims that are required to support the development of prospective rates. These adjustments include trending, large loss adjustments and provisions for catastrophes. Finally, we discuss approaches to incorporate unallocated claims adjustment expense.

Trending

Each of the years of the experience period has a different underlying cost level; our goal is to estimate claims at the cost level of the prospective policy period. Consider, for example, if costs were rising at a rate of 5% per annum. All else equal, the estimated ultimate cost for time $t+2$ would be 1.05^2 times the costs of claims from time t. Trending is the process of adjusting ultimate losses from the cost level of the experience period to prospective cost levels.

Actuaries will often consider separate trends for the frequency of claims and the severity of claims. Actuaries often state past trends separately from future trends. Past trends reflect changes that have taken place between the experience period of the rate calculation and the valuation date. Future trends reflect expectations of change between the valuation date and the prospective policy period.

There are various approaches to estimating severity trend rates. Two common approaches include the estimation of trend rates based on external cost indices and the estimation of trend rates based on claims experience. The former approach generally uses government data such as the consumer price index or components thereof. In the latter approach, actuaries will often fit regression models to discern the rate of change in average claims values over time.

Due to the lack of external indices that would be appropriate as a basis for claims frequency models, actuaries generally either estimate frequency trend based on company experience or assume that the frequency trend is 0%.

It is also common to review pure premium trends directly. Although the pure premium trend is effectively a combination of the frequency and severity trends, direct analysis of pure premiums may mask underlying changes in frequency and severity when they are inversely correlated.

Large Loss and Catastrophe Provisions

Consider, for example, if a five-year experience period included a one-in-20-year event. If we did not adjust the data, we would effectively overestimate claim amounts for that category of claims.

In ratemaking, we remove these unusual large losses and catastrophe losses from the experience period data, and then add a provision consistent with the longer-term average cost of large losses or catastrophes. Although large loss adjustments are commonly based on the insurer's claims experience, the provision for catastrophes is often based on models developed by specialists. Adjustments for catastrophes are more common in property insurance, while adjustments for large losses are more common in liability insurance.

Unallocated Claims Adjustment Expenses

Some insurers include unallocated claims adjustment expenses as a percentage of claims and allocated claims adjustment expenses, while other insurers include unallocated claims adjustment expenses as a percentage of premiums. In our discussion, we use the former approach. For insurers that use the latter approach, the inclusion of a provision for unallocated claims adjustment expenses would follow that described below for other non-claim expenses. Generally, insurers estimate UE by reviewing historical ratios of those payments to claims and allocated claims adjustment expense payments.

10.4 Exposures

In this section, you will learn how to:

- Describe the consideration exposures in the developing pure premiums
- Select an exposure base
- Adjust historical exposures to the level of the prospective period

The denominator of the pure premium equation is "exposure." We use exposures to standardize heterogeneous risks. To explain exposures, we can consider *scale distributions* that we learned about in Chapter 4. To recall a scale distribution, suppose that X has a parametric distribution and define a rescaled version $R = X/E$, $E > 0$. If R is in the same parametric family as X, then the distribution is said to be a scale distribution. As we have seen, the gamma, exponential, and Pareto distributions are examples of scale distributions.

Intuitively, the idea behind exposures is to make risks more comparable to one another. For example, it may be that risks $X_1, \ldots, X_n$ are from different distributions and yet, with the choice of the right exposures, the rates $R_1, \ldots, R_n$ are from the same distribution. Here, we interpret the rate $R_i = C_i/E_i$ as the loss divided by exposure.

Table 10.5.1 provides a few examples.

Table 10.5.1. **Commonly used Exposures in Different Types of Insurance**

Type of Insurance	*Exposure Basis*
Personal Automobile	Earned Car Year, Amount of Insurance Coverage
Homeowners	Earned House Year, Amount of Insurance Coverage
Workers Compensation	Payroll
Commercial General Liability	Sales Revenue, Payroll, Square Footage, Number of Units
Commercial Business Property	Amount of Insurance Coverage
Physician's Professional Liability	Number of Physician Years
Professional Liability	Number of Professionals (e.g., Lawyers or Accountants)
Personal Articles Floater	Value of Item

10.4.1 Criteria for Choosing an Exposure

An exposure base should meet the following criteria. It should:

- be an accurate measure of the quantitative exposure to loss
- be easy for the insurer to determine (at the time the policy is initiated) and not subject to manipulation by the insured,
- be easy to understand by the insured and easy to calculate by the insurer,
- consider any preexisting exposure base established within the industry.

To illustrate, consider personal automobile coverage. Instead of the exposure basis "earned car year," a more accurate measure of the quantitative exposure to loss might be number of miles driven. Historically, this measure had been difficult to determine at the time the policy is issued and subject to potential manipulation by the insured, so it was not typically used. Modern telematic devices that allow for accurate mileage recording support the use of this exposure base in some marketplaces.

As another example, the exposure measure in commercial business property, e.g., fire insurance, is typically the amount of insurance coverage. As property values grow with inflation, so will the amount of insurance coverage. Thus, rates quoted on a per amount of insurance coverage are less sensitive to inflation.

10.4.2 Written and Earned Exposures

In developing premiums and rates, it's important that we use claims information and exposure information that is comparable. Most ratemaking uses an accident year approach. In this approach, we relate claims incurred during a specified period to the premium or exposure "earned" during that same period without consideration of the period in which the underlying policy was written. For example, a 12-month policy issued on 1 July 2019 insures claims events in 2019 or 2020, and the claims are assigned to the year of the event. Generally, we earn premiums and exposures on a pro-rata as to time basis as presented in Table 10.5.2, which displays illustrative calculations for a portfolio of four illustrative policies.

Table 10.5.2. **Exposures for Four 12-Month Policies**

								In-Force
	Effective	Written	Exposure	Earned	Exposure	Unearned	Exposure	Exposure
Policy	Date	2019	2020	2019	2020	1/1/2019	1/1/2020	1/1/2020
A	1 Jan 2019	1.00	0.00	1.00	0.00	0.00	0.00	0.00
B	1 April 2019	1.00	0.00	0.75	0.25	0.25	0.00	1.00
C	1 July 2019	1.00	0.00	0.50	0.50	0.50	0.00	1.00
D	1 Oct 2019	1.00	0.00	0.25	0.75	0.75	0.00	1.00
	Total	4.00	0.00	2.50	1.50	1.50	0.00	3.00

10.4.3 Adjustments to Exposures

Exposure Trend

Sometimes exposure units are inflation sensitive. For example, payroll is a common exposure base for workers compensation coverage. Even if the insured firm does not grow, it's payroll may increase due to wage inflation. We refer to the adjustment applied to inflation sensitive exposures as exposure trend.

10.5 Pure Premiums

In this section, you will learn how to:

- Calculate the expected pure premium

The *pure premium* in equation (10.1) is a random variable, and so, as a baseline, we use the *expected costs* to determine rates. To develop our initial understanding, we will consider the insurer that enters into many contracts with risks that are similar except, by pure chance, in some cases, there are losses on some contracts but not on others. The insurer is obligated to pay the total amount of claim payments for all contracts. If the risks are similar, then all policyholders are equally likely to contribute to the total loss. From probability theory, specifically the law of large numbers, we know that the average of iid risks is close to the expected amount, so we use the expectation as a baseline pricing principle.

In this chapter, we present the development of average premium levels for a portfolio of homogeneous risks. In Chapter 11, we present approaches to develop classification plans which adjust those average premiums to recognize various risk characteristics. In Chapter 15, we present approaches to develop premiums that consider the claim experience of an individual insured.

10.5.1 Experience Period

To develop expected pure premiums, actuaries will typically review claims and exposure experience over a multi-year (typically three to seven years) period. The use of a multi-year period smooths the year-to-year randomness. We refer to this multi-year period as the experience period.

10.5.2 Expected Pure Premium

The expected pure premium is generally calculated as the weighted average of the observations in the experience period. The weights balance responsiveness to more recent experience and the stability of a longer-term average.

$$\text{Pure Premium}^{(t)} = \text{Exposure}_t \times \sum_{i=1}^{n} w_i \frac{\text{Ultimate Claims and Claims Expenses}_i^{(t)}}{\text{Exposure}_i^{(t)}}$$

where:

- w_i is the weight for year i in an n year experience period.

The superscript `(t)` indicates that the ultimate claim estimate for accident year i is adjusted to the level of the prospective program period t. We discussed these adjustments in Section 10.3.2. The following equation demonstrates this process of adjustment.

$$\begin{aligned}
\text{Claims and Claims Expenses}_i^{(t)} &= C_i^{\text{x}LL,\text{x}Cat} \times T_i^{(t)} \times LL \times CP \times UE \\
\text{Exposure}_i^{(t)} &= \text{Exposure}_i \times E_i^{(t)}
\end{aligned}$$

where:

- $C_i^{\mathrm{x}LL,\mathrm{x}Cat}$ is the estimated ultimate claims for year i, excluding large losses and catastrophes
- T_i is a claim trend factor to adjust year i experience to the cost level of year t
- $E_i^{(t)}$ is an exposure trend factor to adjust year i experience to the cost level of year t
- LL is a large loss factor
- CP is a catastrophe provision
- UE is the unallocated claims adjustment expense factor

We discussed these adjustments earlier in this chapter.

10.6 Non-Claim Expenses

In this section, you will learn how to:

- Describe the consideration of operational expenses in the development of premiums

Non-claim insurer Operating expense costs include commissions, premium taxes, and other expenses such as salaries, rent, and inspections.

- Some expenses (such as commissions and premium taxes) vary with premiums are "variable" or "premium variable expenses."
- Other expenses (such as general administrative and head office costs) are not proportional to the premium.

For non-claim expenses, insurers will typically rely on either historical expense ratios, budgeted amounts, or financial forecasts.

We include fixed expenses in our premium equation on a per-exposure basis and we include variable expenses as a rate per unit premium.

10.7 Investment Income

In this section, you will learn how to:

- Describe the consideration of the timing of cash flows in the development of the rate
- Calculate a required provision for underwriting profit

A portion of the required profit is earned from investment income from two sources: policyholder cash flows and investment of the insurer's surplus. To the extent that investment income is insufficient to provide the required rate of return, the premiums will also need to include an underwriting profit provision.

As we described, we presented a simplified premium equation in Section 10.5.2 to promote the understanding of the claims and expense provisions in Section 10.3 and exposures in Section 10.6. We now refine the equation to consider investment income. We now consider the other source of an insurer's profit, investment income.

10.7.1 Investment Income on Policyholder Cash Flows

We first consider policyholder cash flows, i.e., premiums, claims, claim adjustment expenses, and non-claim expenses. We consider investment income on policyholder cash flows by discounting each of the cash flows of each of these components of the premium equation.

- There may be a delay in the insurer's receipt of premium, perhaps because the insurer offers payment plans to the insured.
- Claims and claims adjustment expenses are paid over a period that typically extends beyond the policy term. Generally, property coverages have the shortest payment stream, with all claims being settled and paid over a period that extends between 2 and 5 years, depending on the complexity of the determination of damages. Litigated liability coverages will have intermediate payment streams that range from three to 10 years. Finally, coverages such as workers compensation, offer lifetime benefits that can extend forty years or longer.
- Non-claim expenses are generally paid over the term of the policy period.

We can rewrite our premium equation to capture the discounting. We replace the unity in the denominator with a premium delay factor and we discount the claims and claims adjustments expenses (in the pure premium) and non-claim expenses.

$$\text{Premium} = \frac{\text{Discounted Pure Premiums} + \text{Discounted Fixed Expenses}}{\text{Premium Delay} - \text{Variable Expense Rate} - \text{Profit}}.$$

We recognize that the discounting effect on the numerator is significantly greater than the effect on the denominator. As a result, consideration of investment income on policyholder cash flows serves to reduce the premium.

The consideration of profit serves in the denominator serves to increase the required premium. We now turn to the determination of that profit provision.

10.7.2 Investment Income on Surplus

The insurer's surplus is also comprised of invested assets which provide a rate of return. In ratemaking we assume that the investment income on surplus is earned over the policy term, generally 12 months. Investment income of surplus will reduce the required profit provision. For example, if the insurer were able to earn a rate of return on assets of 5% per annum, then the insurer would realize a return of 2.5% of premiums assuming a 2:1 premium: surplus ratio.

10.7.3 The Underwriting Profit Provisions

An insurer requires capital to support operations. The insured pays a premium for the promise of the insurer to pay a claim in the future. Capital serves as protection for the policyholder in the event that premiums are insufficient to pay claims. The capital provider will reasonably expect to earn a profit to insure the risk and subject its capital to loss. Generally, the required profit is expressed as after-tax return on equity.

If the profit provision in the premium equation were 0, then the premium would equal the present value of the present value of cash flows of the insurance policy. However, as we discussed the insurer will require a return on its capital. Generally, coverages that are riskier, i.e., have more variability, will require more supporting capital. Every claim submitted to the insurer has access to all of the capital of the insurer. In insurance, capital is often referred to as surplus. For ratemaking, we notionally allocate capital to coverage using premium to surplus ratios and we state the required rate of return on an after-tax basis. We have to convert that return to a "percent of premium basis" to include in our premium equation. For example, if we assume a 2:1 premium:surplus ratio, a required after-tax rate of return of 12% and a tax rate of 30%, then the profit provision in the premium would be:

$$\begin{aligned}\frac{12\% \text{ after tax return}}{\text{surplus}} & \times \frac{1 \times \text{surplus}}{2 \times \text{premium}} \times \frac{1 \text{ pre-tax}}{0.7 \text{ after tax}} \\ & = \frac{8.6\% \text{ pre-tax return}}{\text{premium}}.\end{aligned}$$

We can then reduce the required underwriting profit to consider investment income on surplus. Using the example of Section 10.7.2, the resulting required underwriting profit provision would reduce from 8.6% to 6.1%.

10.8 The Premium Equation

In this section, you learn how to:

- Calculate the rate for a class of risk
- Calculate premiums

We can now remove the simplifying assumptions included in Equation (10.1) and provide our final premium equation. The term "pure premium" can be used to refer to rate per exposure unit of provision for claims costs included in the premium for an insured (which may have a quantum of exposure more or less than one exposure unit). In this section, we use the latter definition.

$$\text{Premium} = \frac{\text{Discounted Pure Premium} + \text{Discounted Fixed Expenses}}{\text{Premium Delay} - \text{Variable Expense Rate} - \text{Required Underwriting Profit}}. \tag{10.2}$$

10.9 Pricing Principles

In this section, you learn how to:

- Describe common actuarial pricing principles
- Describe properties of pricing principles
- Choose a pricing principle based on a desired property

Approaches to pricing vary by the type of contract. For example, personal automobile is a widely available product throughout the world and is known as part of the *retail general insurance* market in the United Kingdom. Here, one can expect to do pricing based on a large pool of independent contracts, a situation in which expectations of losses provide an excellent starting point. In contrast, an actuary may wish to price an insurance contract issued to a large employer that covers complex health benefits for thousands of employees.

In this example, knowledge of the entire distribution of potential losses, not just the expected value, is critical for starting the pricing negotiations. To cover a range of potential applications, this section describes general premium principles and their properties that one can use to decide whether or not a specific principle is applicable in a given situation.

10.9.1 Premium Principles

The prior sections of this chapter introduce traditional actuarial pricing principles that provide a price based only target rates of return and the cost to insure the risk; the price does not depend on the demand for insurance.

Assume that the loss X has distribution function $F(\cdot)$ and that there exists some rule (which in mathematics is known as a *functional*), say H, that takes $F(\cdot)$ into the positive real line, denoted as $P = H(F)$. For notation purposes, it is often convenient to substitute the random variable X for its distribution function and write $P = H(X)$. Table 10.8.1 provides several examples.

Table 10.8.1. **Common Premium Principles**

Description	*Definition* $(H(X))$
Net (pure) premium	$\mathrm{E}[X]$
Expected value	$(1+\alpha)\mathrm{E}[X]$
Standard deviation	$\mathrm{E}[X] + \alpha\ SD(X)$
Variance	$\mathrm{E}[X] + \alpha\ \mathrm{Var}(X)$
Zero utility	solution of $u(w) = \mathrm{E}[u(w+P-X)]$
Exponential	$\frac{1}{\alpha}\log \mathrm{E}[e^{\alpha X}]$

A premium principle is similar to a risk measure that is introduced in Section 13.3. Mathematically, both are rules that map the loss rv of interest to a numerical value. From a practical viewpoint, a premium principle provides a guide as to how much an insurer will charge for accepting a risk X. In contrast, a risk measure quantifies the level of uncertainty, or riskiness, that an insurer can use to decide on a capital level to be assured of remaining solvent.

The net, or pure, premium essentially assumes no uncertainty. The expected value, standard deviation, and variance principles each add an explicit loading for uncertainty through the risk parameter $\alpha \geq 0$. For the principle of zero utility, we think of an insurer with utility function $u(\cdot)$ and wealth w as being indifferent to accepting and not accepting risk X. In this case, P is known as an indifference price or, in economics, a reservation price. With exponential utility, the principle of zero utility reduces to the exponential premium principle, that is, assuming $u(x) = (1 - e^{-\alpha x})/\alpha$.

For small values of the risk parameters, the variance principle is approximately equal to exponential premium principle, as illustrated in the following special case.

Special Case: Gamma Distribution. Consider a loss that is gamma distributed with parameters η and θ (we usually use α for the location parameter but, to distinguish it from the risk parameter, for this example we call it η). From the Appendix Chapter 20, the mean is $\eta\ \theta$ and the variance is $\eta\ \theta^2$. Using α_{Var} for the risk parameter, the variance premium is $H_{Var}(X) = \eta\ \theta + \alpha_{Var}\ (\eta\ \theta^2)$. From this appendix, it is straightforward to derive the well-known moment generating function, $M(t) = \mathrm{E}[e^{tX}] = (1-t\theta)^{-\eta}$. With this and a risk parameter α_{Exp}, we may express the exponential premium as

$$H_{Exp}(X) = \frac{-\eta}{\alpha_{Exp}} \log\left(1-\alpha_{Exp}\theta\right).$$

To see the relationship between $H_{Exp}(X)$ and $H_{Var}(X)$, we choose $\alpha_{Exp} = 2\alpha_{Var}$. With an approximation from calculus ($\log(1-x) = -x - x^2/2 - x^3/3 - \cdots$), we write

$$\begin{aligned} H_{Exp}(X) &= \tfrac{-\eta}{\alpha_{Exp}} \log\left(1-\alpha_{Exp}\ \theta\right) = \tfrac{-\eta}{\alpha_{Exp}} \left\{-\alpha_{Exp}\ \theta - (\alpha_{Exp}\ \theta)^2/2 - \cdots\right\} \\ &\approx \eta\ \theta + \tfrac{\alpha_{Exp}}{2}(\eta\ \theta^2) = H_{Var}(X). \end{aligned}$$

10.9.2 Properties of Premium Principles

Properties of premium principles help guide the selection of a premium principle in applications. Table 10.8.2 provides examples of properties of premium principles.

Table 10.8.2. **Common Properties of Premium Principles**

Description	*Definition*
Nonnegative loading	$H(X) \geq \mathrm{E}[X]$
Additivity	$H(X_1+X_2) = H(X_1)+H(X_2)$, for independent X_1, X_2
Scale invariance	$H(cX) = cH(X)$, for $c \geq 0$
Consistency	$H(c+X) = c+H(X)$
No rip-off	$H(X) \leq \max\{X\}$

This is simply a subset of the many properties quoted in the actuarial literature. For example, the review paper of Young (2014) lists 15 properties. See also the properties described as *coherent axioms* that we introduce for risk measures in Section 13.3.

Some of the properties listed in Table 10.8.2 are mild in the sense that they will nearly always be satisfied. For example, the *no rip-off* property indicates that the premium charge will be smaller than the largest or "maximal" value of the loss X (here, we use the notation $\max\{X\}$ for this maximal value which is defined as an "essential supremum" in mathematics). Other properties may not be so mild. For example, for a portfolio of independent risks, the actuary may want the *additivity* property to hold. It is easy to see that this property holds for the expected value, variance, and exponential premium principles but not for the standard deviation principle. Another example is the *consistency* property that does not hold for the expected value principle when the risk loading parameter α is positive.

The *scale invariance* principle is known as *homogeneity of degree one* in economics. For example, it allows us to work in different currencies (e.g., from dollars to Euros) as well as a host of other applications. Although a generally accepted principle, we note that this principle does not hold for a large value of X that may border on a surplus constraint of an insurer; if an insurer has a large probability of becoming insolvent, then that insurer may not wish to use linear pricing. It is easy to check that this principle holds for the expected value and standard deviation principles, although not for the variance and exponential principles.

10.10 Reviewing Rate Adequacy

After establishing the initial premiums, insurance company actuaries will perform rate reviews to measure the current adequacy of those rates. For many regulated coverages (typically, personal lines insurance), actuaries file those rate reviews with the insurance regulator. Actuaries review rates regularly as rate levels require updates to keep pace with inflationary pressures. At times, the required rate will have a decreasing trend; for example with improvements in vehicle safety technology or workplace safety. Of course, the primary purpose of the rate is to test whether the experience of the rate program is consistent with loss and expense assumptions underlying the current rates.

10.10.1 The Loss Ratio Method

The "loss ratio method" is a common approach to assess rate adequacy. The loss ratio is the ratio of loss to the premium.

$$\text{Loss Ratio} = \frac{\text{Loss}}{\text{Premium}}.$$

When determining premiums, it is a bit counter-intuitive to emphasize this ratio because the premium component is built into the denominator. As we will see, the loss ratio method develops rate **changes** rather than rates; we can use rate changes to adjust the current rate to the current costs levels.

We calculate rate changes by comparing the loss ratio of the experience period to the target loss ratio. This adjustment factor is then applied to current rates to determine new indicated rates.

10.10.2 Target Loss Ratio

Let us return to equation (10.2). Noting that the "pure premium" is the provision for loss in the rates, we can start with

$$\text{Premium} \quad = \frac{\text{Discounted Losses} + \text{Discounted Fixed Expenses}}{\text{Premium Delay} - \text{Variable Expense Rate} - \text{Profit}}$$

With this, we have

$$\begin{aligned}
\text{Premium Delay} \quad & \text{- Variable Expense Rate - Profit} \\
& = \frac{\text{Discounted Losses}}{\text{Premium}} + \frac{\text{Discounted Fixed Expenses}}{\text{Premium}} \\
\text{Premium Delay} \quad & \text{ - Variable Expense Rate - Profit} - \frac{\text{Discounted Fixed Expenses}}{\text{Premium}} \\
& = \frac{\text{Discounted Losses}}{\text{Premium}} \\
& = \text{Target Discounted Loss Ratio.}
\end{aligned}$$

For simplification, we will not repeat that the components of the rate change factor are discounted. In the loss ratio method, we compare the projected loss ratio to the target loss ratio. A projected loss ratio that exceeds the target loss ratio implies the need for a rate increase. A projected loss ratio that is less than the target loss ratio implies the need for a rate decrease.

10.10.3 Experience Period Loss Ratios

Earlier in this section, we described the required adjustments to estimate premiums. We apply those same adjustments to the experience period loss ratios.

10.10.4 Adjustments to Loss

- As with the development of pure premiums described above, actuaries will typically review claims experience over a multi-year (typically three to seven years) period to smooth the year-to-year randomness. The years in the experience period are similarly weighted to balance responsiveness to more recent experience and the stability of a longer-term period.

- The numerator of the loss ratio will be *ultimate losses.*
- We will consider the presence of catastrophe and large losses in the claims experience.
- We need to adjust the experience period losses to the cost level of the proposed rate program. We discussed this trend adjustment in Section 10.3.2. We apply the trend factor from the average accident date of the experience period to the average accident date of the proposed rate program. For example, if we are estimating rates that will underlie twelve month policies written in calendar year 2025, the average accident date of the prospective rate program will be 31 December 2025 (sometimes rounded to 1 January 2026). The first policy of the prospective period will be written on 1 January 2025 and expire on 31 December 2025. Assuming even distribution of claim events during the policy, the average accident date (midpoint) of that policy is 1 July 2025. Correspondingly, the last policy of the prospective period will be written on 31 December 2025 and expire on 31 December 2026 with an average accident date (midpoint) of that policy is 1 July 2026. Therefore, the midpoint of all policies written under the proposed rate program is 31 December 2025. To adjust experience for accident year 2022, we apply 3.5 years of trend. The average accident date of accident year 2022 is 1 July 2022 - so 3.5 years is the distance in time to the average accident date of the proposed rate program.

10.10.5 Premium On-Level Adjustment

We also need to adjust premiums for the effect of rate changes. We refer to this adjustment as "on-leveling." There are two common approaches to on-leveling.

- The Parallelogram Method: Premium on-level factors use historical rate change calculations. For example, if the company adopted a +10% rate change on 1 July 2022, then the 2022 earned premium would need to be adjusted by +7.5%. - Policies written prior to 1 July 2022 would need to be adjusted by +10%; - For the premium earned after 1 July 2022, half would be earned on policies written under the old rate levels and require the 10% adjustment and half would be written on policies written under the higher rate levels and require no adjustment. The weighted average of these adjustments if +7.5%.
- Extension of Exposures: The extension of exposures method is a more detailed approach which involves the re-rating of all historical policies at current rates. It is more precise as the parallelogram method relies on rate changes that were calculated as the average rate change given the mix of business at that time. However, the mix of business may change and the rate change effect

on the current mix may be different. The extension of exposures does not rely on those average rate changes and instead relies only on current rates.

10.10.6 Premium Trend

Experience period premiums must also be adjusted for for premium trend, and the basis of premium must match the loss trend. For example, insureds may purchase higher limits of coverage to protect against higher inflation. These higher limits would be reflected in the internal claims experience and may underlie the data used to measure loss trend. If we are considering these changes in the loss trend, then we also need to consider the effect of higher limit purchases in premium trend.

10.10.7 Credibility

Oftentimes, the experience being reviewed is not "fully credible." That is, the predictive value of the data is limited. We, therefore, need to consider an alternative indication of the projected loss ratio to calculate a credibility-weighted loss ratio. We refer to this alternative indicator as the complement of credibility. A common complement is the net loss trend (loss trend/premium trend). The assumption underlying the use of net loss trend as a complement is that in the absence of an alternative indication, we would need to adjust the rate level to consider changes in cost level. Chapter 12 describes credibility in detail.

Example. Loss Ratio Indicated Change Factor. Assume the following information:

- Experience period loss and LAE ratio = 65%
- Experience period credibility = 80%
- Loss Trend = 5%
- Premium On-Level Adjustment = 1.075
- Premium Trend = 2%
- Premium Delay Factor = 0.99
- Projected fixed expense ratio = 6.5%
- Variable expense = 25%
- Target UW profit = 6.1% .

With these assumptions, the indicated change factor can be calculated as

$$\begin{aligned}
&\text{Experience Period Loss Ratio} = 65\% \times \frac{1.05}{1.075 \times 1.02} = 62.2\% \\
&\text{Target Loss Ratio} = 0.99 - 6.5\% - 25\% - 6.1\% = 61.4\% \\
&\text{Complement of Credibility} = 0.614 \times \frac{1.05}{1.02} = 63.2\% \\
&\text{Credibility-weighted loss ratio} = 62.2\% \times 80\% + 63.2\% \times (1 - 80\%) = 62.4\% \\
&\text{Indicated loss ratio} = 62.4\%/61.4\% = 1.016.
\end{aligned}$$

This means that overall average rate level should be increased by 1.6%.

10.11 Further Resources and Contributors

This chapter serves as a bridge between the technical introduction of this book and an introduction to pricing and ratemaking for practicing actuaries. For readers interested in learning practical aspects of pricing, we recommend introductions by the Society of Actuaries in Friedland (2013) and by the Casualty Actuarial Society in Werner and Modlin (2016). For a classic risk management introduction to pricing, see Niehaus and Harrington (2003). See also Finger (2006) and Frees (2014).

Bühlmann (1985) was the first in the academic literature to argue that pricing should be done first at the portfolio level (he referred to this as a *top down* approach) which would be subsequently reconciled with pricing of individual contracts. See also the discussion in Kaas et al. (2008), Chapter 5.

For more background on pricing principles, a classic treatment is by Gerber (1979) with a more modern approach in Kaas et al. (2008). For more discussion of pricing from a financial economics viewpoint, see Bauer et al. (2013).

- **Edward (Jed) Frees**, University of Wisconsin-Madison, and **José Garrido**, Concordia University were the principal authors of the initial version of this chapter.
  - Chapter reviewers included Chun Yong Chew, Curtis Gary Dean, Brian Hartman, and Jeffrey Pai.
- **Rajesh Sahasrabuddhe**, Oliver Wyman, is the author of the second edition of this chapter. Email: rajesh1004@gmail.com for chapter comments and suggested improvements.

TS 10.A. Rate Regulation

Insurance regulation helps to ensure the financial stability of insurers and to protect consumers. Insurers receive premiums in return for promises to pay in the event of a contingent (insured) event. Like other financial institutions such as banks, there is a strong public interest in promoting the continuing viability of insurers.

Market Conduct

To help protect consumers, regulators impose administrative rules on the behavior of market participants. These rules, known as market conduct regulation, provide systems of regulatory controls that require insurers to demonstrate that they are providing fair and reliable services, including rating, in accordance with the statutes and regulations of a jurisdiction.

1. *Product regulation* serves to protect consumers by ensuring that insurance policy provisions are reasonable and fair, and do not contain major gaps in coverage that might be misunderstood by consumers and leave them unprotected.
2. The insurance product is the insurance contract (policy) and the coverage it provides. Insurance contracts are regulated for these reasons:
   a. Insurance policies are complex legal documents that are often difficult to interpret and understand.
   b. Insurers write insurance policies and sell them to the public on a "take it or leave it" basis.

Market conduct includes rules for *intermediaries* such as agents (who sell insurance to individuals) and brokers (who sell insurance to businesses). Market conduct also includes *competition policy regulation*, designed to ensure an efficient and competitive marketplace that offers low prices to consumers.

Rate Regulation

Rate regulation helps guide the development of premiums and so is the focus of this chapter. As with other aspects of market conduct regulation, the intent of these regulations is to ensure that insurers not take unfair advantage of consumers. Rate (and policy form) regulation is common worldwide.

The amount of regulatory scrutiny varies by insurance product. Rate regulation is uncommon in life insurance. Further, in non-life insurance, most commercial lines and reinsurance are free from regulation. Rate regulation is common in automobile insurance, health insurance, workers compensation, medical malpractice, and homeowners insurance. These are markets in which insurance is mandatory or in which universal coverage is thought to be socially desirable.

There are three principles that guide rate regulation: rates should

- be adequate (to maintain insurance company solvency),

- but not excessive (not so high as to lead to exorbitant profits),
- nor unfairly discriminatory (price differences must reflect expected claim and expense differences).

Recently, in auto and home insurance, the twin issues of availability and affordability, which are not explicitly included in the guiding principles, have been assuming greater importance in regulatory decisions.

Rates are Not Unfairly Discriminatory

Some government regulations of insurance restrict the amount, or level, of premium rates. These are based on the first two of the three guiding rate regulation principles, that rates be adequate but not excessive. This type of regulation is discussed further in the following section on types of rate regulation.

Other government regulations restrict the type of information that can be used in risk classification. These are based on the third guiding principle, that rates not be unfairly discriminatory. "Discrimination" in an insurance context has a different meaning than commonly used; for our purposes, discrimination means the ability to distinguish among things or, in our case, policyholders. The real issue is what is meant by the adjective "fair."

In life insurance, it has long been held that it is reasonable and fair to charge different premium rates by age. For example, a life insurance premium differs dramatically between an 80 year old and someone aged 20. In contrast, it is unheard of to use rates that differ by:

- ethnicity or race,
- political affiliation, or
- religion.

It is not a matter of whether data can be used to establish statistical significance among the levels of any of these variables. Rather, it is a societal decision as to what constitutes notions of "fairness."

Different jurisdictions have taken different stances on what constitutes a fair rating variable. For example, in some jurisdictions for some insurance products, gender is no longer a permissible variable. As an illustration, the European Union now prohibits the use of gender for automobile rating. As another example, in the U.S., many discussions have revolved around the use of credit ratings to be used in automobile insurance pricing. Credit ratings are designed to measure consumer financial responsibility. Yet, some argue that credit scores are good proxies for ethnicity and hence should be prohibited.

In an age where more data is being used in imaginative ways, discussions of what constitutes a fair rating variable will only become more important going forward and much of that discussion is beyond the scope of this text. However, it is relevant to the discussion to remark that actuaries and other data analysts can contribute to societal discussions on what constitutes a "fair" rating variable in unique ways by establishing the magnitude of price differences when using variables under discussion.

Types of Rate Regulation

There are several methods, that vary by the level of scrutiny, by which regulators may restrict the rates that insurers offer.

The most restrictive is a government prescribed regulatory system, where the government regulator determines and promulgates the rates, classifications, forms, and so forth, to which all insurers must adhere. Also restrictive are prior approval systems. Here, the insurer must file rates, rules, and so forth, with government regulators. Depending on the statute, the filing becomes effective when a specified waiting period elapses (if the government regulator does not take specific action on the filing, it is deemed approved automatically) or when the government regulator formally approves the filing.

The least restrictive is a no file or *record maintenance* system where the insurer need not file rates, rules, and so forth, with the government regulator. The regulator may periodically examine the insurer to ensure compliance with the law. Another relatively flexible system is the file only system, also known as *competitive* rating, where the insurer simply keeps files to ensure compliance with the law.

In between these two extremes are the (1) file and use, (2) use and file, (3) modified prior approval, and (4) flex rating systems.

1. File and Use: The insurer must file rates, rules, and so forth, with the government regulator. The filing becomes effective immediately or on a future date specified by the filer.
2. Use and File: The filing becomes effective when used. The insurer must file rates, rules, and so forth, with the government regulator within a specified time period after first use.
3. Modified Prior Approval: This is a hybrid of "prior approval" and "file and use" laws. If the rate revision is based solely on a change in loss experience then "file and use" may apply. However, if the rate revision is based on a change in expense relationships or rate classifications, then "prior approval" may apply.
4. Flex (or Band) Rating: The insurer may increase or decrease a rate within a "flex band," or range, without approval of the government

regulator. Generally, either "file and use" or "use and file" provisions apply.

For a broad introduction to government insurance regulation from a global perspective, see the website of the International Association of Insurance Supervisors (IAIS).

11

Risk Classification

Chapter Preview. This chapter motivates the use of risk classification in insurance pricing and introduces readers to Poisson regression as a prominent example of risk classification. In Section 11.1 we explain why insurers need to incorporate various risk characteristics, or rating factors, of individual policyholders in pricing insurance contracts. In Section 11.2, we introduce Poisson regression as a pricing tool to achieve such premium differentials. The concept of exposure is also introduced in this section. As most rating factors are categorical, we show in Section 11.3 how the multiplicative tariff model can be incorporated into a Poisson regression model in practice, along with numerical examples for illustration.

11.1 Introduction

In this section, you learn:

- Why premiums should vary across policyholders with different risk characteristics.
- The meaning of the adverse selection spiral.
- The need for risk classification.

Through insurance contracts, the policyholders effectively transfer their risks to the insurer in exchange for premiums. For the insurer to stay in business, the premium income collected from a pool of policyholders must at least equal the benefit outgo. In general insurance products where a premium is charged for a single period, say annually, the gross insurance premium based on the equivalence principle is stated as

$$\text{Gross Premium} = \text{Expected Losses} + \text{Expected Expenses} + \text{Profit}.$$

Thus, ignoring frictional expenses associated with the administrative expenses and profit, the net or pure premium charged by the insurer should be equal to the expected losses occurring from the risk that is transferred from the policyholder.

If all policyholders in the insurance pool have identical risk profiles, the insurer simply charges the same premium for all policyholders because they have the same expected loss. In reality, however, the policyholders are hardly homogeneous. For example, mortality risk in life insurance depends on the characteristics of the policyholder, such as, age, sex and lifestyle. In auto insurance, those characteristics may include age, occupation, the type or use of the car, and the area where the driver resides. The knowledge of these characteristics or variables can enhance the ability of calculating fair premiums for individual policyholders, as they can be used to estimate or predict the expected losses more accurately.

Adverse Selection. Indeed, if the insurer does not differentiate the risk characteristics of individual policyholders and simply charges the same premium to all insureds based on the average loss in the portfolio, the insurer would face adverse selection, a situation where individuals with a higher chance of loss are attracted in the portfolio and low-risk individuals are repelled.

For example, consider a health insurance where smoking status is an important risk factor for mortality and morbidity. Most health insurers in the market require different premiums depending on smoking status, so smokers pay higher premiums than non-smokers, with other characteristics being identical. Now suppose that there is an insurer, we will call EquitabAll, that offers the same premium to all insureds regardless of smoking status, unlike other competitors. The net premium of EquitabAll is naturally an average mortality loss accounting for both smokers and non-smokers. That is, the net premium is a weighted average of the losses with the weights being the proportions of smokers and non-smokers, respectively. Thus it is easy to see that that a smoker would have a good incentive to purchase insurance from EquitabAll than from other insurers as the offered premium by EquitabAll is relatively lower. At the same time non-smokers would prefer buying insurance from somewhere else where lower premiums, computed from the non-smoker group only, are offered. As a result, there will be more smokers and less non-smokers in the EquitabAll's portfolio, which leads to larger-than-expected losses and hence a higher premium for insureds in the next period to cover the higher costs. With the raised new premium in the next period, non-smokers in EquitabAll will have even greater incentives to switch insurers. As this cycle continues over time, EquitabAll would gradually retain more smokers and less non-smokers in its portfolio with the premium continually raised, eventually leading to business collapse.

In the literature, this phenomenon is known as the adverse selection spiral or death spiral. Therefore, incorporating and differentiating important risk characteristics of individuals in the insurance pricing process are a pertinent component for both the determination of fair premium for individual policyholders and the long term sustainability of insurers.

Rating Factors. In order to incorporate relevant risk characteristics of policyholders in the pricing process, insurers maintain some classification system that assigns each policyholder to one of the risk classes based on a relatively small number of risk characteristics that are deemed most relevant. These characteristics used in the classification system are called rating factors, which are a priori variables in the sense that they are known before the contract begins (e.g., sex, health status, vehicle type, etc, are known during underwriting). All policyholders sharing identical risk factors thus are assigned to the same risk class, and are considered homogeneous from a pricing viewpoint; the insurer consequently charges them the same premium or rate.

Regarding the risk factors and premiums, the *Actuarial Standard of Practice* (ASOP) No. 12 of the Actuarial Standards Board (2018) states that the actuary should select risk characteristics that are related to expected outcomes, and that rates within a risk classification system would be considered equitable if differences in rates reflect material differences in expected cost for risk characteristics. In the process of choosing risk factors, ASOP also requires the actuary to consider the following: relationship of risk characteristics and expected outcomes, causality, objectivity, practicality, applicable law, industry practices, and business practices.

On the quantitative side, an important task for the actuary in building a risk classification framework is to construct a statistical model that can determine the expected loss given various rating factors of a policyholder. The standard approach is to adopt a regression model which produces the expected loss as the output when the relevant risk factors are given as the inputs. In this chapter we learn about Poisson regression, which can be used when the loss is a count variable, as a prominent example of an insurance pricing tool.

11.2 Poisson Regression Model

The Poisson regression model has been successfully used in a wide range of applications and has an advantage of allowing closed-form expressions for important quantities. In this section we introduce Poisson regression as a natural extension of the Poisson distribution.

In this section you will:

- Understand Poisson regression as a convenient tool for combining individual Poisson distributions.
- Sharpen your understanding of the concept of exposure and its importance.
- Formally learn how to formulate a Poisson regression model using indicator variables when the explanatory variables are categorical.

11.2.1 Need for Poisson Regression

Poisson Distribution

To introduce Poisson regression, let us consider a hypothetical health insurance portfolio where all policyholders are of the same age and only one risk factor, smoking status, is relevant. Smoking status thus is a categorical variable with two levels: smoker and non-smoker. As there are two levels for smoking status, we may denote smoker and non-smoker by level 1 and 2, respectively. Here the numbering is arbitrary; smoking status is a nominal categorical variable. (See Section 2.3.1 an introduction to categorical and nominal variables.) Suppose now that we are interested in pricing a health insurance where the premium for each policyholder is determined by the number of outpatient visits to doctor's office during a year. The medical cost for each visit is assumed to be the same regardless of smoking status for simplicity. Thus if we believe that smoking status is a valid risk factor in this health insurance, it is natural to consider observations from smokers separately from non-smokers. In Table 11.1 we present data for this portfolio.

Table 11.1. **Number of Visits to Doctor's Office in Last Year**

Smoker	(level 1)	Non-smoker	(level 2)		Both
Count	Observed	Count	Observed	Count	Observed
0	2213	0	6671	0	8884
1	178	1	430	1	608
2	11	2	25	2	36
3	6	3	9	3	15
4	0	4	4	4	4
5	1	5	2	5	3
Total	2409	Total	7141	Total	9550
Mean	0.0926	Mean	0.0746	Mean	0.0792

As this dataset contains random counts, we try to fit a Poisson distribution for each level.

As introduced in Section 3.2.3, the probability mass function of the Poisson with mean μ is given by

$$\Pr(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, \qquad y = 0, 1, 2, \ldots \tag{11.1}$$

and $\mathrm{E}\,(Y) = \mathrm{Var}\,(Y) = \mu$. In regression contexts, it is common to use μ for mean parameters instead of the Poisson parameter λ although certainly both symbols are suitable. As we saw in Section 3.4.2, the mle of the Poisson distribution is given by the sample mean. Thus if we denote the Poisson mean parameter for each level by $\mu_{(1)}$ (smoker) and $\mu_{(2)}$ (non-smoker), we see from Table 11.1 that $\hat{\mu}_{(1)} = 0.0926$ and $\hat{\mu}_{(2)} = 0.0746$. This simple example shows the basic idea of risk classification. Depending on smoking status, a policyholder will have a different risk characteristic that can be incorporated via varying Poisson mean parameters to compute the fair premium. In this example the ratio of expected loss frequencies is $\hat{\mu}_{(1)}/\hat{\mu}_{(2)} = 1.2402$, implying that smokers tend to visit a doctor's office 24.02% times more frequently compared to non-smokers.

It is also informative to note that if the insurer charges the same premium to all policyholders regardless of smoking status, based on the average characteristic of the portfolio, as was the case for EquitabAll described in Introduction, the expected frequency (or premium) $\hat{\mu}$ is 0.0792, obtained from the last column of Table 11.1. It can be verified that

$$\hat{\mu} = \left(\frac{n_1}{n_1 + n_2}\right)\hat{\mu}_{(1)} + \left(\frac{n_2}{n_1 + n_2}\right)\hat{\mu}_{(2)} = 0.0792, \tag{11.2}$$

where n_i is the number of observations in each level. Clearly, this premium is a weighted average of the premiums for each level with the weight equal to the proportion of insureds in that level.

A simple Poisson regression

In the example above, we have fitted a Poisson distribution for each level separately, but we can actually combine them together in a unified fashion so that a single Poisson model can encompass both smoking and non-smoking statuses. This can be done by relating the Poisson mean parameter with the risk factor. In other words, we make the Poisson mean, which is the expected loss frequency, respond to the change in the smoking status. The conventional approach to deal with a categorical variable is to adopt indicator or dummy variables that take either 1 or 0, so that we turn the switch on for one level and off for others. Therefore we may propose to use

$$\mu = \beta_0 + \beta_1 x_1 \tag{11.3}$$

or, more commonly, a log linear form

$$\log \mu = \beta_0 + \beta_1 x_1, \tag{11.4}$$

where x_1 is an indicator variable with

$$x_1 = \begin{cases} 1 & \text{if smoker,} \\ 0 & \text{otherwise.} \end{cases} \tag{11.5}$$

We generally prefer the log linear relation in (11.4) to the linear one in (11.3) to prevent producing negative μ values, which can happen when there are many different risk factors and levels. The setup in (11.4) and (11.5) then results in different Poisson frequency parameters depending on the level in the risk factor:

$$\log \mu = \begin{cases} \beta_0 + \beta_1 \\ \beta_0 \end{cases} \quad \text{or equivalently,} \quad \mu = \begin{cases} e^{\beta_0+\beta_1} & \text{if smoker (level 1),} \\ e^{\beta_0} & \text{if non-smoker (level 2).} \end{cases} \tag{11.6}$$

This is the simplest form of Poisson regression. Note that we require a single indicator variable to model two levels in this case. Alternatively, it is also possible to use two indicator variables through a different coding scheme. This scheme requires dropping the intercept term so that (11.4) is modified to

$$\log \mu = \beta_1 x_1 + \beta_2 x_2, \tag{11.7}$$

where x_2 is the second indicator variable with

$$x_2 = \begin{cases} 1 & \text{if non-smoker,} \\ 0 & \text{otherwise.} \end{cases}$$

Then we have, from (11.7),

$$\log\mu = \begin{cases}\beta_1 \\ \beta_2\end{cases} \qquad \text{or} \qquad \mu = \begin{cases}e^{\beta_1} & \text{if smoker (level 1),} \\ e^{\beta_2} & \text{if non-smoker (level 2).}\end{cases} \tag{11.8}$$

The numerical result of (11.6) is the same as (11.8) as all coefficients are given as numbers in actual estimation, with the former setup more common in most texts; we also stick to the former.

With this Poisson regression model we can readily understand how the coefficients β_0 and β_1 are linked to the expected loss frequency in each level. According to (11.6), the Poisson mean of the smokers, $\mu_{(1)}$, is given by

$$\mu_{(1)} = e^{\beta_0+\beta_1} = \mu_{(2)}\, e^{\beta_1} \quad \text{or} \quad \mu_{(1)}/\mu_{(2)} = e^{\beta_1}$$

where $\mu_{(2)}$ is the Poisson mean for the non-smokers. This relation between the smokers and non-smokers suggests a useful way to compare the risks embedded in different levels of a given risk factor. That is, the proportional increase in the expected loss frequency of the smokers compared to that of the non-smokers is simply given by a multiplicative factor e^{β_1}. Put another way, if we set the expected loss frequency of the non-smokers as the base value, the expected loss frequency of the smokers is obtained by applying e^{β_1} to the base value.

Dealing with multi-level case

We can readily extend the two-level case to a multi-level one where l different levels are involved for a single rating factor. For this we generally need $l-1$ indicator variables to formulate

$$\log\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_{l-1}x_{l-1}, \tag{11.9}$$

where x_k is an indicator variable that equals 1 if the policy belongs to level k and 0 otherwise, for $k = 1, 2, \ldots, l-1$. By omitting the indicator variable associated with the last level in (11.9) we effectively chose level l as the base case, or reference level, but this choice is arbitrary and does not matter numerically. The resulting Poisson parameter for policies in level k then becomes, from (11.9),

$$\mu = \begin{cases}e^{\beta_0+\beta_k} & \text{if the policy belongs to level } k, (k = 1, 2, ..., l-1), \\ e^{\beta_0} & \text{if the policy belongs to level } l.\end{cases}$$

Thus if we denote the Poisson parameter for policies in level k by $\mu_{(k)}$, we can relate the Poisson parameter for different levels through $\mu_{(k)} = \mu_{(l)}\, e^{\beta_k}$, $k = 1, 2, \ldots, l-1$. This indicates that, just like the two-level case, the expected loss frequency of the kth level is obtained from the base value multiplied by the relative factor e^{β_k}. This relative interpretation becomes more powerful when there are many risk factors with multi-levels, and leads us to a better understanding of the underlying risk and a more accurate prediction of future losses. Finally, we note that the varying Poisson mean is completely driven by the coefficient parameters β_k's, which are to be estimated from the dataset; the procedure of the parameter estimation will be discussed later in this chapter.

11.2.2 Poisson Regression

We now describe Poisson regression in a formal and more general setting. Let us assume that there are n independent policyholders with a set of rating factors characterized by a k-variate vector[1]. The ith policyholder's rating factor is thus denoted by vector $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})'$, and the policyholder has recorded the loss count $y_i \in \{0, 1, 2, \ldots\}$ from the last period of loss observation, for $i = 1, \ldots, n$. In the regression literature, the values $x_{i1}, \ldots, x_{ik}$ are generally known as *explanatory variables*, as these are measurements providing information about the variable of interest y_i. In essence, regression analysis is a method to quantify the relationship between a variable of interest and explanatory variables.

We also assume, for now, that all policyholders have the same one unit period for loss observation, or equal exposure of 1, to keep things simple; we will discuss more details regarding the exposure in the following subsection.

We describe Poisson regression through its mean function. For this we first denote μ_i as the expected loss count of the ith policyholder under the Poisson specification (11.1):

$$\mu_i = \mathrm{E}\ (y_i|\mathbf{x}_i), \qquad y_i \sim Pois(\mu_i),\ i = 1, \ldots, n. \tag{11.10}$$

The condition inside the expectation in equation (11.10) indicates that the loss frequency μ_i is the model expected response to the given set of risk factors or explanatory variables. In principle the conditional mean $\mathrm{E}\ (y_i|\mathbf{x}_i)$ in (11.10) can take different forms depending on how we specify the relationship between $\mathbf{x}$ and y. The standard choice for Poisson regression is to adopt the exponential function, as we mentioned previously, so that

$$\mu_i = \mathrm{E}\ (y_i|\mathbf{x}_i) = e^{\mathbf{x}_i'\beta}, \qquad y_i \sim Pois(\mu_i),\ i = 1, \ldots, n. \tag{11.11}$$

Here $\beta = (\beta_0, \ldots, \beta_k)'$ is the vector of coefficients so that $\mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$. The exponential function in (11.11) ensures that $\mu_i > 0$ for any set of rating factors $\mathbf{x}_i$. Often (11.11) is rewritten as a log linear form

$$\log \mu_i = \log \mathrm{E}\ (y_i|\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}, \qquad y_i \sim Pois(\mu_i),\ i = 1, \ldots, n \tag{11.12}$$

[1] For example, if there are 3 risk factors each of which the number of levels are 2, 3 and 4, respectively, we have $k = (2-1) \times (3-1) \times (4-1) = 6$.

to reveal the relationship when the right side is set as the linear form, $\mathbf{x}_i'\beta$. Again, we see that the mapping works well as both sides of (11.12), $\log \mu_i$ and $\mathbf{x}_i\beta$, can now cover all real values. This is the formulation of Poisson regression, assuming that all policyholders have the same unit period of exposure. When the exposures differ among the policyholders, however, as is the case in most practical cases, we need to revise this formulation by adding an exposure component as an additional term in (11.12).

11.2.3 Incorporating Exposure

Concept of Exposure

We first saw the concept of exposures in Section 10.4. In order to determine the size of potential losses in any type of insurance, one must always know the corresponding exposure. The concept of exposure is an extremely important ingredient in insurance pricing, though we usually take it for granted. For example, when we say the expected claim frequency of a health insurance policy is 0.2, it does not mean much without the specification of the exposure such as, in this case, per month or per year. In fact, all premiums and losses need the exposure precisely specified and must be quoted accordingly; otherwise all subsequent statistical analyses and predictions will be distorted.

In the previous section we assumed the same unit of exposure across all policyholders, but this is hardly realistic in practice. In health insurance, for example, two different policyholders with different lengths of insurance coverage (e.g., 3 months and 12 months, respectively) could have recorded the same number of claim counts. As the expected number of claim counts would be proportional to the length of coverage, we should not treat these two policyholders' loss experiences identically in the modeling process. This motivates the need of the concept of *exposure* in Poisson regression.

The Poisson distribution in (11.1) is parametrized via its mean. To understand the exposure, we alternatively parametrize the Poisson *pmf* in terms of the *rate* parameter λ, based on the definition of the Poisson process:

$$\Pr(Y=y) = \frac{(\lambda t)^y e^{-\lambda t}}{y!}, \qquad y = 0, 1, 2, \ldots \tag{11.13}$$

with E (Y) = Var $(Y) = \lambda t$. Here λ is known as the rate or intensity per unit period of the Poisson process and t represents the length of time or *exposure*, a known constant value. For given λ the Poisson distribution (11.13) produces a larger expected loss count as the exposure t gets larger. Clearly, (11.13) reduces to (11.1) when $t = 1$, which means that the mean and the rate become the same for an exposure of 1, the case we considered in the previous subsection.

In principle, the exposure does not need to be measured in units of time and may represent different things depending the problem at hand. For example:

1. In health insurance, the rate may be the occurrence of a specific disease per 1,000 people and the exposure is the number of people considered in the unit of 1,000.
2. In auto insurance, the rate may be the number of accidents per year of a driver and the exposure is the length of the observed period for the driver in the unit of year.
3. For workers compensation that covers lost wages resulting from an employee's work-related injury or illness, the rate may be the probability of injury in the course of employment per dollar and the exposure is the payroll amount in dollars.
4. In marketing, the rate may be the number of customers who enter a store per hour and the exposure is the number of hours observed.
5. In civil engineering, the rate may be the number of major cracks on the paved road per 10 kms and the exposure is the length of road considered in the unit of 10 kms.
6. In credit risk modelling, the rate may be the number of default events per 1000 firms and the exposure is the number of firms under consideration in the unit of 1,000.

Actuaries may be able to use different exposure bases for a given insurable loss. For example, in auto insurance, both the number of kilometers driven and the number of months covered by insurance can be used as exposure bases. Here the former is more accurate and useful in modelling the losses from car accidents, but more difficult to measure and manage for insurers. Thus, a good exposure base may not be the theoretically best one due to various practical constraints. As a rule, an exposure base must be easy to determine, accurately measurable, legally and socially acceptable, and free from potential manipulation by policyholders.

Incorporating exposure in Poisson regression

As exposures affect the Poisson mean, constructing Poisson regressions requires us to carefully separate the rate and exposure in the modelling process. Focusing on the insurance context, let us denote the rate of the loss event of the ith policyholder by λ_i, the known exposure (the length of coverage) by m_i and the expected loss count under the given exposure by μ_i. Then the Poisson regression formulation in (11.11) and (11.12) should be revised in light of

(11.13) as

$$\mu_i = \mathrm{E}\,(y_i|\mathbf{x}_i) = m_i\,\lambda_i = m_i\,e^{\mathbf{x}_i'\boldsymbol{\beta}}, \qquad y_i \sim Pois(\mu_i),\, i = 1, \ldots, n, \tag{11.14}$$

which gives

$$\log \mu_i = \log m_i + \mathbf{x}_i'\boldsymbol{\beta}, \qquad y_i \sim Pois(\mu_i),\, i = 1, \ldots, \tag{11.15}$$

Adding $\log m_i$ in (11.15) does not pose a problem in fitting as we can always specify this as an extra explanatory variable, as it is a known constant, and fix its coefficient to 1. In the literature the log of exposure, $\log m_i$, is commonly called the offset.

11.3 Categorical Variables and Multiplicative Tariff

In this section you will learn:

- The multiplicative tariff model when the rating factors are categorical.
- How to construct a Poisson regression model based on the multiplicative tariff structure.

11.3.1 Rating Factors and Tariff

In practice most rating factors in insurance are *categorical variables*, meaning that they take one of the predetermined number of possible values. Examples of categorical variables include sex, type of cars, the driver's region of residence and occupation. Continuous variables, such as age or auto mileage, can also be grouped by bands and treated as categorical variables. Thus we can imagine that, with a small number of rating factors, there will be many policyholders falling into the same risk class, charged with the same premium. For the remaining of this chapter we assume that all rating factors are categorical variables.

To illustrate how categorical variables are used in the pricing process, we consider a hypothetical auto insurance with only two rating factors:

- Type of vehicle: Type A (personally owned) and B (owned by corporations). We use index $j = 1$ and 2 to respectively represent each level of this rating factor.

- Age band of the driver: Young (age < 25), middle ($25 \leq$ age < 60) and old age (age ≥ 60). We use index $k = 1, 2$ and 3, respectively, for this rating factor.

From this classification rule, we may create an organized table or list, such as the one shown in Table 11.2, collected from all policyholders. Clearly there are $2 \times 3 = 6$ different risk classes in total. Each row of the table shows a combination of different risk characteristics of individual policyholders. Our goal is to compute six different premiums for each of these combinations. Once the premium for each row has been determined using the given exposure and claim counts, the insurer can replace the last two columns in Table 11.2 with a single column containing the computed premiums. This new table then can serve as a manual to determine the premium for a new policyholder given rating factors during the underwriting process. In non-life insurance, a table (or a set of tables) or list that contains each set of rating factors and the associated premium is referred to as a tariff. Each unique combination of the rating factors in a tariff is called a *tariff cell*; thus, in Table 11.2 the number of tariff cells is six, same as the number of risk classes.

Table 11.2. **Loss Record of the Illustrative Auto Insurer**

Rating	factors	Exposure	Claim count
Type (j)	Age (k)	in year	observed
1	1	89.1	9
1	2	208.5	8
1	3	155.2	6
2	1	19.3	1
2	2	360.4	13
2	3	276.7	6

Let us now look at the loss information in Table 11.2 more closely. The exposure in each row represents the sum of the length of insurance coverages, or in-force times, in years, of all the policyholders in that tariff cell. Similarly the claim counts in each row is the number of claims in each cell. Naturally the exposures and claim counts vary due to the different number of drivers across the cells, as well as different in-force time periods among the drivers within each cell.

In light of the Poisson regression framework, we denote the exposure and claim count of cell (j, k) as m_{jk} and y_{jk}, respectively, and define the claim count per unit exposure as

$$z_{jk} = \frac{y_{jk}}{m_{jk}}, \qquad j = 1, 2;\ k = 1, 2, 3.$$

For example, $z_{12} = 8/208.5 = 0.03837$, meaning that a policyholder in tariff cell

(1,2) would have 0.03837 accidents if insured for a full year on average. The set of z_{ij} values then corresponds to the rate parameter in the Poisson distribution (11.13) as they are the event occurrence rates per unit exposure. That is, we have $z_{jk} = \hat{\lambda}_{jk}$ where λ_{jk} is the Poisson rate parameter. Producing z_{ij} values however does not do much beyond comparing the average loss frequencies across risk classes. To fully exploit the dataset, we will construct a pricing model from Table 11.2 using Poisson regression, for the remaining part of the chapter.

We comment that actual loss records used by insurers typically include many more risk factors, in which case the number of cells grows exponentially. The tariff would then consist of a set of tables, instead of one, separated by some of the basic rating factors, such as sex or territory.

11.3.2 Multiplicative Tariff Model

In this subsection, we introduce the multiplicative tariff model, a popular pricing structure that can be naturally used within the Poisson regression framework. The developments here are based on Table 11.2. Recall that the loss count of a policyholder is described by a Poisson regression model with rate λ and the exposure m, so that the expected loss count becomes $m\lambda$. As m is a known constant, we are essentially concerned with modelling λ, so that it responds to the change in rating factors. Among other possible functional forms, we commonly choose the multiplicative[2] relation to model the Poisson rate λ_{jk} for cell (j, k):

$$\lambda_{jk} = f_0 \times f_{1j} \times f_{2k}, \qquad j = 1, 2;\ k = 1, 2, 3. \tag{11.16}$$

Here $\{f_{1j}, j = 1, 2\}$ are the parameters associated with the two levels in the first rating factor, car type, and $\{f_{2k}, k = 1, 2, 3\}$ associated with the three levels in the age band, the second rating factor. For instance, the Poisson rate for a mid-aged policyholder with a Type B vehicle is given by $\lambda_{22} = f_0 \times f_{12} \times f_{22}$. The first term f_0 is some base value to be discussed shortly. Thus these six parameters are understood as numerical representations of the levels within each rating factor, and are to be estimated from the dataset.

The multiplicative form (11.16) is easy to understand and use, because it clearly shows how the expected loss count (per unit exposure) changes as each rating factor varies. For example, if $f_{11} = 1$ and $f_{12} = 1.2$, then the expected loss count of a policyholder with a vehicle of type B would be 20% larger than type A, when the other factors are the same. In non-life insurance, the

[2] Preferring the multiplicative form to others (e.g., additive one) was already hinted in (11.4).

parameters f_{1j} and f_{2k} are known as relativities as they determine how much expected loss should change relative to the base value f_0. The idea of relativity is quite convenient in practice, as we can decide the premium for a policyholder by simply multiplying a series of corresponding relativities to the base value.

Dropping an existing rating factor or adding a new one is also transparent with this multiplicative structure. In addition, the insurer may adjust the overall premium for all policyholders by controlling the base value f_0 without changing individual relativities. However, by adopting the multiplicative form, we implicitly assume that there is no serious interaction among the risk factors.

When the multiplicative form is used we need to address an identification issue. That is, for any $c > 0$, we can write

$$\lambda_{jk} = f_0 \times \frac{f_{1j}}{c} \times c\, f_{2k}.$$

By comparing with (11.16), we see that the identical rate parameter λ_{jk} can be obtained for very different individual relativities. This over-parametrization, meaning that many different sets of parameters arrive at an identical model, obviously calls for some restriction on f_{1j} and f_{2k}. The standard practice is to make one relativity in each rating factor equal to one. This can be made arbitrarily in theory, but the standard practice is to make the relativity of most common class (base class) equal to one. We will assume that *type A vehicles* and *young drivers* to be the most common classes, that is, $f_{11} = 1$ and $f_{21} = 1$. This way all other relativities are uniquely determined. The tariff cell $(j, k) = (1, 1)$ is then called the base tariff cell, where the rate simply becomes $\lambda_{11} = f_0$, corresponding to the base value according to (11.16). Thus the base value f_0 is generally interpreted as the Poisson rate of the base tariff cell.

Again, (11.16) is log-transformed and rewritten as

$$\log \lambda_{jk} = \log f_0 + \log f_{1j} + \log f_{2k}, \tag{11.17}$$

as it is easier to estimate, similar to (11.12). This log linear form makes the log relativities of the base level in each rating factor equal to zero, i.e., $\log f_{11} = \log f_{21} = 0$, and leads to the following alternative, more explicit expression for (11.17):

$$\log \lambda_{jk} = \begin{cases} \log f_0 + \quad 0 \quad + \quad 0 & \text{for a policy in cell } (1,1), \\ \log f_0 + \quad 0 \quad + \log f_{22} & \text{for a policy in cell } (1,2), \\ \log f_0 + \quad 0 \quad + \log f_{23} & \text{for a policy in cell } (1,3), \\ \log f_0 + \log f_{12} + \quad 0 & \text{for a policy in cell } (2,1), \\ \log f_0 + \log f_{12} + \log f_{22} & \text{for a policy in cell } (2,2), \\ \log f_0 + \log f_{12} + \log f_{23} & \text{for a policy in cell } (2,3). \end{cases} \tag{11.18}$$

This shows that the Poisson rate parameter λ varies across different tariff cells, with the same log linear form used in a Poisson regression framework. In fact the reader may see that (11.18) is an extended version of the early expression (11.6) with multiple risk factors and that the log relativities now play the role of β_i parameters. Therefore all the relativities can be readily estimated via fitting a Poisson regression with a suitably chosen set of indicator variables.

11.3.3 Poisson Regression for Multiplicative Tariff

Indicator Variables for Tariff Cells

We now explain how the relativities can be incorporated into Poisson regression. As seen early in this chapter we use indicator variables to deal with categorical variables. For our illustrative auto insurer, therefore, we define an indicator variable for the first rating factor as

$$x_1 = \begin{cases} 1 & \text{for vehicle type B,} \\ 0 & \text{otherwise.} \end{cases}$$

For the second rating factor, we employ two indicator variables for the age band, that is,

$$x_2 = \begin{cases} 1 & \text{for age band 2,} \\ 0 & \text{otherwise.} \end{cases}$$

and

$$x_3 = \begin{cases} 1 & \text{for age band 3,} \\ 0 & \text{otherwise.} \end{cases}$$

The triple (x_1, x_2, x_3) then can effectively and uniquely determine each risk class. By observing that the indicator variables associated with Type A and Age band 1 are omitted, we see that tariff cell $(j, k) = (1, 1)$ plays the role of the base cell. We emphasize that our choice of the three indicator variables above has been carefully made so that it is consistent with the choice of the base levels in the multiplicative tariff model in the previous subsection (i.e., $f_{11} = 1$ and $f_{21} = 1$).

With the proposed indicator variables we can rewrite the log rate (11.17) as

$$\log \lambda = \log f_0 + \log f_{12} \times x_1 + \log f_{22} \times x_2 + \log f_{23} \times x_3, \qquad (11.19)$$

which is identical to (11.18) when each triple value is actually applied. For example, we can verify that the base tariff cell $(j, k) = (1, 1)$ corresponds to $(x_1, x_2, x_3) = (0, 0, 0)$, and in turn produces $\log \lambda = \log f_0$ or $\lambda = f_0$ in (11.19) as required.

Poisson regression for the tariff model

Under this specification, let us consider n policyholders in the portfolio with the ith policyholder's risk characteristic given by a vector of explanatory variables $\mathbf{x}_i = (1, x_{i1}, x_{i2}, x_{i3})'$, for $i = 1, \ldots, n$. We then recognize (11.19) as

$$\log \lambda_i = \beta_0 + \beta_1 \, x_{i1} + \beta_2 \, x_{i2} + \beta_3 \, x_{i3} = \mathbf{x}_i' \boldsymbol{\beta}, \qquad i = 1, \ldots, n,$$

where $\beta_0, \ldots, \beta_3$ can be mapped to the corresponding log relativities in (11.19). This is exactly the same setup as in (11.15) except for the exposure component. Therefore, by incorporating the exposure in each risk class, a Poisson regression model for this multiplicative tariff model finally becomes

$$\begin{aligned} \log \mu_i &= \log \lambda_i + \log m_i = \log m_i + \beta_0 + \beta_1 \, x_{i1} + \beta_2 \, x_{i2} + \beta_3 \, x_{i3} \\ &= \log m_i + \mathbf{x}_i' \boldsymbol{\beta}, \end{aligned}$$

for $i = 1, \ldots, n$. As a result, the relativities are given by

$$f_0 = e^{\beta_0}, \quad f_{12} = e^{\beta_1}, \quad f_{22} = e^{\beta_2}, \quad \text{and} \quad f_{23} = e^{\beta_3}, \tag{11.20}$$

with $f_{11} = 1$ and $f_{21} = 1$ from the original construction. For the actual dataset, β_i, $i = 0, 1, 2, 3$, is replaced with the *mle* b_i using the method in the technical supplement at the end of this chapter (Section 11.A).

11.3.4 Numerical Examples

We present two numerical examples of Poisson regression. In the first example we construct a Poisson regression model from Table 11.2, which is a dataset of a hypothetical auto insurer. The second example uses an actual industry dataset with more risk factors. As our purpose is to show how a Poisson regression model can be used under a given classification rule, we are not concerned with the quality of the Poisson model fit in this chapter.

Example 11.1: Poisson regression for the illustrative auto insurer. In the last few subsections we considered a dataset of a hypothetical auto insurer with two risk factors, as given in Table 11.2. We now apply a Poisson regression model to this dataset. As done before, we have set $(j, k) = (1, 1)$ as the base tariff cell, so that $f_{11} = f_{21} = 1$. The result of the regression gives the coefficient estimates $(b_0, b_1, b_2, b_3) = (-2.3359, -0.3004, -0.7837, -1.0655)$, which in turn produces the corresponding estimated relativities

$$f_0 = 0.0967, \quad f_{12} = 0.7405, \quad f_{22} = 0.4567 \quad \text{and} \quad f_{23} = 0.3445,$$

from the relation given in (11.20). The `R` script and the output are as follows.

Example 11.2. Poisson regression for Singapore insurance claims data. This actual dataset is a subset of the data used by Frees and Valdez (2008). The data are from the General Insurance Association of Singapore, an organization consisting of non-life insurers in Singapore. These data contains the number of car accidents for $n = 7,483$ auto insurance policies with several categorical explanatory variables and the exposure for each policy. The explanatory variables include four risk factors: the type of the vehicle insured (either automobile (A) or other (O), denoted by `Vtype`), the age of the vehicle in years (`Vage`), gender of the policyholder (`Sex`) and the age of the policyholder (in years, grouped into seven categories, denoted `Age`).

Based on the data description, there are several things to consider before constructing a model. First, there are 3,842 policies with vehicle type A (automobile) and 3,641 policies with other vehicle types. However, age and sex information is available for the policies of vehicle type A only; the drivers of all other types of vehicles are recorded to be aged 21 or less with sex unspecified, except for one policy, indicating that no driver information has been collected for non-automobile vehicles. Second, type A vehicles are all classified as private vehicles and all the other types are not.

When we include these risk factors, we assume all unspecified sex to be male. As the age information is only applicable to type A vehicles, we set the model accordingly. That is, we apply the age variable only to vehicles of type A. Also we used five vehicle age bands, simplifying the original seven bands, by combining vehicle ages 0,1 and 2; the combined band is marked as level 2[3] in the data file. Thus our Poisson model has the following explicit form:

$$\begin{aligned}\log \mu_i = \mathbf{x}_i'\beta + \log m_i &= \beta_0 + \beta_1 I(Sex_i = M) + \sum_{t=2}^{6} \beta_t\, I(Vage_i = t) \\ &+ \sum_{t=7}^{13} \beta_t\, I(Vtype_i = A) \times I(Age_i = t-7) + \log m_i.\end{aligned}$$

The fitting result is given in Table 11.3, for which we have several comments.

- The claim frequency is higher for males by 17.3%, when other rating factors are held fixed. However, this may have been affected by the fact that all unspecified sex has been assigned to male.

- Regarding the vehicle age, the claim frequency gradually decreases as the vehicle age increases, when other rating factors are held fixed. The level starts from 2 for this variable but, again, the numbering is nominal and does

[3] corresponding to `VAgecat1`.

not affect the numerical result.

- The policyholder age variable only applies to type A (automobile) vehicle, and there are no policies in the first age band. We may speculate that younger drivers less than age 21 drive their parents' cars rather than having their own because of high insurance premiums or related regulations. The missing relativity may be estimated by some extrapolation or the professional judgement of the actuary. The claim frequency is the lowest for age band 3 and 4, but gets substantially higher for older age bands, a reasonable pattern seen in many auto insurance loss datasets.

We also note that there is no base level in the policyholder age variable, in the sense that no relativity is equal to 1. This is because the variable is only applicable to vehicle type A. This does not cause a problem numerically, but one may set the base relativity as follows if necessary for other purposes. Since there is no policy in age band 0, we consider band 1 as the base case. Specifically, we treat its relativity as a product of 0.918 and 1, where the former is the common relativity (that is, the common premium reduction) applied to all policies with vehicle type A and the latter is the base value for age band 1. Then the relativity of age band 2 can be seen as $0.917 = 0.918 \times 0.999$, where 0.999 is understood as the relativity for age band 2. The remaining age bands can be treated similarly.

Table 11.3. **Singapore Insurance Claims Data**

Rating factor	Level	Relativity in the tariff	Note
Base value		0.167	f_0
Sex	$1(F)$	1.000	Base level
	$2(M)$	1.173	
Vehicle age	$2(0-2$ yrs)	1.000	Base level
	$3(3-5$ yrs)	0.843	
	$4(6-10$ yrs)	0.553	
	$5(11-15$ yrs)	0.269	
	$6(16+$ yrs)	0.189	
Policyholder age (Only applicable to vehicle type A)	$0(0-21)$	N/A	No policy
	$1(22-25)$	0.918	
	$2(26-35)$	0.917	
	$3(36-45)$	0.758	
	$4(46-55)$	0.632	
	$5(56-65)$	1.102	
	6(65+)	1.179	

Let us try several examples based on Table 11.3. Suppose a male policyholder aged 40 who owns a 7-year-old vehicle of type A. The expected claim frequency for this policyholder is then given by

$$\lambda = 0.167 \times 1.173 \times 0.553 \times 0.758 = 0.082.$$

As another example consider a female policyholder aged 60 who owns a 3-year-old vehicle of type O. The expected claim frequency for this policyholder is

$$\lambda = 0.167 \times 1 \times 0.843 = 0.141.$$

Note that for this policy the age band variable is not used as the vehicle type is not A. The `R` script is given as follows.

As a concluding remark, we comment that Poisson regression is not the only possible count regression model. Actually, the Poisson distribution can be restrictive in the sense that it has a single parameter and its mean and the variance are always equal. There are other count regression models that allow more flexible distributional structure, such as negative binomial regressions and zero-inflated (ZI) regressions; details of these alternative regressions can be found in other texts listed in the next section.

11.4 Risk Classification vs Discrimination

We have so far developed a quantitative model to deal with risk classification. There are however important qualitative aspects of risk classification as well, which have important moral and regulatory and legal implications. We briefly survey various issues related to risk classification in this section; see Frees and Huang (2023) for a more comprehensive treatment.

We start by acknowledging that risk classification, by definition, differentiates or discriminates among insureds or potential buyers based on a wide variety of attributes. That is, insurers divide individuals into subgroups and charge different premiums on the ground that each subgroup, when suitably formed, exhibits a different risk profile and thus produces insurance events (such as medical claims or car accidents) that are different in number and size. In this sense discrimination, of which the meaning is simply treating subgroups differently, is an essential element in insurance business.

Insurers can discriminate among customers in various stages. For example, they may decide not to insure potential customers at the marketing or underwriting stage by excluding particular subgroups intentionally, an issue known as redlining. Also insurers may refuse to renew existing customers, or restrict the insurance coverage. Another form of discrimination can be made by charging unfair prices for certain subgroups. This price discrimination is a standard practice in insurance and not an issue provided that the price differences are made based on the underlying risk level of each subgroup. However, non-risk price discrimination is more problematic in that the price differs for the identical product and coverage. These non-risk rating factors tend to be prohibited in many jurisdictions.

11.4.1 Economic Commodity versus Social Good

While economic arguments view insurance as an economic commodity and thus support insurers' risk-based discrimination, others such as consumer advocates perceive insurance as a social good that should benefit the general public thus argue that discrimination must be avoided especially for disadvantaged groups. These two opposing views can be understood as two extremes of a continual spectrum of fairness when implemented in the real world. To give an idea let us consider the following examples:

- Stock insurance company is located at one end of the spectrum. Here the company issues individual contracts, and insurance is viewed as a collection of separate agreements rather than a collective concept. Actuarially fair pricing is then determined by the expected value of the uncertain event, reflecting the risk transferred from the insured to the insurer. Fairness in this case is defined as each customer paying for their own risk only, supported by economic theory.

- Government-sponsored social insurance is at the other end of the spectrum. Here contracts typically involve subsidies between different subgroups. Governments frequently employ such social policies to redistribute risk or income among individuals, though adherence to the principle of actuarial fairness can vary significantly depending on the target level of the redistribution.
- Group insurance lies in the of the spectrum. For example, consider a disability income contract issued to the employees of an employer. In this case premium differentials by risk factors are not a major issue if the employer pays all or a major portion of the premiums.

Clearly the issue of discrimination and fairness in insurance is a multi-layered problem; it involves not only technical modeling but is also affected by the social consensus depending on the goal and characteristics of the insurance program.

11.4.2 Information Asymmetry

Discrimination in insurance may also arise from information asymmetry which means that insurers and the current or potential customers have unequal knowledge or access to relevant information on the underlying risk.

Adverse selection described in Introduction of this chapter is an example of information asymmetry. For an insurer adverse selection can occur when customers know better about their own risks than the insurer or when other competitors in the market have better knowledge about the risk of the customers, as illustrated at the beginning of this chapter. Generally speaking adverse selection can be reduced as more information on the customers is made available.

Another type of information asymmetry is moral hazard. In a typical case of moral hazard, policyholders become more risk-seeking or less cautious about the risk when the corresponding risk has already been insured. In other words, the insureds have the incentive to take on more risk because of the safety net provided by the insurance, leading to raised costs for the insurer. One remedy to moral hazard is to offer incentives to policyholders so that they can act more responsively or the exposure of the risk itself can be reduced. To illustrate, consider a customer who bought a homeowner's insurance contract and thus becomes less careful about fire and theft. To mitigate such moral hazard, the insurer may offer some premium discount on the condition that fire and security alarms be installed in the house. Some moral hazard applies to insurer's side. For example, insurers may collect protected or sensitive variables and create pricing models to unfairly discriminate customers. The moral hazard of the insurer's side is generally managed and prevented based on insurance regulations and laws.

Another recently emerging area of information asymmetry is the knowledge imbalance created from big data models used by insurers. Consumer advocates point out that the information gap between the customers and insurers will get wider as more big data are available only to insurers. As a result, insurers can cherry pick potentially more profitable customers, and customers without big data tools will be unable to access to such information and thus at disadvantage. This implies that free market competition between insurers may be insufficient to protect policyholders if the insurers collectively monopolize the exclusive knowledge on the customers.

11.4.3 Sensitive Variables and Regulation

Some attributes or variables used in risk classification may be perceived to be unfair or sensitive. In the literature the following list of criteria is often considered in deciding whether an attribute is acceptable or fair as a rating variable.

- Control: An attribute that can be controlled by an insured is generally considered to be an acceptable variable to be used for rating purposes. Smoking status is an example of such attribute. In contrast, race and gender at birth cannot be controlled by insureds.
- Mutability: Some attributes change over time, but they may be used as rating variables if they are deemed fair to everyone. For example, aging applies fairly to us all over the course of a lifetime.
- Statistical Discrimination: If a variable does not have predictive value of an underlying risk, it is generally viewed as unacceptable.
- Causality: A variable known to cause an insured event can be used for rating purposes, but establishing a causal relationship may not be always easy because it requires strong evidence beyond a simple association.
- Limiting or Reversing the Effects of Past Prejudice: If an attribute is related to negative stereotypes or otherwise disadvantaged groups, it may not be used for rating purposes.
- Inhibiting Socially Valuable Behavior: If an insurer's use of an attribute prevents socially desirable behaviors, it may not be used for rating purposes. For example, U.S. laws prohibit insurers from discriminating on the basis of intimate partner violence because such reporting could dissuade victims of violence from seeking needed medical care or police intervention.

In light of these complications, many jurisdictions have so-called rate regulations which prohibit insurers from engaging problematic pricing practices. For example, in the US, the model rating law of the National Association of Insurance Commissioners (NAIC, 2010) says that "rates shall not be excessive, inadequate or unfairly discriminatory." It further notes that "unfair discrimination exists if, after allowing for practical limitations, price differentials fail to reflect equitably the differences in expected losses and expenses." Different jurisdictions maintain different standards on the strictness of rate regulations. In countries where rate regulations are heavily enforced the regulators prescribe the actual rates whereas in other countries the regulators may only require approval of rates.

11.4.4 Big Data Models and Proxy Discrimination

Big data models such as deep learning, machine learning and AI algorithms are now ubiquitous in virtually every area of our society. These models are known to detect new patterns and connections that were previously unknown using advanced algorithms and various data sources.

From the viewpoint of risk classification or rating discrimination in insurance, it is argued that big data models would bring significant changes in privacy and proxy discrimination. The issue of privacy protection is well known. When data from various sources, e.g., the location information from GPS, wearable devices, social networks, credit cards, are combined together, big data models may reveal the identity of individuals or their sensitive information. Given that these collected data consists of seemingly innocuous or voluntarily provided variables, the potential risk of privacy breach and sensitive variable fabrication is becoming a reality.

Proxy discrimination arises when insurers discriminate based on a facially neutral attribute that is highly correlated with a protected and sensitive information. By employing these proxy variables in pricing models insurers could get the same quantitative results that would be obtained from using the protected variables directly. This is problematic because insurers are able to effectively use prohibited variables, without actually violating the rate regulation. Discovery or synthesis of such proxies can be made through statistical and big data models; proxy discrimination is harder to detect in the latter models as their algorithms tend to be less transparent. Though it is impossible to eliminate proxy discrimination completely, several strategies have been suggested to mitigate it:

1. Community Rating: If all policyholders pay the same price, proxy

discrimination can be eliminated. This kind of rating can be found in social insurance programs, but rare in general.

2. Approved Variables: Regulators may specify a set of variables allowed in rating, prohibiting others. For example, in the US individual health insurance market under the Affordable Care Act, insurers are allowed to use only four rating factors: (1) whether a plan covers an individual or family, (2) geographic area, (3) age, and (4) smoking status.
3. Actuarial Justification: In this strategy regulators specify a set of protected or sensitive variables that should not be used in rating. In addition, outside these variables, only variables that are actuarially justifiable or statistically significant are allowed.
4. Limited Prohibitions. Alternatively, regulators specify a set of protected variables and no restriction is made for other variables outside this set.
5. No Restrictions. In this extreme strategy regulators impose no prohibitions on rating variables, which actually is the case for most commercial insurance lines.

In practice the most viable option would be to adopt the third or fourth strategy with suitable modifications, with some disclosure requirement for the pricing model and data source used in the rating process.

11.5 Exercises

11.1. Regarding Table 11.1 answer the following.
(a) Verify the mean values in the table.
(b) Verify the number in equation (11.2).
(c) Produce the fitted Poisson counts for each smoking status in the table.

11.2. In a Poisson regression formulation (11.10), consider using $\mu_i = \mathrm{E}\,(y_i|\mathbf{x}_i) = (\mathbf{x}_i'\beta)^2$, for $i = 1, \ldots, n$, instead of the exponential function. What potential issue would you have?

11.6 Further Resources and Contributors

Further Reading and References

Poisson regression is a special member of a more general regression model class known as the generalized linear model (GLM). The GLM develops a unified regression framework for datasets when the response variables are continuous, binary or discrete. The classical linear regression model with a normally distributed error is also a member of the GLM. There are many standard statistical texts dealing with the GLM, including McCullagh and Nelder (1989). More accessible texts are Dobson and Barnett (2008), Agresti (1996) and Faraway (2016). For actuarial and insurance GLM applications, see Frees (2009), De Jong and Heller (2008). Also, Ohlsson and Johansson (2010) discusses GLM in non-life insurance pricing context with tariff analyses.

In fact there is a notable historical connection between the GLM and an influential actuarial model. In 1960s the actuarial community has developed an auto ratemaking model that produces coherent and consistent rates across subgroups in both additive and multiplicative form. This method, known as Bailey minimum bias method (Bailey and Simon, 1960 and Bailey, 1963), turned out to be equivalent to the solution of a statistical model known as the GLM which made its first appearance in 1970s by Nelder and Wedderburn (1972). This strong connection has helped actuaries use and adopt the GLM models in a wide range of actuarial problems; see Frees, Derrig, Meyers (2014) and references therein for a more detailed historical note on this.

Contributor

- **Joseph H. T. Kim**, Yonsei University, is the principal author of the initial version of this chapter. Email: jhtkim@yonsei.ac.kr for chapter comments and suggested improvements.
- Chapter reviewers include: Chun Yong Chew, Lina Xu, Jeffrey Zheng.

TS 11.A. Estimating Poisson Regression Models

The principles of maximum likelihood estimation (*mle*) are introduced in Sections 3.4.2 and 4.4.2, defined in Section 17.2.2, and theoretically developed in Chapter 19. Here we present the *mle* procedure of Poisson regression so that the reader can see how the explanatory variables are treated in maximizing the likelihood function in the regression setting.

Maximum Likelihood Estimation for Individual Data

In Poisson regression the varying Poisson mean is determined by parameters

β_i's, as shown in (11.15). In this subsection we use the maximum likelihood method to estimate these parameters. Again, we assume that there are n policyholders and the ith policyholder is characterized by $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})'$ with the observed loss count y_i. Then, from (11.14) and (11.15), the log-likelihood function of vector $\beta = (\beta_0, \ldots, \beta_k)$ is given by

$$\begin{aligned}\log L(\beta) = l(\beta) &= \sum_{i=1}^{n} (-\mu_i + y_i \log \mu_i - \log y_i!) \\ &= \sum_{i=1}^{n} (-m_i \exp(\mathbf{x}_i'\beta) + y_i (\log m_i + \mathbf{x}_i'\beta) - \log y_i!) \qquad (11.21)\end{aligned}$$

To obtain the *mle* of $\beta = (\beta_0, \ldots, \beta_k)'$, we differentiate[4] $l(\beta)$ with respect to vector β and set it to zero:

$$\left.\frac{\partial}{\partial \beta} l(\boldsymbol{\beta})\right|_{\beta = \mathbf{b}} = \sum_{i=1}^{n} (y_i - m_i \exp(\mathbf{x}_i'\mathbf{b}))\, \mathbf{x}_i = \mathbf{0}. \qquad (11.22)$$

Numerically solving this equation system gives the *mle* of β, denoted by $\mathbf{b} = (b_0, b_1, \ldots, b_k)'$. Note that, as $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})'$ is a column vector, equation (11.22) is a system of $k+1$ equations with both sides written as column vectors of size $k+1$. If we denote $\hat{\mu}_i = m_i \exp(\mathbf{x}_i'\mathbf{b})$, we can rewrite (11.22) as

$$\sum_{i=1}^{n} (y_i - \hat{\mu}_i)\, \mathbf{x}_i = \mathbf{0}.$$

Since the solution $\mathbf{b}$ satisfies this equation, it follows that the first among the array of $k+1$ equations, corresponding to the first constant element of $\mathbf{x}_i$, yields

$$\sum_{i=1}^{n} (y_i - \hat{\mu}_i) \times 1 = 0,$$

which implies that we must have

$$n^{-1} \sum_{i=1}^{n} y_i = \bar{y} = n^{-1} \sum_{i=1}^{n} \hat{\mu}_i.$$

This is an interesting property saying that the average of the individual losses, $\bar{y}$, is same as the average of the estimated values. That is, the sample mean is preserved under the fitted Poisson regression model.

Maximum Likelihood Estimation for Grouped Data

Sometimes the data are not available at the individual policy level. For example,

[4] We use matrix derivative here.

Table 11.2 provides collective loss information for each risk class after grouping individual policies. When this is the case, y_i and m_i, the quantities needed for the *mle* calculation in (11.22), are unavailable for each i. However this does not pose a problem as long as we have the total loss counts and total exposure for each risk class.

To elaborate, let us assume that there are K different risk classes, and further that, in the kth risk class, we have n_k policies with the total exposure $m_{(k)}$ and the average loss count $\bar{y}_{(k)}$, for $k = 1, \ldots, K$; the total loss count for the kth risk class is then $n_k\,\bar{y}_{(k)}$. We denote the set of indices of the policies belonging to the kth class by C_k. As all policies in a given risk class share the same risk characteristics, we may denote $\mathbf{x}_i = \mathbf{x}_{(k)}$ for all $i \in C_k$. With this notation, we can rewrite (11.22) as

$$\begin{aligned}\sum_{i=1}^{n} (y_i - m_i \exp(\mathbf{x}_i'\mathbf{b}))\,\mathbf{x}_i &= \sum_{k=1}^{K} \left\{ \sum_{i \in C_k} (y_i - m_i \exp(\mathbf{x}_i'\mathbf{b}))\,\mathbf{x}_i \right\} \\ &= \sum_{k=1}^{K} \left\{ \sum_{i \in C_k} \Big(y_i - m_i \exp(\mathbf{x}_{(k)}'\mathbf{b})\Big)\,\mathbf{x}_{(k)} \right\} \\ &= \sum_{k=1}^{K} \left\{ \Big(\sum_{i \in C_k} y_i - \sum_{i \in C_k} m_i \exp(\mathbf{x}_{(k)}'\mathbf{b}) \Big) \mathbf{x}_{(k)} \right\} \\ &= \sum_{k=1}^{K} \Big(n_k\,\bar{y}_{(k)} - m_{(k)} \exp(\mathbf{x}_{(k)}'\mathbf{b}) \Big) \mathbf{x}_{(k)} = 0. \qquad (11.23)\end{aligned}$$

Since $n_k\,\bar{y}_{(k)}$ in (11.23) represents the total loss count for the kth risk class and $m_{(k)}$ is its total exposure, we see that for Poisson regression the *mle* $\mathbf{b}$ is the same whether if we use the individual data or the grouped data.

Information matrix. Section 19.1 defines information matrices. Taking second derivatives to (11.21) gives the information matrix of the *mle* estimators,

$$\mathbf{I}(\boldsymbol{\beta}) = -\mathrm{E}\left(\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} l(\boldsymbol{\beta}) \right) = \sum_{i=1}^{n} m_i \exp(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i\mathbf{x}_i' = \sum_{i=1}^{n} \mu_i \mathbf{x}_i\mathbf{x}_i'. \qquad (11.24)$$

For actual datasets, μ_i in (11.24) is replaced with $\hat{\mu}_i = m_i \exp(\mathbf{x}_i'\mathbf{b})$ to estimate the relevant variances and covariances of the *mle* $\mathbf{b}$ or its functions.

For grouped datasets, we have

$$\mathbf{I}(\boldsymbol{\beta}) = \sum_{k=1}^{K} \left\{ \sum_{i \in C_k} m_i \exp(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i\mathbf{x}_i' \right\} = \sum_{k=1}^{K} m_{(k)} \exp(\mathbf{x}_{(k)}'\boldsymbol{\beta})\mathbf{x}_{(k)}\mathbf{x}_{(k)}'.$$

TS 11.B. Selecting Rating Factors

A complete discussion of rating factor selection is beyond the scope of this book. In addition to technical analyses, you have to think carefully about the type of business (personal, commercial) as well as the regulatory landscape. Nonetheless, a broad overview of some key concerns may serve to ground the reader as one thinks about the pricing of insurance contracts.

Statistical Criteria

From an analyst's perspective, the discussion starts with the statistical significance of a rating factor. If the factor is not statistically significant, then the variable is not even worthy of consideration for inclusion in a rating plan. The statistical significance is judged not only on an in-sample basis but also on how well it fares on an out-of-sample basis, as per our discussion in Chapter 6.

It is common in insurance applications to have many rating factors. Handling multivariate aspects can be difficult with traditional univariate methods. Analysts employ techniques such as *generalized linear models* as described in Section 11.3.

Rating factors are introduced to create cells that contain similar risks. A rating group should be large enough to measure costs with sufficient accuracy. There is an inherent trade-off between theoretical accuracy and homogeneity.

As an example, most insurers charge the same automobile insurance premiums for drivers between the ages of 30 and 50, not varying the premium by age. Presumably costs do not vary much by age, or cost variances are due to other identifiable factors.

Operational Criteria

From a business perspective, statistical criteria only provide a starting point for discussions of potential inclusion of rating factors. Inclusion of a rating factor must also induce economically meaningful results. From an insured's perspective, if differentiation by a factor produces little change in a rate then it is not worth including. From an insurer's perspective, the inclusion of a factor should help segment the marketplace in a way that helps attract the business that they seek.

Rating factors should also be objective, inexpensive to administer, and verifiable. For example, automobile insurance underwriters often talk of “maturity” and “responsibility” as important criteria for youthful drivers. Yet, these are difficult to define objectively and to apply consistently. As another example, in automobile it has long been known that amount of miles (or kilometers) driven is an excellent rating factor. However, insurers have been reluctant to adopt this factor because it is subject to abuse. Historically, driving mileage has not been used because of the difficulty in verifying this variable (it is far too easy to alter the car's odometer to change reported mileage). Going forward, modern day drivers and cars are equipped with global positioning devices and other equipment that allow insurers to use distance driven as a rating factor because it can be verified.

Rating Factors from the Perspective of a Consumer

Insurance companies sell insurance products to a variety of consumers; consequently, companies are affected by public perception. On the one hand, free market competition dictates rating factors that insurers use, as is common in commercial insurance. On the other hand, insurance may be required by law. This is common in personal insurance such as third party automobile liability and homeowners. In these instances, the mandatory and de facto mandatory purchase of insurance may mean that free market competition is insufficient to protect policyholders. Here, the following items affect the social acceptability of using a particular risk characteristic as a rating variable:

- Affordability - introduction of some variables may be mitigated by resulting high costs of insurance.
- Causality - other things being equal, a rating variable is easier to justify if there is a "causal" relationship with losses. A good example is the effects of smoking in life insurance. For many years, this factor was viewed with suspicion by the industry. However, over time, scientific evidence provided overwhelming evidence as this an important predictor of mortality.
- Controllability - A controllable variable is one that is under the control of the insured, e.g., installing burglar alarms. The use of controllable rating variables encourages accident prevention.
- Privacy concerns - people are reluctant to disclose personal information. In today's world with increasing emphasis on social media and the availability of personal information, consumer advocates are concerned that the benefits of big data skew heavily in insurers' favor. They reason that insureds do not have equivalent new tools to compare quality of coverage/policies and performance of insurance companies.

Example: Youthful Drivers. In some cases, a particular risk characteristic may identify a small group of insureds whose risk level is extremely high, and if used as a rating variable, the resulting premium may be unaffordable for that high-risk class. To the extent that this occurs, companies may wish to or be required by regulators to combine classes and introduce subsidies. For example, 16-year-old drivers are generally higher risk than 17-year-old drivers. Some companies have chosen to use the same rates for 16- and 17-year-old drivers to minimize the affordability issues that arise when a family adds a 16-year-old to the auto policy.

Societal Effects of Rating Factors

With public discussions of rating factors, it is also important to think about the societal effects of classification.

For example, does a rating variable encourage "good" behavior? As an example, we return to the use of distance driven as a rating factor. Many people advocate for including this variable as a factor. The motivation is that if insurance, like fuel, is priced based on distance driven, this will induce consumers to reduce the amount driven, thereby benefiting society.

One can consider other aspects of societal effects of classification, see, for example, Niehaus and Harrington (2003):

- Re-distributive Effects - provide a cross-subsidy from e.g., high risks to low risks
- Classification Costs - Money spent by society, insurers, to classify people appropriately.

Legal Criteria

For example, some states have statutes prohibiting the use of gender in rating insurance while others permit it as a rating variable. As a result, an insurer writing in multiple states may include gender as a rating variable in those states where it is permitted, but not include it in a state that prohibits its use for rating.

If allowed by law, the company may continue to charge the average rate but utilize the characteristic to identify, attract, and select the lower-risk insureds that exist in the insured population; this is called *skimming the cream*. See Frees and Huang (2023) for a broad discussion of the discrimination in pricing.

12

Experience Rating Using Credibility Theory

Chapter Preview. This chapter introduces credibility theory as an important actuarial tool for estimating pure premiums, frequencies, and severities for individual risks or classes of risks. Credibility theory provides a convenient framework for combining the experience for an individual risk or class with other data to produce more stable and accurate estimates. Several models for calculating credibility estimates will be discussed including Bühlmann, Bühlmann-Straub, limited fluctuation, and nonparametric and semiparametric credibility methods. The chapter will also show a connection between credibility theory and Bayesian estimation which was introduced in Chapter 9, Bayesian Inference and Modeling.

12.1 Introduction to Applications of Credibility Theory

What premium should be charged to provide insurance? The answer depends upon the exposure to the risk of loss. A common method to compute an insurance premium is to rate an insured using a classification rating plan. A classification plan is used to select an insurance rate based on an insured's rating characteristics such as geographic territory, age, etc. All classification rating plans use a limited set of criteria to group insureds into a "class" and there will be variation in the risk of loss among insureds within the class.

An experience rating plan attempts to capture some of the variation in the risk of loss among insureds within a rating class by using the insured's own loss experience to complement the rate from the classification rating plan. One way to do this is to use a credibility weight Z with $0 \leq Z \leq 1$ to compute

$$\hat{R} = Z\bar{X} + (1 - Z)M,$$

$$\begin{aligned} \hat{R} &= \text{credibility weighted rate for risk,} \\ \bar{X} &= \text{average loss for the risk over a specified time period,} \\ M &= \text{the rate for the classification group, often called the manual rate.} \end{aligned}$$

For a risk whose loss experience is stable from year to year, Z might be close to 1. For a risk whose losses vary widely from year to year, Z may be close to 0.

Credibility theory is also used for computing rates for individual classes within a classification rating plan. When classification plan rates are being determined, some or many of the groups may not have sufficient data to produce stable and reliable rates. The actual loss experience for a group will be assigned a credibility weight Z and the complement of credibility $1 - Z$ may be given to the average experience for risks across all classes. Or, if a class rating plan is being updated, the complement of credibility may be assigned to the current class rate. Credibility theory can also be applied to the calculation of expected frequencies and severities.

Computing numeric values for Z requires analysis and understanding of the data. What are the variances in the number of losses and sizes of losses for risks? What is the variance between expected values across risks?

12.2 Bühlmann Credibility

In this section, you learn how to:

- Compute a credibility-weighted estimate for the expected loss for a risk or group of risks.
- Determine the credibility Z assigned to observations.
- Calculate the values required in Bühlmann credibility including the Expected Value of the Process Variance (EPV), Variance of the Hypothetical Means (VHM) and collective mean μ.

A classification rating plan groups policyholders together into classes based on risk characteristics. Although policyholders within a class have similarities, they are not identical and their expected losses will not be exactly the same. An experience rating plan can supplement a class rating plan by credibility weighting an individual policyholder's loss experience with the class rate to produce a more accurate rate for the policyholder. Chapter 15 Experience Rating using Bonus-Malus provides examples of rating plans that adjust a policyholder's rate to recognize their loss experience.

The Bühlmann credibility model introduced in this section is often called greatest accuracy credibility, least-squares credibility, or Bayesian credibility.

In this presentation a risk parameter θ will be assigned to each policyholder. Losses X for the policyholder with parameter θ will have a pdf $f_{X|\Theta=\theta}(x)$ and mean

$$\mu(\theta) = \mathrm{E}_X(X|\theta) = \int x f_{x|\Theta=\theta}(x)\,dx \tag{12.1}$$

and variance

$$\sigma^2(\theta) = \mathrm{Var}_X(X|\theta) = \int (x-\mu(\theta))^2 f_{x|\Theta=\theta}(x)\,dx. \tag{12.2}$$

The integrals are over the support for the distributions. Losses X can represent pure premiums, aggregate losses, number of claims, claim severities, or some other measure of loss for a period of time, often one year. Risk parameter θ may be continuous or discrete and may be multivariate depending on the model. For a randomly selected risk the risk parameter θ is unknown but the probability density function for θ is modeled with $f_\Theta(\theta)$. Averaging across the policyholders in the class the collective mean loss is

$$\mu = \mathrm{E}_\Theta[\mathrm{E_X}(X|\theta)] = \int f_\Theta(\theta)\mu(\theta)d\theta = \int f_\Theta(\theta)\int x f_{x|\Theta=\theta}(x)\,dxd\theta. \tag{12.3}$$

Example 12.2.1. The number of claims X for an insured in a class has a Poisson distribution with mean $\theta > 0$. The risk parameter θ is exponentially distributed within the class with pdf $f(\theta) = e^{-\theta}$. What is the expected number of claims for an insured chosen at random from the class?

Example Solution. Random variable X is Poisson with mean θ so $\mu(\theta) = \mathrm{E}_X(X|\theta) = \theta$. The expected number of claims for a randomly chosen insured is $\mu = \mathrm{E}_\Theta(\mu(\theta)) = \mathrm{E}_\Theta(\mathrm{E}_X(X|\theta)) = \mathrm{E}_\Theta(\theta) = \int_0^\infty \theta e^{-\theta}d\theta = 1$.

In the prior example the risk parameter θ is a continuous random variable with an exponential distribution. In the next example there are three types of risks and the risk parameter has a discrete distribution.

Example 12.2.2. For any risk (policyholder) in a population the number of losses N in a year has a Poisson distribution with parameter λ. Individual loss amounts X_i for a risk are independent of N and are iid with Type II Pareto distribution $F(x) = 1 - [\theta/(x+\theta)]^\alpha$. There are three types of risks in the population as follows:

Risk Type	Percentage of Population	Poisson Parameter	Pareto Parameters
A	50%	$\lambda = 0.5$	$\theta = 1000, \alpha = 2.0$
B	30%	$\lambda = 1.0$	$\theta = 1500, \alpha = 2.0$
C	20%	$\lambda = 2.0$	$\theta = 2000, \alpha = 2.0$

If a risk is selected at random from the population, what is the expected aggregate loss in a year?

Example Solution. The expected number of claims for a risk is $\mathrm{E}_N(N|\lambda)=\lambda$. The expected value for a Pareto distributed random variable is $\mathrm{E}_X(X|\theta,\alpha)=\theta/(\alpha-1)$. The expected value of the aggregate loss random variable $S = X_1 + \cdots + X_N$ for a risk with parameters λ, α, and θ assuming independence of N and X_i's is $\mathrm{E}(S|\lambda,\theta,\alpha) = \mathrm{E}_N(N|\lambda)\mathrm{E}_X(X|\theta,\alpha) = \lambda\theta/(\alpha-1)$. The expected aggregate loss for a risk of type A is $\mathrm{E}(S_\mathrm{A})$=(0.5)(1000)/(2-1)=500. The expected aggregate loss for a risk selected at random from the population is $\mathrm{E}(S) =$ 0.5[(0.5)(1000)]+0.3[(1.0)(1500)]+0.2[(2.0)(2000)]=1500.

What is the risk parameter for a risk (policyholder) in the prior example? One could say that the risk parameter has three components $(\lambda, \theta, \alpha)$ with possible values (0.5,1000,2.0), (1.0,1500,2.0), and (2.0,2000,2.0) depending on the type of risk.

Note that in both of the examples the risk parameter is a random quantity with its own probability distribution. We do not know the value of the risk parameter for a randomly chosen risk.

12.2.1 Credibility-Weighted Estimate for the Expected Loss

If a policyholder with risk parameter θ has losses $x_1, \ldots, x_n$ during n time periods then the goal is to find $\mathrm{E}_\Theta(\mu(\theta)|x_1, \ldots, x_n)$, the conditional expectation of $\mu(\theta)$ given observations $x_1, \ldots, x_n$. Section 12.3, Bayesian Inference and Bühlmann Credibility explains how to evaluate $\mathrm{E}_\Theta(\mu(\theta)|x_1, \ldots, x_n)$ using Bayesian inference.

The Bühlmann credibility model calculates a linear approximation $\hat{\mu}(\theta) = Z\bar{x} + (1-Z)\mu$ to estimate $\mathrm{E}_\Theta(\mu(\theta)|x_1, \ldots, x_n)$ with $\bar{x} = (x_1 + \ldots + x_n)/n$. We can rewrite this as $\hat{\mu}(\theta) = a + b\bar{x}$ which makes it obvious that the credibility estimate is a linear function of the mean.

In the Bühlmann model, $\mathrm{E}_\Theta(\mu(\theta)|X_1, \ldots, X_n)$ is approximated by the linear function $a + b\bar{X}$ and constants a and b are calculated to minimize the square of the difference between these two quantities

$$G(a,b) = \mathrm{E}_X([\mathrm{E}_\Theta(\mu(\theta)|X_1,\ldots,X_n) - (a+b\bar{X})]^2), \tag{12.4}$$

hence the alternative name least-squares credibility. Minimizing the expectation yields $b = n/(n+K)$ and $a = (1-b)\mu$. Quantity n is the number of observations and $\mu = \mathrm{E}_\Theta(\mu(\theta))$ is the population mean. For the moment we will assign K the mysterious equation K = (Expected Value of the Process Variance) / (Variance of the Hypothetical Means)=EPV/VHM and will clarify the meaning at the beginning of the next section. More details about this model and calculation of a and b can be found in references (Bühlmann, 1967), (Bühlmann and Gisler, 2005), (Klugman et al., 2012), and (Tse, 2009).

The Bühlmann credibility-weighted estimate for $\mathrm{E}_\Theta(\mu(\theta)|x_1,\ldots,x_n)$ for the policyholder is

$$\hat{\mu}(\theta) = Z\bar{x} + (1-Z)\mu \tag{12.5}$$

with

θ = a risk parameter that identifies a policyholder's risk level

$\hat{\mu}(\theta)$ = estimated expected loss for a policyholder with parameter θ and loss experience $\bar{x}$

$\bar{x}$ = $(x_1 + \cdots + x_n)/n$ is the average of n observations of the policyholder

Z = credibility assigned to n observations $= n/(n+K)$

K = EPV/VHM

μ = the expected loss for a randomly chosen policyholder in the class.

For a selected policyholder, random variables X_j are assumed to be iid for $j = 1,\ldots,n$ because it is assumed that the policyholder's exposure to loss is not changing through time and $\mathrm{E}_X(\bar{X}|\theta) = \mathrm{E}_X(X_j|\theta) = \mu(\theta)$.

If a policyholder is randomly chosen from the class and there is no loss information about the risk then the expected loss is $\mu = \mathrm{E}_\Theta(\mu(\theta))$ where the expectation is taken over all θ's in the class. In this situation $Z = 0$ and the expected loss is $\hat{\mu}(\theta) = \mu$ for the risk. The quantity μ can also be written as $\mu = \mathrm{E}(X_j)$ or $\mu = \mathrm{E}(\bar{X})$ and is referred to as the overall mean, population mean, or collective mean. Note that $\mathrm{E}(X_j)$ is evaluated with the law of total expectation: $\mathrm{E}(X_j) = \mathrm{E}_\Theta[\mathrm{E}_\mathrm{X}(X_j|\theta)]$.

Although formula (12.5) was introduced using experience rating as an example, the Bühlmann credibility model has wider application. Suppose that a rating plan has multiple classes. Credibility formula (12.5) can be used to determine individual class rates. The overall mean μ would be the average loss for all classes combined, $\bar{x}$ would be the experience for the individual class, and $\hat{\mu}(\theta)$ would be the estimated loss for the class.

12.2.2 Credibility *Z*, *EPV*, and *VHM*

When computing the credibility estimate $\hat{\mu}(\theta) = Z\bar{X} + (1 - Z)\mu$, how much weight Z should go to experience $\bar{X}$ and how much weight $(1 - Z)$ to the overall mean μ? In Bühlmann credibility there are three factors that need to be considered:

1. How much variation is there in a single observation X_j for a selected risk? With $\bar{X} = (X_1 + \cdots + X_n)/n$ and assuming that the observations are iid conditional on θ, it follows that $\text{Var}_X(\bar{X}|\theta) = \text{Var}_X(X_j|\theta)/n$. For larger $\text{Var}_X(\bar{X}|\theta)$ less credibility weight Z should be given to experience $\bar{X}$. The Expected Value of the Process Variance, abbreviated EPV, is the expected value of $\text{Var}_X(X_j|\theta)$ across all risks:
$$EPV = \text{E}_\Theta(\text{Var}_X(X_j|\theta)).$$
Because $\text{Var}_X(\bar{X}|\theta) = \text{Var}_X(X_j|\theta)/n$ it follows that $\text{E}_\Theta(\text{Var}_X(\bar{X}|\theta)) = EPV/n$.
2. How homogeneous is the population of risks whose experience was combined to compute the overall mean μ? If all the risks are similar in loss potential then more weight $(1 - Z)$ would be given to the overall mean μ because μ is the average for a group of similar risks whose means $\mu(\theta)$ are not far apart. The homogeneity or heterogeneity of the population is measured by the Variance of the Hypothetical Means with abbreviation VHM:
$$VHM = \text{Var}_\Theta(\text{E}_X(X_j|\theta)) = \text{Var}_\Theta(\text{E}_X(\bar{X}|\theta)).$$
Note that we used $\text{E}_X(\bar{X}|\theta) = \text{E}_X(X_j|\theta)$ for the second equality.
3. How many observations n were used to compute $\bar{X}$? A larger sample would infer a larger Z.

Example 12.2.3. The number of claims N in a year for a risk in a population has a Poisson distribution with mean $\lambda > 0$. The risk parameter λ is uniformly distributed over the interval $(0, 2)$. Calculate the EPV and VHM for the population.

Example Solution. Random variable N is Poisson with parameter λ so $\text{Var}(N|\lambda) = \lambda$. The Expected Value of the Process variance is $EPV =$

$\mathrm{E}(\mathrm{Var}(N|\lambda)) = \mathrm{E}(\lambda) = \int_0^2 \lambda \frac{1}{2} d\lambda = 1$. The Variance of the Hypothetical Means is $VHM = \mathrm{Var}(\mathrm{E}(N|\lambda)) = \mathrm{Var}(\lambda) = \mathrm{E}(\lambda^2) - (\mathrm{E}(\lambda))^2 = \int_0^2 \lambda^2 \frac{1}{2} d\lambda - (1)^2 = \frac{1}{3}$.

The Bühlmann credibility formula includes values for n, EPV, and VHM:

$$Z = \frac{n}{n+K} \quad , \quad K = \frac{EPV}{VHM}. \tag{12.6}$$

If the VHM increases then Z increases. If the EPV increases then Z gets smaller. Credibility Z asymptotically approaches 1 as the number of observations n goes to infinity.

If you multiply the numerator and denominator of the Z formula by (VHM/n) then Z can be rewritten as

$$Z = \frac{VHM}{VHM + (EPV/n)}.$$

The number of observations n is captured in the term (EPV/n).

Example 12.2.4. The law of total variance can be written as $\mathrm{Var}(Y) = \mathrm{E}(\mathrm{Var}[Y|X]) + \mathrm{Var}(\mathrm{E}[Y|X])$. Show that $\mathrm{Var}(\bar{X}) = VHM + (EPV/n)$ and derive a formula for Z in terms of $\bar{X}$.

Example Solution. The quantity $\mathrm{Var}(\bar{X})$ is called the unconditional variance or the total variance of $\bar{X}$. The law of total variance says

$$\mathrm{Var}(\bar{X}) = \mathrm{E}_\Theta(\mathrm{Var}_X(\bar{X}|\theta)) + \mathrm{Var}_\Theta(\mathrm{E}_X(\bar{X}|\theta)).$$

In bullet (1) at the beginning of this section we showed $\mathrm{E}_\Theta(\mathrm{Var}_X(\bar{X}|\theta)) = EPV/n$. In bullet (2), $\mathrm{Var}_\Theta(\mathrm{E}_X(\bar{X}|\theta)) = VHM$. Reordering the right hand side gives $\mathrm{Var}(\bar{X}) = VHM + (EPV/n)$. Another way to write the formula for credibility Z is $Z = \mathrm{Var}_\Theta(\mathrm{E}_X(\bar{X}|\theta))/\mathrm{Var}(\bar{X})$. This implies $(1-Z) = \mathrm{E}_\Theta(\mathrm{Var}_X(\bar{X}|\theta))/\mathrm{Var}(\bar{X})$.

The following long example and solution demonstrate how to compute the credibility-weighted estimate with frequency and severity data.

Example 12.2.5. For any risk in a population the number of losses N in a year has a Poisson distribution with parameter λ. Individual loss amounts X for a selected risk are independent of N and are *iid* with exponential distribution $F(x) = 1 - e^{-x/\beta}$. There are three types of risks in the population as shown below. A risk was selected at random from the population and all losses were recorded over a five-year period. The total amount of losses over the five-year period was 5,000. Use Bühlmann credibility to estimate the annual expected aggregate loss for the risk.

Risk Type	Percentage of Population	Poisson Parameter	Exponential Parameter
A	50%	$\lambda = 0.5$	$\beta = 1000$
B	30%	$\lambda = 1.0$	$\beta = 1500$
C	20%	$\lambda = 2.0$	$\beta = 2000$

Example Solution. Because individual loss amounts X are exponentially distributed, $\mathrm{E}(X|\beta) = \beta$ and $\mathrm{Var}(X|\beta) = \beta^2$. For aggregate loss $S = X_1 + \cdots + X_N$, the mean is $\mathrm{E}(S) = \mathrm{E}(N)\mathrm{E}(X)$ and process variance is $\mathrm{Var}(S) = \mathrm{E}(N)\mathrm{Var}(X) + [\mathrm{E}(X)]^2\mathrm{Var}(N)$.

With Poisson frequency and exponentially distributed loss amounts, $\mathrm{E}(S|\lambda, \beta) = \lambda\beta$ and $\mathrm{Var}(S|\lambda, \beta) = \lambda\beta^2 + \beta^2\lambda = 2\lambda\beta^2$.

Population mean μ: Risk means are μ(A)=0.5(1000)=500; μ(B)=1.0(1500)=1500; μ(C)=2.0(2000)=4000; and μ=0.50(500)+0.30(1500)+0.20(4000)=1,500.

VHM: VHM=0.50(500 $-$ 1500)^2 + 0.30(1500 $-$ 1500)^2 + 0.20(4000 $-$ 1500)^2=1,750,000.

EPV: Process variances are $\sigma^2(A) = 2(0.5)(1000)^2 = 1,000,000$; $\sigma^2(B) = 2(1.0)(1500)^2 = 4,500,000$; $\sigma^2(C) = 2(2.0)(2000)^2 = 16,000,000$; and EPV=0.50(1,000,000)+0.30(4,500,000)+0.20(16,000,000)=5,050,000. $\bar{\mathbf{X}}$: $\bar{X}_5 = 5,000/5$=1,000.

K: $K = 5,050,000/1,750,000$=2.89.

Z: There are five years of observations so $n = 5$. $Z = 5/(5 + 2.89)$=0.63.

$\boldsymbol{\hat{\mu}(\theta)}$: $\hat{\mu}(\theta) = 0.63(1,000) + (1 - 0.63)1,500 = \boxed{\mathbf{1,185.00}}$.

In real world applications of Bühlmann credibility the value of $K = EPV/VHM$ must be estimated. Sometimes a value for K is selected using judgment. A smaller K makes estimator $\hat{\mu}(\theta)$ more responsive to actual experience $\bar{X}$ whereas a larger K produces a more stable estimate by giving more weight to μ. Judgment may be used to balance responsiveness and stability. Section 12.5 in this chapter will discuss methods for determining K from data.

12.3 Bayesian Inference and Bühlmann Credibility

In this section, you learn how to:

- Calculate formulas for expected outcomes for beta-binomial and gamma-Poisson models using Bayes Theorem or Bühlmann credibility .
- Understand the connection between the Bayesian and Bühlmann estimates for conjugate families.

Chapter 9 presents Bayesian inference and modeling and it is assumed that the reader is familiar with that material, in particular, Section 9.3 which discusses conjugate families. This section will compare Bayesian inference with Bühlmann credibility and show connections between the two models.

First we will look at a Bayesian model. Suppose a risk has n observed losses $x_1, x_2, ..., x_n$. These losses will be represented by the vector $\mathbf{x} = (x_1, x_2, ..., x_n)$ which are realizations of the random variables $\mathbf{X} = (X_1, X_2, ..., X_n)$ which we will assume are iid.

A risk with risk parameter θ has expected loss $\mu(\theta) = \mathrm{E}_X(X|\theta)$. If the risk had losses $\mathbf{x}$ then $\mathrm{E}_\Theta(\mu(\theta)|\mathbf{x})$ is the conditional expectation of $\mu(\theta)$ given outcomes $\mathbf{x}$. The expected loss is updated to reflect the observations.

The expectation $\mathrm{E}_\Theta(\mu(\theta)|\mathbf{x})$ can be calculated using the conditional density function $f_{X|\Theta=\theta}(x|\theta)$ and the posterior distribution $f_{\Theta|\mathbf{X}=\mathbf{x}}(\theta|\mathbf{x})$

$$\begin{aligned} \mu(\theta) &= \mathrm{E}_\Theta(X|\theta) = \int x f_{X|\Theta=\theta}(x|\theta)dx \\ \mathrm{E}_\Theta(\mu(\theta)|\mathbf{x}) &= \int \mu(\theta) f_{\Theta|\mathbf{X}=\mathbf{x}}(\theta|\mathbf{x})d\theta. \end{aligned}$$

The integrations are over the support of the distributions. The posterior distribution comes from Bayes theorem

$$f_{\Theta|\mathbf{X}=\mathbf{x}}(\theta) = \frac{f_{\mathbf{X}|\Theta=\theta}(\mathbf{x})\, f_\Theta(\theta)}{f_{\mathbf{X}}(\mathbf{x})}.$$

The first function $f_{\mathbf{X}|\Theta=\theta}(\mathbf{x})$ in the numerator is the likelihood function and the second term $f_\Theta(\theta)$ is the prior distribution. The denominator $f_{\mathbf{X}}(\mathbf{x})$ is the joint density function for n losses $\mathbf{x} = (x_1, \ldots, x_n)$.

Now we turn to the Bühlmann model. The Bühlmann credibility estimate for

$\mathrm{E}_\Theta(\mu(\theta)|\mathbf{x})$ is $\hat{\mu}(\theta) = Z\bar{x} + (1 - Z)\mu$. This model requires credibility Z and collective mean μ which can be computed from the distributions used in the Bayesian model described above, if the distributions are known.

Example 12.3.1. Using n, conditional density function $f_{X|\Theta=\theta}(x|\theta)$, and prior distribution $f_\Theta(\theta)$, calculate credibility Z and collective mean μ for the Bühlmann credibility estimate $\hat{\mu}(\theta)$.

Example Solution. The collective mean is

$$\mu = \int\int x f_{X|\Theta=\theta}(x|\theta) f_\Theta(\theta) dx d\theta.$$

The Variance of the Hypothetical Means (VHM) is calculated as follows

$$\begin{aligned} \mu(\theta) &= \int x f_{X|\Theta=\theta}(x|\theta) dx, \\ \text{VHM} &= \int (\mu(\theta) - \mu)^2 f_\Theta(\theta) d\theta. \end{aligned}$$

The Expected Value of the Process Value (EPV) is calculated as follows

$$\begin{aligned} \text{Var}_X(x|\theta) &= \int (x - \mu(\theta))^2 f_{X|\Theta=\theta}(x|\theta) dx, \\ \text{EPV} &= \int \text{Var}_X(x|\theta) f_\Theta(\theta) d\theta. \end{aligned}$$

The credibility is $Z = n/(n + K)$ with $K = EPV/VHM$.

12.3.1 Beta-Binomial Model

Section 9.3.1 of the chapter Bayesian Inference and Modeling analyzes the beta-binomial model.

The number of successes x in m Bernoulli trials with unknown probability of success q is given by the binomial distribution

$$p_{X|Q=q}(x) = \binom{m}{x} q^x (1 - q)^{m-x}, \quad x \in \{0, 1, ..., m\}.$$

The probability of success q is modeled with the conjugate prior for the binomial distribution: the beta distribution with parameters a and b. The pdf of the beta distribution is

$$f_Q(q) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} q^{a-1} (1 - q)^{b-1}, \quad q \in [0, 1].$$

Given x successes in m Bernoulli trials the posterior distribution for q was shown in 9.3.1 to be

$$f_{Q|X=x}(q) = \frac{\Gamma(a+b+m)}{\Gamma(a+x)\Gamma(b+m-x)} q^{a+x-1}(1-q)^{b+m-x-1},$$

which is a beta distribution with parameters $a+x$ and $b+m-x$.

The mean for the beta distribution with parameters a and b is $\mathrm{E}(Q) = a/(a+b)$. Given x successes in m trials in the beta-binomial model the mean of the posterior distribution is

$$\mathrm{E}_Q(Q|x) = \frac{a+x}{a+b+m}.$$

The Bühlmann credibility estimate for $\mathrm{E}_Q(Q|x)$ exactly matches the Bayesian estimate as demonstrated in the following example.

Example 12.3.2. The probability that a coin toss will yield heads is q. The prior distribution for probability q is beta with parameters a and b. On m tosses of the coin there were exactly x heads. Use Bühlmann credibility to estimate the expected value of q.

Example Solution. Define random variables Y_j such that $Y_j = 1$ if the j^{th} coin toss is heads and $Y_j = 0$ if tails for $j = 1, \ldots, m$. Random variables Y_j are iid conditional on q with $\Pr(Y = 1|q) = q$ and $\Pr(Y = 0|q) = 1 - q$ The number of heads in m tosses can be represented by the random variable $X = Y_1 + \cdots + Y_m$.

We want to estimate $q = E(Y_j|X = x)$ using Buhlmann credibility: $\hat{q} = Z\bar{x} + (1-Z)\mu$. The overall mean is $\mu = \mathrm{E}(\mathrm{E}(Y_j|Q)) = \mathrm{E}(Q) = a/(a+b)$. The sample mean is $\bar{x} = x/m$.

The credibility is $Z = m/(m+K)$ and $K = EPV/VHM$. With $\mathrm{Var}(Y_j|q) = q(1-q)$ it follows that $EPV = \mathrm{E}(\mathrm{Var}[Y_j|Q]) = \mathrm{E}(Q(1-Q))$. Because $\mathrm{E}(Y_j|q) = q$ then $VHM = \mathrm{Var}((\mathrm{E}(Y_j|Q)) = \mathrm{Var}(Q)$. For the beta distribution

$$\mathrm{E}(Q) = \frac{a}{a+b}, \mathrm{E}(Q^2) = \frac{a(a+1)}{(a+b)(a+b+1)}, \text{ and } \mathrm{Var}(Q) = \frac{ab}{(a+b)^2(a+b+1)}.$$

Parameter $K = EPV/VHM$=$[\mathrm{E}(Q) - \mathrm{E}(Q^2)]/\mathrm{Var}(Q)$. With some algebra this reduces to $K = a+b$. The Buhlmann credibility-weighted estimate is

$$\hat{q} = \frac{m}{m+a+b}\left(\frac{x}{m}\right) + \left(1 - \frac{m}{m+a+b}\right)\frac{a}{a+b}$$
$$\hat{q} = \frac{a+x}{a+b+m}$$

which is the same as the Bayesian posterior mean.

12.3.2 Gamma-Poisson Model

The chapter Bayesian Inference and Modeling also analyzes the gamma-Poisson conjugate family. The results are summarized below.

Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a sample of iid Poisson random variables with

$$p_{X_i|\Lambda=\lambda}(x_i) = \frac{\lambda^{x_i}\, e^{-\lambda}}{x_i!}, \quad x_i \in \mathbb{R}_+.$$

Define the prior distribution for Λ to be gamma with parameters α and θ,

$$f_\Lambda(\lambda) = \frac{1}{\Gamma(\alpha)\theta^\alpha}\lambda^{\alpha-1}\, e^{-\frac{\lambda}{\theta}}, \quad \lambda \in \mathbb{R}_+.$$

Given a sample of n observations $\mathbf{x} = (x_1, x_2, ..., x_n)$, the posterior distribution of Λ is

$$f_{\Lambda|\mathbf{X}=\mathbf{x}}(\lambda) = \frac{1}{\Gamma(\alpha+x)\left(\frac{\theta}{n\theta+1}\right)^{\alpha+x}}\lambda^{\alpha+x-1}\, e^{-\frac{\lambda\,(n\theta+1)}{\theta}},$$

where $x = \sum_{i=1}^n x_i$, which is a gamma distribution with parameters $\alpha + x$ and $\frac{\theta}{n\theta+1}$.

We are going to make a minor change to the formulas above. Instead of a scale parameter θ, we will substitute a rate parameter $\beta = 1/\theta$. The posterior distribution becomes

$$f_{\Lambda|\mathbf{X}=\mathbf{x}}(\lambda) = \frac{(\beta+n)^{(\alpha+x)}}{\Gamma(\alpha+x)}\lambda^{\alpha+x-1}\, e^{-\lambda(\beta+n)}.$$

The posterior distribution is gamma and the expected value for Λ given observations $\mathbf{x}$ is easy to calculate:

$$\mathrm{E}_\Lambda(\Lambda|x_1, \ldots, x_n) = \frac{\alpha+x}{\beta+n}.$$

Prior to collecting a sample, $\mathrm{E}(\Lambda) = \alpha/\beta$ using parameters from the prior distribution.

The Bühlmann credibility model will give the same result as seen in the following example.

Example 12.3.3 The number of claims X each year for a risk has a Poisson distribution $p(x) = \lambda^x e^{-\lambda}/x!$. Each risk in a class has a constant risk parameter λ. Parameter λ is gamma distributed across the class with pdf $f(\lambda) = \beta^\alpha\lambda^{\alpha-1}e^{-\lambda\beta}/\Gamma(\alpha)$. A risk was selected at random from the population and observed for n years. The claims counts were $\mathbf{x} = (x_1, x_2, ..., x_n)$. Use Bühlmann credibility to calculate the expected value of λ for the selected risk.

Example Solution. The variance for a Poisson distribution with parameter λ is λ so $EPV = \mathrm{E}(\mathrm{Var}(X|\lambda)) = \mathrm{E}(\lambda) = \alpha/\beta$. The mean number of claims per year for the risk is λ so $VHM = \mathrm{Var}(\mathrm{E}(X|\lambda)) = \mathrm{Var}(\lambda) = \alpha/\beta^2$. The credibility parameter is $K = EPV/VHM = (\alpha/\beta)/(\alpha/\beta^2) = \beta$. The overall mean is $\mathrm{E}(\mathrm{E}(X|\lambda)) = \mathrm{E}(\lambda) = \alpha/\beta$. Letting $x = \sum_{j=1}^{n} x_j$, the sample mean is $\bar{x} = x/n$. The credibility-weighted estimate for the expected number of claims for the risk is

$$\hat{\mu} = \frac{n}{n+\beta}\bar{x} + \left(1 - \frac{n}{n+\beta}\right)\frac{\alpha}{\beta} = \frac{\alpha + x}{\beta + n}.$$

We will leave it to the reader to compare the Bayesian and Bühlmann models for the normal-normal conjugate family.

12.3.3 Exact Credibility

As demonstrated in the prior section, the Bühlmann credibility estimates for the beta-binomial and gamma-Poisson models exactly match the Bayesian analysis results. The term exact credibility is applied in these situations. Exact credibility may occur if the probability distribution for X_j is in the linear exponential family and the prior distribution is a conjugate prior. Besides these two models, examples of exact credibility also include Gamma-Exponential and Normal-Normal models.

If the conditional mean $\mathrm{E}_\Theta(\mu(\theta)|X_1, ..., X_n)$ is linear in the mean of the observations, then the Bühlmann credibility estimate will coincide with the Bayesian estimate. More information about exact credibility can be found in (Bühlmann and Gisler, 2005), (Klugman et al., 2012), and (Tse, 2009).

12.4 Bühlmann-Straub Credibility

In this section, you learn how to:

- Compute a credibility-weighted estimate for the expected loss for a risk or group of risks using the Bühlmann-Straub model.
- Determine the credibility Z assigned to observations.
- Calculate required values including the Expected Value of the Process Variance (EPV), Variance of the Hypothetical Means (VHM) and collective mean μ.

- Recognize situations when the Bühlmann-Straub model is appropriate.

With standard Bühlmann credibility as described in the prior section, losses $X_1, \ldots, X_n$ arising from a selected policyholder are assumed to be iid. If the subscripts indicate year 1, year 2 and so on up to year n, then the iid assumption means that the policyholder has the same exposure to loss every year. For commercial insurance this assumption is frequently violated.

Consider a commercial policyholder that uses a fleet of vehicles in its business. In year 1 there are m_1 vehicles in the fleet, m_2 vehicles in year 2, .., and m_n vehicles in year n. The exposure to loss from ownership and use of this fleet is not constant from year to year. The annual losses for the fleet are not iid.

Define Y_{jk} to be the loss for the k^{th} vehicle in the fleet for year j. Then, the total losses for the fleet in year j are $Y_{j1} + \cdots + Y_{jm_j}$ where we are adding up the losses for each of the m_j vehicles. In the Bühlmann-Straub model it is assumed that random variables Y_{jk} are iid across all vehicles and years for the policyholder. With this assumption the means $\mathrm{E}_Y(Y_{jk}|\theta) = \mu(\theta)$ and variances $\mathrm{Var}_Y(Y_{jk}|\theta) = \sigma^2(\theta)$ are the same for all vehicles and years. The quantity $\mu(\theta)$ is the expected loss and $\sigma^2(\theta)$ is the variance in the loss for one year for one vehicle for a policyholder with risk parameter θ.

If X_j is the average loss per unit of exposure in year j, $X_j = (Y_{j1} + \cdots + Y_{jm_j})/m_j$, then $\mathrm{E}_Y(X_j|\theta) = \mu(\theta)$ and $\mathrm{Var}_Y(X_j|\theta) = \sigma^2(\theta)/m_j$ for a policyholder with risk parameter θ. Note that we used the fact that the Y_{jk} are iid for a given policyholder. The average loss per vehicle for the entire n-year period is

$$\bar{X} = \frac{1}{m}\sum_{j=1}^{n} m_j X_j \quad , \quad m = \sum_{j=1}^{n} m_j .$$

It follows that $\mathrm{E}_Y(\bar{X}|\theta) = \mu(\theta)$ and $\mathrm{Var}_Y(\bar{X}|\theta) = \sigma^2(\theta)/m$ where $\mu(\theta)$ and $\sigma^2(\theta)$ are the mean and variance for a single vehicle for one year for the policyholder.

Example 12.4.1. Prove that $\mathrm{Var}_Y(\bar{X}|\theta) = \sigma^2(\theta)/m$ for a risk with risk parameter θ.

Example Solution.

$$\begin{aligned}
\mathrm{Var}_Y(\bar{X}|\theta) &= \mathrm{Var}_Y\left(\frac{1}{m}\sum_{j=1}^{n} m_j X_j|\theta\right) \\
&= \frac{1}{m^2}\sum_{j=1}^{n}\mathrm{Var}_Y(m_j X_j|\theta) = \frac{1}{m^2}\sum_{j=1}^{n} m_j^2 \mathrm{Var}_Y(X_j|\theta) \\
&= \frac{1}{m^2}\sum_{j=1}^{n} m_j^2(\sigma^2(\theta)/m_j) = \frac{\sigma^2(\theta)}{m^2}\sum_{j=1}^{n} m_j = \sigma^2(\theta)/m.
\end{aligned}$$

The Buhlmann-Straub credibility estimate is:

$$\hat{\mu}(\theta) = Z\bar{x} + (1-Z)\mu \tag{12.7}$$

with

θ = a risk parameter that identifies a policyholder's risk level

$\hat{\mu}(\theta)$ = estimated expected loss for one exposure for the policyholder with loss experience $\bar{X}$

$\bar{x}$ = $\frac{1}{m}\sum_{j=1}^{n} m_j x_j$ is the average loss per exposure for m exposures.
x_j is the average loss per exposure and m_j is the number of exposures in year j.

Z = credibility assigned to m exposures

μ = expected loss for one exposure for randomly chosen policyholder from population.

Note that $\hat{\mu}(\theta)$ is the estimator for the expected loss for one exposure. If the policyholder has m_j exposures then the expected loss is $m_j\hat{\mu}(\theta)$.

In Example 12.2.4, it was shown that $Z = \mathrm{Var}_\Theta(\mathrm{E}_X(\bar{X}|\theta))/\mathrm{Var}(\bar{X})$ where $\bar{X}$ is the average loss for n observations. In equation (12.7) the $\bar{X}$ is the average loss for m exposures and the same Z formula can be used:

$$Z = \frac{\mathrm{Var}_\Theta(\mathrm{E}_Y(\bar{X}|\theta))}{\mathrm{Var}(\bar{X})} = \frac{\mathrm{Var}_\Theta(\mathrm{E}_Y(\bar{X}|\theta))}{\mathrm{E}_\Theta(\mathrm{Var}_Y(\bar{X}|\theta)) + \mathrm{Var}_\Theta(\mathrm{E}_Y(\bar{X}|\theta))}.$$

(Note that X_j is a sum of Y_{jk}'s and $\bar{X}$ is an average of Y_{jk}'s.) The denominator was expanded using the law of total variance. As noted above $\mathrm{E}_Y(\bar{X}|\theta) = \mu(\theta)$

so $\mathrm{Var}_\Theta(\mathrm{E}_Y(\bar{X}|\theta)) = \mathrm{Var}_\Theta(\mu(\theta)) = VHM$. Because $\mathrm{Var}_Y(\bar{X}|\theta) = \sigma^2(\theta)/m$ it follows that $\mathrm{E}_\Theta(\mathrm{Var}_Y(\bar{X}|\theta)) = \mathrm{E}_\Theta(\sigma^2(\theta))/m = EPV/m$. Making these substitutions and using a little algebra gives

$$Z = \frac{m}{m+K} \quad , \quad K = \frac{EPV}{VHM}. \tag{12.8}$$

This is the same Z as for Bühlmann credibility except number of exposures m replaces number of years or observations n.

Example 12.4.2. A commercial automobile policyholder had the following exposures and claims over a three-year period:

Year	Number of Vehicles	Number of Claims
1	9	5
2	12	4
3	15	4

- The number of claims in a year for each vehicle in the policyholder's fleet is Poisson distributed with the same mean (parameter) λ.
- Parameter λ is distributed among the policyholders in the population with *pdf* $f(\lambda) = 6\lambda(1-\lambda)$ with $0 < \lambda < 1$.

The policyholder has 18 vehicles in its fleet in year 4. Use Bühlmann-Straub credibility to estimate the expected number of policyholder claims in year 4.

Example Solution. The expected number of claims for one vehicle for a randomly chosen policyholder is $\mu = \mathrm{E}(\lambda) = \int_0^1 \lambda[6\lambda(1-\lambda)]d\lambda = 1/2$. The average number of claims per vehicle for the policyholder is $\bar{X}$=13/36. The expected value of the process variance for a single vehicle is $EPV = \mathrm{E}(\lambda) = 1/2$.

The variance of the hypothetical means across policyholders is $VHM = \mathrm{Var}(\lambda) =$ $\mathrm{E}(\lambda^2)$-$(\mathrm{E}(\lambda))^2 = \int_0^1 \lambda^2[6\lambda(1-\lambda)]d\lambda - (1/2)^2 = (3/10) - (1/4) = (6/20) - (5/20) =$ 1/20. So, $K = EPV/VHM$=(1/2)/(1/20)=10. The number of exposures in the experience period is $m = 9+12+15 = 36$. The credibility is $Z = 36/(36+10) =$ 18/23.

The credibility-weighted estimate for the number of claims for one vehicle is $\hat{\mu}(\theta) = Z\bar{X} + (1-Z)\mu$=(18/23)(13/36)+(5/23)(1/2)=9/23. With 18 vehicles in the fleet in year 4 the expected number of claims is 18(9/23)=162/23=7.04 .

12.5 Estimating Credibility Parameters

In this section, you learn how to:

- Perform nonparametric estimation with the Bühlmann and Bühlmann-Straub credibility models.
- Identify situations when semiparametric estimation is appropriate.
- Use data to approximate the EPV and VHM.

The examples in this chapter have provided assumptions for calculating credibility parameters. In actual practice the actuary must use real world data and judgment to determine credibility parameters.

12.5.1 Nonparametric Estimation for Bühlmann and Bühlmann-Straub Models

Bayesian analysis as described previously requires assumptions about a prior distribution and likelihood. It is possible to produce estimates without these assumptions and these methods are often referred to as empirical Bayes methods. Bühlmann and Bühlmann-Straub credibility with parameters estimated from the data are included in the category of empirical Bayes methods.

Bühlmann Model. First we will address the simpler Bühlmann model. Assume that there are r risks in a population. For risk i with risk parameter θ_i the losses for n periods are $X_{i1},\ldots,X_{in}$. The losses for a given risk are *iid* across periods as assumed in the Bühlmann model. For risk i the sample mean is $\bar{X}_i=\sum_{j=1}^{n}X_{ij}/n$ and the unbiased sample process variance is $s_i^2=\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2/(n-1)$. An unbiased estimator for the EPV can be calculated by taking the average of s_i^2 for the r risks in the population:

$$\widehat{EPV}=\frac{1}{r}\sum_{i=1}^{r}s_i^2=\frac{1}{r(n-1)}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2. \tag{12.9}$$

The individual risk means $\bar{X}_i$ for $i=1,\ldots,r$ can be used to estimate the VHM. An unbiased estimator of $\mathrm{Var}(\bar{X}_i)$ is

$$\widehat{\mathrm{Var}}(\bar{X}_i)=\frac{1}{r-1}\sum_{i=1}^{r}(\bar{X}_i-\bar{X})^2 \text{ and } \bar{X}=\frac{1}{r}\sum_{i=1}^{r}\bar{X}_i,$$

but $\mathrm{Var}(\bar{X}_i)$ is not the VHM. The total variance formula or *unconditional variance* formula is

$$\mathrm{Var}(\bar{X}_i)=\mathrm{E}_X(\mathrm{Var}_\Theta(\bar{X}_i|\theta_i))+\mathrm{Var}_\Theta(\mathrm{E}_X(\bar{X}_i|\theta_i)).$$

The VHM is the second term on the right because $\mu(\theta_i)=\mathrm{E}_X(\bar{X}_i|\theta_i)$ is the hypothetical mean for risk i. So,

$$VHM=\mathrm{Var}(\bar{X}_i)-\mathrm{E}_\Theta(\mathrm{Var}_X(\bar{X}_i|\theta_i)).$$

As discussed previously in Section 12.2.2, $EPV/n = \mathrm{E}_\Theta(\mathrm{Var}_X[\bar{X}_i|\theta_i])$ and using the above estimators gives an estimator for the VHM:

$$\widehat{VHM} = \frac{1}{r-1}\sum_{i=1}^{r}(\bar{X}_i - \bar{X})^2 - \frac{\widehat{EPV}}{n}. \tag{12.10}$$

Although the expected loss for a risk with parameter θ_i is $\mu(\theta_i)$=$\mathrm{E}_X(\bar{X}_i|\theta_i)$, the variance of the sample mean $\bar{X}_i$ is greater than or equal to the variance of the hypothetical means: $\mathrm{Var}(\bar{X}_i) \geq \mathrm{Var}(\mu(\theta_i))$. The variance in the sample means $\mathrm{Var}(\bar{X}_i)$ includes both the variance in the hypothetical means plus a process variance term.

In some cases formula (12.10) can produce a negative value for $\widehat{VHM}$ because of the subtraction of $\widehat{EPV}/n$, but a variance cannot be negative. The process variance within risks is so large that it overwhelms the measurement of the variance in means between risks. In this case we cannot use this method to determine the values needed for Bühlmann credibility.

Example 12.5.1. Two policyholders had claims over a three-year period as shown in the table below. Estimate the expected number of claims for each policyholder using Bühlmann credibility and calculating necessary parameters from the data.

Year	Risk A	Risk B
1	0	2
2	1	1
3	0	2

Example Solution. $\bar{x}_A = \frac{1}{3}(0+1+0) = \frac{1}{3}$, $\bar{x}_B = \frac{1}{3}(2+1+2) = \frac{5}{3}$

$\bar{x} = \frac{1}{2}(\frac{1}{3} + \frac{5}{3}) = 1$

$s_A^2 = \frac{1}{3-1}\left[(0-\frac{1}{3})^2 + (1-\frac{1}{3})^2 + (0-\frac{1}{3})^2\right] = \frac{1}{3}$

$s_B^2 = \frac{1}{3-1}\left[(2-\frac{5}{3})^2 + (1-\frac{5}{3})^2 + (2-\frac{5}{3})^2\right] = \frac{1}{3}$

$\widehat{EPV} = \frac{1}{2}\left(\frac{1}{3} + \frac{1}{3}\right) = \frac{1}{3}$

$\widehat{VHM} = \frac{1}{2-1}\left[(\frac{1}{3}-1)^2 + (\frac{5}{3}-1)^2\right] - \frac{1/3}{3} = \frac{7}{9}$

$K = \frac{1/3}{7/9} = \frac{3}{7}$

$Z = \frac{3}{3+(3/7))} = \frac{7}{8}$

$\hat{\mu}_A = \frac{7}{8}\left(\frac{1}{3}\right) + (1-\frac{7}{8})1 = \frac{5}{12}$

$\hat{\mu}_B = \frac{7}{8}\left(\frac{5}{3}\right) + (1 - \frac{7}{8})1 = \frac{19}{12}$

Example 12.5.2. Two policyholders had claims over a three-year period as shown in the table below. Calculate the nonparametric estimate for the VHM.

Year	Risk A	Risk B
1	3	3
2	0	0
3	0	3

Example Solution. $\bar{x}_A = \frac{1}{3}(3+0+0) = 1$, $\bar{x}_B = \frac{1}{3}(3+0+3) = 2$

$\bar{x} = \frac{1}{2}(1+2) = \frac{3}{2}$

$s_A^2 = \frac{1}{3-1}\left[(3-1)^2 + (0-1)^2 + (0-1)^2\right] = 3$

$s_B^2 = \frac{1}{3-1}\left[(3-2)^2 + (0-2)^2 + (3-2)^2\right] = 3$

$\widehat{EPV} = \frac{1}{2}(3+3) = 3$

$\widehat{VHM} = \frac{1}{2-1}\left[(1-\frac{3}{2})^2 + (2-\frac{3}{2})^2\right] - \frac{3}{3} = -\frac{1}{2}$.

The process variance is so large that it is not possible to estimate the VHM.

Bühlmann-Straub Model Empirical formulas for EPV and VHM in the Bühlmann-Straub model are more complicated because a risk's number of exposures can change from one period to another. Also, the number of experience periods does not have to be constant across the population. First some definitions:

- X_{ij} is the losses per exposure for risk i in period j. Losses can refer to number of claims or amount of loss. There are r risks so $i = 1, \ldots, r$.
- n_i is the number of observation periods for risk i
- m_{ij} is the number of exposures for risk i in period j for $j = 1, \ldots, n_i$

Risk i with risk parameter θ_i has m_{ij} exposures in period j which means that the losses per exposure random variable can be written as $X_{ij} = (Y_{i1} + \cdots + Y_{im_{ij}})/m_{ij}$. Random variable Y_{ik} is the loss for one exposure. For risk i losses Y_{ik} are iid with mean $\mathrm{E}_Y(Y_{ik}|\theta_i) = \mu(\theta_i)$ and process variance $\mathrm{Var}_Y(Y_{ik}|\theta_i) = \sigma^2(\theta_i)$. It follows that $\mathrm{Var}_Y(X_{ij}|\theta_i) = \sigma^2(\theta_i)/m_{ij}$.

Two more important definitions are:

- $\bar{X}_i = \frac{1}{m_i}\sum_{j=1}^{n_i} m_{ij}X_{ij}$ with $m_i = \sum_{j=1}^{n_i} m_{ij}$. $\bar{X}_i$ is the average loss per exposure for risk i for all observation periods combined.
- $\bar{X} = \frac{1}{m}\sum_{i=1}^{r} m_i\bar{X}_i$ with $m = \sum_{i=1}^{r} m_i$. $\bar{X}$ is the average loss per exposure for all risks for all observation periods combined.

An unbiased estimator for the process variance $\sigma^2(\theta_i)$ of one exposure for risk i is

$$s_i{}^2 = \frac{\sum_{j=1}^{n_i} m_{ij}(X_{ij}-\bar{X}_i)^2}{n_i - 1}.$$

The weights m_{ij} are applied to the squared differences because the X_{ij} are the averages of m_{ij} exposures. The weighted average of the sample variances $s_i{}^2$ for each risk i in the population with weights proportional to the number of $(n_i - 1)$ observation periods will produce the expected value of the process variance (EPV) estimate

$$\widehat{EPV} = \frac{\sum_{i=1}^{r}(n_i-1)s_i{}^2}{\sum_{i=1}^{r}(n_i-1)} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i} m_{ij}(X_{ij}-\bar{X}_i)^2}{\sum_{i=1}^{r}(n_i-1)}.$$

The quantity $\widehat{EPV}$ is an unbiased estimator for the expected value of the process variance of one exposure for a risk chosen at random from the population.

To calculate an estimator for the variance in the hypothetical means (VHM) the squared differences of the individual risk sample means $\bar{X}_i$ and population mean $\bar{X}$ are used. An unbiased estimator for the VHM is

$$\widehat{VHM} = \frac{\sum_{i=1}^{r} m_i(\bar{X}_i-\bar{X})^2 - (r-1)\widehat{EPV}}{m - \frac{1}{m}\sum_{i=1}^{r} m_i^2}.$$

This complicated formula is necessary because of the varying number of exposures. Proofs that the EPV and VHM estimators shown above are unbiased can be found in several references mentioned at the end of this chapter including (Bühlmann and Gisler, 2005), (Klugman et al., 2012), and (Tse, 2009).

Example 12.5.3. Two policyholders had claims shown in the table below. Estimate the expected number of claims per vehicle for each policyholder using Bühlmann-Straub credibility and calculating parameters from the data.

Policyholder		Year 1	Year 2	Year 3	Year 4
A	Number of claims	0	2	2	3
A	Insured vehicles	1	2	2	2
B	Number of claims	0	0	1	2
B	Insured vehicles	0	2	3	4

Example Solution. $\bar{x}_A = \frac{0+2+2+3}{1+2+2+2} = 1$

$\bar{x}_B = \frac{0+1+2}{2+3+4} = \frac{1}{3}$

$\bar{x} = \frac{7(1)+9(1/3)}{7+9} = \frac{5}{8}$

$s_A^2 = \frac{1}{4-1}\left[1(0-1)^2 + 2(1-1)^2 + 2(1-1)^2 + 2(\frac{3}{2}-1)^2\right] = \frac{1}{2}$

$s_B^2 = \frac{1}{3-1}\left[2(0-\frac{1}{3})^2 + 3(\frac{1}{3}-\frac{1}{3})^2 + 4(\frac{1}{2}-\frac{1}{3})^2\right] = \frac{1}{6}$

$\widehat{EPV} = \left[3\left(\frac{1}{2}\right) + 2\left(\frac{1}{6}\right)\right]/(3+2) = \frac{11}{30} = 0.3667$

$\widehat{VHM} = \left[(7(1-\frac{5}{8})^2 + 9(\frac{1}{3}-\frac{5}{8})^2 - (2-1)\frac{11}{30}\right] / \left[16 - \left(\frac{1}{16}\right)(7^2+9^2)\right] = 0.1757$

$K = \frac{0.3667}{0.1757} = 2.0871$

$m_A = 7$, $m_B = 9$

$Z_A = \frac{7}{7+2.0871} = 0.7703$, $Z_B = \frac{9}{9+2.0871} = 0.8118$

$\hat{\mu}_A = 0.7703(1) + (1-0.7703)(5/8) = 0.9139$

$\hat{\mu}_B = 0.8118(1/3) + (1-0.8118)(5/8) = 0.3882.$

12.5.2 Semiparametric Estimation for Bühlmann and Bühlmann-Straub Models

In the prior section on nonparametric estimation, there were no assumptions about the distribution of the losses per exposure X_{ij}. Assuming that the X_{ij} have a particular distribution and using properties of the distribution along with the data to determine credibility parameters is referred to as semiparametric estimation.

An example of semiparametric estimation would be the assumption of a Poisson distribution when estimating claim frequencies. The Poisson distribution has the property that the mean and variance are identical and this property can simplify calculations. The following simple example comes from the prior section but now includes a Poisson assumption about claim frequencies.

Example 12.5.4. Two policyholders had claims over a three-year period as shown in the table below. Assume that the number of claims for each risk has a Poisson distribution. Estimate the expected number of claims for each policyholder using Bühlmann credibility and calculating necessary parameters from the data.

Year	Risk A	Risk B
1	0	2
2	1	1
3	0	2

Example Solution. $\bar{x}_A = \frac{1}{3}(0+1+0) = \frac{1}{3}$, $\bar{x}_B = \frac{1}{3}(2+1+2) = \frac{5}{3}$

$\bar{x} = \frac{1}{2}(\frac{1}{3} + \frac{5}{3}) = 1$

With Poisson assumption the estimated variance for risk A is $\hat{\sigma}_A^2 = \bar{x}_A = \frac{1}{3}$

Similarly, $\hat{\sigma}_B^2 = \bar{x}_B = \frac{5}{3}$

$\widehat{EPV} = \frac{1}{2}(\frac{1}{3}) + \frac{1}{2}(\frac{5}{3}) = 1$. This is also $\bar{x}$ because of Poisson assumption.

$\widehat{VHM} = \frac{1}{2-1}\left[(\frac{1}{3} - 1)^2 + (\frac{5}{3} - 1)^2\right] - \frac{1}{3} = \frac{5}{9}$

$K = \frac{1}{5/9} = \frac{9}{5}$

$Z_A = Z_B = \frac{3}{3+(9/5)} = \frac{5}{8}$

$\hat{\mu}_A = \frac{5}{8}\left(\frac{1}{3}\right) + (1 - \frac{5}{8})1 = \frac{7}{12}$

$\hat{\mu}_B = \frac{5}{8}\left(\frac{5}{3}\right) + (1 - \frac{5}{8})1 = \frac{17}{12}$.

Although we assumed that the number of claims for each risk was Poisson distributed in the prior example, we did not need this additional assumption because there was enough information to use nonparametric estimation. In fact, the Poisson assumption might not be appropriate because for risk B the sample mean is not equal to the sample variance: $\bar{x}_B = \frac{5}{3} \neq s_B^2 = \frac{1}{3}$.

The following example is commonly used to demonstrate a situation where semiparametric estimation is needed. There is insufficient information for nonparametric estimation but with the Poisson assumption, estimates can be calculated.

Example 12.5.5. A portfolio of 2,000 policyholders generated the following claims profile during a five-year period:

Number of Claims In 5 Years	Number of policies
0	923
1	682
2	249
3	70
4	51
5	25

In your model you assume that the number of claims for each policyholder has a Poisson distribution and that a policyholder's expected number of claims is constant through time. Use Bühlmann credibility to estimate the annual expected number of claims for policyholders with 3 claims during the five-year period.

Example Solution. Let θ_i be the risk parameter for the i^{th} risk in the portfolio with mean $\mu(\theta_i)$ and variance $\sigma^2(\theta_i)$. With the Poisson assumption $\mu(\theta_i) = \sigma^2(\theta_i)$. The expected value of the process variance is $EPV = \mathrm{E}(\sigma^2(\theta_i))$ where the expectation is taken across all risks in the population. Because of the Poisson assumption for all risks it follows that $EPV = \mathrm{E}(\sigma^2(\theta_i)) = \mathrm{E}(\mu(\theta_i))$. An estimate for the annual expected number of claims is $\hat{\mu}(\theta_i)$= (observed number of claims)/5. This can also serve as the estimate for the expected value of the process variance for a risk. Weighting the process variance estimates (or means) by the number of policies in each group gives the estimators

$$\widehat{EPV} = \bar{x} = \frac{923(0) + 682(1) + 249(2) + 70(3) + 51(4) + 25(5)}{(5)(2000)} = 0.1719.$$

Using the formula (
eqrefeq:VHM-estimate), the VHM estimator is

$$\begin{aligned}
\widehat{VHM} &= \frac{1}{2000-1}[923(0-0.1719)^2 + 682(0.20-0.1719)^2 + 249(0.40-0.1719)^2 \\
&\quad +70(0.60-0.1719)^2 + 51(0.80-0.1719)^2 + 25(1-0.1719)^2] - \frac{0.1719}{5} \\
&= 0.0111 \\
\hat{K} &= \widehat{EPV}/\widehat{VHM} = 0.1719/0.0111 = 15.49 \\
\hat{Z} &= \frac{5}{5+15.49} = 0.2440 \\
\hat{\mu}_{3\text{ claims}} &= 0.2440(3/5) + (1-0.2440)0.1719 = 0.2764.
\end{aligned}$$

12.6 Limited Fluctuation Credibility

In this section, you learn how to:

- Calculate full credibility standards for number of claims, average size of claims, and aggregate losses.
- Learn how the relationship between means and variances of underlying distributions affects full credibility standards.
- Determine credibility-weight Z using the square-root partial credibility formula.

Limited fluctuation credibility, also called "classical credibility" and "American credibility," was given this name because the method explicitly attempts to limit fluctuations in estimates for claim frequencies, severities, or losses. For example, suppose that you want to estimate the expected number of claims N for a group of risks in an insurance rating class. How many risks are needed in the class to ensure that a specified level of accuracy is attained in the estimate? First the question will be considered from the perspective of how many claims are needed.

12.6.1 Full Credibility for Claim Frequency

Let N be a random variable representing the number of claims for a group of risks, for example, risks within a particular rating classification. The observed number of claims will be used to estimate $\mu_N = \mathrm{E}[N]$, the expected number of claims. How big does μ_N need to be to get a good estimate? One way to quantify the accuracy of the estimate would be with a statement like: "The observed value of N should be within 5% of μ_N at least 90% of the time." Writing this as a mathematical expression would give $\Pr[0.95\mu_N \leq N \leq 1.05\mu_N] \geq 0.90$. Generalizing this statement by letting the range parameter k replace 5% and probability level p replace 0.90 gives the equation

$$\Pr[(1-k)\mu_N \leq N \leq (1+k)\mu_N] \geq p. \tag{12.11}$$

The expected number of claims required for the probability on the left-hand side of (12.11) to equal p is called the full credibility standard.

If the expected number of claims is greater than or equal to the full credibility standard then full credibility can be assigned to the data so $Z = 1$. Usually the expected value μ_N is not known so full credibility will be assigned to the data if the actual observed number of claims n is greater than or equal to the full credibility standard. The k and p values must be selected and the actuary may rely on experience, judgment, and other factors in making the choices.

Subtracting μ_N from each term in (12.11) and dividing by the standard deviation σ_N of N gives

$$\Pr\left[\frac{-k\mu_N}{\sigma_N} \leq \frac{N-\mu_N}{\sigma_N} \leq \frac{k\mu_N}{\sigma_N}\right] \geq p. \tag{12.12}$$

In limited fluctuation credibility the standard normal distribution is used to approximate the distribution of $(N-\mu_N)/\sigma_N$. If N is the sum of many claims from a large group of similar risks and the claims are independent, then the approximation may be reasonable.

Let y_p be the value such that

$$\Pr[-y_p \leq \frac{N-\mu_N}{\sigma_N} \leq y_p] = \Phi(y_p) - \Phi(-y_p) = p$$

where $\Phi()$ is the cumulative distribution function of the standard normal. Because $\Phi(-y_p) = 1 - \Phi(y_p)$, the equality can be rewritten as $2\Phi(y_p) - 1 = p$. Solving for y_p gives $y_p = \Phi^{-1}((p+1)/2)$ where $\Phi^{-1}()$ is the inverse of $\Phi()$.

Equation (12.12) will be satisfied if $k\mu_N/\sigma_N \geq y_p$ assuming the normal approximation. First we will consider this inequality for the case when N has a Poisson distribution: $\Pr[N=n] = \lambda^n \mathrm{e}^{-\lambda}/n!$. Because $\lambda = \mu_N = \sigma_N^2$ for the Poisson, taking square roots yields $\mu_N^{1/2} = \sigma_N$. So, $k\mu_N/\mu_N^{1/2} \geq y_p$ which is equivalent to $\mu_N \geq (y_p/k)^2$. Let's define λ_{kp} to be the value of μ_N for which equality holds. Then the full credibility standard for the Poisson distribution is

$$\lambda_{kp} = \left(\frac{y_p}{k}\right)^2 \text{ with } y_p = \Phi^{-1}((p+1)/2). \tag{12.13}$$

If the expected number of claims μ_N is greater than or equal to λ_{kp} then equation (12.11) is assumed to hold and full credibility can be assigned to the data. As noted previously, because μ_N is usually unknown, full credibility is given if the observed number of claims n satisfies $n \geq \lambda_{kp}$.

Example 12.6.1. The full credibility standard is set so that the observed number of claims is to be within 5% of the expected value with probability $p = 0.95$. If the number of claims has a Poisson distribution find the number of claims needed for full credibility.

Example Solution. Referring to a standard normal distribution table, $y_p = \Phi^{-1}((p+1)/2) = \Phi^{-1}((0.95+1)/2)$=$\Phi^{-1}(0.975) = 1.960$. Using this value and $k = .05$ then $\lambda_{kp} = (y_p/k)^2 = (1.960/0.05)^2 = 1,536.64$. After rounding up the full credibility standard is 1,537.

If claims are not Poisson distributed then equation (12.12) does not imply (12.13). Setting the upper bound of $(N - \mu_N)/\sigma_N$ in (12.12) equal to y_p gives $k\mu_N/\sigma_N = y_p$. Squaring both sides and moving everything to the right side except for one of the μ_N's gives $\mu_N = (y_p/k)^2(\sigma_N^2/\mu_N)$. This is the full credibility standard for frequency and will be denoted by n_f,

$$n_f = \left(\frac{y_p}{k}\right)^2 \left(\frac{\sigma_N^2}{\mu_N}\right) = \lambda_{kp}\left(\frac{\sigma_N^2}{\mu_N}\right). \tag{12.14}$$

This is the same equation as the Poisson full credibility standard except for the (σ_N^2/μ_N) multiplier. When the claims distribution is Poisson this extra term is one because the variance equals the mean.

Example 12.6.2. The full credibility standard is set so that the total number of claims is to be within 5% of the observed value with probability $p = 0.95$. The number of claims has a negative binomial distribution,

$$\Pr(N = x) = \binom{x+r-1}{x}\left(\frac{1}{1+\beta}\right)^r \left(\frac{\beta}{1+\beta}\right)^x,$$

with $\beta = 1$. Calculate the full credibility standard.

Example Solution. From the prior example, $\lambda_{kp} = 1,536.64$. The mean and variance for the negative binomial are $\mathrm{E}(N) = r\beta$ and $\mathrm{Var}(N) = r\beta(1+\beta)$ so $(\sigma_N^2/\mu_N) = (r\beta(1+\beta)/(r\beta)) = 1+\beta$ which equals 2 when $\beta = 1$. So, $n_f = \lambda_{kp}(\sigma_N^2/\mu_N) = 1,536.64(2) = 3,073.28$ and rounding up gives a full credibility standard of 3,074.

We see that the negative binomial distribution with $(\sigma_N^2/\mu_N) > 1$ requires more claims for full credibility than a Poisson distribution for the same k and p values. The next example shows that a binomial distribution which has $(\sigma_N^2/\mu_N) < 1$ will need fewer claims for full credibility.

Example 12.6.3. The full credibility standard is set so that the total number of claims is to be within 5% of the observed value with probability $p = 0.95$. The number of claims has a binomial distribution

$$\Pr(N = x) = \binom{m}{x} q^x (1-q)^{m-x}.$$

Calculate the full credibility standard for $q = 1/4$.

Example Solution. From the first example in this section $\lambda_{kp} = 1,536.64$. The mean and variance for a binomial are $\mathrm{E}(N) = mq$ and $\mathrm{Var}(N) = mq(1-q)$ so $(\sigma_N^2/\mu_N) = (mq(1-q)/(mq)) = 1 - q$ which equals 3/4 when $q = 1/4$. So, $n_f = \lambda_{kp}(\sigma_N^2/\mu_N) = 1,536.64(3/4) = 1,152.48$ and rounding up gives a full credibility standard of 1,153.

Rather than using expected number of claims to define the full credibility standard, the number of exposures can be used for the full credibility standard. An exposure is a measure of risk. For example, one car insured for a full year would be one car-year. Two cars each insured for exactly one-half year would also result in one car-year. Car-years attempt to quantify exposure to loss. Two car-years would be expected to generate twice as many claims as one car-year if the vehicles have the same risk of loss. To translate a full credibility standard denominated in terms of number of claims to a full credibility standard denominated in exposures one needs a reasonable estimate of the expected number of claims per exposure.

Example 12.6.4. The full credibility standard should be selected so that the observed number of claims will be within 5% of the expected value with probability $p = 0.95$. The number of claims has a Poisson distribution. If one exposure is expected to have about 0.20 claims per year, find the number of exposures needed for full credibility.

Example Solution. With $p = 0.95$ and $k = .05$, $\lambda_{kp} = (y_p/k)^2 = (1.960/0.05)^2 = 1,536.64$ claims are required for full credibility. The claims frequency rate is 0.20 claims per exposure. To convert the full credibility standard to a standard denominated in exposures the calculation is: (1,536.64 claims)/(0.20 claims/exposures) = 7,683.20 exposures. This can be rounded up to 7,684.

Frequency can be defined as the number of claims per exposure. Letting m denote the number of exposures. Then, if observed claim frequency N/m is used to estimate $\mathrm{E}(N/m)$:

$$\Pr[(1-k)\mathrm{E}(N/m) \leq N/m \leq (1+k)\mathrm{E}(N/m)] \geq p.$$

Because the number of exposures is not a random variable, $\mathrm{E}(N/m) = \mathrm{E}(N)/m = \mu_N/m$ and the prior equation becomes

$$\Pr\left[(1-k)\frac{\mu_N}{m} \leq \frac{N}{m} \leq (1+k)\frac{\mu_N}{m}\right] \geq p.$$

Multiplying through by m results in equation (12.11) at the beginning of the section. The full credibility standards that were developed for estimating expected number of claims also apply to frequency.

12.6.2 Full Credibility for Aggregate Losses and Pure Premium

Aggregate losses are the total of all loss amounts for a risk or group of risks. Letting S represent aggregate losses

$$S = X_1 + X_2 + \cdots + X_N.$$

The random variable N represents the number of losses and random variables $X_1, X_2, \ldots, X_N$ are the individual loss amounts. In this section it is assumed that N is independent of the loss amounts and that $X_1, X_2, \ldots, X_N$ are iid.

The mean and variance of S are

$$\mu_S = \mathrm{E}(S) = \mathrm{E}(N)\mathrm{E}(X) = \mu_N \mu_X$$

and

$$\sigma_S^2 = \mathrm{Var}(S) = \mathrm{E}(N)\mathrm{Var}(X) + [\mathrm{E}(X)]^2\mathrm{Var}(N) = \mu_N\sigma_X^2 + \mu_X^2\sigma_N^2,$$

where X is the amount of a single loss. See the discussion on *collective risk models* in Section 7.3 for more discussion of this framework.

Observed losses S will be used to estimate expected losses $\mu_S = \mathrm{E}(S)$. As with the frequency model in the previous section, the observed losses must be close to the expected losses as quantified in the equation

$$\Pr[(1-k)\mu_S \le S \le (1+k)\mu_S] \ge p.$$

After subtracting the mean and dividing by the standard deviation,

$$\Pr\left[\frac{-k\mu_S}{\sigma_S} \le (S-\mu_S)/\sigma_S \le \frac{k\mu_S}{\sigma_S}\right] \ge p.$$

As done in the previous section the distribution for $(S-\mu_S)/\sigma_S$ is assumed to be standard normal and $k\mu_S/\sigma_S = y_p = \Phi^{-1}((p+1)/2)$. This equation can be rewritten as $\mu_S^2 = (y_p/k)^2\sigma_S^2$. Using the prior formulas for μ_S and σ_S^2 gives $(\mu_N\mu_X)^2 = (y_p/k)^2(\mu_N\sigma_X^2 + \mu_X^2\sigma_N^2)$. Dividing both sides by $\mu_N\mu_X^2$ and reordering terms on the right side results in a full credibility standard n_S for aggregate losses

$$n_S = \left(\frac{y_p}{k}\right)^2\left[\left(\frac{\sigma_N^2}{\mu_N}\right) + \left(\frac{\sigma_X}{\mu_X}\right)^2\right] = \lambda_{kp}\left[\left(\frac{\sigma_N^2}{\mu_N}\right) + \left(\frac{\sigma_X}{\mu_X}\right)^2\right]. \qquad (12.15)$$

Example 12.6.5. The number of claims has a Poisson distribution. Individual loss amounts are independently and identically distributed with a Pareto distribution $F(x) = 1 - [\theta/(x+\theta)]^{\alpha}$. The number of claims and loss amounts are independent. If observed aggregate losses should be within 5% of the expected value with probability $p = 0.95$, how many losses are required for full credibility?

Example Solution. Because the number of claims is Poisson, $(\sigma_N^2/\mu_N) = 1$. The mean of the Pareto is $\mu_X = \theta/(\alpha - 1)$ and the variance is $\sigma_X^2 = \theta^2\alpha/[(\alpha-1)^2(\alpha-2)]$ so $(\sigma_X/\mu_X)^2 = \alpha/(\alpha-2)$. Combining the frequency and severity terms gives $[(\sigma_N^2/\mu_N) + (\sigma_X/\mu_X)^2] = 2(\alpha-1)/(\alpha-2)$. From a standard normal distribution table $y_p = \Phi^{-1}((0.95+1)/2) = 1.960$. The full credibility standard is $n_S = (1.96/0.05)^2[2(\alpha-1)/(\alpha-2)] = 3,073.28(\alpha-1)/(\alpha-2)$. Suppose $\alpha = 3$ then $n_S = 6,146.56$ for a full credibility standard of 6,147. Note that considerably more claims are needed for full credibility for aggregate losses than frequency alone.

When the number of claims is Poisson distributed then equation (12.15) can be simplified using $(\sigma_N^2/\mu_N) = 1$. It follows that

$$[(\sigma_N^2/\mu_N) + (\sigma_X/\mu_X)^2] = [1 + (\sigma_X/\mu_X)^2] = [(\mu_X^2 + \sigma_X^2)/\mu_X^2] = \mathrm{E}(X^2)/\mathrm{E}(X)^2$$

using the relationship $\mu_X^2 + \sigma_X^2 = \mathrm{E}(X^2)$. The full credibility standard is $n_S = \lambda_{kp}\,\mathrm{E}(X^2)/\mathrm{E}(X)^2$.

The pure premium PP is equal to aggregate losses S divided by exposures m: $PP = S/m$. The full credibility standard for pure premium will require

$$\Pr\left[(1-k)\mu_{PP} \le PP \le (1+k)\mu_{PP}\right] \ge p.$$

The number of exposures m is assumed fixed and not a random variable so $\mu_{PP} = \mathrm{E}(S/m) = \mathrm{E}(S)/m = \mu_S/m$.

$$\Pr\left[(1-k)\left(\frac{\mu_S}{m}\right) \le \left(\frac{S}{m}\right) \le (1+k)\left(\frac{\mu_S}{m}\right)\right] \ge p.$$

Multiplying through by m returns the bounds for losses

$$\Pr[(1-k)\mu_S \le S \le (1+k)\mu_S] \ge p.$$

This means that the full credibility standard n_{PP} for the pure premium is the same as that for aggregate losses

$$n_{PP} = n_S = \lambda_{kp}\left[\left(\frac{\sigma_N^2}{\mu_n}\right) + \left(\frac{\sigma_X}{\mu_X}\right)^2\right].$$

12.6.3 Full Credibility for Severity

Let X be a random variable representing the size of one claim. Claim severity is $\mu_X = \mathrm{E}(X)$. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of n claims that will be used to estimate claim severity μ_X. The claims are assumed to be *iid.* The average value of the sample is

$$\bar{X} = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right).$$

How big does n need to be to get a good estimate? Note that n is not a random variable whereas it is in the aggregate loss model.

In Section 12.6.1 the accuracy of an estimator for frequency was defined by requiring that the number of claims lie within a specified interval about the mean number of claims with a specified probability. For severity this requirement is

$$\Pr[(1-k)\mu_X \le \bar{X} \le (1+k)\mu_X] \ge p,$$

where k and p need to be specified. Following the steps in Section 12.6.1, the mean claim severity μ_X is subtracted from each term and the standard deviation of the claim severity estimator $\sigma_{\bar{X}}$ is divided into each term yielding

$$\Pr\left[\frac{-k\ \mu_X}{\sigma_{\bar{X}}} \le (\bar{X} - \mu_X)/\sigma_{\bar{X}} \le \frac{k\ \mu_X}{\sigma_{\bar{X}}}\right] \ge p.$$

As in prior sections, it is assumed that $(\bar{X} - \mu_X)/\sigma_{\bar{X}}$ is approximately normally distributed and the prior equation is satisfied if $k\mu_X/\sigma_{\bar{X}} \ge y_p$ with $y_p = \Phi^{-1}((p+1)/2)$. Because $\bar{X}$ is the average of individual claims $X_1, X_2, \ldots, X_n$, its standard deviation is equal to the standard deviation of an individual claim divided by $\sqrt{n}$: $\sigma_{\bar{X}} = \sigma_X/\sqrt{n}$. So, $k\mu_X/(\sigma_X/\sqrt{n}) \ge y_p$ and with a little algebra this can be rewritten as $n \ge (y_p/k)^2(\sigma_X/\mu_X)^2$. The full credibility standard for severity is

$$n_X = \left(\frac{y_p}{k}\right)^2 \left(\frac{\sigma_X}{\mu_X}\right)^2 = \lambda_{kp} \left(\frac{\sigma_X}{\mu_X}\right)^2. \qquad (12.16)$$

Note that the term σ_X/μ_X is the coefficient of variation for an individual claim. Even though λ_{kp} is the full credibility standard for frequency given a Poisson distribution, there is no assumption about the distribution for the number of claims.

Example 12.6.6. Individual loss amounts are independently and identically distributed with a Type II Pareto distribution $F(x) = 1 - [\theta/(x+\theta)]^\alpha$. How many claims are required for the average severity of observed claims to be within 5% of the expected severity with probability $p = 0.95$?

Example Solution. The mean of the Pareto is $\mu_X = \theta/(\alpha-1)$ and the variance is $\sigma_X^2 = \theta^2\alpha/[(\alpha-1)^2(\alpha-2)]$ so $(\sigma_X/\mu_X)^2 = \alpha/(\alpha-2)$. From a standard normal distribution table $y_p = \Phi^{-1}((0.95+1)/2) = 1.960$. The full credibility standard is $n_X = (1.96/0.05)^2[\alpha/(\alpha-2)] = 1,536.64\alpha/(\alpha-2)$. Suppose $\alpha = 3$ then $n_X = 4,609.92$ for a full credibility standard of 4,610.

12.6.4 Partial Credibility

In prior sections full credibility standards were calculated for estimating frequency (n_f), pure premium (n_{PP}), and severity (n_X) - in this section these full credibility standards will be denoted by n_0. In each case the full credibility standard was the expected number of claims required to achieve a defined level of accuracy when using empirical data to estimate an expected value. If the observed number of claims is greater than or equal to the full credibility standard then a full credibility weight $Z = 1$ is given to the data.

In limited fluctuation credibility, credibility weights Z assigned to data are

$$Z = \begin{cases} \sqrt{n/n_0} & \text{if } n < n_0 \\ 1 & \text{if } n \geq n_0, \end{cases}$$

where n_0 is the full credibility standard. The quantity n is the number of claims for the data that is used to estimate the expected frequency, severity, or pure premium.

Example 12.6.7. The number of claims has a Poisson distribution. Individual loss amounts are independently and identically distributed with a Type II Pareto distribution $F(x) = 1 - [\theta/(x+\theta)]^\alpha$. Assume that $\alpha = 3$. The number of claims and loss amounts are independent. The full credibility standard is that the observed pure premium should be within 5% of the expected value with probability $p = 0.95$. What credibility Z is assigned to a pure premium computed from 1,000 claims?

Example Solution. Because the number of claims is Poisson,

$$\frac{\mathrm{E}(X^2)}{[\mathrm{E}\ (X)]^2} = \frac{\sigma_N^2}{\mu_N} + \left(\frac{\sigma_X}{\mu_X}\right)^2.$$

The mean of the Pareto is $\mu_X = \theta/(\alpha-1)$ and the second moment is $\mathrm{E}(X^2) = 2\theta^2/[(\alpha-1)(\alpha-2)]$ so $\mathrm{E}(X^2)/[\mathrm{E}\ (X)]^2 = 2(\alpha-1)/(\alpha-2)$. From a standard normal distribution table, $y_p = \Phi^{-1}((0.95+1)/2) = 1.960$. The full credibility standard is

$$n_{PP} = (1.96/0.05)^2[2(\alpha-1)/(\alpha-2)] = 3,073.28(\alpha-1)/(\alpha-2)$$

and if $\alpha = 3$ then $n_0 = n_{PP} = 6,146.56$ or 6,147 if rounded up. The credibility assigned to 1,000 claims is $Z = (1,000/6,147)^{1/2} = 0.40$.

Limited fluctuation credibility uses the formula $Z = \sqrt{n/n_0}$ to limit the fluctuation in the credibility-weighted estimate to match the fluctuation allowed for data with expected claims at the full credibility standard. Variance or standard deviation is used as the measure of fluctuation. Next we show an example to explain why the square-root formula is used.

Suppose that average claim severity is being estimated from a sample of size n that is less than the full credibility standard $n_0 = n_X$. Applying credibility theory, the estimate $\hat{\mu}_X$ would be

$$\hat{\mu}_X = Z\bar{X} + (1-Z)M_X,$$

with $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$ and *iid* random variables X_i representing the sizes of individual claims. The complement of credibility is applied to M_X which could be last year's estimated average severity adjusted for inflation, the average severity for a much larger pool of risks, or some other relevant quantity selected by the actuary. It is assumed that the variance of M_X is zero or negligible. With this assumption

$$\text{Var}(\hat{\mu}_X) = \text{Var}(Z\bar{X}) = Z^2\text{Var}(\bar{X}) = \frac{n}{n_0}\text{Var}(\bar{X}).$$

Because $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$ it follows that $\text{Var}(\bar{X}) = \text{Var}(X_i)/n$ where random variable X_i is one claim. So,

$$\text{Var}(\hat{\mu}_X) = \frac{n}{n_0}\text{Var}(\bar{X}) = \frac{n}{n_0}\frac{\text{Var}(X_i)}{n} = \frac{\text{Var}(X_i)}{n_0}.$$

The last term is exactly the variance of a sample mean $\bar{X}$ when the sample size is equal to the full credibility standard $n_0 = n_X$.

12.6.5 Full Credibility Standard for Limited Fluctuation Credibility

Limited-fluctuation credibility requires a full credibility standard. The general formula for aggregate losses or pure premium, as obtained in formula (12.15), is

$$n_S = \left(\frac{y_p}{k}\right)^2 \left[\left(\frac{\sigma_N^2}{\mu_N}\right) + \left(\frac{\sigma_X}{\mu_X}\right)^2\right],$$

with N representing number of claims and X the size of claims. If one assumes $\sigma_X = 0$ then the full credibility standard for frequency results. If $\sigma_N = 0$ then the full credibility formula for severity follows. Probability p and k value are often selected using judgment and experience.

In practice it is often assumed that the number of claims is Poisson distributed so that $\sigma_N^2/\mu_N = 1$. In this case the formula can be simplified to

$$n_S = \left(\frac{y_p}{k}\right)^2 \left[\frac{\mathrm{E}(X^2)}{(\mathrm{E}(X))^2}\right].$$

An empirical mean and second moment for the sizes of individual claim losses can be computed from past data, if available.

12.7 Balancing Credibility Estimators

The credibility weighted model $\hat{\mu}(\theta_i) = Z_i\bar{X}_i + (1 - Z_i)\bar{X}$, where $\bar{X}_i$ is the loss per exposure for risk i and $\bar{X}$ is loss per exposure for the population, can be used to estimate the expected loss for risk i. The overall mean is $\bar{X} = \sum_{i=1}^{r}(m_i/m)\bar{X}_i$ where m_i and m are number of exposures for risk i and population, respectively.

For the credibility weighted estimators to be in balance we want

$$\bar{X} = \sum_{i=1}^{r}(m_i/m)\bar{X}_i = \sum_{i=1}^{r}(m_i/m)\hat{\mu}(\theta_i).$$

If this equation is satisfied then the estimated losses for each risk will add up to the population total, an important goal in ratemaking, but this may not happen if the complement of credibility is applied to $\bar{X}$.

To achieve balance, we will set $\hat{M}_X$ as the amount that is applied to the complement of credibility and thus analyze the following equation:

$$\sum_{i=1}^{r}(m_i/m)\bar{X}_i = \sum_{i=1}^{r}(m_i/m)\left\{Z_i\bar{X}_i + (1 - Z_i)\cdot\hat{M}_X\right\}.$$

A little algebra gives

$$\sum_{i=1}^{r}m_i\bar{X}_i = \sum_{i=1}^{r}m_iZ_i\bar{X}_i + \hat{M}_X\sum_{i=1}^{r}m_i(1 - Z_i),$$

and

$$\hat{M}_X = \frac{\sum_{i=1}^{r}m_i(1 - Z_i)\bar{X}_i}{\sum_{i=1}^{r}m_i(1 - Z_i)}.$$

Using this value for $\hat{M}_X$ will bring the credibility weighted estimators into balance.

If credibilities Z_i were computed using the Bühlmann-Straub model, then $Z_i = m_i/(m_i + K)$. The prior formula can be simplified using the following relationship

$$m_i(1 - Z_i) = m_i \left(1 - \frac{m_i}{m_i + K}\right) = m_i \left(\frac{(m_i + K) - m_i}{m_i + K}\right) = KZ_i.$$

Therefore, an amount when applied to the complement of credibility that will bring the credibility-weighed estimators into balance with the overall mean loss per exposure is

$$\hat{M}_X = \frac{\sum_{i=1}^{r} Z_i \bar{X}_i}{\sum_{i=1}^{r} Z_i}.$$

Example 12.7.1. An example from the nonparametric Bühlmann-Straub section had the following data for two risks. Find an amount for the complement of credibility $\hat{M}_X$ that will produce credibility-weighted estimates that are in balance.

Policyholder		Year 1	Year 2	Year 3	Year 4
A	Number of claims	0	2	2	3
A	Insured vehicles	1	2	2	2
B	Number of claims	0	0	1	2
B	Insured vehicles	0	2	3	4

Example Solution. The credibilities from the prior example are $Z_A = \frac{7}{7+2.0871} = 0.7703$ and $Z_B = \frac{9}{9+2.0871} = 0.8118$. The sample means are $\bar{x}_A = 1$ and $\bar{x}_B = 1/3$. The balanced complement of credibility is

$$\hat{M}_X = \frac{0.7703(1) + 0.8118(1/3)}{0.7703 + 0.8118} = 0.6579.$$

The updated credibility estimates are $\hat{M}_{X_A} = 0.7703(1) + (1 - 0.7703)(.6579) = 0.9214$ versus the previous 0.9139 and $\hat{M}_{X_B} = 0.8118(1/3) + (1 - 0.8118)(.6579) = 0.3944$ versus the previous 0.3882. Checking the balance on the new estimators: (7/16)(0.9214)+(9/16)(0.3944)=0.6250. This exactly matches $\bar{X} = 10/16 = 0.6250$.

12.8 Further Resources and Contributors

Contributor

- **Gary Dean**, Ball State University is the author of the initial version of this chapter. Email: cgdean@bsu.edu for chapter comments and suggested improvements.
- Chapter reviewers include: Liang (Jason) Hong, Ambrose Lo, Ranee Thiagarajah, Hongjuan Zhou.

13

Insurance Portfolio Management including Reinsurance

Chapter Preview. An insurance portfolio is simply a collection of insurance contracts. To help manage the uncertainty of the portfolio, this chapter

- quantifies unusually large obligations by examining the tail of the distribution,
- quantifies the overall riskiness by introducing summaries known as risk measures, and
- discusses options of spreading portfolio risk through reinsurance, the purchase of insurance protection by an insurer.

13.1 Introduction to Insurance Portfolios

In previous chapters, our analyses primarily focused on the contract level, which represents agreements between policyholders and insurers. Insurers maintain and manage portfolios, which are essentially collections of these individual contracts. Conceptually, one can liken an insurance company to nothing more than a collection, or portfolio, of insurance contracts. Similar to banking and investments, there are management decisions that are made exclusively at the portfolio level. Within this chapter, we address three crucial actuarial tasks: quantifying the impact of extreme events, determining overall portfolio risk, and managing insurance portfolios through reinsurance.

Insurance portfolios, representing the obligations of insurers, pique our interest primarily due to the probabilities associated with significant outcomes. These outcomes often translate to unusually large obligations. To illustrate, within property and casualty insurance, large obligations frequently stem from unforeseen consequences of **climate-related risks**. For instance, consider the freezing rain event of 1998 that swept through eastern Ontario and southwestern Quebec, lasting six days. This calamity resulted in double the typical precipitation for the region during an ice storm and gave rise to a catastrophe, triggering over 840,000 insurance claims. Astonishingly, this number exceeded the claims filed in the wake of Hurricane Andrew, one of North America's most extensive natural disasters. The catastrophe led to insurance settlements exceeding 1.44 billion Canadian dollars, marking the highest loss burden in Canada's history.

Such incidents are not isolated; similar catastrophic events, like Hurricane Harvey, Superstorm Sandy, the 2011 Japanese earthquake and tsunami, have also caused extreme insurance losses. In our exploration of extreme events in insurance, we introduce the concept of heavy-tailed distributions in Section 13.2.

Insurance companies engage in the buying and selling of risks as if they were commodities. As we explored in Chapter 10, greater uncertainty associated with risks typically translates into higher prices. In that chapter, pricing principles were introduced to quantify the magnitude of these risks. Furthermore, insurance portfolios represent the obligations of a company and, although they are not traded on a marketplace, they require careful management. One crucial aspect of this management is aligning the size of the obligations with an equivalent amount of assets. The subsequent Chapter 14 on loss reserves offers practical methods for achieving this alignment. Additionally, insurers need to assess the extent of their obligations for purposes such as capacity planning, policy formulation, and maintaining a balanced product portfolio that fosters revenue growth while managing volatility. To facilitate these tasks, Section 13.3 introduces risk measures that succinctly capture the uncertainty inherent in the distribution of an insurance portfolio.

Similar to individuals, insurance companies manage their risk portfolios by acquiring insurance, in this case, risk protection from reinsurers, which are insurance companies serving insurers. Just as individuals can structure the amount of risk they retain through mechanisms like deductibles and policy limits, insurers employ similar strategies to structure their risk portfolios. This practice of sharing insurance portfolio risk is detailed in Section 13.4, where we delve into the concept of reinsurance.

These three actuarial tasks, quantifying the impact of extreme events, determining overall portfolio risk, and managing insurance portfolios through reinsurance, are based on the distribution of insurance portfolios. In Chapter 7, we delved into modeling the distribution of insurance portfolios as the sum of individual contracts where we used S for aggregate losses. Now, this chapter is dedicated to the direct exploration of portfolio distributions and so we revert to the traditional X notation.

13.2 Tails of Distributions

In this section, you learn how to:

- Describe a heavy tail distribution intuitively.
- Classify the heaviness of a distribution's tails based on moments.
- Compare the tails of two distributions.

For extreme events such as those due to climate risks, a few major events hitting a portfolio and then converting into losses usually represent the greatest part of the indemnities paid by insurance companies. The aforementioned losses, also called 'extremes', are quantitatively modeled by the tails of the associated probability distributions. From the quantitative modeling standpoint, relying on probabilistic models having lengthy tails can be daunting. For instance, periods of financial stress may appear with a higher frequency than expected, and insurance losses may occur with worse severity. Therefore, the study of probabilistic behavior in the tail portion of actuarial models is important in quantitative risk management. For this reason, this section introduces a few mathematical notions that describe the tail weight of random variables. These notions will benefit us in the construction and selection of appropriate models with desired mathematical properties in the tail portion.

Formally, define X to be the random obligations that arise from a collection (portfolio) of insurance contracts. At the portfolio level, we are particularly interested in studying the right tail of the distribution of X which represents the occurrence of large losses. Informally, *a random variable is said to be heavy-tailed if high probabilities are assigned to large values.* This does not imply that the probability density/mass function increases as the value of X goes to infinity. Indeed, for a real-valued random variable, the pdf/pmf must diminish at infinity in order to guarantee the total probability to be equal to one. Instead, what we are concerned about is the *rate* of decay of the pdf/pmf. Unwelcome outcomes are more likely to occur for an insurance portfolio that is described by a loss random variable possessing a heavier (right) tail. Tail weight can be an absolute or a relative concept. Specifically, for the former, we may consider a random variable to be heavy-tailed if certain mathematical properties of the probability distribution are met. For the latter, we can say the tail of one distribution is heavier/lighter than the other if some tail measures are larger/smaller.

Several quantitative approaches have been proposed to classify and compare tail weights. For most of these approaches, the survival function serves as the building block. In what follows, we introduce two simple yet useful tail classification methods both of which are based on the behavior of the survival function of X.

13.2.1 Classification Based on Moments

One way of classifying the tail weight of a distribution is by determining whether or not a raw moment is finite. Because our major interest lies in the right tail of a distribution, we henceforth assume the obligation or loss random variable X to be non-negative. At the outset, the $k-$th raw moment of a continuous random variable X, introduced in Section 4.1, can be expressed as

$$\mu_k' = \int_0^\infty x^k f(x)\ dx = k \int_0^\infty x^{k-1} S(x)\ dx,$$

where $S(\cdot)$ denotes the survival function of X. This expression emphasizes that the finiteness of the raw moments depends on the asymptotic behavior of the survival function at infinity. Namely, the faster the survival function decays to zero, the higher is the order (k) is which the associated random variable may be finite. To capture this idea, we can formally define $k^* = \sup\{k > 0 : \mu_k' < \infty\}$, where *sup* represents the supremum operator. You may interpret k^* to be the largest value of k so that the moment is finite.

This definition leads us to a moment-based tail weight classification method which is defined as follows.

Definition 13.1. Consider a non-negative loss random variable X.

- If all the positive raw moments exist, namely the maximal order of finite moment $k^* = \infty$, then X is said to be **light tailed** based on the moment method.
- If $k^* < \infty$, then X is said to be heavy tailed based on the moment method.
- Moreover, for two positive loss random variables X_1 and X_2 with maximal orders of moment k_1^* and k_2^* respectively, we say X_1 has a **heavier (right) tail** than X_2 if $k_1^* \leq k_2^*$.

The first part of Definition 13.1 is an absolute concept of tail weight, while the second part is a relative concept of tail weight which compares the (right) tails between two distributions. Next, we present a few examples that illustrate the applications of the moment-based method for comparing tail weight.

Example 13.2.1. Light tail nature of the gamma distribution. Let $X \sim gamma(\alpha, \theta)$, with $\alpha > 0$ and $\theta > 0$. Show that $\mu_k' < \infty$ for all $k > 0$.

Example Solution. From the probability density functions expression in Section refS:ContinuousDistributions, the kth raw moment is

$$\begin{aligned} \mu'_k &= \int_0^\infty x^k \frac{x^{\alpha-1}e^{-x/\theta}}{\Gamma(\alpha)\theta^\alpha} dx \\ &= \int_0^\infty (y\theta)^k \frac{(y\theta)^{\alpha-1}e^{-y}}{\Gamma(\alpha)\theta^\alpha} \theta dy \\ &= \frac{\theta^k}{\Gamma(\alpha)}\Gamma(\alpha+k) < \infty. \end{aligned}$$

Because all the positive moments exist, we have $k^* = \infty$. Thus, in accordance with the moment-based classification method in Definition 13.1, the gamma distribution is light-tailed.

Example 13.2.2. Light tail nature of the Weibull distribution. Let $X \sim Weibull(\theta, \tau)$, with $\theta > 0$ and $\tau > 0$. Show that $\mu'_k < \infty$ for all $k > 0$.

Example Solution. From the probability density functions expression in Section refS:ContinuousDistributions, the kth raw moment is

$$\begin{aligned} \mu'_k &= \int_0^\infty x^k \frac{\tau x^{\tau-1}}{\theta^\tau} e^{-(x/\theta)^\tau} dx \\ &= \int_0^\infty \frac{y^{k/\tau}}{\theta^\tau} e^{-y/\theta^\tau} dy \\ &= \theta^k \Gamma(1 + k/\tau) < \infty. \end{aligned}$$

Again, due to the existence of all the positive moments, the Weibull distribution is light-tailed.

The gamma and Weibull distributions are used extensively in the actuarial practice. Applications of these two distributions are vast which include, but are not limited to, insurance claim severity modeling, solvency assessment, loss reserving, aggregate risk approximation, reliability engineering and failure analysis. We have thus far seen two examples of using the moment-based method to analyze light-tailed distributions. We document a heavy-tailed example in what follows.

Example 13.2.3. Heavy tail nature of the Pareto distribution. Let

$X \sim Pareto(\alpha, \theta)$, with $\alpha > 0$ and $\theta > 0$. Then, for $k > 0$,

$$\begin{aligned} \mu_k^{'} &= \int_0^\infty x^k \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}} dx \\ &= \alpha\theta^\alpha \int_\theta^\infty (y-\theta)^k y^{-(\alpha+1)} dy. \end{aligned}$$

From basic calculus, recall that

$$INT_k = \int_\theta^\infty y^{k-\alpha-1} dy = \begin{cases} < \infty, & \text{for } k < \alpha; \\ = \infty, & \text{for } k \geq \alpha. \end{cases}$$

Also note that:

$$\lim_{y\to\infty} \frac{(y-\theta)^k y^{-(\alpha+1)}}{y^{k-\alpha-1}} = \lim_{y\to\infty} (1-\theta/y)^k = 1.$$

Application of the limit comparison theorem for improper integrals yields μ_k' is finite if and only if INT_k is finite. Hence we can conclude that the raw moments of Pareto random variables exist only up to $k < \alpha$, i.e., $k^* = \alpha$, and thus the distribution is heavy-tailed.

What is more, the maximal order of finite moment depends only on the shape parameter α and it is an increasing function of α. In other words, based on the moment method, the tail weight of Pareto random variables is solely manipulated by α – the smaller the value of α, the heavier the tail weight becomes. Since $k^* < \infty$, the tail of Pareto distribution is heavier than those of the gamma and Weibull distributions.

Despite its simple implementation and intuitive interpretation, there are certain circumstances in which the application of the moment-based method is not suitable.

- 1. For more complicated probabilistic models, the k-th raw moment may not be simple to derive, and thus the identification of the maximal order of finite moment can be challenging.

- 2. The moment-based method does not well comply with main body of the well established heavy tail theory in the literature. Specifically, the existence of moment generating functions is arguably the most popular method for classifying heavy tail versus light tail within the community of academic actuaries. However, for some random variables such as the lognormal random variables, their moment generating functions do not exist even though all the positive moments are finite. In these cases, applications of the moment-based methods can lead to different tail weight assessment.

- 3. When we need to compare the tail weight between two light-tailed distributions (where both have all finite positive moments), the moment-based method is no longer informative (see, e.g., Examples 13.2.1 and 13.2.2).

13.2.2 Comparison Based on Limiting Tail Behavior

In order to resolve the aforementioned issues of the moment-based classification method, an alternative approach for comparing tail weight is to directly study the limiting behavior of the survival functions.

Definition 13.2. For two random variables X and Y, let

$$\gamma = \lim_{t\to\infty} \frac{S_X(t)}{S_Y(t)}.$$

We say that

- X has a **heavier right tail** than Y if $\gamma = \infty$,
- X and Y are **proportionally equivalent in the right tail** if $\gamma = c \in (0, \infty)$, and
- X has a **lighter right tail** than Y if $\gamma = 0$.

Example 13.2.4. Comparison of Pareto to Weibull distributions. Let $X \sim Pareto(\alpha, \theta)$ and $Y \sim Weibull(\tau, \theta)$, for $\alpha > 0$, $\tau > 0$, and $\theta > 0$. Show that the Pareto has a heavier right tail than the Weibull.

Example Solution.

$$\begin{aligned}
\lim_{t\to\infty} \frac{S_X(t)}{S_Y(t)} &= \lim_{t\to\infty} \frac{(1+t/\theta)^{-\alpha}}{\exp\{-(t/\theta)^{\tau}\}} \\
&= \lim_{t\to\infty} \frac{\exp\{t/\theta^{\tau}\}}{(1+t^{1/\tau}/\theta)^{\alpha}} \\
&= \lim_{t\to\infty} \frac{\sum_{i=0}^{\infty} \left(\frac{t}{\theta^{\tau}}\right)^{i} /i!}{(1+t^{1/\tau}/\theta)^{\alpha}} \\
&= \lim_{t\to\infty} \sum_{i=0}^{\infty} \left(t^{-i/\alpha} + \frac{t^{(1/\tau - i/\alpha)}}{\theta} \right)^{-\alpha} /\theta^{\tau i} i! \\
&= \infty.
\end{aligned}$$

Therefore, the Pareto distribution has a heavier tail than the Weibull distribution.

One may also realize that exponentials go to infinity faster than polynomials, thus the aforementioned limit must be infinite.

For some distributions of which the survival functions do not admit explicit expressions, we may find the following alternative formula useful:

$$\begin{aligned}\lim_{t\to\infty}\frac{S_X(t)}{S_Y(t)} &= \lim_{t\to\infty}\frac{S_X^{'}(t)}{S_Y^{'}(t)}\\ &= \lim_{t\to\infty}\frac{-f_X(t)}{-f_Y(t)}\\ &= \lim_{t\to\infty}\frac{f_X(t)}{f_Y(t)},\end{aligned}$$

given that the density functions exist. This is an application of L'Hôpital's Rule from calculus.

Example 13.2.5. Comparison of Pareto to gamma distributions. Let $X \sim Pareto(\alpha,\theta)$ and $Y \sim gamma(\alpha,\theta)$, for $\alpha > 0$ and $\theta > 0$. Show that the Pareto has a heavier right tail than the gamma.

Example Solution.

$$\begin{aligned}\lim_{t\to\infty}\frac{f_X(t)}{f_Y(t)} &= \lim_{t\to\infty}\frac{\alpha\theta^{\alpha}(t+\theta)^{-\alpha-1}}{t^{\tau-1}e^{-t/\lambda}\lambda^{-\tau}\Gamma(\tau)^{-1}}\\ &\propto \lim_{t\to\infty}\frac{e^{t/\lambda}}{(t+\theta)^{\alpha+1}t^{\tau-1}}\\ &= \infty,\end{aligned}$$

as exponentials go to infinity faster than polynomials.

13.3 Risk Measures

In this section, you learn how to:

- Define the value-at-risk and calculate this quantity for a given distribution.
- Define the expected shortfall and calculate this quantity for a given distribution.

- Define the idea of *coherence* and determine whether or not a risk measure is coherent.

In the previous section, we studied two methods for classifying the weight of distribution tails. We may claim that the risk associated with one distribution is more dangerous (asymptotically) than the other if the tail is heavier. However, knowing that one risk is more dangerous than the other may not provide sufficient information for risk management purposes and, in addition, one is also interested in quantifying how much more. In fact, the magnitude of risk associated with a given loss distribution is an essential input for many insurance applications, such as actuarial pricing, reserving, hedging, insurance regulatory oversight, and so forth.

The literature on risk measures has been growing rapidly in popularity and importance. In the next two subsections, we introduce two indices which have earned interest among theoreticians, practitioners, and regulators. They are namely the *Value-at-Risk* (VaR) and the *Expected Shortfall* (ES) measures. The rationale underpinning these two risk measures is similar to that for the tail classification methods – we hope to capture the uncertainty of extreme losses.

13.3.1 Value-at-Risk

In Section 4.4.1, we defined the quantile of a distribution. We now look to a special case of this and offer the formal definition of the value-at-risk, or *VaR*.

Definition 13.3. Consider an insurance loss random variable X. The value-at-risk measure of X with confidence level $q \in (0,1)$ is formulated as

$$VaR_q[X] = \inf\{x : F_X(x) \geq q\}. \tag{13.1}$$

Here, inf is the infimum operator so that the VaR measure outputs the smallest value of x such that the associated cdf exceeds or equates to q. This is simply the quantile that was introduced in Section 4.1.2.

Here is how we should interpret VaR in the context of actuarial applications. The VaR is a measure of the 'maximal' probable loss for an insurance product/portfolio or a risky investment occurring $q \times 100\%$ of times, over a specific time horizon (typically, one year). For instance, if we let X be the annual loss random variable of an insurance product, then $VaR_{0.95}[X] = 100$ million means that there is no more than a 5% chance that the loss will exceed 100 million over a given year. Owing to this meaningful interpretation, VaR has become the industry standard for measuring financial and insurance risks since the 1990's. Financial conglomerates, regulators, and academics often utilize

VaR to measure risk capital, ensure the compliance with regulatory rules, and disclose the financial positions.

Next, we present a few examples concerning the computation of VaR.

Example 13.3.1. VaR for the exponential distribution. Consider an insurance loss random variable X with an exponential distribution having parameter θ for $\theta > 0$, then the *cdf* of X is given by

$$F_X(x) = 1 - e^{-x/\theta}, \text{ for } x > 0.$$

Give a closed-form expression for the VaR.

Example Solution. Because exponential distribution is a continuous distribution, the smallest value such that the *cdf* first exceeds or equates to $q \in (0, 1)$ must be at the point x_q satisfying

$$q = F_X(x_q) = 1 - \exp\{-x_q/\theta\}.$$

Thus

$$VaR_q[X] = F_X^{-1}(q) = -\theta[\log(1-q)].$$

The result reported in Example 13.3.1 can be generalized to any continuous random variables having a strictly increasing *cdf*. Specifically, the VaR of any continuous random variables is simply the inverse of the corresponding *cdf*. Let us consider another example of continuous random variable which has the support from negative infinity to positive infinity.

Example 13.3.2. VaR for the normal distribution. Consider an insurance loss random variable $X \sim Normal(\mu, \sigma^2)$ with $\sigma > 0$. In this case, one may interpret the negative values of X as profit or revenue. Give a closed-form expression for the VaR.

Example Solution. Because normal distribution is a continuous distribution, the VaR of X must satisfy

$$\begin{aligned} q &= F_X(VaR_q[X]) \\ &= \Pr\left[(X-\mu)/\sigma \leq (VaR_q[X]-\mu)/\sigma\right] \\ &= \Phi((VaR_q[X]-\mu)/\sigma). \end{aligned}$$

Therefore, we have

$$VaR_q[X] = \Phi^{-1}(q)\ \sigma + \mu.$$

In many insurance applications, we have to deal with transformations of random variables. For instance, in Example 13.3.2, the loss random variable $X \sim Normal(\mu, \sigma^2)$ can be viewed as a linear transformation of a standard normal random variable $Z \sim Normal(0, 1)$, namely $X = Z\sigma + \mu$. By setting $\mu = 0$ and $\sigma = 1$, it is straightforward for us to check $VaR_q[Z] = \Phi^{-1}(q)$. A useful finding revealed from Example 13.3.2 is that the VaR of a linear transformation of the normal random variables is equivalent to the linear transformation of the VaR of the original random variables. This finding can be further generalized to any random variables as long as the transformations are strictly increasing.

Example 13.3.3. VaR **for transformed variables.** Consider an insurance loss random variable Y with a lognormal distribution with parameters $\mu \in \mathbf{R}$ and $\sigma^2 > 0$. Give an expression of the VaR of Y in terms of the standard normal inverse *cdf*.

Example Solution. Note that $\log Y \sim Normal(\mu, \sigma^2)$, or equivalently let $X \sim Normal(\mu, \sigma^2)$, then $Y \overset{d}{=} e^X$ which is strictly increasing transformation. Here, the notation '$\overset{d}{=}$' means equality in distribution. The VaR of Y is thus given by the exponential transformation of the VaR of X. Precisely, for $q \in (0, 1)$,

$$VaR_q[Y] = e^{VaR_q[X]} = \exp\{\Phi^{-1}(q)\ \sigma + \mu\}.$$

We have thus far seen a number of examples about the VaR for continuous random variables, let us consider an example concerning the VaR for a discrete random variable.

Example 13.3.4. VaR **for a discrete random variable.** Consider an insurance loss random variable with the following probability distribution:

$$\Pr[X = x] = \begin{cases} 0.75, & \text{for } x = 1 \\ 0.20, & \text{for } x = 3 \\ 0.05, & \text{for } x = 4. \end{cases}$$

Determine the VaR at $q = 0.6, 0.9, 0.95, 0.95001$.

Example Solution. The corresponding *cdf* of X is

$$F_X(x) = \begin{cases} 0, & x < 1 \\ 0.75, & 1 \le x < 3 \\ 0.95, & 3 \le x < 4 \\ 1, & 4 \le x. \end{cases}$$

By the definition of VaR, we thus have

$VaR_{0.6}[X] = 1$, $VaR_{0.9}[X] = 3$, $VaR_{0.95}[X] = 3$, and $VaR_{0.950001}[X] = 4$.

Let us now conclude the current subsection by an open discussion of the VaR measure. Some advantages of utilizing VaR include

- possessing a practically meaningful interpretation, and
- relatively simple to compute for many distributions with closed-form distribution functions.

On the other hand, the limitations of VaR can be particularly pronounced for some risk management practices. We report some of them herein:

- the selection of the confidence level $q \in (0, 1)$ is highly subjective, while the VaR can be very sensitive to the choice of q (e.g., in Example 13.3.4, $VaR_{0.95}[X] = 3$ and $VaR_{0.950001}[X] = 4$);
- the scenarios/loss information that are above the $(1 - q) \times 100\%$ worst event, are completely neglected;
- as will be seen in Section 13.3.3, the VaR is not a coherent risk measure.

The VaR represents the $(1 - q) \times 100\%$ chance maximal loss. One major drawback of the VaR measure is that it does not reflect the extremal losses occurring beyond the $(1 - q) \times 100\%$ chance worst scenario. For illustrative purposes, let us consider the following slightly unrealistic yet inspiring example.

Example 13.3.5. Consider two loss random variable's $X \sim Uniform[0, 100]$, and Y with an exponential distribution having parameter $\theta = 31.71$. We use VaR at 95% confidence level to measure the riskiness of X and Y. Simple calculation yields (see, also, Example 13.3.1),

$$VaR_{0.95}[X] = VaR_{0.95}[Y] = 95,$$

and thus these two loss distributions have the same level of risk according to $VaR_{0.95}$. However, Y is riskier than X if extremal losses are of major concern since X is bounded above while Y is unbounded. Simply quantifying risk by using VaR at a specific confidence level could be misleading and may not reflect the true nature of risk.

13.3.2 Expected Shortfall

Another commonly used risk measure is the **expected shortfall**, ES. Mathematically, we can express this as

$$ES_q(X) = \frac{1}{1-q} \int_q^1 VaR_a(X) da. \tag{13.2}$$

That is, the ES is the average of $VaR_\alpha[X]$ with varying degree of confidence level over $\alpha \in [q, 1]$. Thus, it is also known as the *average value at risk.* In this respect, one can see that for any given $q \in (0, 1)$

$$ES_q[X] \geq VaR_q[X].$$

The ES effectively resolves most of the limitations of VaR outlined in the previous subsection. First, due to the averaging effect, the ES may be less sensitive to the change of confidence level compared with VaR. Second, all the extremal losses that are above the $(1 - q) \times 100\%$ worst probable event are taken in account.

There are a few other forms of the ES that will be useful to us. For notional convenience, we write $\pi_q = VaR_q[X]$ and have

$$ES_q(X) = \left\{ \begin{array}{cl} \frac{1}{1-q} \int_q^1 VaR_a(X) da & \text{Expected Shortfall} \\ \pi_q + \frac{1}{1-q} \left\{ \mathrm{E}[X] - \mathrm{E}[X \wedge \pi_q] \right\} & \text{Tail VaR} \\ \mathrm{E}(X|X > \pi_q) & \text{Conditional VaR.} \end{array} \right. \tag{13.3}$$

The different expressions in Display (13.3) hold under some additional (mild) assumptions on the continuity of the distribution function at the point π_q. As we are interested in applications to portfolios, we employ such assumptions in this chapter which allows us to describe alternative ways of thinking about these measures. For example, from the third expression, we see that ES can also be interpreted to be the expected amount given that the loss exceeds the VaR_q.

Naturally, analysts may work with distributions where the assumptions of continuity do not hold, such as discrete distributions (see the examples Chapter 3). For these distributions, Display (13.3) provides a definition for some alternative risk measures, the *Tail value-at-risk* and the *Conditional value-at-risk.* You can learn more about these alternative risk measures in the references given in Section 13.6.

To see the connections between the second and third equalities, use a variable substitution, $z = VaR_a(X) = F^{-1}(a)$ so that $F(z) = a$ and $f(z)dz = da$. With this, we have

$$\begin{aligned} \int_a^b VaR_a(X)\ da &= \int_a^b F_a^{-1}\ da = \int_{F_a^{-1}}^{F_b^{-1}} zf(z)dz \\ &= -z[1 - F(z)]\Big|_{F_a^{-1}}^{F_b^{-1}} + \int_{F_a^{-1}}^{F_b^{-1}} [1 - F(z)]dz \\ &= F_a^{-1}(1 - a) - F_b^{-1}(1 - b) + \\ &\qquad [\mathrm{E}(X \wedge F_b^{-1}) - \mathrm{E}(X \wedge F_a^{-1})]. \end{aligned}$$

Thus,

$$\begin{aligned} \tfrac{1}{1-q}\int_q^1 VaR_a(X)da &= \tfrac{1}{1-q}\left\{\pi_q(1-q) + [\mathrm{E}(X) - \mathrm{E}(X \wedge \pi_q)]\right\} \\ &= \pi_q + \tfrac{1}{1-q}[\mathrm{E}(X) - \mathrm{E}(X \wedge \pi_q)], \end{aligned}$$

as claimed.

Using the third expression in Display (13.3), the computation of ES consists of two major steps - the VaR and the average of losses that are above the VaR. From this and a change of variables, the ES can be computed via

$$ES_q[X] = \frac{1}{(1-q)}\int_{\pi_q}^{\infty} x f_X(x)dx. \tag{13.4}$$

Example 13.3.6. ES for a normal distribution. Consider an insurance loss random variable $X \sim Normal(\mu, \sigma^2)$ with $\mu \in \mathbf{R}$ and $\sigma > 0$. Give an expression for ES.

Example Solution. Let Z be the standard normal random variable. For $q \in (0,1)$, the ES of X can be computed via

$$\begin{aligned} ES_q[X] &= \mathrm{E}[X|X > VaR_q[X]] \\ &= \mathrm{E}[\sigma Z + \mu | \sigma Z + \mu > VaR_q[X]] \\ &= \sigma \mathrm{E}[Z|Z > (VaR_q[X] - \mu)/\sigma] + \mu \\ &\overset{(1)}{=} \sigma \mathrm{E}[Z|Z > VaR_q[Z]] + \mu, \end{aligned}$$

where '$\overset{(1)}{=}$' holds because of the results reported in Example 13.3.2. Next, we turn to study $ES_q[Z] = \mathrm{E}[Z|Z > VaR_q[Z]]$. Let $\omega(q) = [\Phi^{-1}(q)]^2/2$, we have

$$\begin{aligned} (1-q)\ ES_q[Z] &= \int_{\Phi^{-1}(q)}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &= \int_{\omega(q)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x} dx \\ &= \frac{1}{\sqrt{2\pi}} e^{-\omega(q)} \\ &= \phi[\Phi^{-1}(q)]. \end{aligned}$$

Thus,

$$ES_q[X] = \sigma \frac{\phi[\Phi^{-1}(q)]}{1-q} + \mu.$$

We mentioned earlier in the previous subsection that the VaR of a strictly increasing function of random variable is equal to the function of VaR of the original random variable. Motivated by the results in Example 13.3.6, one can show that the ES of a strictly increasing linear transformation of random variable is equal to the function of VaR of the original random variable. This is due to the linearity property of expectations. However, the aforementioned finding cannot be extended to non-linear functions. The following example of lognormal random variable serves as a counter example.

Example 13.3.7. ***ES*** **of a lognormal distribution.** Consider an insurance loss random variable X with a lognormal distribution having parameters $\mu \in \mathbf{R}$ and $\sigma > 0$. Show that

$$ES_q[X] = \frac{e^{\mu+\sigma^2/2}}{(1-q)}\Phi(\Phi^{-1}(q) - \sigma).$$

Example Solution. Recall that the *pdf* of lognormal distribution is formulated as

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi x}} \exp\{-(\log x - \mu)^2/2\sigma^2\}, \text{ for } x > 0.$$

Fix $q \in (0,1)$, then the expected shortfall can be computed via

$$\begin{aligned} ES_q[X] &= \frac{1}{(1-q)}\int_{\pi_q}^{\infty} x f_X(x)dx \\ &= \frac{1}{(1-q)}\int_{\pi_q}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\} dx \\ &\overset{(1)}{=} \frac{1}{(1-q)}\int_{\omega(q)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}w^2+\sigma w+\mu} dw \\ &= \frac{e^{\mu+\sigma^2/2}}{(1-q)}\int_{\omega(q)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(w-\sigma)^2} dw \\ &= \frac{e^{\mu+\sigma^2/2}}{(1-q)}\Phi(\omega(q) - \sigma), \qquad (13.5) \end{aligned}$$

where $\overset{(1)}{=}$ holds by applying change of variable $w = (\log x - \mu)/\sigma$, and $\omega(q) = (\log \pi_q - \mu)/\sigma$. Evoking the formula of VaR for lognormal random variable reported in Example 13.3.2, we can simplify the expression
eqrefeq:cte-normal into

$$ES_q[X] = \frac{e^{\mu+\sigma^2/2}}{(1-q)}\Phi(\Phi^{-1}(q) - \sigma).$$

Clearly, the ES of lognormal random variable is not the exponential of the ES of normal random variable.

For distributions of which the survival distribution functions are more tractable to work with, we may apply the integration by parts technique (assuming the mean is finite) to rewrite equation (13.4) as

$$\begin{aligned} ES_q[X] &= \left[-xS_X(x)\Big|_{\pi_q}^{\infty} + \int_{\pi_q}^{\infty} S_X(x)dx \right] \frac{1}{(1-q)} \\ &= \pi_q + \frac{1}{(1-q)} \int_{\pi_q}^{\infty} S_X(x)dx. \end{aligned}$$

Example 13.3.8. ***ES*** **of an exponential distribution.** Consider an insurance loss random variable X with an exponential distribution having parameter θ for $\theta > 0$. Give an expression for the ES.

Example Solution. We have seen from the previous subsection that

$$\pi_q = -\theta[\log(1-q)].$$

Let us now consider the ES:

$$\begin{aligned} ES_q[X] &= \pi_q + \int_{\pi_q}^{\infty} e^{-x/\theta} dx/(1-q) \\ &= \pi_q + \theta e^{-\pi_q/\theta}/(1-q) \\ &= \pi_q + \theta. \end{aligned}$$

The second expression in Display (13.3) shows how to express the ES in terms of limited expected values. For many commonly used parametric distributions, the formulas for calculating $\mathrm{E}[X]$ and $\mathrm{E}[X \wedge \pi_q]$ can be found in a table of distributions.

Example 13.3.9. ***ES*** **of a Pareto distribution.** Consider a loss random variable $X \sim Pareto(\theta, \alpha)$ with $\theta > 0$ and $\alpha > 0$. The *cdf* of X is given by

$$F_X(x) = 1 - \left(\frac{\theta}{\theta + x} \right)^{\alpha}, \text{ for } x > 0.$$

Fix $q \in (0,1)$ and set $F_X(\pi_q) = q$, we readily obtain

$$\pi_q = \theta \left[(1-q)^{-1/\alpha} - 1 \right]. \tag{13.6}$$

From Section 20.2, we know that $\mathrm{E}[X] = \frac{\theta}{\alpha - 1}$, and

$$\mathrm{E}[X \wedge \pi_q] = \frac{\theta}{\alpha - 1} \left[1 - \left(\frac{\theta}{\theta + \pi_q} \right)^{\alpha - 1} \right].$$

The second expression in Display (13.3) yields

$$\begin{aligned} ES_q[X] &= \pi_q + \frac{\theta}{\alpha - 1} \frac{[\theta/(\theta + \pi_q)]^{\alpha - 1}}{(\theta/(\theta + \pi_q))^{\alpha}} \\ &= \pi_q + \frac{\theta}{\alpha - 1} \left(\frac{\pi_q + \theta}{\theta} \right) \\ &= \pi_q + \frac{\pi_q + \theta}{\alpha - 1}, \end{aligned}$$

where π_q is given by (13.6).

13.3.3 Coherent Risk Measures

The VaR and ES are widely used risk measures but how does the analyst know which one to employ? Broadly speaking, we seek a function that maps the loss random variable of interest to a numerical value indicating the level of riskiness, which is termed the risk measure. Put mathematically, the risk measure simply summarizes the distribution function of a random variable as a single number.

The VaR and ES are risk measures but one might also consider two simpler alternatives, the mean $\mathrm{E}[X]$ and the standard deviation $\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$. In addition, other classical special cases include the *standard deviation principle*

$$H_{\mathrm{SD}}(X) = \mathrm{E}[X] + \alpha \mathrm{SD}(X), \text{ for } \alpha \geq 0, \tag{13.7}$$

and the *variance principle*

$$H_{\mathrm{Var}}(X) = \mathrm{E}[X] + \alpha \mathrm{Var}(X), \text{ for } \alpha \geq 0.$$

One can check that all the aforementioned functions are risk measures in which we input the loss random variable and the functions output a numerical value. In contrast, the function $H^*(X) = \alpha X^{\beta}$ for any real-valued $\alpha, \beta \neq 0$, is not a risk measure because H^* produces another random variable rather than a single numerical value.

Because risk measures are scalar measures which aim to describe the stochastic uncertainty of loss random variables distributions, it is not surprising that no risk measure can capture all the risk information of the associated random variables. Therefore, when seeking useful risk measures, it is important for us to keep in mind that the measures should be:

- interpretable practically,
- computable conveniently, and

- able to reflect the most critical information of risk underpinning the loss distribution.

Several risk measures have been developed in the literature. Unfortunately, there is no best risk measure that can outperform the others, and the selection of appropriate risk measure depends on the application questions at hand. In this respect, there are multiple approaches to assess the uncertainty. However, for many risk management applications, there is a wide agreement that economically grounded risk measures should satisfy four major axioms, described as follows.

Consider a risk measure $H(\cdot)$. It is said to be a coherent risk measure for two random variables X and Y if the following axioms are satisfied.

- **Axiom 1.** *Subadditivity:* $H(X+Y) \leq H(X) + H(Y)$.
  - The economic implication of this axiom is that diversification benefits exist if different risks are combined.

- **Axiom 2.** *Monotonicity:* if $\Pr[X \leq Y] = 1$, then $H(X) \leq H(Y)$.
  - Recall that X and Y are random variables representing losses, the underlying economic implication is that higher losses essentially leads to a higher level of risk.

- **Axiom 3.** *Positive homogeneity:* $H(cX) = cH(X)$ for any positive constant c.
  - A potential economic implication about this axiom is that risk measure should be independent of the monetary units in which the risk is measured. For example, let c be the currency exchange rate between the US and Canadian dollars, then the risk of random losses measured in terms of US dollars (i.e., X) and Canadian dollars (i.e., cX) should be different only up to the exchange rate c (i.e., $cH(x) = H(cX)$).

- **Axiom 4.** *Translation invariance:* $H(X+c) = H(X) + c$ for any positive constant c.
  - If the constant c is interpreted as risk-free cash and X is an insurance portfolio, then adding cash to a portfolio only increases the portfolio risk by the amount of cash.

Verifying these properties can be straightforward but can be also be challenging at times. For example, it is a simple matter to check that the mean is a coherent risk measure.

Special Case. The Mean is a Coherent Risk Measure.

For any pair of random variables X and Y having finite means and constant $c > 0$,

- validation of *subadditivity*: $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$;
- validation of *monotonicity*: if $\Pr[X \leq Y] = 1$, then $\mathrm{E}[X] \leq \mathrm{E}[Y]$;
- validation of *positive homogeneity*: $\mathrm{E}[cX] = c\mathrm{E}[X]$;
- validation of *translation invariance*: $\mathrm{E}[X+c] = \mathrm{E}[X] + c$

With a little more effort, we can determine the following.

Special Case. The Standard Deviation is not a Coherent Risk Measure.

Verification of the Special Case. To see that the standard deviation is not a coherent risk measure, start by checking that the standard deviation satisfies

Validation of *Subadditivity*.

$$\begin{aligned} \mathrm{SD}[X+Y] &= \sqrt{\mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)} \\ &\leq \sqrt{\mathrm{SD}(X)^2 + \mathrm{SD}(Y)^2 + 2\mathrm{SD}(X)\mathrm{SD}(Y)} \\ &= \mathrm{SD}(X) + \mathrm{SD}(Y); \end{aligned}$$

Validation of *Positive Homogeneity*: $\mathrm{SD}[cX] = c\ \mathrm{SD}[X]$. However, the standard deviation does not comply with translation invariance property as for any positive constant c,

$$\mathrm{SD}(X+c) = \mathrm{SD}(X) < \mathrm{SD}(X) + c.$$

Moreover, the standard deviation also does not satisfy the monotonicity property. To see this, consider the following two random variables:

$$X = \begin{cases} 0, & \text{with probability } 0.25 \\ 4, & \text{with probability } 0.75, \end{cases} \tag{13.8}$$

and Y is a degenerate random variable such that

$$\Pr[Y = 4] = 1. \tag{13.9}$$

You can check that $\Pr[X \leq Y] = 1$, but $\mathrm{SD}(X) = \sqrt{4^2 \cdot 0.25 \cdot 0.75} = \sqrt{3} > \mathrm{SD}(Y) = 0$.

We have so far checked that $\mathrm{E}[\cdot]$ is a coherent risk measure and that $\mathrm{SD}(\cdot)$ is not. Exercise 13.1 asks you to study the coherent property for the standard deviation principle (13.7) which is a linear combination of coherent and incoherent risk measures.

It turns out that the VaR is not a coherent risk measure. Specifically, the VaR measure does not satisfy the subadditivity axiom, meaning that diversification benefits may not be fully reflected.

In contrast, ES is a coherent risk measure and thus is able to more accurately capture the diversification effects of insurance portfolio. Herein, we do not intend to provide the proof of the coherent feature for ES, which is considered to be challenging technically.

13.4 Reinsurance

In this section, you learn how to:

- Define basic reinsurance treaties including proportional, quota share, non-proportional, stop-loss, excess of loss, and surplus share.
- Interpret the optimality of quota share for reinsurers and compute optimal quota share agreements.
- Interpret the optimality of stop-loss for insurers.
- Interpret and calculate optimal excess of loss retention limits.

Recall from Section 5.1.4 that reinsurance is simply insurance purchased by an insurer. Insurance purchased by non-insurers is sometimes known as primary insurance to distinguish it from reinsurance. Reinsurance differs from personal insurance purchased by individuals, such as auto and homeowners insurance, in contract flexibility. Like insurance purchased by major corporations, reinsurance programs are generally tailored more closely to the buyer. For contrast, in personal insurance buyers typically cannot negotiate on the contract terms although they may have a variety of different options (contracts) from which to choose.

The two broad types are proportional and non-proportional reinsurance. A proportional reinsurance contract is an agreement between a reinsurer and a ceding company (also known as the reinsured) in which the reinsurer assumes a given percent of losses and premium. A reinsurance contract is also known as a treaty. Non-proportional agreements are simply everything else. As examples of non-proportional agreements, this chapter focuses on stop-loss and excess of loss contracts. For all types of agreements, we split the total risk X into the portion taken on by the reinsurer, $Y_{reinsurer}$, and that retained by the insurer, $Y_{insurer}$, that is, $X = Y_{insurer} + Y_{reinsurer}$.

The mathematical structure of a basic reinsurance treaty is the same as the coverage modifications of personal insurance introduced in Chapter 5. For a proportional reinsurance, the transformation $Y_{insurer} = cX$ is identical to a coinsurance adjustment in personal insurance. For stop-loss reinsurance, the transformation $Y_{reinsurer} = \max(0, X - M)$ is the same as an insurer's payment with deductible M and $Y_{insurer} = \min(X, M) = X \wedge M$ is equivalent to what a policyholder pays with deductible M. For practical applications of the mathematics, in personal insurance the focus is generally upon the expectation as this is a key ingredient used in pricing. In contrast, for reinsurance the focus is on the entire distribution of the risk, as the extreme events are a primary concern of the financial stability of the insurer and reinsurer.

This section describes the foundational and most basic of reinsurance treaties: Section 13.4.1 for proportional and Section 13.4.2 for non-proportional reinsurance. Section 13.4.3 gives a flavor of more complex contracts.

13.4.1 Proportional Reinsurance

The simplest example of a proportional treaty is called quota share.

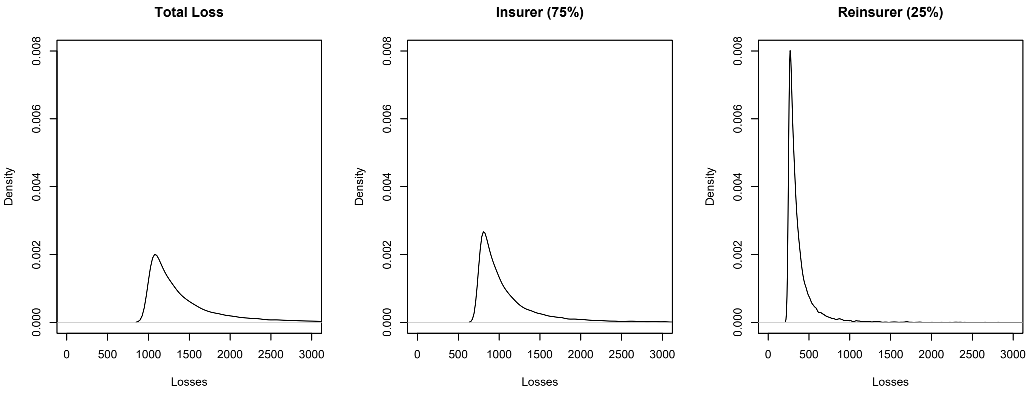
- In a quota share treaty, the reinsurer receives a flat percent, say 50%, of the premium for the book of business reinsured.
- In exchange, the reinsurer pays 50% of losses, including allocated loss adjustment expenses.
- The reinsurer also pays the ceding company a ceding commission which is designed to reflect the differences in underwriting expenses incurred.

The amounts paid by the primary insurer and the reinsurer are summarized as

$$Y_{insurer} = cX \quad \text{and} \quad Y_{reinsurer} = (1-c)X,$$

where $c \in (0,1)$ denotes the proportion retained by the insurer. Note that $Y_{insurer} + Y_{reinsurer} = X$.

Example 13.4.1. Distribution of losses under quota share. To develop an intuition for the effect of quota-share agreement on the distribution of losses, the following is a short `R` demonstration using simulation. The accompanying figure provides the relative shapes of the distributions of total losses, the retained portion (of the insurer), and the reinsurer's portion.



Quota Share is Desirable for Reinsurers

The quota share contract is particularly desirable for the reinsurer. To see this, suppose that an insurer and reinsurer wish to enter a contract to share total losses X such that

$$Y_{insurer} = g(X) \quad \text{and} \quad Y_{reinsurer} = X - g(X),$$

for some generic function $g(\cdot)$ (known as the *retention* function). So that the insurer does not retain more than the loss, we consider only functions so that $g(x) \leq x$. Suppose further that the insurer only cares about the variability of retained claims and is indifferent to the choice of g as long as $\mathrm{Var}(Y_{insurer})$ stays the same and equals, say, Q. Then, the following result shows that the quota share reinsurance treaty minimizes the reinsurer's uncertainty as measured by $\mathrm{Var}(Y_{reinsurer})$.

Proposition. Suppose that $\mathrm{Var}(Y_{insurer}) = Q$ and assume that $Q \leq \mathrm{Var}(X)$. Then, $\mathrm{Var}[(1-c)X] \leq \mathrm{Var}(Y_{reinsurer})$ for all $g(\cdot)$.

Proof of the Proposition. With $Y_{reinsurer} = X - Y_{insurer}$ and the law of total variation

$$\begin{aligned}
\mathrm{Var}(Y_{reinsurer}) &= \mathrm{Var}(X - Y_{insurer}) \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y_{insurer}) - 2Cov(X, Y_{insurer}) \\
&= \mathrm{Var}(X) + Q - 2Corr(X, Y_{insurer}) \times \sqrt{Q}\sqrt{\mathrm{Var}(X)}.
\end{aligned}$$

In this expression, we see that Q and $\mathrm{Var}(X)$ do not change with the choice of the retention function g. Thus, we can minimize $\mathrm{Var}(Y_{reinsurer})$ by maximizing the correlation $Corr(X, Y_{insurer})$. If we use a quota share reinsurance agreement, then $Corr(X, Y_{insurer}) = Corr(X, cX) = 1$, the maximum possible correlation. This establishes the proposition.

The proposition is intuitively appealing - with quota share insurance, the insurer and reinsurer share the responsibility for very large claims in the tail of the distribution. This is in contrast to non-proportional agreements where reinsurers take responsibility for the very large claims.

Optimizing Quota Share Agreements for Insurers

Now assume n risks in the portfolio, $X_1, \ldots, X_n$, so that the portfolio sum is $X = X_1 + \cdots + X_n$. For simplicity, we focus on the case of independent risks (extensions to dependence is the subject of Chapter 16). Each risk X_i may represent risk of an individual policy, claim, or a sub-portfolio, depending on the application. As an example of the latter, the insurer may subdivide its portfolio into subportfolios consisting of lines of business such as (1) personal auto, (2) commercial auto, (3) homeowners, (4) workers' compensation, and so forth.

In general, let us consider a variation of the basic quota share agreement where the amount retained by the insurer may vary with each risk, say c_i. Thus, the insurer's portion of the portfolio risk is $Y_{insurer} = \sum_{i=1}^n c_i X_i$. What is the best choice of the proportions c_i?

To formalize this question, we seek to find those values of c_i that minimize $\mathrm{Var}(Y_{insurer})$ subject to the constraint that $\mathrm{E}(Y_{insurer}) = K$. The requirement that $\mathrm{E}(Y_{insurer}) = K$ suggests that the insurers wishes to retain a revenue in at least the amount of the constant K. Subject to this revenue constraint, the insurer wishes to minimize the uncertainty of the retained risks as measured by the variance.

The Optimal Retention Proportions. Minimizing $\mathrm{Var}(Y_{insurer})$ subject to $\mathrm{E}(Y_{insurer}) = K$ is a constrained optimization problem. We can use the method of Lagrange multipliers, a calculus technique, to solve this. To this end, define the Lagrangian

$$\begin{aligned} L &= \mathrm{Var}(Y_{insurer}) - \lambda(\mathrm{E}(Y_{insurer}) - K) \\ &= \sum_{i=1}^n c_i^2 \ \mathrm{Var}(X_i) - \lambda(\sum_{i=1}^n c_i \ \mathrm{E}(X_i) - K) \end{aligned}$$

Taking a partial derivative with respect to λ and setting this equal to zero simply means that the constraint, $\mathrm{E}(Y_{insurer}) = K$, is enforced and we have to choose the proportions c_i to satisfy this constraint. Moreover, taking the partial derivative with respect to each proportion c_i yields

$$\frac{\partial}{\partial c_i} L = 2c_i \ \mathrm{Var}(X_i) - \lambda \ \mathrm{E}(X_i) = 0$$

so that

$$c_i = \frac{\lambda}{2} \frac{\mathrm{E}(X_i)}{\mathrm{Var}(X_i)}.$$

With our constraint, we may determine λ as the solution of

$$\begin{aligned} K &= \textstyle\sum_{i=1}^n c_i \mathrm{E}(X_i) \\ &= \tfrac{\lambda}{2} \textstyle\sum_{i=1}^n \frac{\mathrm{E}(X_i)^2}{\mathrm{Var}(X_i)} \end{aligned}$$

and use this value of λ to determine the proportions.

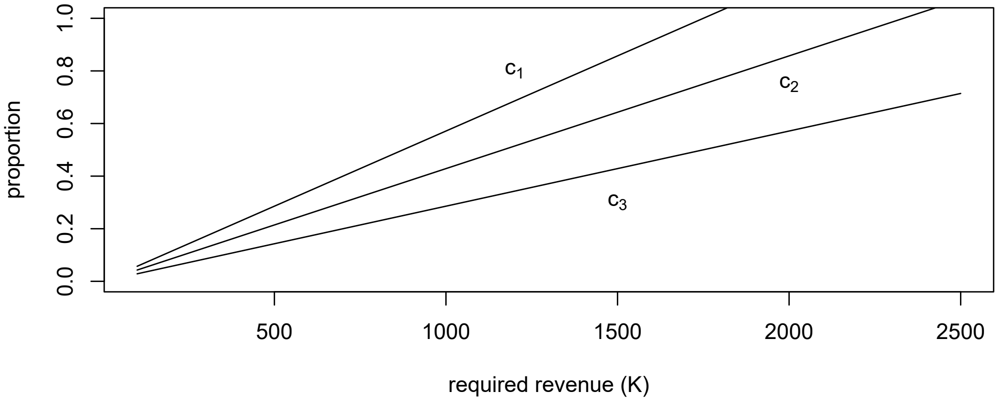
From the math, it turns out that the constant for the ith risk, c_i is proportional to $\frac{\mathrm{E}(X_i)}{\mathrm{Var}(X_i)}$. This is intuitively appealing. Other things being equal, a higher revenue as measured by $\mathrm{E}(X_i)$ means a higher value of c_i. In the same way, a higher value of uncertainty as measured by $\mathrm{Var}(X_i)$ means a lower value of c_i. The proportional scaling factor is determined by the revenue requirement $\mathrm{E}(Y_{insurer}) = K$. The following example helps to develop a feel for this relationship.

Example 13.4.2. Three Pareto risks. Consider three risks that have a Pareto distribution, each having a different set of parameters (so they are independent but non-identical). Specifically, use the parameters:

- $\alpha_1 = 3$, $\theta_1 = 1000$ for the first risk X_1,
- $\alpha_2 = 3$, $\theta_2 = 2000$ for the second risk X_2, and
- $\alpha_3 = 4$, $\theta_3 = 3000$ for the third risk X_3.

Provide a graph that gives values of c_1, c_2, and c_3 for a required revenue K. Note that these values increase linearly with K.

Solution.



13.4.2 Non-Proportional Reinsurance

The Optimality of Stop-Loss Insurance

Under a stop-loss arrangement, the insurer sets a retention level $M(> 0)$ and pays in full total claims for which $X \leq M$. Further, for claims for which $X > M$, the primary insurer pays M and the reinsurer pays the remaining amount $X - M$. That is, the insurer retains an amount M of the risk and the reinsurer pays the excess. Summarizing this mathematically, the amounts paid by the primary insurer and the reinsurer are

$$Y_{insurer} = \begin{cases} X & \text{for } X \leq M \\ M & \text{for } X > M \end{cases} \quad = \min(X, M) = X \wedge M$$

and

$$Y_{reinsurer} = \begin{cases} 0 & \text{for } X \leq M \\ X - M & \text{for } X > M \end{cases} \quad = \max(0, X - M).$$

As before, note that $Y_{insurer} + Y_{reinsurer} = X$.

The stop-loss type of contract is particularly desirable for the insurer. Similar to earlier, suppose that an insurer and reinsurer wish to enter a contract so that $Y_{insurer} = g(X)$ and $Y_{reinsurer} = X - g(X)$ for some generic retention function $g(\cdot)$. Suppose further that the insurer only cares about the variability of retained claims and is indifferent to the choice of g as long as $\mathrm{Var}(Y_{insurer})$ can be minimized. Again, we impose the constraint that $\mathrm{E}(Y_{insurer}) = K$; the insurer needs to retain a revenue K. Subject to this revenue constraint, the insurer wishes to minimize uncertainty of the retained risks (as measured by the variance). Then, the following result shows that the stop-loss reinsurance treaty minimizes the insurer's uncertainty.

Proposition. Suppose that $\mathrm{E}(Y_{insurer}) = K$ and choose M such that $\mathrm{E}(X \wedge M) = K$. Then, $\mathrm{Var}(X \wedge M) \leq \mathrm{Var}[g(X)]$ for all $g(.)$ such that $\mathrm{E}[g(X)] = K$.

Proof of the Proposition. Add and subtract a constant M and expand the square to get

$$\begin{aligned}\mathrm{Var}[g(X)] &= \mathrm{E}[g(X) - K]^2 = \mathrm{E}(g(X) - M + M - K)^2 \\ &= \mathrm{E}[g(X) - M]^2 + (M - K)^2 + 2\mathrm{E}[g(X) - M](M - K) \\ &= \mathrm{E}[g(X) - M]^2 - (M - K)^2,\end{aligned}$$

because $\mathrm{E}[g(X)] = K$.

Now, for any retention function, we have $g(X) \leq X$, that is, the insurer's retained claims are less than or equal to total claims. Using the notation $g_{SL}(X) = X \wedge M$ for stop-loss insurance, we have

$$\begin{aligned}M - g_{SL}(X) &= M - (X \wedge M) \\ &= \max(M - X, 0) \\ &\leq \max(M - g(X), 0).\end{aligned}$$

Squaring each side yields

$$[M - g_{SL}(X)]^2 \leq \max([M - g(X)]^2, 0) \leq [M - g(X)]^2.$$

Returning to our expression for the variance, we have

$$\begin{aligned}\mathrm{Var}[g_{SL}(X)] &= \mathrm{E}[g_{SL}(X) - M]^2 - (M - K)^2 \\ &\leq \mathrm{E}[g(X) - M]^2 - (M - K)^2 = \mathrm{Var}[g(X)],\end{aligned}$$

for any retention function g. This establishes the proposition.

The proposition is intuitively appealing - with stop-loss insurance, the reinsurer takes the responsibility for very large claims in the tail of the distribution, not the insurer.

Excess of Loss

A closely related form of non-proportional reinsurance is the excess of loss coverage. Under this contract, we assume that the total risk X can be thought of as composed as n separate risks $X_1, \ldots, X_n$ and that each of these risks are subject to an upper limit, say, M_i. So the insurer retains

$$Y_{insurer} = \sum_{i=1}^{n} Y_{i,insurer}, \qquad \text{where} \qquad Y_{i,insurer} = X_i \wedge M_i.$$

and the reinsurer is responsible for the excess, $Y_{reinsurer} = X - Y_{insurer}$. The retention limits may vary by risk or may be the same for all risks, that is, $M_i = M$, for all i.

Optimal Choice for Excess of Loss Retention Limits

What is the best choice of the excess of loss retention limits M_i? To formalize this question, we seek to find those values of M_i that minimize $\mathrm{Var}(Y_{insurer})$ subject to the constraint that $\mathrm{E}(Y_{insurer}) = K$. Subject to this revenue constraint, the insurer wishes to minimize the uncertainty of the retained risks (as measured by the variance).

The Optimal Retention Limits. Minimizing $\mathrm{Var}(Y_{insurer})$ subject to $\mathrm{E}(Y_{insurer}) = K$ is a constrained optimization problem. We can use the method of Lagrange multipliers, a calculus technique, to solve this. As before, define the Lagrangian

$$\begin{array}{ll} L & = \mathrm{Var}(Y_{insurer}) - \lambda(\mathrm{E}(Y_{insurer}) - K) \\ & = \sum_{i=1}^n \ \mathrm{Var}(X_i \wedge M_i) - \lambda(\sum_{i=1}^n \ \mathrm{E}(X_i \wedge M_i) - K). \end{array}$$

We first recall the relationships

$$\mathrm{E}(X \wedge M) = \int_0^M \ (1 - F(x))dx$$

and

$$\mathrm{E}(X \wedge M)^2 = 2\int_0^M \ x(1 - F(x))dx.$$

Taking a partial derivative of L with respect to λ and setting this equal to zero simply means that the constraint, $\mathrm{E}(Y_{insurer}) = K$, is enforced and we have to choose the limits M_i to satisfy this constraint. Moreover, taking the partial derivative with respect to each limit M_i yields

$$\begin{array}{ll} \frac{\partial}{\partial M_i} L & = \frac{\partial}{\partial M_i} \ \mathrm{Var}(X_i \wedge M_i) - \lambda \frac{\partial}{\partial M_i} \ \mathrm{E}(X_i \wedge M_i) \\ & = \frac{\partial}{\partial M_i} \left(\mathrm{E}(X_i \wedge M_i)^2 - (\mathrm{E}(X_i \wedge M_i))^2\right) - \lambda(1 - F_i(M_i)) \\ & = 2M_i(1 - F_i(M_i)) - 2\mathrm{E}(X_i \wedge M_i)(1 - F_i(M_i)) - \lambda(1 - F_i(M_i)). \end{array}$$

Setting $\frac{\partial}{\partial M_i} L = 0$ and solving for λ, we get

$$\lambda = 2(M_i - \mathrm{E}(X_i \wedge M_i)).$$

From the math, it turns out that the retention limit less the expected insurer's claims, $M_i - \mathrm{E}(X_i \wedge M_i)$, is the same for *all* risks. This is intuitively appealing.

Example 13.4.3. Excess of loss for three Pareto risks. Consider three risks that have a Pareto distribution, each having a different set of parameters (so they are independent but non-identical). Use the same set of parameters as in Example 13.4.2. For this example:

a. Show numerically that the optimal retention limits M_1, M_2, and M_3 resulting retention limit minus expected insurer's claims, $M_i - \mathrm{E}(X_i \wedge M_i)$, is the same for all risks, as we derived theoretically.
b. Further, graphically compare the distribution of total risks to that retained by the insurer and by the reinsurer.

Solution

a. We first optimize the Lagrangian using the `R` package `alabama` for *Augmented Lagrangian Adaptive Barrier Minimization Algorithm*.

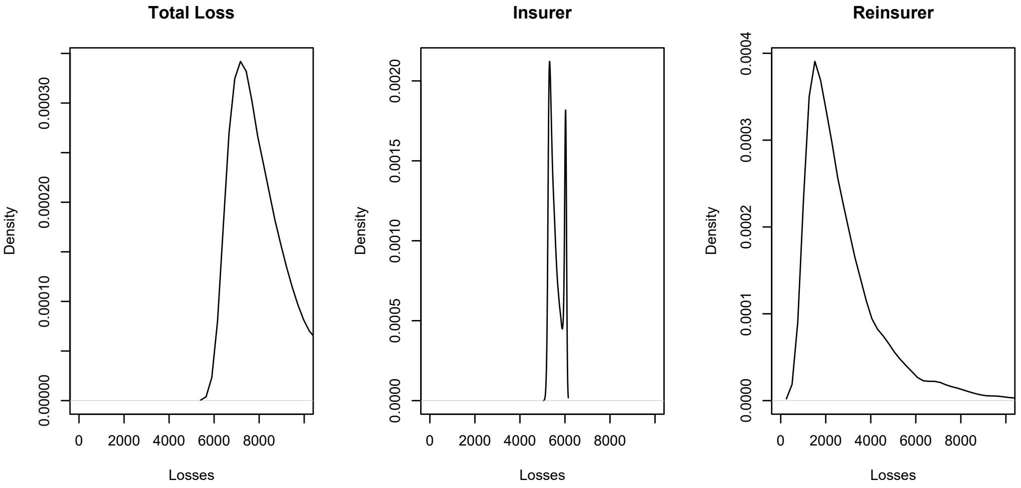
The optimal retention limits M_1, M_2, and M_3 resulting retention limit minus expected insurer's claims, $M_i - \mathrm{E}(X_i \wedge M_i)$, is the same for all risks, as we derived theoretically.

```
[1] 1344.13508

[1] 1344.13325

[1] 1344.13349
```

b. We graphically compare the distribution of total risks to that retained by the insurer and by the reinsurer.



13.4.3 Additional Reinsurance Treaties

Surplus Share Proportional Treaty

Another proportional treaty is known as surplus share; this type of contract is common in commercial property insurance.

- A surplus share treaty allows the reinsured to limit its exposure on a risk to a given amount (the retained line).
- The reinsurer assumes a part of the risk in proportion to the amount that the insured value exceeds the retained line, up to a given limit (expressed as a multiple of the retained line, or number of lines).

For example, let the retained line be 100,000 and the given limit be 4 lines (400,000). Then, if X is the loss, the reinsurer's portion is $\min(400000,(X-100000)_+)$.

Layers of Coverage

One can also extend non-proportional stop-loss treaties by introducing additional parties to the contract. For example, instead of simply an insurer and reinsurer or an insurer and a policyholder, think about the situation with all three parties, a policyholder, insurer, and reinsurer, who agree on how to share a risk. More generally, we consider k parties. If $k=3$, it could be an insurer and two different reinsurers.

Example 13.4.4. Layers of coverage for three parties.

- Suppose that there are $k=3$ parties. The first party is responsible for the first 100 of claims, the second responsible for claims from 100 to 3000, and the third responsible for claims above 3000.
- If there are four claims in the amounts 50, 600, 1800 and 4000, then they would be allocated to the parties as follows:

Layer	Claim 1	Claim 2	Claim 3	Claim 4	Total
(0, 100]	50	100	100	100	350
(100, 3000]	0	500	1700	2900	5100
(3000, ∞)	0	0	0	1000	1000
Total	50	600	1800	4000	6450

To handle the general situation with k groups, partition the positive real line into k intervals using the cut-points

$$0=M_0<M_1<\cdots<M_{k-1}<M_k=\infty.$$

Note that the jth interval is $(M_{j-1}, M_j]$. Now let Y_j be the amount of risk shared by the jth party. To illustrate, if a loss x is such that $M_{j-1} < x \le M_j$, then

$$\left(\begin{array}{c} Y_1 \\ Y_2 \\ \vdots \\ Y_j \\ Y_{j+1} \\ \vdots \\ Y_k \end{array}\right) = \left(\begin{array}{c} M_1 - M_0 \\ M_2 - M_1 \\ \vdots \\ x - M_{j-1} \\ 0 \\ \vdots \\ 0 \end{array}\right)$$

More succinctly, we can write

$$Y_j = \min(X, M_j) - \min(X, M_{j-1}).$$

With the expression $Y_j = \min(X, M_j) - \min(X, M_{j-1})$, we see that the jth party is responsible for claims in the interval $(M_{j-1}, M_j]$. With this, you can check that $X = Y_1 + Y_2 + \cdots + Y_k$. As emphasized in the following example, we also remark that the parties need not be different.

Example 13.4.5.

- Suppose that a policyholder is responsible for the first 100 of claims and all claims in excess of 100,000. The insurer takes claims between 100 and 100,000.
- Then, we would use $M_1 = 100$, $M_2 = 100000$.
- The policyholder is responsible for $Y_1 = \min(X, 100)$ and $Y_3 = X - \min(X, 100000) = \max(0, X - 100000)$.

For additional reading, see the Wisconsin Property Fund site for an example on layers of reinsurance.

Portfolio Management Example

Many other variations of the foundational contracts are possible. For one more illustration, consider the following.

Example 13.4.6. Portfolio Management. You are the Chief Risk Officer of a telecommunications firm. Your firm has several property and liability risks. We will consider:

- X_1 - buildings, modeled using a gamma distribution with mean 200 and scale parameter 100.
- X_2 - motor vehicles, modeled using a gamma distribution with mean 400 and scale parameter 200.

- X_3 - directors and executive officers risk, modeled using a Pareto distribution with mean 1000 and scale parameter 1000.
- X_4 - cyber risks, modeled using a Pareto distribution with mean 1000 and scale parameter 2000.

Denote the total risk as $X = X_1 + X_2 + X_3 + X_4$. For simplicity, you assume that these risks are independent. (Later, in Section 16.6, we will consider the more complex case of dependence.)

To manage the risk, you seek some insurance protection. You wish to manage internally small building and motor vehicles amounts, up to M_1 and M_2, respectively. You seek insurance to cover all other risks. Specifically, the insurer's portion is

$$Y_{insurer} = (X_1 - M_1)_+ + (X_2 - M_2)_+ + X_3 + X_4,$$

so that your retained risk is $Y_{retained} = X - Y_{insurer} = \min(X_1, M_1) + \min(X_2, M_2)$. Using deductibles $M_1 = 100$ and $M_2 = 200$:

a. Determine the expected claim amount of (i) that retained, (ii) that accepted by the insurer, and (iii) the total overall amount.
b. Determine the 80th, 90th, 95th, and 99th percentiles for (i) that retained, (ii) that accepted by the insurer, and (iii) the total overall amount.
c. Compare the distributions by plotting the densities for (i) that retained, (ii) that accepted by the insurer, and (iii) the total overall amount.

Solution.

In preparation, here is the code needed to set the parameters.

With these parameters, we can now simulate realizations of the portfolio risks.

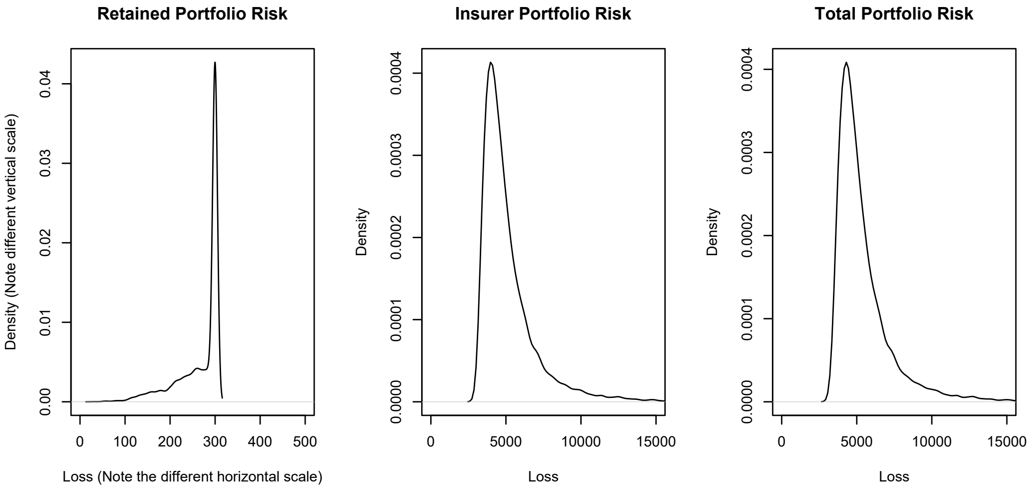
(a) Here are the results for the expected claim amounts.

```
     Retained Insurer   Total
[1,]   269.05 5274.41 5543.46
```

(b) Here are the results for the quantiles.

```
             80%     90%     95%      99%
Retained  300.00  300.00  300.00   300.00
Insurer  6075.67 7399.80 9172.69 14859.02
Total    6351.35 7675.04 9464.20 15159.02
```

(c) Here are the results for the density plots of the retained, insurer, and total portfolio risk.



13.5 Exercises

Theoretical Exercise

Exercise 13.1. In this exercise, you will demonstrate that only under specific circumstances can the standard deviation principle (13.7) be considered a coherent risk measure.

- **a.** Show that subadditivity, positive homogeneity, and translation invariance, hold for the standard deviation principle.
- **b.** Assume that $0 \leq \alpha \leq 1/\sqrt{3}$. Show that for these values of α that monotonicity holds for standard deviation principle. Thus, for these values of α, the standard deviation principle is coherent.
- **c.** For $\alpha > 1/\sqrt{3}$, show that monotonicity does not hold and so the standard deviation principle can not be considered coherent in general.

Verification of the Special Case. To this end, for a given $\alpha > 0$, we check the four axioms for $H_{\mathrm{SD}}(X+Y)$ one by one:

a1 *validation of subadditivity:*

$$\begin{aligned} H_{\mathrm{SD}}(X+Y) &= \mathrm{E}[X+Y] + \alpha \mathrm{SD}(X+Y) \\ &\leq \mathrm{E}[X] + \mathrm{E}[Y] + \alpha[\mathrm{SD}(X) + \mathrm{SD}(Y)] \\ &= H_{\mathrm{SD}}(X) + H_{\mathrm{SD}}(Y); \end{aligned}$$

a2 *validation of positive homogeneity:*

$$H_{\mathrm{SD}}(cX) = c\ \mathrm{E}[X] + c\ \alpha\ \mathrm{SD}(X) = c\ H_{\mathrm{SD}}(X);$$

a3 *validation of translation invariance:*

$$H_{\mathrm{SD}}(X + c) = \mathrm{E}[X] + c + \alpha\ \mathrm{SD}(X) = H_{\mathrm{SD}}(X) + c.$$

b/c *validation of monotonicity*

It only remains to verify the monotonicity property, which may or may not be satisfied depending on the value of α. To see this, consider again the setup of eqrefeq:special-x and
eqrefeq:special-y in which $\Pr[X \leq Y] = 1$. Let $\alpha = 0.1 \cdot \sqrt{3}$, then $H_{\mathrm{SD}}(X) = 3 + 0.3 = 3.3 < H_{\mathrm{SD}}(Y) = 4$ and the monotonicity condition is met. On the other hand, let $\alpha = \sqrt{3}$, then $H_{\mathrm{SD}}(X) = 3+3 = 6 > H_{\mathrm{SD}}(Y) = 4$ and the monotonicity condition is not satisfied. More precisely, by setting

$$H_{\mathrm{SD}}(X) = 3 + \alpha\sqrt{3} \leq 4 = H_{\mathrm{SD}}(Y),$$

we find that the monotonicity condition is only satisfied for $0 \leq \alpha \leq 1/\sqrt{3}$, and thus the standard deviation principle H_{SD} is coherent.

This result appears to be very intuitive since the standard deviation principle H_{SD} is a linear combination of two risk measures of which one is coherent and the other is incoherent. If $\alpha \leq 1/\sqrt{3}$, then the coherent measure dominates the incoherent one, thus the resulting measure H_{SD} is coherent and vice versa. Note that the aforementioned conclusion may not be generalized to any pair of random variables X and Y.

Exercises with a Practical Focus

Exercise 13.2. Property Fund. Consider commercial property claims from the Wisconsin Property Fund, introduced in Section 1.3. This exercise is based on 1,377 claims from 2010 for damages to state government properties and their building contents. You will use these data to estimate an empirical distribution function, without reference to a parametric model.

- **a.** Use the empirical distribution function to estimate VaR over several confidence levels. Produce a graph similar to the left-hand panel of Figure 13.1.
- **b.** Use the empirical distribution function to estimate ES over several confidence levels. Produce a graph similar to the middle panel of Figure 13.1.
- **c.** Compare the two measures from parts (a) and (b) to produce a graph similar to the right-hand panel of Figure 13.1. This comparison shows, for any given level of confidence, that the ES measure far exceeds the VaR.

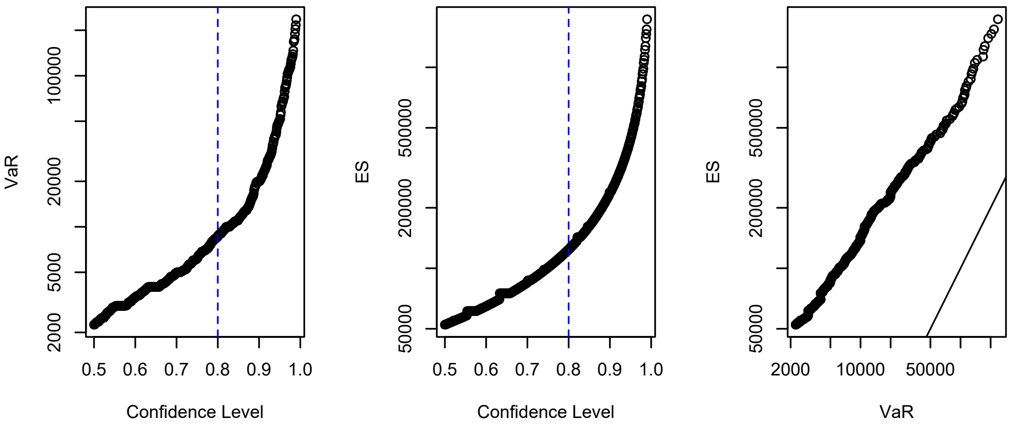


FIGURE 13.1: **Property Fund VaR and ES Plots.** The left-hand panel shows the value at risk VaR for several confidence levels and the middle panel gives similar information for the expected shortfall (ES). The confidence level $\alpha = 0.80$ is marked with a blue dashed vertical line. Note that the vertical axes differ. This is emphasized by direct comparison in the right-hand panel where the 45 degree solid line falls below the empirical values.

Exercise 13.3. Risk Measures with Stop-Loss. Consider the stop-loss arrangement with retention level M described in Section 13.4.2.

a. Show that the value at risk for the retained portion can be expressed as

$$VaR_\alpha[X \wedge M] = \begin{cases} F_\alpha^{-1} & \text{if } \alpha < F(M) \\ M & \text{if } \alpha \geq F(M) \end{cases},$$

where $F_\alpha^{-1} = VaR_\alpha(X)$ is a quantile for a random variable X.

b. Show that the expected shortfall for the retained portion can be expressed as

$$ES_\alpha[X \wedge M] = \begin{cases} F_\alpha^{-1} + \frac{1}{1-\alpha}\left\{\mathrm{E}(X \wedge M) - \mathrm{E}(X \wedge F_\alpha^{-1})\right\} & \text{if } \alpha < F(M) \\ M & \text{if } \alpha \geq F(M) \end{cases}.$$

c. Let us continue Exercise 13.2 where we examined empirical estimates of the distribution using 1,377 property damage claims. We now impose an upper limit M. A confidence level of $\alpha = 0.99$ is used for this illustration. Provide a plot of the value at risk for retained losses under the stop-loss arrangement versus the upper limit M. The plot should be comparable to the left-hand panel of Figure 13.2 where a blue dashed vertical line marks the $\widehat{VaR}_{0.99} = 236427$.

d. Provide a plot of the expected shortfall for retained losses under the stop-loss arrangement versus the upper limit M The plot should be comparable to the right-hand panel of Figure 13.2.

By displaying the figures side-by-side in Figure 13.2, we learn that the ES is smoother at this point when compared to the VaR.

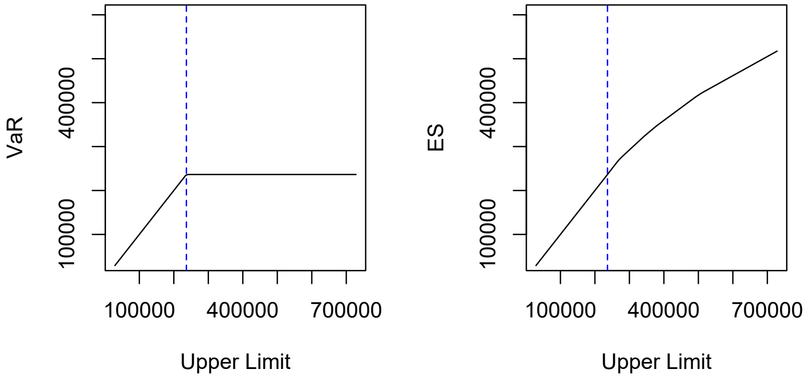


FIGURE 13.2: **Property Fund VaR and ES Plots for Various Upper Limits.** The left-hand panel shows the retained risk VaR over different upper limits and the right-hand panel gives similar information for the expected shortfall (ES). The blue dashed vertical line marks $\widehat{VaR}_{\alpha}$.

Example Solution. a. The distribution function for the limited random variable $X \wedge M$ is

$$\Pr[X \wedge M \leq z] = F_{X \wedge M}(z) = \begin{cases} F(z) & \text{if } z < M \\ 1 & \text{if } z \geq M \end{cases} .$$

(Draw a graph of this function.) From this, if $\alpha \geq F(M)$, then $F^{-1}_{X \wedge M}(\alpha) = M$. In the same way, if $\alpha < F(M)$, then $F^{-1}_{X \wedge M}(\alpha) = F^{-1}_{\alpha}$. This is sufficient for part (a).

b. From equation
eqrefeq:ESExpressions, the expected shortfall for retained risks can be expressed

as

$$
\begin{aligned}
ES_{\alpha}[X \wedge M] &= F^{-1}_{X \wedge M}(\alpha) + \tfrac{1}{1-\alpha}\Big\{\mathrm{E}[X \wedge M] - \mathrm{E}[X \wedge M \wedge F^{-1}_{X \wedge M}(\alpha)]\Big\} \\
&= \begin{cases} F^{-1}_{\alpha} + \tfrac{1}{1-\alpha}\left\{\mathrm{E}[X \wedge M] - \mathrm{E}[X \wedge M \wedge F^{-1}_{\alpha}]\right\} & \text{if } \alpha < F(M) \\ M + \tfrac{1}{1-\alpha}\left\{\mathrm{E}[X \wedge M] - \mathrm{E}[X \wedge M \wedge M]\right\} & \text{if } \alpha \geq F(M) \end{cases} . \\
&= \begin{cases} F^{-1}_{\alpha} + \tfrac{1}{1-\alpha}\left\{\mathrm{E}[X \wedge M] - \mathrm{E}[X \wedge F^{-1}_{\alpha}]\right\} & \text{if } \alpha < F(M) \\ M & \text{if } \alpha \geq F(M) \end{cases} ,
\end{aligned}
$$

as desired.

c/d.

You can learn more about the R code for this example at the online version of this book, Actuarial Community (2025).

13.6 Further Resources and Contributors

We refer the interested reader to Denuit et al. (2006) and Hardy (2006) for more comprehensive discussions of alternative risk measures for both discrete and continuous random variables. Note, however, that the definition in Denuit et al. (2006) of "expected shortfall" differs from the one in this text. We use the definition of expected shortfall from Wang and Zitikis (2022).

As summarized in Wang and Zitikis (2022), both VaR and ES have solid axiomatic foundations and "appear in the banking regulation frameworks of Basel III/IV, as well as in the insurance regulation frameworks of Solvency II and the Swiss Solvency Test." In addition to the coherence properties introduced in Section 13.3.3, this paper introduces economic axioms to motivate the use of ES. Thus, their usefulness in determining adequate solvency for banks and insurers motivates our emphasis in Section 13.3 of these measures.

Concepts of pricing individual risks were introduced in Chapter 10. For a comprehensive treatment of pricing portfolios, we refer to Mildenhall and Major (2022).

There are many superb treatments of reinsurance in the literature. An outstanding book-long introduction is Albrecher et al. (2017).

Some of the examples from this chapter were borrowed from Clark (1996), Klugman et al. (2012), and Bahnemann (2015). These resources provide excellent sources for additional discussions and examples.

- **Edward (Jed) Frees**, University of Wisconsin-Madison, and **Jianxi Su**, Purdue University were the principal authors of the initial version of this chapter.
  - Chapter reviewers include: Fei Huang, Hirokazu (Iwahiro) Iwasawa, Peng Shi, Ranee Thiagarajah, Ping Wang, and Chengguo Weng.
- **Edward (Jed) Frees**, University of Wisconsin-Madison and Australian National University, is the author of the second edition of this chapter. Email: jfrees@bus.wisc.edu for chapter comments and suggested improvements.
  - Chapter reviewers include Chengguo Weng.

14

Loss Reserving

Chapter Preview. This chapter introduces loss reserving (also known as claims reserving) for property and casualty (P&C, or general, non-life) insurance products. In particular, the chapter sketches some basic, though essential, analytic tools to assess the reserves on a portfolio of P&C insurance products. First, Section 14.1 motivates the need for loss reserving, then Section 14.2 studies the available data sources and introduces some formal notation to tackle loss reserving as a prediction challenge. Next, Section 14.3 covers the chain-ladder method and Mack's distribution-free chain-ladder model. Section 14.4 then develops a fully stochastic approach to determine the outstanding reserve with generalized linear models (GLMs), including the technique of bootstrapping to obtain a predictive distribution of the outstanding reserve via simulation.

14.1 Motivation

Our starting point is the lifetime of a P&C insurance claim. Figure 14.1 pictures the development of such a claim over time and identifies the events of interest:

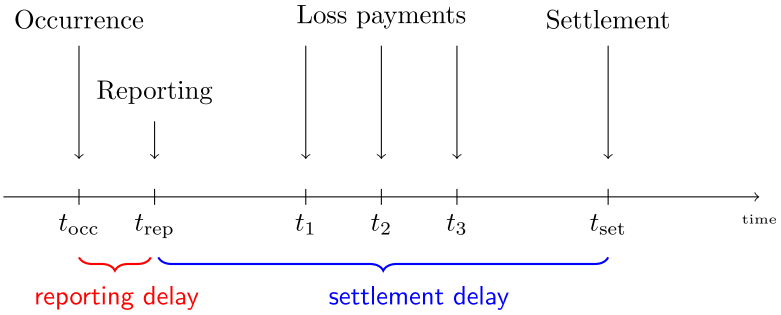


FIGURE 14.1: **Lifetime or Run-off of a Claim**

The insured event or accident occurs at time t_{occ}. This incident is reported to the insurance company at time t_{rep}, after some delay. If the filed claim is accepted by the insurance company, payments will follow to reimburse the financial loss of the policyholder. In this example the insurance company compensates the incurred loss with loss payments at times t_1, t_2 and t_3. Eventually, the claim settles or closes at time t_{set}.

Often claims will not settle immediately due to the presence of delay in the reporting of a claim, delay in the settlement process or both. The reporting delay is the time that elapses between the occurrence of the insured event and the reporting of this event to the insurance company. The time between reporting and settlement of a claim is known as the settlement delay. For example, it is very intuitive that a material or property damage claim settles quicker than a bodily injury claim involving a complex type of injury. Closed claims may also reopen due to new developments, e.g. an injury that requires extra treatment. Put together, the development of a claim typically takes some time. The presence of this delay in the run-off of a claim requires the insurer to hold capital in order to settle these claims in the future.

14.1.1 Closed, IBNR, and RBNS Claims

Based on the status of the claim's run-off we distinguish three types of claims in the books of an insurance company. A first type of claim is a closed claim. For these claims the complete development has been observed. With the red line in Figure 14.2 indicating the present moment, all events from the claim's development take place before the present moment. Hence, these events are observed at the present moment. For convenience, we will assume that a closed claim can not reopen.

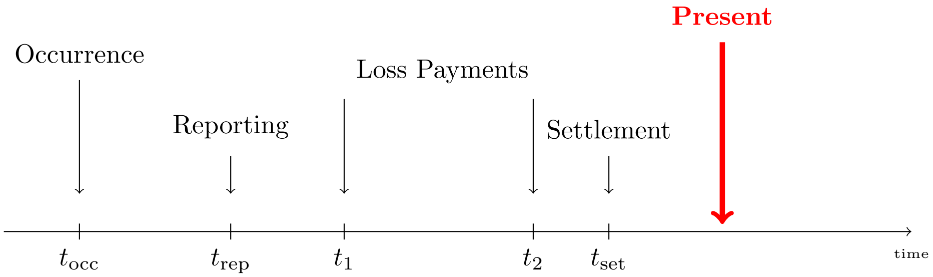


FIGURE 14.2: **Lifetime of a Closed Claim**

An RBNS claim is one that has been **R**eported, **B**ut is **N**ot fully **S**ettled at the present moment or the moment of evaluation (the valuation date), that is, the moment when the reserves should be calculated and booked by the insurer. Occurrence, reporting and possibly some loss payments take place before the present moment, but the closing of the claim happens in the future, beyond the present moment.

An IBNR claim is one that has **I**ncurred in the past **B**ut is **N**ot yet **R**eported.

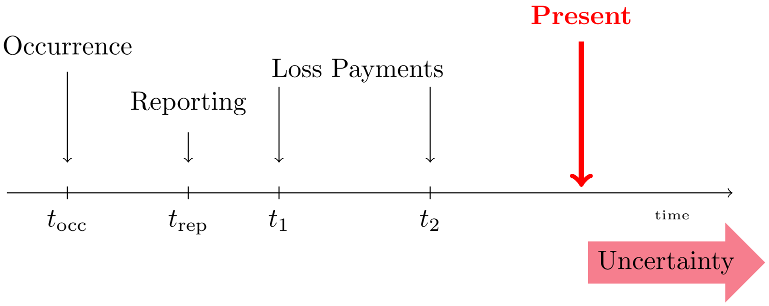


FIGURE 14.3: **Lifetime of an RBNS Claim**

For such a claim the insured event took place, but the insurance company is not yet aware of the associated claim. This claim will be reported in the future and its complete development (from reporting to settlement) takes place in the future.

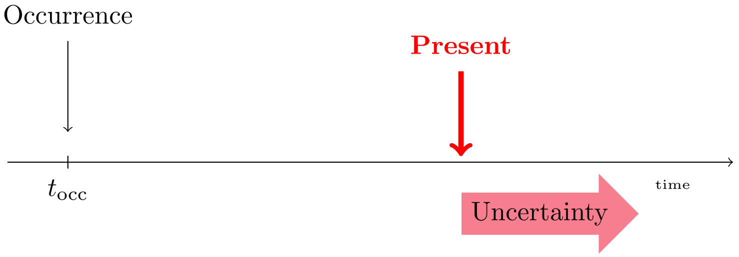


FIGURE 14.4: **Lifetime of an IBNR Claim**

Insurance companies will reserve capital to fulfill their future liabilities with respect to both RBNS as well as IBNR claims. The future development of such claims is uncertain and predictive modeling techniques will be used to calculate appropriate reserves, from the historical development data observed on similar claims.

14.1.2 Why Reserving?

The inverted production cycle of the insurance market and the claim dynamics pictured in Section 14.1.2 motivate the need for reserving and the design of predictive modeling tools to estimate reserves. In insurance, the premium income precedes the costs. An insurer will charge a client a premium, before actually knowing how costly the insurance policy or contract will become. In typical manufacturing industry this is not the case and the manufacturer knows - before selling a product - what the cost of producing this product was. At a specified evaluation moment τ the insurer will predict outstanding liabilities with respect to contracts sold in the past. This is the claims reserve or loss reserve; it is the capital necessary to settle open claims from past exposures. It is a very important element on the balance sheet of the insurer, more specifically on the liabilities side of the balance sheet.

14.2 Loss Reserve Data

14.2.1 From Micro to Macro

We now shed light on the data available to estimate the outstanding reserve for a portfolio of P&C contracts. Insurance companies typically register data on the development of an individual claim as sketched in the timeline on the left hand side of Figure 14.5. We refer to data registered at this level as **granular or micro-level** data. Typically, an actuary aggregates the information registered on the individual development of claims across all claims in a portfolio. This aggregation leads to data structured in a triangular format as shown on the right hand side of Figure 14.5. Such data are called **aggregate or macro-level** data because each cell in the triangle displays information obtained by aggregating the development of multiple claims.

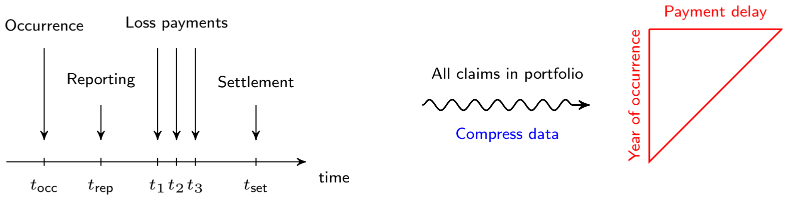


FIGURE 14.5: **From Granular Data to Run-off Triangle**

The triangular display used in loss reserving is called a **run-off or development** triangle. On the vertical axis the triangle lists the accident or occurrence years during which a portfolio is followed. The loss payments booked for a specific claim are connected to the year during the which the insured event occurred. The horizontal axis indicates the payment delay since occurrence of the insured event.

14.2.2 Run-off Triangles

A first example of a run-off triangle with incremental payments is displayed in Figure 14.6 (taken from Wüthrich and Merz (2008), Table 2.2, also used in Wüthrich and Merz (2015), Table 1.4). Accident years (or years of occurrence) are shown on the vertical axis and run from 2004 up to 2013. These refer to the year during which the insured event occurred. The horizontal axis indicates the payment delay in years since occurrence of the insured event. *0 delay* is used for payments made in the year of occurrence of the accident or insured event. *One year* of delay is used for payments made in the year after occurrence of the accident.

accident	payment delay (in years)									
year	0	1	2	3	4	5	6	7	8	9
2004	5,947.0	3,721.2	895.7	207.8	206.7	621.2	658.1	148.5	111.3	158.1
2005	6,346.8	3,246.4	723.2	151.8	678.2	366.0	527.5	111.9	116.5	
2006	6,269.1	2,976.2	8470.5	262.8	152.7	654.4	535.5	892.4		
2007	5,863	2,683.2	722.5	190.7	133.0	883.4	433.3			
2008	5,778.9	2,745.2	653.9	273.4	230.3	105.2				
2009	6,184.8	2,828.3	572.8	244.9	105.0					
2010	5,600.2	2,893.2	563.1	225.5						
2011	5,288.1	2,440.1	528.0							
2012	5,290.8	2,357.9								
2013	5,675.6									

FIGURE 14.6: **A Run-off Triangle with Incremental Payment Data.** *Source: Wüthrich and Merz (2008), Table 2.2.*

For example, cell $(2004, 0)$ in the above triangle displays the number $5,947$, the total amount paid in the year 2004 for all claims occurring in year 2004. Thus, it is the total amount paid with 0 years of delay on all claims that occurred in the year 2004. Similarly, the number in cell $(2012, 1)$ displays the total $2,357.9$ paid in the year 2013 for all claims that occurred in year 2012.

accident	payment delay (in years)									
year	0	1	2	3	4	5	6	7	8	9
2004	5,947	9,668	10,564	10,772	10,978	11,041	11,106	11,121	11,132	11,148
2005	6,347	9,593	10,316	10,468	10,536	10,573	10,625	10,637	10,648	
2006	6,269	9,245	10,092	10,355	10,508	10,573	10,627	10,636		
2007	5,863	8,546	9,269	9,459	9,592	9,681	9,724			
2008	5,779	8,524	9,178	9,451	9,682	9,787				
2009	6,185	9,013	9,586	9,831	9,936					
2010	5,600	8,493	9,057	9,282						
2011	5,288	7,728	8,256							
2012	5,291	7,649								
2013	5,676									

FIGURE 14.7: **A Run-off Triangle with Cumulative Payment Data.** *Source: Wüthrich and Merz (2008), Table 2.2.*

Whereas the triangle in Figure 14.6 displays incremental payment data, the Figure 14.7 shows the same information in cumulative format. Now, cell $(2004, 1)$ displays the total claim amount paid *up to* payment delay 1 for all claims that occurred in year 2004. Therefore, it is the sum of the amount paid in 2004 and the amount paid in 2005 on accidents that occurred in 2004.

Different pieces of information can be stored in run-off triangles as those shown in Figure 14.6 and Figure 14.7. Depending on the kind of data stored, the triangle will be used to estimate different quantities.

For example, in incremental format a cell may display:

- the claim payments, as motivated before
- the number of claims that occurred in a specific year and were reported with a certain delay, when the goal is to estimate the number of IBNR claims
- the change in incurred amounts, where incurred claim amounts are the sum of cumulative paid claims and the case estimates. The case estimate is the claims handler's expert estimate of the outstanding amount on a claim.

In cumulative format a cell may display:

- the cumulative paid amount, as motivated before
- the total number of claims from an occurrence year, reported up to a certain delay
- the incurred claim amounts.

Other sources of information are potentially available, e.g. covariates (like the type of claim), external information (like inflation, change in regulation). Most claims reserving methods designed for run-off triangles are rather based on a single source of information, although recent contributions focus on the use of more detailed data for loss reserving.

14.2.3 Loss Reserve Notation

Run-off Triangles

To formalize the displays shown in Figures 14.6 and 14.7, we let i refer to the occurrence or accident year, the year in which the insured event happened. In our notation the first accident year considered in the portfolio is denoted with 1 and the latest, most recent accident year is denoted with I. Then, j refers to the payment delay or development year, where a delay equal to 0 corresponds to the accident year itself. Figure 14.8 shows a triangle where the same number of years is considered in both the vertical as well as the horizontal direction, hence j runs from 0 up to $J = I - 1$.

The random variable X_{ij} denotes the incremental claims paid in development period j on claims from accident year i. Thus, X_{ij} is the total amount paid in development year j for all claims that happened in occurrence year i. These payments are actually paid out in accounting or calendar year $i + j$. Taking a cumulative point of view, C_{ij} is the cumulative amount paid up until (and including) development year j for accidents that occurred in year i. Ultimately, a total amount C_{iJ} is paid in the final development year J for claims that

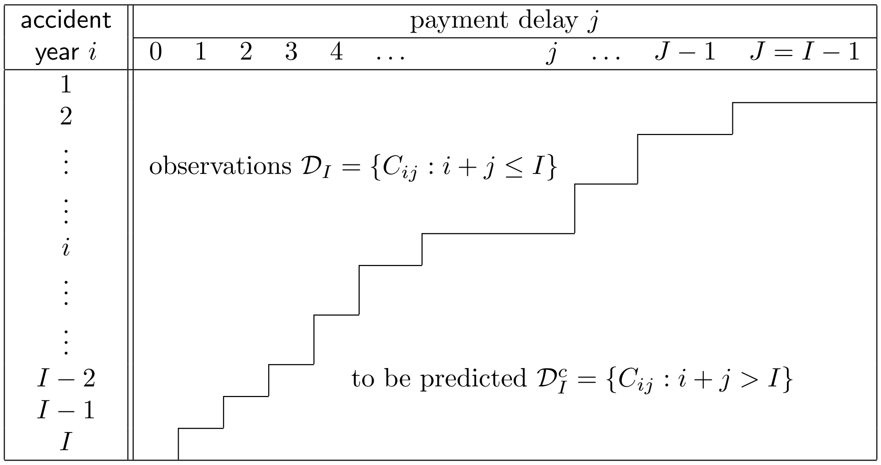

accident year i
payment delay j
0
1
2
3
4
...
j
...
$J-1$
$J=I-1$
1
2
$\vdots$
$\vdots$
i
$\vdots$
$\vdots$
$I-2$
$I-1$
I
observations $\mathcal{D}_I = \{C_{ij} : i+j \leq I\}$
to be predicted $\mathcal{D}_I^c = \{C_{ij} : i+j > I\}$

FIGURE 14.8: **Mathematical notation for a run-off triangle.** *Source:* Wüthrich and Merz (2008)

occurred in accident year i. In this chapter time is expressed in years, though other time units can be used as well, e.g. six-month periods or quarters.

The Loss Reserve

At the evaluation moment τ, the data in the upper triangle have been observed, whereas the lower triangle has to be predicted. Here, the evaluation moment is the end of accident year I which implies that a cell (i,j) with $i+j \leq I$ is observed, and a cell (i,j) with $i+j > I$ belongs to the future and has to be predicted. Thus, for a cumulative run-off triangle, the goal of a loss reserving method is to predict $C_{i,I-1}$, the ultimate claim amount for occurrence year i, corresponding to the final development period $I-1$ in Figure 14.7. We assume that - beyond this period - no further payments will follow, although this assumption can be relaxed.

Since $C_{i,I-1}$ is cumulative, it includes both an observed part as well as a part that has to be predicted. Therefore, the outstanding liability or loss reserve for accident year i is

$$\mathcal{R}_i^{(0)} = \sum_{\ell=I-i+1}^{I-1} X_{i\ell} = C_{i,I} - C_{i,I-i}.$$

We express the reserve either as a sum of incremental data, the $X_{i\ell}$, or as a difference between cumulative numbers. In the latter case the outstanding amount is the ultimate cumulative amount $C_{i,I}$ minus the most recently observed cumulative amount $C_{i,I-i}$. Following Wüthrich and Merz (2015), the notation $\mathcal{R}_i^{(0)}$ refers to the reserve for occurrence year i where $i = 1, \ldots, I$. The superscript (0) refers to the evaluation of the reserve at the present moment, say $\tau = 0$. We understand $\tau = 0$ at the end of occurrence year I, the most recent calendar year for which data are observed and registered.

14.2.4 R Code to Summarize Loss Reserve Data

We use the ChainLadder package (Gesmann et al., 2019) to import run-off triangles in R and to explore the trends present in these triangles. The package's vignette nicely documents its functions for working with triangular data. First, we explore two ways to import a triangle.

Long Format Data

The dataset triangle_W_M_long.txt stores the cumulative run-off triangle from Wüthrich and Merz (2008) (Table 2.2) in long format. That is: each cell in the triangle is one row in this data set, and three features are stored: the payment size (cumulative, in this example), the year of occurrence (i) and the payment delay (j). We import the .txt file and store the resulting data frame as my_triangle_long:

```r
my_triangle_long <- read.table("Data/triangle_W_M_long.txt", header = TRUE)
```

We use the as.triangle function from the ChainLadder package to transform the data frame into a triangular display. The resulting object my_triangle is now of type triangle.

```r
my_triangle <- as.triangle(my_triangle_long, origin = "origin", dev = "dev", value = "payment")
```

We display the triangle and recognize the numbers (in thousands) in Figure 14.7. Cells in the lower triangle are indicated as *not available*, NA.

Triangular Format Data

Alternatively, the triangle may be stored in a .csv file with the occurrence years in the rows and the development years in the column cells. We import this .csv file and transform the resulting my_triangle_csv to a matrix.

```r
my_triangle_csv <- read.csv2("Data/triangle_W_M.csv", header = FALSE)
my_triangle_matrix <- as.matrix(my_triangle_csv)
dimnames(my_triangle_matrix) <- list(origin = 2004:2013, dev = 0:(ncol(my_triangle_matrix) -
    1))
```

We inspect the triangle:

From Cumulative to Incremental, and vice versa

The R functions cum2incr() and incr2cum() enable us to switch from cumulative to incremental displays, and vice versa, in an easy way.

```
my_triangle_incr <- cum2incr(my_triangle)
```

We recognize the incremental triangle from Figure 14.6.

Visualizing Triangles

To explore the evolution of the cumulative payments per occurrence year, Figure 14.9 shows `my_triangle` using the `plot` function available for objects of type `triangle` in the `ChainLadder` package. Each line in this plot depicts an occurrence year (from 2004 to 2013, labelled as 1 to 10). Development periods are labelled from 1 to 10 (instead of 0 to 9, as used above).

```
plot(my_triangle)
```

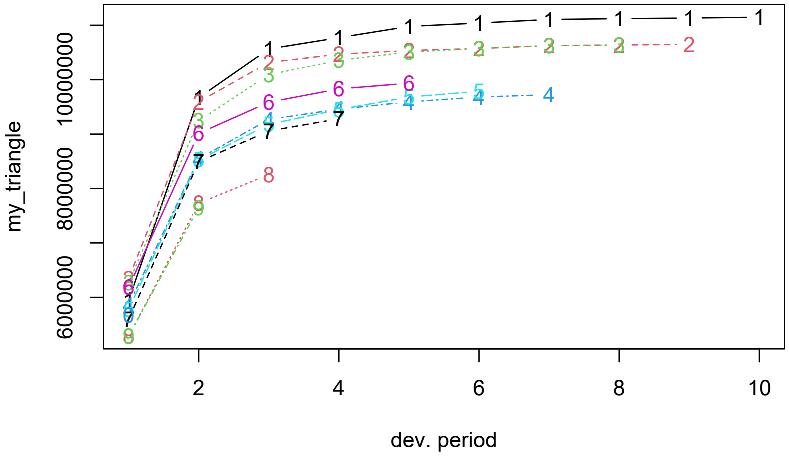
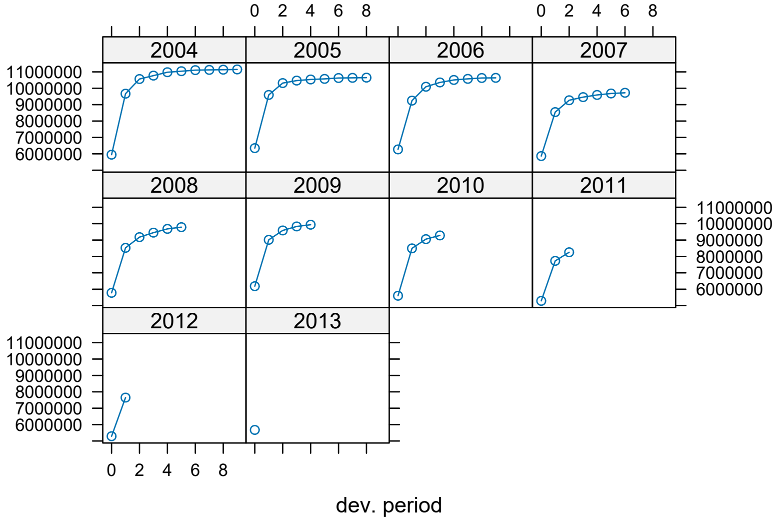


FIGURE 14.9: **Claim Development by Occurrence Year**

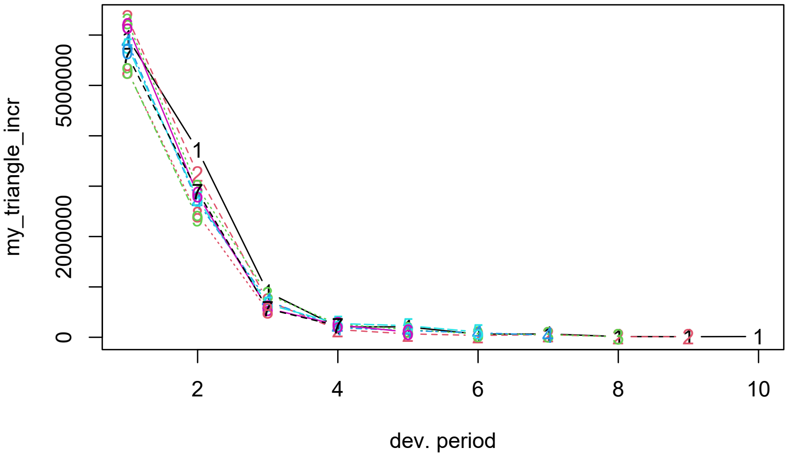
Alternatively, the `lattice` argument creates one plot per occurrence year.

```
plot(my_triangle, lattice = TRUE)
```

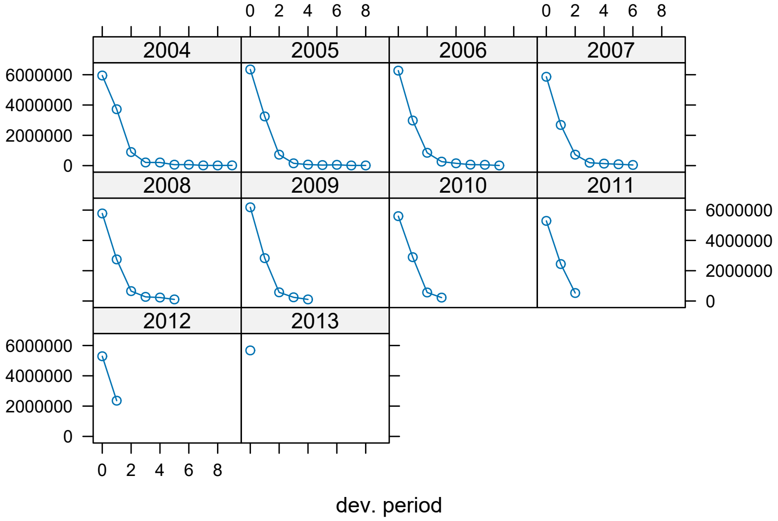


Instead of plotting the cumulative triangle stored in `my_triangle`, we can plot the incremental run-off triangle.

```
plot(my_triangle_incr)
```



```
plot(my_triangle_incr, lattice = TRUE)
```



14.3 The Chain-Ladder Method

The most widely used method to estimate outstanding loss reserves is the so-called chain-ladder method. The origins of this method are obscure but was firmly entrenched in practical applications by the early 1970's, Taylor (1986). As will be seen, the name refers to the chaining of a sequence of (year-to-year development) factors into a ladder of factors; immature losses climb toward maturity when multiplied by this concatenation of ratios, hence the apt descriptor *chain-ladder method.* We will start with exploring the chain-ladder method in its deterministic or algorithmic version, hence without making any stochastic assumptions. Then we will describe Mack's distribution-free chain-ladder model.

14.3.1 The Deterministic Chain-Ladder

The deterministic chain-ladder method focuses on the run-off triangle in cumulative form. Recall that a cell (i, j) in this triangle displays the cumulative amount paid up until development period j for claims that occurred in year i. The chain-ladder method assumes that **development factors** f_j (also called age-to-age factors, link ratios or chain-ladder factors) exist such that

$$C_{i,j+1} = f_j \times C_{i,j}.$$

Thus, the development factor tells you how the cumulative amount in development year j grows to the cumulative amount in year $j + 1$. We highlight the cumulative amount in period 0 in blue and the cumulative amount in period 1 in red on the Figure 14.10 taken from Wüthrich and Merz (2008) (Table 2.2, also used in Wüthrich and Merz (2015), Table 1.4).

accident	payment delay (in years)									
year	0	1	2	3	4	5	6	7	8	9
1	5,947	9,668	10,564	10,772	10,978	11,041	11,106	11,121	11,132	11,148
2	6,347	9,593	10,316	10,468	10,536	10,573	10,625	10,637	10,648	
3	6,269	9,245	10,092	10,355	10,508	10,573	10,627	10,636		
4	5,863	8,546	9,269	9,459	9,592	9,681	9,724			
5	5,779	8,524	9,178	9,451	9,682	9,787				
6	6,185	9,013	9,586	9,831	9,936					
7	5,600	8,493	9,057	9,282						
8	5,288	7,728	8,256							
9	5,291	7,649								
10	5,676									

FIGURE 14.10: **A Run-off Triangle with Cumulative Payment Data Highlighting the Cumulative Amount in Period 0 in Blue and the Cumulative Amount in Period 1 in Red.** *Source: Wüthrich and Merz (2008), Table 2.2.*

The chain-ladder method then presents an intuitive recipe to estimate or calculate these development factors. Since the first development factor f_0 describes the development of the cumulative claim amount from development period 0 to development period 1, it can be estimated as the ratio of the cumulative amounts in red and the cumulative amounts in blue, highlighted in the Figure 14.10. We then obtain the following estimate $\hat{f}_0^{CL}$ for the first development factor f_0, given observations $\mathcal{D}_I$:

$$\hat{f}_0^{CL} = \frac{\sum_{i=1}^{10-0-1} C_{i,0+1}}{\sum_{i=1}^{10-0-1} C_{i0}} = 1.4925.$$

Note that the index i, used in the sums in the numerator and denominator, runs from the first occurrence period (1) to the last occurrence period (9) for which both development periods 0 and 1 are observed. As such, this development factor measures how the data in blue grow to the data in red, averaged across all occurrence periods for which both periods are observed. The chain-ladder method then uses this development factor estimator to predict the cumulative amount $C_{10,1}$ (i.e. the cumulative amount paid up until and including development year 1 for accidents that occurred in year 10). This prediction is obtained by multiplying the most recent observed cumulative claim amount for occurrence period 10 (i.e. $C_{10,0}$ with development period 0) with the estimated development factor $\hat{f}_0^{CL}$:

$$\hat{C}_{10,1} = C_{10,0} \cdot \hat{f}_0^{CL} = 5,676 \cdot 1.4925 = 8,471.$$

Going forward with this reasoning, the next development factor f_1 can be estimated. Since f_1 captures the development from period 1 to period 2, it can be estimated as the ratio of the numbers in red and the numbers in blue as highlighted in Figure 14.11.

accident	payment delay (in years)									
year	0	1	2	3	4	5	6	7	8	9
1	5,947	9,668	10,564	10,772	10,978	11,041	11,106	11,121	11,132	11,148
2	6,347	9,593	10,316	10,468	10,536	10,573	10,625	10,637	10,648	
3	6,269	9,245	10,092	10,355	10,508	10,573	10,627	10,636		
4	5,863	8,546	9,269	9,459	9,592	9,681	9,724			
5	5,779	8,524	9,178	9,451	9,682	9,787				
6	6,185	9,013	9,586	9,831	9,936					
7	5,600	8,493	9,057	9,282						
8	5,288	7,728	8,256							
9	5,291	7,649								
10	5,676									

FIGURE 14.11: **A Run-off Triangle with Cumulative Payment Data Highlighting the Cumulative Amount in Period 1 in Blue and the Cumulative Amount in Period 2 in Red.** *Source: Wüthrich and Merz (2008), Table 2.2.*

The mathematical notation of the estimate $\hat{f}_1^{CL}$ for the next development factor f_1, given observations $\mathcal{D}_I$, equals:

$$\hat{f}_1^{CL} = \frac{\sum_{i=1}^{10-1-1} C_{i,1+1}}{\sum_{i=1}^{10-1-1} C_{i1}} = 1.0778.$$

Consequently, this factor measures how the cumulative paid amount in development period 1 grows to period 2, averaged across all occurrence periods for which both periods are observed. The index i now runs from period 1 to 8, since these are the occurrence periods for which both development periods 1 and 2 are observed. This estimate for the second development factor is then used to predict the missing, unobserved cells in development period 2:

$$\begin{aligned} \hat{C}_{10,2} &= C_{10,0} \cdot \hat{f}_0^{CL} \cdot \hat{f}_1^{CL} = \hat{C}_{10,1} \cdot \hat{f}_1^{CL} = 8,471 \cdot 1.0778 = 9,130 \\ \hat{C}_{9,2} &= C_{9,1} \cdot \hat{f}_1^{CL} = 7,649 \cdot 1.0778 = 8,244. \end{aligned}$$

Note that for $\hat{C}_{10,2}$ you actually use the estimate $\hat{C}_{10,1}$ and multiply it with the estimated development factor $\hat{f}_1^{CL}$.

We continue analogously and obtain following predictions, printed in italics in the Figure 14.12.

Eventually we need to estimate the values in the final column. The last development factor f_8 measures the growth from development period 8 to

accident	payment delay (in years)									
year	0	1	2	3	4	5	6	7	8	9
1	5,947	9,668	10,564	10,772	10,978	11,041	11,106	11,121	11,132	11,148
2	6,347	9,593	10,316	10,468	10,536	10,573	10,625	10,637	10,648	
3	6,269	9,245	10,092	10,355	10,508	10,573	10,627	10,636	*10,647*	
4	5,863	8,546	9,269	9,459	9,592	9,681	9,724	*9,735*	*9,745*	
5	5,779	8,524	9,178	9,451	9,682	9,787	*9,837*	*9,848*	*9,858*	
6	6,185	9,013	9,586	9,831	9,936	*10,005*	*10,057*	*10,067*	*10,078*	
7	5,600	8,493	9,057	9,282	*9,420*	*9,485*	*9,534*	*9,545*	*9,555*	
8	5,288	7,728	8,256	*8,445*	*8,570*	*8,630*	*8,675*	*8,684*	*8,693*	
9	5,291	7,649	*8,243*	*8,432*	*8,557*	*8,617*	*8,661*	*8,671*	*8,680*	
10	5,676	*8,471*	*9,130*	*9,339*	*9,477*	*9,543*	*9,592*	*9,603*	*9,613*	
$\hat{f}^{CL}$	1.493	1.078	1.023	1.015	1.007	1.005	1.001	1.001		

FIGURE 14.12: **A Run-off Triangle with Cumulative Payment Data Including Predictions in Italic** *Source: Wüthrich and Merz (2008), Table 2.2.*

development period 9 in the triangle. Since only the first row in the triangle has both cells observed, this last factor is estimated as the ratio of the value in red and the value in blue in Figure 14.13.

accident	payment delay (in years)									
year	0	1	2	3	4	5	6	7	8	9
1	5,947	9,668	10,564	10,772	10,978	11,041	11,106	11,121	11,132	11,148
2	6,347	9,593	10,316	10,468	10,536	10,573	10,625	10,637	10,648	
3	6,269	9,245	10,092	10,355	10,508	10,573	10,627	10,636	*10,647*	
4	5,863	8,546	9,269	9,459	9,592	9,681	9,724	*9,735*	*9,745*	
5	5,779	8,524	9,178	9,451	9,682	9,787	*9,837*	*9,848*	*9,858*	
6	6,185	9,013	9,586	9,831	9,936	*10,005*	*10,057*	*10,067*	*10,078*	
7	5,600	8,493	9,057	9,282	*9,420*	*9,485*	*9,534*	*9,545*	*9,555*	
8	5,288	7,728	8,256	*8,445*	*8,570*	*8,630*	*8,675*	*8,684*	*8,693*	
9	5,291	7,649	*8,243*	*8,432*	*8,557*	*8,617*	*8,661*	*8,671*	*8,680*	
10	5,676	*8,471*	*9,130*	*9,339*	*9,477*	*9,543*	*9,592*	*9,603*	*9,613*	
$\hat{f}^{CL}$	1.493	1.078	1.023	1.015	1.007	1.005	1.001	1.001		

FIGURE 14.13: **A Run-off Triangle with Cumulative Payment Data Highlighting the Cumulative Amount in Period 8 in Blue and the Cumulative Amount in Period 9 in Red.** *Source: Wüthrich and Merz (2008), Table 2.2.*

Given observations $\mathcal{D}_I$, this factor estimate $\hat{f}_8^{CL}$ is equal to:

$$\hat{f}_8^{CL} = \frac{\sum_{i=1}^{10-8-1} C_{i,8+1}}{\sum_{i=1}^{10-8-1} C_{i8}} = 1.001.$$

Typically this last development factor is close to 1 and hence the cash flows paid in the final development period are minor. Using this development factor estimate, we can now estimate the remaining cumulative claim amounts in the column by multiplying the values for development year 8 with this factor.

The general math notation for the chain ladder predictions for the lower triangle $(i + j > I)$ is as follows:

$$\begin{aligned} \hat{C}_{ij}^{CL} &= C_{i,I-i} \cdot \prod_{l=I-i}^{j-1} \hat{f}_l^{CL} \\ \hat{f}_j^{CL} &= \frac{\sum_{i=1}^{I-j-1} C_{i,j+1}}{\sum_{i=1}^{I-j-1} C_{ij}}, \end{aligned}$$

where $C_{i,I-i}$ is on the last observed diagonal. It is clear that an important assumption of the chain-ladder method is that the proportional developments of claims from one development period to the next are similar for all occurrence years.

This yields the following Figure 14.14:

accident	payment delay (in years)									
year	0	1	2	3	4	5	6	7	8	9
1	5,947	9,668	10,564	10,772	10,978	11,041	11,106	11,121	11,132	11,148
2	6,347	9,593	10,316	10,468	10,536	10,573	10,625	10,637	10,648	*10,663*
3	6,269	9,245	10,092	10,355	10,508	10,573	10,627	10,636	*10,647*	*10,662*
4	5,863	8,546	9,269	9,459	9,592	9,681	9,724	*9,735*	*9,745*	*9,759*
5	5,779	8,524	9,178	9,451	9,682	9,787	*9,837*	*9,848*	*9,858*	*9,872*
6	6,185	9,013	9,586	9,831	9,936	*10,005*	*10,057*	*10,067*	*10,078*	*10,092*
7	5,600	8,493	9,057	9,282	*9,420*	*9,485*	*9,534*	*9,545*	*9,555*	*9,568*
8	5,288	7,728	8,256	*8,445*	*8,570*	*8,630*	*8,675*	*8,684*	*8,693*	*8,705*
9	5,291	7,649	*8,243*	*8,432*	*8,557*	*8,617*	*8,661*	*8,671*	*8,680*	*8,692*
10	5,676	*8,471*	*9,130*	*9,339*	*9,477*	*9,543*	*9,592*	*9,603*	*9,613*	*9,626*
$\hat{f}^{CL}$	1.493	1.078	1.023	1.015	1.007	1.005	1.001	1.001	1.001	

FIGURE 14.14: **A Run-off Triangle with Cumulative Payment Data Including Predictions in Italic** *Source: Wüthrich and Merz (2008), Table 2.2.*

The numbers in the last column show the estimates for the ultimate claim amounts. The estimate for the outstanding claim amount $\hat{\mathcal{R}}_i^{CL}$ for a particular occurrence period $i = I - J + 1, \ldots, I$ is then given by the difference between the ultimate claim amount and the cumulative amount as observed on the most recent diagonal:

$$\hat{\mathcal{R}}_i^{CL} = \hat{C}_{iJ}^{CL} - C_{i,I-i}.$$

This is the chain-ladder estimate for the reserve necessary to fulfill future liabilities with respect to claims that occurred in this particular occurrence period. These reserves per occurrence period and for the total summed over all occurrence periods are summarized in Figure 14.15.

14.3.2 Mack's Distribution-Free Chain-Ladder Model

At this stage, the traditional chain-ladder method provides a point estimator $\hat{C}_{iJ}^{CL}$ for the forecast of C_{iJ}, using the information $\mathcal{D}_I$. Since the chain-ladder

	$C_{i,I-i}$	Dev.To.Date	$\hat{C}_{i,J}^{CL}$	$\hat{\mathcal{R}}_i^{CL}$
1	11,148,123	1.000	11,148,123	0
2	10,648,192	0.999	10,663,317	15,125
3	10,635,750	0.998	10,662,007	26,257
4	9,724,069	0.996	9,758,607	34,538
5	9,786,915	0.991	9,872,216	85,301
6	9,935,752	0.984	10,092,245	156,493
7	9,282,022	0.970	9,568,142	286,120
8	8,256,212	0.948	8,705,378	449,166
9	7,648,729	0.880	8,691,971	1,043,242
10	5,675,568	0.590	9,626,383	3,950,815
totals	92,741,332.00	0.94	98,788,390.50	6,047,058.50

FIGURE 14.15: **Reserves per Occurrence Period and for Total**

method is a purely deterministic and intuitively natural algorithm to complete a run-off triangle, we are not able to determine how reliable that point estimator is or to model the variation of the future payments. To answer such questions an underlying stochastic model that reproduces the chain-ladder reserve estimates is needed.

In this section we will focus on the distribution-free chain-ladder model as an underlying stochastic model, introduced in Mack (1993). This method allows us to estimate the standard errors of the chain-ladder predictions. In the next Section 14.4, generalized linear models are used to develop a fully stochastic approach for predicting the outstanding reserve.

In Mack's approach the following conditions (without assuming a distribution) hold:

- Cumulative claims $(C_{ij})_{j=0,\dots,J}$ are independent over different occurrence periods i.
- There exist fixed constants $f_0, \dots, f_{J-1}$ and $\sigma_0^2, \dots, \sigma_{J-1}^2$ such that for all $i = 1, \dots, I$ and $j = 0, \dots, J-1$:

$$\begin{aligned} E[C_{i,j+1}|C_{i0}, \dots, C_{ij}] &= f_j \cdot C_{ij} \\ \mathrm{Var}(C_{i,j+1}|C_{ij}) &= \sigma_j^2 \cdot C_{ij}. \end{aligned}$$

  This means that the cumulative claims $(C_{ij})_{j=0,\dots,J}$ are Markov processes (in the development periods j) and hence the future only depends on the present.

Under these assumptions, the expected value of the ultimate claim amount $C_{i,J}$, given the available data in the upper triangle, is the cumulative amount on the most recent diagonal $(C_{i,I-1})$ multiplied with appropriate development factors f_j. In mathematical notation we obtain for known development factors f_j and observations $\mathcal{D}_I$:

$$E[C_{iJ}|\mathcal{D}_I] = C_{i,I-i} \prod_{j=I-i}^{J-1} f_j.$$

This is exactly what the deterministic chain-ladder method does, as explained in Section 14.3.1. In practice, the development factors are not known and need to be estimated from the data that is available in the upper triangle. In Mack's approach we obtain exactly the same expression for estimating the development factors f_j at time I as in the deterministic chain-ladder algorithm:

$$\hat{f}_j^{CL} = \frac{\sum_{j=1}^{I-j-1} C_{i,j+1}}{\sum_{i=1}^{I-j-1} C_{ij}}.$$

The predictions for the cells in the lower triangle (i.e. for cells $C_{i,j} $where $i + j > I$) are then obtained by replacing the unknown factors f_j by their corresponding estimates $\hat{f}_j^{CL}$:

$$\hat{C}_{ij}^{CL} = C_{i,I-i} \prod_{l=I-i}^{j-1} \hat{f}_l^{CL}.$$

To quantify the prediction error that comes with the chain-ladder predictions, Mack also introduced variance parameters σ_j^2. To gain insight in the estimation of these variance parameters, so-called individual development factors $f_{i,j}$ are introduced (which are specific to occurrence period i):

$$f_{i,j} = \frac{C_{i,j+1}}{C_{ij}}.$$

These individual development also describe how the cumulative amount grows from period j to period $j+1$, but they consider the ratio of only two cells (instead of taking the ratio of two sums over all available occurrence periods). Note that the development factors can be written as a weighted average of individual development factors:

$$\hat{f}_j^{CL} = \sum_{i=1}^{I-j-1} \frac{C_{ij}}{\sum_{i=1}^{I-j-1} C_{ij}} f_{i,j},$$

where the weights are equal to the cumulative claims C_{ij}.

Let us now estimate the variance parameters σ^2 by writing Mack's variance assumption in equivalent ways. First, the variance of the ratio of $C_{i,j+1}$ and $c_{i,j}$ conditional on $C_{i,0}, \ldots, C_{i,j}$ is proportional to the inverse of $C_{i,j}$:

$$\text{Var}[C_{i,j+1}/C_{ij}|C_{i0}, \ldots, C_{ij}] \ \propto \ \frac{1}{C_{ij}}.$$

This reminds us of a typical *weighted least squares* setting where the weights are the inverse of the variability of a response. Therefore, a more volatile or imprecise response variable will get less weight. The $C_{i,j}$ play the role of the weights. Using the unknown variance parameter σ_j^2 this variance assumption can be written as:

$$\mathrm{Var}[C_{i,j+1}|C_{i0},\ldots,C_{ij}] = \sigma_j^2 \cdot C_{ij},$$

The connection with weighted least squares then directly leads to an unbiased estimate for the unknown variance parameter σ_j^2 in the form of a weighted residual sum of squares:

$$\hat{\sigma}_j^2 = \frac{1}{I-j-2}\sum_{i=1}^{I-j-1} C_{ij}\left(\frac{C_{i,j+1}}{C_{ij}} - \hat{f}_j^{CL}\right)^2.$$

The weights are again equal to $C_{i,j}$ and the residuals are the differences between the ratios $C_{i,j+1}/C_{i,j}$ and the individual development factors.

We now have all ingredients required to calibrate the distribution-free chain-ladder model to the data. The next step is then to analyze the prediction uncertainty and the prediction error. Hereto we use the chain-ladder predictor where we replace the unknown development factors with their estimators:

$$\hat{C}_{iJ}^{CL} = C_{i,I-i}\prod_{l=I-i}^{J-1}\hat{f}_l^{CL}$$

We use this expression either as an estimator for the conditional expectation of the ultimate claim amount (given the observed upper triangle) or as a predictor for the ultimate claim amount as a random variable (given the observed upper triangle).

In statistics the simplest measure to analyze the uncertainty that comes with a point estimate or prediction is the *Mean Squared Error of Prediction* (*MSEP*). Here we consider a conditional *MSEP*, conditional on the data observed in the upper triangle:

$$MSEP_{C_{iJ}|\mathcal{D}_I}\left(\hat{C}_{iJ}^{CL}\right) = E\left[\left(C_{iJ} - \hat{C}_{iJ}^{CL}\right)^2|\mathcal{D}_I\right].$$

This conditional *MSEP* measures:

- the distance between the (true) ultimate claim C_{iJ} and its chain-ladder predictor $\hat{C}_{iJ}^{CL}$ at time I, and
- the total prediction uncertainty over the entire run-off of the nominal ultimate claim C_{iJ}. It does not consider time value of money, a risk margin nor any dynamics in claim development.

The *MSEP* that comes with the estimate for the ultimate cumulative claim amount is equal to the *MSEP* that measures the squared distance between the true and the estimated reserve:

$$\begin{aligned} MSEP_{\hat{\mathcal{R}}_i^I|\mathcal{D}_I}(\hat{\mathcal{R}}_i^I) &= E[(\hat{\mathcal{R}}_i^I - \mathcal{R}_i^I)^2|\mathcal{D}_I] \\ &= E[(\hat{C}_{iJ}^{CL} - C_{iJ})^2|\mathcal{D}_I] = MSEP(\hat{C}_{iJ}). \end{aligned}$$

The reason for this equivalence is the fact that the reserve is the ultimate claim amount minus the most recently observed claim amount. The latter is observed and used in both $\mathcal{R}_i^I$ and $\hat{\mathcal{R}}_i^I$.

It is interesting to decompose this *MSEP* into a component that captures *process variance* and a component that captures *parameter estimation variance*:

$$\begin{aligned} MSEP_{C_{iJ}|\mathcal{D}_I}\left(\hat{C}_{iJ}^{CL}\right) &= E\left[\left(C_{iJ} - \hat{C}_{iJ}\right)^2 |\mathcal{D}_I\right] \\ &= \mathrm{Var}(C_{iJ}|\mathcal{D}_I) + \left(E[C_{iJ}|\mathcal{D}_I] - \hat{C}_{iJ}^{CL}\right)^2 \\ &= \text{process variance} + \text{parameter estimation variance,} \end{aligned}$$

for a $\mathcal{D}_I$ measurable estimator/predictor $\hat{C}_{iJ}$. The process variance component captures the volatility or uncertainty in the random variable $C_{i,J}$ and the parameter estimation variance measures the error that arises from replacing the unknown development factors f_j with their estimated values. This result follows immediately from following equality about the variance of a shifted random variable X where the shift a is deterministic:

$$\mathrm{E}\ (X-a)^2 = \mathrm{Var}(X) + [\mathrm{E}(X) - a]^2\,.$$

Applied to the expression of the *MSEP* you treat $\hat{C}_{i,J}$ as fixed because you work conditionally on the data in the upper triangle and $\hat{C}_{i,J}$ only uses information from this upper triangle.

Mack (1993) then derived the important formula for the conditional *MSEP* in the distribution-free chain-ladder model for a single occurrence period i:

$$\widehat{MSEP}_{C_{iJ}|\mathcal{D}_I} = \left(\hat{C}_{iJ}^{CL}\right)^2 \sum_{j=I-i}^{J-1} \left[\frac{\hat{\sigma}_j^2}{(\hat{f}_j^{CL})^2}\left(\frac{1}{\hat{C}_{ij}^{CL}} + \frac{1}{\sum_{n=1}^{I-j-1} C_{nj}}\right)\right].$$

For the derivation of this popular formula, we refer to his paper. Note that it is an estimate of the *MSEP* since the unknown parameters f_j and σ_j need to be estimated as the estimation error cannot be calculated explicitly.

Mack also derived a formula for the *MSEP* for the total reserve, across all occurrence periods:

$$\begin{aligned} &\widehat{MSEP}_{\sum_{i=1}^I \hat{C}_{iJ}^{CL}}\left(\textstyle\sum_{i=1}^I \hat{C}_{iJ}^{CL}\right) \\ &\quad \textstyle\sum_{i=1}^I \widehat{MSEP}_{C_{iJ}|\mathcal{D}_I}\left(\hat{C}_{iJ}^{CL}\right) + 2\sum_{1\le i<k\le I} \hat{C}_{iJ}^{CL}\hat{C}_{kJ}^{CL} \sum_{j=I-i}^{J-1} \frac{\hat{\sigma}_j^2/\left(\hat{f}_j^{\mathrm{CL}}\right)^2}{\sum_{n=1}^{I-j-1} C_{nj}}. \end{aligned}$$

The result is the sum of the *MSEP*s per occurrence period plus a covariance term. This covariance term is added because the MSEPs for different occurrence periods i use the same parameter estimates $\hat{f}_j^{CL}$ of f_j for different accident years i.

14.3.3 R code for Chain-Ladder Predictions

We use the object `my_triangle` of type `triangle` that was created in Section 14.2.4. The distribution-free chain-ladder model of Mack (1993) is implemented in the `ChainLadder` package (Gesmann et al., 2019) (as a special form of weighted least squares) and can be applied on the data `my_triangle` to predict outstanding claim amounts and to estimate the standard error around those forecasts.

```
CL <- MackChainLadder(my_triangle)
```

The development factors are obtained as follows:

```
round(CL$f, digits = 4)
```

```
 [1] 1.4925 1.0778 1.0229 1.0148 1.0070 1.0051 1.0011 1.0010 1.0014 1.0000
```

We can also print the complete run-off triangle (including predictions).

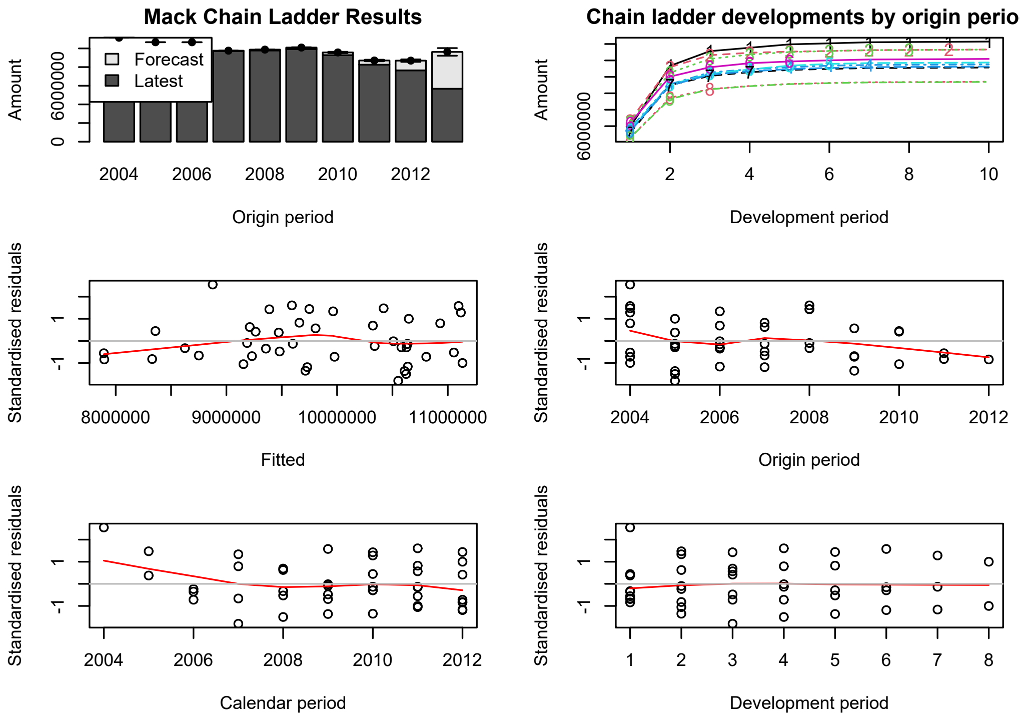
The MSEP for the total reserve across all occurrence periods is given by:

```
CL$Total.Mack.S.E^2
```

```
           9
214348469061
```

It is strongly advised to validate Mack's assumptions by checking that there are no trends in the residual plots. The last four plots that we obtain with the following command show respectively the standardized residuals versus the fitted values, the origin period, the calendar period and the development period.

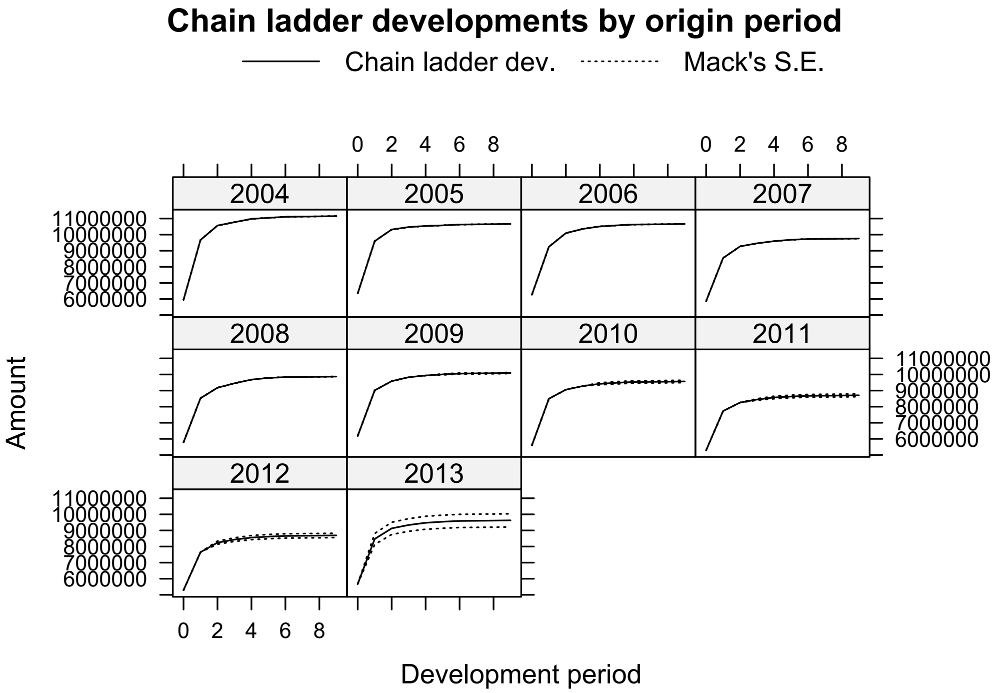
```
plot(CL)
```



The top left-hand plot is a bar-chart of the latest claims position plus IBNR and Mack's standard error by occurrence period. The top right-hand plot shows the forecasted development patterns for all occurrence periods (starting with 1 for the oldest occurrence period).

When setting the argument `lattice=TRUE` we obtain a plot of the development, including the prediction and estimated standard errors by occurrence period:

```
plot(CL, lattice = TRUE)
```



14.4 GLMs and Bootstrap for Loss Reserves

This section is being written and is not yet complete nor edited. It is here to give you a flavor of what will be in the final version.

This section covers regression models to analyze run-off triangles. When analyzing the data in a run-off triangle with a regression model, the standard toolbox for model building, estimation and prediction becomes available. Using these tools we are able to go beyond the point estimate and standard error as derived in Section 14.3. More specifically, we build a generalized linear model (GLM) for the incremental payments X_{ij} in Figure 14.6. Whereas the chain-ladder method works with cumulative data, typical GLMs assume the response variables to be independent and therefore work with incremental run-off triangles.

14.4.1 Model Specification

Let X_{ij} denote the incremental payment in cell (i,j) of the run-off triangle. We assume the X_{ij}s to be independent with a density $f(x_{ij};\theta_{ij},\phi)$ from the exponential family of distributions. We identify

- $\mu_{ij}=E[X_{ij}]$ the expected value of cell X_{ij}
- ϕ the dispersion parameter and $\text{Var}[X_{ij}]=\phi\cdot V(\mu_{ij})$, where $V(.)$ is the variance function
- η_{ij} the linear predictor such that $\eta_{ij}=g(\mu_{ij})$ with g the link function.

Distributions from the exponential family and their default link functions are listed on http://stat.ethz.ch/R-manual/R-patched/library/stats/html/family.html. We now discuss three specific GLMs widely used for loss reserving.

First, the Poisson regression model was introduced in Section 11.2. In this model, we assume that X_{ij} has a Poisson distribution with parameter

$$\mu_{ij}=\pi_i\cdot\gamma_j,$$

a cross-classified structure that captures a multiplicative effect of the occurrence year i and the development period j. The proposed model structure is not identifiable without an additional constraint on the parameters, e.g. $\sum_{j=0}^{J}\gamma_j=1$. This constraint gives an explicit interpretation to π_i (with $i=1,\ldots,I$) as the exposure or volume measure for occurrence year i and γ_j as the fraction of the total volume paid out with delay j. However, when calibrating GLMs in `R` alternative constraints such as $\pi_1=1$ or $\gamma_1=1$, or a reparametrization where $\mu_{ij}=\exp\left(\mu+\alpha_i+\beta_j\right)$ are easier to implement. We continue with the latter specification, including $\alpha_1=\beta_0=0$, the so-called corner constraints. This GLM treats the occurrence year and the payment delay as factor variables and fits a parameter per level, next to an intercept μ. The corner constraints put the effect of the first level of a factor variable equal to zero. The Poisson assumption is particularly useful for a run-off triangle with numbers of reported claims, often used in the estimation of the number of IBNR claims (see Section 14.2).

Second, an interesting modification of the basic Poisson regression model is the **over-dispersed Poisson** regression model where Z_{ij} has a Poisson distribution with parameter μ_{ij}/ϕ and

$$\begin{aligned}X_{ij}&\sim\phi\cdot Z_{ij}\\ \mu_{ij}&=\exp\left(\mu+\alpha_i+\beta_j\right).\end{aligned}$$

Consequently, X_{ij} has the same specification for the mean as in the basic Poisson regression model, but now

$$\text{Var}[X_{ij}]=\phi^2\cdot\text{Var}[Z_{ij}]=\phi\cdot\exp\left(\mu+\alpha_i+\beta_j\right).$$

This construction allows for under (when $\phi < 1$) and over-dispersion (with $\phi > 1$). Because X_{ij} no longer follows a well-known distribution, this approach is referred to as quasi-likelihood. It is particularly useful to model a run-off triangle with incremental payments, as these typically reveal over-dispersion.

Third, the **gamma** regression model is relevant to model a run-off triangle with claim payments. Recall from Section 4.2.1 (see also the Appendix Chapter 20) that the gamma distribution has shape parameter α and scale parameter θ. From these, we reparameterize and define a new parameter $\mu = \alpha \cdot \theta$ while retaining the scale parameter θ. Further, assume that X_{ij} has a gamma distribution and allow μ to vary by ij such that

$$\mu_{ij} = \exp\left(\mu + \alpha_i + \beta_j\right).$$

14.4.2 Model Estimation and Prediction

We now estimate the regression parameters μ, α_i and β_j in the proposed GLMs. In `R` the `glm` function is readily available to estimate these parameters via maximum likelihood estimation (mle) or quasi-likelihood estimation (in the case of the over-dispersed Poisson). Having the parameter estimates $\hat{\mu}$, $\hat{\alpha}_i$ and $\hat{\beta}_j$ available, a point estimate for each cell in the upper triangle follows

$$\hat{X}_{ij} = E[\hat{X}_{ij}] = \exp\left(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j\right), \text{ with } i+j \leq I.$$

Similarly, a cell in the lower triangle will be predicted as

$$\hat{X}_{ij} = E[\hat{X}_{ij}] = \exp\left(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j\right), \text{ with } i+j > I.$$

Point estimates for outstanding reserves (per occurrence year i or the total reserve) then follow by summing the cell-specific estimates. By combining the observations in the upper triangle with their point estimates, we can construct properly defined residuals and use these for residual inspection.

14.4.3 Bootstrap

14.5 Further Resources and Contributors

Contributors

- **Katrien Antonio**, KU Leuven and University of Amsterdam, **Jan Beirlant**, KU Leuven, and **Tim Veerdonck**, University of Antwerp, are the principal authors of the initial version of this chapter. Email: katrien.antonio@kuleuven.be for chapter comments and suggested improvements.

Further Readings and References

As displayed in Figure 14.1, similar timelines and visualizations are discussed (among others) in Wüthrich and Merz (2008), Antonio and Plat (2014) and Wüthrich and Merz (2015).

Over time actuaries started to think about possible underlying models and we mention some important contributions:

- Kremer (1982): two-way ANOVA
- Kremer (1984), Mack (1991): Poisson model
- Mack (1993): distribution-free chain-ladder model
- Renshaw (1989); Renshaw and Verrall (1998): over-dispersed Poisson model
- Gisler (2006); Gisler and Wüthrich (2008); Bühlmann et al. (2009): Bayesian chain-ladder model.

The various stochastic models proposed in actuarial literature rely on different assumptions and have different model properties, but have in common that they provide exactly the chain-ladder reserve estimates. For more information we also refer to Mack and Venter (2000) and to the lively discussion that was published in *ASTIN Bulletin: Journal of the International Actuarial Association* in 2006 (Venter, 2006).

To read more about exponential families and generalized linear models, see, for example, McCullagh and Nelder (1989) and Wüthrich and Merz (2008). We refer to (Kremer, 1982), (Renshaw and Verrall, 1998) and (England and Verrall, 2002), and the overviews in (Taylor, 2000), (Wüthrich and Merz, 2008) and (Wüthrich and Merz, 2015) for more details on the discussed GLMs. XXX presents alternative distributional assumptions and specifications of the linear predictor.

15

Experience Rating using Bonus-Malus

Chapter Preview. This chapter introduces bonus-malus system used in motor insurance ratemaking. In particular, the chapter discusses the features of bonus-malus system and studies its modelling and properties via basic statistical techniques. Section 15.1 introduces the use of bonus-malus system as an experience rating scheme, followed by Section 15.2 which describes its practical implementation in several countries. Section 15.3 covers its modelling setup by a discrete time Markov Chain. Next, Section 15.4 studies a number of simple relevant properties associated with the stationary distribution of bonus-malus system. Section 15.5 focuses on the determination of *a posteriori* premium rating to complement *a priori* ratemaking.

15.1 Introduction

In this section, you learn how to:

- Use bonus-malus system as an experience rating scheme.
- Compare bonus-malus system with risk classification (Chapter 8) and credibility premium (Chapter 9).

Bonus-malus system, which is used interchangeably as “no-fault discount”, “merit rating”, “experience rating” or “no-claim discount” in different countries, is based on penalizing insureds who are responsible for one or more claims by a premium surcharge (malus), and rewarding insureds with a premium discount (bonus) if they do not have any claims. Insurers use bonus-malus system (*BMS*) for two main purposes: to encourage drivers to drive more carefully in a policy year without any claims, and to ensure insureds to pay premiums proportional to their risks based on their claims experience via an experience rating mechanism.

BMS is an experience rating system commonly used in motor insurance. It represents an attempt to categorize insureds into homogeneous groups who pay premiums based on their claims experience. Depending on the rules in the scheme, new policyholders may be required to pay full premium initially, and obtain discounts in the future years as a result of claim-free years. *BMS* rewards policyholders for not making any claims during a policy year. In other words, it grants a bonus to a careful driver. This bonus principle may affect policy holders' decisions whether to claim or not to claim, especially when involving accidents with slight damages, which is known as the 'hunger for bonus' phenomenon. The 'hunger for bonus' under a *BMS* may reduce insurers' claim costs, and may be able to offset the expected decrease in premium income.

In motor insurance, *BMS* is a form of *a posteriori* rating to complement the use of *a priori* risk classification described in Chapter 11. The *a priori* risk classification divides portfolio of drivers into a number of homogeneous risk classes based on rating factors, such that policyholders in the same risk class pay the same premium. The ideal *a posteriori* mechanism is the credibility premium developed in Chapter 12, whereby premiums are derived on an individual basis for each policyholder by incorporating both the *a priori* and *a posteriori* information. However, such individual premium determination is overly complex from a commercial standpoint for practical implementations by motor insurers. For this reason, *BMS* is the preferred solution and it consists of three elements: bonus-malus classes, transition rules, and premium levels (also known as premium relativities). The advantage of using *BMS* is that the bonus-malus classes and the transition rules are pre-specified in advance by insurers. The bonus-malus classes and transition rules will be discussed in the next section.

15.2 *BMS* in Several Countries

In this section, you learn how to:

- Use *BMS* in Malaysia and other countries.
- Determine a transition rule.

Many countries around the world have adopted some form of *BMS* in their automobile insurance. The specifics of these systems can vary from country to country, but the general idea is to reward safe driving behavior by reducing premiums for policyholders who do not make claims, and increasing premiums for those who do. Some of the countries that have implemented or adopted the *BMS* are France, Germany, Italy, Spain and United Kingdom from Europe, and Malaysia, Hong Kong, Taiwan, Singapore and Korea from Asia. Please refer to Lemaire and Hongmin (1994), Lemaire (1998) and Park et. al (2010) for implementation of other *BMS* around the world.

15.2.1 *BMS* in Malaysia

Before the liberalization of Motor Tariff on 1st July 2017, the rating of motor insurance in Malaysia was governed by the Motor Tariff. Under the tariff, the rate charged should not be lower than the rates specified under the classes of risks. The basic risk classes considered were scope of insurance, cubic capacity of vehicle and estimated value of vehicle (or sum insured, whichever is lower). The final premium to be paid is adjusted by the policyholder's claim experience, or equivalently, his bonus-malus entitlement.

Effective on 1st July 2017, the premium rates for motor insurance are liberalized, or de-tariffed. The pricing of premium is now determined by individual insurers and takaful operators, and the consumers are able to enjoy a wider choice of motor insurance products at competitive prices. Since tariff liberalization encourages innovation and competition among insurers and takaful operators, the premiums are based on broader risk factors other than the risk classes specified in the Motor Tariff. Other rating factors may be defined in the risk profile of an insured, such as age of vehicle, age of driver, safety and security features of vehicle, geographical location of vehicle and traffic offences of driver. However, the bonus-malus entitlement from the Motor Tariff remains 'unchanged' and continue to exist, and is 'transferable' from one insurer, or from one takaful operator, to another.

The discounts in the Malaysian *BMS* are divided into six classes, starting from the initial class of 0% discount, followed by classes of 25%, 30%, 38.3%, 45% and 55% discounts. A claim-free policy year indicates that a policyholder is entitled to move one-step forward to the next discount class, such as from a 0% discount to a 25% discount in the renewal year. If a policyholder is already at the highest class, which is at a 55% discount, a claim-free policy year indicates that the policyholder remains in the same class. On the other hand, if one or more claims are made within the policy year, the discount will be forfeited and the policyholder has to start at 0% discount in the renewal year. This set of transition rules can also be summarized as a rule of -1/Top, that is, a class of

TABLE 15.1: **Transition table for bonus-malus classes (Malaysia)**

Classes	Discounts (%)
0	0.00
1	25.00
2	30.00
3	38.33
4	45.00
5 (and above)	55.00

bonus for a claim-free year, and moving to the highest class after having one or more claims. For an illustration purpose, Table 15.1 and Figure 15.1 show the classes and the transition diagram for the Malaysian *BMS*.

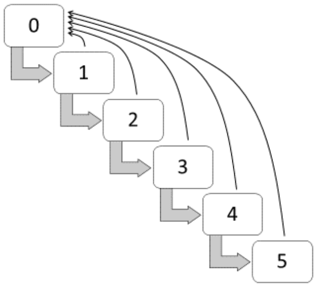


FIGURE 15.1: **Transition diagram for bonus-malus classes (Malaysia)**

15.2.2 *BMS* in Other Countries

The *BMS* in Brazil are subdivided into seven classes, with the following premium levels (Lemaire and Zi, 1994): 100, 90, 85, 80, 75, 70, and 65. These premium levels are entitled to the following discounts: 0%, 10%, 15%, 20%, 25%, 30% and 35%. New policyholders have to start at 0% discount, or at premium level 100. A claim-free policy year indicates that a policyholder can move forward at a one-class discount. If one or more claims incurred within the policy year, the policyholder has to move one-class backward for each claim. Table 15.2 and Figure 15.2 show the classes and the transition diagram for the *BMS* in Brazil. This set of transition rules can also be summarized as a rule of -1/+1, that is, a class of bonus for a claim-free policy year, and a class of malus for each claim reported.

TABLE 15.2: **Transition table for bonus-malus classes (Brazil)**

Classes	Discounts (%)
0	0
1	10
2	15
3	20
4	25
5	30
6 (and above)	35

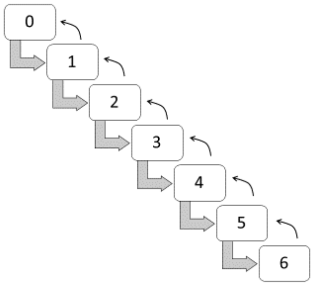


FIGURE 15.2: **Transition diagram for bonus-malus classes (Brazil)**

The *BMS* in Switzerland are subdivided into twenty-two classes, with the following premium levels: 270, 250, 230, 215, 200, 185, 170, 155, 140, 130, 120, 110, 100, 90, 80, 75, 70, 65, 60, 55, 50 and 45 (Lemaire and Zi, 1994). The premium levels 270, 250, 230, 215, 200, 185, 170, 155, 140, 130, 120 and 110 are the premiums with the following loadings (malus): 170%, 150%, 130%, 115%, 100%, 85%, 70%, 55%, 40%, 30%, 20%, and 10%. On the other hand, the premium levels 100, 90, 80, 75, 70, 65, 60, 55, 50 and 45 are the premiums with the following discounts (bonus): 0%, 10%, 20%, 25%, 30%, 35%, 40%, 45%, 50% and 55%. New policyholders have to start at 0% discount, or at premium level 100, and a claim-free policy year indicates that a policyholder can move one-class forward. If one or more claims incurred within the policy year, the policyholder has to move four-classes backward for each claim. Table 15.3 and Figure 15.3 respectively show the classes and the transition diagram for the *BMS* in Switzerland. This set of transition rule can be summarized as a rule of -1/+4. It should be noted that the entry level is at class 12, which is at premium level 100 (or 0% discount).

Table 15.3. Bonus-malus classes (Switzerland)

Classes	Loadings (%)	Classes	Discounts (%)
0	170	12	0
1	150	13	10
2	130	14	20
3	115	15	25
4	100	16	30
5	85	17	35
6	70	18	40
7	55	19	45
8	40	20	50
9	30	21	55
10	20		
11	10		

15.3 *BMS* and Markov Chain Model

In this section, you learn how to:

- Represent bonus-malus classes using transition probabilities.
- Use year to year transition matrix.

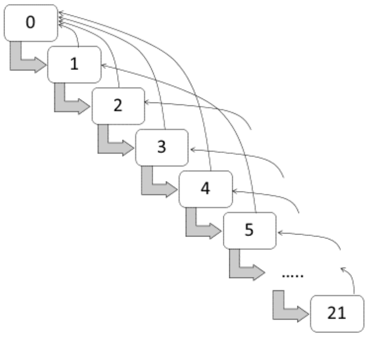


FIGURE 15.3: **Transition diagram for bonus-malus classes (Switzerland)**

A *BMS* can be represented by a discrete time Markov chain. A stochastic process is said to possess the *Markov property* if the evolution of the process in the future depends only on the present state but not the past. A discrete time Markov Chain is a Markov process with discrete state space.

15.3.1 Transition Probability

A Markov Chain is determined by its transition probabilities. The *transition probability* from state i (at time n) to state j (at time $n+1$) is called a one-step transition probability, and is denoted by $p_{ij}(n, n+1) = Pr(X_{n+1} = j|X_n = i)$, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, k$. For general transition from time m to time n, for $m < n$, by conditioning on X_o for $m \leq o \leq n$, we have the Chapman-Kolmogorov equation of

$$p_{ij}(m, n) = \sum_{l \in S} p_{il}(m, o) p_{lj}(o, n).$$

A time-homogeneous Markov Chain satisfies the property of $p_{ij}(n, n+t) = p_{ij}^{(t)}$ for all n. For instance, we have $p_{ij}(n, n+1) = p_{ij}^{(1)} \equiv p_{ij}$. In this case, the Chapman-Kolmogorov equation can be written as

$$p_{ij}(0, m+n) = \sum_{l \in S} p_{il}(0, m) p_{lj}(m, m+n) = \sum_{l \in S} p_{il}^{(m)} p_{lj}^{(n)}.$$

In the context of *BMS*, the transition of the bonus-malus classes is governed by the transition probability in a given policy year. The transition of the bonus-malus classes is also a time-homogeneous Markov Chain since the set of transition rules is fixed and independent of time. We can represent the one-step transition probabilities by a $k \times k$ *transition matrix* $\mathbf{P} = (p_{ij})$ that corresponds to bonus-malus classes $0, 1, 2, \ldots, k-1$.

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0k-1} \\ p_{10} & p_{11} & \cdots & p_{1k-1} \\ \vdots & \ddots & & \vdots \\ p_{k-10} & p_{k-11} & \cdots & p_{k-1k-1} \end{bmatrix}$$

Here, its (i, j)-th element is the transition probability from state i to state j. In other words, each row of the transition matrix represents the transition of flowing out of state, whereas each column represents the transition of flowing into the state. The summation of transition probabilities of flowing out of state must equal to 1, or each row of the matrix must sum to 1, i.e. $\sum_j p_{ij} = 1$. All probabilities must be non-negative, i.e. $p_{ij} \geq 0$.

15.3.2 Some Applications

Consider the Malaysian *BMS*. Let $\{X_t : t = 0, 1, 2, \ldots\}$ be the bonus-malus class occupied by a policyholder at time t with state space $S = \{0, 1, 2, 3, 4, 5\}$. Therefore, the transition probability in a no-claim policy year is equal to the probability of transition from state i to state $i + 1$, i.e. p_{ii+1}. If an insured has one or more claims within the policy year, the probability of transitioning back to state 0 is represented by $p_{i0} = 1 - p_{ii+1}$. Hence, the Malaysian *BMS* can be represented by the following 6×6 transition matrix:

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{01} & 0 & 0 & 0 & 0 \\ p_{10} & 0 & p_{12} & 0 & 0 & 0 \\ p_{20} & 0 & 0 & p_{23} & 0 & 0 \\ p_{30} & 0 & 0 & 0 & p_{34} & 0 \\ p_{40} & 0 & 0 & 0 & 0 & p_{45} \\ p_{50} & 0 & 0 & 0 & 0 & p_{55} \end{bmatrix} = \begin{bmatrix} 1-p_{01} & p_{01} & 0 & 0 & 0 & 0 \\ 1-p_{12} & 0 & p_{12} & 0 & 0 & 0 \\ 1-p_{23} & 0 & 0 & p_{23} & 0 & 0 \\ 1-p_{34} & 0 & 0 & 0 & p_{34} & 0 \\ 1-p_{45} & 0 & 0 & 0 & 0 & p_{45} \\ 1-p_{55} & 0 & 0 & 0 & 0 & p_{55} \end{bmatrix}$$

Example 15.3.1. Provide the transition matrix for the *BMS* in Brazil.

Example Solution. Based on the bonus-malus classes and the transition diagram shown in Figure 15.2, the probability of a no-claim policy year is equal to the probability of moving one-class forward, whereas the probability of having one or more claims within the policy year is equal to the probability of moving one-class backward for each claim. Therefore, each row can contain two or more transition probabilities; one probability for advancing to the next state, and one or more probabilities for moving one-class backward. The transition matrix is:

$$\mathbf{P} = \begin{bmatrix} 1-p_{01} & p_{01} & 0 & 0 & 0 & 0 & 0 \\ 1-p_{12} & 0 & p_{12} & 0 & 0 & 0 & 0 \\ 1-\sum_j p_{2j} & p_{21} & 0 & p_{23} & 0 & 0 & 0 \\ 1-\sum_j p_{3j} & p_{31} & p_{32} & 0 & p_{34} & 0 & 0 \\ 1-\sum_j p_{4j} & p_{41} & p_{42} & p_{43} & 0 & p_{45} & 0 \\ 1-\sum_j p_{5j} & p_{51} & p_{52} & p_{53} & p_{54} & 0 & p_{56} \\ 1-\sum_j p_{6j} & p_{61} & p_{62} & p_{63} & p_{64} & p_{65} & p_{66} \end{bmatrix}$$

Example 15.3.2. Provide the transition matrix for the *BMS* in Switzerland.

Example Solution. From Table 15.3 and Figure 15.3, the probability of a no-claim policy year is equal to the probability of moving one-class forward, whereas the probability of having one or more claims within the policy year is equal to the probability of moving four-classes backward for each claim. The transition matrix is:

$$\mathbf{P} = \begin{bmatrix} 1-p_{01} & p_{01} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 1-p_{12} & 0 & p_{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 1-p_{23} & 0 & 0 & p_{23} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 1-p_{34} & 0 & 0 & 0 & p_{34} & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 1-p_{45} & 0 & 0 & 0 & 0 & p_{45} & 0 & 0 & 0 & 0 & 0 & \cdots \\ 1-\sum\limits_j p_{5j} & p_{51} & 0 & 0 & 0 & 0 & p_{56} & 0 & 0 & 0 & 0 & \cdots \\ 1-\sum\limits_j p_{6j} & 0 & p_{62} & 0 & 0 & 0 & 0 & p_{67} & 0 & 0 & 0 & \cdots \\ 1-\sum\limits_j p_{7j} & 0 & 0 & p_{73} & 0 & 0 & 0 & 0 & p_{78} & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\ 1-\sum\limits_j p_{19,j} & 0 & 0 & p_{19,3} & 0 & 0 & 0 & p_{19,7} & 0 & 0 & 0 & \cdots \\ 1-\sum\limits_j p_{20,j} & 0 & 0 & 0 & p_{20,4} & 0 & 0 & 0 & p_{20,8} & 0 & 0 & \cdots \\ 1-\sum\limits_j p_{21,j} & p_{21,1} & 0 & 0 & 0 & p_{21,5} & 0 & 0 & 0 & p_{21,9} & 0 & \cdots \end{bmatrix}$$

15.4 *BMS* and Stationary Distribution

In this section, you learn how to:

- Calculate stationary probabilities.
- Observe a premium evolution.
- Measure the convergence rate.

15.4.1 Stationary Distribution

A stationary probability, which is also known as a steady-state probability, is a probability of being in a state at equilibrium or in the long run. In a Markov chain, each state has a corresponding stationary probability. These probabilities do not change over time once the Markov chain has achieved its steady state. Stationary probability is important for understanding the long-term behavior, equilibrium states, and predictive aspects of a system (such as *BMS*) which are modeled using a Markov chain. In this section, we introduce a stationary probability because it offers some practical applications of the *BMS*.

Stationary probabilities can be represented by a row vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)$ with the following properties:

$$\begin{aligned}
&0 \leq \pi_j \leq 1, \\
&\sum_j \pi_j = 1, \\
&\pi_j = \sum_i \pi_i p_{ij}.
\end{aligned}$$

The last equation can be written in terms of matrix and vector, which is $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$, where $\boldsymbol{\pi}$ is the stationary probability vector and $\mathbf{P}$ is the transition matrix. The first two conditions are necessary for the probability distribution, whereas the last property indicates that $\boldsymbol{\pi}$ is invariant (i.e. unchanged) by the one-step transition matrix. In other words, once the Markov Chain has reached stationary state, its probability distribution will stay stationary over time. Mathematically, the stationary vector $\boldsymbol{\pi}$ can also be obtained by finding the left eigenvector of the one-step transition matrix.

Example 15.4.1. Find the stationary distribution for the *BMS* in Malaysia assuming that the probability of a no-claim policy year for all bonus-malus classes are equal, and it is equivalent to p_0.

Example Solution. The transition matrix can be re-written as:

$$\mathbf{P} = \begin{bmatrix} 1-p_0 & p_0 & 0 & 0 & 0 & 0 \\ 1-p_0 & 0 & p_0 & 0 & 0 & 0 \\ 1-p_0 & 0 & 0 & p_0 & 0 & 0 \\ 1-p_0 & 0 & 0 & 0 & p_0 & 0 \\ 1-p_0 & 0 & 0 & 0 & 0 & p_0 \\ 1-p_0 & 0 & 0 & 0 & 0 & p_0 \end{bmatrix}$$

The stationary distribution can be calculated using $\pi_j = \sum_i \pi_i p_{ij}$ or $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$, where $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_5.)$ The solutions are:

$$\begin{aligned}
&\pi_0 = \sum_i \pi_i p_{i0} = (1-p_0)\sum_i \pi_i = 1-p_0 \\
&\pi_1 = \sum_i \pi_i p_{i1} = \pi_0 p_{01} = (1-p_0)p_0 \\
&\pi_2 = \sum_i \pi_i p_{i2} = \pi_1 p_{12} = (1-p_0){p_0}^2 \\
&\pi_3 = \sum_i \pi_i p_{i3} = \pi_2 p_{23} = (1-p_0){p_0}^3 \\
&\pi_4 = \sum_i \pi_i p_{i4} = \pi_3 p_{34} = (1-p_0){p_0}^4 \\
&\pi_5 = \sum_i \pi_i p_{i5} = \pi_4 p_{45} + \pi_5 p_{55} = (1-p_0){p_0}^5 + \pi_5 p_0 \\
&\therefore \pi_5 = \tfrac{(1-p_0){p_0}^5}{(1-p_0)} = {p_0}^5
\end{aligned}$$

The stationary distribution shown in Example 15.4.1 represents the asymptotic distribution of the *BMS*, or the distribution in the long run. As an example, assuming that the probability of a no-claim policy year is $p_0 = 0.90$, the *stationary probabilities* are:

$$\begin{aligned}
&\pi_0 = 1-p_0 = 0.1000 \\
&\pi_1 = (1-p_0)p_0 = 0.0900 \\
&\pi_2 = (1-p_0){p_0}^2 = 0.0810 \\
&\pi_3 = (1-p_0){p_0}^3 = 0.0729 \\
&\pi_4 = (1-p_0){p_0}^4 = 0.0656 \\
&\pi_5 = {p_0}^5 = 0.5905
\end{aligned}$$

In other words, $\pi_0 = 0.10$ indicates that 10% of insureds will eventually belong to class 0, $\pi_1 = 0.09$ indicates that 9% of insureds will eventually belong to class 1, and so forth, until $\pi_5 = 0.59$, which indicates that 59% of insureds will eventually belong to class 5.

15.4.2 R Code for a Stationary Distribution

We can use the left eigenvector of a transition matrix to calculate a stationary distribution. The following R code can be used to calculate a stationary distribution in two stages:

1. Create a Transition Matrix
2. Find a stationary distribution using left eigenvector.

1. Create a Transition Matrix

```
#create transition matrix for Malaysian data
P=matrix(data=0,nrow=6,ncol=6)
P[1,2]=P[2,3]=P[3,4]=P[4,5]=P[5,6]=P[6,6]=0.9
P[,1]=0.1
P

#output
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]  0.1  0.9  0.0  0.0  0.0  0.0
[2,]  0.1  0.0  0.9  0.0  0.0  0.0
[3,]  0.1  0.0  0.0  0.9  0.0  0.0
[4,]  0.1  0.0  0.0  0.0  0.9  0.0
[5,]  0.1  0.0  0.0  0.0  0.0  0.9
[6,]  0.1  0.0  0.0  0.0  0.0  0.9
```

2. Find a stationary distribution using left eigenvector

```
#for left eigenvector, use eigenvector of transpose of transition matrix
#then divide entry of 1st column by sum of 1st column so that all entries sum to 1
#provide stationary distribution in terms of a row vector (use transpose)
#provide stationary distribution with numeric/real values (use function Re())

Re(t(eigen(t(P))$vectors[,1]/sum(eigen(t(P))$vectors[,1])))

#output
     [,1] [,2]  [,3]   [,4]    [,5]    [,6]
[1,]  0.1 0.09 0.081 0.0729 0.06561 0.59049
```

Example 15.4.2. Consider the *BMS* in Brazil where the transition rule is -1/+1. Let the probability of a no-claim policy year (probability of one-class forward) equal to p_0, the probability of one or more claims in a policy year (probability of one-class backward) equal to p_1, the probability of one or more claims in the next policy year (probability of two-classes backward) equal to p_2, and so on and so forth. Find the stationary distribution assuming that p_k

is distributed as Poisson with probability

$$p_k = \frac{e^{-0.1}(0.1)^k}{k!}, k = 0, 1, 2, \ldots$$

Example Solution. The transition matrix for the BMS in Brazil can be written as:

$$\mathbf{P} = \begin{bmatrix} 1-p_0 & p_0 & 0 & 0 & 0 & 0 & 0 \\ 1-p_0 & 0 & p_0 & 0 & 0 & 0 & 0 \\ 1-\sum_i p_i & p_1 & 0 & p_0 & 0 & 0 & 0 \\ 1-\sum_i p_i & p_2 & p_1 & 0 & p_0 & 0 & 0 \\ 1-\sum_i p_i & p_3 & p_2 & p_1 & 0 & p_0 & 0 \\ 1-\sum_i p_i & p_4 & p_3 & p_2 & p_1 & 0 & p_0 \\ 1-\sum_i p_i & p_5 & p_4 & p_3 & p_2 & p_1 & p_0 \end{bmatrix}$$

$$= \begin{bmatrix} 0.0952 & 0.9048 & 0 & 0 & 0 & 0 & 0 \\ 0.0952 & 0 & 0.9048 & 0 & 0 & 0 & 0 \\ 0.0047 & 0.0905 & 0 & 0.9048 & 0 & 0 & 0 \\ 0.0002 & 0.0045 & 0.0905 & 0 & 0.9048 & 0 & 0 \\ 0.0000 & 0.0002 & 0.0045 & 0.0905 & 0 & 0.9048 & 0 \\ 0.0000 & 0.0000 & 0.0002 & 0.0045 & 0.0905 & 0 & 0.9048 \\ 0.0000 & 0.0000 & 0.0000 & 0.0002 & 0.0045 & 0.0905 & 0.9048 \end{bmatrix}$$

Using the earlier ‘R‘ codes on this transition matrix $\mathbf{P}$, the stationary probabilities are obtained as:

$$\begin{bmatrix} \pi_0 \\ \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \\ \pi_5 \\ \pi_6 \end{bmatrix} = \begin{bmatrix} 0.0000 \\ 0.0000 \\ 0.0003 \\ 0.0022 \\ 0.0145 \\ 0.0936 \\ 0.8894 \end{bmatrix}.$$

The probabilities indicate that 89% of insureds will eventually belong to class 6, 9% of insureds will eventually belong to class 5, and 1.5% of insureds will eventually belong to class 4. Other classes would have less than 1% of insureds in the long run.

Example 15.4.3. Using the results from Example 15.4.2, find the mean premium under the steady state condition assuming that the premium prior to implementing the bonus-malus discount is 1000.

Example Solution. After using the discount, the premium for each class is

$$= \ (1000) \times (1 \text{ - } \% \text{ of discount})$$

Using stationary probabilities from Example 15.4.2, the mean premium under steady state condition is:

$$\begin{aligned}
&= \sum_j (\text{proportion in class } j \text{ in the long run}) \times (\$1000) \times (1 \text{ - } \% \text{ of discount}) \\
&= \$1000 \times [\pi_0(1) + \pi_1(1-0.1) + \pi_2(1-0.15) + \ldots + \pi_6(1-0.35)] \\
&= \$1000 \times [0 + 0 + (0.0003)(0.85) + \ldots + (0.8894)(0.65)] \\
&= \$656.5
\end{aligned}$$

The results indicate that the final premium reduce from 1000 to 656.5 in the long run under stationary condition if the discount is considered. From a financial standpoint, this implies that the collected premium is insufficient to cover the expected claim cost of 1000. This result is not surprising because none of the *BMS* classes in Brazil impose a malus loading for the policyholders. More importantly, it indicates that the *BMS* will only be financially balanced if there are both bonus and malus classes and the premium levels are re-calculated such that the expected premium under the stationary distribution equals to 1000.

15.4.3 Premium Evolution

We may be interested to find out the evolution of the mean premium after n years (or n steps). Under the *BMS*, the n-step transition probability, $p_{ij}^{(n)} = \Pr(X_n = j|X_0 = i)$, can be used to calculate the evolution of the mean premium. The probability $p_{ij}^{(n)}$ can be obtained as the (i,j)-th element of the n-th power of transition matrix $\mathbf{P}$, that is, $\mathbf{P}^n$.

Example 15.4.4. Consider the *BMS* in Malaysia where the transition rule is -1/Top. Let the probability of a no-claim policy year equal to p_0 (probability of one-class forward) and the probability of one or more claims in the policy year equal to $1 - p_0$ (probability of moving back to class 1). Observe the mean premiums in 20 years assuming that p_k is distributed as Poisson with probability $p_k = \frac{e^{-0.1}(0.1)^k}{k!}$, $k = 0$. Let the mean premium prior to implementing the bonus-malus discount equals to 1000.

Example Solution. Under the Malaysian BMS, the transition matrix in the first year is:

$$\mathbf{P^{(1)}} = \begin{bmatrix} 0.0952 & 0.9048 & 0 & 0 & 0 & 0 \\ 0.0952 & 0 & 0.9048 & 0 & 0 & 0 \\ 0.0952 & 0 & 0 & 0.9048 & 0 & 0 \\ 0.0952 & 0 & 0 & 0 & 0.9048 & 0 \\ 0.0952 & 0 & 0 & 0 & 0 & 0.9048 \\ 0.0952 & 0 & 0 & 0 & 0 & 0.9048 \end{bmatrix}$$

The mean premium in the first year, after implementing the discount, is:

$$\begin{aligned} &= \sum_j (\$1000) \times (\text{average proportion in class } j) \times (1 \text{ - \% of discount}) \\ &= \$1000 \times \left[\frac{\sum_i p_{i0}}{6}(1) + \frac{\sum_i p_{i1}}{6}(1 - 0.25) + \ldots + \frac{\sum_i p_{i5}}{6}(1 - 0.55) \right] \\ &= \$1000 \times [0.0952(1) + 0.1508(0.75) + \cdots + 0.3016(0.45)] \\ &= \$625.5 \end{aligned}$$

Using similar steps, the mean premium in the n-th year for $n = 2, ..., 20$ can be observed. From 'R', the mean premiums in 20 years are:

625.5, 598.7, 580.6, 570.6, 565.8, 565.8, 565.8, 565.8, 565.8, 565.8,

565.8, 565.8, 565.8, 565.8, 565.8, 565.8, 565.8, 565.8, 565.8, 565.8.

15.4.4 R Code for Premium Evolution

The following R code can be used to find the premium in the n-th year and the premiums in 20 years under the *BMS* in Malaysia (to find the solution in Example 15.4.4). A function for transition matrix is created so that it can be used in the later sections (Section 15.5).

1. Create a function for transition matrix
2. Create a function for the n-th power of a square matrix
3. Create a function for premium in n-th year
4. Provide premiums for 20 years,

1. Create a function for transition matrix

```
#create transition probability using function of \lambda
TP=function(\lambda)
{ P=matrix(data=0,nrow=6,ncol=6)
  P[1,2]=P[2,3]=P[3,4]=P[4,5]=P[5,6]=P[6,6]=exp(-\lambda)
  P[,1]=1-exp(-\lambda)
  P}
```

```
TP(0.1)

#output
           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 0.09516258 0.9048374 0.0000000 0.0000000 0.0000000 0.0000000
[2,] 0.09516258 0.0000000 0.9048374 0.0000000 0.0000000 0.0000000
[3,] 0.09516258 0.0000000 0.0000000 0.9048374 0.0000000 0.0000000
[4,] 0.09516258 0.0000000 0.0000000 0.0000000 0.9048374 0.0000000
[5,] 0.09516258 0.0000000 0.0000000 0.0000000 0.0000000 0.9048374
[6,] 0.09516258 0.0000000 0.0000000 0.0000000 0.0000000 0.9048374
```

2. Create a function for the n-th power of a square matrix

```
#create function for nth power of square matrix
powA = function(n,\lambda)
{ if (n==1)  return (TP(\lambda))
  if (n==2)  return (TP(\lambda)%*%TP(\lambda))
  if (n>2) return (TP(\lambda)%*%powA(n-1,\lambda))}
powA(3,0.1)

#output
           [,1]       [,2]       [,3]      [,4]      [,5]      [,6]
[1,] 0.09516258 0.08610666 0.07791253 0.7408182 0.0000000 0.0000000
[2,] 0.09516258 0.08610666 0.07791253 0.0000000 0.7408182 0.0000000
[3,] 0.09516258 0.08610666 0.07791253 0.0000000 0.0000000 0.7408182
[4,] 0.09516258 0.08610666 0.07791253 0.0000000 0.0000000 0.7408182
[5,] 0.09516258 0.08610666 0.07791253 0.0000000 0.0000000 0.7408182
[6,] 0.09516258 0.08610666 0.07791253 0.0000000 0.0000000 0.7408182
```

3. Create a function for premium in n-th year

```
#define BMS discount for Malaysia
BMS=c(1,0.75,0.7,0.6167,0.55,0.45)

#create function for mean premium in nth year, when prior premium is $1000
prem=function(n,\lambda)
{ p=numeric(0)
  for (i in 1:length(BMS))
  p[i]=mean(powA(n,\lambda)[,i])
  1000*sum(p*BMS)}

#example for premium in 3rd year
prem(3,0.1)

#output
[1] 580.5789
```

4. Provide premiums for 20 years

```
#provide mean premium for all 20 years
allprem=function(n,\lambda)
{ p=numeric(0)
```

```
    for (i in 1:n)
    p[i]=prem(i,\lambda)
    p}
  round(allprem(20,0.1),1)

  #output
  [1] 625.5 598.7 580.6 570.6 565.8 565.8 565.8 565.8 565.8 565.8 565.8
  565.8 565.8 565.8 565.8 565.8 565.8 565.8 565.8 565.8
```

Example 15.4.5. Using the results from Example 15.4.2, observe the mean premiums in 20 years under the *BMS* in Brazil, assuming that the premium prior to implementing the bonus-malus discount is 1000.

Example Solution. From Example 15.4.2,the transition matrix for the *BMS* in Brazil is:

$$\mathbf{P} = \begin{bmatrix} 0.0952 & 0.9048 & 0 & 0 & 0 & 0 & 0 \\ 0.0952 & 0 & 0.9048 & 0 & 0 & 0 & 0 \\ 0.0047 & 0.0905 & 0 & 0.9048 & 0 & 0 & 0 \\ 0.0002 & 0.0045 & 0.0905 & 0 & 0.9048 & 0 & 0 \\ 0.0000 & 0.0002 & 0.0045 & 0.0905 & 0 & 0.9048 & 0 \\ 0.0000 & 0.0000 & 0.0002 & 0.0045 & 0.0905 & 0 & 0.9048 \\ 0.0000 & 0.0000 & 0.0000 & 0.0002 & 0.0045 & 0.0905 & 0.9048 \end{bmatrix}$$

Using ‘R‘, the premiums in 20 years are:

766.9, 737.6, 713.1, 693.8, 679.2, 669.3, 664.0, 660.5, 658.8, 657.8,

657.2, 656.9, 656.7, 656.6, 656.6, 656.6, 656.6, 656.5, 656.5, 656.5.

The results in Examples 15.4.4-15.4.5 allow us to observe the evolution of mean premium for the *BMS* in Malaysia and Brazil. The evolution of premiums for both countries are provided in Table 15.4, and are shown graphically in Figure 15.4.

Table 15.4. Evolution of Premium (Malaysia and Brazil)

Year	Premium Malaysia	Premium Brazil	Year	Premium Malaysia	Premium Brazil
1	625.5	766.9	11	565.8	657.2
2	598.7	737.6	12	565.8	656.9
3	580.6	713.1	13	565.8	656.7
4	570.6	693.8	14	565.8	656.6
5	565.8	679.2	15	565.8	656.6
6	565.8	669.3	16	565.8	656.6
7	565.8	664.0	17	565.8	656.6
8	565.8	660.5	18	565.8	656.5
9	565.8	658.8	19	565.8	656.5
10	565.8	657.7	20	565.8	656.5

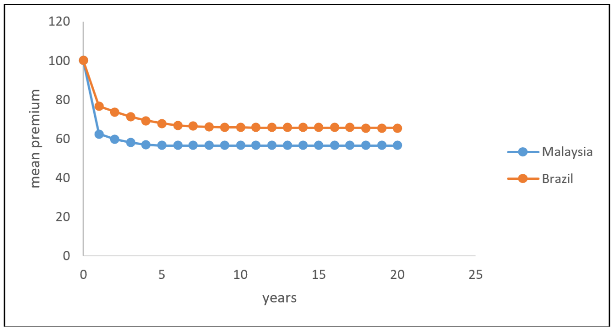


FIGURE 15.4: **Evolution of Premium (Malaysia and Brazil)**

The following R code can be used to create a function for the transition matrix under the BMS in Brazil. The function can be used to find the mean premium in the n-th year and the mean premiums in 20 years. The function can also be used in the later section (Section 15.5).

1. Create a function for transition matrix

```
#create function for transition matrix without 1st column
TM=function(\lambda)
{ P=matrix(data=0,nrow=7,ncol=7)
  P[1,2]=P[2,3]=P[3,4]=P[4,5]=P[5,6]=P[6,7]=P[7,7]=exp(-\lambda)
  P[3,2]=P[4,3]=P[5,4]=P[6,5]=P[7,6]=\lambda*exp(-\lambda)
  P[4,2]=P[5,3]=P[6,4]=P[7,5]=(1/2)*(\lambda^2)*exp(-\lambda)
  P[5,2]=P[6,3]=P[7,4]=(1/6)*(\lambda^3)*exp(-\lambda)
  P[6,2]=P[7,3]=(1/24)*(\lambda^4)*exp(-\lambda)
  P[7,2]=(1/120)*(\lambda^5)*exp(-\lambda)
  P}

#add 1st column in transition matrix
```

```
TP=function(\lambda)
{P=TM(\lambda)
for (i in 1:7)
 P[i,1]=1-sum(TM(\lambda)[i,-1])
 P}

#provide transition matrix (in 4 decimal places)
round(TP(0.1),4)

#output
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
[1,] 0.0952 0.9048 0.0000 0.0000 0.0000 0.0000 0.0000
[2,] 0.0952 0.0000 0.9048 0.0000 0.0000 0.0000 0.0000
[3,] 0.0047 0.0905 0.0000 0.9048 0.0000 0.0000 0.0000
[4,] 0.0002 0.0045 0.0905 0.0000 0.9048 0.0000 0.0000
[5,] 0.0000 0.0002 0.0045 0.0905 0.0000 0.9048 0.0000
[6,] 0.0000 0.0000 0.0002 0.0045 0.0905 0.0000 0.9048
[7,] 0.0000 0.0000 0.0000 0.0002 0.0045 0.0905 0.9048
```

2. Recall the n^{th} power of a square matrix

```
round(powA(n=3,\lambda=0.1),4)

  #output
         [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
  [1,] 0.0211 0.1602 0.0779 0.7408 0.0000 0.0000 0.0000
  [2,] 0.0174 0.0157 0.2261 0.0000 0.7408 0.0000 0.0000
  [3,] 0.0024 0.0233 0.0112 0.2222 0.0000 0.7408 0.0000
  [4,] 0.0010 0.0023 0.0226 0.0111 0.2222 0.0000 0.7408
  [5,] 0.0001 0.0009 0.0022 0.0226 0.0111 0.2222 0.7408
  [6,] 0.0000 0.0001 0.0009 0.0024 0.0262 0.0815 0.8890
  [7,] 0.0000 0.0000 0.0002 0.0017 0.0127 0.0927 0.8927
```

3. Recall the premium in *n*-th year

```
#define BMS discount for Brazil
BMS=c(1,0.9,0.85,0.8,0.75,0.7,0.65)

#call function for premium
prem(n=3,\lambda=0.1)

#output
[1] 713.117
```

4. Recall all premiums in 20 years

```
round(allprem(n=20,\lambda=0.1),1)

#output
[1] 766.9 737.6 713.1 693.8 679.2 669.3 664.0 660.5 658.8 657.7 657.2 656.9
 656.7 656.6 656.6 656.6 656.5 656.5 656.5 656.5
```

15.4.5 Convergence Rate

We may also be interested to determine the variation between the probability in the n-th year, $p_{ij}^{(n)}$, and the stationary probability, π_j. The variation between the probabilities can be measured using:

$$\left| average(p_{ij}^{(n)}) - \pi_j \right|.$$

Therefore, the total variation can be measured by the sum of variation in all classes:

$$\sum_j \left| average(p_{ij}^{(n)}) - \pi_j \right|.$$

The total variation is also called the convergence rate because it measures the convergence rate after n years (or n transitions). A lower total variation implies a better convergence rate between the n-step transition probabilities and the stationary distribution.

Example 15.4.6. Using the results from Example 15.4.4, provide the total variations (convergence rate) in 20 years under the *BMS* in Malaysia.

Example Solution. Using ‘R‘, the stationary probabilities are:

$$\begin{bmatrix} \pi_0 \\ \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \\ \pi_5 \end{bmatrix} = \begin{bmatrix} 0.0952 \\ 0.0861 \\ 0.0779 \\ 0.0705 \\ 0.0638 \\ 0.6064 \end{bmatrix}$$

The transition matrix in the first year is:

$$\mathbf{P}^{(1)} = \begin{bmatrix} 0.0952 & 0.9048 & 0 & 0 & 0 & 0 \\ 0.0952 & 0 & 0.9048 & 0 & 0 & 0 \\ 0.0952 & 0 & 0 & 0.9048 & 0 & 0 \\ 0.0952 & 0 & 0 & 0 & 0.9048 & 0 \\ 0.0952 & 0 & 0 & 0 & 0 & 0.9048 \\ 0.0952 & 0 & 0 & 0 & 0 & 0.9048 \end{bmatrix}$$

The variation can be computed as:

$$\begin{aligned}
\left|\sum_i \tfrac{p_{i0}}{6} - \pi_0\right| &= 0\\
\left|\sum_i \tfrac{p_{i1}}{6} - \pi_1\right| &= 0.0647\\
&\vdots\\
\left|\sum_i \tfrac{p_{i5}}{6} - \pi_5\right| &= .3048
\end{aligned}$$

Therefore, the total variation in the first year is

$$\sum_j \left|\sum_i \frac{p_{ij}}{6} - \pi_j\right| = 0.6096.$$

Using ‘R‘, the total variations (or convergence rate) in 20 years are:

0.6096, 0.3941, 0.2252, 0.0958, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.

15.4.6 `R` Code for Convergence Rate

The following `R` code can be used to calculate the total variation in the nth year, and the total variations (convergence rates) in 20 years under the *BMS* in Malaysia (the solution in Example 15.4.6).

1. Recall the Transition Matrix
2. Create a function for stationary probabilities
3. Create a function for total variation in** **the n-th year
4. Provide total variations (convergence rate) in 20 years

1. Recall the Transition Matrix

```
TP(0.1)

#output
           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 0.09516258 0.9048374 0.0000000 0.0000000 0.0000000 0.0000000
[2,] 0.09516258 0.0000000 0.9048374 0.0000000 0.0000000 0.0000000
[3,] 0.09516258 0.0000000 0.0000000 0.9048374 0.0000000 0.0000000
[4,] 0.09516258 0.0000000 0.0000000 0.0000000 0.9048374 0.0000000
[5,] 0.09516258 0.0000000 0.0000000 0.0000000 0.0000000 0.9048374
[6,] 0.09516258 0.0000000 0.0000000 0.0000000 0.0000000 0.9048374
```

2. Create a function for stationary probabilities

```
#create function for stationary probabilities
ST=function(\lambda)
{Re(t(eigen(t(TP(\lambda)))$vectors[,1]/sum(eigen(t(TP(\lambda)))$vectors[,1])))}
ST(0.1)

#output
           [,1]       [,2]       [,3]       [,4]       [,5]      [,6]
[1,] 0.09516258 0.08610666 0.07791253 0.07049817 0.06378939 0.6065307
```

3. Create a function for total variation in the *n*-th year

```
#create function for total variation in nth year
TV=function(n,\lambda)
{ dif=numeric(0)
  for (j in 1:length(ST(\lambda)))
  dif[j]=abs(mean(powA(n,\lambda)[,j])-ST(\lambda)[j])
  sum(dif)}

#example for n=2
TV(2,0.1)

#output
[1] 0.3943306
```

4. Provide total variations (convergence rate) in 20 years

```
#provide total variation in each year for n years (4 decimal places)
tot.var=function(n,\lambda)
{ q=numeric(0)
  for (t in 1:n)
  q[t]=TV(t,\lambda)
  q}
round(tot.var(20,0.1),4)

#output
[1] 0.6098 0.3943 0.2253 0.0959 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
0.0000
[13] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
```

Example 15.4.7. Provide the total variations (or convergence rate) in 20 years under the *BMS* in Brazil using the results from example 15.4.5.

Example Solution. Using ‘R‘ code, the total variations (or convergence rates) in 20 years for the BMS in Brazil are:

1.2617, 1.0536, 0.8465, 0.6412, 0.4362, 0.2316, 0.1531, 0.0747, 0.0480, 0.0232,

0.0145, 0.0071, 0.0043, 0.0021, 0.0013, 0.0006, 0.0004, 0.0002, 0.0001, 0.0001.

Examples 15.4.6-15.4.7 provide the degree of convergence for two different *BMS* (two different countries). The Malaysian *BMS* reaches full stationary only after five years, while the *BMS* in Brazil takes a longer period. As mentioned in Lemaire (1998), a more sophisticated *BMS* would converge more slowly, and is considered as a drawback as it takes a longer period to stabilize. The main objective of a *BMS* is to separate the good drivers from the bad drivers, and thus, it is desirable to have a classification process that can be finalized (or stabilized) as soon as possible.

15.5 *BMS* and Premium Rating

In this section, you learn how to:

- Integrate priori information into optimal relativities.
- Calculate probability of staying in BMS level.
- Calculate constrained optimal relativities.
- Calculate unconstrained optimal relativities.

15.5.1 Premium Rating

In motor insurance ratemaking, BMS is a form of *a posteriori* rating mechanism to complement the use of *a priori* risk classification as described in Chapter 11. The *a priori* ***risk segmentation*** introduced in Section 11.1 divides portfolio of drivers into a number of homogeneous risk classes based on observable rating factors (see also Section 11.3.1), such that policyholders in the same risk class pay the same *a priori* premium. The underlying reason for utilizing BMS that relies on claims experience information is to deal with the *residual heterogeneity* within each homogeneous risk class (e.g., see the discussion before Example 2.6.2) since the observable variables are far from perfect in predicting the riskiness of driving behaviors.

The ideal *a posteriori* mechanism is the credibility premium framework developed in Chapter 12 (see also Dionne and Vanasse (1989)), whereby premiums are derived on an individual basis for each policyholder by incorporating both the *a priori* and *a posteriori* information. However, such individual premium determination is overly complex from a commercial standpoint for practical implementations by motor insurers. For this reason, *BMS* is the preferred solution and it consists of the following three building blocks: (a) *BMS* classes; (b) transition rules; (c) premium levels (also known as premium relativities or premium adjustment coefficients).

The first two building blocks are pre-specified in advance and have been discussed in previous sections, whereas the determination (instead of predetermined as discussed in the cases of Malaysian, Brazilian and Swiss systems) of premium relativities are important for motor insurers precisely because of its complementary and correction nature to account for the imperfection or inaccuracies in the *a priori* risk classification. Note that the premium relativities under BMS are different from the relativity measure defined in Section 11.3.2, which is the ratio between the expected risk of a given risk class to a baseline risk class, both of which are calculated from observable rating factors. In the following subsections, we briefly introduce the required modelling setup to study the determination of optimal relativities. We refer interested readers to Denuit et al. (2007) for a fuller discussion on the technical details.

15.5.2 A Priori Risk Classification

Let us consider a portfolio of n policies, where the risk exposure (see Section 11.2.3) of driver i is denoted as m_i and the number of claims reported is represented by Y_i, following from the notations used in Section 11.3.3. Let $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \ldots, x_{iq})$ be the vector of observable variables for $i = 1, 2, \ldots, n$. The Poisson regression as developed in Section 11.3.3 is commonly chosen to model Y_i under the generalized linear models (GLM) framework, see Section 13.3.2.2 and also McCullagh and Nelder (1989).

We can then express the predicted *a priori* expected claim frequency for policyholder i as

$$\mu_i = m_i\lambda_i = m_i \exp\left(\hat{\beta}_0 + \sum_{m=1}^{q} \hat{\beta}_m x_{im}\right),$$

where $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_q$ are the estimated regression coefficients. In other words, $\lambda_i = \frac{\mu_i}{m_i}$ is the expected claim frequency per unit exposure, which is the main focus of the *a priori* risk classification.

15.5.3 Modelling of Residual Heterogeneity

Since unobserved factors that may affect driving behaviors are not taken into account in estimating the expected claim frequency, insurers would have to account for the residual heterogeneity within each *a priori* risk class by introducing a random effect component Θ_i into the conditional distribution of

Y_i. Given $\Theta_i = \theta$, Y_i follows a Poisson distribution with mean $\lambda_i\theta$, that is,

$$\Pr(Y_i = k|\Theta_i = \theta) = \exp(-\lambda_i\theta)\frac{(\lambda_i\theta)^k}{k!}, k = 0, 1, 2, ...$$

Following from the setup of gamma-Poisson model in Section 9.3.2, we further assume that all the Θ_i's are independent and follow a gamma (a, a) distribution with the following density function as introduced in Appendix 20.2

$$f(\theta) = \frac{1}{\Gamma(a)}a^a\theta^{a-1}\exp(-a\theta), \quad \theta > 0,$$

where the use of gamma-Poisson mixture produces a negative binomial distribution for Y_i (see Section 4.3.2). With these specifications, we obtain $\mathrm{E}(\Theta_i) = 1$ and hence $\mathrm{E}(Y_i) = \mathrm{E}(\mathrm{E}(Y_i|\Theta_i)) = \mathrm{E}(\lambda_i\Theta_i) = \lambda_i$. by the law of iterated expectation in Appendix 18.2.

Furthermore, it can be shown that the posterior distribution of $\Theta|y_1 = k_1, y_2 = k_2, \ldots, y_n = k_n$ is gamma distributed with parameters $a + \sum_{j=1}^n k_j$ and $a + n\lambda_i$ and therefore the Bayesian premium is given as

$$\mathrm{E}(\lambda_i\Theta|y_1 = k_1, \ldots, y_n = k_n) = \lambda_i \times \frac{a + \sum_{j=1}^n k_j}{a + n\lambda_i}.$$

On the other hand, applying the Bühlmann credibility-weighted estimate in Section 12.2 to the gamma-Poisson model in Section 9.3.2, we obtain

$$\begin{array}{cl}
EPV & = \mathrm{E}(\mathrm{Var}(Y|\lambda_i)) = \mathrm{E}(\lambda_i\Theta) = \lambda_i, \\
VHM & = \mathrm{Var}(\mathrm{E}(Y|\lambda_i)) = \mathrm{Var}(\lambda_i\Theta) = \frac{\lambda_i^2}{a}, \\
K & = \frac{EPV}{VHM} = \frac{\lambda_i}{\frac{\lambda_i^2}{a}} = \frac{a}{\lambda_i}, \\
Z & = \frac{n}{n+K} = \frac{n\lambda_i}{n\lambda_i+a}, \\
\bar{Y} & = \frac{\sum_{j=1}^n y_j}{n} = \frac{\sum_{j=1}^n k_j}{n}, \\
\mu & = \mathrm{E}(\mathrm{E}(Y_i|\lambda_i)) = \mathrm{E}(\lambda_i\Theta) = \lambda_i,
\end{array}$$

and hence the credibility-weighted estimate as

$$\begin{aligned}
\mathrm{E}[\mathrm{E}(Y|\lambda_i)|y_1 = k_1, ..., y_n = k_n] & = \mathrm{E}[\lambda_i\Theta|y_i = k_1, ..., y_n = k_n] \\
& = Z\bar{Y} + (1 - Z)\mu \\
& = \frac{n\lambda_i}{n\lambda_i+a}\frac{\sum_{j=1}^n k_j}{n} + \frac{a}{n\lambda_i+a}\lambda_i \\
& = \frac{\lambda_i(a+\sum_{j=1}^n k_j)}{a+n\lambda_i}
\end{aligned}$$

that is, the Bühlmann credibility premium exactly matches the Bayesian premium.

Despite the fact that the credibility premium derived on an individual basis above is the ideal *a posteriori* premium, in practice insurers make use of BMS as a discrete approximation to the Bayesian premium, due to the relatively simpler structure of BMS as compared to the individual calculations of credibility premium.

15.5.4 Stationary Distribution Allowing for Residual Heterogeneity

Suppose that a driver is selected at random from the portfolio that has been classified into h risk classes via the use of observed *a priori* variables. The true expected claim frequency for this driver is given by $\Lambda\Theta$, where Λ is the unknown *a priori* expected claim frequency and Θ is the random residual heterogeneity. Let us further denote w_g as the proportion of drivers in the g-th risk class, that is, $w_g = \Pr(\Lambda = \lambda_g) = \frac{n_g}{n}$ where n_g is the number of drivers classified in the g-th risk class. Note that since there are two different concepts of risk classes (from *a priori* risk classification) and *BMS* (or *NCD*) classes (for *a posteriori* rating mechanism), for the rest of this chapter we will refer *BMS* classes as *BMS* levels instead to avoid unnecessary confusion.

Let $p_{ij}^{\lambda}(\lambda\theta)$ be the transition probability of moving from BMS level i to level j for a driver with expected claim frequency $\lambda\theta$ belonging to the risk class with predicted claim frequency of λ. In other words, the one-step transition matrix can be written as $\mathbf{P}(\lambda\theta;\lambda) = \{p_{ij}^{\lambda}(\lambda\theta)\}$. The row vector of the stationary distribution $\boldsymbol{\pi} = (\pi_0^{\lambda}(\lambda\theta), \pi_1^{\lambda}(\lambda\theta), \ldots, \pi_{k-1}^{\lambda}(\lambda\theta))$ can be obtained by solving the following conditions:

$$\begin{aligned} \boldsymbol{\pi}(\lambda\theta;\lambda)\mathbf{P}(\lambda\theta;\lambda) &= \boldsymbol{\pi}(\lambda\theta;\lambda) \\ \boldsymbol{\pi}(\lambda\theta;\lambda)\mathbf{1} &= 1 \end{aligned}$$

where $\mathbf{1}$ is the column vector of 1's and $\pi_\ell^{\lambda}(\lambda\theta)$ is the stationary probability for a driver with true expected claim frequency of $\lambda\theta$ to be in level ℓ when the equilibrium steady state is reached in the long run.

Note that the equation for stationary distribution that allows for residual heterogeneity in this section, $\boldsymbol{\pi}(\lambda\theta)\mathbf{P}(\lambda\theta) = \boldsymbol{\pi}(\lambda\theta)$, is similar to the equation of stationary distribution in Section 15.4.1 where $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$. The only difference is that the stationary distribution $\boldsymbol{\pi}(\lambda\theta)$ and the transition matrix $\boldsymbol{P}(\lambda\theta)$ are written in terms of a function of $\boldsymbol{\lambda}\theta$.

With these setup, the probability of drivers staying in *BMS* level $L = \ell$ for

$\ell = 0, 1, \ldots, k-1$ in the context of the entire portfolio can be obtained as

$$\begin{aligned}
\Pr(L=\ell) &= \sum_{g=1}^{h} \Pr(L=\ell|\Lambda=\lambda_g)\Pr(\Lambda=\lambda_g) \\
&= \sum_{g=1}^{h} \Pr(\Lambda=\lambda_g)\int_0^\infty \Pr(L=\ell|\Lambda=\lambda_g, \Theta=\theta)f(\theta)d\theta \\
&= \sum_{g=1}^{h} w_g \int_0^\infty \pi_\ell^{\lambda_g}(\lambda_g\theta)f(\theta)d\theta.
\end{aligned}$$

From previous section (section 15.5.3), Θ_i is the random effect component. As an example, if we assume that all Θ_i's are independent and follow a gamma (a, a) distribution, then $f(\theta)$ is the density function of a gamma (a, a) distribution.

For further understanding, we provide `R` program to calculate the probability of staying in level $L = \ell$, $\Pr(L = \ell)$, for the Malaysian BMS. From previous sections, the functions for transition matrix and stationary distribution are functions of λ. Therefore, the functions allow us to write `R` program to calculate $\Pr(L = \ell)$. Similar `R` programs can be developed for the Brazilian BMS.

Example 15.5.1. Consider the *BMS* levels and the transition rules of the Malaysian system (-1/Top). Assume that the following 3 values of *a priori* expected claim frequency are given; $\lambda_1 = 0.1$, $\lambda_2 = 0.3$, $\lambda_3 = 0.5$, with the following proportions (weights); $\Pr(\Lambda = \lambda_1) = 0.6$, $\Pr(\Lambda = \lambda_2) = 0.3$, $\Pr(\Lambda = \lambda_3) = 0.1$. We also assume that the gamma parameter is fixed at $a = 1.5$. Calculate the probability of staying in level $L = \ell$, $\Pr(L = \ell)$.

Example Solution. We can use `R` program to calculate $\Pr(L = \ell)$.

1. Define the Parameters

```
#define parameters
a.hat=1.5
\lambda.hat=c(0.1,0.3,0.5)
weight=c(0.6,0.3,0.1)
```

2. Recall the Transition Matrix

```
#As an example, \lambda=0.1, theta=1, so that \lambda*theta=0.1
TP(0.1)

#output
           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 0.09516258 0.9048374 0.0000000 0.0000000 0.0000000 0.0000000
[2,] 0.09516258 0.0000000 0.9048374 0.0000000 0.0000000 0.0000000
[3,] 0.09516258 0.0000000 0.0000000 0.9048374 0.0000000 0.0000000
[4,] 0.09516258 0.0000000 0.0000000 0.0000000 0.9048374 0.0000000
[5,] 0.09516258 0.0000000 0.0000000 0.0000000 0.0000000 0.9048374
[6,] 0.09516258 0.0000000 0.0000000 0.0000000 0.0000000 0.9048374
```

3. Recall the Stationary Distribution

```
#As an example, \lambda=0.1, theta=1, so that \lambda*theta=0.1
ST(0.1)

#output
           [,1]       [,2]       [,3]       [,4]       [,5]      [,6]
[1,] 0.09516258 0.08610666 0.07791253 0.07049817 0.06378939 0.6065307
```

4. Calculate $\Pr(L = \ell)$

```
#create function for pi*fdist
int1=function(theta,s,a,\lambda)
{   a=a.hat
    \lambda=\lambda.hat[j]
    f.dist=gamma(a)^(-1)*a^a*theta^(a-1)*exp(-a*theta)
    p=ST(\lambda*theta)
    return(p[1,s+1]*f.dist)}

#create matrix for integral of pi*fdist
#there are 3 rating classes (each with \lambda=0.1,0.3,0.5), so we need a 3x6 matrix
P1=matrix(nrow=3,ncol=6,data=0)
for (j in 1:3)
{for (i in 0:5) P1[j,i+1]=integrate(Vectorize(int1),lower=0,upper=Inf,s=i)$value}
P1

#output
           [,1]      [,2]      [,3]       [,4]       [,5]      [,6]
[1,] 0.09226953 0.0789042 0.0681005 0.05926000 0.05194672 0.6495191
[2,] 0.23927422 0.1570442 0.1095757 0.08001922 0.06053327 0.3535534
[3,] 0.35048094 0.1847610 0.1112046 0.07298753 0.05092619 0.2296397

#calculate probability of L (weight*matrix P1)
prob.L=t(weight)%*%P1
prob.L

#output
          [,1]      [,2]       [,3]       [,4]       [,5]      [,6]
[1,] 0.1621921 0.1129319 0.08485348 0.06686052 0.05442063 0.5187414
```

The results indicate that in the long run and under residual heterogeneity, 16% of insureds will eventually belong to level $\ell = 0$, 11% of insureds will eventually belong to level $\ell = 1$, and so forth. The majority of insureds (more than half, i.e. 52%) will eventually occupy the highest level which is level $\ell = 5$.

15.5.5 Determination of Optimal Relativities

The *optimal relativity* for each *BMS* level was first derived by Norberg (1976) through the minimization of the following objective function, which is more commonly known as the Norberg's criterion:

$$\min \mathrm{E}((\bar{\lambda}\Theta - \bar{\lambda} r_L)^2) = \min \mathrm{E}((\Theta - r_L)^2),$$

where $\bar{\lambda}$ is the constant expected claim frequency for all policyholders in the absence of *a priori* risk classification and r_L is the premium relativity for BMS level L. It should be noted that when there is no risk classification, all λ's are equal and they are represented by a constant $\bar{\lambda}$.

Pitrebois et al. (2003) then incorporated the information of *a priori* risk classification into the optimization of the same objective function of

$$\min \mathrm{E}((\Theta - r_L)^2)$$

to derive r_L analytically. Tan et al. (2015) further proposed the minimization of the following objective function

$$\min \mathrm{E}((\Lambda\Theta - \Lambda r_L)^2), \text{ subject to } \mathrm{E}(r_L) = 1$$

under a financial balanced constraint (that is, the expected premium relativity equals 1) to determine the optimal relativities of a *BMS* given pre-specified *BMS* levels and transition rules, where

$$\begin{aligned}
\min \mathrm{E}[(\Lambda\Theta - \Lambda \mathrm{r_L})^2] &= \sum_{\ell=0}^{k-1} \mathrm{E}[(\Lambda\Theta - \Lambda r_L)^2 | L = \ell] \Pr(L = \ell) \\
&= \sum_{\ell=0}^{k-1} \mathrm{E}(\mathrm{E}[(\Lambda\Theta - \Lambda r_L)^2 | L = \ell, \Lambda) | L = \ell] \Pr(L = \ell) \\
&= \sum_{\ell=0}^{k-1} \sum_{g-1}^{h} \mathrm{E}((\Lambda\Theta - \Lambda r_L)^2 | L = \ell, \Lambda = \lambda_g) \Pr(\Lambda = \lambda_g | L = \ell) \Pr(L = \ell) \\
&= \sum_{\ell=0}^{k-1} \sum_{g=1}^{h} \textstyle\int_0^\infty (\lambda_g \theta - \lambda_g r_\ell)^2 \pi_\ell(\lambda_g \theta) w_g f(\theta) d\theta \\
&= \sum_{g=1}^{h} w_g \textstyle\int_0^\infty \sum_{\ell=0}^{k-1} (\lambda_g \theta - \lambda_g r_\ell)^2 \pi_\ell(\lambda_g \theta) f(\theta) d\theta.
\end{aligned}$$

It is crucially important that the optimal relativity has an average of 100%, so that the bonuses and maluses exactly offset each other to result in a financial equilibrium condition. Note that the approach considered by Pitrebois et al. (2003) does not require the financial balanced constraint because the analytical solution to its objective function is given by $r_\ell = \mathrm{E}(\Theta | L = \ell)$, so it follows that $\mathrm{E}(r_L) = \mathrm{E}\left(\mathrm{E}(\Theta | L)\right) = \mathrm{E}(\Theta) = 1$ with the specific choice of gamma (a, a) distribution for the random effect component Θ.

In this case, the optimization problem can be solved by specifying the Lagrangian as

$$\begin{aligned}
\boldsymbol{\mathcal{L}}(\mathbf{r}, \alpha) &= \mathrm{E}((\Lambda\Theta - \Lambda r_L)^2) + \alpha(\mathrm{E}(r_L) - 1) \\
&= \sum_{\ell=0}^{k-1} \mathrm{E}((\Lambda\Theta - \Lambda r_L)^2 | L = \ell) \Pr(L = \ell) + \alpha(\sum_{\ell=0}^{k-1} r_\ell \Pr(L = \ell) - 1),
\end{aligned}$$

where $\mathbf{r} = (r_0, r_1, \ldots, r_{k-1})^T$. The required first order conditions are given as follows

$$\begin{aligned} \Pr(L=\ell)(2\mathrm{E}(\Lambda^2\Theta - \Lambda^2 r_L|L=\ell) - \alpha) &= 0, \qquad \ell = 0,1,...,k-1 \\ \sum_{\ell=0}^{k-1} r_\ell \Pr(L=\ell) - 1 &= 0. \end{aligned}$$

Finally, the solution set for α and $r_\ell, \ell = 0, 1, \ldots, k-1$ is obtained as

$$\begin{aligned} \alpha &= \frac{(\sum_{\ell=0}^{k-1} \frac{\mathrm{E}(\Lambda^2\Theta|L=\ell)}{\mathrm{E}(\Lambda^2|L=\ell)}) - 1}{\sum_{\ell=0}^{k-1} \frac{\Pr(L=\ell)}{2\mathrm{E}(\Lambda^2|L=\ell)}}, \\ r_\ell &= \frac{\mathrm{E}(\Lambda^2\Theta|L=\ell)}{\mathrm{E}(\Lambda^2|L=\ell)} - \frac{\alpha}{2\mathrm{E}(\Lambda^2|L=\ell)}, \end{aligned}$$

where

$$\begin{aligned} \Pr(L=\ell) &= \sum_{g=1}^{h} w_g \int_0^\infty \pi_\ell^{\lambda_g}(\lambda_g\theta) f(\theta) d\theta, \\ \mathrm{E}(\Lambda^2\Theta|L=\ell) &= \frac{\sum_{g=1}^{h} w_g \int_0^\infty \lambda_g^2 \theta \pi_\ell^{\lambda_g}(\lambda_g\theta) f(\theta) d\theta}{\sum_{g=1}^{h} w_g \int_0^\infty \pi_\ell^{\lambda_g}(\lambda_g\theta) f(\theta) d\theta}, \\ \mathrm{E}(\Lambda^2|L=\ell) &= \frac{\sum_{g=1}^{h} w_g \int_0^\infty \lambda_g^2 \pi_\ell^{\lambda_g}(\lambda_g\theta) f(\theta) d\theta}{\sum_{g=1}^{h} w_g \int_0^\infty \pi_\ell^{\lambda_g}(\lambda_g\theta) f(\theta) d\theta}. \end{aligned}$$

If we perform the optimization without the financial balanced constraint, then we obtain

$\alpha^{\text{unconstrained}} = 0$, and $r_\ell^{\text{unconstrained}} = \frac{\mathrm{E}(\Lambda^2\Theta|L=\ell)}{\mathrm{E}(\Lambda^2|L=\ell)}$.

It should be noted that the optimal relativity r_ℓ of each level ℓ is the scale that determines the premium's discount/loading to the insureds. If $r_\ell < 1$, then the insured receives a discount based on his favorable past performance. If $r_\ell > 1$, then the insured is penalized and has to pay additional loading based on his past performance. The concept is similar to the discount under the *BMS* in Malaysia and Brazil which were discussed in previous sections. The difference is that the discounts under the *BMS* in Malaysia and Brazil were pre-determined (r_ℓ =100%, 75%, 70%, 61.67%, 55%, 45% respectively for ℓ =0,1,2,3,4,5 for *BMS* in Malaysia; r_ℓ =100%, 90%, 85%, 80%, 75%, 70%, 65% respectively for ℓ =0,1,2,3,4,5,6 for *BMS* in Brazil), whereas the relativities r_ℓ under heterogeneous residual are determined using optimization (minimization of an objective function which is subjected to a constraint).

For further understanding, we provide R program to calculate the optimal relativity r_ℓ under *unconstrained method* which allows for residual heterogeneity for the Malaysian *BMS*. From previous sections, the functions for transition matrix and stationary distribution are functions of λ. Therefore, the functions allow us to write `R` program to calculate r_ℓ . Similar `R` program can be developed for the Brazilian BMS.

It should be noted that the calculation of optimal relativity under *constrained method* needs more formulas and codes. The reader can create the codes on their own by referring to the codes under the *unconstrained method* provided in Example 15.5.2.

Example 15.5.2. Consider the *BMS* levels and the transition rules of the Malaysian system (-1/Top) in Example 15.5.1. Calculate the the optimal relativity r_ℓ under *unconstrained method* which allows for residual heterogeneity.

Example Solution. We can use R program to calculate r_ℓ under *unconstrained method*.

1. Recall the parameters from Example 15.5.1

```
a.hat=1.5
    \lambda.hat=c(0.1,0.3,0.5)
    weight=c(0.6,0.3,0.1)
```

2. Recall the Transition Matrix for Malaysian BMS

```
TP(0.1)

#output
           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 0.09516258 0.9048374 0.0000000 0.0000000 0.0000000 0.0000000
[2,] 0.09516258 0.0000000 0.9048374 0.0000000 0.0000000 0.0000000
[3,] 0.09516258 0.0000000 0.0000000 0.9048374 0.0000000 0.0000000
[4,] 0.09516258 0.0000000 0.0000000 0.0000000 0.9048374 0.0000000
[5,] 0.09516258 0.0000000 0.0000000 0.0000000 0.0000000 0.9048374
[6,] 0.09516258 0.0000000 0.0000000 0.0000000 0.0000000 0.9048374
```

3. Recall the Stationary Distribution

```
ST(0.1)

#output
           [,1]       [,2]       [,3]       [,4]       [,5]      [,6]
[1,] 0.09516258 0.08610666 0.07791253 0.07049817 0.06378939 0.6065307
```

4. Recall P1 from Example 15.5.1 (matrix for integral of pi*fdist)

```
P1
```

```
#output
           [,1]      [,2]      [,3]       [,4]       [,5]      [,6]
[1,] 0.09226953 0.0789042 0.0681005 0.05926000 0.05194672 0.6495191
[2,] 0.23927422 0.1570442 0.1095757 0.08001922 0.06053327 0.3535534
[3,] 0.35048094 0.1847610 0.1112046 0.07298753 0.05092619 0.2296397
```

5. Calculate the optimal relativities under heterogeneity r_ℓ

```
#create function for theta*pi*fdist
int2=function(theta,s,a,\lambda)
{   a=a.hat
    \lambda=\lambda.hat[j]
    f.dist=gamma(a)^(-1)*a^a*theta^(a-1)*exp(-a*theta)
    p=ST(\lambda*theta)
    return(theta*p[1,s+1]*f.dist)}

#create matrix for integral of theta*pi*fdist
#there are 3 rating classes (each with \lambda=0.1,0.3,0.5, so we need a 3x6 matrix)
P2=matrix(nrow=3,ncol=6,data=0)
for (j in 1:3)
{for (i in 0:5) P2[j,i+1]=integrate(Vectorize(int2),lower=0,upper=Inf,s=i)$value}
P2

#output
          [,1]      [,2]       [,3]       [,4]       [,5]       [,6]
[1,] 0.1490027 0.1196800 0.09737915 0.08014937 0.06664948 0.48713928
[2,] 0.3660619 0.2027370 0.12238497 0.07876803 0.05327145 0.17677669
[3,] 0.5128607 0.2082845 0.10207811 0.05653418 0.03412764 0.08611487

#calculate relativities, r
r=(t(weight*(\lambda.hat)^2)%*%P2)/(t(weight*(\lambda.hat)^2)%*%P1)
r

#output
         [,1]     [,2]     [,3]      [,4]      [,5]      [,6]
[1,] 1.495888 1.221421 1.047722 0.9263311 0.8360428 0.5133785
```

The results show that under *unconstrained method*, the first 3 levels ($\ell = 0,1,2$) have premium loadings ($r_\ell = 150\%,\ 122\%,\ 105\%$), and the last 3 levels ($\ell = 3,4,5$) have premium discounts ($r_\ell = 93\%,\ 84\%,\ 51\%$).

As mentioned above, the expected premium relativity equals 1, $\mathrm{E}(r_L) = 1$, under a financial balanced constraint (*constrained method*). As expected, the *unconstrained method* (in Example 15.5.2) does not provide expected premium relativity equals 1. We can use R program to find the expected premium relativity $\mathrm{E}(r_L)$.

Example 15.5.3. Consider Example 15.5.2. Find the expected premium relativity $\mathrm{E}(r_L)$.

Solution. We can use R program to find the expected premium relativity $\mathrm{E}(r_L)$.

1. Recall stationary probabilities from Example 15.5.1 and relativities from Example 15.5.2

```
rbind(prob.L,r)

#output
          [,1]      [,2]       [,3]       [,4]       [,5]      [,6]
[1,] 0.1621921 0.1129319 0.08485348 0.06686052 0.05442063 0.5187414
[2,] 1.4958885 1.2214214 1.04772189 0.92633112 0.83604284 0.5133785
```

2. Calculate the expected premium relativity $\mathrm{E}(r_L)$.

```
#calculate E(r)
expected.r=sum(prob.L*r)
expected.r

#output
[1] 0.8432052
```

The results show that under unconstrained method, the expected premium relativity is 84.32% (which is less than 100%).

15.5.6 Numerical Illustrations

In this section, we present two numerical illustrations that integrate *a priori* information into the determination of optimal relativities. We consider the *BMS* levels and the transition rules of both Malaysian and Brazilian systems but choose to calculate the set of optimal relativities instead of the specified premium levels given earlier. In our illustrations, by referring to Example 15.5.1-Example 15.5.2, the following 3 values of *a priori* expected claim frequency are given:

$$\lambda_1 = 0.1, \lambda_2 = 0.3, \lambda_3 = 0.5$$

with the following proportions:

$$\Pr(\Lambda = \lambda_1) = 0.6, \Pr(\Lambda = \lambda_2) = 0.3, \Pr(\Lambda = \lambda_3) = 0.1.$$

The gamma parameter is fixed at $a = 1.5$. Note that while these modelling assumptions are simple, the purpose here is to demonstrate the determination of optimal relativities under a relatively simple setup, and that the optimization procedure for the *BMS* remains the same even if the *a priori* risk classification is performed extensively. We refer interested readers to the motor vehicle claims data as documented in De Jong and Heller (2008) to conduct the *a priori* risk segmentation before proceeding to the determination of optimal relativities.

For the Malaysian *BMS* with 6 levels and the transition rule of -1/Top, the obtained numerical values of optimal relativities are presented in Table 15.5 together with the stationary probabilities. We find that around half of the policyholders will occupy the highest *BMS* level with the lowest premium relativity over the long run when the stationary state has been reached. We also observe that the constrained optimal relativities are higher than the unconstrained counterparts because of the need to satisfy the financial balanced constraint ($\mathrm{E}(r_L) = 100\%$).

Table 15.5. Optimal Relativities with $k = 6$ levels and transition rule of -1/Top

Level ℓ	$\Pr(L = \ell)$	r_ℓ	$r_\ell^{\text{unconstrained}}$
0	16.22%	131.99%	149.59%
1	11.29%	127.33%	122.14%
2	8.49%	120.64%	104.77%
3	6.69%	113.93%	92.63%
4	5.44%	107.79%	83.60%
5	51.87%	78.06%	51.34%
$\mathrm{E}(r_L)$		100%	84.32%

Moreover, we see that except for the highest *BMS* level (level 5), other *BMS* levels will impose malus surcharges to policyholders occupying those levels. This finding is not surprising since our theoretical framework here is to determine optimal relativities given the calculation of *a priori* base premiums by solely relying on claim frequency information but not claim severity. In practice, insurers could afford to introduce NCD levels with only discounts (bonuses) but not loadings (maluses) because the *a priori* base premiums have been inflated accordingly taking into account both the information of claim frequency and claim severity.

For the Brazilian *BMS* with 7 levels and the transition rule of -1/+1, the corresponding numerical values of optimal relativities are shown in Table 15.6. We find that around three quarters of the policyholders will occupy the highest *BMS* level with the lowest premium relativity in the stationary state. This finding is mainly due to the less severe penalty in the transition rule of -1/+1 in comparison to the rule of -1/Top, so more policyholders are expected to occupy the highest *BMS* level. Similar to the earlier example, we find that the unconstrained optimal relativities are lower and result in a lower value of $\mathrm{E}(r_L)$.

Table 15.6. Optimal Relativities with $k = 7$ levels and transition rule of -1/+1

Level ℓ	$\Pr(L = \ell)$	r_ℓ	$r_\ell^{\text{unconstrained}}$
0	3.28%	234.94%	228.65%
1	2.21%	196.24%	189.27%
2	2.00%	168.36%	160.59%
3	2.38%	145.96%	137.03%
4	4.02%	125.53%	114.63%
5	10.38%	106.25%	91.12%
6	75.74%	85.89%	61.74%
$\mathrm{E}(r_L)$		100%	78.97%

Note that the obtained values of optimal relativities may not be desirable for commercial implementations because of the possibility of irregular differences between adjacent *BMS* levels. To alleviate this problem, insurers could consider imposing linear optimal relativities in the form of $r_L^{\text{linear}} = a + bL$ by solving the following constrained optimization with an inequality constraint

$$\min \mathrm{E}\left((\Lambda\Theta - \Lambda a - \Lambda bL)^2\right) \text{ subject to } a + b\mathrm{E}(L) \geq 1.$$

We refer interested readers to Tan (2016) for a discussion on how to incorporate further commercial constraints and also on the solution to this optimization problem involving Kuhn-Tucker conditions.

15.6 Further Resources and Contributors

Further Reading and References

Note that our discussions in Section 15.5 focus on the classical frequency-driven BMS, which implicitly assume that the information of frequency and severity are independent, consistent with the collective risk model as discussed in Section 7.3. However, a number of recent empirical studies (Frees et al. (2016a); Garrido et al. (2016)) point towards the need due to their significant dependence structure. In this regard, Oh et al. (2020a) and Oh et al. (2020c) propose recent BMS framework that allows for such frequency-severity dependence based on the bivariate random effect model, where the former utilize both frequency and severity information in the specification of transition rule.

On the other hand, the framework presented in Section 15.5 is found to suffer from a double-counting problem, which results in biased premiums due to the dual role of the *a priori* rating factors in affecting both the *a priori* risk classification as well as *a posteriori* experience rating. We refer interested readers to Oh et al. (2020b) who propose to incorporate the estimation of *a priori* rate (in addition to the *a posteriori* rate) under a full optimization process to resolve the double-counting problem.

Contributors

- **Noriszura Ismail**, Universiti Kebangsaan Malaysia and **Chong It Tan**, Macquarie University, are the principal authors of the initial version of this chapter.
- **Noriszura Ismail**, Universiti Kebangsaan Malaysia is the principal author of the second edition of this chapter. Email: <ni@ukm.edu.my> for chapter comments and suggested improvements.

16

Quantifying Dependence

Chapter Preview. Dependence modeling involves using statistical models to describe the dependence structure between random variables and enables us to understand the relationships between variables in a dataset. This chapter introduces readers to techniques for modeling and quantifying dependence or association of multivariate distributions. Section 16.1 elaborates basic measures for modeling the dependence between variables.

Section 16.2 introduces an approach to modeling dependence using copulas which is reinforced with practical illustrations in Section 16.3. The types of copula families and basic properties of copula functions are explained in Section 16.4. The chapter concludes by explaining why the study of dependence modeling is important in Section 16.6.

16.1 Classic Measures of Scalar Associations

In this section, you learn how to:

- Estimate correlation using the Pearson method
- Use rank based measures like Spearman, Kendall to estimate correlation
- Measure tail dependency

In this chapter, we consider the first two variables from an insurance dataset of sample size ($n = 1500$) introduced in Frees and Valdez (1998) and is now readily available in the `copula` package; *losses* and *expenses*.

- `LOSS`, general liability claims from the Insurance Services Office, Inc. (ISO)
- `ALAE`, specifically attributable to the settlement of individual claims (e.g. lawyer's fees, claims investigation expenses)

We would like to know whether the distribution of `LOSS` depends on the distribution of `ALAE` or whether they are statistically independent. To visualize the relationship between losses and expenses, the scatterplots in Figure 16.1 are created on dollar and log dollar scales. It is difficult to see any relationship between the two variables in the left-hand panel. Their dependence is more evident when viewed on the log scale, as in the right-hand panel. This section elaborates basic measures for modeling the dependence between variables.

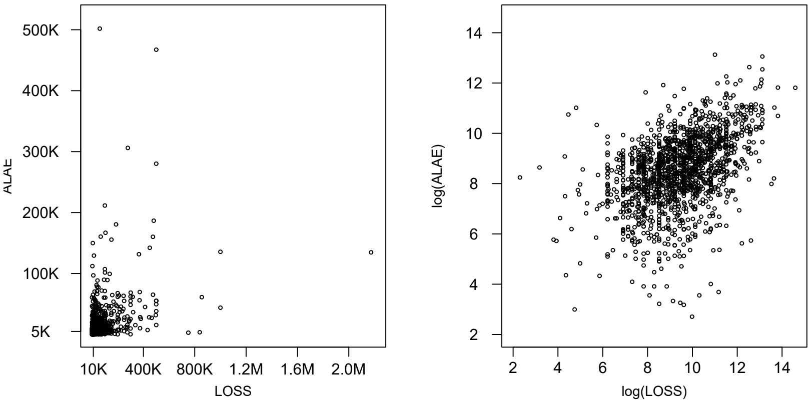


FIGURE 16.1: **Scatter Plot of LOSS and ALAE**

16.1.1 Association Measures for Quantitative Variables

For this section, consider a pair of random variables (X, Y) having joint distribution function $F(\cdot)$ and a random sample $(X_i, Y_i), i = 1, \ldots, n$. For the continuous case, suppose that $F(\cdot)$ has absolutely continuous marginals with marginal density functions.

Pearson Correlation

Define the sample covariance function $\widehat{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$, where $\bar{X}$ and $\bar{Y}$ are the sample means of X and Y, respectively. Then, the product-moment (Pearson) correlation can be written as

$$r = \frac{\widehat{Cov}(X,Y)}{\sqrt{\widehat{Cov}(X,X)\widehat{Cov}(Y,Y)}} = \frac{\widehat{Cov}(X,Y)}{\sqrt{\widehat{Var}(X)}\sqrt{\widehat{Var}(Y)}}.$$

The correlation statistic r is widely used to capture linear association between random variables. It is a (nonparametric) estimator of the correlation parameter ρ, defined to be the covariance divided by the product of standard deviations.

This statistic has several important features. Unlike regression estimators, it is symmetric between random variables, so the correlation between X

and Y equals the correlation between Y and X. It is unchanged by linear transformations of random variables (up to sign changes) so that we can multiply random variables or add constants as is helpful for interpretation. The range of the statistic is $[-1, 1]$ which does not depend on the distribution of either X or Y.

Further, in the case of independence, the correlation coefficient r is 0. However, it is well known that zero correlation does not in general imply independence, one exception is the case of normally distributed random variables. The correlation statistic r is also a (maximum likelihood) estimator of the association parameter for the bivariate normal distribution. So, for normally distributed data, the correlation statistic r can be used to assess independence. For additional interpretations of this well-known statistic, readers will enjoy Lee Rodgers and Nicewander (1998).

You can obtain the Pearson correlation statistic r using the `cor()` function in `R` and selecting the `pearson` method. This is demonstrated below by using the `LOSS` rating variable in millions of dollars and `ALAE` amount variable in dollars from the dataset in Figure 16.1.

From the `R` output above, $r = 0.4$, which indicates a positive association between `LOSS` and `ALAE`. This means that as the loss amount of a claim increases we expect expenses to increase.

16.1.2 Rank Based Measures

Spearman's Rho

The Pearson correlation coefficient does have the drawback that it is not invariant to nonlinear transforms of the data. For example, the correlation between X and $\log Y$ can be quite different from the correlation between X and Y. As we see from the `R` code for the Pearson correlation statistic above, the correlation statistic r between the `ALAE` variable in logarithmic dollars and the `LOSS` amounts variable in dollars is 0.33 as compared to 0.4 when we calculate the correlation between the `ALAE` variable in dollars and the `LOSS` amounts variable in dollars. This limitation is one reason for considering alternative statistics.

Alternative measures of correlation are based on ranks of the data. Let $R(X_j)$ denote the rank of X_j from the sample $X_1, \ldots, X_n$ and similarly for $R(Y_j)$. Let $R(X) = (R(X_1), \ldots, R(X_n))'$ denote the vector of ranks, and similarly for $R(Y)$. For example, if $n = 3$ and $X = (24, 13, 109)$, then $R(X) = (2, 1, 3)$. A comprehensive introduction of rank statistics can be found in, for example, Hettmansperger (1984). Also, ranks can be used to obtain the empirical distribution function, refer to Section 4.4.1 for more on the empirical distribution function.

With this, the correlation measure of Spearman (1904) is simply the product-moment correlation computed on the ranks:

$$r_S = \frac{\widehat{Cov}(R(X), R(Y))}{\sqrt{\widehat{Cov}(R(X), R(X))\widehat{Cov}(R(Y), R(Y))}} = \frac{\widehat{Cov}(R(X), R(Y))}{(n^2-1)/12}.$$

You can obtain the Spearman correlation statistic r_S using the `cor()` function in `R` and selecting the `spearman` method. From below, the Spearman correlation between the `LOSS` variable and `ALAE` variable is 0.45.

We can show that the Spearman correlation statistic is invariant under strictly increasing transformations. From the `R` Code for the Spearman correlation statistic above, $r_S = 0.45$ between the `ALAE` variable in logarithmic dollars and `LOSS` amount variable in dollars.

Example 16.1.1. Calculation by Hand. You are given the following six observations:

Observation	x **value**	y **value**
1	15	19
2	9	7
3	5	13
4	3	15
5	21	17
6	12	11

Calculate the sample Spearman's ρ.

Example Solution. The Spearman correlation is simply the product-moment correlation computed on the ranks:

$$r_S = \frac{\widehat{Cov}(R(X), R(Y))}{\sqrt{\widehat{Cov}(R(X), R(X))\widehat{Cov}(R(Y), R(Y))}} = \frac{\widehat{Cov}(R(X), R(Y))}{(n^2-1)/12}.$$

where $\widehat{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$.

Then we have:

Obs	x **value**	y **value**	**Rank of** x_i $(R(X))$	**Rank of** y_i $(R(Y))$	$R(X)_i - R(\bar{X})$	$R(Y)_i - R(\bar{Y})$	$(R(X)_i - R(\bar{X}))\times$ $(R(Y)_i - R(\bar{Y}))$
1	15	19.	2	1	$2-3.5=-1.5$	$1-3.5=-2.5$	3.75
2	9	7	4	6	$4-3.5=0.5$	$6-3.5=2.5$	1.25
3	5	13	5	4	$5-3.5=1.5$	$4-3.5=0.5$	0.75
4	3	15	6	3	$6-3.5=2.5$	$3-3.5=-0.5$	−1.25
5	21	17	1	2	$1-3.5=-2.5$	$2-3.5=-1.5$	3.75
6	12	11	3	5	$3-3.5=-0.5$	$5-3.5=1.5$	−0.75
Total							7.5

Note that: $R(\bar{X}) = R(\bar{Y}) = \frac{1+2+3+4+5+6}{6} = 3.5$.

Then,

$$\widehat{Cov}(R(X),R(Y)) = \frac{1}{n}\sum_{i=1}^{n}(R(X)_i - R(\bar{X}))(R(Y)_i - R(\bar{Y})) = \frac{7.5}{6} = 1.25.$$

Applying the formula

$$r_S = \frac{\widehat{Cov}(R(X),R(Y))}{(n^2-1)/12} = \frac{1.25}{(6^2-1)/12} = 0.42857.$$

Kendall's Tau

An alternative measure that uses ranks is based on the concept of *concordance*. An observation pair (X,Y) is said to be concordant (discordant) if the observation with a larger value of X has also the larger (smaller) value of Y. Then $\Pr(concordance) = \Pr[(X_1 - X_2)(Y_1 - Y_2) > 0]$, $\Pr(discordance) = \Pr[(X_1 - X_2)(Y_1 - Y_2) < 0]$, $\Pr(tie) = \Pr[(X_1 - X_2)(Y_1 - Y_2) = 0]$ and

$$\begin{aligned} \tau(X,Y) &= \Pr(concordance) - \Pr(discordance) \\ &= 2\Pr(concordance) - 1 + \Pr(tie). \end{aligned}$$

Thus, the population parameter Kendall's tau, $\tau = \tau(X,Y)$, measures whether higher values of one variable generally correspond to higher values of another variables, regardless of the actual values of those variables.

To estimate this, the pairs (X_i, Y_i) and (X_j, Y_j) are said to be concordant if the product $sgn(X_j - X_i)sgn(Y_j - Y_i)$ equals 1 and discordant if the product equals -1. Here, $sgn(x) = 1, 0, -1$ as $x > 0$, $x = 0$, $x < 0$, respectively. With this, we can express the (statistical) association measure of Kendall (1938), known as Kendall's tau, as

$$\begin{aligned} \hat{\tau} &= \tfrac{2}{n(n-1)}\textstyle\sum_{i<j}\ sgn(X_j - X_i) \times sgn(Y_j - Y_i) \\ &= \tfrac{2}{n(n-1)}\textstyle\sum_{i<j}\ sgn(R(X_j) - R(X_i)) \times sgn(R(Y_j) - R(Y_i)). \end{aligned}$$

Interestingly, Hougaard (2000), page 137, attributes the original discovery of this statistic to Fechner (1897), noting that Kendall's discovery was independent and more complete than the original work.

You can obtain Kendall's tau using the `cor()` function in `R` and selecting the `kendall` method. From below, $\hat{\tau} = 0.32$ between the `LOSS` variable in dollars and the `ALAE` variable in dollars. When there are ties in the data, the `cor()` function computes *Kendall's tau_b* as proposed by Kendall (1945).

Also, to show that the Kendall's tau is invariant under strictly increasing transformations, we see that $\hat{\tau} = 0.32$ between the `ALAE` variable in logarithmic dollars and the `LOSS` amount variable in dollars.

Example 16.1.2. Calculation by Hand. You are given the following six observations:

Observation	x **value**	y **value**
1	15	19
2	9	7
3	5	13
4	3	15
5	21	17
6	12	11

Calculate the sample Kendall's τ.

Example Solution. We can obtain the Kendall's tau using:

$$\hat{\tau} = \frac{2}{n(n-1)} \sum_{i<j} sgn(X_j - X_i) \times sgn(Y_j - Y_i)$$

Here, $sgn(x) = 1, 0, -1$ as $x > 0$, $x = 0$, $x < 0$, respectively. For each pair of observations i, j so that $i < j$, the pairs (X_i, Y_i) and (X_j, Y_j) are said to be concordant if the product $sgn(X_j - X_i)sgn(Y_j - Y_i)$ equals 1 and discordant if the product equals -1. This is summarized in the table below, where a 1 indicates concordance, -1 sign indicates discordance. Note: The pairs compared are in the upper triangle.

i/j	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$
$i=1$		1	1	1	-1	1
$i=2$			-1	-1	1	1
$i=3$				-1	1	-1
$i=4$					1	-1
$i=5$						1
$i=6$						

There are 9 concordant pairs and 6 discordant pairs. Applying the formula

$$\hat{\tau} = \frac{2}{n(n-1)} \sum_{i<j} sgn(X_j - X_i) \times sgn(Y_j - Y_i) = \frac{2}{6(6-1)}(3) = 0.2.$$

16.1.3 Tail Dependence Coefficients

Tail dependence is a statistical concept that measures the strength of the dependence between two variables in the tails of their distribution. Specifically, tail dependence measures the correlation or dependence between the extreme values of two variables beyond a certain threshold, that is, the dependence in the corner of the lower-left quadrant or upper-right quadrant of the bivariate distribution. Tail dependence is essential in many areas of finance, economics, and risk management. For example, it is relevant in analyzing extreme events, such as financial crashes, natural disasters, and pandemics. In these situations, tail dependence can help to determine the likelihood of joint extreme events occurring and to develop strategies to manage the associated risks.

In Figure 16.2, the concept of tail dependence is demonstrated through an example. The figure showcases two randomly generated variables with a Kendall's Tau of 0.7. On the left side of Figure 16.2, the variables (X and Y) are simulated using the bivariate normal distribution, while on the right side, they are generated using the bivariate t-distribution. Although both sides display a Kendall's Tau of 0.7, there is a difference in the upper right quadrant (above the dashed lines) on each panel. In the left panel, the values in the upper right quadrant (upper tails of X and Y) are independent, while in the right panel, the upper tail values appear to be correlated (the upper right corners of the right panel contain 4 points). This suggests that the probability of Y occurring above a high threshold (e.g., the dashed line in the figure) when X exceeds the same threshold is higher in the right panel than in the left panel of Figure 16.2.

Consider a pair of random variables (X, Y), from definitions provided in Joe (1997), the upper tail dependent coefficient denoted by λ_{up} is given by:

$$\lambda_{\text{up}} = \lim_{u \to 1^-} \Pr\left\{ X > F_X^{-1}(u) \mid Y > F_Y^{-1}(u) \right\},$$

in case the limit exists. Here, $F_X^{-1}(u)$ and $F_Y^{-1}(u)$ denote the quantiles of X and Y at the level u. Then, X and Y are said to be upper tail-dependent if $\lambda_{\text{up}} \in (0, 1]$ and upper tail-independent if $\lambda_{\text{up}} = 0$. When a variable reaches an extreme high value, the upper tail-dependent condition indicates that the other variable also reaches an extremely high value. On the other hand, the

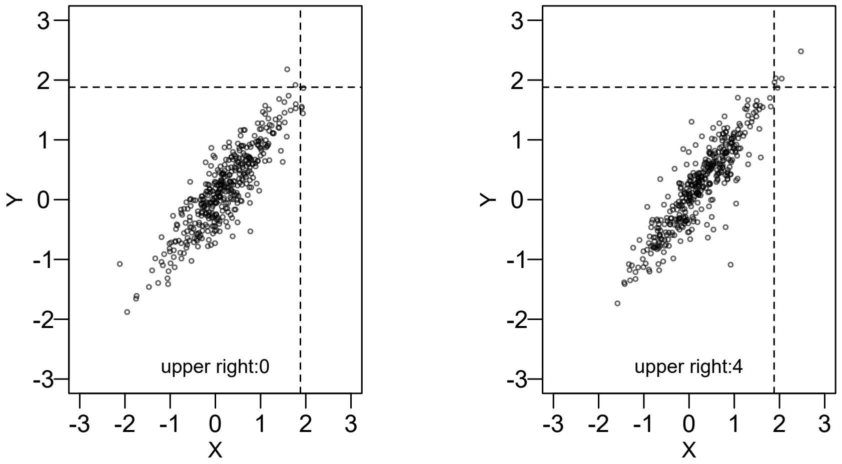


FIGURE 16.2: **Left Panel: Upper tails of X and Y are independent. Right Panel: Upper tails of X and Y appear to be dependent.**

upper tail-independent suggests that the extreme values of the two variables are not related to each other. Similarly, the lower tail dependence coefficient, λ_{lo}, is defined as:

$$\lambda_{\text{lo}}= \lim_{u\to 0^+} \Pr\left\{X \le F_X^{-1}(u) \mid Y \le F_Y^{-1}(u)\right\}.$$

Let $R(X_j)$ and $R(Y_j)$ denote the rank of X_j and Y_j, $j = 1, \dots, n$, respectively. From Schmidt (2005), non-parametric estimates of λ_{up} and λ_{lo} are given by:

$$\hat{\lambda}_{\text{lo}} = \frac{1}{k}\sum_{j=1}^{n} I\left\{R(X_j) \le k, R(Y_j) \le k\right\},$$

and

$$\hat{\lambda}_{\text{up}} = \frac{1}{k}\sum_{j=1}^{n} I\left\{R(X_j) > n-k, R(Y_j) > n-k\right\},$$

where $k \in 1, ..., n$ is the threshold rank and a parameter to be chosen by the analyst, $k = k(n) \to \infty$ and $k/n \to 0$ as $n \to \infty$, . Here, n is the sample size, and $I\{\cdot\}$ takes the value of 1 if the condition is satisfied, and 0 otherwise.

Figure 16.3 shows the scatter plot of the ranks of the `LOSS` variable and the `ALAE` variable. You can obtain the upper and lower tail dependent coefficient using the `tdc()` function from the `FRAPO` package in `R`. From below, $\hat{\lambda}_{\text{up}} = 0.39$, at

$k = 75$ (note that $n = 1500$), between the `LOSS` variable and the `ALAE` variable and $\hat{\lambda}_{\text{lo}} = 0.13$. The results implies the losses and expenses variables appear to be more upper-tailed dependent.

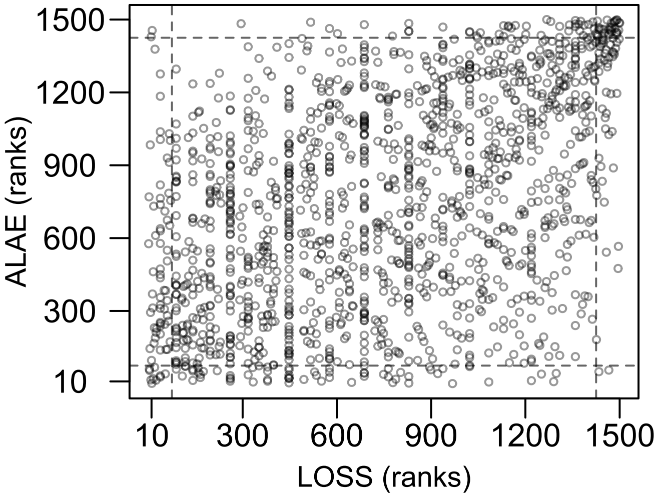


FIGURE 16.3: **Scatter Plot of Ranks of LOSS and ALAE**

16.2 Introduction to Copulas

In this section, you learn how to:

- Describe a multivariate distribution function in terms of a copula function.

16.2.1 Definition of a Copula

Copulas are widely used in insurance and many other fields to model the dependence among multivariate outcomes as they expresses the dependence between the variables explicitly. Recall that the joint cumulative distribution function (*cdf*) for two variables Y_1 and Y_2 is given by:

$$F(y_1, y_2) = \Pr(Y_1 \le y_1, Y_2 \le y_2).$$

For the multivariate case in p dimensions, we have:

$$F(y_1, \ldots, y_p) = \Pr(Y_1 \le y_1, \ldots, Y_p \le y_p).$$

The joint distribution considers both the marginal distributions and how the variables are related to each other. However, it expresses this dependence implicitly. Copulas offer a different method that allows us to break down the joint distribution of variables into individual components (the marginal distributions and a copula) that can be adjusted separately.

A copula is a multivariate distribution function with uniform marginals. Specifically, let $\{U_1, \ldots, U_p\}$ be p uniform random variables on $(0, 1)$. Their distribution function

$$C(u_1, \ldots, u_p) = \Pr(U_1 \le u_1, \ldots, U_p \le u_p),$$

is a copula. We seek to use copulas in applications that are based on more than just uniformly distributed data. Thus, consider arbitrary marginal distribution functions $F_1(y_1)$,. . . ,$F_p(y_p)$. Then, we can define a multivariate distribution function using the copula such that

$$F(y_1, \ldots, y_p) = C(F_1(y_1), \ldots, F_p(y_p)). \tag{16.1}$$

Here, F is a multivariate distribution function, and the resulting value from the copula function is limited to a range of $[0, 1]$ as it relates to probabilities. Sklar (1959) showed that *any* multivariate distribution function F, can be written in the form of equation (16.1), that is, using a copula representation.

Sklar also showed that, if the marginal distributions are continuous, then there is a unique copula representation. Hence, copulas can be used instead of joint distribution functions. In order to be considered valid, they must meet the necessary requirements of a valid joint cumulative distribution function. In this chapter we focus on copula modeling with continuous variables. A copula C is considered to be **absolutely continuous** if the **density**

$$c(u_1, \ldots, u_p) = \frac{\partial^p}{\partial_{u_p} \ldots \partial_{u_1}} C(u_1, \ldots, u_p),$$

exists. For the discrete case, readers can see Joe (2014) and Genest and Nešlohva (2007). For the bivariate case where $p = 2$, we can write a copula and the distribution function of two random variables as

$$C(u_1,\, u_2) = \Pr(U_1 \le u_1,\, U_2 \le u_2)$$

and

$$F(y_1,\, y_2) = C(F_1(y_1), F_p(y_2)).$$

One example of a bivariate copula is the product copula, also called the independence copula, as it captures the property of independence of the two variables Y_1 and Y_2. The copula (distribution function) is

$$F(y_1,\, y_2) = C(F_1(y_1), F_p(y_2)) = F_1(y_1)F_p(y_2) = u_1 u_2 = \Pi(u).$$

In Figure 16.4, both the distribution function and scatter plot of observations generated from the independence copula are displayed. The scatter plot indicates that there is no correlation between the two components, U_1 and U_2.

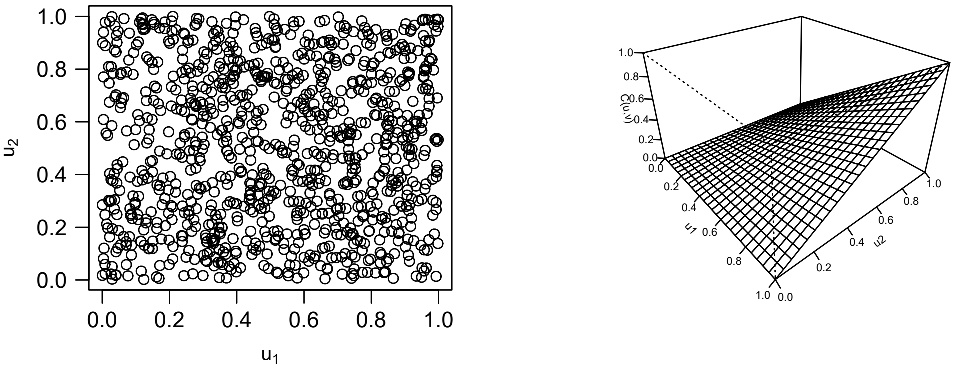


FIGURE 16.4: **Independence Copula.** Left: Scatterplot of observations from Independence Copula. Right: Plot for distribution function for Independence Copula.

There is another type of copula that is frequently utilized, known as Frank's Copula (Frank, 1979). This copula can represent both positive and negative dependence and has a straightforward analytic structure. The copula (distribution function) is

$$C(u_1, u_2) = \frac{1}{\gamma} \log\left(1 + \frac{(\exp(\gamma u_1) - 1)(\exp(\gamma u_2) - 1)}{\exp(\gamma) - 1}\right). \tag{16.2}$$

This is a bivariate distribution function with its domain on the unit square $[0, 1]^2$. Here γ is the dependence parameter, that is, the range of dependence is controlled by the parameter γ. Positive association increases as γ increases. As we will see, this positive association can be summarized with Spearman's rho (ρ_S) and Kendall's tau (τ).

In Figure 16.5, we can see scatterplots of data generated from the Frank's copula. As γ value changes, we observe that components U_1 and U_2 become positively or negatively dependent. When θ approaches 0, (16.2) transforms into an independence copula. Also, Figure 16.6 provides the distribution and density functions for Frank's copula when $\gamma = 12$. In Section 16.4, we will explore copula functions other than the commonly used Frank's copula.

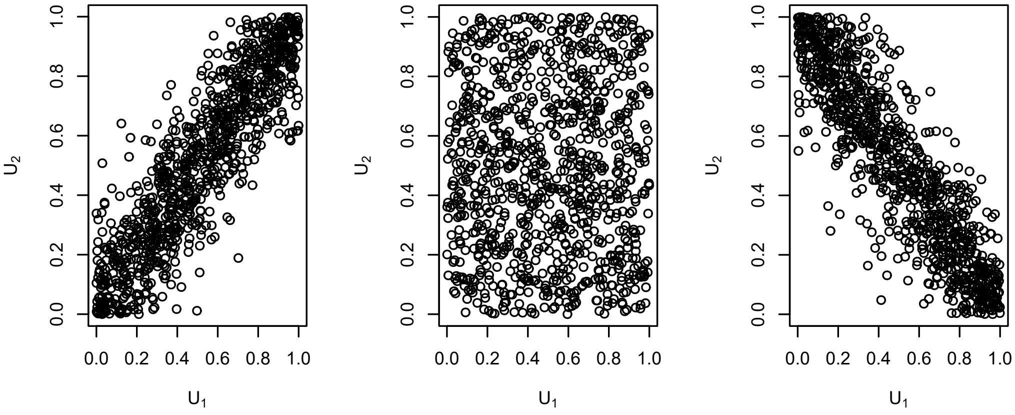


FIGURE 16.5: **Scatterplot of Observations from Frank's Copula**. $\gamma = 12$ (left), $\gamma = 0$ (middle) and $\gamma = -12$ (right).

Example 16.2.1. Copula Representation Example

Suppose we have a variable X that follows a Pareto distribution with a scale parameter of $\theta = 10$ and a shape parameter of $\alpha = 1.6$. Additionally, let Y be an exponential variable with a mean value of 8. Write $F_{X,Y}(7.2, 4.1)$ in the form $C(u, v)$.

Note: $F_{X,Y}(x, y)$ is the joint distribution function and $C(u, v)$ is the copula that links X and Y.

Example Solution. Denote the marginal distribution functions of X and Y as $F_X(x)$ and $F_Y(y)$, respectively. Since $F_{X,Y}(7.2, 4.1) = C[F_X(7.2), F_Y(4.1)]$, we can use the marginal distribution functions to obtain the arguments of the copula function:

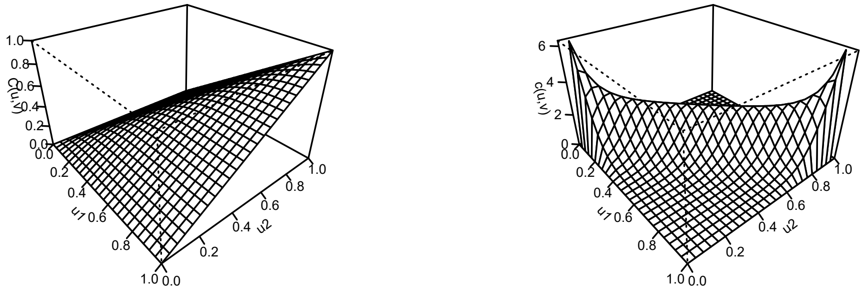


FIGURE 16.6: Left: Plot for distribution function for Frank's Copula ($\gamma = 12$). Right: Plot for the density function for Frank's Copula ($\gamma = 12$).

For the Pareto variable, X:

$$F_X(7.2) = 1 - \left(\frac{10}{7.2 + 10}\right)^{1.6} = 0.58;$$

For the Exponential variable, Y:

$$F_Y(4.1) = 1 - e^{-0.125 \times 4.1} = 0.40.$$

Hence, $F_{X,Y}(7.2, 4.1) = C[0.58, 0.40]$.

16.2.2 Sklar's Theorem

In Sklar (1959), Sklar showcased how copulas can capture the dependence structure of a group of random variables. This principle has since been referred to as Sklar's Theorem and serves as the cornerstone of copula theory. Dependence modeling with copulas for continuous multivariate distributions allows for the separation of modeling the univariate marginals and the dependence structure, where a copula can represent the dependence structure.

1. For a p-variate distribution F, with marginal cumulative distribution

functions $F_1, \ldots, F_p$, the copula associated with F is a distribution function $C : [0,1]^p \to [0,1]$ with $U(0,1)$ margins that satisfies:

$$F(\mathbf{y}) = C(F_1(y_1), \ldots, F_p(y_p)), \quad \mathbf{y} = \{y_1 \ldots y_p\} \in \mathbb{R}^p. \tag{16.3}$$

If F is a continuous p-variate distribution function with univariate margins $F_1, \ldots, F_p$ and quantile functions $F_1^{-1}, \ldots, F_p^{-1}$, then:

$$C(\mathbf{u}) = F\left(F_1^{-1}(u_1), \ldots, F_p^{-1}(u_p)\right), \quad \mathbf{u} \in [0,1]^p,$$

is the unique choice.

2. The converse also holds: If C is a copula and $F_1, \ldots, F_p$ are univariate cumulative distribution functions, then the function F defined by (16.3) is a joint cumulative distribution function with marginal cumulative distribution functions $F_1, \ldots, F_p$.

Proof. Suppose we have $Y_1, \ldots, Y_p \sim F$, $U \sim \mathrm{U}(0,1)$ and assume $Y_1, \ldots, Y_p$ is continuous, which implies $F_1^{-1}(U_i) = Y_i$. Then, the random variables have a multivariate distribution function C, given by:

$$\begin{aligned} C(u_1, \ldots, u_p) &= \Pr(U_1 \le u_1, \ldots, U_p \le u_p) \\ &= \Pr(F_1(Y_1) \le u_1, \ldots, F_p(Y_p) \le u_p) \\ &= \Pr\left(Y_1 \le F_1^{-1}(u_1), \ldots, Y_p \le F_p^{-1}(u_p)\right) \\ &= F\left(F_1^{-1}(u_1), \ldots, F_p^{-1}(u_p)\right) \\ &= F(y_1, \ldots, y_p). \end{aligned}$$

Hence:

$$C(F_1(y_1), \ldots, F_p(y_p)) = F(y_1, \ldots, y_p).$$

According to the first part of Sklar's theorem, there is a unique underlying copula that is unknown, and it can be estimated from the data available. After estimating the margins and copula, they are usually combined using (16.3) to give the estimated multivariate distribution function. Also, Sklar's theorem's second part enables the construction of adaptable multivariate distribution functions with specified univariate margins. These functions are useful in more intricate models, like pricing models.

16.3 Application Using Copulas

In this section, you learn how to:

- Discover dependence structure between random variables
- Model the dependence with a copula function

This section analyzes the insurance losses and expenses data with the statistical program R. The data set is visualized in Figure 16.1. The model fitting process is started by marginal modeling of each of the two variables, LOSS and ALAE. Then we model the joint distribution of these marginal outcomes.

16.3.1 Marginal Models

We first examine the marginal distributions of losses and expenses before going through the joint modeling. The histograms show that both LOSS and ALAE are right-skewed and fat-tailed. Because of these features, for both marginal distributions of losses and expenses, we consider a Pareto distribution, distribution function of the form

$$F(y) = 1 - \left(\frac{\theta}{y+\theta} \right)^{\alpha} .$$

Here, θ is a scale parameter and α is a shape parameter. Section 20.2 provides details of this distribution.

The marginal distributions of losses and expenses are fit using the method of maximum likelihood. Specifically, we use the vglm function from the R VGAM package. Firstly, we fit the marginal distribution of ALAE. Parameters are summarized in Table 16.6.

We repeat this procedure to fit the marginal distribution of the LOSS variable. Because the loss variable also seems right-skewed and heavy-tailed data, we also model the marginal distribution with the Pareto distribution (although with different parameters).

Table 16.6. Summary of Pareto Maximum Likelihood Fitted Parameters from the LGPIF Data

	Shape $\hat{\theta}$	Scale $\hat{\alpha}$
ALAE	15133.60360	2.22304
LOSS	16228.14797	1.23766

To visualize the fitted distribution of LOSS and ALAE variables, one can use the estimated parameters and plot the corresponding distribution function and density function. For more details on the selection of marginal models, see Chapter 6.

16.3.2 Probability Integral Transformation

When studying simulation, in Section 8.1.2 we learned about the inverse transform method. This is a way of mapping a $U(0,1)$ random variable into a random variable X with distribution function F via the inverse of the distribution, that is, $X = F^{-1}(U)$. The probability integral transformation goes in the other direction, it states that $F(X) = U$. Although the inverse transform result is available when the underlying random variable is continuous, discrete or a hybrid combination of the two, the probability integral transform is mainly useful when the distribution is continuous. That is the focus of this chapter.

We use the probability integral transform for two purposes: (1) for diagnostic purposes, to check that we have correctly specified a distribution function and (2) as an input into the copula function in equation (16.1).

For the first purpose, we can check to see whether the Pareto is a reasonable distribution to model our marginal distributions. Given the fitted Pareto distribution, the variable ALAE is transformed to the variable u_1, which follows a uniform distribution on $[0,1]$:

$$u_1 = \hat{F}_1(ALAE) = 1 - \left(\frac{\hat{\theta}}{\hat{\theta} + ALAE}\right)^{\hat{\alpha}}.$$

After applying the probability integral transformation to the ALAE variable, we plot the histogram of *Transformed* ALAE in Figure 16.7. This plot appears reasonably close to what we expect to see with a uniform distribution, suggesting that the Pareto distribution is a reasonable specification.

In the same way, the variable LOSS is also transformed to the variable u_2, which follows a uniform distribution on $[0,1]$. The left-hand panel of Figure 16.8 shows a plot the histogram of *Transformed* ALAE, again reinforcing the Pareto distribution specification. For another way of looking at the data, the variable u_2 can be transformed to a *normal score* with the quantile function of standard normal distribution. As we see in Figure 16.8, normal scores of the variable LOSS are approximately marginally standard normal. This figure is helpful because analysts are used to looking for patterns of approximate normality (which seems to be evident in the figure). The logic is that, if the Pareto distribution is correctly specified, then transformed losses u_2 should be

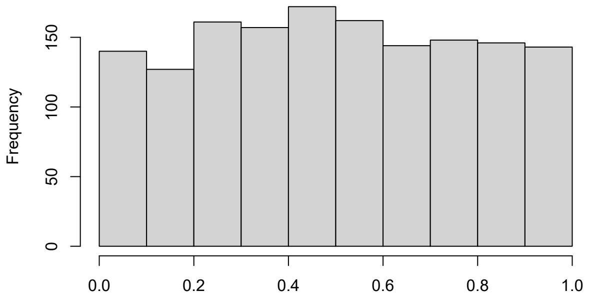


FIGURE 16.7: **Histogram of Transformed ALAE**

approximately normal, and the normal scores $\Phi^{-1}(u_2)$, should be approximately normal. (Here, Φ is the cumulative standard normal distribution function.)

16.3.3 Joint Modeling with Copula Function

Before jointly modeling losses and expenses, we draw the scatterplot of transformed variables (U_1, U_2) and the scatterplot of normal scores in Figure 16.9. The left-hand panel is a plot of U_1 versus U_2, where $U_1 = \hat{F}_1(ALAE)$ and $U_2 = \hat{F}_2(LOSS))$. Then we transform each one using an inverse standard normal distribution function, $\Phi^{-1}(\cdot)$, or `qnorm` in `R` to get normal scores. As in Figure 16.1, it is difficult to see patterns in the left-hand panel. However, with rescaling, patterns are evident in the right-hand panel. To learn more details about normal scores and their applications in copula modeling, see Joe (2014).

The right-hand panel of Figure 16.1 shows us there is a positive dependency between these two random variables. This can be summarized using, for example, Spearman's rho that turns out to be 0.451. As we learned in Section 16.1.2, this statistic depends only on the order of the two variables through their respective ranks. Therefore, the statistic is the same for (1) the original data in Figure 16.1, (2) the data transformed to uniform scales in the left-hand panel of Figure 16.9, and (3) the normal scores in the right-hand panel of Figure 16.9.

The next step is to calculate estimates of the copula parameters. One option is to use traditional maximum likelihood and determine all the parameters at the same time which can be computationally burdensome. Even in our simple example, this means maximizing a (log) likelihood function over five parameters, two for the marginal `ALAE` distribution, two for the marginal `LOSS` distribution, and one for the copula. A widely alternative, known as the

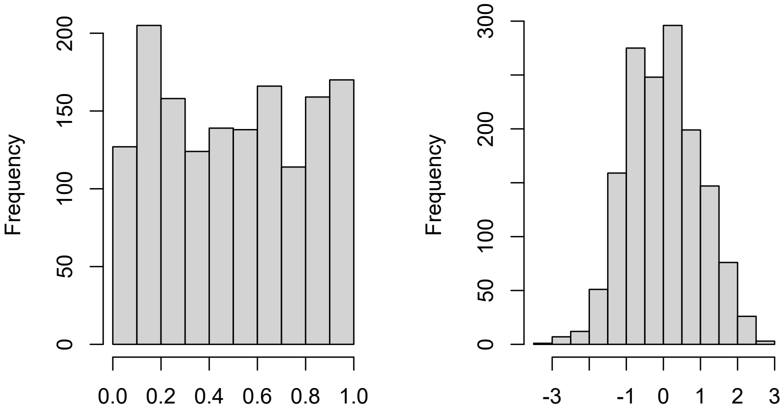


FIGURE 16.8: **Histogram of Transformed Loss.** The left-hand panel shows the distribution of probability integral transformed losses. The right-hand panel shows the distribution for the corresponding normal scores.

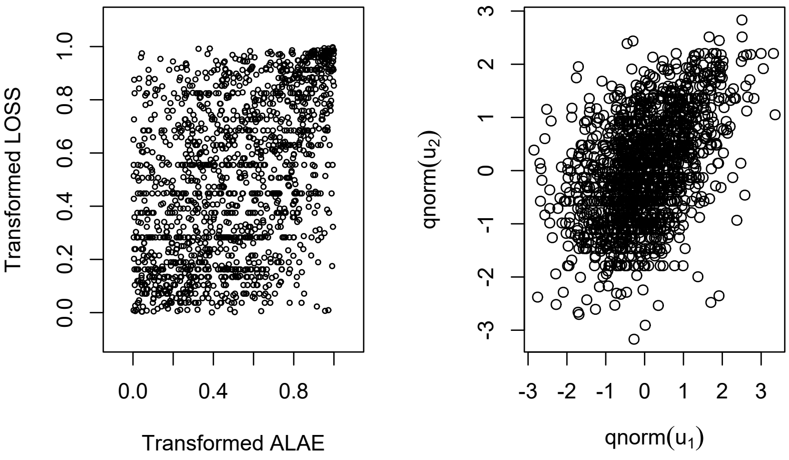


FIGURE 16.9: Left: Scatter plot for transformed variables. Right:Scatter plot for normal scores

inference for margins (IFM) approach, is to simply use the fitted marginal distributions, u_1 and u_2, as inputs when determining the copula. This is the approach taken here. In the following code, you will see that the fitted copula parameter becomes $\hat{\gamma} = 3.114$.

To visualize the fitted Frank's copula, the distribution function and density function perspective plots are drawn in Figure 16.10.

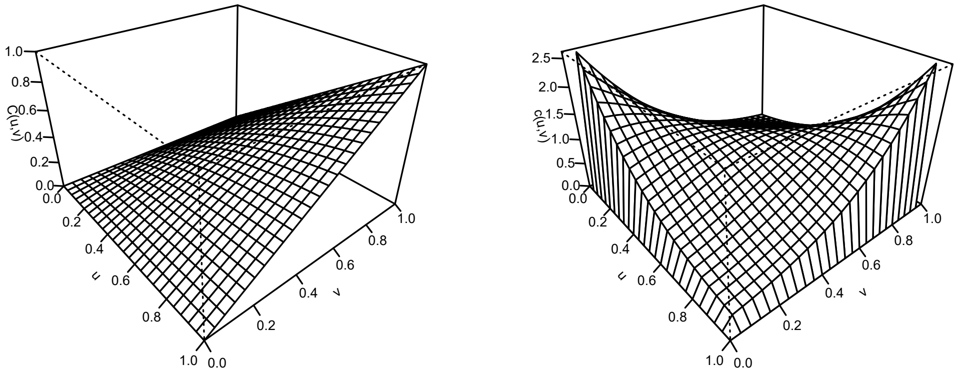


FIGURE 16.10: **Frank's Copula**. Left: Plot for distribution function for Frank's Copula. Right:Plot for density function for Frank's Copula

We can estimate the anticipated expenses when losses surpass a specific threshold by utilizing the fitted Frank copula based on the data on losses and expenses. For instance, according to the data, the mean expense when losses exceed $\$200,000$ is $\$58,807$. However, when we apply the fitted Frank copula, the projected expenses when losses exceed $\$200,000$ is $\$26,767$. This suggests that the Frank copula doesn't provide an accurate estimate and may not be suitable for this dataset. We will now explore other copula types.

16.4 Types of Copulas

In this section, you learn how to:

- Define the basic types of elliptical copulas, including the normal, t
- Define basic types of Archimedean copulas

There are several families of copulas that have been described in the literature. Two main families of the copula families are the **Archimedean** and **Elliptical** copulas.

16.4.1 Normal (Gaussian) Copulas

We started our study with Frank's copula in equation (16.2) because it can capture both positive and negative dependence and has a readily understood analytic form. However, extensions to multivariate cases where $p > 2$ are not easy and so we look to alternatives. In particular, the normal, or Gaussian, distribution has been used for many years in empirical work, starting with Gauss in 1887. So, it is natural to turn to this distribution as a benchmark for understanding multivariate dependencies.

For a multivariate normal distribution, think of p normal random variables, each with mean zero and standard deviation one. Their dependence is controlled by $\mathbf{\Sigma}$, a correlation matrix, with ones on the diagonal. The number in the ith row and jth column, say $\mathbf{\Sigma}_{ij}$, gives the correlation between the ith and jth normal random variables. This collection of random variables has a multivariate normal distribution with probability density function

$$\phi_N(\mathbf{z}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \mathbf{\Sigma}}} \exp\left(-\frac{1}{2}\mathbf{z}'\mathbf{\Sigma}^{-1}\mathbf{z}\right). \tag{16.4}$$

To develop the corresponding copula version, it is possible to start with equation (16.1), evaluate this using normal variables, and go through a bit of calculus. Instead, we simply state as a definition, the normal (Gaussian) **copula** density function is

$$c_N(u_1, \ldots, u_p) = \phi_N\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p)\right) \prod_{j=1}^{p} \frac{1}{\phi(\Phi^{-1}(u_j))}.$$

Here, we use Φ and ϕ to denote the standard normal distribution and density functions. Unlike the usual probability density function ϕ_N, the copula density function has its domain on the hyper-cube $[0,1]^p$. For contrast, Figure 16.11 compares these two density functions.

16.4.2 *t*- and Elliptical Copulas

Another copula used widely in practice is the t- copula. Both the t- and the normal copula are special cases of a family known as *elliptical* copulas, so we introduce this general family first, then specialize to the case of the t- copula.

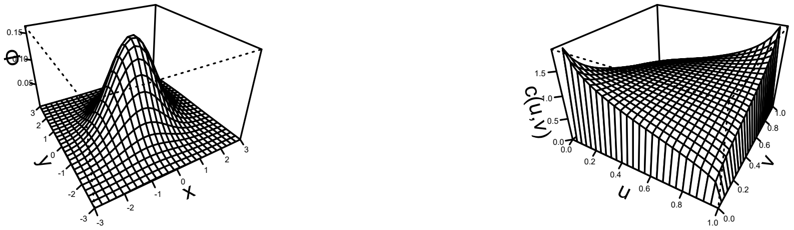


FIGURE 16.11: **Bivariate Normal Probability Density Function Plots.** The left-hand panel is a traditional bivariate normal probability density function. The right-hand plot is a plot of the copula density for the normal distribution.

The normal and the t- distributions are examples of symmetric distributions. More generally, elliptical distributions is a class of distributions that are symmetric and can be multivariate. In short, an elliptical distribution is a type of symmetric, multivariate distribution. The multivariate normal and multivariate t- are special types of elliptical distributions.

Elliptical copulas are constructed from elliptical distributions. This copula decomposes a (multivariate) elliptical distribution into their univariate elliptical marginal distributions by Sklar's theorem. Properties of elliptical copulas can be obtained from the properties of the corresponding elliptical distributions, see for example, Hofert et al. (2018).

In general, a p-dimensional vector of random variables has an *elliptical distribution* if the density can be written as

$$h_E(\mathbf{z}) = \frac{k_p}{\sqrt{\det \boldsymbol{\Sigma}}} g_p \left(\frac{1}{2} (\mathbf{z} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}) \right),$$

for $\mathbf{z} \in R^p$ and k_p is a constant, determined so the density integrates to one. The function $g_p(\cdot)$ is called a *generator* because it can be used to produce different distributions. Table 16.7 summarizes a few choices used in actuarial practice. The choice $g_p(x) = \exp(-x)$ gives rises to the normal *pdf* in equation (16.4). The choice $g_p(x) = \exp(-(1+2x/r)^{-(p+r)/2})$ gives rise to a multivariate

t- distribution with r degrees of freedom with *pdf*

$$h_{t_r}(\mathbf{z}) = \frac{k_p}{\sqrt{\det \boldsymbol{\Sigma}}} \exp\left[- \left(1 + \frac{(\mathbf{z}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})}{r} \right)^{-(p+r)/2} \right].$$

Table 16.7. Generator Functions ($g_p(\cdot)$) for Selected Elliptical Distributions

	Generator
Distribution	$g_p(x)$
Normal distribution	e^{-x}
$t-$ distribution with r degrees of freedom	$(1+2x/r)^{-(p+r)/2}$
Cauchy	$(1+2x)^{-(p+1)/2}$
Logistic	$e^{-x}/(1+e^{-x})^2$
Exponential power	$\exp(-rx^s)$

We can use elliptical distributions to generate copulas. Because copulas are concerned primarily with relationships, we may restrict our considerations to the case where $\mu = \mathbf{0}$ and $\boldsymbol{\Sigma}$ is a correlation matrix. With these restrictions, the marginal distributions of the multivariate elliptical copula are identical; we use H to refer to this marginal distribution function and h is the corresponding density. This marginal density is $h(z) = k_1 g_1(z^2/2)$. For example, in the normal case we have $H(\cdot) = \Phi(\cdot)$ and $h(\cdot) = \phi(\cdot)$.

We are now ready to define the *pdf* of the *elliptical copula*, a function defined on the unit cube $[0,1]^p$ as

$$c_E(u_1, \ldots, u_p) = h_E\left(H^{-1}(u_1), \ldots, H^{-1}(u_p)\right) \prod_{j=1}^{p} \frac{1}{h(H^{-1}(u_j))}.$$

As noted above, most empirical work focuses on the normal copula and t-copula. Specifically, t-copulas are useful for modeling the dependency in the tails of bivariate distributions, especially in financial risk analysis applications. The t-copulas with same association parameter in varying the degrees of freedom parameter show us different tail dependency structures. For more information about t-copulas, readers can see Joe (2014) and Hofert et al. (2018).

We used the same approach as with the fitted Frank copula to fit the Normal and t copula. The `R` code below fits the Normal and t copula and estimates the expected level of expenses when losses exceed \$200,000. The results show that the estimated expenses using the fitted Normal copula when losses exceed \$200,000 is \$35,411. However, this is not a good fit compared to the mean expense of $\$58,807$ from the losses and expenses data. When losses exceed $\$200,000$, the fitted t copula estimates expenses to be $\$47,354$, making it a better fit than the Normal copula.

16.4.3 Archimedean Copulas

This class of copulas is also constructed from a *generator* function. For Archimedean copulas, we assume that $g(\cdot)$ is a convex, decreasing function with domain [0,1] and range $[0,\infty)$ such that $g(0) = 0$. Use g^{-1} for the inverse function of g. Then the function

$$C_g(u_1, \ldots, u_p) = g^{-1}\left(g(u_1) + \cdots + g(u_p)\right)$$

is said to be an *Archimedean* copula distribution function.

For the bivariate case, $p = 2$, an Archimedean copula function can be written by the function

$$C_g(u_1,\, u_2) = g^{-1}\left(g(u_1) + g(u_2)\right).$$

Some important special cases of Archimedean copulas include the Frank, Clayton/Cook-Johnson, and Gumbel/Hougaard copulas. Each copula class is derived from different generator functions. As another useful special case, recall the Frank's copula described in Sections 16.2 and 16.3. To illustrate, we now provide explicit expressions for the Clayton and Gumbel/Hougaard copulas.

Clayton Copula

For $p = 2$, the Clayton copula with parameter $\gamma \in [-1, \infty)$ is defined by

$$C_\gamma^C(u) = \max\{u_1^{-\gamma} + u_2^{-\gamma} - 1, 0\}^{1/\gamma}, \quad u \in [0,1]^2.$$

This is a bivariate distribution function defined on the unit square $[0,1]^2$. The range of dependence is controlled by the parameter γ, similar to Frank's copula.

Gumbel-Hougaard Copula

The Gumbel-Hougaard copula is parametrized by $\gamma \in [1, \infty)$ and defined by

$$C_\gamma^{GH}(u) = \exp\left(-\left(\sum_{i=1}^{2}(-\log u_i)^\gamma\right)^{1/\gamma}\right), \quad u \in [0,1]^2.$$

For more information on Archimedean copulas, see Joe (2014), Frees and Valdez (1998), and Genest and Mackay (1986).

We used the same approach to fit the Clayton and Gumbel-Hougaard copulas as we did for the fitted Frank copula. The `R` code below fits these two copulas and determines the expected expense level for losses higher than $\$200,000$. Our analysis shows that the estimated expenses for losses exceeding $\$200,000$ using the fitted Clayton copula are $\$14,209$, while the fitted Gumbel-Hougaard copula predicts $\$58,554$. Of all the copula types considered, the Gumbel-Hougaard copula provides the best fit for this data. For more on Goodness-of-fit tests, see Hofert et al. (2018).

16.5 Properties of Copulas

In this section, you learn how to:

- Interpret bounds that limit copula distribution functions as the amount of dependence varies
- Calculate measures of association for different copulas and interpret their properties
- Interpret tail dependency for different copulas

With many choices of copulas available, it is helpful for analysts to understand general features of how these alternatives behave.

16.5.1 Bounds on Association

Any distribution function is bounded below by zero and from above by one. Additional types of bounds are available in multivariate contexts. These bounds are useful when studying dependencies. That is, as an analyst thinks about variables as being extremely dependent, one has available bounds that cannot be exceeded, regardless of the dependence. The most widely used bounds in dependence modeling are known as the *Fréchet-Höeffding* bounds, given as

$$\max(u_1 + \cdots + u_p - p + 1, 0) \leq C(u_1, \ldots, u_p) \leq \min(u_1, \ldots, u_p).$$

To see the right-hand side of this equation, note that

$$C(u_1, \ldots, u_p) = \Pr(U_1 \leq u_1, \ldots, U_p \leq u_p) \leq \Pr(U_j \leq u_j),$$

for $j = 1, \ldots, p$. The bound is achieved when $U_1 = \cdots = U_p$. To see the left-hand side when $p = 2$, consider $U_2 = 1 - U_1$. In this case, if $1 - u_2 < u_1$ then

$$\Pr(U_1 \leq u_1, U_2 \leq u_2) = \Pr(1 - u_2 \leq U_1 < u_1) = u_1 + u_2 - 1.$$

See, for example, Nelson (1997) for additional discussion.

To see how these bounds relate to the concept of dependence, consider the case of $p = 2$. As a benchmark, first note that the product copula, $C(u_1, u_2) = u_1 \cdot u_2$, is the result of assuming independence between random variables. Now, from the above discussion, we see that the lower bound is achieved when the two random variables are perfectly negatively related ($U_2 = 1 - U_1$). Further, it is clear that the upper bound is achieved when they are perfectly positively related ($U_2 = U_1$). To emphasize this, the Frechet-Hoeffding bounds for two random variables appear in Figure 16.12.

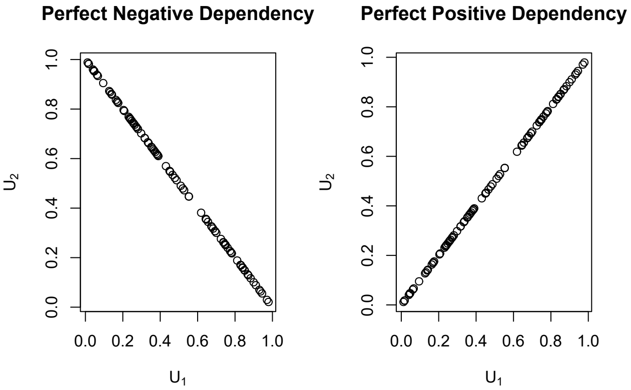


FIGURE 16.12: **Perfect Positive and Perfect Negative Dependence Plots**

Let's assign the Fréchet-Höeffding lower bound as W and the upper bound as M. That is, $W = \max(u_1 + \cdots + u_p - p + 1, 0)$ and $M = \min(u_1, \ldots, u_p)$. It's important to note that W is a copula only if $p = 2$, while M is a copula for all $p \geq 2$. In dimension two, $W = \max(u_1 + u_2 - 1, 0)$ is known as the **counter-monotonic copula**. It captures the inverse relationship between two variables, that is, two random variables that are perfectly negatively related. On the other hand, $M = \min(u_1, u_2)$ in dimension two is known as the **comonotone copula**. It captures the relationship between two variables where one is related to the other by a strictly increasing function, that is, two random variables that are perfectly positively dependent. The co-monotonic copulas can be extended to the multivariate case. However, it's not possible to extend the counter-monotonic copula because it's not possible to have three or more variables where each pair has a direct inverse relationship.

Example 16.5.1. Largest Possible Value Example

Suppose we have a variable X that follows a Pareto distribution with a scale parameter of $\theta = 10$ and a shape parameter of $\alpha = 1.6$. Additionally, let Y be an exponential variable with a mean value of 8. Let $F_{X,Y}(x, y)$ be the joint distribution function. What is the largest possible value of $F_{X,Y}(7.2, 4.1)$?

Example Solution. Let $C(u, v)$ is the copula that links X and Y. Denote the marginal distribution functions of X and Y as $F_X(x)$ and $F_Y(y)$, respectively. Since $F_{X,Y}(7.2, 4.1) = C[F_X(7.2), F_Y(4.1)]$, we can use the marginal distribution functions to obtain the arguments of the copula function:

For the Pareto variable, X:

$$F_X(7.2) = 1 - \left(\frac{10}{7.2 + 10}\right)^{1.6} = 0.58.$$

For the Exponential variable, Y:

$$F_Y(4.1) = 1 - e^{-0.125\times 4.1} = 0.40.$$

Hence, $F_{X,Y}(7.2, 4.1) = C[0.58, 0.40]$. Now:

The largest value of the joint distribution function is obtained when the dependence structure is comonotonic, or $C(u, v) = \min(u, v)$. Hence, the answer required is $\min(0.58, 0.40) = 0.40$.

16.5.2 Measures of Association

Empirical versions of Spearman's rho and Kendall's tau were introduced in Section 16.1.2, respectively. The interesting thing about these expressions is that these summary measures of association are based **only** on the ranks of each variable. Thus, any strictly increasing transform does not affect these measures of association. Specifically, consider two random variables, Y_1 and Y_2, and let m_1 and m_2 be strictly increasing functions. Then, the association, when measured by Spearman's rho or Kendall's tau, between $m_1(Y_1)$ and $m_2(Y_2)$ does not change regardless of the choice of m_1 and m_2. For example, this allows analysts to consider dollars, Euros, or log dollars, and still retain the same essential dependence. As we have seen in Section 16.1, this is not the case with the Pearson's measure of correlation.

Schweizer et al. (1981) established that the copula accounts for all the dependence in the sense that the way Y_1 and Y_2 "move together" is captured by the copula, regardless of the scale in which each variable is measured. They also showed that (population versions of) the two standard nonparametric measures of association could be expressed solely in terms of the copula function. Spearman's correlation coefficient is given by

$$\rho_S = 12 \int_0^1 \int_0^1 \{C(u,v) - uv\}\, dudv. \tag{16.5}$$

Kendall's tau is given by

$$\tau = 4 \int_0^1 \int_0^1 C(u,v)\ dC(u,v) - 1.$$

For these expressions, we assume that Y_1 and Y_2 have a jointly continuous distribution function.

Example. Loss versus Expenses. Earlier, in Section 16.3, we saw that the Spearman's correlation was 0.452, calculated with the `rho` function. Then, we fit Frank's copula to these data, and estimated the dependence parameter to be $\hat{\gamma} = 3.114$. As an alternative, the following code shows how to use the empirical version of equation (16.5). In this case, the Spearman's correlation coefficient is 0.462, which is close to the sample Spearman's correlation coefficient, 0.452.

16.5.3 Tail Dependency

As discussed in Section 16.1.3, there are applications in which it is useful to distinguish the part of the distribution in which the association is strongest. For example, in insurance it is helpful to understand association among the largest losses, that is, association in the right tails of the data. This subsection defines upper and lower tail dependency in terms of copulas.

To capture this type of dependency, we use the *right-tail concentration* function, defined as

$$R(z) = \frac{\Pr(U_1 > z, U_2 > z)}{1-z} = \Pr(U_1 > z | U_2 > z) = \frac{1 - 2z + C(z,z)}{1-z}.$$

As a benchmark, $R(z)$ will be equal to z under independence. Joe (1997) uses the term "upper tail dependence parameter" for $R = \lim_{z \to 1} R(z)$.

In the same way, one can define the *left-tail concentration* function as

$$L(z) = \frac{\Pr(U_1 \le z, U_2 \le z)}{z} = \Pr(U_1 \le z | U_2 \le z) = \frac{C(z,z)}{z},$$

with the lower tail dependence parameter $L = \lim_{z \to 0} L(z)$. A tail dependency concentration function captures the probability of two random variables simultaneously having extreme values.

It is of interest to see how well a given copula can capture tail dependence. To this end, we calculate the left and right tail concentration functions for four different types of copulas; Normal, Frank, Gumbel and t- copulas. The results are summarized for concentration function values for these four copulas in Table 16.8. As in Venter (2002), we show $L(z)$ for $z \leq 0.5$ and $R(z)$ for $z > 0.5$ in the tail dependence plot in Figure 16.13. We interpret the tail dependence plot to mean that both the Frank and Normal copula exhibit no tail dependence whereas the t- and the Gumbel do so. The t- copula is symmetric in its treatment of upper and lower tails.

Table 16.8. Tail Dependence Parameters for Four Copulas

Copula	Lower	Upper
Frank	0	0
Gumbel	0	0.74
Normal	0	0
$t-$	0.10	0.10

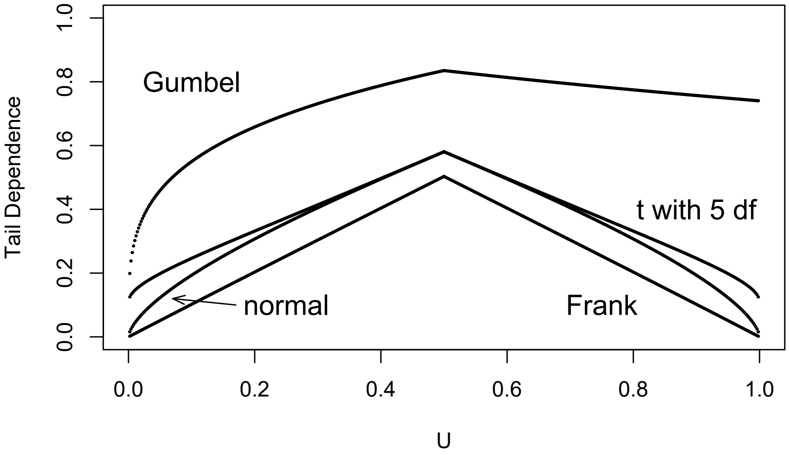


FIGURE 16.13: **Tail Dependence Plots**

Example 16.5.2. Lower Tail Dependence Coefficient Example

The bivariate distribution function $C(u, v) = uv$. What is the lower tail dependence coefficient of this copula?

Example Solution.

$$\lambda_{lo} = \lim_{u \to 0^+} \frac{C(u,u)}{u} = \lim_{u \to 0^+} \frac{u^2}{u} = \lim_{u \to 0^+} u = 0$$

16.6 Importance of Dependence Modeling

In this section, you learn how to:

- Explain the importance of dependence modeling
- Explain the importance of copulas for regression applications

16.6.1 Why is Dependence Modeling Important?

Dependence modeling is important because it enables us to understand the dependence structure by defining the relationship between variables in a dataset. In insurance, ignoring dependence modeling may not impact pricing but could lead to misestimation of required capital to cover losses. For instance, from Section 16.3 , it is seen that there was a positive relationship between `LOSS` and `ALAE`. This means that, if there is a large loss then we expect expenses to be large as well and ignoring this relationship could lead to mis-estimation of reserves.

To illustrate the importance of dependence modeling, we refer you back to portfolio management Example 13.4.6 that assumed that the property and liability risks are independent. Now, we incorporate dependence by allowing the four lines of business to depend on one another through a Gaussian copula. In Table 16.9, we show that dependence affects the portfolio quantiles (VaR_q), although not the expected values. For instance, the $VaR_{0.99}$ for total risk which is the amount of capital required to ensure, with a 99% degree of certainty that the firm does not become technically insolvent is higher when we incorporate dependence. This leads to less capital being allocated when dependence is ignored and can cause unexpected solvency problems.

Table 16.9. Results for Portfolio Expected Value and Quantiles (VaR_q)

Independent	Expected Value	$VaR_{0.9}$	$VaR_{0.95}$	$VaR_{0.99}$
Retained	269	300	300	300
Insurer	2,274	4,400	6,173	11,859
Total	2,543	4,675	6,464	12,159
Gaussian Copula	Expected Value	$VaR_{0.9}$	$VaR_{0.95}$	$VaR_{0.99}$
Retained	269	300	300	300
Insurer	2,340	4,988	7,339	14,905
Total	2,609	5,288	7,639	15,205

It should be noted that there are various methods of conducting dependence modeling, but copulas are effective for many actuarial applications. It's important to stress that each copula function captures a distinct dependency structure based on its functional form and dependence parameters. Therefore, utilizing copulas without comprehending their limitations and properties can lead to biased and statistically incorrect results. Since selecting the right copula involves extensive effort, here are some general tips that can assist:

1. When analyzing data, diagnostic and exploratory analysis can provide insight into the dependence structure of the data, which can help determine suitable copula functions. For Archimedean Copulas specifically, understanding the dependence structure can narrow down the appropriate type of copula function. For instance, the Gumbel-Hougaard copula is not suitable for negative dependency, but the Frank Copula can effectively capture three distinct types of dependency in the data.
2. Researchers cannot rely on Normal copula or Frank copula functions to capture the upper and lower tail dependency in data. Instead, a t copula with low degrees of freedom works well for both tails. The Gumbel-Hougaard copula shows some upper tail dependence but less or no lower tail dependence, while the Clayton copula exhibits strong lower tail dependence.

16.6.2 Copula Regression

In regression studies, the response variable is determined by a group of explanatory variables. This is often one of the initial statistical methods used to understand the connection between the response and explanatory variables. However, Linear Models and Generalized Linear Models can impose constraints on the selection of distributions for the response variables, which can be restrictive for practical data scenarios. For example, insurance claim amounts and financial asset returns typically exhibit heavy-tailed and skewed distributions, and may not adhere to normality patterns, with the possibility of having extreme values.

The use of copulas in regression is gaining attention in the field of actuarial science. Copula regression separates the dependency structure from the selection of marginal distributions, allowing for greater flexibility in choosing distributions for actuarial applications. The parameters for the marginal distributions and the copula distribution can be estimated either separately or together. The maximum likelihood method is often effective for estimating the parameters. However, for copula regression parameter estimation, the inference for margins method (IFM) is commonly used. Copula functions preserve the marginals and make predictions using the dependent variable's conditional mean given the covariates. See Krämer et al. (2013), Parsa and Klugman (2011); for detailed examples on copula regression.

16.7 Further Resources and Contributors

Contributors

- **Edward (Jed) Frees** and **Nii-Armah Okine**, University of Wisconsin-Madison, and **Emine Selin Sarıdaş**, Mimar Sinan University, are the principal authors of the initial version of this chapter.
  - Chapter reviewers include: Runhuan Feng, Fei Huang, Himchan Jeong, Min Ji, and Toby White.
- **Nii-Armah Okine**, Appalachian State University, and **Emine Selin Sarıdaş**, Mimar Sinan University, are the principal authors of the second edition of this chapter. Email: okinean@appstate.edu and selin.saridas@msgsu.edu.tr for chapter comments and suggested improvements.
  - Chapter reviewers include Mélina Mailhot.

TS 16.A. Other Classic Measures of Scalar Associations

TS 16.A.1. Blomqvist's Beta

Blomqvist (1950) developed a measure of dependence now known as Blomqvist's beta, also called the *median concordance coefficient* and the *medial correlation coefficient*. Using distribution functions, this parameter can be expressed as

$$\beta_B = 4F\left(F_X^{-1}(1/2), F_Y^{-1}(1/2)\right) - 1.$$

That is, first evaluate each marginal at its median ($F_X^{-1}(1/2)$ and $F_Y^{-1}(1/2)$, respectively). Then, evaluate the bivariate distribution function at the two medians. After rescaling (multiplying by 4 and subtracting 1), the coefficient turns out to have a range of $[-1,1]$, where 0 occurs under independence.

Like Spearman's rho and Kendall's tau, an estimator based on ranks is easy to provide. First write $\beta_B = 4C(1/2,1/2) - 1 = 2\Pr((U_1 - 1/2)(U_2 - 1/2)) - 1$ where U_1, U_2 are uniform random variables. Then, define

$$\hat{\beta}_B = \frac{2}{n}\sum_{i=1}^{n} I\left((R(X_i) - \frac{n+1}{2})(R(Y_i) - \frac{n+1}{2}) \geq 0\right) - 1.$$

See, for example, Joe (2014), page 57 or Hougaard (2000), page 135, for more details.

Because Blomqvist's parameter is based on the center of the distribution, it is particularly useful when data are censored; in this case, information in extreme parts of the distribution are not always reliable. How does this affect a choice of association measures? First, recall that association measures are based on a bivariate distribution function. So, if one has knowledge of a good approximation of the distribution function, then calculation of an association measure is straightforward in principle. Second, for censored data, bivariate extensions of the univariate Kaplan-Meier distribution function estimator are available. For example, the version introduced in Dabrowska (1988) is appealing. However, because of instances when large masses of data appear at the upper range of the data, this and other estimators of the bivariate distribution function are unreliable. This means that, summary measures of the estimated distribution function based on Spearman's rho or Kendall's tau can be unreliable. For this situation, Blomqvist's beta appears to be a better choice as it focuses on the center of the distribution. Hougaard (2000), Chapter 14, provides additional discussion.

You can obtain the Blomqvist's beta, using the `betan()` function from the `copula` library in `R`. From below, $\beta_B = 0.3$ between the `Coverage` rating variable in millions of dollars and `Claim` amount variable in dollars.

In addition, to show that the Blomqvist's beta is invariant under strictly increasing transformations, $\beta_B = 0.3$ between the `Coverage` rating variable in logarithmic millions of dollars and `Claim` amount variable in dollars.

TS 16.A.2. Nonparametric Approach Using Spearman Correlation with Tied Ranks

For the first variable, the average rank of observations in the sth row is

$$r_{1s} = n_{m_1\bullet} + \cdots + n_{s-1,\bullet} + \frac{1}{2}\left(1 + n_{s\bullet}\right)$$

and similarly $r_{2t} = \frac{1}{2}\left[(n_{\bullet m_1} + \cdots + n_{\bullet,s-1} + 1) + (n_{\bullet m_1} + \cdots + n_{\bullet s})\right]$. With this, we have Spearman's rho with tied rank is

$$\hat{\rho}_S = \frac{\sum_{s=m_1}^{m_2} \sum_{t=m_1}^{m_2} n_{st}(r_{1s} - \bar{r})(r_{2t} - \bar{r})}{\left[\sum_{s=m_1}^{m_2} n_{s\bullet}(r_{1s} - \bar{r})^2 \sum_{t=m_1}^{m_2} n_{\bullet t}(r_{2t} - \bar{r})^2\right]^2}$$

where the average rank is $\bar{r} = (n+1)/2$.

Special Case: Binary Data. Here, $m_1 = 0$ and $m_2 = 1$. For the first variable ranks, we have $r_{10} = (1 + n_{0\bullet})/2$ and $r_{11} = (n_{0\bullet} + 1 + n)/2$. Thus, $r_{10} - \bar{r} = (n_{0\bullet} - n)/2$ and $r_{11} - \bar{r} = n_{0\bullet}/2$. This means that we have $\sum_{s=0}^{1} n_{s\bullet}(r_{1s} - \bar{r})^2 = n(n - n_{0\bullet})n_{0\bullet}/4$ and similarly for the second variable. For the numerator, we have

$$\begin{array}{rl}
\displaystyle\sum_{s=0}^{1}\sum_{t=0}^{1} & n_{st}(r_{1s} - \bar{r})(r_{2t} - \bar{r}) \\
= & n_{00}\dfrac{n_{0\bullet} - n}{2}\dfrac{n_{\bullet 0} - n}{2} + n_{01}\dfrac{n_{0\bullet} - n}{2}\dfrac{n_{\bullet 0}}{2} + n_{10}\dfrac{n_{0\bullet}}{2}\dfrac{n_{\bullet 0} - n}{2} + n_{11}\dfrac{n_{0\bullet}}{2}\dfrac{n_{\bullet 0}}{2} \\
= & \dfrac{1}{4}(n_{00}(n_{0\bullet} - n)(n_{\bullet 0} - n) + (n_{0\bullet} - n_{00})(n_{0\bullet} - n)n_{\bullet 0} \\
 & \quad +(n_{\bullet 0} - n_{00})n_{0\bullet}(n_{\bullet 0} - n) + (n - n_{\bullet 0} - n_{0\bullet} + n_{00})n_{0\bullet}n_{\bullet 0}) \\
= & \dfrac{1}{4}(n_{00}n^2 - n_{0\bullet}(n_{0\bullet} - n)n_{\bullet 0} \\
 & \quad +n_{\bullet 0}n_{0\bullet}(n_{\bullet 0} - n) + (n - n_{\bullet 0} - n_{0\bullet})n_{0\bullet}n_{\bullet 0}) \\
= & \dfrac{1}{4}(n_{00}n^2 - n_{0\bullet}n_{\bullet 0}(n_{0\bullet} - n + n_{\bullet 0} - n + n - n_{\bullet 0} - n_{0\bullet}) \\
= & \dfrac{n}{4}(nn_{00} - n_{0\bullet}n_{\bullet 0}).
\end{array}$$

This yields

$$\begin{array}{rcl}
\hat{\rho}_S & = & \dfrac{n(nn_{00} - n_{0\bullet}n_{\bullet 0})}{4\sqrt{(n(n - n_{0\bullet})n_{0\bullet}/4)(n(n - n_{\bullet 0})n_{\bullet 0}/4)}} \\
 & = & \dfrac{nn_{00} - n_{0\bullet}n_{\bullet 0}}{\sqrt{n_{0\bullet}n_{\bullet 0}(n - n_{0\bullet})(n - n_{\bullet 0})}} \\
 & = & \dfrac{n_{00} - n(1 - \hat{\pi}_X)(1 - \hat{\pi}_Y)}{\sqrt{\hat{\pi}_X(1 - \hat{\pi}_X)\hat{\pi}_Y(1 - \hat{\pi}_Y)}}
\end{array}$$

where $\hat{\pi}_X = (n - n_{0\bullet})/n$ and similarly for $\hat{\pi}_Y$. Note that this is same form as the Pearson measure. From this, we see that the joint count n_{00} drives this association measure.

You can obtain the ties-corrected Spearman correlation statistic r_S using the `cor()` function in `R` and selecting the `spearman` method. From below $\hat{\rho}_S = -0.09$.

17

Appendix A: Review of Statistical Inference

Chapter Preview. The appendix gives an overview of concepts and methods related to statistical inference on the population of interest, using a random sample of observations from the population. In the appendix, Section 17.1 introduces the basic concepts related to the population and the sample used for making the inference. Section 17.2 presents the commonly used methods for point estimation of population characteristics. Section 17.3 demonstrates interval estimation that takes into consideration the uncertainty in the estimation, due to use of a random sample from the population. Section 17.4 introduces the concept of hypothesis testing for the purpose of variable and model selection.

17.1 Basic Concepts

In this section, you learn the following concepts related to statistical inference.

- Random sampling from a population that can be summarized using a list of items or individuals within the population
- Sampling distributions that characterize the distributions of possible outcomes for a statistic calculated from a random sample
- The central limit theorem that guides the distribution of the mean of a random sample from the population

Statistical inference is the process of making conclusions on the characteristics of a large set of items/individuals (i.e., the **population**), using a representative set of data (e.g., a **random sample**) from a list of items or individuals from the population that can be sampled. While the process has a broad spectrum of applications in various areas including science, engineering, health, social, and economic fields, statistical inference is important to insurance companies that use data from their existing policy holders in order to make inference on the characteristics (e.g., risk profiles) of a specific segment

TABLE 17.1: **Wisconsin Property Fund Summary Statistics**

	Minimum	First Quartile	Median	Mean	Third Quartile	Maximum	Standard Deviation
Claim	1	789	2,250	26,623	6,171	12,922,218	368,030
Logarithmic Claims	0	6.67	7.719	7.804	8.728	16.374	1.683

of target customers (i.e., the population) whom the insurance companies do not directly observe.

Example – Wisconsin Property Fund. Assume there are 1,377 *individual* claims from the 2010 experience. Summary statistics are in Table 17.1 and a visual summary is in Figure 17.1.

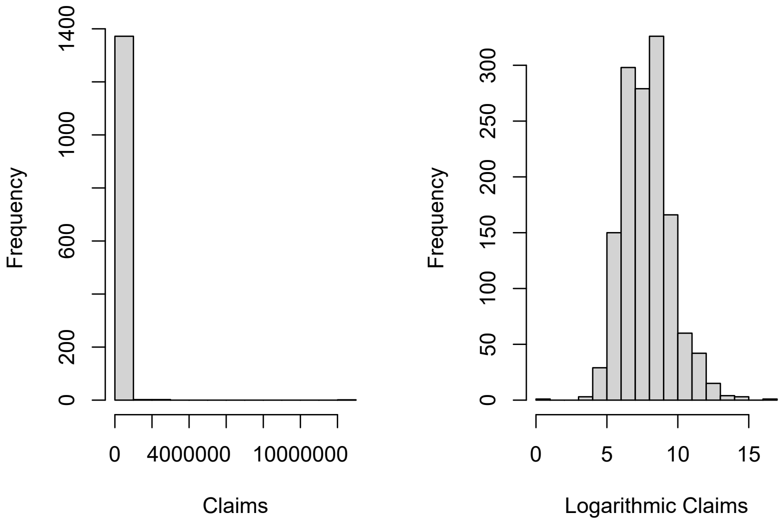


FIGURE 17.1: **Distribution of Claims for Wisconsin Property Fund**

Using the 2010 claim experience (the sample), the Wisconsin Property Fund may be interested in assessing the severity of all claims that could potentially occur, such as 2010, 2011, and so forth (the population). This process is important in the contexts of ratemaking or claim predictive modeling. In order for such inference to be valid, we need to assume that

- the set of 2010 claims is a *random sample* that is representative of the population,
- the *sampling distribution* of the average claim amount can be estimated, so

that we can quantify the bias and uncertainty in the estimation due to use of a finite sample.

17.1.1 Random Sampling

In statistics, a sampling **error** occurs when the **sampling frame**, the list from which the sample is drawn, is not an adequate approximation of the population of interest. A sample must be a representative subset of a population, or universe, of interest. If the sample is not representative, taking a larger sample does not eliminate bias, as the same mistake is repeated over again and again. Thus, we introduce the concept for random sampling that gives rise to a simple **random sample** that is representative of the population.

We assume that the random variable X represents a draw from a population with a distribution function $F(\cdot)$ with mean $\mathrm{E}[X] = \mu$ and variance $\mathrm{Var}[X] = \mathrm{E}[(X - \mu)^2]$, where $E(\cdot)$ denotes the expectation of a random variable. In **random sampling**, we make a total of n such draws represented by $X_1, \ldots, X_n$, each unrelated to one another (i.e., *statistically independent*). We refer to $X_1, \ldots, X_n$ as a **random sample** (*with replacement*) from $F(\cdot)$, taking either a parametric or nonparametric form. Alternatively, we may say that $X_1, \ldots, X_n$ are identically and independently distributed (*iid*) with distribution function $F(\cdot)$.

17.1.2 Sampling Distribution

Using the random sample $X_1, \ldots, X_n$, we are interested in making a conclusion on a specific attribute of the population distribution $F(\cdot)$. For example, we may be interested in making an inference on the population mean, denoted μ. It is natural to think of the **sample mean**, $\bar{X} = \sum_{i=1}^{n} X_i$, as an estimate of the population mean μ. We call the sample mean as a **statistic** calculated from the random sample $X_1, \ldots, X_n$. Other commonly used summary statistics include sample standard deviation and sample quantiles.

When using a statistic (e.g., the sample mean $\bar{X}$) to make statistical inference on the population attribute (e.g., population mean μ), the quality of inference is determined by the bias and uncertainty in the estimation, owing to the use of a sample in place of the population. Hence, it is important to study the distribution of a statistic that quantifies the bias and variability of the statistic. In particular, the distribution of the sample mean, $\bar{X}$ (or any other statistic), is called the **sampling distribution**. The sampling distribution depends on the sampling process, the statistic, the sample size n and the population distribution $F(\cdot)$. The central limit theorem gives the large-sample (sampling) distribution of the sample mean under certain conditions.

17.1.3 Central Limit Theorem

In statistics, there are variations of the central limit theorem (CLT) ensuring that, under certain conditions, the sample mean will approach the population mean with its sampling distribution approaching the normal distribution as the sample size goes to infinity. We give the Lindeberg–Levy CLT that establishes the asymptotic sampling distribution of the sample mean $\bar{X}$ calculated using a random sample from a universe population having a distribution $F(\cdot)$.

Lindeberg–Levy CLT. Let $X_1, \ldots, X_n$ be a random sample from a population distribution $F(\cdot)$ with mean μ and variance $\sigma^2 < \infty$. The difference between the sample mean $\bar{X}$ and μ, when multiplied by $\sqrt{n}$, converges in distribution to a normal distribution as the sample size goes to infinity. That is,

$$\sqrt{n}(\bar{X} - \mu) \rightarrow^d N(0, \sigma).$$

Note that the CLT does not require a parametric form for $F(\cdot)$. Based on the CLT, we may perform statistical inference on the population mean (we *infer*, not *deduce*). The types of inference we may perform include **estimation** of the population, **hypothesis testing** on whether a null statement is true, and **prediction** of future samples from the population.

17.2 Point Estimation and Properties

In this section, you learn how to

- estimate population parameters using method of moments estimation
- estimate population parameters based on maximum likelihood estimation

The population distribution function $F(\cdot)$ can usually be characterized by a limited (finite) number of terms called **parameters**, in which case we refer to the distribution as a **parametric distribution**. In contrast, in **nonparametric** analysis, the attributes of the sampling distribution are not limited to a small number of parameters.

For obtaining the population characteristics, there are different attributes related to the population distribution $F(\cdot)$. Such measures include the mean, median, percentiles (i.e., 95th percentile), and standard deviation. Because these summary measures do not depend on a specific parametric reference, they are **nonparametric** summary measures.

In **parametric** analysis, on the other hand, we may assume specific families of distributions with specific parameters. For example, people usually think of logarithm of claim amounts to be normally distributed with mean μ and standard deviation σ. That is, we assume that the claims have a *lognormal* distribution with parameters μ and σ. Alternatively, insurance companies commonly assume that claim severity follows a gamma distribution with a shape parameter α and a scale parameter θ. Here, the normal, lognormal, and gamma distributions are examples of parametric distributions. In the above examples, the quantities of μ, σ, α, and θ are known as *parameters.* For a given parametric distribution family, the distribution is uniquely determined by the values of the parameters.

One often uses θ to denote a summary attribute of the population. In parametric models, θ can be a parameter or a function of parameters from a distribution such as the normal mean and variance parameters. In nonparametric analysis, it can take a form of a nonparametric summary such as the population mean or standard deviation. Let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be a function of the sample that provides a proxy, or an **estimate**, of θ. It is referred to as a **statistic**, a function of the sample $X_1, \ldots, X_n$.

Example – Wisconsin Property Fund. The sample mean 7.804 and the sample standard deviation 1.683 can be either deemed as nonparametric estimates of the population mean and standard deviation, or as parametric estimates of μ and σ of the normal distribution concerning the logarithmic claims. Using results from the lognormal distribution, we may estimate the expected claim, the lognormal mean, as 10,106.8 ($= \exp(7.804 + 1.683^2/2)$).

For the Wisconsin Property Fund data, we may denote $\hat{\mu} = 7.804$ and $\hat{\sigma} = 1.683$, with the hat notation denoting an **estimate** of the parameter based on the sample. In particular, such an estimate is referred to as a **point estimate**, a single approximation of the corresponding parameter. For point estimation, we introduce the two commonly used methods called the method of moments estimation and maximum likelihood estimation.

17.2.1 Method of Moments Estimation

Before defining the method of moments estimation, we define the the concept of **moments**. Moments are population attributes that characterize the distribution function $F(\cdot)$. Given a random draw X from $F(\cdot)$, the expectation $\mu_k = \mathrm{E}[X^k]$ is called the k**th moment** of X, $k = 1, 2, 3, \ldots$ For example, the population mean μ is the *first* moment. Furthermore, the expectation $\mathrm{E}[(X - \mu)^k]$ is called a k**th central moment**. Thus, the variance is the second central moment.

Using the random sample $X_1, \ldots, X_n$, we may construct the corresponding sample moment, $\hat{\mu}_k = (1/n)\sum_{i=1}^n X_i^k$, for estimating the population attribute μ_k. For example, we have used the sample mean $\bar{X}$ as an estimator for the population mean μ. Similarly, the second central moment can be estimated as $(1/n)\sum_{i=1}^n (X_i - \bar{X})^2$. Without assuming a parametric form for $F(\cdot)$, the sample moments constitute nonparametric estimates of the corresponding population attributes. Such an estimator based on matching of the corresponding sample and population moments is called a **method of moments estimator** (*mme*).

While the *mme* works naturally in a nonparametric model, it can be used to estimate parameters when a specific parametric family of distribution is assumed for $F(\cdot)$. Denote by $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)$ the vector of parameters corresponding to a parametric distribution $F(\cdot)$. Given a distribution family, we commonly know the relationships between the parameters and the moments. In particular, we know the specific forms of the functions $h_1(\cdot), h_2(\cdot), \cdots, h_m(\cdot)$ such that $\mu_1 = h_1(\boldsymbol{\theta})$, $\mu_2 = h_2(\boldsymbol{\theta})$, $\cdots$, $\mu_m = h_m(\boldsymbol{\theta})$. Given the *mme* $\hat{\mu}_1, \ldots, \hat{\mu}_m$ from the random sample, the *mme* of the parameters $\hat{\theta}_1, \cdots, \hat{\theta}_m$ can be obtained by solving the equations of

$$\begin{aligned}
\hat{\mu}_1 \quad &= h_1(\hat{\theta}_1, \cdots, \hat{\theta}_m) \\
\hat{\mu}_2 \quad &= h_2(\hat{\theta}_1, \cdots, \hat{\theta}_m) \\
\vdots \quad &\quad \vdots \\
\hat{\mu}_m \quad &= h_m(\hat{\theta}_1, \cdots, \hat{\theta}_m).
\end{aligned}$$

Example – Wisconsin Property Fund. Assume that the claims follow a lognormal distribution, so that logarithmic claims follow a normal distribution. Specifically, assume $\log(X)$ has a normal distribution with mean μ and variance σ^2, denoted as $\log(X) \sim N(\mu, \sigma^2)$. It is straightforward that the *mme* $\hat{\mu} = \bar{X}$ and $\hat{\sigma} = \sqrt{(1/n)\sum_{i=1}^n (X_i - \bar{X})^2}$. For the Wisconsin Property Fund example, the method of moments estimates are $\hat{\mu} = 7.804$ and $\hat{\sigma} = 1.683$.

17.2.2 Maximum Likelihood Estimation

When $F(\cdot)$ takes a parametric form, the maximum likelihood method is widely used for estimating the population parameters $\boldsymbol{\theta}$. Maximum likelihood estimation is based on the likelihood function, a function of the parameters given the observed sample. Denote by $f(x_i|\boldsymbol{\theta})$ the probability function of X_i evaluated at $X_i = x_i$ $(i = 1, 2, \cdots, n)$; it is the probability mass function in the case of a discrete X and the probability density function in the case of a continuous X. Assuming independence, the **likelihood function** of $\boldsymbol{\theta}$ associated with the observation $(X_1, X_2, \cdots, X_n) = (x_1, x_2, \cdots, x_n) = \mathbf{x}$ can be written as

$$L(\boldsymbol{\theta}|\mathbf{x}) = \prod_{i=1}^n f(x_i|\boldsymbol{\theta}),$$

with the corresponding **log-likelihood function** given by

$$l(\boldsymbol{\theta}|\mathbf{x}) = \log(L(\boldsymbol{\theta}|\mathbf{x})) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta}).$$

The maximum likelihood estimator (*mle*) of $\boldsymbol{\theta}$ is the set of values of $\boldsymbol{\theta}$ that maximize the likelihood function (log-likelihood function), given the observed sample. That is, the *mle* $\hat{\boldsymbol{\theta}}$ can be written as

$$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}\in\Theta} l(\boldsymbol{\theta}|\mathbf{x}),$$

where Θ is the parameter space of $\boldsymbol{\theta}$, and $\text{argmax}_{\boldsymbol{\theta}\in\Theta} l(\boldsymbol{\theta}|\mathbf{x})$ is defined as the value of $\boldsymbol{\theta}$ at which the function $l(\boldsymbol{\theta}|\mathbf{x})$ reaches its maximum.

Given the analytical form of the likelihood function, the *mle* can be obtained by taking the first derivative of the log-likelihood function with respect to $\boldsymbol{\theta}$, and setting the values of the partial derivatives to zero. That is, the *mle* are the solutions of the equations of

$$\begin{aligned} \frac{\partial l(\hat{\boldsymbol{\theta}}|\mathbf{x})}{\partial \hat{\theta}_1} &= 0 \\ \frac{\partial l(\hat{\boldsymbol{\theta}}|\mathbf{x})}{\partial \hat{\theta}_2} &= 0 \\ \cdots & \\ \frac{\partial l(\hat{\boldsymbol{\theta}}|\mathbf{x})}{\partial \hat{\theta}_m} &= 0, \end{aligned}$$

provided that the second partial derivatives are negative.

For parametric models, the *mle* of the parameters can be obtained either analytically (e.g., in the case of normal distributions and linear estimators), or numerically through iterative algorithms such as the Newton-Raphson method and its adaptive versions (e.g., in the case of generalized linear models with a non-normal response variable).

Normal distribution. Assume $(X_1, X_2, \cdots, X_n)$ to be a random sample from the normal distribution $N(\mu, \sigma^2)$. With an observed sample $(X_1, X_2, \cdots, X_n) = (x_1, x_2, \cdots, x_n)$, we can write the likelihood function of μ, σ^2 as

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right],$$

with the corresponding log-likelihood function given by

$$l(\mu, \sigma^2) = -\frac{n}{2}[\log(2\pi) + \log(\sigma^2)] - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 .$$

By solving

$$\frac{\partial l(\hat{\mu}, \sigma^2)}{\partial \hat{\mu}} = 0,$$

we obtain $\hat{\mu} = \bar{x} = (1/n)\sum_{i=1}^{n} x_i$. It is straightforward to verify that $\frac{\partial l^2(\hat{\mu},\sigma^2)}{\partial \hat{\mu}^2}|_{\hat{\mu}=\bar{x}} < 0$. Since this works for arbitrary x, $\hat{\mu} = \bar{X}$ is the *mle* of μ. Similarly, by solving

$$\frac{\partial l(\mu, \hat{\sigma}^2)}{\partial \hat{\sigma}^2} = 0,$$

we obtain $\hat{\sigma}^2 = (1/n)\sum_{i=1}^{n}(x_i - \mu)^2$. Further replacing μ by $\hat{\mu}$, we derive the *mle* of σ^2 as $\hat{\sigma}^2 = (1/n)\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Hence, the sample mean $\bar{X}$ and $\hat{\sigma}^2$ are both the *mme* and MLE for the mean μ and variance σ^2, under a normal population distribution $F(\cdot)$. More details regarding the properties of the likelihood function are given in Appendix Section 19.1.

17.3 Interval Estimation

In this section, you learn how to

- derive the exact sampling distribution of the *mle* of the normal mean
- obtain the large-sample approximation of the sampling distribution using the large sample properties of the *mle*
- construct a confidence interval of a parameter based on the large sample properties of the *mle*

Now that we have introduced the *mme* and *mle*, we may perform the first type of statistical inference, **interval estimation** that quantifies the uncertainty resulting from the use of a finite sample. By deriving the sampling distribution of *mle*, we can estimate an interval (a confidence interval) for the parameter. Under the frequentist approach (e.g., that based on maximum likelihood estimation), the confidence intervals generated from the same random sampling frame will cover the true value the majority of times (e.g., 95% of the times), if we repeat the sampling process and re-calculate the interval over and over again. Such a process requires the derivation of the sampling distribution for the *mle*.

17.3.1 Exact Distribution for Normal Sample Mean

Due to the **additivity** property of the normal distribution (i.e., a sum of normal random variables that follows a multivariate normal distribution still follows a normal distribution) and that the normal distribution belongs to the **location–scale family** (i.e., a location and/or scale transformation of a normal random variable has a normal distribution), the sample mean $\bar{X}$ of a random sample from a normal $F(\cdot)$ has a normal sampling distribution for any finite n. Given $X_i \sim^{iid} N(\mu, \sigma^2)$, $i = 1, \ldots, n$, the *mle* of μ has an exact distribution

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Hence, the sample mean is an unbiased estimator of μ. In addition, the uncertainty in the estimation can be quantified by its variance σ^2/n, that decreases with the sample size n. When the sample size goes to infinity, the sample mean will approach a single mass at the true value.

17.3.2 Large-sample Properties of *MLE*

For the *mle* of the mean parameter and any other parameters of other parametric distribution families, however, we usually cannot derive an exact sampling distribution for finite samples. Fortunately, when the sample size is sufficiently large, *mle*s can be approximated by a normal distribution. Due to the general maximum likelihood theory, the *mle* has some nice large-sample properties.

- The *mle* $\hat{\theta}$ of a parameter θ, is a **consistent** estimator. That is, $\hat{\theta}$ converges in probability to the true value θ, as the sample size n goes to infinity.
- The *mle* has the **asymptotic normality** property, meaning that the estimator will converge in distribution to a normal distribution centered around the true value, when the sample size goes to infinity. Namely,

  $$\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d N\left(0,\, V\right), \quad \text{as} \quad n \rightarrow \infty,$$

  where V is the inverse of the Fisher Information. Hence, the *mle* $\hat{\theta}$ approximately follows a normal distribution with mean θ and variance V/n, when the sample size is large.
- The *mle* is **efficient**, meaning that it has the smallest asymptotic variance V, commonly referred to as the **Cramer–Rao lower bound**. In particular, the Cramer–Rao lower bound is the inverse of the Fisher information defined as $\mathcal{I}(\theta) = -\mathrm{E}(\partial^2 \log f(X; \theta)/\partial\theta^2)$. Hence, $\mathrm{Var}(\hat{\theta})$ can be estimated based on the observed Fisher information that can be written as $-\sum_{i=1}^{n} \partial^2 \log f(X_i; \theta)/\partial\theta^2$.

For many parametric distributions, the Fisher information may be derived analytically for the *mle* of parameters. For more sophisticated parametric models, the Fisher information can be evaluated numerically using numerical integration for continuous distributions, or numerical summation for discrete distributions. More details regarding maximum likelihood estimation are given in Appendix Section 19.2.

17.3.3 Confidence Interval

Given that the *mle* $\hat{\theta}$ has either an exact or an approximate normal distribution with mean θ and variance $\mathrm{Var}(\hat{\theta})$, we may take the square root of the variance and plug-in the estimate to define $se(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}$. A **standard error** is an estimated standard deviation that quantifies the uncertainty in the estimation resulting from the use of a finite sample. Under some regularity conditions governing the population distribution, we may establish that the statistic

$$\frac{\hat{\theta} - \theta}{se(\hat{\theta})}$$

converges in distribution to a Student-t distribution with degrees of freedom (a parameter of the distribution) $n - p$, where p is the number of parameters in the model other than the variance. For example, for the normal distribution case, we have $p = 1$ for the parameter μ; for a linear regression model with an independent variable, we have $p = 2$ for the parameters of the intercept and the independent variable. Denote by $t_{n-p}(1-\alpha/2)$ the $100 \times (1-\alpha/2)$-th percentile of the Student-t distribution that satisfies $\Pr\left[t < t_{n-p}\left(1-\alpha/2\right)\right] = 1 - \alpha/2$. We have,

$$\Pr\left[-t_{n-p}\left(1 - \frac{\alpha}{2}\right) < \frac{\hat{\theta} - \theta}{se(\hat{\theta})} < t_{n-p}\left(1 - \frac{\alpha}{2}\right)\right] = 1 - \alpha,$$

from which we can derive a **confidence interval** for θ. From the above equation we can derive a pair of statistics, $\hat{\theta}_1$ and $\hat{\theta}_2$, that provide an interval of the form $[\hat{\theta}_1, \hat{\theta}_2]$. This interval is a $1 - \alpha$ confidence interval for θ such that $\Pr\left(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2\right) = 1 - \alpha$, where the probability $1 - \alpha$ is referred to as the **confidence level**. Note that the above confidence interval is not valid for small samples, except for the case of the normal mean.

Normal distribution. For the normal population mean μ, the *mle* has an exact sampling distribution $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$, in which we can estimate $se(\hat{\theta})$ by $\hat{\sigma}/\sqrt{n}$. Based on the **Cochran's theorem**, the resulting statistic has an exact Student-t distribution with degrees of freedom $n - 1$. Hence, we can derive the lower and upper bounds of the confidence interval as

$$\hat{\mu}_1 = \hat{\mu} - t_{n-1}\left(1 - \frac{\alpha}{2}\right)\frac{\hat{\sigma}}{\sqrt{n}}$$

and

$$\hat{\mu}_2 = \hat{\mu} + t_{n-1}\left(1 - \frac{\alpha}{2}\right)\frac{\hat{\sigma}}{\sqrt{n}}.$$

When $\alpha = 0.05$, $t_{n-1}(1 - \alpha/2) \approx 1.96$ for large values of n. Based on the Cochran's theorem, the confidence interval is valid regardless of the sample size.

Example – Wisconsin Property Fund. For the lognormal claim model, (7.715235, 7.893208) is a 95% confidence interval for μ.

More details regarding interval estimation based the *mle* of other parameters and distribution families are given in Appendix Chapter 19.

17.4 Hypothesis Testing

In this section, you learn how to

- understand the basic concepts in hypothesis testing including the level of significance and the power of a test
- perform hypothesis testing such as a Student-t test based on the properties of the *mle*
- construct a likelihood ratio test for a single parameter or multiple parameters from the same statistical model
- use information criteria such as the Akaike's information criterion or the Bayesian information criterion to perform model selection

---

For the parameter(s) $\boldsymbol{\theta}$ from a parametric distribution, an alternative type of statistical inference is called **hypothesis testing** that verifies whether a hypothesis regarding the parameter(s) is true, under a given probability called the **level of significance** α (e.g., 5%). In hypothesis testing, we reject the null hypothesis, a restrictive statement concerning the parameter(s), if the probability of observing a random sample as extremal as the observed one is smaller than α, if the null hypothesis were true.

17.4.1 Basic Concepts

In a statistical test, we are usually interested in testing whether a statement regarding some parameter(s), a **null hypothesis** (denoted H_0), is true given the observed data. The null hypothesis can take a general form $H_0 : \theta \in \Theta_0$, where Θ_0 is a subset of the parameter space Θ of θ that may contain multiple parameters. For the case with a single parameter θ, the null hypothesis usually takes either the form $H_0 : \theta = \theta_0$ or $H_0 : \theta \leq \theta_0$. The opposite of the null hypothesis is called the **alternative hypothesis** that can be written as $H_a : \theta \neq \theta_0$ or $H_a : \theta > \theta_0$. The statistical test on $H_0 : \theta = \theta_0$ is called a **two-sided** as the alternative hypothesis contains two inequalities of $H_a : \theta < \theta_0$ or $\theta > \theta_0$. In contrast, the statistical test on either $H_0 : \theta \leq \theta_0$ or $H_0 : \theta \geq \theta_0$ is called a **one-sided** test.

A statistical test is usually constructed based on a statistic T and its exact or large-sample distribution. The test typically rejects a two-sided test when either $T > c_1$ or $T < c_2$, where the two constants c_1 and c_2 are obtained based on the sampling distribution of T at a probability level α called the **level of significance**. In particular, the level of significance α satisfies

$$\alpha = \Pr(\text{reject } H_0 | H_0 \text{ is true}),$$

meaning that if the null hypothesis were true, we would reject the null hypothesis only 5% of the times, if we repeat the sampling process and perform the test over and over again.

Thus, the level of significance is the probability of making a **type I error** (error of the first kind), the error of incorrectly rejecting a true null hypothesis. For this reason, the level of significance α is also referred to as the type I error rate. Another type of error we may make in hypothesis testing is the **type II error** (error of the second kind), the error of incorrectly accepting a false null hypothesis. Similarly, we can define the **type II error rate** as the probability of not rejecting (accepting) a null hypothesis given that it is not true. That is, the type II error rate is given by

$$\Pr(\text{accept } H_0 | H_0 \text{ is false}).$$

Another important quantity concerning the quality of the statistical test is called the **power** of the test β, defined as the probability of rejecting a false null hypothesis. The mathematical definition of the power is

$$\beta = \Pr(\text{reject } H_0 | H_0 \text{ is false}).$$

Note that the power of the test is typically calculated based on a specific alternative value of $\theta = \theta_a$, given a specific sampling distribution and a given sample size. In real experimental studies, people usually calculate the required sample size in order to choose a sample size that will ensure a large chance of obtaining a statistically significant test (i.e., with a prespecified statistical power such as 85%).

17.4.2 Student-*t* test based on *mle*

Based on the results from Section 17.3.1, we can define a Student-t test for testing $H_0 : \theta = \theta_0$. In particular, we define the test statistic as

$$t\text{-stat} = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})},$$

which has a large-sample distribution of a student-t distribution with degrees of freedom $n - p$, when the null hypothesis is true (i.e., when $\theta = \theta_0$).

For a given **level of significance** α, say 5%, we reject the null hypothesis if the event t-stat $< -t_{n-p}(1 - \alpha/2)$ or t-stat $> t_{n-p}(1 - \alpha/2)$ occurs (the **rejection region**). Under the null hypothesis H_0, we have

$$\Pr\left[t\text{-stat} < -t_{n-p}\left(1 - \frac{\alpha}{2}\right)\right] = \Pr\left[t\text{-stat} > t_{n-p}\left(1 - \frac{\alpha}{2}\right)\right] = \frac{\alpha}{2}.$$

In addition to the concept of rejection region, we may reject the test based on the p**-value** defined as $2\Pr(T > |t\text{-stat}|)$ for the aforementioned two-sided test, where the random variable $T \sim T_{n-p}$. We reject the null hypothesis if p-value is smaller than and equal to α. For a given sample, a p-value is defined to be the smallest significance level for which the null hypothesis would be rejected.

Similarly, we can construct a one-sided test for the null hypothesis $H_0 : \theta \leq \theta_0$ (or $H_0 : \theta \geq \theta_0$). Using the same test statistic, we reject the null hypothesis when t-stat $> t_{n-p}(1 - \alpha)$ (or t-stat $< -t_{n-p}(1 - \alpha)$ for the test on $H_0 : \theta \geq \theta_0$). The corresponding p-value is defined as $\Pr(T > |t\text{-stat}|)$ (or $\Pr(T < |t\text{-stat}|)$ for the test on $H_0 : \theta \geq \theta_0$). Note that the test is not valid for small samples, except for the case of the test on the normal mean.

One-sample t Test for Normal Mean. For the test on the normal mean of the form $H_0 : \mu = \mu_0$, $H_0 : \mu \leq \mu_0$ or $H_0 : \mu \geq \mu_0$, we can define the test statistic as

$$t\text{-stat} = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}},$$

for which we have an exact sampling distribution t-stat $\sim T_{n-1}$ from the Cochran's theorem, with T_{n-1} denoting a Student-t distribution with degrees of freedom $n - 1$. According to the Cochran's theorem, the test is valid for both small and large samples.

TABLE 17.2: **Wisconsin Property Fund Parameter Estimates**

	Parameter Estimate	Standard Error	t-stat
Gamma	10.190	0.050	203.831
	-1.236	0.030	-41.180
Lognormal	7.804	0.045	172.089
	0.520	0.019	27.303
Pareto	7.733	0.093	82.853
	-0.001	0.054	-0.016
GB2	2.831	1.000	2.832
	1.203	0.292	4.120
	6.329	0.390	16.220
	1.295	0.219	5.910

Example – Wisconsin Property Fund. Assume that mean logarithmic claims have historically been approximately by $\mu_0 = \log(5000) = 8.517$. We might want to use the 2010 data to assess whether the mean of the distribution has changed (a two-sided test), or whether it has increased (a one-sided test). Given the actual 2010 average $\hat{\mu} = 7.804$, we may use the one-sample t test to assess whether this is a significant departure from $\mu_0 = 8.517$ (i.e., in testing $H_0 : \mu = 8.517$). The test statistic $t\text{-stat} = (8.517 - 7.804)/(1.683/\sqrt{1377}) = 15.72 > t_{1376}(0.975)$. Hence, we reject the two-sided test at $\alpha = 5\%$. Similarly, we will reject the one-sided test at $\alpha = 5\%$.

Example – Wisconsin Property Fund. For numerical stability and extensions to regression applications, statistical packages often work with transformed versions of parameters. Table 17.2 provides estimates based on the **R** package **VGAM** (the function). More details on the *mle* of other distribution families are given in Appendix Chapter 19.

17.4.3 Likelihood Ratio Test

In the previous subsection, we have introduced the Student-t test on a single parameter, based on the properties of the *mle*. In this section, we define an alternative test called the **likelihood ratio test** (LRT). The LRT may be used to test multiple parameters from the same statistical model.

Given the likelihood function $L(\theta|\mathbf{x})$ and $\Theta_0 \subset \Theta$, the likelihood ratio test statistic for testing $H_0 : \theta \in \Theta_0$ against $H_a : \theta \notin \Theta_0$ is given by

$$L = \frac{\sup_{\theta \in \Theta_0} L(\theta|\mathbf{x})}{\sup_{\theta \in \Theta} L(\theta|\mathbf{x})},$$

and that for testing $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$ is

$$L = \frac{L(\theta_0|\mathbf{x})}{\sup_{\theta\in\Theta} L(\theta|\mathbf{x})}.$$

The *LRT* rejects the null hypothesis when $L < c$, with the threshold depending on the level of significance α, the sample size n, and the number of parameters in θ. Based on the **Neyman–Pearson Lemma**, the *LRT* is the **uniformly most powerful** test for testing $H_0 : \theta = \theta_0$ versus $H_a : \theta = \theta_a$. That is, it provides the largest power β for a given α and a given alternative value θ_a.

Based on the **Wilks's Theorem**, the likelihood ratio test statistic $-2\log(L)$ converges in distribution to a Chi-square distribution with the degree of freedom being the difference between the dimensionality of the parameter spaces Θ and Θ_0, when the sample size goes to infinity and when the null model is nested within the alternative model. That is, when the null model is a special case of the alternative model containing a restricted sample space, we may approximate c by $\chi^2_{p_1-p_2}(1-\alpha)$, the $100\times(1-\alpha)$ th percentile of the Chi-square distribution, with $p_1 - p_2$ being the degrees of freedom, and p_1 and p_2 being the numbers of parameters in the alternative and null models, respectively. Note that the *LRT* is also a large-sample test that will not be valid for small samples.

17.4.4 Information Criteria

In real-life applications, the *LRT* has been commonly used for comparing two nested models. The *LRT* approach as a model selection tool, however, has two major drawbacks: 1) It typically requires the null model to be nested within the alternative model; 2) models selected from the *LRT* tends to provide in-sample over-fitting, leading to poor out-of-sample prediction. In order to overcome these issues, model selection based on information criteria, applicable to non-nested models while taking into consideration the model complexity, is more widely used for model selection. Here, we introduce the two most widely used criteria, the Akaike's information criterion and the Bayesian information criterion.

In particular, the **Akaike's information criterion** (*AIC*) is defined as

$$AIC = -2\log L(\hat{\boldsymbol{\theta}}) + 2p,$$

where $\hat{\boldsymbol{\theta}}$ denotes the *mle* of $\boldsymbol{\theta}$, and p is the number of parameters in the model. The additional term $2p$ represents a penalty for the complexity of the model. That is, with the same maximized likelihood function, the *AIC* favors model with less parameters. We note that the *AIC* does not consider the impact from the sample size n.

Alternatively, people use the **Bayesian information criterion** (BIC) that takes into consideration the sample size. The BIC is defined as

$$BIC = -2 \log L(\hat{\boldsymbol{\theta}}) + p \ \log(n).$$

We observe that the BIC generally puts a higher weight on the number of parameters. With the same maximized likelihood function, the BIC will suggest a more parsimonious model than the AIC.

Example – Wisconsin Property Fund. Both the AIC and BIC statistics suggest that the *GB2* is the best fitting model whereas gamma is the worst.

Distribution	AIC	BIC
Gamma	28,305.2	28,315.6
Lognormal	26,837.7	26,848.2
Pareto	26,813.3	26,823.7
GB2	26,768.1	26,789.0

In Figure 17.2,

- black represents actual (smoothed) logarithmic claims
- Best approximated by green which is fitted GB2
- Pareto (purple) and Lognormal (lightblue) are also pretty good
- Worst are the exponential (in red) and gamma (in dark blue)

```
Sample size:  6258
```

You can learn more about the R code for this example at the online version of this book, Actuarial Community (2025).

Contributors

- **Lei (Larry) Hua**, Northern Illinois University, and **Edward (Jed) Frees**, University of Wisconsin-Madison, are the principal authors of the initial version of this chapter. Email: lhua@niu.edu or jfrees@bus.wisc.edu for chapter comments and suggested improvements.

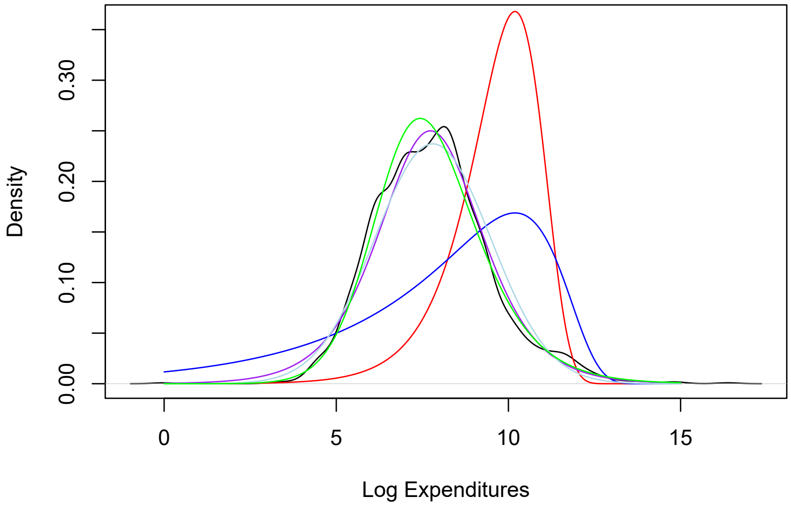


FIGURE 17.2: **Fitted Claims Distribution**

18

Appendix B: Iterated Expectations

This appendix introduces the laws related to iterated expectations. In particular, Section 18.1 introduces the concepts of conditional distribution and conditional expectation. Section 18.2 introduces the Law of Iterated Expectations and the Law of Total Variance.

In some situations, we only observe a single outcome but can conceptualize an outcome as resulting from a two (or more) stage process. Such types of statistical models are called **two-stage**, or **hierarchical** models. Some special cases of hierarchical models include:

- models where the parameters of the distribution are random variables;
- mixture distribution, where Stage 1 represents the draw of a subpopulation and Stage 2 represents a random variable from a distribution that is determined by the subpopulation drew in Stage 1;
- an aggregate distribution, where Stage 1 represents the draw of the number of events and Stage 2 represents the loss amount occurred per event.

In these situations, the process gives rise to a conditional distribution of a random variable (the Stage 2 outcome) given the other (the Stage 1 outcome). The Law of Iterated Expectations can be useful for obtaining the unconditional expectation or variance of a random variable in such cases.

18.1 Conditional Distribution and Conditional Expectation

In this section, you learn

- the concepts related to the conditional distribution of a random variable given another
- how to define the conditional expectation and variance based on the conditional distribution function

The iterated expectations are the laws regarding calculation of the expectation and variance of a random variable using a conditional distribution of the variable given another variable. Hence, we first introduce the concepts related to the conditional distribution, and the calculation of the conditional expectation and variance based on a given conditional distribution.

18.1.1 Conditional Distribution

Here we introduce the concept of conditional distribution respectively for discrete and continuous random variables.

Discrete Case

Suppose that X and Y are both discrete random variables, meaning that they can take a finite or countable number of possible values with a positive probability. The **joint probability (mass) function** of (X, Y) is defined as

$$p(x, y) = \Pr[X = x, Y = y].$$

When X and Y are **independent** (the value of X does not depend on that of Y), we have

$$p(x, y) = p(x)p(y),$$

with $p(x) = \Pr[X = x]$ and $p(y) = \Pr[Y = y]$ being the **marginal probability functions** of X and Y, respectively.

Given the joint probability function, we may obtain the marginal probability function of Y as

$$p(y) = \sum_x p(x, y),$$

where the summation is over all possible values of x, and the marginal probability function of X can be obtained in a similar manner.

The **conditional probability (mass) function** of $(Y|X)$ is defined as

$$p(y|x) = \Pr[Y = y|X = x] = \frac{p(x, y)}{\Pr[X = x]},$$

where we may obtain the conditional probability function of $(X|Y)$ in a similar manner. In particular, the above conditional probability represents the probability of the event $Y = y$ given the event $X = x$. Hence, even in cases where $\Pr[X = x] = 0$, the function may be given as a particular form, in real applications.

Continuous Case

For continuous random variables X and Y, we may define their joint probability (density) function based on the joint cumulative distribution function. The

joint cumulative distribution function of (X, Y) is defined as

$$F(x, y) = \Pr[X \leq x, Y \leq y].$$

When X and Y are *independent*, we have

$$F(x, y) = F(x)F(y),$$

with $F(x) = \Pr[X \leq x]$ and $F(y) = \Pr[Y \leq y]$ being the **cumulative distribution functions** (cdfs) of X and Y, respectively. The random variable X is referred to as a **continuous** random variable if its cdf is continuous on x.

When the cdf $F(x)$ is continuous on x, then we define $f(x) = \partial F(x)/\partial x$ as the **(marginal) probability density function** (pdf) of X. Similarly, if the joint cdf $F(x, y)$ is continuous on both x and y, we define

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

as the **joint probability density function** of (X, Y), in which case we refer to the random variables as **jointly continuous**.

When X and Y are *independent*, we have

$$f(x, y) = f(x)f(y).$$

Given the joint density function, we may obtain the marginal density function of Y as

$$f(y) = \int_x f(x, y)\, dx,$$

where the integral is over all possible values of x, and the marginal probability function of X can be obtained in a similar manner.

Based on the joint pdf and the marginal pdf, we define the **conditional probability density function** of $(Y|X)$ as

$$f(y|x) = \frac{f(x, y)}{f(x)},$$

where we may obtain the conditional probability function of $(X|Y)$ in a similar manner. Here, the conditional density function is the density function of y given $X = x$. Hence, even in cases where $\Pr[X = x] = 0$ or when $f(x)$ is not defined, the function may be given in a particular form in real applications.

18.1.2 Conditional Expectation and Conditional Variance

Now we define the conditional expectation and variance based on the conditional distribution defined in the previous subsection.

Discrete Case

For a discrete random variable Y, its **expectation** is defined as $\mathrm{E}[Y] = \sum_y y\, p(y)$ if its value is finite, and its **variance** is defined as $\mathrm{Var}[Y] = \mathrm{E}\{(Y - \mathrm{E}[Y])^2\} = \sum_y y^2\, p(y) - \{\mathrm{E}[Y]\}^2$ if its value is finite.

For a discrete random variable Y, the **conditional expectation** of the random variable Y given the event $X = x$ is defined as

$$\mathrm{E}[Y|X = x] = \sum_y y\, p(y|x),$$

where X does not have to be a discrete variable, as far as the conditional probability function $p(y|x)$ is given.

Note that the conditional expectation $\mathrm{E}[Y|X = x]$ is a fixed number. When we replace x with X on the right-hand side of the above equation, we can define the expectation of Y given the random variable X as

$$\mathrm{E}[Y|X] = \sum_y y\, p(y|X),$$

which is still a *random variable*, and the randomness comes from X.

In a similar manner, we can define the **conditional variance** of the random variable Y given the event $X = x$ as

$$\mathrm{Var}[Y|X = x] = \mathrm{E}[Y^2|X = x] - \{\mathrm{E}[Y|X = x]\}^2 = \sum_y y^2\, p(y|x) - \{\mathrm{E}[Y|X = x]\}^2.$$

The variance of Y given X, $\mathrm{Var}[Y|X]$ can be defined by replacing x by X in the above equation, and $\mathrm{Var}[Y|X]$ is still a random variable and the randomness comes from X.

Continuous Case

For a continuous random variable Y, its **expectation** is defined as $\mathrm{E}[Y] = \int_y y\, f(y)dy$ if the integral exists, and its **variance** is defined as $\mathrm{Var}[Y] = \mathrm{E}\{(X - \mathrm{E}[Y])^2\} = \int_y y^2\, f(y)dy - \{\mathrm{E}[Y]\}^2$ if its value is finite.

For jointly continuous random variables X and Y, the **conditional expectation** of the random variable Y given $X = x$ is defined as

$$\mathrm{E}[Y|X = x] = \int_y y\, f(y|x)dy.$$

where X does not have to be a continuous variable, as far as the conditional probability function $f(y|x)$ is given.

Similarly, the conditional expectation $\mathrm{E}[Y|X = x]$ is a fixed number. When we replace x with X on the right-hand side of the above equation, we can define the expectation of Y given the random variable X as

$$\mathrm{E}[Y|X] = \int_y y\, p(y|X)\, dy,$$

which is still a *random variable*, and the randomness comes from X.

In a similar manner, we can define the **conditional variance** of the random variable Y given the event $X = x$ as

$$\mathrm{Var}[Y|X=x] = \mathrm{E}[Y^2|X=x] - \{\mathrm{E}[Y|X=x]\}^2 = \int_y y^2\, f(y|x)\, dy - \{\mathrm{E}[Y|X=x]\}^2.$$

The variance of Y given X, $\mathrm{Var}[Y|X]$ can then be defined by replacing x by X in the above equation, and similarly $\mathrm{Var}[Y|X]$ is also a random variable and the randomness comes from X.

18.2 Iterated Expectations and Total Variance

In this section, you learn

- the Law of Iterated Expectations for calculating the expectation of a random variable based on its conditional distribution given another random variable
- the Law of Total Variance for calculating the variance of a random variable based on its conditional distribution given another random variable
- how to calculate the expectation and variance based on an example of a two-stage model

18.2.1 Law of Iterated Expectations

Consider two random variables X and Y, and $h(X, Y)$, a random variable depending on the function h, X and Y.

Assuming all the expectations exist and are finite, the **Law of Iterated Expectations** states that

$$\mathrm{E}[h(X,Y)] = \mathrm{E}\left\{\mathrm{E}\left[h(X,Y)|X\right]\right\}, \tag{18.1}$$

where the first (inside) expectation is taken with respect to the random variable Y and the second (outside) expectation is taken with respect to X.

For the Law of Iterated Expectations, the random variables may be discrete, continuous, or a hybrid combination of the two. We use the example of discrete variables of X and Y to illustrate the calculation of the unconditional expectation using the Law of Iterated Expectations. For continuous random variables, we only need to replace the summation with the integral, as illustrated earlier in the appendix.

Given $p(y|x)$ the conditional pmf of X and Y, the conditional expectation of $h(X, Y)$ given the event $X = x$ is defined as

$$\mathrm{E}\left[h(X,Y)|X=x\right] = \sum_y h(x,y)p(y|x),$$

and the conditional expectation of $h(X, Y)$ given X being a *random variable* can be written as

$$\mathrm{E}\left[h(X,Y)|X\right] = \sum_y h(X,y)p(y|X).$$

The unconditional expectation of $h(X, Y)$ can then be obtained by taking the expectation of $\mathrm{E}\left[h(X,Y)|X\right]$ with respect to the random variable X. That is, we can obtain $\mathrm{E}[h(X,Y)]$ as

$$\begin{aligned} \mathrm{E}\left\{\mathrm{E}\left[h(X,Y)|X\right]\right\} &= \sum_x \left\{\sum_y h(x,y)p(y|x)\right\} p(x) \\ &= \sum_x \sum_y h(x,y)p(y|x)p(x) \\ &= \sum_x \sum_y h(x,y)p(x,y) = \mathrm{E}[h(X,Y)] \end{aligned} .$$

The Law of Iterated Expectations for the continuous and hybrid cases can be proved in a similar manner, by replacing the corresponding summation(s) by integral(s).

18.2.2 Law of Total Variance

Assuming that all the variances exist and are finite, the **Law of Total Variance** states that

$$\mathrm{Var}[h(X,Y)] = \mathrm{E}\left\{\mathrm{Var}\left[h(X,Y)|X\right]\right\} + \mathrm{Var}\left\{\mathrm{E}\left[h(X,Y)|X\right]\right\}, \qquad (18.2)$$

where the first (inside) expectation/variance is taken with respect to the random variable Y and the second (outside) expectation/variance is taken with respect to X. Thus, the unconditional variance equals to the expectation of the conditional variance plus the variance of the conditional expectation.

In order to verify this rule, first note that we can calculate a conditional variance as

$$\mathrm{Var}\left[h(X,Y)|X\right] = \mathrm{E}[h(X,Y)^2|X] - \left\{\mathrm{E}\left[h(X,Y)|X\right]\right\}^2 .$$

From this, the expectation of the conditional variance is

$$\begin{aligned}\mathrm{E}\{\mathrm{Var}\left[h(X,Y)|X\right]\} &= \mathrm{E}\left\{\mathrm{E}\left[h(X,Y)^2|X\right]\right\} - \mathrm{E}\left(\left\{\mathrm{E}\left[h(X,Y)|X\right]\right\}^2\right) \\ &= \mathrm{E}\left[h(X,Y)^2\right] - \mathrm{E}\left(\left\{\mathrm{E}\left[h(X,Y)|X\right]\right\}^2\right) . \qquad (18.3)\end{aligned}$$

Further, note that the conditional expectation, $\mathrm{E}\left[h(X,Y)|X\right]$, is a function of X, denoted $g(X)$. Thus, $g(X)$ is a random variable with mean $\mathrm{E}[h(X,Y)]$ and variance

$$\begin{aligned}\mathrm{Var}\left\{\mathrm{E}\left[h(X,Y)|X\right]\right\} &= \mathrm{Var}[g(X)] \\ &= \mathrm{E}[g(X)^2] \ - \left\{\mathrm{E}[g(X)]\right\}^2 \\ &= \mathrm{E}\left(\left\{\mathrm{E}\left[h(X,Y)|X\right]\right\}^2\right) - \left\{\mathrm{E}[h(X,Y)]\right\}^2 . \qquad (18.4)\end{aligned}$$

Thus, adding Equations
eqrefeq:AppBEV1 and
eqrefeq:AppBVE2 leads to the unconditional variance $\mathrm{Var}\left[h(X,Y)\right]$.

18.2.3 Application

To apply the Law of Iterated Expectations and the Law of Total Variance, we generally adopt the following procedure.

1. Identify the random variable that is being conditioned upon, typically a stage 1 outcome (that is not observed).
2. Conditional on the stage 1 outcome, calculate summary measures such as a mean, variance, and the like.
3. There are several results of the step 2, one for each stage 1 outcome. Then, combine these results using the iterated expectations or total variance rules.

Mixtures of Finite Populations. Suppose that the random variable N_1 represents a realization of the number of claims in a policy year from the population of good drivers and N_2 represents that from the population of bad drivers. For a specific driver, there is a probability α that (s)he is a good driver.

For a specific draw N, we have

$$N = \begin{cases} N_1, & \text{if (s)he is a good driver;} \\ N_2, & \text{otherwise.} \end{cases}$$

Let T be the indicator whether (s)he is a good driver, with $T = 1$ representing that the driver is a good driver with $\Pr[T = 1] = \alpha$ and $T = 2$ representing that the driver is a bad driver with $\Pr[T = 2] = 1 - \alpha$.

From equation (18.1), we can obtain the expected number of claims as

$$\mathrm{E}[N] = \mathrm{E}\left\{\mathrm{E}\left[N|T\right]\right\} = \mathrm{E}[N_1] \times \alpha + \mathrm{E}[N_2] \times (1 - \alpha).$$

From equation (18.2), we can obtain the variance of N as

$$\mathrm{Var}[N] = \mathrm{E}\left\{\mathrm{Var}\left[N|T\right]\right\} + \mathrm{Var}\left\{\mathrm{E}\left[N|T\right]\right\}.$$

To be more concrete, suppose that N_j follows a Poisson distribution with the mean λ_j, $j = 1, 2$. Then we have

$$\mathrm{Var}[N|T = j] = \mathrm{E}[N|T = j] = \lambda_j, \quad j = 1, 2.$$

Thus, we can derive the expectation of the conditional variance as

$$\mathrm{E}\left\{\mathrm{Var}\left[N|T\right]\right\} = \alpha\lambda_1 + (1 - \alpha)\lambda_2$$

and the variance of the conditional expectation as

$$\mathrm{Var}\left\{\mathrm{E}\left[N|T\right]\right\} = (\lambda_1 - \lambda_2)^2\alpha(1 - \alpha).$$

Note that the later is the variance for a Bernoulli with outcomes λ_1 and λ_2, and the binomial probability α.

Based on the Law of Total Variance, the unconditional variance of N is given by

$$\mathrm{Var}[N] = \alpha\lambda_1 + (1 - \alpha)\lambda_2 + (\lambda_1 - \lambda_2)^2\alpha(1 - \alpha).$$

18.3 Conjugate Distributions

As described in Section 9.3, for conjugate distributions the posterior and the prior come from the same family of distributions. In insurance applications, this broadly occurs in a "family of distribution families" known as the linear exponential family which we introduce first.

18.3.1 Linear Exponential Family

Definition. The distribution function of the *linear exponential family* is

$$f(x;\gamma,\theta)=\exp\left(\frac{x\gamma-b(\gamma)}{\theta}+S\left(x,\theta\right)\right).$$

Here, x is a dependent variable and γ is the parameter of interest. The quantity θ is a scale parameter. The term $b(\gamma)$ depends only on the parameter γ, not the dependent variable. The statistic $S\left(x,\theta\right)$ is a function of the dependent variable and the scale parameter, not the parameter γ.

The dependent variable x may be discrete, continuous or a hybrid combination of the two. Thus, $f\left(\cdot\right)$ may be interpreted to be a density or mass function, depending on the application. Table 18.1 provides several examples, including the normal, binomial and Poisson distributions.

Table 18.1. Selected Distributions of the Linear Exponential Family

Distribution	Parameters	Density or Mass Function	Components
General	$\gamma,\ \theta$	$\exp\left(\frac{x\gamma-b(\gamma)}{\theta}+S\left(x,\theta\right)\right)$	$\gamma,\ \theta, b(\gamma), S(x,\theta)$
Normal	μ,σ^2	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\mu,\sigma^2,\frac{\gamma^2}{2},-\left(\frac{x^2}{2\theta}+\frac{\log(2\pi\theta)}{2}\right)$
Binomal	π	$\binom{n}{x}\pi^x(1-\pi)^{n-x}$	$\log\left(\frac{\pi}{1-\pi}\right), 1, n\log(1+e^{\gamma}),$ $\log\binom{n}{x}$
Poisson	λ	$\frac{\lambda^x}{x!}\exp(-\lambda)$	$\log\lambda, 1, e^{\gamma}, -\log(x!)$
Negative Binomial*	r,p	$\frac{\Gamma(x+r)}{x!\Gamma(r)}p^r(1-p)^x$	$\log(1-p), 1, -r\log(1-e^{\gamma}),$ $\log\left[\frac{\Gamma(x+r)}{x!\Gamma(r)}\right]$
Gamma	α,γ	$\frac{1}{\Gamma(\alpha)\gamma^{\alpha}}x^{\alpha-1}\exp(-x/\gamma)$	$-\frac{\gamma}{\alpha},\frac{1}{\alpha},-\log(-\gamma),-\gamma^{-1}\log\gamma$ $-\log\left(\Gamma(\gamma^{-1})\right)+(\gamma^{-1}-1)\log x$

*This assumes that the parameter r is fixed but need not be an integer.

The Tweedie (see Section **??**) and inverse Gaussian distributions are also members of the linear exponential family. The linear exponential family of distribution families is extensively used as the basis of generalized linear models as described in, for example, Frees (2009).

18.3.2 Conjugate Distributions

Now assume that the parameter γ is random with distribution $\pi(\gamma,\tau)$, where τ is a vector of parameters that describe the distribution of γ. In Bayesian models, the distribution π is known as the prior and reflects our belief or information about γ. The likelihood $f(x|\gamma)$ is a probability conditional on γ.

The distribution of γ with knowledge of the random variables, $\pi(\gamma,\tau|x)$, is called the posterior distribution. For a given likelihood distribution, priors and posteriors that come from the same parametric family are known as conjugate families of distributions.

For a linear exponential likelihood, there exists a natural conjugate family. Specifically, consider a likelihood of the form $f(x|\gamma) = \exp\{(x\gamma - b(\gamma))/\theta\}\exp\{S(x,\theta)\}$. For this likelihood, define the prior distribution

$$\pi(\gamma,\tau) = C\exp\{\gamma a_1(\tau) - b(\gamma)a_2(\tau))\},$$

where C is a normalizing constant. Here, $a_1(\tau) = a_1$ and $a_2(\tau) = a_2$ are functions of the parameters τ although we simplify the notation by dropping explicit dependence on τ. The joint distribution of x and γ is given by $f(x,\gamma) = f(x|\gamma)\pi(\gamma,\tau)$. Using Bayes Theorem, the posterior distribution is

$$\pi(\gamma,\tau|x) = C_1\exp\left\{\gamma\left(a_1 + \frac{x}{\theta}\right) - b(\gamma)\left(a_2 + \frac{1}{\theta}\right)\right\},$$

where C_1 is a normalizing constant. Thus, we see that $\pi(\gamma,\tau|x)$ has the same form as $\pi(\gamma,\tau)$.

Special case. Gamma-Poisson Model. Consider a Poisson likelihood so that $b(\gamma) = e^\gamma$ and scale parameter (θ) equals one. Thus, we have

$$\pi(\gamma,\tau) = C\exp\{\gamma a_1 - a_2 e^\gamma\} = C\ \left(e^\gamma\right)^{a_1}\exp\left(-a_2 e^\gamma\right).$$

From the table of exponential family distributions, we recognize this to be a gamma distribution. That is, we have that the prior distribution of $\lambda = e^\gamma$ is a gamma distribution with parameters $\alpha_{prior} = a_1 + 1$ and $\theta_{prior}^{-1} = a_2$. The posterior distribution is a gamma distribution with parameters $\alpha_{post} = a_1 + x + 1 = \alpha_{prior} + x$ and $\theta_{post}^{-1} = a_2 + 1 = \theta_{prior}^{-1} + 1$.

Special case. Normal-Normal Model. Consider a normal likelihood so that $b(\gamma) = \gamma^2/2$ and the scale parameter is σ^2. Thus, we have

$$\pi(\gamma,\tau) = C\exp\left\{\gamma a_1 - \frac{\gamma^2}{2}a_2\right\} = C_1(\tau)\exp\left\{-\frac{a_2}{2}\left(\gamma - \frac{a_1}{a_2}\right)^2\right\},$$

The prior distribution of γ is normal with mean a_1/a_2 and variance a_2^{-1}. The posterior distribution of γ given x is normal with mean $(a_1 + x/\sigma^2)/(a_2 + \sigma^{-2})$ and variance $(a_2 + \sigma^{-2})^{-1}$.

Special case. Beta-Binomial Model. Consider a binomial likelihood so that $b(\gamma) = n \log(1+e^{\gamma})$ and scale parameter equals one. Thus, we have

$$\pi(\gamma,\tau) = C \exp\left\{\gamma a_1 - n a_2 \log(1+e^{\gamma})\right\} = C \left(\frac{e^{\gamma}}{1+e^{\gamma}}\right)^{a_1} \left(1-\frac{e^{\gamma}}{1+e^{\gamma}}\right)^{-na_2+a_1}.$$

This is a beta distribution. As in the other cases, prior parameters a_1 and a_2 are updated to become posterior parameters $a_1 + x$ and $a_2 + 1$.

Contributors

- **Lei (Larry) Hua**, Northern Illinois University, and **Edward (Jed) Frees**, University of Wisconsin-Madison, are the principal authors of the initial version of this chapter. Email: lhua@niu.edu or jfrees@bus.wisc.edu for chapter comments and suggested improvements.
  - The chapter was reviewed by Benjamin Côté.

19

Appendix C: Maximum Likelihood Theory

Chapter Preview. Appendix Chapter 17 introduced the maximum likelihood theory regarding estimation of parameters from a parametric family. This appendix gives more specific examples and expands some of the concepts. Section 19.1 reviews the definition of the likelihood function, and introduces its properties. Section 19.2 reviews the maximum likelihood estimators, and extends their large-sample properties to the case where there are multiple parameters in the model. Section 19.3 reviews statistical inference based on maximum likelihood estimators, with specific examples on cases with multiple parameters.

19.1 Likelihood Function

In this section, you learn

- the definitions of the likelihood function and the log-likelihood function
- the properties of the likelihood function

From Appendix Chapter 17, the likelihood function is a function of parameters given the observed data. Here, we review the concepts of the likelihood function, and introduces its properties that are bases for maximum likelihood inference.

19.1.1 Likelihood and Log-likelihood Functions

Here, we give a brief review of the likelihood function and the log-likelihood function from Appendix Chapter 17. Let $f(\cdot|\boldsymbol{\theta})$ be the probability function of X, the probability mass function (pmf) if X is discrete or the probability density function (pdf) if it is continuous. The likelihood is a function of the parameters ($\boldsymbol{\theta}$) given the data ($\mathbf{x}$). Hence, it is a function of the parameters with the data being fixed, rather than a function of the data with the parameters being fixed. The vector of data $\mathbf{x}$ is usually a realization of a *random sample* as defined in Appendix Chapter 17.

Given a realized random sample $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ of size n, the **likelihood function** is defined as

$$L(\boldsymbol{\theta}|\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}),$$

with the corresponding **log-likelihood function** given by

$$l(\boldsymbol{\theta}|\mathbf{x}) = \log L(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta}),$$

where $f(\mathbf{x}|\boldsymbol{\theta})$ denotes the joint probability function of $\mathbf{x}$. The log-likelihood function leads to an additive structure that is easy to work with.

In Appendix Chapter 17, we have used the normal distribution to illustrate concepts of the likelihood function and the log-likelihood function. Here, we derive the likelihood and corresponding log-likelihood functions when the population distribution is from the Pareto distribution family.

Example – Pareto Distribution. Suppose that $X_1, \ldots, X_n$ represents a random sample from a single-parameter Pareto distribution with the **cumulative distribution function** given by

$$F(x) = \Pr(X_i \leq x) = 1 - \left(\frac{500}{x}\right)^{\alpha}, \qquad x > 500,$$

with parameter $\theta = \alpha$.

The corresponding probability density function is $f(x) = 500^{\alpha} \alpha x^{-\alpha-1}$ and the log-likelihood function can be derived as

$$l(\boldsymbol{\alpha}|\mathbf{x}) = \sum_{i=1}^{n} \log f(x_i;\alpha) = n\alpha \log 500 + n \log \alpha - (\alpha+1) \sum_{i=1}^{n} \log x_i.$$

19.1.2 Properties of Likelihood Functions

In mathematical statistics, the first derivative of the log-likelihood function with respect to the parameters, $u(\boldsymbol{\theta}) = \partial l(\boldsymbol{\theta}|\mathbf{x})/\partial \boldsymbol{\theta}$, is referred to as the **score function**, or the **score vector** when there are multiple parameters in $\boldsymbol{\theta}$. The score function or score vector can be written as

$$u(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{x}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(x_i;\boldsymbol{\theta}),$$

where $u(\boldsymbol{\theta}) = (u_1(\boldsymbol{\theta}), u_2(\boldsymbol{\theta}), \cdots, u_p(\boldsymbol{\theta}))$ when $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)$, with the element $u_k(\boldsymbol{\theta}) = \partial l(\boldsymbol{\theta}|\mathbf{x})/\partial\theta_k$ being the partial derivative with respect to θ_k $(k = 1, 2, \cdots, p)$.

The likelihood function has the following properties:

- One basic property of the likelihood function is that the expectation of the score function with respect to $\mathbf{x}$ is 0. That is,

$$\mathrm{E}[u(\boldsymbol{\theta})] = \mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{x})\right] = \mathbf{0}.$$

  To illustrate this, we have

$$\begin{aligned}\mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{x})\right] &= \mathrm{E}\left[\frac{\frac{\partial}{\partial\boldsymbol{\theta}} f(\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{x};\boldsymbol{\theta})}\right] = \int \frac{\partial}{\partial\boldsymbol{\theta}} f(\mathbf{y};\boldsymbol{\theta}) d\mathbf{y} \\ &= \frac{\partial}{\partial\boldsymbol{\theta}} \int f(\mathbf{y};\boldsymbol{\theta}) d\mathbf{y} = \frac{\partial}{\partial\boldsymbol{\theta}} 1 = \mathbf{0}.\end{aligned}$$

- Denote by $\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} l(\boldsymbol{\theta}|\mathbf{x})$ the second derivative of the log-likelihood function. This is a $p \times p$ matrix of second derivatives known as the hessian of the log-likelihood. Another basic property of the likelihood function is that the sum of the expectation of the hessian matrix and the expectation of the Kronecker product of the score vector and its transpose is $\mathbf{0}$. That is,

$$\mathrm{E}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} l(\boldsymbol{\theta}|\mathbf{x})\right) + \mathrm{E}\left(\frac{\partial l(\boldsymbol{\theta}|\mathbf{x})}{\partial\boldsymbol{\theta}} \frac{\partial l(\boldsymbol{\theta}|\mathbf{x})}{\partial\boldsymbol{\theta}'}\right) = \mathbf{0}.$$

- Define the **Fisher information matrix** as

$$\mathcal{I}(\boldsymbol{\theta}) = \mathrm{E}\left(\frac{\partial l(\boldsymbol{\theta}|\mathbf{x})}{\partial\boldsymbol{\theta}} \frac{\partial l(\boldsymbol{\theta}|\mathbf{x})}{\partial\boldsymbol{\theta}'}\right) = -\mathrm{E}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} l(\boldsymbol{\theta}|\mathbf{x})\right).$$

As the sample size n goes to infinity, the score function (vector) converges in distribution to a **normal distribution** (or **multivariate normal distribution** when $\boldsymbol{\theta}$ contains multiple parameters) with mean $\mathbf{0}$ and variance (or covariance matrix in the multivariate case) given by $\mathcal{I}(\boldsymbol{\theta})$.

19.2 Maximum Likelihood Estimators

In this section, you learn

- the definition and derivation of the maximum likelihood estimator (*mle*) for parameters from a specific distribution family
- the properties of maximum likelihood estimators that ensure valid large-sample inference of the parameters
- why using the *mle*-based method, and what caution that needs to be taken

---

In statistics, maximum likelihood estimators are values of the parameters $\boldsymbol{\theta}$ that are most likely to have been produced by the data.

19.2.1 Definition and Derivation of *MLE*

Based on the definition given in Appendix Chapter 17, the value of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}_{mle}$, that maximizes the likelihood function, is called the *maximum likelihood estimator* (*mle*) of $\boldsymbol{\theta}$.

Because the log function $\log(\cdot)$ is a one-to-one function, we can also determine $\hat{\boldsymbol{\theta}}_{mle}$ by maximizing the log-likelihood function, $l(\boldsymbol{\theta}|\mathbf{x})$. That is, the *mle* is defined as

$$\hat{\boldsymbol{\theta}}_{mle} = \text{argmax}_{\boldsymbol{\theta}\in\Theta}\ l(\boldsymbol{\theta}|\mathbf{x}).$$

Given the analytical form of the likelihood function, the *mle* can be obtained by taking the first derivative of the log-likelihood function with respect to $\boldsymbol{\theta}$, and setting the values of the partial derivatives to zero. That is, the *mle* are the solutions of the equations of

$$\frac{\partial l(\hat{\boldsymbol{\theta}}|\mathbf{x})}{\partial \hat{\boldsymbol{\theta}}} = \mathbf{0}.$$

Example. Course C/Exam 4. May 2000, 21. You are given the following five observations: 521, 658, 702, 819, 1217. You use the single-parameter Pareto with cumulative distribution function:

$$F(x) = 1 - \left(\frac{500}{x}\right)^{\alpha}, \qquad x > 500.$$

Calculate the maximum likelihood estimate of the parameter α.

Example Solution. With $n = 5$, the log-likelihood function is

$$l(\alpha|\mathbf{x}) = \sum_{i=1}^{5} \log f(x_i;\alpha) = 5\alpha \log 500 + 5 \log \alpha - (\alpha + 1) \sum_{i=1}^{5} \log x_i.$$

Solving for the root of the score function yields

$$\begin{aligned} \tfrac{\partial}{\partial\alpha} l(\alpha|\mathbf{x}) &= 5\log 500 + 5/\alpha - \textstyle\sum_{i=1}^{5} \log x_i \\ &=_{set} 0 \Rightarrow \hat{\alpha}_{mle} = \tfrac{5}{\sum_{i=1}^{5} \log x_i - 5\log 500} = 2.453. \end{aligned}$$

19.2.2 Asymptotic Properties of *MLE*

From Appendix Chapter 17, the MLE has some nice large-sample properties, under certain regularity conditions. We presented the results for a single parameter in Appendix Chapter 17, but results are true for the case when $\boldsymbol{\theta}$ contains multiple parameters. In particular, we have the following results, in a general case when $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_p)$.

- The *mle* of a parameter $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{mle}$, is a **consistent** estimator. That is, the *mle* $\hat{\boldsymbol{\theta}}_{mle}$ converges in probability to the true value $\boldsymbol{\theta}$, as the sample size n goes to infinity.
- The *mle* has the **asymptotic normality** property, meaning that the estimator will converge in distribution to a multivariate normal distribution centered around the true value, when the sample size goes to infinity. Namely,

  $$\sqrt{n}(\hat{\boldsymbol{\theta}}_{mle} - \boldsymbol{\theta}) \rightarrow N\left(\mathbf{0},\, \boldsymbol{V}\right), \quad \text{as} \quad n \rightarrow \infty,$$

  where $\boldsymbol{V}$ denotes the asymptotic variance (or covariance matrix) of the estimator. Hence, the *mle* $\hat{\boldsymbol{\theta}}_{mle}$ has an approximate normal distribution with mean $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{V}/n$, when the sample size is large.
- The *mle* is **efficient**, meaning that it has the smallest asymptotic variance $\boldsymbol{V}$, commonly referred to as the **Cramer–Rao lower bound**. In particular, the Cramer–Rao lower bound is the inverse of the Fisher information (matrix) $\mathcal{I}(\boldsymbol{\theta})$ defined earlier in this appendix. Hence, $\mathrm{Var}(\hat{\boldsymbol{\theta}}_{mle})$ can be estimated based on the observed Fisher information.

Based on the above results, we may perform statistical inference based on the procedures defined in Appendix Chapter 17.

Example. Course C/Exam 4. Nov 2000, 13. A sample of ten observations comes from a parametric family $f(x,;\theta_1,\theta_2)$ with log-likelihood function

$$l(\theta_1,\theta_2) = \sum_{i=1}^{10} f(x_i;\theta_1,\theta_2) = -2.5\theta_1^2 - 3\theta_1\theta_2 - \theta_2^2 + 5\theta_1 + 2\theta_2 + k,$$

where k is a constant. Determine the estimated covariance matrix of the maximum likelihood estimator, $\hat{\theta_1}, \hat{\theta_2}$.

Example Solution. Denoting $l = l(\theta_1, \theta_2)$, the hessian matrix of second derivatives is

$$\begin{pmatrix} \frac{\partial^2}{\partial \theta_1^2} l & \frac{\partial^2}{\partial \theta_1 \partial \theta_2} l \\ \frac{\partial^2}{\partial \theta_1 \partial \theta_2} l & \frac{\partial^2}{\partial \theta_1^2} l \end{pmatrix} = \begin{pmatrix} -5 & -3 \\ -3 & -2 \end{pmatrix}$$

Thus, the information matrix is:

$$\mathcal{I}(\theta_1, \theta_2) = -\mathrm{E}\left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} l(\boldsymbol{\theta}|\mathbf{x})\right) = \begin{pmatrix} 5 & 3 \\ 3 & 2 \end{pmatrix}$$

and

$$\mathcal{I}^{-1}(\theta_1, \theta_2) = \frac{1}{5(2) - 3(3)} \begin{pmatrix} 2 & -3 \\ -3 & 5 \end{pmatrix} = \begin{pmatrix} 2 & -3 \\ -3 & 5 \end{pmatrix}.$$

19.2.3 Use of Maximum Likelihood Estimation

The method of maximum likelihood has many advantages over alternative methods such as the method of moments introduced in Appendix Chapter 17.

- It is a general tool that works in many situations. For example, we may be able to write out the closed-form likelihood function for censored and truncated data. Maximum likelihood estimation can be used for regression models including covariates, such as survival regression, generalized linear models and mixed models, that may include covariates that are time-dependent.
- From the efficiency of the *mle*, it is optimal, the best, in the sense that it has the smallest variance among the class of all unbiased estimators for large sample sizes.
- From the results on the asymptotic normality of the *mle*, we can obtain a large-sample distribution for the estimator, allowing users to assess the variability in the estimation and perform statistical inference on the parameters. The approach is less computationally extensive than re-sampling methods that require a large number of fittings of the model.

Despite its numerous advantages, *mle* has its drawback in cases such as generalized linear models when it does not have a closed analytical form. In such cases, maximum likelihood estimators are computed iteratively using numerical optimization methods. For example, we may use the Newton-Raphson iterative algorithm or its variations for obtaining the *mle*. Iterative algorithms require starting values. For some problems, the choice of a close starting value is critical, particularly in cases where the likelihood function has local minimums or maximums. Hence, there may be a convergence issue when the starting value is far from the maximum. It is important to start from different values across the parameter space and compare the maximized likelihood or log-likelihood to make sure the algorithms have converged to a global maximum.

19.3 Statistical Inference Based on Maximum Likelihood Estimation

In this section, you learn how to

- perform hypothesis testing based on *mle* for cases where there are multiple parameters in $\boldsymbol{\theta}$
- perform likelihood ratio test for cases where there are multiple parameters in $\boldsymbol{\theta}$

In Appendix Chapter 17, we have introduced maximum likelihood based methods for statistical inference when $\boldsymbol{\theta}$ contains a single parameter. Here, we will extend the results to cases where there are multiple parameters in $\boldsymbol{\theta}$.

19.3.1 Hypothesis Testing

In Appendix Chapter 17, we defined hypothesis testing concerning the null hypothesis, a statement on the parameter(s) of a distribution or model. One important type of inference is to assess whether a parameter estimate is statistically significant, meaning whether the value of the parameter is zero or not.

We have learned earlier that the *mle* $\hat{\boldsymbol{\theta}}_{mle}$ has a large-sample normal distribution with mean $\boldsymbol{\theta}$ and the variance-covariance matrix $\mathcal{I}^{-1}(\boldsymbol{\theta})$. Based on the multivariate normal distribution, the jth element of $\hat{\boldsymbol{\theta}}_{mle}$, say $\hat{\theta}_{MLE,j}$, has a large-sample univariate normal distribution.

Define $se(\hat{\theta}_{MLE,j})$, the standard error (estimated standard deviation) to be the square root of the jth diagonal element of $\mathcal{I}^{-1}(\boldsymbol{\theta})_{mle}$. To assess the null hypothesis that $\theta_j = \theta_0$, we define the t-statistic or t-ratio to be $t(\hat{\theta}_{MLE,j}) = (\hat{\theta}_{MLE,j} - \theta_0)/se(\hat{\theta}_{MLE,j})$.

Under the null hypothesis, it has a Student-t distribution with degrees of freedom equal to $n - p$, with p being the dimension of $\boldsymbol{\theta}$.

For most actuarial applications, we have a large sample size n, so the t-distribution is very close to the (standard) normal distribution. In the case when n is very large or when the standard error is known, the t-statistic can be referred to as a z-statistic or z-score.

Based on the results from Appendix Chapter 17, if the t-statistic $t(\hat{\theta}_{MLE,j})$ exceeds a cut-off (in absolute value), then the test for the j parameter θ_j is said to be statistically significant. If θ_j is the regression coefficient of the j th independent variable, then we say that the jth variable is statistically significant.

For example, if we use a 5% significance level, then the cut-off value is 1.96 using a normal distribution approximation for cases with a large sample size. More generally, using a $100\alpha\%$ significance level, then the cut-off is a $100(1-\alpha/2)\%$ quantile from a Student-t distribution with the degree of freedom being $n-p$.

Another useful concept in hypothesis testing is the p-value, shorthand for probability value. From the mathematical definition in Appendix Chapter 17, a p-value is defined as the smallest significance level for which the null hypothesis would be rejected. Hence, the p-value is a useful summary statistic for the data analyst to report because it allows the reader to understand the strength of statistical evidence concerning the deviation from the null hypothesis.

19.3.2 *MLE* and Model Validation

In addition to hypothesis testing and interval estimation introduced in Appendix Chapter 17 and the previous subsection, another important type of inference is selection of a model from two choices, where one choice is a special case of the other with certain parameters being restricted. For such two models with one being nested in the other, we have introduced the likelihood ratio test (LRT) in Appendix Chapter 17. Here, we will briefly review the process of performing a LRT based on a specific example of two alternative models.

Suppose that we have a (large) model under which we derive the maximum likelihood estimator, $\hat{\boldsymbol{\theta}}_{mle}$. Now assume that some of the p elements in $\boldsymbol{\theta}$ are equal to zero and determine the maximum likelihood estimator over the remaining set, with the resulting estimator denoted $\hat{\boldsymbol{\theta}}_{Reduced}$.

Based on the definition in Appendix Chapter 17, the statistic, $LRT = 2\left(l(\hat{\boldsymbol{\theta}}_{mle}) - l(\hat{\boldsymbol{\theta}}_{Reduced})\right)$, is called the likelihood ratio statistic. Under the null hypothesis that the reduced model is correct, the likelihood ratio has a chi-square distribution with degrees of freedom equal to d, the number of variables set to zero.

Such a test allows us to judge which of the two models is more likely to be correct, given the observed data. If the statistic LRT is large relative to the critical value from the chi-square distribution, then we reject the reduced model in favor of the larger one. Details regarding the critical value and alternative methods based on information criteria are given in Appendix Chapter 17.

Contributors

- **Lei (Larry) Hua**, Northern Illinois University, and **Edward (Jed) Frees**, University of Wisconsin-Madison, are the principal authors of the initial version of this chapter. Email: lhua@niu.edu or jfrees@bus.wisc.edu for chapter comments and suggested improvements.
  - The chapter was reviewed by Benjamin Côté.

20

Appendix D: Summary of Distributions

User Notes

- The R functions are from the packages actuar and invgamma.
- Tables appear when first loaded by the browser. To hide them, click on one of the distributions, e.g., *Poisson*, and then click on the *Hide* button.
- More information on the R codes is available at the R Codes for Loss Data Analytics site.

20.1 Discrete Distributions

Overview. This section summarizes selected discrete probability distributions used throughout *Loss Data Analytics*. Relevant functions and R code are provided.

20.1.1 The *(a,b,0)* Class

Poisson

Functions

Name	Function
Parameter assumptions	$\lambda > 0$
p_0	$e^{-\lambda}$
Probability mass function p_k	$\frac{e^{-\lambda}\lambda^k}{k!}$
Expected value $\mathrm{E}[N]$	λ
Variance	λ
Probability generating function $P(z)$	$e^{\lambda(z-1)}$
a and b for recursion	$a = 0$ $b = \lambda$

R Commands

Function Name	R Command
Probability mass function	dpois($x =, lambda = \lambda$)
Distribution function	ppois($p =, lambda = \lambda$)
Quantile function	qpois($q =, lambda = \lambda$)
Random sampling function	rpois($n =, lambda = \lambda$)

Geometric

Functions

Name	Function
Parameter assumptions	$\beta > 0$
p_0	$\frac{1}{1+\beta}$
Probability mass function p_k	$\frac{\beta^k}{(1+\beta)^{k+1}}$
Expected value E$[N]$	β
Variance	$\beta(1+\beta)$
Probability generating function $P(z)$	$[1-\beta(z-1)]^{-1}$
a and b for recursion	$a = \frac{\beta}{1+\beta}$ $b = 0$

R Commands

Function Name	R Command
Probability mass function	dgeom($x =, prob = \frac{1}{1+\beta}$)
Distribution function	pgeom($p =, prob = \frac{1}{1+\beta}$)
Quantile function	qgeom($q =, prob = \frac{1}{1+\beta}$)
Random sampling function	rgeom($n =, prob = \frac{1}{1+\beta}$)

Binomial

Functions

Name	Function
Parameter assumptions	$0 < q < 1,$ m is an integer $0 \le k \le m$
p_0	$(1-q)^m$
Probability mass function p_k	$\binom{m}{k} q^k (1-q)^{m-k}$
Expected value $\mathrm{E}[N]$	mq
Variance	$mq(1-q)$
Probability generating function $P(z)$	$[1+q(z-1)]^m$
a and b for recursion	$a = \frac{-q}{1-q}$ $b = \frac{(m+1)q}{1-q}$

R Commands

Function Name	R Command
Probability mass function	dbinom$(x =, size = m, prob = q)$
Distribution function	pbinom$(p =, size = m, prob = q)$
Quantile function	qbinom$(q =, size = m, prob = q)$
Random sampling function	rbinom$(n =, size = m, prob = q)$

Negative Binomial

Functions

Name	Function
Parameter assumptions	$r > 0, \beta > 0$
p_0	$(1+\beta)^{-r}$
Probability mass function p_k	$\frac{r(r+1)\cdots(r+k-1)\beta^k}{k!(1+\beta)^{r+k}}$
Expected value $\mathrm{E}[N]$	$r\beta$
Variance	$r\beta(1+\beta)$
Probability generating function $P(z)$	$[1-\beta(z-1)]^{-r}$
a and b for recursion	$a = \frac{\beta}{1+\beta}$ $b = \frac{(r-1)\beta}{1+\beta}$

R Commands

Function Name	R Command
Probability mass function	dnbinom$(x =, size = r, prob = \frac{1}{1+\beta})$
Distribution function	pnbinom$(p =, size = r, prob = \frac{1}{1+\beta})$
Quantile function	qnbinom$(q =, size = r, prob = \frac{1}{1+\beta})$
Random sampling function	rnbinom$(n =, size = r, prob = \frac{1}{1+\beta})$

20.1.2 The *(a,b,1)* Class

Zero Truncated Poisson

Functions

Name	Function
Parameter assumptions	$\lambda > 0$
p_1^T	$\frac{\lambda}{e^\lambda-1}$
Probability mass function p_k^T	$\frac{\lambda^k}{k!(e^\lambda-1)}$
Expected value $\mathrm{E}[N]$	$\frac{\lambda}{1-e^{-\lambda}}$
Variance	$\frac{\lambda[1-(\lambda+1)e^{-\lambda}]}{(1-e^{-\lambda})^2}$
Probability generating function $P(z)$	$\frac{e^{\lambda z}-1}{e^\lambda-1}$
a and b for recursion	$a = 0$ $b = \lambda$

R Commands

Function Name	R Command
Probability mass function	dztpois$(x =, lambda = \lambda)$
Distribution function	pztpois$(p =, lambda = \lambda)$
Quantile function	qztpois$(q =, lambda = \lambda)$
Random sampling function	rztpois$(n =, lambda = \lambda)$

Zero Truncated Geometric

Functions

Name	Function
Parameter assumptions	$\beta > 0$
p_1^T	$\frac{1}{1+\beta}$
Probability mass function p_k^T	$\frac{\beta^{k-1}}{(1+\beta)^k}$
Expected value $\mathrm{E}[N]$	$1+\beta$
Variance	$\beta(1+\beta)$
Probability generating function $P(z)$	$\frac{[1-\beta(z-1)]^{-1}-(1+\beta)^{-1}}{1-(1+\beta)^{-1}}$
a and b for recursion	$a=\frac{\beta}{1+\beta}$ $b=0$

R Commands

Function Name	R Command
Probability mass function	dztgeom$(x=, prob=\frac{1}{1+\beta})$
Distribution function	pztgeom$(p=, prob=\frac{1}{1+\beta})$
Quantile function	qztgeom$(q=, prob=\frac{1}{1+\beta})$
Random sampling function	rztgeom$(n=, prob=\frac{1}{1+\beta})$

Zero Truncated Binomial

Functions

Name	Function
Parameter assumptions	$0<q<1,$ m is an integer $0 \leq k \leq m$
p_1^T	$\frac{m(1-q)^{m-1}q}{1-(1-q)^m}$
Probability mass function p_k^T	$\frac{\binom{m}{k}q^k(1-q)^{m-k}}{1-(1-q)^m}$
Expected value $\mathrm{E}[N]$	$\frac{mq}{1-(1-q)^m}$
Variance	$\frac{mq[(1-q)-(1-q+mq)(1-q)^m]}{[1-(1-q)^m]^2}$
Probability generating function $P(z)$	$\frac{[1+q(z-1)^m]-(1-q)^m}{1-(1-q)^m}$
a and b for recursion	$a=\frac{-q}{1-q}$ $b=\frac{(m+1)q}{1-q}$

R Commmands

Function Name	R Command
Probability mass function	dztbinom$(x =, size = m, prob = p)$
Distribution function	pztbinom$(p =, size = m, prob = p)$
Quantile function	qztbinom$(q =, size = m, prob = p)$
Random sampling function	rztbinom$(n =, size = m, prob = p)$

Zero Truncated Negative Binomial

Functions

Name	Function
Parameter assumptions	$r > -1, r \neq 0$
p_1^T	$\frac{r\beta}{(1+\beta)^{r+1}-(1+\beta)}$
Probability mass function p_k^T	$\frac{r(r+1)\cdots(r+k-1)}{k![(1+\beta)^r-1]}(\frac{\beta}{1+\beta})^k$
Expected value $\mathrm{E}[N]$	$\frac{r\beta}{1-(1+\beta)^{-r}}$
Variance	$\frac{r\beta[(1+\beta)-(1+\beta+r\beta)(1+\beta)^{-r}]}{[1-(1+\beta)^{-r}]^2}$
Probability generating function $P(z)$	$\frac{[1-\beta(z-1)]^{-r}-(1+\beta)^{-r}}{1-(1+\beta)^{-r}}$
a and b for recursion	$a = \frac{\beta}{1+\beta}$ $b = \frac{(r-1)\beta}{1+\beta}$

R Commands

Function Name	R Command
Probability mass function	dztnbinom$(x =, size = r, prob = \frac{1}{1+\beta})$
Distribution function	pztnbinom$(p =, size = r, prob = \frac{1}{1+\beta})$
Quantile function	qztnbinom$(q =, size = r, prob = \frac{1}{1+\beta})$
Random sampling function	rztnbinom$(n =, size = r, prob = \frac{1}{1+\beta})$

Logarithmic

Functions

Name	Function
Parameter assumptions	$\beta > 0$
p_1^T	$\frac{\beta}{(1+\beta) ln(1+\beta)}$
Probability mass function p_k^T	$\frac{\beta^k}{k(1+\beta)^k \ln(1+\beta)}$
Expected value $\mathrm{E}[N]$	$\frac{\beta}{\ln(1+\beta)}$
Variance	$\frac{\beta[1+\beta-\frac{\beta}{ln(1+\beta)}]}{\ln(1+\beta)}$
Probability generating function $P(z)$	$1-\frac{ln[1-\beta(z-1)]}{\ln(1+\beta)}$
a and b for recursion	$a=\frac{\beta}{1+\beta}$ $b=\frac{-\beta}{1+\beta}$

`R` Commands

Function Name	R Command
Probability mass function	dnbinom($x=, prob=\frac{\beta}{1+\beta}$)
Distribution function	pnbinom($p=, prob=\frac{\beta}{1+\beta}$)
Quantile function	qnbinom($q=, prob=\frac{\beta}{1+\beta}$)
Random sampling function	rnbinom($n=, prob=\frac{\beta}{1+\beta}$)

20.2 Continuous Distributions

Overview. This section summarizes selected continuous probability distributions used throughout *Loss Data Analytics*. Relevant functions, `R` code, and illustrative graphs are provided.

20.2.1 One Parameter Distributions

Exponential

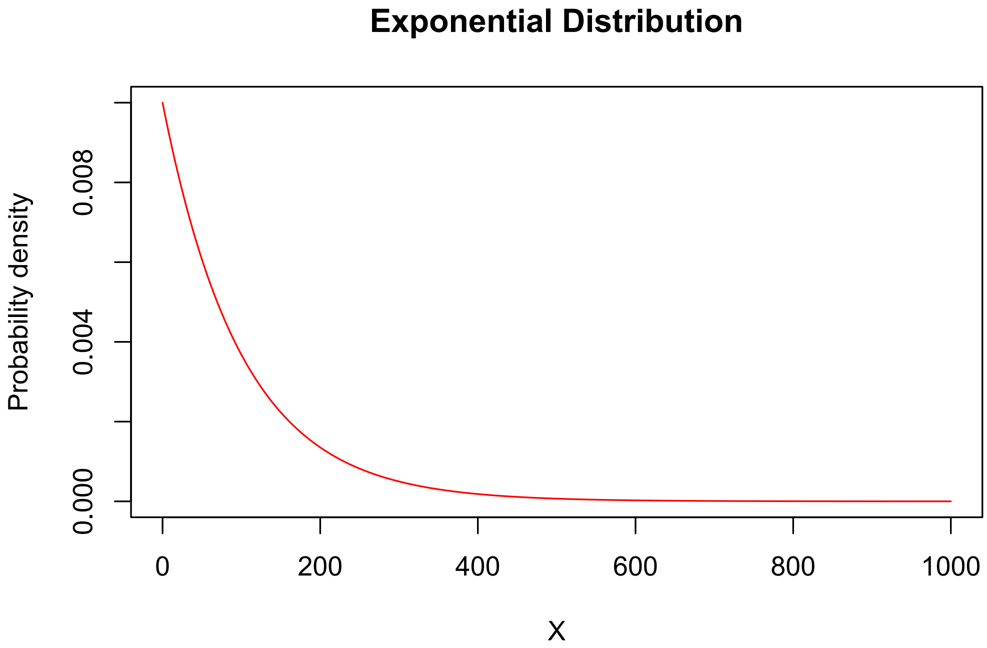
Functions

Name	Function
Parameter assumptions	$\theta > 0$
Probability density function $f(x)$	$\frac{1}{\theta}e^{-x/\theta}$
Distribution function $F(x)$	$1 - e^{-x/\theta}$
k^{th} raw moment $\mathrm{E}[X^k]$	$\theta^k\Gamma(k+1)$ $k > -1$
$VaR_p(x)$	$-\theta\ln(1-p)$
Limited Expected Value $\mathrm{E}[X \wedge x]$	$\theta(1 - e^{-x/\theta})$

R Commands

Function Name	R Command
Density function	dexp($x =, rate = 1/\theta$)
Distribution function	pexp($p =, rate = 1/\theta$)
Quantile function	qexp($q =, rate = 1/\theta$)
Random sampling function	rexp($n =, rate = 1/\theta$)

Illustrative Graph



Inverse Exponential

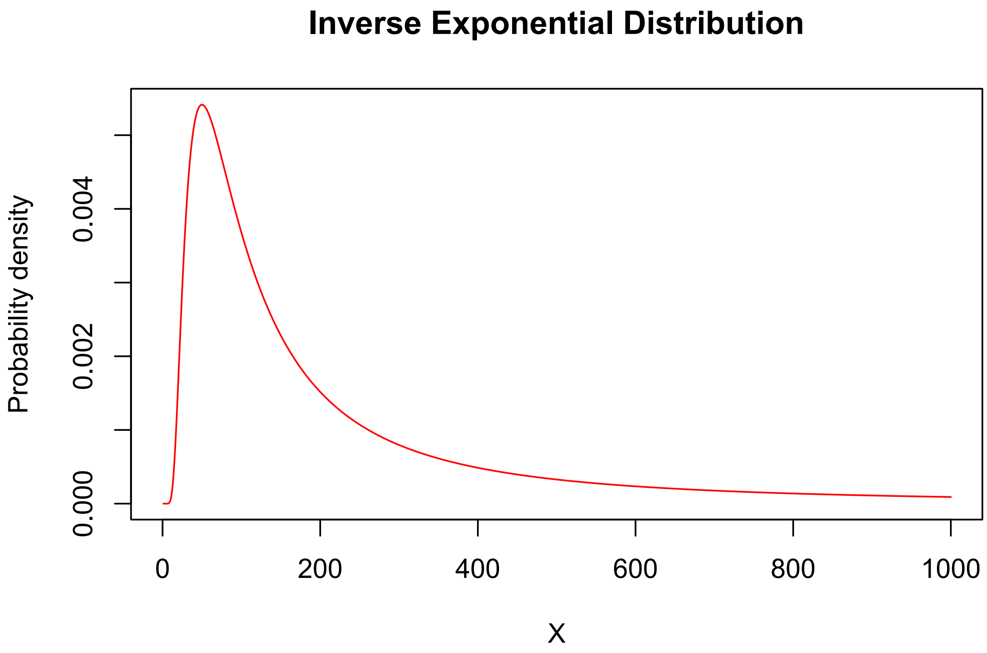
Functions

Name	Function
Parameter assumptions	$\theta > 0$
Probability density function $f(x)$	$\frac{\theta e^{-\theta/x}}{x^2}$
Distribution function $F(x)$	$e^{-\theta/x}$
k^{th} raw moment $\mathrm{E}[X^k]$	$\theta^k\Gamma(1-k)$ $k < 1$
$\mathrm{E}[(X \wedge x)^k]$	$\theta^k G(1-k;\theta/x) + x^k(1-e^{-\theta/x})$

R Commands

Function Name	R Command
Density function	dinvexp($x =, scale = \theta$)
Distribution function	pinvexp($p =, scale = \theta$)
Quantile function	qinvexp($q =, scale = \theta$)
Random sampling function	rinvexp($n =, scale = \theta$)

Illustrative Graph



Single Parameter Pareto

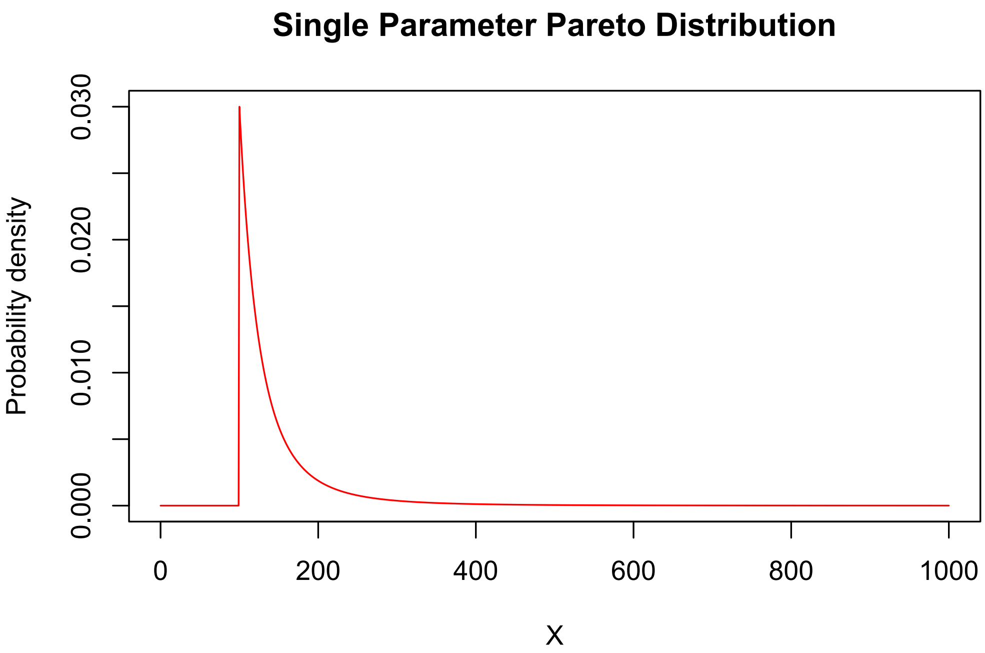
Functions

Name	Function
Parameter assumptions	θ is known, $x > \theta, \alpha > 0$
Probability density function $f(x)$	$\frac{\alpha\theta^\alpha}{x^{\alpha+1}}$
Distribution function $F(x)$	$1-(\theta/x)^\alpha$
k^{th} raw moment $\mathrm{E}[X^k]$	$\frac{\alpha\theta^k}{\alpha-k}$ $k<\alpha$
$\mathrm{E}[(X \wedge x)^k]$	$\frac{\alpha\theta^k}{\alpha-k}-\frac{k\theta^\alpha}{(\alpha-k)x^{\alpha-k}}$ $x \geq \theta$

R Commands

Function Name	R Command
Density function	$\text{dpareto1}(x =, shape = \alpha, min = \theta)$
Distribution function	$\text{ppareto1}(p =, shape = \alpha, min = \theta)$
Quantile function	$\text{qpareto1}(q =, shape = \alpha, min = \theta)$
Random sampling function	$\text{rpareto1}(n =, shape = \alpha, min = \theta)$

Illustrative Graph



20.2.2 Two Parameter Distributions

Pareto

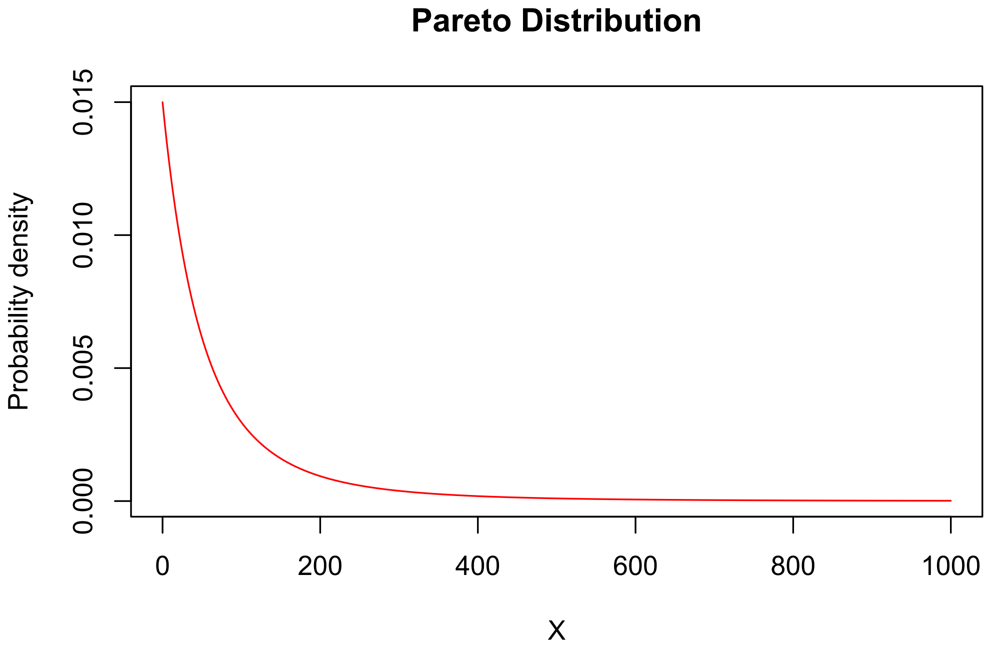
Functions

Name	Function
Parameter assumptions	$\theta > 0, \alpha > 0$
Probability density function $f(x)$	$\frac{\alpha\theta^{\alpha}}{(x+\theta)^{\alpha+1}}$
Distribution function $F(x)$	$1 - \left(\frac{\theta}{x+\theta}\right)^{\alpha}$
k^{th} raw moment $\mathrm{E}[X^k]$	$\frac{\theta^k\Gamma(k+1)\Gamma(\alpha-k)}{\Gamma(\alpha)}$ $-1 < k < \alpha$
Limited Expected Value: $\alpha \neq 1$ $\mathrm{E}[X \wedge x]$	$\frac{\theta}{\alpha-1}\left[1 - \left(\frac{\theta}{x+\theta}\right)^{\alpha-1}\right]$
Limited Expected Value: $\alpha = 1$ $\mathrm{E}[X \wedge x]$	$-\theta \ln\left(\frac{\theta}{x+\theta}\right)$
$\mathrm{E}[(X \wedge x)^k]$	$\frac{\theta^k\Gamma(k+1)\Gamma(\alpha-k)}{\Gamma(\alpha)}\beta(k+1, \alpha-k; \frac{x}{x+\theta}) + x^k(\frac{\theta}{x+\theta})^{\alpha}$

R Commands

Function Name	R Command
Density function	dpareto$(x =, shape = \alpha, scale = \theta)$
Distribution function	ppareto$(p =, shape = \alpha, scale = \theta)$
Quantile function	qpareto$(q =, shape = \alpha, scale = \theta)$
Random sampling function	rpareto$(n =, shape = \alpha, scale = \theta)$

Illustrative Graph



Inverse Pareto

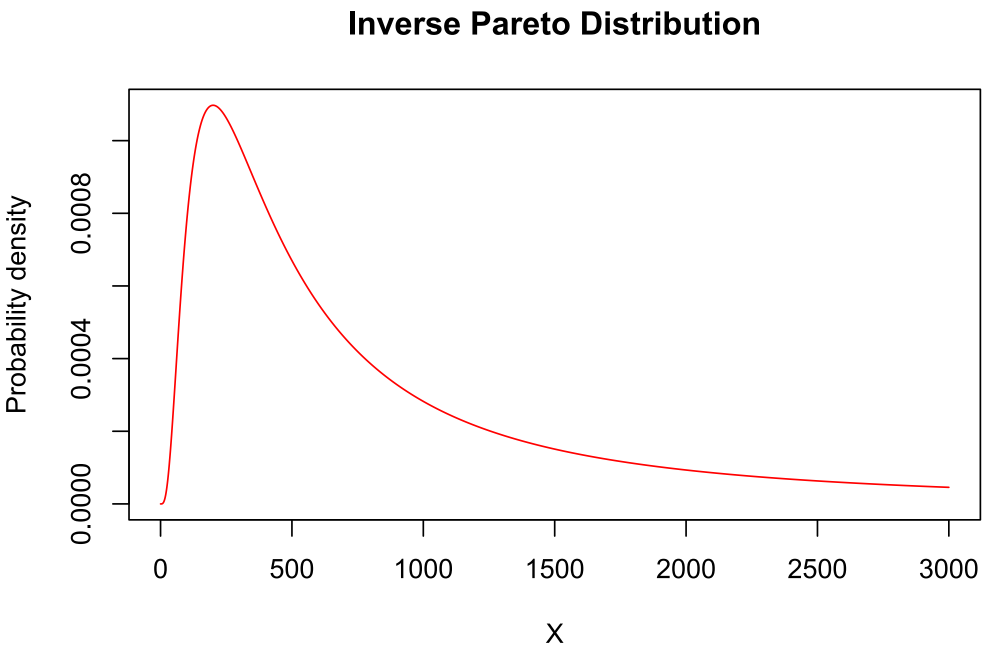
Functions

Name	Function
Parameter assumptions	$\theta > 0, \tau > 0$
Probability density function $f(x)$	$\frac{\tau\theta x^{\tau-1}}{(x+\theta)^{\tau}-1}$
Distribution function $F(x)$	$\left(\frac{x}{x+\theta}\right)^{\tau}$
k^{th} raw moment $\mathrm{E}[X^k]$	$\frac{\theta^k\Gamma(\tau+k)\Gamma(1-k)}{\Gamma(\tau)}$ $-\tau < k < 1$
$\mathrm{E}[(X \wedge x)^k]$	$\theta^k\tau\int_0^{x/(x+\theta)} \; y^{\tau+k-1}(1-y)^{-k}dy + x^k[1-\left(\frac{x}{x+\theta}\right)^{\tau}]$ $k > -\tau$

`R` Commands

Function Name	R Command
Density function	dinvpareto$(x =, shape = \tau, scale = \theta)$
Distribution function	pinvpareto$(p =, shape = \tau, scale = \theta)$
Quantile function	qinvpareto$(q =, shape = \tau, scale = \theta)$
Random sampling function	rinvpareto$(n =, shape = \tau, scale = \theta)$

Illustrative Graph

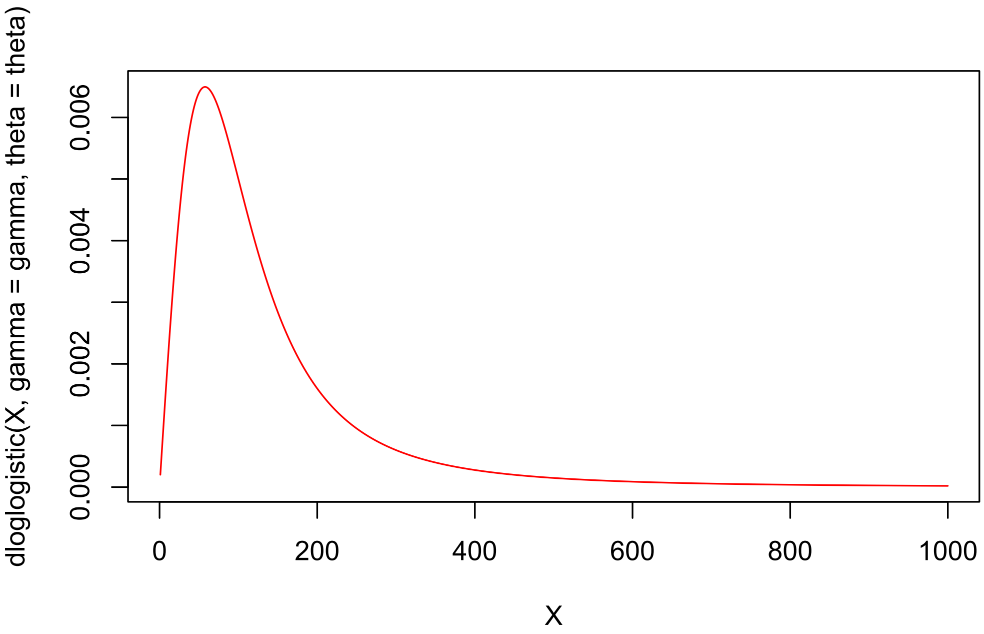


Loglogistic

Functions

Name	Function
Parameter assumptions	$\theta > 0, \gamma > 0, u = \frac{(x/\theta)^\gamma}{1+(x/\theta)^\gamma}$
Probability density function $f(x)$	$\frac{\gamma(x/\theta)^\gamma}{x[1+(x/\theta)^\gamma]^2}$
Distribution function $F(x)$	u
k^{th} raw moment $\mathrm{E}[X^k]$	$\theta^k\Gamma(1+(k/\gamma))\Gamma(1-(k/\gamma))$ $-\gamma < k < \gamma$
$\mathrm{E}[(X \wedge x)^k]$	$\theta^k\Gamma(1+(k/\gamma))\Gamma(1-(k/\gamma))\beta(1+(k/\gamma), 1-(k/\gamma); u) + x^k(1-u)$ $k > -\gamma$

Illustrative Graph



Paralogistic

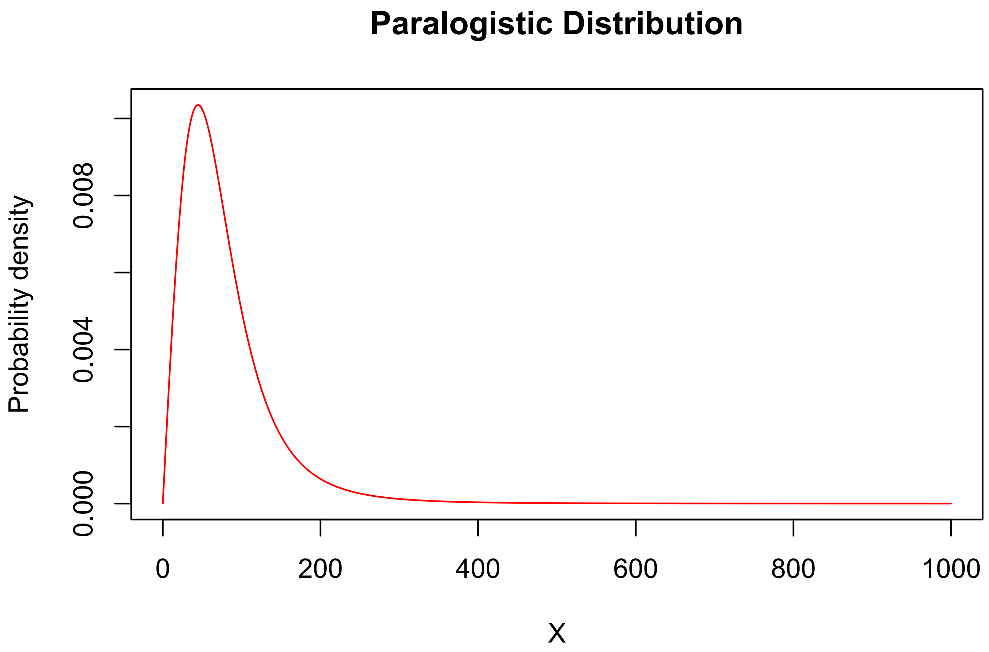
Functions

Name	Function
Parameter assumptions	$\theta > 0, \alpha > 0, u = \frac{1}{1+(x/\theta)^{\alpha}}$
Probability density function $f(x)$	$\frac{\alpha^2 (x/\theta)^{\alpha}}{x[1+(x/\theta)^{\alpha}]^{\alpha+1}}$
Distribution function $F(x)$	$1 - u^{\alpha}$
k^{th} raw moment $\mathrm{E}[X^k]$	$\frac{\theta^k \Gamma(1+(k/\alpha))\Gamma(\alpha-(k/\alpha))}{\Gamma(\alpha)}$ $-\alpha < k < \alpha^2$
$\mathrm{E}[(X \wedge x)^k]$	$\frac{\theta^k \Gamma(1+(k/\alpha))\Gamma(\alpha-(k/\alpha))}{\Gamma(\alpha)} \beta(1+(k/\alpha), \alpha-(k/\alpha); 1-u) + x^k u^{\alpha}$ $k > -\alpha$

R Commands

Function Name	R Command
Density function	dparalogis($x =, shape = \alpha, scale = \theta$)
Distribution function	pparalogis($p =, shape = \alpha, scale = \theta$)
Quantile function	qparalogis($q =, shape = \alpha, scale = \theta$)
Random sampling function	rparalogis($n =, shape = \alpha, scale = \theta$)

Illustrative Graph



Gamma

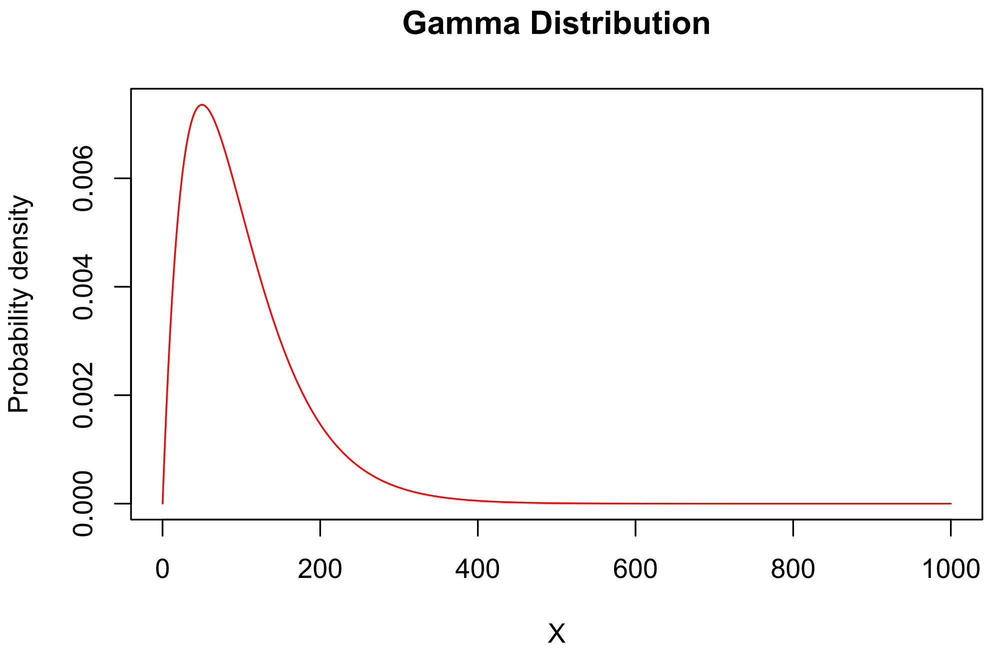
Functions

Name	Function
Parameter assumptions	$\theta > 0,\ \alpha > 0$
Probability density function $f(x)$	$\frac{1}{\theta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\theta}$
Distribution function $F(x)$	$\Gamma(\alpha; \frac{x}{\theta})$
k^{th} raw moment $\mathrm{E}[X^k]$	$\theta^k \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}$ $k > -\alpha$
$\mathrm{E}[X \wedge x]^k$	$\frac{\theta^k \Gamma(k+\alpha)}{\Gamma(\alpha)} \Gamma(k+\alpha; x/\theta) + x^k[1 - \Gamma(\alpha; x/\theta)]$ $k > -\alpha$

R Commands

Density function	dgamma$(x =, shape = \alpha, scale = \theta)$
Distribution function	pgamma$(p =, shape = \alpha, scale = \theta)$
Quantile function	qgamma$(q =, shape = \alpha, scale = \theta)$
Random sampling function	rgamma$(n =, shape = \alpha, scale = \theta)$

Illustrative Graph



Inverse Gamma

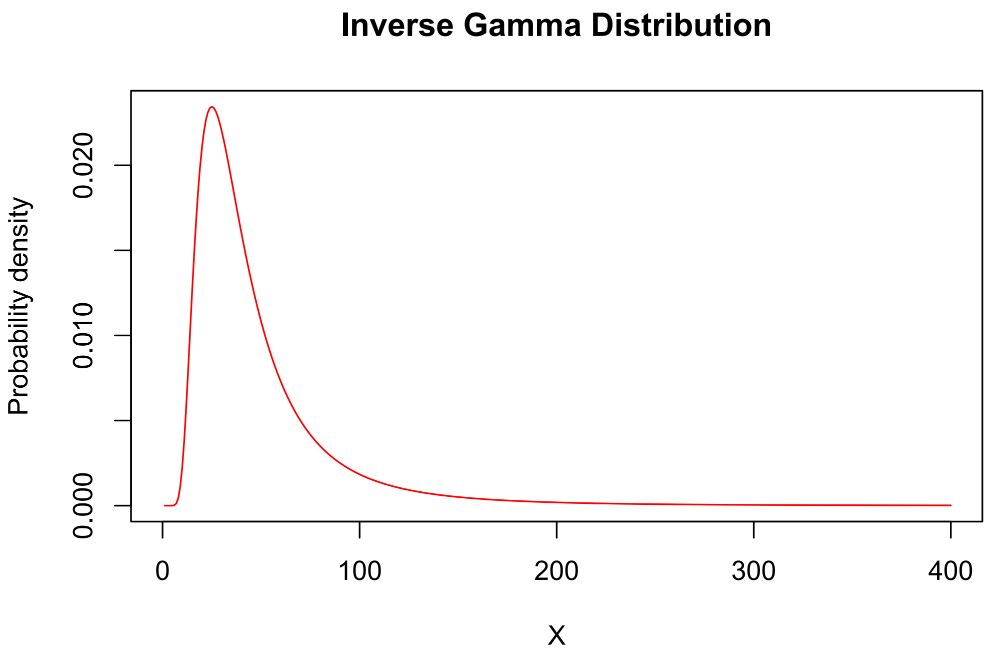
Functions

Name	Function
Probability density function $f(x)$	$\frac{(\theta/x)^{\alpha} e^{-\theta/x}}{x\Gamma(\alpha)}$
Distribution function $F(x)$	$1 - \Gamma(\alpha; \theta/x)$
k^{th} raw moment $\mathrm{E}[X^k]$	$\frac{\theta^k \Gamma(\alpha-k)}{\Gamma(\alpha)}$ $k < \alpha$
$\mathrm{E}[(X \wedge x)^k]$	$\frac{\theta^k \Gamma(\alpha-k)}{\Gamma(\alpha)}[1 - \Gamma(\alpha - k; \theta/x)] + x^k \Gamma(\alpha; \theta/x)$

R Commands

Function Name	R Command
Density function	dinvgamma($x =, shape = \alpha, scale = \theta$)
Distribution function	pinvgamma($p =, shape = \alpha, scale = \theta$)
Quantile function	qinvgamma($q =, shape = \alpha, scale = \theta$)
Random sampling function	rinvgamma($n =, shape = \alpha, scale = \theta$)

Illustrative Graph



Weibull

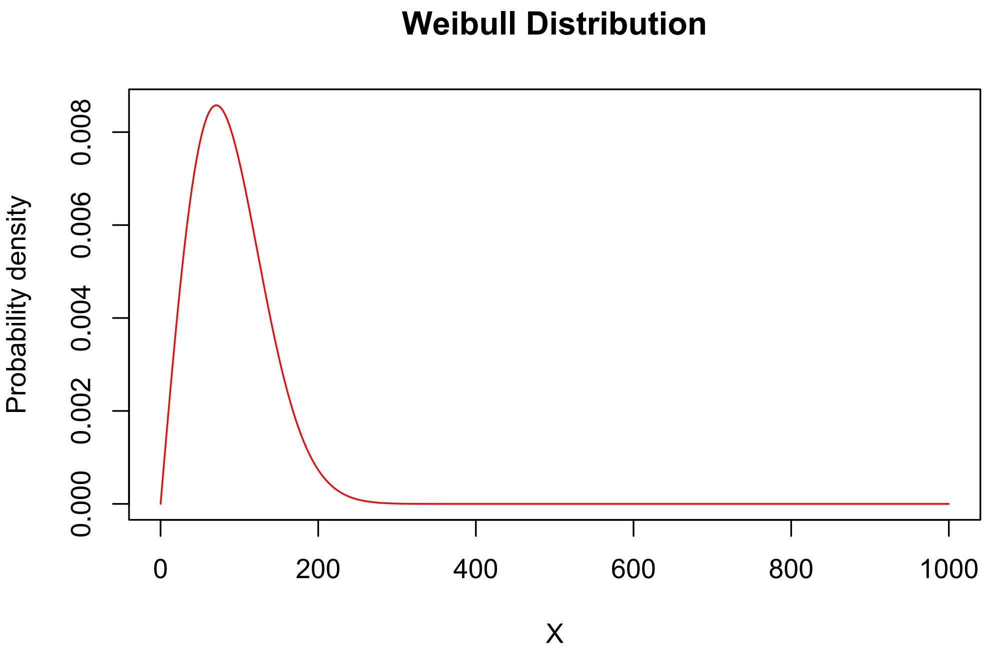
Functions

Name	Function
Parameter assumptions	$\theta > 0, \alpha > 0$
Probability density function $f(x)$	$\dfrac{\alpha \left(\frac{x}{\theta}\right)^{\alpha} \exp\left(-\left(\frac{x}{\theta}\right)^{\alpha}\right)}{x}$
Distribution function $F(x)$	$1 - \exp\left(-\left(\frac{x}{\theta}\right)^{\alpha}\right)$
k^{th} raw moment $\mathrm{E}[X^k]$	$\theta^k \Gamma(1 + \frac{k}{\alpha})$ $k > -\alpha$
$\mathrm{E}[(X \wedge x)^k]$	$\theta^k \Gamma(1 + \frac{k}{\alpha}) \Gamma\left[1 + \frac{k}{\alpha}; \left(\frac{x}{\theta}\right)^{\alpha}\right] + x^k \exp\left(-\left(\frac{x}{\theta}\right)^{\alpha}\right)$ $k > -\alpha$

R Commands

Function Name	R Command
Density function	dweibull$(x =, shape = \alpha, scale = \theta)$
Distribution function	pweibull$(p =, shape = \alpha, scale = \theta)$
Quantile function	qweibull$(q =, shape = \alpha, scale = \theta)$
Random sampling function	rweibull$(n =, shape = \alpha, scale = \theta)$

Illustrative Graph



Inverse Weibull

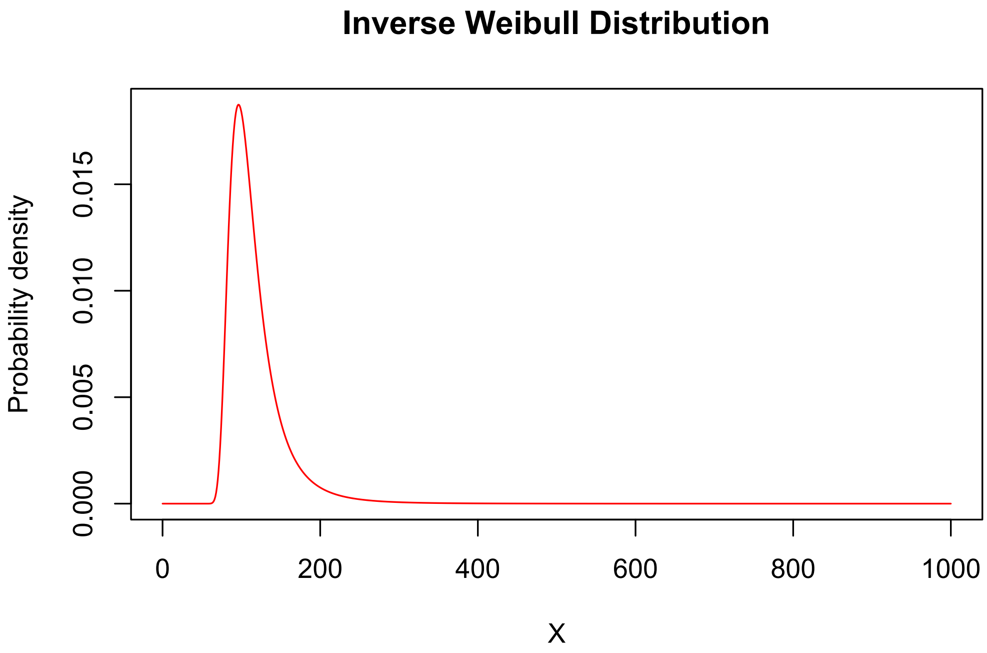
Functions

Name	Function
Parameter assumptions	$\theta > 0, \tau > 0$
Probability density function $f(x)$	$\frac{\tau(\theta/x)^{\tau} \exp\left(-\left(\frac{\theta}{x}\right)^{\tau}\right)}{x}$
Distribution function $F(x)$	$\exp\left(-\left(\frac{\theta}{x}\right)^{\tau}\right)$
k^{th} raw moment $\mathrm{E}[X^k]$	$\theta^k\Gamma(1-(k/\tau))$ $k < \tau$
$\mathrm{E}[(X \wedge x)^k]$	$\theta^k\Gamma(1-(k/\tau))[1-\Gamma(1-(k/\tau);(\theta/x)^{\tau})]+x^k[1-e^{-(\theta/x)^{\tau}}]$

R Commands

Function Name	R Command
Density function	dinvweibull($x =, shape = \tau, scale = \theta$)
Distribution function	pinvweibull($p =, shape = \tau, scale = \theta$)
Quantile function	qinvweibull($q =, shape = \tau, scale = \theta$)
Random sampling function	rinvweibull($n =, shape = \tau, scale = \theta$)

Illustrative Graph



Uniform

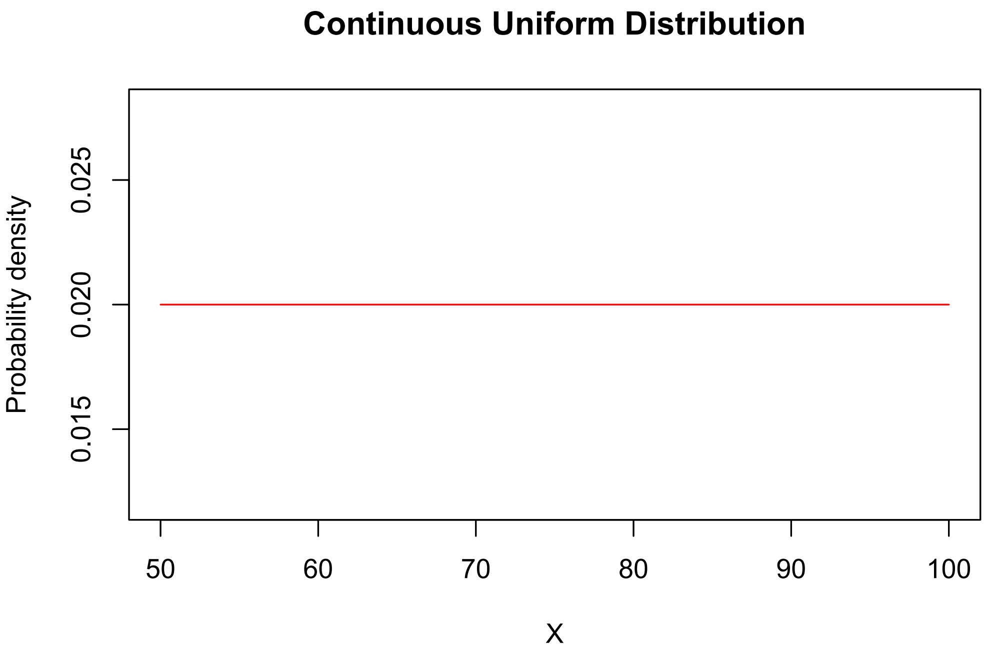
Functions

Name	Function
Parameter assumptions	$-\infty < \alpha < \beta < \infty$
Probability density f(x)	$\frac{1}{\beta-\alpha}$
Distribution function $F(x)$	$\frac{x-\alpha}{\beta-\alpha}$
Mean E[X]	$\frac{\beta+\alpha}{2}$
Variance $E[(X-\mu)^2]$	$\frac{(\beta-\alpha)^2}{12}$
E$[(X-\mu)^k]$	$\mu_k = 0 \quad \text{for odd } k$ $\mu_k = \frac{(\beta-\alpha)^k}{2^k(k+1)} \quad \text{for even } k$

R Commands

Function Name	R Command
Density function	dunif($x =, min = a, max = b$)
Distribution function	punif($p =, min = a, max = b$)
Quantile function	qunif($q =, min = a, max = b$)
Random sampling function	runif($n =, min = a, max = b$)

Illustrative Graph



Normal

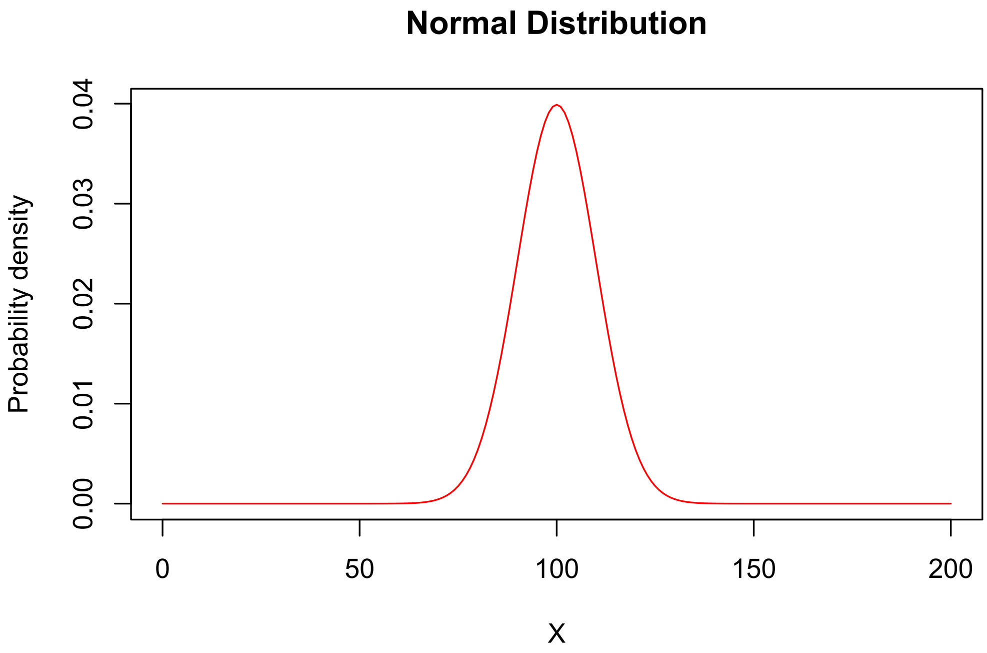
Functions

Name	Function
Parameter assumptions	$-\infty < \mu < \infty,\ \sigma > 0$
Probability density f(x)	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
Distribution function $F(x)$	$\Phi\left(\frac{x-\mu}{\sigma}\right)$
Mean E[X]	μ
Variance $E[(X-\mu)^2]$	σ^2
E$[(x-\mu)^k]$	$\mu_k = 0$ for even k $\mu_k = \frac{k!\sigma^2}{(\frac{k}{2})!2^{k/2}}$ for odd k

R Commands

Function Name	R Command
Density function	dnorm$(x =, mean = \mu, sd = \sigma)$
Distribution function	pnorm$(p =, mean = \mu, sd = \sigma)$
Quantile function	qnorm$(q =, mean = \mu, sd = \sigma)$
Random sampling function	rnorm$(n =, mean = \mu, sd = \sigma)$

Illustrative Graph



Cauchy

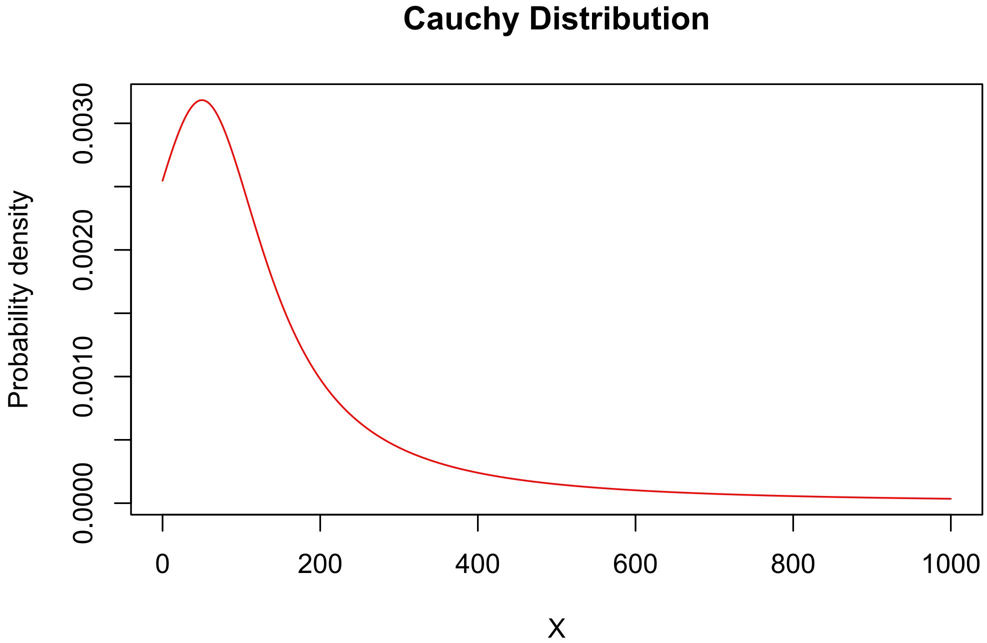
Functions

Name	Function
Parameter assumptions	$-\infty < \alpha < \infty, \beta > 0$
Probability density function $f(x)$	$\frac{1}{\pi\beta}[1 + \left(\frac{x-\alpha}{\beta}\right)^2]^{-1}$

R Commands

Function Name	R Command
Density function	dcauchy$(x =, location = \alpha, scale = \beta)$
Distribution function	pcauchy$(p =, location = \alpha, scale = \beta)$
Quantile function	qcauchy$(q =, location = \alpha, scale = \beta)$
Random sampling function	rcauchy$(n =, location = \alpha, scale = \beta)$

Illustrative Graph



20.2.3 Three Parameter Distributions

Generalized Pareto

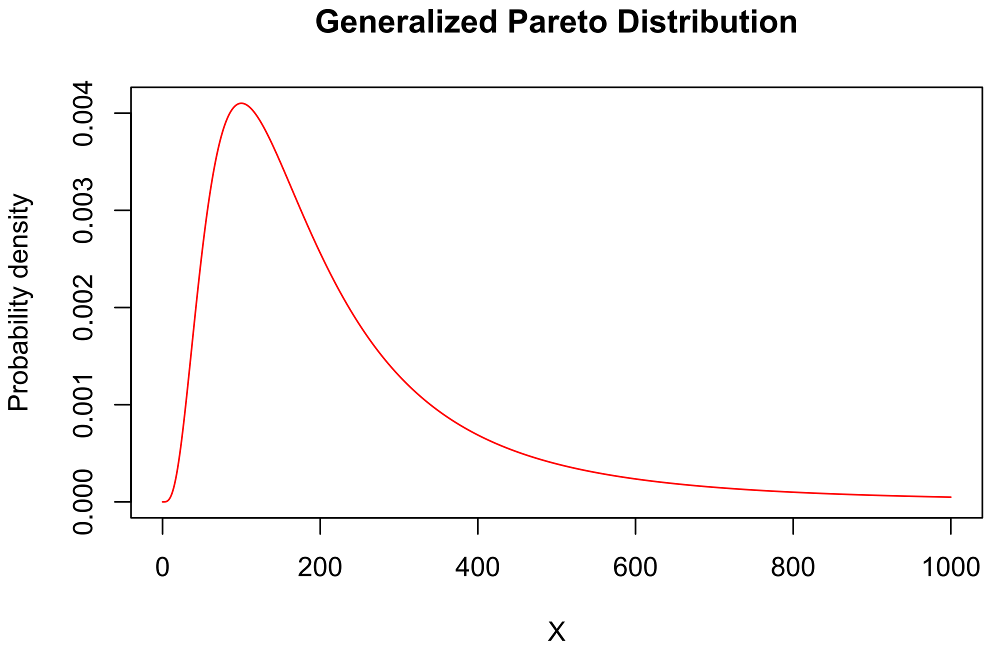
Functions

Name	Function
Parameter assumptions	$\theta > 0, \alpha > 0, \tau > 0, u = \frac{x}{x+\theta}$
Probability density function $f(x)$	$\frac{\Gamma(\alpha+\tau)}{\Gamma(\alpha)\Gamma(\tau)} \frac{\theta^\alpha x^{\tau-1}}{(x+\theta)^{\alpha+\tau}}$
Distribution function $F(x)$	$\beta(\tau, \alpha; u)$
k^{th} raw moment $\mathrm{E}[X^k]$	$\frac{\theta^k \Gamma(\tau+1)\Gamma(\alpha-k)}{\Gamma(\alpha)\Gamma(\tau)}$ $-\tau < k < \alpha$
$\mathrm{E}[(X \wedge x)^k]$	$\frac{\theta^k \Gamma(\tau+k)\Gamma(\alpha-k)}{\Gamma(\alpha)\Gamma(\tau)} \beta(\tau+k, \alpha-k; u) + x^k[1-\beta(\tau, \alpha; u)]$ $k > -\tau$

R Commands

Function Name	R Command
Density function	dgenpareto($x =, shape1 = \alpha, shape2 = \tau, scale = \theta$)
Distribution function	pgenpareto($q =, shape1 = \alpha, shape2 = \tau, scale = \theta$)
Quantile function	qgenpareto($p =, shape1 = \alpha, shape2 = \tau, scale = \theta$)
Random sampling function	rgenpareto($r =, shape1 = \alpha, shape2 = \tau, scale = \theta$)

Illustrative Graph



Burr

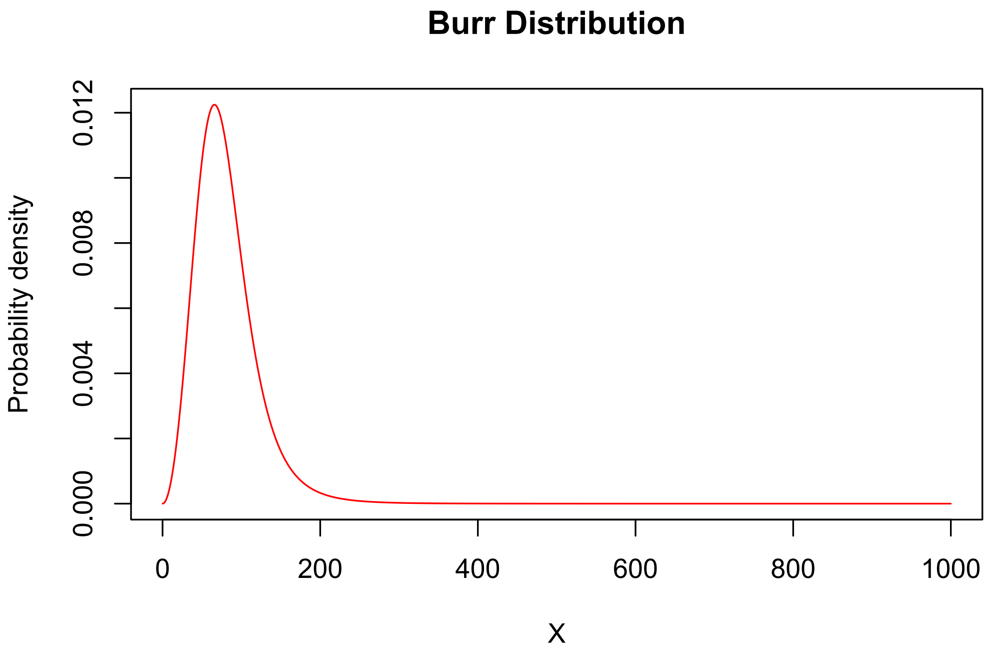
Functions

Name	Function
Parameter assumptions	$\theta > 0, \alpha > 0, \gamma > 0, u = \frac{1}{1+(x/\theta)^{\gamma}}$
Probability density function $f(x)$	$\frac{\alpha\gamma(x/\theta)^{\gamma}}{x[1+(x/\theta)^{\gamma}]^{\alpha+1}}$
Distribution function $F(x)$	$1 - u^{\alpha}$
k^{th} raw moment $\mathrm{E}[X^k]$	$\frac{\theta^k\Gamma(1+(k/\gamma))\Gamma(\alpha-(k/\gamma))}{\Gamma(\alpha)}$ $-\gamma < k < \alpha\gamma$
$\mathrm{E}[(X \wedge x)^k]$	$\frac{\theta^k\Gamma(1+(k/\gamma))\Gamma(\alpha-(k/\gamma))}{\Gamma(\alpha)}\beta(1+(k/\gamma), \alpha-(k/\gamma); 1-u) + x^k u^{\alpha}$ $k > -\gamma$

R Commands

Function Name	R Command
Density function	dburr$(x =, shape1 = \alpha, shape2 = \gamma, scale = \theta)$
Distribution function	pburr$(p =, shape1 = \alpha, shape2 = \gamma, scale = \theta)$
Quantile function	qburr$(q =, shape1 = \alpha, shape2 = \gamma, scale = \theta)$
Random sampling function	rburr$(n =, shape1 = \alpha, shape2 = \gamma, scale = \theta)$

Illustrative Graph



Inverse Burr

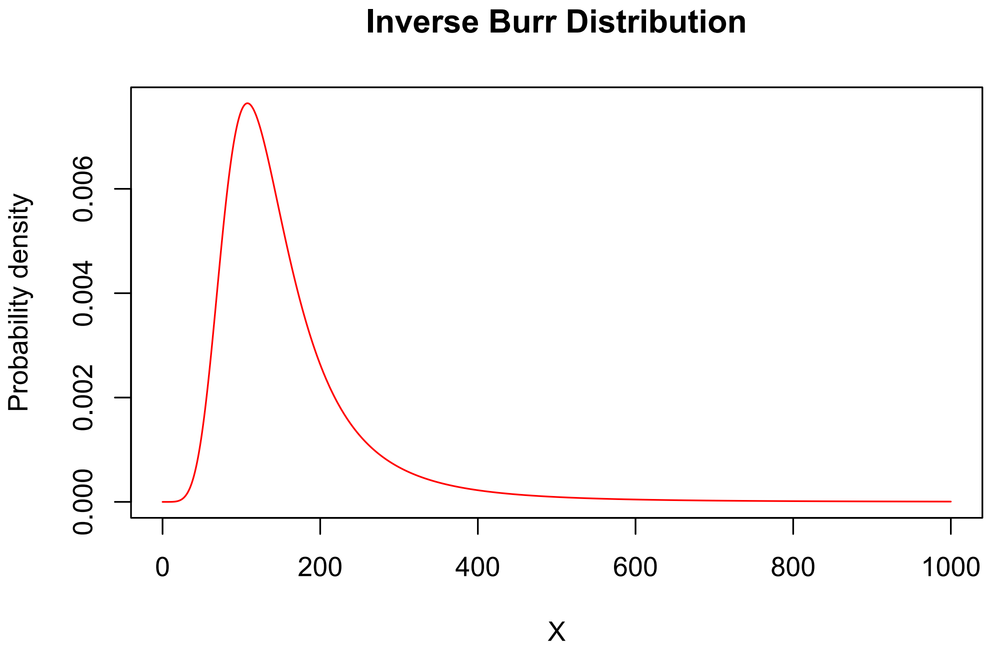
Functions

Name	Function
Parameter assumptions	$\theta > 0, \tau > 0, \gamma > 0, u = \frac{(x/\theta)^\gamma}{1+(x/\theta)^\gamma}$
Probability density function $f(x)$	$\frac{\tau\gamma(x/\theta)^{\tau\gamma}}{x[1+(x/\theta)^\gamma]^{\tau+1}}$
Distribution function $F(x)$	u^τ
k^{th} raw moment $\mathrm{E}[X^k]$	$\frac{\theta^k\Gamma(\tau+(k/\gamma))\Gamma(1-(k/\gamma))}{\Gamma(\tau)}$ $-\tau\gamma < k < \gamma$
$\mathrm{E}[(X \wedge x)^k]$	$\frac{\theta^k\Gamma(\tau+(k/\gamma))\Gamma(1-(k/\gamma))}{\Gamma(\tau)}\beta(\tau + (k/\gamma), 1 - (k/\gamma); u) + x^k[1 - u^\tau]$ $k > -\tau\gamma$

R Commands

Function Name	R Command
Density function	dinvburr$(x =, shape1 = \tau, shape2 = \gamma, scale = \theta)$
Distribution function	pinvburr$(p =, shape1 = \tau, shape2 = \gamma, scale = \theta)$
Quantile function	qinvburr$(q =, shape1 = \tau, shape2 = \gamma, scale = \theta)$
Random sampling function	rinvburr$(n =, shape1 = \tau, shape2 = \gamma, scale = \theta)$

Illustrative Graph



20.2.4 Four Parameter Distribution

Generalized Beta of the Second Kind (GB2)

Functions

Name	Function
Parameter assumptions	$\theta > 0, \alpha_1 > 0, \alpha_2 > 0, \sigma > 0$
Probability density function $f(x)$	$\frac{(x/\theta)^{\alpha_2/\sigma}}{x\sigma\ \mathrm{B}(\alpha_1,\alpha_2)\left[1+(x/\theta)^{1/\sigma}\right]^{\alpha_1+\alpha_2}}$
k^{th} raw moment $\mathrm{E}[X^k]$	$\frac{\theta^k\ \mathrm{B}(\alpha_1+k\sigma,\alpha_2-k\sigma)}{\mathrm{B}(\alpha_1,\alpha_2)}$ $k > 0$

R Commands

Please see the R Codes for Loss Data Analytics site for information about this distribution.

20.2.5 Other Distributions

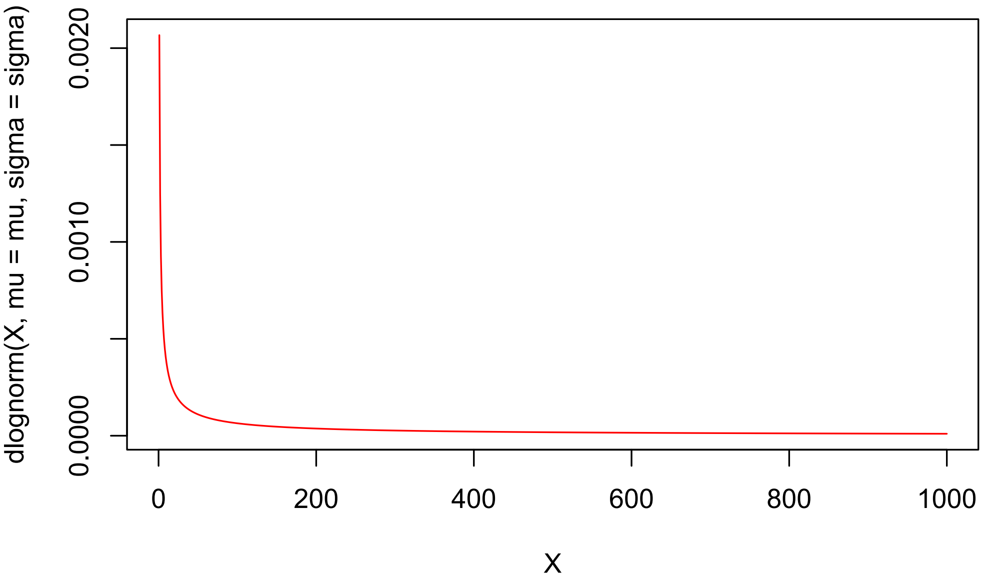
–>

Lognormal

Functions

Name	Function
Parameter assumptions	$-\infty < \mu < \infty, \sigma > 0$
Probability density function $f(x)$	$\frac{1}{x\sqrt{2\pi}\sigma} \exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right)$
Distribution function $F(x)$	$\Phi\left(\frac{\ln(x)-\mu}{\sigma}\right)$
k^{th} raw moment $\mathrm{E}[X^k]$	$\exp(k\mu + \frac{k^2\sigma^2}{2})$
Limited Expected Value $\mathrm{E}[X \wedge x]$	$\exp\left(k\mu + \frac{k^2\sigma^2}{2}\right)\Phi\left(\frac{\ln(x)-\mu-k\sigma^2}{\sigma}\right) + x^k\left[1 - \Phi\left(\frac{\ln(x)-\mu}{\sigma}\right)\right]$

Illustrative Graph



Inverse Gaussian

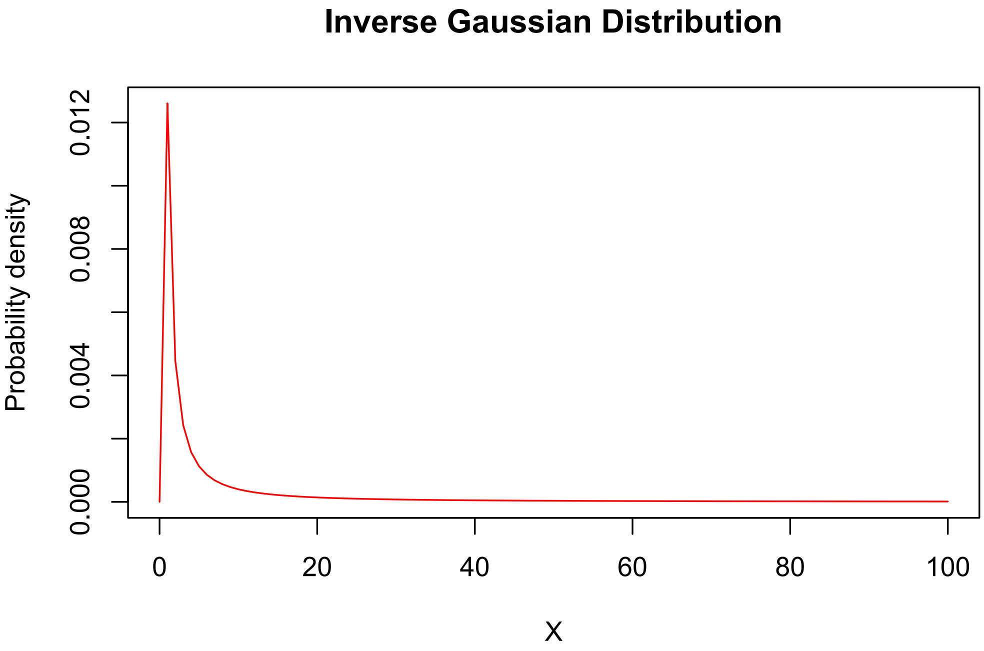
Functions

Name	Function
Parameter assumptions	$\theta > 0, \mu > 0, z = \frac{x-\mu}{\mu}\ ,\ y = \frac{x+\mu}{\mu}$
Probability density function $f(x)$	$\left(\frac{\theta}{2\pi x^3}\right)^{1/2} \exp\left(\frac{-\theta z^2}{2x}\right)$
Distribution function $F(x)$	$\Phi\left[z\left(\frac{\theta}{x}\right)^{1/2}\right] + \exp\left(\frac{2\theta}{\mu}\right)\Phi\left[-y\left(\frac{\theta}{x}\right)^{1/2}\right]$
Mean E[X]	μ
Var[X]	$\frac{\mu^3}{\theta}$
$\mathrm{E}[(X \wedge x)^k]$	$x - \mu x\Phi\left[z\left(\frac{\theta}{x}\right)^{1/2}\right] - (\mu y)\exp\left(\frac{2\theta}{\mu}\right)\Phi\left[-y\left(\frac{\theta}{x}\right)^{1/2}\right]$

`R` Commands

Function Name	R Command
Density function	dinvgauss($x =, mean = \mu, dispersion = \theta$)
Distribution function	pinvgauss($p =, mean = \mu, dispersion = \theta$)
Quantile function	qinvgauss($q =, mean = \mu, dispersion = \theta$)
Random sampling function	rinvgauss($n =, mean = \mu, dispersion = \theta$)

Illustrative Graph



20.2.6 Distributions with Finite Support

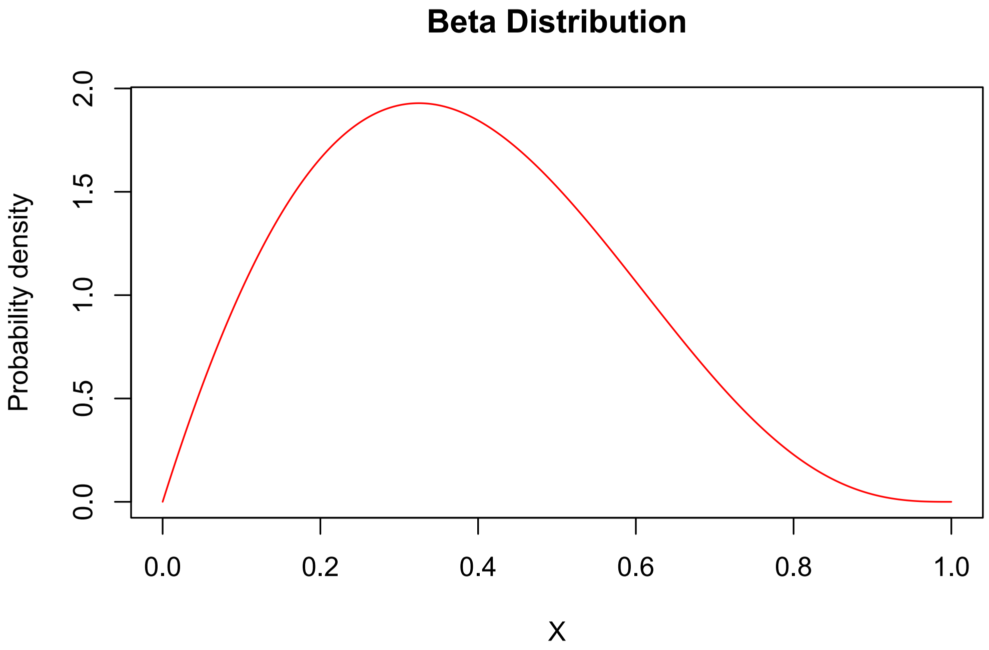
Beta

Functions

Name	Function
Parameter assumptions	$\theta > 0,\ a > 0,\ b > 0, u = \frac{x}{\theta},\ 0 < x < \theta$
Probability density function $f(x)$	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^a (1-u)^{b-1} \frac{1}{x}$
Distribution function $F(x)$	$\beta(a, b; u)$
k^{th} raw moment $\mathrm{E}[X^k]$	$\frac{\theta^k \Gamma(a+b)\Gamma(a+k)}{\Gamma(a)\Gamma(a+b+k)}$ $k > -a$
$\mathrm{E}[X \wedge x]^k$	$\frac{\theta^k a(a+1)\cdots(a+k-1)}{(a+b)(a+b+1)\cdots(a+b+k-1)} \beta(a+k, b; u) + x^k[1 - \beta(a, b; u)]$

R Commands

Function Name	R Command
Density function	dbeta($x =, shape1 = a, shape2 = b, ncp = \theta$)
Distribution function	pbeta($p =, shape1 = a, shape2 = b, ncp = \theta$)
Quantile function	qbeta($q =, shape1 = a, shape2 = b, ncp = \theta$)
Random sampling function	rbeta($n =, shape1 = a, shape2 = b, ncp = \theta$)



Generalized Beta

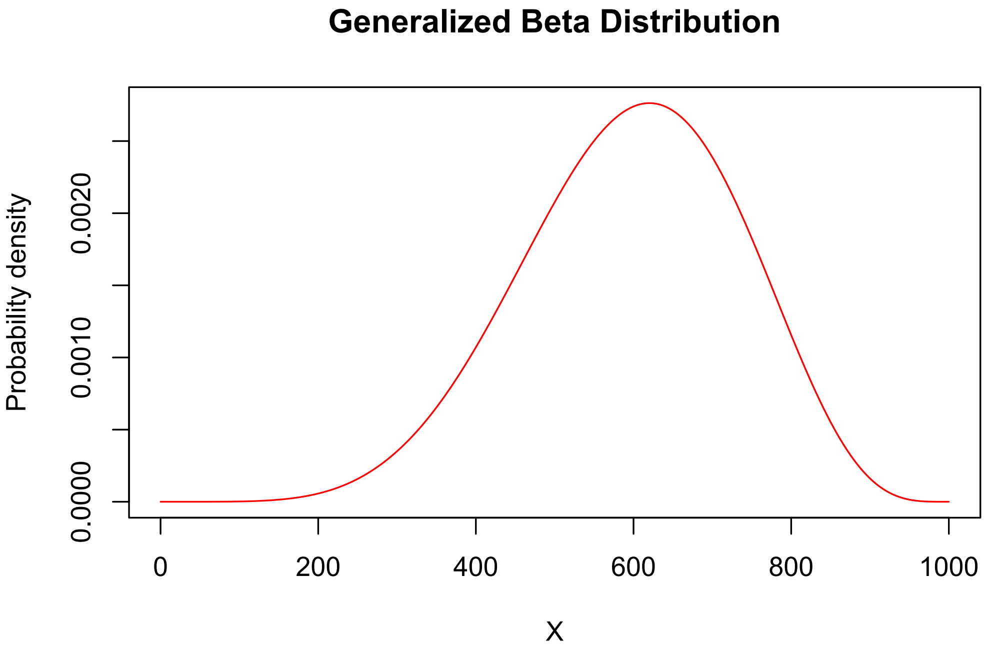
Functions

Name	Function
Parameter assumptions	$\theta > 0, a > 0, b > 0, \tau > 0, 0 < x < \theta \ , \ u = (x/\theta)^{\tau}$
Probability density function $f(x)$	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{\alpha}(1-u)^{b-1}\frac{\tau}{x}$
Distribution function $F(x)$	$\beta(a, b; u)$
k^{th} raw moment $\mathrm{E}[X^k]$	$\frac{\theta^k\Gamma(a+b)\Gamma(a+(k/\tau))}{\Gamma(a)\Gamma(a+b+(k/\tau))}$ $k > -\alpha\tau$
$\mathrm{E}[(X \wedge x)^k]$	$\frac{\theta^k\Gamma(a+b)\Gamma(a+(k/\tau))}{\Gamma(a)\Gamma(a+b+(k/\tau))}\beta(a+(k/\tau), b; u) + x^k[1-\beta(a, b; u)]$

R Commands

Function Name	R Command
Density function	dgenbeta$(x =, shape1 = a, shape2 = b, shape3 = \tau, scale = \theta)$
Distribution function	pgenbeta$(p =, shape1 = a, shape2 = b, shape3 = \tau, scale = \theta)$
Quantile function	qgenbeta$(q =, shape1 = a, shape2 = b, shape3 = \tau, scale = \theta)$
Random sampling function	rgenbeta$(n =, shape1 = a, shape2 = b, shape3 = \tau, scale = \theta)$

Illustrative Graph



20.3 Limited Expected Values

Overview. This section summarizes limited expected values for selected continuous distributions.

Functions

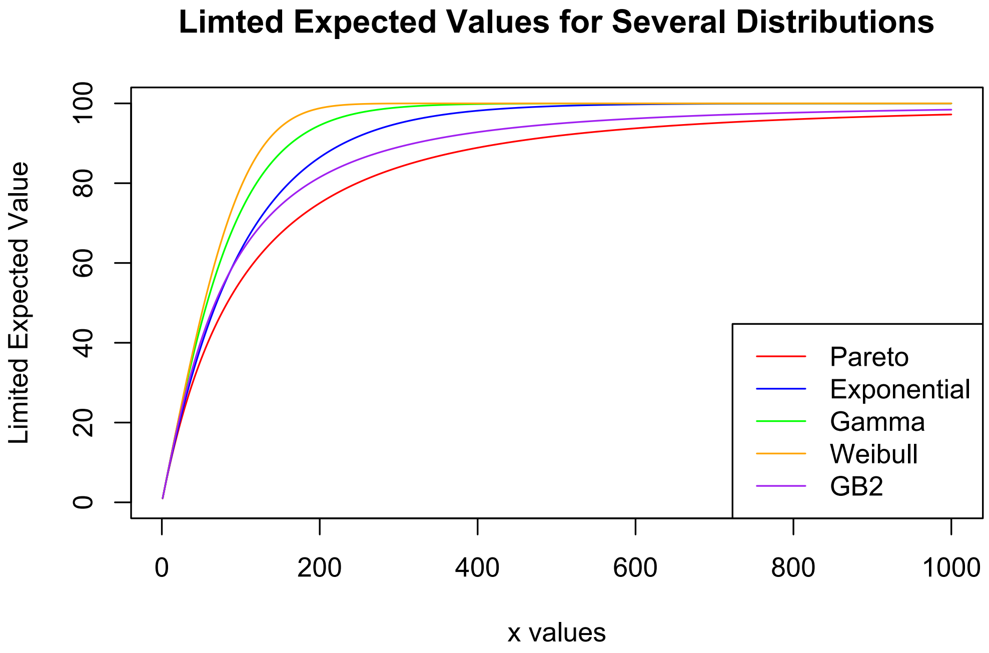
Limited Expected Value Functions

Distribuion	Function
GB2	$\frac{\theta\Gamma(\tau+1)\Gamma(\alpha-1)}{\Gamma(\alpha)\Gamma(\tau)}\beta(\tau+1,\alpha-1;\frac{x}{x+\beta})+x[1-\beta(\tau,\alpha;\frac{x}{x+\beta})]$
Burr	$\frac{\theta\Gamma(1+\frac{1}{\gamma})\Gamma(\alpha-\frac{1}{\gamma})}{\Gamma(\alpha)}\beta(1+\frac{1}{\gamma},\alpha-\frac{1}{\gamma};1-\frac{1}{1+(x/\theta)^\gamma})+x\left(\frac{1}{1+(x/\theta)^\gamma}\right)^\alpha$
Inverse Burr	$\frac{\theta\Gamma(\tau+(1/\gamma))\Gamma(1-(1/\gamma))}{\Gamma(\tau)}\beta(\tau+\frac{1}{\gamma},1-\frac{1}{\gamma};\frac{(x/\theta)^\gamma}{1+(x/\theta)^\gamma})+x[1-\left(\frac{(x/\theta)^\gamma}{1+(x/\theta)^\gamma}\right)^\tau]$
Pareto $\alpha=1$ $\alpha\neq 1$	 $-\theta\ln\left(\frac{\theta}{x+\theta}\right)$ $\frac{\theta}{\alpha-1}[1-\left(\frac{\theta}{x+\theta}\right)^{\alpha-1}]$
Inverse Pareto	$\theta\tau\int_0^{x/(x+\theta)} y^\tau(1-y)^{-1}dy+x[1-\left(\frac{x}{x+\theta}\right)^\tau]$
Loglogistic	$\theta\Gamma(1+\frac{1}{\gamma})\Gamma(1-\frac{1}{\gamma})\beta(1+\frac{1}{\gamma},1-\frac{1}{\gamma};\frac{(x/\theta)^\gamma}{1+(x/\theta)^\gamma})+x(1-\frac{(x/\theta)^\gamma}{1+(x/\theta)^\gamma})$
Paralogistic	$\frac{\theta\Gamma(1+\frac{1}{\alpha})\Gamma(\alpha-\frac{1}{\alpha})}{\Gamma(\alpha)}\beta(1+\frac{1}{\alpha},\alpha-\frac{1}{\alpha};1-\frac{1}{1+(x/\theta)^\alpha})+x\left(\frac{1}{1+(x/\theta)^\alpha}\right)^\alpha$
Inverse Paralogistic	$\frac{\theta\Gamma(\tau+\frac{1}{\tau})\Gamma(1-\frac{1}{\tau})}{\Gamma(\tau)}\beta(\tau+\frac{1}{\tau},1-\frac{1}{\tau};\frac{(x/\theta)^\tau}{1+(x/\theta)^\tau})+x[1-\left(\frac{(x/\theta)^\tau}{1+(x/\theta)^\tau}\right)^\tau]$
Gamma	$\frac{\theta\Gamma(\alpha+1)}{\Gamma(\alpha)}\Gamma(\alpha+1;\frac{x}{\theta})+x[1-\Gamma(\alpha;\frac{x}{\theta})]$
Inverse Gamma	$\frac{\theta\Gamma(\alpha-1)}{\Gamma(\alpha)}[1-\Gamma(\alpha-1;\frac{\theta}{x})]+x\Gamma(\alpha;\frac{\theta}{x})$
Weibull	$\theta\Gamma(1+\frac{1}{\alpha})\Gamma(1+\frac{1}{\alpha};\left(\frac{x}{\theta}\right)^\alpha)+x*\exp(-(x/\theta)^\alpha)$
Inverse Weibull	$\theta\Gamma(1-\frac{1}{\alpha})[1-\Gamma(1-\frac{1}{\alpha};\left(\frac{\theta}{x}\right)^\alpha)]+x[1-\exp(-(\theta/x)^\alpha)]$
Exponential	$\theta(1-\exp(-(x/\theta)))$
Inverse Exponential	$\theta G(0;\frac{\theta}{x})+x(1-\exp(-(\theta/x)))$
Lognormal	$\exp(\mu+\sigma^2/2)\Phi\left(\frac{\ln(x)-\mu-\sigma^2}{\sigma}\right)+x[1-\Phi\left(\frac{\ln(x)-\mu}{\sigma}\right)]$
Inverse Gaussian	$x-\mu\left(\frac{x-\mu}{\mu}\right)\Phi\left[\left(\frac{x-\mu}{\mu}\right)\left(\frac{\theta}{x}\right)^{1/2}\right]-\mu\left(\frac{x+\mu}{\mu}\right)\exp\left(\frac{2\theta}{\mu}\right)\Phi\left[-\left(\frac{x+\mu}{\mu}\right)\left(\frac{\theta}{x}\right)^{1/2}\right]$
Single-Parameter Pareto	$\frac{\alpha\theta}{\alpha-1}-\frac{\theta^\alpha}{(\alpha-1)x^{\alpha-1}}$
Generalized Beta	$\frac{\theta\Gamma(a+b)\Gamma(a+\frac{1}{\tau})}{\Gamma(a)\Gamma(a+b+\frac{1}{\tau})}\beta(a+\frac{1}{\tau},b;\left(\frac{x}{\theta}\right)^\tau)+x\left[1-\beta(a,b;\left(\frac{x}{\theta}\right)^\tau)\right]$
Beta	$\frac{\theta a}{(a+b)}\beta(a+1,b;\frac{x}{\theta})+x[1-\beta(a,b;\frac{x}{\theta})]$

Illustrative Graph

Comparison of Limited Expected Values for Selected Distributions

Distribution	Parameters	E$[X]$	$E[X \wedge 100]$	$E[X \wedge 250]$	$E[X \wedge 500]$	$E[X \wedge 1000]$
Pareto	$\alpha = 3, \theta = 200$	100	55.55	80.25	91.84	97.22
Exponential	$\theta = 100$	100	63.21	91.79	99.33	99.99
Gamma	$\alpha = 2, \theta = 50$	100	72.93	97.64	99.97	100
Weibull	$\tau = 2, \theta = \frac{200}{\sqrt{\pi}}$	100	78.99	99.82	100	100
GB2	$\alpha = 3, \tau = 2, \theta = 100$	100	62.50	86.00	94.91	98.42



21

Appendix E: Conventions for Notation

Chapter Preview. Loss Data Analytics serves as a bridge between actuarial problems and methods and widely accepted statistical concepts and tools. Thus, the notation should be consistent with standard usage employed in probability and mathematical statistics. See, for example, (Halperin et al., 1965) for a description of one standard.

21.1 General Conventions

- Random variables are denoted by upper-case italicized Roman letters, with X or Y denoting a claim size variable, N a claim count variable, and S an aggregate loss variable. Realizations of random variables are denoted by corresponding lower-case italicized Roman letters, with x or y for claim sizes, n for a claim count, and s for an aggregate loss.
- Probability events are denoted by upper-case Roman letters, such as Pr(A) for the probability that an outcome in the event ''A" occurs.
- Cumulative probability functions are denoted by $F(z)$ and probability density functions by the associated lower-case Roman letter: $f(z)$.
- For distributions, parameters are denoted by lower-case Greek letters. A caret or ''hat" indicates a sample estimate of the corresponding population parameter. For example, $\hat{\beta}$ is an estimate of β .
- The arithmetic mean of a set of numbers, say, $x_1, \ldots, x_n$, is usually denoted by $\bar{x}$; the use of x, of course, is optional.
- Use upper-case boldface Roman letters to denote a matrix other than a vector. Use lower-case boldface Roman letters to denote a (column) vector. Use a superscript prime "'" for transpose. For example, $\mathbf{x}'\mathbf{A}\mathbf{x}$ is a quadratic form.
- Acronyms are to be used sparingly, given the international focus of our audience. Introduce acronyms commonly used in statistical nomenclature but limit the number of acronyms introduced. For example, *pdf* for probability density function is useful but *GS* for Gini statistic is not.

21.2 Abbreviations

Here is a list of abbreviations that we adopt. We italicize these acronyms. For example, we can discuss the goodness of fit in terms of the *AIC* criterion.

AIC	Akaike information criterion
BIC	(Schwarz) Bayesian information criterion
cdf	cumulative distribution function
df	degrees of freedom
iid	independent and identically distributed
GLM	generalized linear model
mle	maximum likelihood estimate/estimator
ols	ordinary least squares
pdf	probability density function
pmf	probability mass function

21.3 Common Statistical Symbols and Operators

Here is a list of commonly used statistical symbols and operators, including the latex code that we use to generate them (in the parens).

$I(\cdot)$	binary indicator function (I). For example, $I(A)$ is one A if an outcome in event occurs and is 0 otherwise.
$\Pr(\cdot)$	probability (`\Pr`)
$\mathrm{E}(\cdot)$	expectation operator (`\mathrm{E}`). For example, $\mathrm{E}(X) = \mathrm{E}\ X$ is the expected value of the random variable $X,$ commonly denoted by μ.
$\mathrm{Var}(\cdot)$	variance operator (`\mathrm{Var}`). For example, $\mathrm{Var}(X) = \mathrm{Var}\ X$ is the variance of the random variable $X,$commonly denoted by σ^2.
$\mu_k = \mathrm{E}\ X^k$	kth moment of the random variable X. For k=1, use $\mu = \mu_1$.
$\mathrm{Cov}(\cdot, \cdot)$	covariance operator (`\mathrm{Cov}`). For example, $\mathrm{Cov}(X,Y) = \mathrm{E}\{(X - \mathrm{E}\ X)(Y - \mathrm{E}\ Y)\} = \mathrm{E}(XY) - (\mathrm{E}\ X)(\mathrm{E}\ Y)$ is the covariance between random variables X and Y.
$\mathrm{E}(X \mid \cdot)$	conditional expectation operator. For example, $\mathrm{E}(X \mid Y = y)$ is the conditional expected value of a random variable X given that the random variable Y equals y.
$\Phi(\cdot)$	standard normal cumulative distribution function (`\Phi`)
$\phi(\cdot)$	standard normal probability density function (`\phi`)
$\sim$	means is distributed as (`\sim`). For example, $X \sim F$ means that the random variable X has distribution function F.
$se(\hat{\beta})$	standard error of the parameter estimate $\hat{\beta}$ (`\hat{\beta}`), usually an estimate of the standard deviation of $\hat{\beta},$ which is $\sqrt{Var(\hat{\beta})}$.
H_0	null hypothesis
H_a or H_1	alternative hypothesis

21.4 Common Mathematical Symbols and Functions

Here is a list of commonly used mathematical symbols and functions, including the latex code that we use to generate them (in the parens).

$\equiv$	identity, equivalence (`\equiv`)
$\implies$	implies (`\implies`)
$\iff$	if and only if (`\iff`)
$\to, \longrightarrow$	converges to (`\to`, `\longrightarrow`)
$\mathbb{N}$	natural numbers $1, 2, \dots$ (`\mathbb{N}`)
$\mathbb{R}$	real numbers (`\mathbb{R}`)
$\in$	belongs to (`\in`)
$\notin$	does not belong to (`\notin`)
$\subseteq$	is a subset of (`\subseteq`)
$\subset$	is a proper subset of (`\subset`)
$\cup$	union (`\cup`)
$\cap$	intersection (`\cap`)
$\emptyset$	empty set (`\emptyset`)
A^c	complement of A
$g * f$	convolution $(g * f)(x) = \int_{-\infty}^{\infty} g(y)f(x-y)dy$
$\exp$	exponential (`\exp`)
$\log$	natural logarithm (`\log`)
$\log_a$	logarithm to the base a
$!$	factorial
$\mathrm{sgn}(x)$	sign of x(`sgn`)
$\lfloor x \rfloor$	integer part of x, that is, largest integer $\leq x$ (`\lfloor`, `\rfloor`)
$\lvert x \rvert$	absolute value of scalar x
$\varGamma(x)$	gamma (generalized factorial) function (`\varGamma`), satisfying $\varGamma(x+1) = x\varGamma(x)$
$B(x, y)$	beta function, $\varGamma(x)\varGamma(y)/\varGamma(x+y)$

21.5 Further Readings

To make connections to other literatures, see (Abadir and Magnus, 2002) http://www.janmagnus.nl/misc/notation.zip for a summary of notation from the econometrics perspective. This reference has a terrific feature that many latex symbols are defined in the article. Further, there is a long history of discussion and debate surrounding actuarial notation; see (Boehm et al., 1975) for one contribution.

22

Appendix. Data Resources

This appendix section describes the datasets used in this book and others that you may wish to explore.

For each set of data, we provide download buttons so that you can easily access the data in standard .csv (comma separated value) format. This allows you replicate and experiment with the methods developed in the book as well as sharpen your understanding through exercises.

We provide the source of each dataset. We also recommend, for deeper understanding, that you occasionally refer to these original sources to further develop your appreciation of the data underpinning the analytics developed in this book.

22.1 Wisconsin Property Fund

Description: The Wisconsin Local Government Property Insurance Fund (LGPIF) is an insurance pool administered by the Wisconsin Office of the Insurance Commissioner. The LGPIF was established to provide property insurance for local government entities that include counties, cities, towns, villages, school districts, and library boards. The fund insures local government property such as government buildings, schools, libraries, and motor vehicles. It covers all property losses except those resulting from flood, earthquake, wear and tear, extremes in temperature, mold, war, nuclear reactions, and embezzlement or theft by an employee.

The data are available using this download button: Download the Wisconsin Property Fund Data

TABLE 22.1: **Variables in the Wisconsin Property Fund Dataset**

Variable	Description
PolicyNum	Policy number
Year	Contract year
Premium	Premium
Deduct	Deductible
BCcov	Coverage for building and contents
Freq	Number of claims during the year (frequency)
Fire5	Binary variable to indicate the fire class is below 5
NoClaimCredit	Binary variable to indicate no claims in the past two years
EntityType	Categorical variable that is one of six types: 1=Village, 2=City,3=County, 4=Misc, 5=School, or Town)
AlarmCredit	Categorical variable that is one of four types: (0, 5, 10, or 15) for automatic smoke alarms in main rooms
BCClaim	Builing and contents claims

TABLE 22.2: **Wisconsin Property Fund First Five Rows**

PolicyNum	Year	Premium	Deduct	BCcov	Freq	Fire5	NoClaimCredit	EntityType	AlarmCredit	BCClaim
120002	2006	9313	1000	22714456	0	1	0	3	1	0
120002	2007	8767	1000	25046646	0	1	0	3	1	0
120002	2008	7090	1000	20851525	0	1	1	3	1	0
120002	2009	8522	1000	21852696	0	1	1	3	1	0
120002	2010	7994	1000	23511493	1	1	1	3	1	6839

TABLE 22.3: **Wisconsin Property Fund Last Five Rows**

PolicyNum	Year	Premium	Deduct	BCcov	Freq	Fire5	NoClaimCredit	EntityType	AlarmCredit	BCClaim
180787	2010	199	500	285000	0	1	1	4	1	0
180788	2010	58344	100000	416739800	1	1	0	4	1	168304
180789	2010	295	500	500988	1	1	0	4	1	1034
180790	2010	2077	1000	3580665	0	1	0	4	4	0
180791	2010	81	500	118800	0	1	0	4	1	0

22.2 ANU Corporate Travel Data

Universities purchase corporate travel policies to cover employees and students traveling on official university business for a wide variety of accidents and incidents while away from the campus or primary workplace. This broad coverage includes medical care and evacuation, loss of personal property, extraction for political and weather related reasons, and more. See Frees and Butt (2022) for more information about this coverage.

There are 2107 observations in this dataset. The variable names are described in Table 22.4 and the first and last five observations are in Table 22.6.

Data are available using this button: Download Corporate Travel Claims Data.

TABLE 22.4: **Variables in the Corporate Travel Dataset**

Variable	Description
UW Year	Underwriting Year
Loss Date	Date that the loss occurred
Reported Date	Date that the loss was reported
Last Trans Date	Last date in which there was a transaction regarding the loss
Paid Loss	Cumulative amount paid on the loss
Outstanding Reserve	Estimate of the loss amount yet to be paid
Incurred Loss	Sum of the amount paid and the estimate of future payments
Status	An indicator as to whether the claim has been deemed settled (closed) or not settled (open)

TABLE 22.5: **Corporate Travel Data First Five Rows**

UW.Year	Loss.Date	Reported.Date	Last.Trans.Date	Paid.Loss	Outstanding.Reserve	Incurred.Loss	Status
2021	19/12/2021	20/12/2021	24/12/2021	10000	0	10000	Closed
2021	9/4/2022	29/04/2022	30/05/2022	423	0	423	Closed
2021	2/5/2022	4/5/2022		0	500	500	Open
2021	5/5/2022	17/05/2022		0	562	562	Open
2021	30/04/2022	27/05/2022	10/6/2022	1500	0	1500	Closed

TABLE 22.6: **Corporate Travel Data Last Five Rows**

UW.Year	Loss.Date	Reported.Date	Last.Trans.Date	Paid.Loss	Outstanding.Reserve	Incurred.Loss	Status
2006	1/11/2006	19/06/2007		0	0	0	Closed
2006	24/06/2007	26/06/2007	8/1/2008	6278	0	6278	Closed
2006	4/7/2007	6/7/2007	11/9/2007	114	0	114	Closed
2006	20/05/2007	26/06/2007	14/07/2007	136	0	136	Closed
2006	15/02/2007	27/06/2007	14/07/2007	1208	0	1208	Closed

Source: Frees, Edward and Butt, Adam (2022). "ANU Corporate Travel Insurance Claims 2022". Australian National University Data Commons. DOI https://doi.org/10.25911/vrdw-9f32.

22.3 ANU Group Personal Accident Data

Group personal accident insurance offers financial protection in case of injury or death resulting from an incident that occurs on the job. Like workers' compensation, group personal accident offers insurance coverage and liability insurance protection against accidental death or injury. Unlike workers' compensation, group personal accident covers students and ANU's voluntary workers. See Frees and Butt (2022) for more information about this coverage.

There are 148 observations in this dataset. The variable names are described in Table 22.7 and the first and last five observations are in Table 22.9.

Data are available using this button: Download Group Personal Accident Claims Data.

TABLE 22.7: **Variables in the Group Personal Accident Dataset**

Variable	Description
UW Year	Underwriting Year
Loss Date	Date that the loss occurred
Last Trans Date	Last date in which there was a transaction regarding the loss.
Paid Loss	Cumulative amount paid on the loss
Outstanding Reserve	Estimate of the loss amount yet to be paid
Incurred Loss	Sum of the amount paid and the estimate of future payments
Status	An indicator as to whether the claim has been deemed settled (closed) or not settled (open)

TABLE 22.8: **Group Personal Accident Data First Five Rows**

UW.Year	Loss.Date	Last.Trans.Date	Paid.Loss	Outstanding.Reserve	Incurred.Loss	Status
2021	6/12/2021	3/6/2022	805	0	805	Closed
2021	15/11/2021		0	0	0	Closed
2021	15/11/2021		0	0	0	Closed
2021	22/03/2022	4/5/2022	396	0	396	Closed
2021	11/4/2022	2/8/2022	740	360	1100	Open

TABLE 22.9: **Group Personal Accident Data Last Five Rows**

UW.Year	Loss.Date	Last.Trans.Date	Paid.Loss	Outstanding.Reserve	Incurred.Loss	Status
2010	6/3/2011	26/07/2011	776	0	776	Closed
2010	22/07/2011	23/01/2012	4625	0	4625	Closed
2010	5/6/2011	30/01/2012	1504	0	1504	Closed
2007	11/1/2008	23/02/2008	0	0	0	Closed
2007	29/08/2008		0	0	0	Closed

Source: Frees, Edward and Butt, Adam (2022). “ANU Group Personal Accident Claims 2022”. Australian National University Data Commons. https://doi.org/10.25911/jcfx-zj56.

22.4 ANU Motor Vehicle Data

This policy covers ANU’s vehicles including cars, vans, utilities, and motorcycles. See Frees and Butt (2022) for more information about this coverage.

There are 318 observations in this dataset. The variable names are described in Table 22.10 and the first and last five observations are in Table 22.12.

Data are available using this button: Download Motor Vehicle Claims Data.

TABLE 22.10: **Variables in the Motor Vehicle Dataset**

Variable	Description
Policy Term Start Date	Start date of the contract year in which the loss occurred
Loss Date	Date that the loss occurred
Reported Date	Date that the loss was reported
Motor Fault	Party responsible for the loss
Driver Age	Age of the driver
Vehicle Description	Type of vehicle
Loss Postcode	Postal code where the loss occurred
Excess	The deductible applied to the loss
Motor Net Paid	Amount paid to the insured (ANU)
Outstanding Estimate	Estimate of the loss amount yet to be paid
Motor Net Incurred	Sum of the amount paid and the estimate of future payments
Third Party Identified	Indicates whether a responsible third party could be identified
Third Party Insured	Indicates whether a responsible third party was insured

TABLE 22.11: **Motor Vehicle Data First Five Rows**

Policy.Term.Start.Date	Loss.Date	Reported.Date	Motor.Fault	Driver.Age	Vehicle.Description	Loss.Postcode	Excess	Motor.Net.Paid	Outstanding.Estimate	Motor.Net.Incurred	Third.Party.Identified	Third.Party.Insured
1/11/2011	6/6/2012	4/10/2012	THIRD PARTY RESPONSIBLE	NA	FORD TRANSIT VAN	2600	1000	385	0	385	IDENTIFIED	
1/11/2011	16/08/2012	14/11/2013	INSURED RESPONSIBLE	39	TOYOTA HIACE	2612	1000	901	0	901		
1/11/2011	4/9/2012	17/01/2013	INSURED RESPONSIBLE	52	HYUNDAI IX35	2600	1000	1226	0	1226		
1/11/2011	21/09/2012	28/09/2012	THIRD PARTY RESPONSIBLE	59	HOLDEN COMMODORE	2518	NA	1672	0	1672	IDENTIFIED	NOT INSURED
1/11/2011	22/09/2012	12/10/2012	INSURED RESPONSIBLE	NA	SUBARU FORESTER	2612	1000	3419	0	3419		INSURED

Source: Frees, Edward and Butt, Adam (2022). "ANU Motor Vehicle Claims 2022". Australian National University Data Commons. DOI https://doi.org/10.25911/g7e4-9e46.

TABLE 22.12: **Motor Vehicle Data Last Five Rows**

Policy.Term.Start.Date	Loss.Date	Reported.Date	Motor.Fault	Driver.Age	Vehicle.Description	Loss.Postcode
1/11/2021	4/4/2022	5/4/2022	INSURED RESPONSIBLE	66	VOLKSWAGEN TIGUAN	2604
11/1/2021	11/4/2022	9/5/2022	INSURED RESPONSIBLE	27	TOYOTA HILUX	2540
1/11/2021	11/4/2022	9/5/2022	INSURED RESPONSIBLE	27	TOYOTA HILUX	2540
11/1/2021	15/04/2022	11/7/2022	INSURED RESPONSIBLE	21	TOYOTA HILVX	2601
1/11/2021	18/07/2022	18/07/2022	NO-ONE RESPONSIBLE	NA	TOYOTA HILUX	2601

Excess	Motor.Net.Paid	Outstanding.Estimate	Motor.Net.Incurred	Third.Party.Identified	Third.Party.Insured
0	2373	1056	3429		
0	210	25000	25210		
0	0	31927	31927		
0	0	2750	2750		
0	0	299	299		

22.5 Spanish Personal Insurance Data

This dataset consists of 10,000 insurance private customers of a real portfolio of insurance policy holders in Spain with a motor insurance and a homeowners insurance contract for policy year 2014. The data contain information on each customer, policies and yearly claims by type of contract.

The data are available using this download button: Download the Spanish Personal Insurance Data

The description of the data appears in Table 22.13.

TABLE 22.13: Variable and Description of Spanish Personal Insurance Data

Variable	Description
gender	1 for male and 0 for female
Age_client	the age of the customer in years
year	Policy year. Equals 5 corresponding to 2014.
age_of_car_M	the number of years since the vehicle was bought by the customer
Car_power_M	the power of the vehicle
Car_2ndDriver_M	1 if the customer has informed the insurance company that a second occasional driver uses the vehicle, and 0 otherwise
num_policiesC	the total number of policies held by the same customer in the insurance company
metro_code	1 for urban or metropolitan and 0 for rural
Policy_PaymentMethodA	1 for annual payment and 0 for monthly payment in the motor policy
Policy_PaymentMethodH	1 for annual payment and 0 for monthly payment in the homeowners policy
Insuredcapital_content_re	the value of content in homeowners insurance
Insuredcapital_continent_re	the value of building in homeowners insurance
appartment	1 if the homeowners insurance correspond to an apartment and 0 otherwise
Client_Seniority	the number of years that the customer has been in the company
Retention	1 if the policy is renewed and 0 otherwise
NClaims1	the number of claims in the motor insurance policy for the corresponding year
NClaims2	the number of claims in the homeowners insurance policy for the corresponding year
Claims1	the sum of claims cost in the motor insurance policy for the corresponding year
Claims2	the sum of claims cost in the homeowners insurance policy for the corresponding year
Types	1 when neither an auto nor a home claim, it is equal to 2 when the customer has an auto but not a home claim, it is equal to 3 when the customer does not have not an auto but a home claim and it is equal to 4 when both an auto and a home claim.
PolID	Policy Identification Number

All monetary units are expressed in Euros. In motor insurance, only claims at fault are considered.

These data were drawn from a larger database of 40,284 insurance private customers. These customers are tracked from 2010 to 2014. Some customers do not renew their policies, so that they do not stay in the sample for five years. For the smaller data, only the 2014 policy year was used and from this, a random sample of 10,000 customers was drawn.

TABLE 22.14: **Spanish Personal Insurance Data First Five Rows**

gender	Age.client	year	age.of.car.M	Car.power.M	Car.2ndDriver.M	num.policiesC
1	47	5	12	163	0	0
1	52	5	13	80	0	1
0	66	5	7	97	0	1
1	70	5	17	95	0	1
1	67	5	13	110	0	1

metro.code	Policy.PaymentMethodA	Policy.PaymentMethodH	Insuredcapital.content.re	Insuredcapital.continent.re	appartment
0	1	1	10	12	1
0	1	1	10	11	0
1	1	1	9	11	1
0	1	1	10	11	1
0	1	1	11	12	0

Client.Seniority	Retention	NClaims1	NClaims2	Claims1	Claims2	Types	PolID
7	1	0	0	0	0	1	12476
18	1	0	0	0	0	1	29232
15	1	0	0	0	0	1	23770
16	1	0	1	0	58	3	8228
6	1	0	0	0	0	1	37088

TABLE 22.15: **Spanish Personal Insurance Data Last Five Rows**

gender	Age.client	year	age.of.car.M	Car.power.M	Car.2ndDriver.M	num.policiesC
1	66	5	8	143	0	1
1	55	5	18	125	1	1
0	41	5	10	190	0	1
1	50	5	5	140	0	1
1	55	5	12	90	0	1

metro.code	Policy.PaymentMethodA	Policy.PaymentMethodH	Insuredcapital.content.re	Insuredcapital.continent.re	appartment
0	1	1	10	11	1
0	1	1	11	11	1
0	1	1	9	12	1
0	1	1	10	12	0
1	1	1	11	13	0

Client.Seniority	Retention	NClaims1	NClaims2	Claims1	Claims2	Types	PolID
20	1	0	0	0	0	1	2967
15	1	0	0	0	0	1	9387
6	1	0	0	0	0	1	36519
8	1	0	0	0	0	1	33276
6	1	0	0	0	0	1	25370

See Frees et al. (2021) for more information about this dataset. The larger database contains 122935 rows and is freely available at:

Source: Guillen, Montserrat; Bolancé, Catalina; Frees, Edward W.; Valdez, Emiliano A. (2021), "Insurance data for homeowners and motor insurance customers monitored over five years", Mendeley Data, V1, DOI https://doi.org/10.17632/vfchtm5y7j.1

22.6 ‘R’ Package CASdatasets

The R package CASdatasets provides a convenient way to access many well-known insurance datasets. This package was originally created to support the book *Computational Actuarial Science with R*, edited by Arthur Charpentier, Charpentier (2014).

To install the package, here is a bit of R code:

```r
install.packages("CASdatasets", repos = "http://cas.uqam.ca/pub/", type = "source")
library(CASdatasets)
`?`(CASdatasets)
`?`(sgautonb  # See the documentation of the Singapore Auto Data
)
`?`(lossalae  # See the documentation of the Loss and Expense Data
)
```

Note that this package assumes that you have already installed a few other packages, including *xts*, *sp*, and *zoo*.

To illustrate,

- in Chapter 3 we use the Singapore data (referred to as sgautonb in the package) and
- in Chapter 16 we use the loss and expense data (referred to as lossalae in the package).

22.7 Other Data Sources

There exists man other (non-actarial) data sources. First, data can be obtained from university-based researchers who collect primary data. Second, data can be obtained from organizations that are set up for the purpose of releasing secondary data for the general research community. Third, data can be obtained from national and regional statistical institutes that collect data. Finally, companies have corporate data that can be obtained for research purposes.

While it might be difficult to obtain data to address a specific research problem or answer a business question, it is relatively easy to obtain data to test a model or an algorithm for data analysis. In the modern era, readers can obtain datasets from the Internet. The following is a list of some websites to obtain real-world data:

- **UCI Machine Learning Repository.** This website (url: http://archive.ics.uci.edu/ml/index.php) maintains more than 400 datasets that can be used to test machine learning algorithms.
- **Kaggle.** The Kaggle website (url: https://www.kaggle.com/) include real-world datasets used for data science competitions. Readers can download data from Kaggle by registering an account.
- **DrivenData.** DrivenData aims at bringing cutting-edge practices in data science to solve some of the world's biggest social challenges. In its website (url: https://www.drivendata.org/), readers can participate in data science competitions and download datasets.
- **Analytics Vidhya.** This website (url: https://datahack.analyticsvidhya.com/contest/all/) allows you to participate and download datasets from practice problems and hackathon problems.
- **KDD Cup.** KDD Cup is the annual Data Mining and Knowledge Discovery competition organized by the ACM Special Interest Group on Knowledge Discovery and Data Mining. This website (url: http://www.kdd.org/kdd-cup) contains the datasets used in past KDD Cup competitions since 1997.
- **U.S. Government's open data.** This website (url: https://www.data.gov/) contains about 200,000 datasets covering a wide range of areas including climate, education, energy, and finance.
- **AWS Public Datasets.** In this website (url: https://aws.amazon.com/datasets/), Amazon provides a centralized repository of public datasets, including some huge datasets.

23

Glossary

Term	Definition	Section
analytics	Analytics is the process of using data to make decisions.	1.1
renters insurance	Renters insurance is an insurance policy that covers the contents of an apartment or house that you are renting.	1.1
automobile insurance	An insurance policy that covers damage to your vehicle, damage to other vehicles in the accident, as well as medical expenses of those injured in the accident.	1.1
casualty insurance	Causalty insurance is a form of liability insurance providing coverage for negligent acts and omissions. examples include workers compensation, errors and omissions, fidelity, crime, glass, boiler, and various malpractice coverages.	1.1
commercial insurance		1.1
term	The duration of an insurance contract	1.1
insurance claim	An insurance claim is the compensation provided by the insurer for incurred hurt, loss, or damage that is covered by the policy.	1.1
homeowners insurance	Homeowners insurance is an insurance policy that covers the contents and property of a building that is owned by you or a friend.	1.1
property insurance	Property insurance is a policy that protects the insured against loss or damage to real or personal property. the cause of loss might be fire, lightening, business interruption, loss of rents, glass breakage, tornado, windstorm, hail, water damage, explosion, riot, civil commotion, rain, or damage from aircraft or vehicles.	1.1
non-life	Non-life insurance is any type of insurance where payments are not based on the death (or survivorship) of a named insured. examples include automobile, homeowners, and so on. also known as property and casualty or general insurance.	1.1

(continued)

Term	Definition	Section
life insurance	Life insurance is a contract where the insurer promises to pay upon the death of an insured person. the person being paid is the beneficiary.	1.1
personal insurance	Insurance purchased by a person	1.1
loss adjustment expenses	Loss adjustment expenses are costs to the insurer that are directly attributable to settling a claims. for example, the cost of an adjuster is someone who assess the claim cost or a lawyer who becomes involve in settling an insurer's legal obligation on a claim	1.2
unallocated	Unallocated loss adjustment expenses are costs that can only be indirectly attributed to claim settlement; for example, the cost of an office to support claims staff	1.2
allocated	Allocated loss adjustment expenses, sometimes known by the acronym alea, are costs that can be directly attributed to settling a claim; for example, the cost of an adjuster	1.2
underwriting	Underwriting is the process where the company makes a decision as to whether or not to take on a risk.	1.2
loss reserving	A loss reserve is an estimate of liability indicating the amount the insurer expects to pay for claims that have not yet been realized. this includes losses incurred but not yet reported (ibnr) and those claims that have been reported claims that haven't been paid (known by the acronym rbns for reported but not settled).	1.2
risk classification	Risk classification is the process of grouping policyholders into categories, or classes, where each insured in the class has a risk profile that is similar to others in the class.	1.2
retrospective premiums	The process of determining the cost of an insurance policy based on the actual loss experience determined as an adjustment to the initial premium payment.	1.2
claims adjustment	Claims adjustment is the process of determining coverage, legal liability, and settling claims.	1.2
claims leakage	Claims leakage respresents money lost through claims management inefficiencies.	1.2
adjuster	An adjuster is a person who investigates claims and recommends settlement options based on estimates of damage and insurance policies held.	1.2

(continued)

Term	Definition	Section
dividends	A dividend is the refund of a portion of the premium paid by the insured from insurer surplus.	1.2
indemnification	Indemnification is the compensation provided by the insurer.	1.3
rating variables	Rating variables are the components of an insurance pricing formula. they can include numeric variables (like values, revenue, or area) and classification variables (like location, type of vehicle, or type of occupancy.)	1.3
frequency	Count random variables that represent the number of claims	2.1
severity	The amount, or size, of each payment for an insured event	2.1
probability mass function (pmf)	A function that gives the probability that a discrete random variable is exactly equal to some value	2.1
distribution function	The chance that the random variable is less than or equal to x, as a function of x	2.1
mean	Average	2.1
moments	The rth moment of a list is the average value of the random variable raised to the rth power	2.1
survival function	The probability that the random variable takes on a value greater than a number x	2.1
moment generating function (mgf)	The mgf of random variable n is defined the expectation of exp(tn), as a function of t	2.2
probability generating function (pgf)	For a random variable n, its pgf is defined as the expectation of s^n, as a function of s	2.2
convex hulls	The convex hull of a set of points x is the smallest convex set that contains x	2.2
risk classes	The formation of different premiums for the same coverage based on each homogeneous group's characteristics.	2.2
binomial distribution	A random variable has a binomial distribution (with parameters m and q) if it is the number of "successes" in a fixed number m of independent random trials, all of which have the same probability q of resulting in "success."	2.2
binary outcomes	Outcomes whose unit can take on only two possible states, traditionally labeled as 0 and 1	2.2

(continued)

Term	Definition	Section
m-convolution	The addition of m independent random variables	2.2
poisson distribution	A discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event	2.2
negative binomial distribution	The number of successes until we observe the rth failure in independent repetitions of an experiment with binary outcomes	2.2
overdispersed	The presence of greater variability (statistical dispersion) in a data set than would be expected based on a given statistical model	2.2
underdispersed	There was less variation in the data than predicted	2.2
(a, b, 0) class	The poisson, binomial and negative binomial distributions	2.3
maximum likelihood estimator (mle)	The possible value of the parameter for which the chance of observing the data largest	2.4
local extrema	The largest and smallest value of the function within a given range	2.4
central limit theorem (clt)	In some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed.	2.4
newton's method	A root-finding algorithm which produces successively better approximations to the roots of a real-valued function	2.4
robust	Resistant to errors in the results, produced by deviations from assumptions	2.4
explanatory variables	In regression, the explanatory variable is the one that is supposed to "explain" the other.	2.5
regression analysis	A set of statistical processes for estimating the relationships among variables	2.5
homogeneous	Units of exposure that face approximately the same expected frequency and severity of loss.	2.5
(a,b,1)	A count distribution with probabilities satisfying p_k/p_{k-1}=a+b/k, for some some constants a and b and k>=2	2.5

(continued)

Term	Definition	Section
zero truncation	Zero modification of a count distribution such that it assigns zero probability to zero count	2.5
degenerate distribution	A deterministic distribution and takes only a single value	2.5
convex combination	A linear combination of points where all coefficients are non-negative and sum to 1	2.5
convex function	A real-valued function defined on an interval is called convex if the line segment between any two points on the graph of the function lies above or on the graph.	2.6
mixture distribution	The probability distribution of a random variable that is derived from a collection of other random variables as follows: first, a random variable is selected by chance from the collection according to given probabilities of selection, and then the value of the selected random variable is realized	2.6
chi-square distribution	The chi-squared distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables	2.7
aic	A goodness of fit measure of a statistical model that describes how well it fits a set of observations.	2.7
pearson’s chi-square test	A statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance	2.7
multinomial likelihood	The multinomial distribution models the probability of counts for rolling a k-sided die n times	2.7
aggregate losses	Aggregate claims, or total claims observed in the time period	3
liability insurance	Insurance that compensates an insured for loss due to legal liability towards others	3
mixture distribution	A weighted average of other distributions, which may be continuous or discrete	3
continuous random variable	Random variable which can take infinitely many values in its specified domain	3.1
raw moment	The kth moment of a random variable x is the average (expected) value of x^k	3.1
central moment	The kth central moment of a random variable x is the expected value of (x-its mean)^k	3.1
skewness	Measure of the symmetry of a distribution, 3rd central moment/standard deviation^3	3.1

(continued)

Term	Definition	Section
kurtosis	Measure of the peaked-ness of a distribution, 4th central moment/standard deviation^4	3.1
expected value	Average	3.1
exponential distribution	A single parameter continous probability distribution that is defined by its rate parameter	3.1
independent	Two variables are independent if conditional information given about one variable provides no information regarding the other variable	3.1
percentile	The pth percentile of a random variable x is the smallest value x_p such that the probability of not exceeding it is p%	3.1
chi-square distribution	A common distribution used in chi-square tests for determining goodness of fit of observed data to a theorized distribution	3.2
light tailed distribution	A distribution with thinner tails than the benchmark exponential distribution	3.2
pareto distribution	A heavy-tailed and positively skewed distribution with 2 parameters	3.2
hazard function	Ratio of the probability density function and the survival function: f(x)/s(x), and represents an instantaneous probability within a small time frame	3.2
weibull distribution	A positively skewed continuous distribution with 2 parameters that can have an increasing or decreasing hazard function depending on the shape parameter	3.2
generalized beta distribution of the second kind	A 4-parameter flexible distribution that encompasses many common distributions	3.2
parametric distributions	Probability distribution defined by a fixed set of parameters	3.3
transformation	A function or method that turns one distribution into another	3.3
distribution function technique	A transformation technique that involves finding the cdf of the transformed distribution through its relation with the original cdf	3.3
change-of-variable technique	A transformation technique that involves finding the pdf of the transformed distribution through its relation with the original pdf using inverse functions	3.3

(continued)

Term	Definition	Section
moment-generating function technique	A transformation technique that uses moment generating functions properties to determine the mgf of a linear combination of variables	3.3
lognormal distribution	A heavy-tailed, positively skewed 2-parameter continuous distribution such that the natural log of the random variable is normally distributed with the same parameter values	3.3
reliability data	A dataset consisting of failure times for failed units and run times for units still functioning	3.3
power transformation	A transformation type that involves raising a random variable to a power	3.3
exponential transformation	A transformation type that involves raising a random variable in the exponent	3.3
mixing parameters	Proportion weight given to each subpopulation in a mixture	3.3
heterogeneous population	A dataset where the subpopulations are represented by separate distinct distributions	3.3
finite mixture	A mixture distribution with a finite k number of subpopulations	3.3
continuous mixture	A mixture distribution with an infinite number of subpopulations, where the mixing parameter is itself a continuous distribution	3.3
conditional distribution	A probability distribution that applies to a subpopulation satisfying the condition	3.3
unconditional distribution	A probability distribution independent of any another imposed conditions	3.3
prior distribution	A probability distribution assigned prior to observing additional data	3.3
scale distribution	A distribution with the property that multiplying all values by a constant leads to the same distribution family with only the scale parameter changed	3.3
moral hazard	Situation where an insured is more likely to be risk seeking if they do not bear sufficient consequences for a loss	3.4
payment per loss	Amount insurer pays when a loss occurs and can be 0	3.4
payment per payment	Amount insurer pays given a payment is needed and is greater than 0	3.4

(continued)

Term	Definition	Section
left censored	Values below a threshold d are not ignored but converted to 0	3.4
left truncated	Values below a threshold d are not reported and unknown	3.4
loss elimination ratio (ler)	% decrease of the expected payment by the insurer as a result of the deductible	3.4
franchise deductible	Insurer pays nothing for losses below the deductible, but pays the full amount for any loss above the deductible	3.4
limit of coverage	Policy limit, or maximum contractual financial obligation of the insurer for a loss	3.4
group insurance	Insurance provided to groups of people to take advantage of lower administrative costs vs. individual policies	3.4
growth factor	Multiplicative factor applied to a distribution to account for the impact of inflation, typically (1+rate)	3.4
cedent	Party that is transferring the risk to a reinsurer	3.4
excess of loss coverage	Contract where an insurer pays all claims up to a specified amount and then the reinsurer pays claims in excess of stated reinsurance deductible	3.4
retention	Maximum amount payable by the primary insurer in a reinsurance arrangement	3.4
right censored variable	Values above a threshold u are not ignored but converted to u	3.4
reinsurance	A transaction where the primary insurer buys insurance from a re-insurer who will cover part of the losses and/or loss adjustment expenses of the primary insurer	3.4
method of maximum likelihood	Statistical method used to derive the parameter values from data that maximize the probability of observing the data given the parameters	3.5
grouped data	Data bucketed into categories with ranges, such as for use in histograms or frequency tables	3.5
large-sample properties	Asymptotic properties of a distribution as the amount of data increases towards infinity	3.5
asymptotic variance	Variability of the distribution of an estimator as the amount of data increases towards infinity	3.5
delta method	Statistical method used to approximate the asymptotic variance for a function based on parameters whose asymptotic variance can be determined	3.5

(continued)

Term	Definition	Section
log-likelihood function	Natural log of the likelihood function	3.5
covariance matrix	Matrix where the (i,j)^th element represents the covariance between the ith and jth random variables	3.5
complete data	Data where each individual observation is known, and no values are censored, truncated, or grouped	3.5
parametric	Distributional assumptions made on the population from which the data is drawn, with properties defined using parameters.	4.1
nonparametric	No distributional assumptions are made on the population from which the data is drawn.	4.1
sampling scheme	How the data is obtained from the population and what data is observed.	4.1
unbiased	An estimator that has no bias, that is, the expected value of an estimator equals the parameter being estimated.	4.1
plug-in principle	The plug-in principle or analog principle of estimation proposes that population parameters be estimated by sample statistics which have the same property in the sample as the parameters do in the population.	4.1
indicator	A categorical variable that has only two groups. the numerical values are usually taken to be one to indicate the presence of an attribute, and zero otherwise. another name for a binary variable.	4.1
empirical distribution function	The empirical distribution is a non-parametric estimate of the underlying distribution of a random variable. it directly uses the data observations to construct the distribution, with each observed data point in a size-n sample having probability 1/n.	4.1
first quartile	The 25th percentile; the number such that approximately 25% of the data is below it.	4.1
third quartile	The 75th percentile; the number such that approximately 75% of the data is below it.	4.1
quantile	The q-th quantile is the point(s) at which the distribution function is equal to q, i.e. the inverse of the cumulative distribution function.	4.1
smoothed empirical quantile	A quantile obtained by linear interpolation between two empirical quantiles, i.e. data points.	4.1

(continued)

Term	Definition	Section
bandwidth	A small positive constant that defines the width of the steps and the degree of smoothing.	4.1
kernel density estimator	A nonparametric estimator of the density function of a random variable.	4.1
bias-variance tradeoff	The tradeoff between model simplicity (underfitting; high bias) and flexibility (overfitting; high variance).	4.1
model diagnostics	Procedures to assess the validity of a model	4.1
probability-probability (pp) plot	A plot that compares two models through their cumulative probabilities.	4.1
quantile-quantile (qq) plot	A plot that compares two models through their quantiles.	4.1
goodness of fit statistics	A measure used to assess how well a statistical model fits the data, usually by summarizing the discrepancy between the observations and the expected values under the model.	4.1
goodness of fit	A measure used to assess how well a statistical model fits the data, usually by summarizing the discrepancy between the observations and the expected values under the model.	4.1
method of moments	The estimation of population parameters by approximating parametric moments using empirical sample moments.	4.1
percentile matching	The estimation of population parameters by approximating parametric percentiles using empirical quantiles.	4.1
percentile	A 100p-th percentile is the number such that 100 times p percent of the data is below it.	4.1
gini index	A measure for assessing income inequality. it measures the discrepancy between the income and population distributions and is calculated from the lorenz curve.	4.2
model selection	The process of selecting a statistical model from a set of candidate models using data.	4.2
in-sample	A dataset used for analysis and model development. also known as a training dataset.	4.2
out-of-sample	A dataset used for model validation. also known as a test dataset.	4.2

(continued)

Term	Definition	Section
cross-validation	A model validation procedure in which the data sample is partitioned into subsamples, where splits are formed by separately taking each subsample as the out-of-sample dataset.	4.2
model validation	The process of confirming that the proposed model is appropriate.	4.2
data-snooping	Repeatedly fitting models to a data set without a prior hypothesis of interest.	4.2
predictive inference	Preditive inference is the process of using past data observations to predict future observations.	4.2
likelihood function	A function of the likeliness of the parameters in a model, given the observed data.	4.3
ogive estimator	A nonparametric estimator for the distribution function in the presence of grouped data.	4.3
product-limit estimator	A nonparametric estimator of the survival function in the presence of incomplete data. also known as the kaplan-meier estimator.	4.3
risk set	The number of observations that are active (not censored) at a specific point.	4.3
nelson-aalen	A nonparametric estimator of the cumulative hazard function in the presence of incomplete data.	4.3
credibility	An actuarial method of balancing an individual's loss experience and the experience in the overall portfolio to improve ratemaking estimates.	4.4
bayesian	A type of statistical inference in which the model parameters and the data are random variables.	4.4
predictive distribution	The distribution of new data, conditional on a base set of data, under the bayesian framework.	4.4
least squares	A technique for estimating parameters in linear regression. it is a standard approach in regression analysis to the approximate solution of overdetermined systems. in this technique, one determines the parameters that minimize the sum of squared differences between each observation and the corresponding linear combination of explanatory variables.	4.4
markov chain monte carlo (mcmc) simulation	The class of numerical methods that use markov chains to generate draws from a posterior distribution.	4.4

(continued)

Term	Definition	Section
improper prior	A prior distribution in which the sum or integral of the distribution is not finite.	4.4
confidence interval	Another term for interval estimate. unlike a point estimate, it gives a range of reliability for approximating a parameter of interest.	4.4
decision analysis	Bayesian decision theory is the study of an agent's choices, which is informed by bayesian probability.	4.4
conjugate distributions	Distributions such that the posterior and the prior come from the same family of distributions.	4.4
credibility interval	A summary of the posterior distribution of parameters under the bayesian framework.	4.4
prior distribution	The distribution of the parameters prior to observing data under the bayesian framework.	4.4
exposure	A measure of the rating units for which rates are applied to determine the premium. for example, exposures may be measured on a per unit basis (e.g. a family with auto insurance under one contract may have an exposure of 2 cars) or per $1,000 of value (e.g. homeowners insurance).	5.1
inflation	Inflation is a sustained increase in the general price level of goods and services over a period of time.	5.1
business line		5.1
individual risk model	A modeling approach for aggregate losses in which the loss from each individual contract is considered.	5.1
collective risk model	A modeling approach for aggregate losses in which the aggregate loss is represented in terms of a frequency distribution and a severity distribution.	5.1
coverage	Insurance coverage is the amount of risk or liability that is covered for an individual or entity by an insurance policy.	5.1
frequency distribution	The random number of claims that occur under the collective risk model.	5.1
severity distribution	The randomly distributed amount of each loss under the collective risk model.	5.1
central limit theorem	Given certain conditions, the arithmetic mean of a large number of replications of independent random variables, each with a finite mean and variance, will be approximately normally distributed, regardless of the underlying distribution.	5.2

(continued)

Term	Definition	Section
term life insurance	A term life insurance policy is payable only if death of the insured occurs within a specified time, such as 5 or 10 years, or before a specified age.	5.2
pure endowment	A pure endowment is an insurance policy that is payable at the end of the policy period if the insured is still alive. if the insured has died, there is nothing paid in the form of benefits.	5.2
support	The set of all outcomes for a random variable following some distribution. for example, exponentially distributed random variable x has support x>0.	5.2
convolution	The convolution of probability distributions is the distribution corresponding to the addition of independent random variables.	5.2
law of iterated expectations	A decomposition of the expected value of a random variable into conditional components. specifically, for random variables x and y, e(x) = e[e(x|y)].	5.3
compound distribution	A random variable follows a compound distribution if it is parameterized and contains at least one parameter that is itself a random variable. for example, the tweedie distribution is a compound distribution.	5.3
tweedie distribution	A compound distribution that is a poisson sum of gamma random variables. because it can accommodate a discrete probability mass at zero and a continuous positive component, it is suitable for modeling aggregate insurance claims.	5.3
shape parameter	A numerical parameter of a parametric distribution affecting the shape of a distribution rather than simply shifting it (as a location parameter does) or stretching/shrinking it (as a scale parameter does).	5.3
scale parameter	A numerical parameter of a parametric distribution that stretches/shrinks the distribution without changing its location or shape. the larger the scale parameter, the more spread out the distribution. the scale parameter is also the reciprocal of the rate parameter. for example, the normal distribution has scale parameter \sigma.	5.3
exponential dispersion	A set of distributions that represents a generalisation of the natural exponential family and also plays an important role in generalized linear models.	5.3

(continued)

Term	Definition	Section
generalized linear models	Commonly known by the acronym glm. an extension of the linear regression model where the dependent variable is a member of the linear exponential family. glm encompasses linear, binary, count, and long-tailed, regressions all as special cases.	5.3
exponential family	A family of parametric distributions that are practical for modeling the underlying response variable in generalized linear models. this family includes the normal, bernoulli, poisson, and tweedie distributions as special cases, among many others.	5.3
monte carlo simulation	A computerized statistical model that simulates the effects of various types of uncertainty.	5.4
empirical distribution	The empirical distribution is a non-parametric estimate of the underlying distribution of a random variable. it directly uses the data observations to construct the distribution, with each observed data point in a size-n sample having probability 1/n.	5.4
converge	A type of stochastic convergence for a sequence of random variables x_1,. . . , x_n that approaches some other distribution as n approaches \infty.	5.4
policy limits	A policy limit is the maximum value covered by a policy.	5.5
ground-up loss	The total amount of loss sustained before policy adjustments are made (i.e. before deductions are applied for coinsurance, deductibles, and/or policy limits.)	5.5
per-loss basis	Due to policy modifications (e.g. deductibles), not all losses that occur result in payment. the per-loss basis considers every loss that occurs.	5.5
per-payment basis	Due to policy modifications (e.g. deductibles), not all losses that occur result in payment. the per-payment basis which considers only the losses that result in some payment to the insured.	5.5
memoryless	The memoryless property means that a given probability distribution is independent of its history and what has already elapsed. specifically, random variable x is memoryless if pr(x > s+t | x >= s) = pr(x > t). note that it does not mean x > s+t and x >= s are independent events.	5.5

(continued)

Term	Definition	Section
central limit theorem	The sample mean and sample sum of a random sample of n from a population will converge to a normal curve as the sample size n grows	6.1
simulations	A computer generation of various hypothetical conditions and outputs, based on the model structure provided	6.1
linear congruential generator	Algorithm that yields pseudo-randomized numbers calculated using a linear recursive relationship and a starting seed value	6.1
pseudo-random numbers	Values that appear random but can be replicated by formula	6.1
inverse transform method	Samples a uniform number between 0 and 1 to represent the randomly selected percentile, then uses the inverse of the cumulative density function of the desired distribution to simulate from in order to find the simulated value from the desired distribution	6.1
quantile function	Inverse function for the cumulative density function which takes a percentile value in [0,1] as the input, and outputs the corresponding value in the distribution	6.1
greatest lower bound	Largest value that is less than or equal to a specified subset of values/elements	6.1
universal life insurance	Type of cash value life insurance where the policy's cash value is the excess of premium payments over the cost of insurance, accumulated with interest, with adjustable premiums and coverage over time	6.1
variable life insurance	Type of life insurance whose face value and coverage term can vary depending upon the performance of underlying invested securities	6.1
sampling variability	How much an estimate can vary between samples	6.1
cauchy distribution	A continuous distribution that represents the distribution of the ratio of two independent normally random variables, where the denominator distribution has mean zero	6.1
kolmogorov-smirnov test	A nonparametric statistical test used to determine if a data sample could come from a hypothesized continuous probability distribution	6.1

(continued)

Term	Definition	Section
bootstrap	A method of sampling with replacement from the original dataset to create additional simulated datasets of the same size as the original	6.2
nonparametric approach	A statistical method where no assumption is made about the distribution of the population	6.2
parametric approach	A statistical method where a prior assumption is made about the distribution or model form	6.2
bias	The difference between the expected value of an estimator and the parameter being estimated. bias is an estimation error that does not become smaller as one observes larger sample sizes.	6.2
bias-corrected estimator	If an estimator is known to be consistently biased in a manner, it can be corrected using a factor to be come less biased or unbiased	6.2
jensen inequality	For a convex function f(x), f(expected value of x) <= expected value of f(x)	6.2
natural estimator	An estimator that uses the sample moments as the estimators for the population	6.2
percentile bootstrap interval	Confidence interval for the parameter estimates determined using the actual percentile results from the bootstrap sampling approach, as every bootstrap sample has an associated parameter estimate(s) that can be ranked against the others	6.2
k-fold cross-validation	A type of validation method where the data is randomly split into k groups, and each of the k groups is held out as a test dataset in turn, while the other k-1 gropus are used for distribution or model fitting, with the process repeated k times in total	6.3
leave-one-out cross validation	A special case of k-fold cross validation, where each single data point gets a turn in being the lone hold-out test data point, and n separate models in total are built and tested	6.3
jackknife statistics	To calculate an estimator, leave out each observation in turn, calculate the sample estimator statistic each time, and average over the n separate estimates	6.3
accept-reject mechanism	A sampling method that is used where the random sample is discarded if not within a certain pre-specified range [a, b] and is commonly used when the traditional inverse transform method cannot be easily used	6.4

(continued)

Term	Definition	Section
importance sampling mechanism	Type of sampling method where values in the region of interest can be over-sampled or values outside the region of interest can be under-sampled	6.4
ergodic theorem	Ergodic theory studies the behavior of a dynamical system when it is allowed to run for an extended time	6.5
markov process	A stochastic (time dependent) process that satisfies memorylessness, meaning future predictions of the process can be made solely based on its present state and not the historical path	6.5
invariant measure	Any mathematical measure that is preserved by a function (the mean is an example)	6.5
composants	Component (smaller, self-contained part of larger entity)	6.5
hastings metropolis	A markov chain monte carlo (mcmc) method for random sampling from a probability distribution where values are iteratively generated, with the distribution of the next sample dependent only on the current sample value, and at each iteration, the candidate sample can be either accepted or rejected	6.5
premium	Amount of money an insurer charges to provide the coverage described in the policy	7.1
ratemaking	Process used by insurers to calculate insurance rates, which drive insurance premiums	7.1
insurance rates	Amount of money needed to cover losses, expenses, and profit per one unit of exposure	7.1
insured contingent event	A condition that results in an insurance claim	7.1
expected costs	The cost to an insurer of payments to the insured and allocated loss adjustment expenses (alaes). overhead and profit are not included	7.1
underwriting profit	Profit an insurer derives from providing coverage, excluding investment income	7.1
experience rating	A type of rating plan that uses the insured's historical loss experience as part of the premium determination	7.1
price	A quantity, usually of money, that is exchanged for a good or service	7.1
rates	A rate is the price, or premium, charged per unit of exposure. a rate is a premium expressed in standardized units.	7.1
technical prices		7.1

(continued)

Term	Definition	Section
loss cost	The sum of losses divided by an exposure; it is also known as the pure premium.	7.2
profit loading	A factor or percentage applied to the premium calculation to account for insurer profit in a policy	7.2
indicated change factor	A factor calculated from the loss ratio method that calculates how the rates should change, with factors $>$ 1 indicating an increase and vice versa	7.2
indicated rate	In a rate filing, the amount that the loss experience suggests that the insurer should charge to cover costs.	7.2
credibility	Weight assigned to observed data vs. that assigned to an external or broader-based set of data	7.4
parametric distribution	Model assumption that the sample data comes from a population that can be modeled by a probability distribution with a fixed set of parameters	7.4
commercial business property	Line of business that insures against damage to their buildings and contents due to a covered cause of loss	7.4
continuous variables	Type of variable that can take on any real value	7.4
discrimination	Process of determining premiums on the basis of likelihood of loss. insurance laws prohibit "unfair discrimination".	7.4
rating factor	A rating factor, or rating variable, is a characteristic of the policyholder or risk being insured by which rates vary.	7.4
rating variable	A rating factor, or rating variable, is a characteristic of the policyholder or risk being insured by which rates vary.	7.4
factor	A variable that varies by groups or categories.	7.4
relativity	The difference of the expected risk between a specific level of a rating factor and an accepted baseline value. this difference may be arithmetic or proportional.	7.4
scale distribution	Suppose that y = c x, where x comes from a parametric distribution family and c is a positive constant. the distribution is said to be a scale distribution if (i) the distributions of y and x come from the same family and (ii) only a single parameter differs and that by a factor of c.	7.4

(continued)

Term	Definition	Section
written exposures	Exposure is based off policies written/issued	7.5
earned exposures	Exposure is based off amount exposed to loss for which coverage has been provided	7.5
unearned exposures	Exposure amount for which coverage has not yet been provided	7.5
in force exposures	Exposure amount subject to loss at a particular point in time	7.5
calendar year method	Experience for rating is aggregated based on calendar year, as opposed to other methods such as when a policy term began	7.5
accident date	Date of loss occurrence that gives rise to a claim	7.5
report date	Date when insurer is notified of the claim	7.5
open claim	A claim that has been reported but not yet closed	7.5
mix of business	Different types of policies in an insurer’s portfolio	7.5
on-level earned premium	Earned premium of historical policies using the current rate structure	7.5
experience loss ratio	Ratio of experience loss to on-level earned premium in the experience period	7.5
claim	The amount paid to an individual or corporation for the recovery, under a policy of insurance, for loss that comes within that policy.	7.5
incurred but not reported	A claim is said to be incurred but not reported if the insured event occurs prior to a valuation date (and hence the insurer is liable for payment) but the event has not been reported to the insurer.	7.5
closed	A claim is said to be closed when the company deems its financial obligations on the claim to be resolved.	7.5
valuation date	A valuation date is the date at which a company summarizes its financial position, typically quarterly or annually.	7.5
policy year	This is the period between a policy’s anniversary dates.	7.5
gini index	The gini index is twice the area between a lorenz curve and a 45 degree line.	7.6
line of equality	45 degree line equating x and y, that represents a perfect alignment in the sample and population distribution	7.6
pp plot	Statistical plot used to assess how close a data sample matches a theorized distribution	7.6

(continued)

Term	Definition	Section
performance curve	A concentration curve is a graph of the distribution of two variables, where both variables are ordered by only one of variables. for insurance applications, it is a graph of distribution of losses versus premiums, where both losses and premiums are ordered by premiums.	7.6
community rating	This generally refers to the premium principle where all risks pay the same amount.	7.6
market conduct regulation	Regulation that ensures consumers obtain fair and reasonable insurance prices and coverage	7.7
government prescribed	Government sets the entire rating system including coverages	7.7
prior approval	Regulator must approve rates, forms, rules filed by insurers before use	7.7
no file	Insurers may use new rates, forms, rules without approval from regulators	7.7
file only	Insurers must file rates, forms, rules for record keeping and use immediately	7.7
rating factors	Characteristics of a risk that help price the insurance contract	8
multiplicative tariff model	A rating method where each rating factor is the product of parameters associated with that rating factor	8
risk characteristics	The distinguishing features of a policy that help determine the expected loss on the policy	8.1
gross insurance premium	Sum of expected losses and expenses and profit on a policy	8.1
adverse selection	A pricing structure that entices riskier individuals to purchase and discourages low-risk individuals from purchasing	8.1
adverse selection spiral	Phenomenon where a book of business deteriorates as it attracts ever-riskier individuals when forced to increase premiums due to losses	8.1
a priori variables	Variables which the insurer has prior knowledge of before the policy inception	8.1
closed-form expressions	A mathematical expression that can be well defined with a formula that has a finite number of operations	8.2
levels	Different outcomes of a categorical variable	8.2
nominal	A categorical variable where the categories do not have a natural order and any numbering is arbitrary	8.2

(continued)

Term	Definition	Section
dummy variables	A variable that takes on a value of 0 or 1 to indicate the absence or presence of a categorical characteristic	8.2
log linear form	Linear regression model where the response variable is the natural log of the expected response value	8.2
base case	The categorical level chosen as the default with all dummy variable indicators of 0	8.2
workers compensation	A no-fault insurance system prescribed by state law where benefits are provided by an employer to an employee due to a job-related injury, including death, resulting from an accident or occupational disease	8.2
exposure bases	The unit of measurement chosen to represent the exposure for a particular risk	8.2
offset	Natural log of the exposure amount that is added to a regression model to account for varying exposures	8.2
tariff	A table or list that contains the rating factors and associated premiums and other risk information	8.3
in-force times	The timeframe during which a policy is active and the insurer is bound by the contractual obligation	8.3
rate parameter	Parameter in certain distributions, such as the exponential, that indicate how quickly the function decays, and it is the reciprocal of the scale parameter	8.3
functional forms	The algebraic relationship between a dependent variable and explanatory variables	8.3
multiplicative form	Relationship where the dependent variable is a product of the explanatory variables	8.3
base tariff cell	The chosen set of rating categories where the rate equals the intercept of the model (the base value)	8.3
relativities	A numerical estimate of value in one category relative to the value in a base classification, typically expressed as a factor	8.3
non-automobile vehicles	Motorized vehicles which are not autos, such as atvs, off-road vehicles, go-carts, etc.	8.3
distributional structure	The manner in which a statistical distribution is parameterized	8.3
information matrix	Matrix that measures the amount of information that an observable random variable x carries about an unknown parameter of a distribution, and is used to calculate covariance matrices of maximum likelihood estimators	8.5

(continued)

Term	Definition	Section
classification rating plan	A rating plan that uses an insured's risk characteristics to determine premium	9.1
credibility weight	The weight assigned to an insured's historical loss experience for the purposes of determining their premium in an experience rating plan	9.1
complement of credibility	The remainder of the weight not assigned to an insured's historical loss experience in the experience rating plan	9.1
class rate	Average rate per exposure for an insured in a particular classification group	9.1
full credibility standard	The threshold of experience necessary to assign 100% credibility to the insured's own experience	9.2
limited fluctuation credibility	A credibility method that attempts to limit fluctuations in its estimates	9.2
cumulative distribution function of the standard normal	Cumulative density function for the normal distribution with mean 0 and standard deviation 1	9.2
buhlmann credibility	A credibility method that uses the amount of experience, expected value of the process variance, and variance of the hypothetical means to determine the credibility weight	9.3
collective mean	The mean estimate of a risk when no loss information about the risk is known	9.3
law of total expectation	The expected value of the conditional expected value of x given y is the same as the expected value of x	9.3
risk parameter	Parameter in a distribution whose value reflects the risk categorization	9.3
expected value of the process variance	Average of the natural variability of observations from within each risk	9.3
variance of the hypothetical means	Variance of the means across different classes, used to determine how similar or different the classes are from one another	9.3
buhlmann-straub credibility	An extension of the buhlmann credibility model that allows for varying exposure by year	9.4

(continued)

Term	Definition	Section
bayes theorem	A probability law that expresses conditional probability of the event a given the event b in terms of the conditional probability of the event b given the event a and the unconditional probability of a	9.5
bayesian inference	A branch of statistics that leverages bayes theorem to update the distribution as more experience becomes available	9.5
gamma-poisson model	A statistical model that assumes the frequency of claims is poisson whose mean has a prior distribution that is a gamma distribution	9.5
exact credibility	A situation where the bayesian credibility estimate matches that of the buhlmann credibility estimate	9.5
beta-binomial model	A statistical model for modeling the probability of an event using the binomial distribution with a probability that has a prior distribution from a beta distribution	9.5
nonparametric estimation	Statistical method that allows the functional form of a fit from data to have no assumed prior distribution, constraints, or parameters	9.5
empirical bayes methods	Credibility methods that estimate the credibility weight without using any assumptions about prior distributions or likelihoods, instead relying only on empirical data	9.5
semiparametric estimation	Credibility method that assumes a distribution for the loss per exposure random variable and otherwise uses empirical data	9.5
portfolios	A collection of contracts	10.1
insurance portfolios	A collection, or aggregation, of insurance contracts	10.1
reinsurers	A company that sells reinsurance	10.1
heavy tailed	A rv is said to be heavy tailed if high probabilities are assigned to large values	10.2
survival function	One minus the distribution function. it gives the probability that a rv exceeds a specific value.	10.2
coherent risk measure	A risk measure that is is subadditive, monontonic, has positive homogeneity, and is translation invariant.	10.3
mean excess loss function	The expected value of a loss in excess of a quantity, given that the loss exceeds the quantity	10.3
risk measure	A measure that summarizes the riskiness, or uncertainty, of a distribution	10.3
value-at-risk	A risk measure based on a quantile function	10.3

(continued)

Term	Definition	Section
ceding company	A company that purchases reinsurance (also known as the reinsured)	10.4
excess of loss	Under an excess of loss arrangement, the insurer sets a retention level for each claim and pays claim amounts less than the level with the reinsurer paying the excess.	10.4
primary insurance	Insurance purchased by a non-insurer	10.4
proportional reinsurance	An agreement between a reinsurer and a ceding company (also known as the reinsured) in which the reinsurer assumes a given percent of losses and premium	10.4
quota share	A proportional treaty where the reinsurer receives a flat percent of the premium for the book of business reinsured and pays a percentage of losses, including allocated loss adjustment expenses. the reinsurer may also pays the ceding company a ceding commission which is designed to reflect the differences in underwriting expenses incurred.	10.4
reinsured	A company that purchases reinsurance (also known as the ceding company)	10.4
retained line	The amount of exposure that the the reinsured retains on a given line in a surplus share reinsurance agreement.	10.4
retention function	A function that maps the insurer portfolio loss into the amount of loss retained by the insurer.	10.4
stop-loss	Under a stop-loss arrangement, the insurer sets a retention level and pays in full total claims less than the level with the reinsurer paying the excess.	10.4
surplus share	A proportional reinsurance treaty that is common in commercial property insurance. a surplus share treaty allows the reinsured to limit its exposure on any one risk to a given amount (the retained line). the reinsurer assumes a part of the risk in proportion to the amount that the insured value exceeds the retained line, up to a given limit (expressed as a multiple of the retained line, or number of lines).	10.4
treaty	A reinsurance contract that applies to a designated book of business or exposures.	10.4
bonus-malus system	A type of rating mechanism where insured premiums are adjusted based on their individual loss experience history	12.1

(continued)

Term	Definition	Section
no claim discount (ncd) system	A type of experience rating where insureds obtain discounts on future years' premiums based on claims-free experience	12.1
hunger for bonus	Phenomenon where insureds under an experience rating system are dissuaded from filing minor claims in order to keep their no-claims discount	12.1
takaful	Co-operative system of reimbursement or repayment in case of loss as an insurance alternative	12.2
markov chain	A stochastic model (time dependent) where the probability of each event depends only on the current state and not the historical path	12.3
transition matrix	Matrix that represents all probabilities for transition from one state to another (could be same state) for a markov chain	12.3
stationary distribution	Probability distribution remains unchanged in the markov chain as time progresses	12.4
ergodic	Irreducible markov chain where it is eventually possible to move from any state to any other state, with positive probability	12.4
irreversible	A markov chain where there does not exist a probability distribution that allows for the chain to be walked backwards in time	12.4
eigenvector	A non-zero vector that changes by only a scalar factor when that linear transformation is applied	12.4
n-step transition probability	Probability of ending in a state j after n periods, starting in state i, where i and j can be the same state	12.4
convergence rate	After n transitions, the sum of variation between the probability in each state vs. the stationary probability	12.4
poisson regression model	Type of regression model used for fitting data with an integral (count) response variable with mean equal to the variance	12.5
negative binomial regression model	Type of regression model used for fitting data with an integral (count) response variable and can account for variance greater than the mean	12.5
overdispersion	Phenomenon where the variance of data is larger than what is modeled	12.5
cross-classified rating classes	Table that combines the effects of multiple rating classifications	12.5

(continued)

Term	Definition	Section
structured data	Data that can be organized into a repository format, typically a database	13.1
unstructured data	Data that is not in a predefined format, most notably text, audio visual	13.1
qualitative data	Data which is non numerical in nature	13.1
quantitative data	Data which is numerical in nature	13.1
ordinal data	Data field with a natural ordering	13.1
interval data	Continuous data which is broken into interval bands with a natural ordering	13.1
key-value databases	Data storage method that stores amd finds records using a unique key hash	13.1
column-oriented databases	Data storage method that stores records by column instead of by row	13.1
document databases	Data storage method that uses the document metadata for search and retrieval, also known as semi-structured data	13.1
data decay	Corruption of data due to hardware failure in the storage device	13.1
reverification	Manual process of checking the integrity of data	13.1
data element analysis	Analysis of the format and definition of each field	13.1
structural analysis	Statistical analysis of the structured data present to detect irregularities	13.1
robust	Statistics which are more unaffected by outliers or small departures from model assumptions	13.2
exploratory data analysis	Approach to analyzing data sets to summarize their main characteristics, using visual methods, descriptive statistics, clustering, dimension reduction	13.2
confirmatory data analysis	Process used to challenge assumptions about the data through hypothesis tests, significance testing, model estimation, prediction, confidence intervals, and inference	13.2
supervised learning methods	Model that predicts a response target variable using explanatory predictors as input	13.2
unsupervised learning methods	Models that work with explanatory variables only to describe patterns or groupings	13.2

(continued)

Term	Definition	Section
classification methods	Supervised learning method where the response is a categorical variable	13.2
regression methods	Classical supervised learning method where the response may be continuous, binary, or a mixture of discrete and continuous	13.2
model flexibility	A measure of model complexity, typically based on the number of estimated parameters	13.2
explanatory modeling	Process where the modeling goal is to identify variables with meaningful and statistically significant relationships and test hypotheses	13.2
predictive modeling	Process where the modeling goal is to predict new observations	13.2
data modeling	Assumes data generated comes from a stochastic data model	13.2
algorithmic modeling	Assumes data generated comes from unknown algorithmic models	13.2
predictive accuracy	Quantitative measure of how well the explanatory variables predict the response outcome	13.2
scripts	A program or sequence of instructions that is executed by another program	13.2
reproducible analysis	Modeling practice where data, code, analyses are published together in a manner so that others may verify the findings	13.2
literate programming	Coding practice where documentation and code are written together	13.2
data ownership	Governance process that details legal ownership of enterprise-wide data and outlines who has ability to create, edit, modify, share and restrict access to the data	13.2
machine learning	Study of algorithms and statistical models that perform a specific task without using explicit instructions, relying on patterns and inference	13.3
pattern recognition	Automated recognition of patterns and regularities in data	13.3
data mining	Process of collecting, cleaning, processing, analyzing, and discovering patterns and useful insights from large data sets	13.3
principal component analysis	Dimension reduction technique that uses orthogonal transformations to convert a set of possibly correlated variables into a set of linearly uncorrelated variables	13.3

(continued)

Term	Definition	Section
cluster analysis	Unsupervised learning method that aims to splot data into homogenous groups using a similarity measure	13.3
k-means algorithm	Type of clustering that aims to partition data into k mutually exclusive clusters by assigning observations to the cluster with the nearest centroid	13.3
linear regression	Supervised model that uses a linear function to approximate the relationship between the target and explanatory variables	13.3
generalized linear model	Supervised model that generalizes linear regression by allowing the linear component to be related to the response variable via a link function and by allowing the variance of each measurement to be a function of its predicted value	13.3
systematic component	The linear combination of explanatory variables component in a glm	13.3
link function	Function that relates between the linear predictor component to the mean of the target variable	13.3
decision trees	Modeling technique that uses a tree-like model of decisions to divide the sample space into non-overlapping regions to make predictions	13.3
categorical variable	A variable whose values are qualitative groups and can have no natural ordering (nominal) or an ordering (ordinal)	14.1
variables	A variable is any characteristics, number, or quantity that can be measured or counted.	14.1
interval variable	An ordinal variable with the additional property that the magnitudes of the differences between two values are meaningful	14.1
spatial data	Data and information having an implicit or explicit association with a location relative to the earth	14.1
high dimensional	Data set is high dimensional when it has many variables. In many applications, the number of variables may be larger than the sample size.	14.1
qualitative	This is a type of variable in which the measurement denotes membership in a set of groups, or categories	14.1
nominal variable	This is a type of qualitative/ categorical variable which has two or more categories without having any kind of natural order.	14.1

(continued)

Term	Definition	Section
ordinal variable	This is a type of qualitative/ categorical variable which has two or more ordered categories.	14.1
binary variable	Is a special type of categorical variable where there are only two categories.	14.1
quantitative variable	A quantitative variable is a type of variable in which numerical level is a realization from some scale so that the distance between any two levels of the scale takes on meaning.	14.1
continuous variable	A continuous variable is a quantitative variable that can take on any value within a finite interval.	14.1
policyholder	Person in actual possession of insurance policy; policy owner.	14.1
discrete variable	A discrete variable is quantitative variable that takes on only a finite number of values in any finite interval.	14.1
count variable	A count variable is a discrete variable with values on nonnegative integers.	14.1
circular data	In a circular data, all values around the circle are equally likely. Example, imagine an analog picture of a clock.	14.1
insurers	An insurance company authorized to write insurance under the laws of any state.	14.1
multivariate	Multivariate variable involves taking many measurements on a single entity.	14.1
workers compensation	Insurance that covers an employer’s liability for injuries, disability or death to persons in their employment, without regard to fault, as prescribed by state or federal workers’ compensation laws and other statutes.	14.1
univariate	Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable.	14.1
missing data	Missing data occur when no data value is stored for a variable in an observation. Missing data can occur because of nonresponse: no information is provided for one or more items or for a whole unit or subject.	14.1
censored	Censored data have unknown values beyond a bound on either end of the number line or both. Here, the data is observed but the values (measurements) are not known completely.	14.1

(continued)

Term	Definition	Section
truncated	Truncation occurs when values beyond a boundary are either excluded when gathered or excluded when analyzed. An object can be detected only if its value is greater than some number.	14.1
stochastic process	Stochastic process is defined as a collection of random variables that is indexed by some mathematical set, meaning that each random variable of the stochastic process is uniquely associated with an element in the set.	14.1
deductibles	A deductible is a parameter specified in the contract. Typically, losses below the deductible are paid by the policyholder whereas losses in excess of the deductible are the insurer's responsibility (subject to policy limits and coninsurance).	14.1
rank based measures	Statistical dependence between the rankings of two variables	14.2
odds ratio	A statistic quantifying the strength of the association between two events, a and b, which is defined as the ratio of the odds of a in the presence of b and the odds of a in the absence of b	14.2
likelihood ratio test	A statistical test of the goodness-of-fit between two models	14.2
pearson correlation	A measure of the linear correlation between two variables	14.2
product-moment (pearson) correlation	Pearson correlation, a measure of the linear correlation between two variables	14.2
kendall tau	A statistic used to measure the ordinal association between two measured quantities	14.2
concordant	An observation pair (x,y) is said to be concordant if the observation with a larger value of x has also the larger value of y	14.2
discordant	An observation pair (x,y) is said to be discordant if the observation with a larger value of x has the smaller value of y	14.2
pearson chi-square statistic	A statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance	14.2

(continued)

Term	Definition	Section
tetrachoric correlation	A technique for estimating the correlation between two theorised normally distributed continuous latent variables, from two observed binary variables	14.2
polychoric correlation	A technique for estimating the correlation between two theorised normally distributed continuous latent variables, from two observed ordinal variables	14.2
polyserial correlation	The correlation between two continuous variables with a bivariate normal distribution, where one variable is observed directly, and the other is unobserved	14.2
biserial correlation	A correlation coefficient used when one variable is dichotomous	14.2
normal score	Transformed data which closely resemble a standard normal distribution	14.2
copula	A multivariate distribution function with uniform marginals	14.3
spearmans rho	A nonparametric measure of rank correlation	14.3
marginal distributions	The probability distribution of the variables contained in the subset of a collection of random variables	14.4
fat-tailed	A fat-tailed distribution is a probability distribution that exhibits a large skewness or kurtosis, relative to that of either a normal distribution or an exponential distribution	14.4
probability integral transformation	Any continuous variable can be mapped to a uniform random variable via its distribution function	14.4
elliptical copulas	The copulas of elliptical distributions	14.5
correlation matrix	A table showing correlation coefficients between variables	14.5
elliptical distributions	Any member of a broad family of probability distributions that generalize the multivariate normal distribution	14.5
tail dependency	A measure of their comovements in the tails of the distributions	14.5
frechet-hoeffding bounds	Bounds of multivariate distribution functions	14.5
blomqvists beta	A dependence measure based on the center of the distribution	14.7

(continued)

Term	Definition	Section
reinsurance	Insurance purchased by an insurer	1.1, 10.4
deductible	A deductible is a parameter specified in the contract. typically, losses below the deductible are paid by the policyholder whereas losses in excess of the deductible are the insurer's responsibility (subject to policy limits and coninsurance).	1.2, 5.3
coinsurance	Coinsurance is an arrangement whereby the insured and insurer share the covered losses. typically, a coinsurance parameter specified means that both parties receive a proportional share, e.g., 50%, of the loss.	1.2, 5.5
pure premium	Pure premium is the total severity divided by the number of claims. it does not include insurance company expenses, premium taxes, contingencies, nor an allowance for profits. also called loss costs. some definitions include allocated loss adjustment expenses (alae).	1.3, 7.1, 7.2
standard deviation	The square-root of variance	2.1, 3.1
variance	Second central moment of a random variable x, measuring the expected squared deviation of between the variable and its mean	2.1, 3.1
aggregate claims	The sum of all claims observed in a period of time	2.1, 5.1, 14.1
median	50th percentile of a definition, or middle value where half of the distribution lies below	3.1, 4.1
lorenz curve	A graph of the proportion of a population on the horizontal axis and a distribution function of interest on the vertical axis.	4.1, 7.6
law of total variance	A decomposition of the variance of a random variable into conditional components. specifically, for random variables x and y on the same probability space, var(x) = e[var(y|x)] + var[e(x|y)].	5.3, 9.4
tail value-at-risk	The expected value of a risk given that the risk exceeds a value-at-risk	6.2, 10.3
expected shortfall	The average value at risk	6.2, 10.3

(continued)

Term	Definition	Section
coefficient of variation	Standard deviation divided by the mean of a distribution, to measure variability in terms of units of the mean	6.3, 9.2
loss ratio	The sum of losses divided by the premium.	7.1, 7.2
homogeneous risks	Risks that have the same distribution, that is, the distributions are identical.	7.1, 7.2
heterogeneous	Heterogeneous risks have different distributions. often, we can attribute differences to varying exposures or risk factors.	7.1, 7.4
exposure	A type of rating variable that is so important that premiums and losses are often quoted on a "per exposure" basis. that is, premiums and losses are commonly standardized by exposure variables.	7.2, 7.4
loss	The amount of damages sustained by an individual or corporation, typically as the result of an insurable event.	7.5, 14.1
iid	Independent and identically distributed	
pdf	Probability density function	
aic	Akaike's information criterion	
bic	Bayesian information criterion	
pmf	Probability mass function	
mcmc	Markov Chain Monte Carlo	
cdf	Cumulative distribution function	
df	Degrees of freedom	
glm	Generalized linear model	
mle	Maximum likelihood estimate	
ols	Ordinary least squares	
pf	Probability function	
rv	Random variable	
reporting delay	The time that elapses between the occurrence of the insured event and the reporting of this event to the insurance company.	11.1
settlement delay	The time between reporting and settlement of a claim.	11.1
rbns	Reported, But is Not fully Settled	11.1
ibnr	Incurred in the past But is Not yet Reported. For such a claim the insured event took place, but the insurance company is not yet aware of the associated claim.	11.1
granular		11.1

(continued)

Term	Definition	Section
case estimates	The claims handlers expert estimate of the outstanding amount on a claim.	11.1
.csv	Comma separated value file	11.2
.txt	Text file	11.2
run-off triangle	Triangular display of loss reserve data. Accident or occurrence periods on one axis (often vertical) with development periods on the other (often horizontal). Also known as a development triangle.	11.2
development triangle	Triangular display of loss reserve data. Accident or occurrence periods on one axis (often vertical) with development periods on the other (often horizontal). Also known as a run-off triangle.	11.2
msep	Mean Squared Error of Prediction	
chain-ladder method	An algorithm for predicting incomplete losses to their ultimate cumulative value. The name refers to the chaining of a sequence of (year-to-year development) factors into a ladder of factors.	11.3
wls	weighted least squares	11.3
glm	Generalized linear model	
frequentist	Type of statistical inference based in frequentist probability, which treats probability in equivalent terms to frequency and draws conclusions from sample-data by means of emphasizing the frequency or proportion of findings in the data.	9
posterior distribution	The posterior distribution is the updated probability distribution of a parameter after incorporating prior information and observed data through Bayesian inference.	9.1
bayes' rule	A probability law that expresses conditional probability of the event a given the event b in terms of the conditional probability of the event b given the event a and the unconditional probability of a	9.1
informative	An informative prior, in statistics, is a prior probability distribution that is chosen deliberately to incorporate specific information or beliefs about a parameter before observing new data.	9.2

(continued)

Term	Definition	Section
weakly informative	A weakly informative prior is a prior probability distribution that introduces some general constraints or vague beliefs about a parameter, without heavily influencing the final inference.	9.2
noninformative	A noninformative prior is a prior probability distribution that intentionally avoids incorporating specific information or strong beliefs about a parameter.	9.2
improper	An improper prior is a prior probability distribution that does not integrate to a finite value over the entire parameter space.	9.2
conjugate distributions	Conjugate distributions are specific pairs of prior and likelihood functions that result in a posterior distribution within the same family of probability distributions as the prior.	9.3
hyperparameters	Hyperparameters are parameters that define the distribution of a prior distribution	9.3
gibbs sampler	The Gibbs sampler is an iterative algorithm in statistics used for simulating samples from complex probability distributions. It's particularly useful in Bayesian analysis for drawing samples from multivariate distributions by updating one variable at a time while keeping others fixed.	9.4
metropolis–hastings algorithm	The Metropolis–Hastings algorithm is a method to generate samples from complex distributions by proposing new samples and deciding whether to accept them, making it valuable for Bayesian analysis and complex modeling.	9.4
precision	Precision is the inverse of variance and is often used to quantify the amount of uncertainty or variability in a prior or posterior distribution.	9.3

Bibliography

Aalen, Odd (1978). “Nonparametric inference for a family of counting processes,” *The Annals of Statistics*, Vol. 6, pp. 701–726.

Abadir, Karim and Jan Magnus (2002). “Notation in econometrics: a proposal for a standard,” *The Econometrics Journal*, Vol. 5, pp. 76–90.

Abbott, Dean (2014). *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*, Hoboken, NJ. Wiley.

Abdullah, Mohammad F. and Kamsuriah Ahmad (2013). “The mapping process of unstructured data to structured data,” in *2013 International Conference on Research and Innovation in Information Systems (ICRIIS)*, pp. 151–155.

Actuarial Community (2025). *Loss Data Analytics*, URL: https://openacttexts.github.io/Loss-Data-Analytics/index.html.

Actuarial Standards Board (2018). “Actuarial Standards of Practice,” American Academy of Actuaries, URL: http://www.actuarialstandardsboard.org/standards-of-practice/, [Retrieved on Oct 3, 2018].

Aggarwal, Charu C. (2015). *Data Mining: The Textbook*, New York, NY. Springer.

Agresti, Alan (1996). *An Introduction to Categorical Data Analysis*. Wiley New York.

Albrecher, Hansjörg, Jan Beirlant, and Jozef L Teugels (2017). *Reinsurance: Actuarial and Statistical Aspects*. John Wiley & Sons.

Antonio, K. and R. Plat (2014). “Micro–level stochastic loss reserving for general insurance,” *Scandinavian Actuarial Journal*, Vol. 7, pp. 649–669.

Bahnemann, David (2015). *Distributions for Actuaries*, No. 2, URL: https://www.casact.org/pubs/monographs/papers/02-Bahnemann.pdf.

Bailey, Robert A. and J. Simon LeRoy (1960). “Two studies in automobile ratemaking,” *Proceedings of the Casualty Actuarial Society Casualty Actuarial Society*, Vol. XLVII.

Bauer, Daniel, Richard D. Phillips, and George H. Zanjani (2013). “Financial pricing of insurance,” in *Handbook of Insurance*. Springer, pp. 627–645.

Bégin, Jean-François (2019). “Economic scenario generator and parameter uncertainty: A Bayesian approach,” *ASTIN Bulletin*, Vol. 49, pp. 335–372.

——— (2021). “On complex economic scenario generators: Is less more?” *ASTIN Bulletin*, Vol. 51, pp. 779–812.

——— (2023). “Ensemble economic scenario generators: Unity makes strength,” *North American Actuarial Journal*, Vol. 27, pp. 444–471.

Bermúdez, Lluís and Dimitris Karlis (2011). “Bayesian multivariate Poisson models for insurance ratemaking,” *Insurance: Mathematics and Economics*, Vol. 48, pp. 226–236.

Bernardo, José M and Adrian FM Smith (2009). *Bayesian Theory.* John Wiley & Sons: New York, NY, United States of America.

Bignozzi, Valeria and Andreas Tsanakas (2016). “Parameter uncertainty and residual estimation risk,” *Journal of Risk and Insurance*, Vol. 83, pp. 949–978.

Billingsley, Patrick (2008). *Probability and measure.* John Wiley & Sons.

Bishop, Christopher M. (2007). *Pattern Recognition and Machine Learning*, New York, NY. Springer.

Blomqvist, Nils (1950). “On a measure of dependence between two random variables,” *The Annals of Mathematical Statistic*, pp. 593–600.

Boehm, C, J Engelfriet, M Helbig, A IM Kool, P Leepin, E Neuburger, and AD Wilkie (1975). “Thoughts on the harmonization of some proposals for a new International actuarial notation,” *Blätter der DGVFM*, Vol. 12, pp. 99–129.

Bowers, Newton L., Hans U. Gerber, James C. Hickman, Donald A. Jones, and Cecil J. Nesbitt (1986). *Actuarial Mathematics.* Society of Actuaries Itasca, Ill.

Box, George E. P. (1980). “Sampling and Bayes’ inference in scientific modelling and robustness,” *Journal of the Royal Statistical Society. Series A (General)*, pp. 383–430.

Breiman, Leo (2001). “Statistical modeling: The two cultures,” *Statistical Science*, Vol. 16, pp. 199–231.

Breiman, Leo, Jerome Friedman, Charles J. Stone, and R.A. Olshen (1984). *Classification and Regression Trees*, Raton Boca, FL. Chapman and Hall/CRC.

Bühlmann, Hans (1967). “The complement of credibility,” pp. 199–207.

Bühlmann, Hans (1985). "Premium calculation from top down," *ASTIN Bulletin: The Journal of the IAA*, Vol. 15, pp. 89–101.

Bühlmann, Hans, Massimo De Felice, Alois Gisler, Franco Moriconi, and Mario V Wüthrich (2009). "Recursive credibility formula for chain ladder factors and the claims development result," *ASTIN Bulletin: The Journal of the IAA*, Vol. 39, pp. 275–306.

Bühlmann, Hans and Alois Gisler (2005). *A Course in Credibility Theory and its Applications*. ACTEX Publications.

Buttrey, Samuel E. and Lyn R. Whitaker (2017). *A Data Scientist's Guide to Acquiring, Cleaning, and Managing Data in R*, Hoboken, NJ. Wiley.

Cairns, Andrew JG (2000). "A discussion of parameter and model uncertainty in insurance," *Insurance: Mathematics and Economics*, Vol. 27, pp. 313–330.

Cairns, Andrew JG, David Blake, and Kevin Dowd (2006). "A two-factor model for stochastic mortality with parameter uncertainty: Theory and Calibration," *Journal of Risk and Insurance*, Vol. 73, pp. 687–718.

Charpentier, Arthur (2014). *Computational Actuarial Science with R*. CRC press.

Chen, Min, Shiwen Mao, Yin Zhang, and Victor CM Leung (2014). *Big Data: Related Technologies, Challenges and Future Prospects*, New York, NY. Springer.

Cheung, Eric CK, Weihong Ni, Rosy Oh, and Jae-Kyung Woo (2021). "Bayesian credibility under a bivariate prior on the frequency and the severity of claims," *Insurance: Mathematics and Economics*, Vol. 100, pp. 274–295.

Clark, David R (1996). *Basics of reinsurance pricing*, pp.41–43, URL: https://www.soa.org/files/edu/edu-2014-exam-at-study-note-basics-rein.pdf.

Cowles, Mary Kathryn (2013). *Applied Bayesian Statistics: With R and OpenBUGS Examples*. Springer Science & Business Media: New York, NY, United States of America.

Cummins, J. David and Richard A. Derrig (2012). *Managing the Insolvency Risk of Insurance Companies: Proceedings of the Second International Conference on Insurance Solvency*, Vol. 12. Springer Science & Business Media.

Dabrowska, Dorota M. (1988). "Kaplan-meier estimate on the plane," *The Annals of Statistics*, pp. 1475–1489.

Daroczi, Gergely (2015). *Mastering Data Analysis with R*, Birmingham, UK. Packt Publishing.

De Jong, Piet and Gillian Z. Heller (2008). *Generalized Linear Models for Insurance Data*. Cambridge University Press, Cambridge.

Denuit, Michel, Jan Dhaene, Marc Goovaerts, and Rob Kaas (2006). *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. John Wiley & Sons.

Denuit, Michel, Xavier Maréchal, Sandra Pitrebois, and Jean-François Walhin (2007). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. John Wiley & Sons, Chichester.

Derrig, Richard A, Krzysztof M Ostaszewski, and Grzegorz A Rempala (2001). "Applications of resampling methods in actuarial practice," in *Proceedings of the Casualty Actuarial Society*, Vol. 87, pp. 322–364, Casualty Actuarial Society.

Dickson, David C. M., Mary Hardy, and Howard R. Waters (2013). *Actuarial Mathematics for Life Contingent Risks*. Cambridge University Press.

Dionne, Georges and Charles Vanasse (1989). "A generalization of automobile insurance rating models: the negative binomial distribution with a regression component," *ASTIN Bulletin*, Vol. 19(2), pp. 199–212.

Dobson, Annette J and Adrian Barnett (2008). *An Introduction to Generalized Linear Models*. CRC press.

Earnix (2013). "2013 Insurance Predictive Modeling Survey," Earnix and Insurance Services Office, Inc. URL: https://www.verisk.com/archived/2013/majority-of-north-american-insurance-companies-use-predictive-analytics-to-enhance-business-performance-new-earnix-iso-survey-shows/, [Retrieved on July 23, 2020].

Efron, Bradley (1979). "Bootstrap methods: Another look at the bootstrap," *The Annals of Statistics*, Vol. 7, pp. 1–26.

——— (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM.

——— (1992). *Bootstrap Methods: Another Look at the Jackknife*, pp. 569–593. Springer New York, URL: https://doi.org/10.1007/978-1-4612-4380-9_41, DOI: http://dx.doi.org/10.1007/978-1-4612-4380-9_41.

England, P. and R. Verrall (2002). "Stochastic claims reserving in general insurance," *British Actuarial Journal*, Vol. 8/3, pp. 443–518.

Faraway, Julian J (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Vol. 124. CRC press.

Fechner, G. T (1897). “Kollektivmasslehre,” *Wilhelm Englemann, Leipzig.*

Fellingham, Gilbert W, Athanasios Kottas, and Brian M Hartman (2015). “Bayesian nonparametric predictive modeling of group health claims,” *Insurance: Mathematics and Economics*, Vol. 60, pp. 1–10.

Finger, Robert J. (2006). “Risk classification,”, pp. 231–276.

Forte, Rui Miguel (2015). *Mastering Predictive Analytics with R*, Birmingham, UK. Packt Publishing.

Frank, Maurice J (1979). “On the simultaneous associativity of F(x, y) and x+y-F(x, y),” *Aequationes mathematicae*, Vol. 19, pp. 194–226.

Frees, Edward W (2009). *Regression Modeling with Actuarial and Financial Applications*. Cambridge University Press, URL: https://doi.org/10.1017/CBO9780511814372.

——— (2014). “Frequency and severity models,” in Edward W Frees, Glenn Meyers, and Richard Derrig eds. *Predictive Modeling Applications in Actuarial Science*, Vol. 1, pp. 138–164. Cambridge University Press Cambridge, URL: https://doi.org/10.1017/CBO9781139342674.

——— (2015). “Analytics of insurance markets,” *Annual Review of Financial Economics*, Vol. 7, pp. 253–277.

Frees, Edward W, Catalina Bolancé, Montserrat Guillen, and Emiliano A Valdez (2021). “Dependence modeling of multivariate longitudinal hybrid insurance data with dropout,” *Expert Systems with Applications*, Vol. 185, p. 115552.

Frees, Edward W and Adam Butt (2022). “ANU Insurable Risks,” URL: https://doi.org/10.25911/0SE7-N746.

Frees, Edward W and Lisa Gao (2019). “Predictive analytics and medical malpractice,” *North American Actuarial Journal*, pp. 1–17, URL: https://doi.org/10.1080/10920277.2019.1634597, DOI: http://dx.doi.org/10.1080/10920277.2019.1634597.

Frees, Edward W and Fei Huang (2021). “The discriminating (pricing) actuary,” *North American Actuarial Journal*, pp. 1–23, URL: https://www.tandfonline.com/doi/pdf/10.1080/10920277.2021.1951296.

——— (2023). “The discriminating (pricing) actuary,” *North American Actuarial Journal*, Vol. 27:1, pp. 2–24, URL: https://www.tandfonline.com/doi/pdf/10.1080/10920277.2021.1951296.

Frees, Edward W, Gee Lee, and Lu Yang (2016a). “Multivariate frequency-severity regression models in insurance," *Risks*, Vol. 4, p. 4, URL: https://doi.org/10.3390/risks4010004.

——— (2016b). "Multivariate frequency-severity regression models under insurance," *Risks*, Vol. 4(1), p. 4.

Frees, Edward W and Emiliano A Valdez (1998). "Understanding relationships using copulas," *North American Actuarial Journal*, Vol. 2, pp. 1–25.

Frees, Edward W and Emiliano A. Valdez (2008). "Hierarchical insurance claims modeling," *Journal of the American Statistical Association*, Vol. 103, pp. 1457–1469.

Friedland, Jacqueline (2013). *Fundamentals of General Insurance Actuarial Analysis*. Society of Actuaries.

Gan, Guojun (2011). *Data Clustering in C++: An Object-Oriented Approach*, Data Mining and Knowledge Discovery Series, Boca Raton, FL, USA. Chapman & Hall/CRC Press, DOI: http://dx.doi.org/10.1201/b10814.

Gan, Guojun, Chaoqun Ma, and Jianhong Wu (2007). *Data Clustering: Theory, Algorithms, and Applications*, Philadelphia, PA. SIAM Press, DOI: http://dx.doi.org/10.1137/1.9780898718348.

Garrido, Jose, Christian Genest, and Juliana Schulz (2016). "Generalized linear models for dependent frequency and severity of insurance claims," *Insurance: Mathematics and Economics*, Vol. 70, pp. 205–215.

Gelfand, Alan E and Adrian FM Smith (1990). "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, Vol. 85, pp. 398–409.

Gelman, Andrew and Donald B Rubin (1992). "Inference from iterative simulation using multiple sequences," *Statistical Science*, pp. 457–472.

Genest, Christian and Josh Mackay (1986). "The joy of copulas: Bivariate distributions with uniform marginals," *The American Statistician*, Vol. 40, pp. 280–283.

Genest, Christian and Johanna Nešlohva (2007). "A primer on copulas for count data," *Journal of the Royal Statistical Society*, pp. 475–515.

Gerber, Hans U (1979). *An Introduction to Mathematical Risk Theory, vol. 8 of SS Heubner Foundation Monograph Series*. University of Pennsylvania Wharton School SS Huebner Foundation for Insurance Education.

Gesmann, Markus, Daniel Murphy, Yanwei Zhang, Alessandro Carrato, Mario Wuthrich, Fabio Concina, and Eric Dal Moro (2019). *ChainLadder: Statistical*

Methods and Models for Claims Reserving in General Insurance, URL: https://CRAN.R-project.org/package=ChainLadder, R package version 0.2.10.

Gisler, Alois (2006). "The estimation error in the chain-ladder reserving method: a Bayesian approach," *ASTIN Bulletin: The Journal of the IAA*, Vol. 36, pp. 554–565.

Gisler, Alois and Mario V Wüthrich (2008). "Credibility for the chain ladder reserving method," *ASTIN Bulletin: The Journal of the IAA*, Vol. 38, pp. 565–600.

Goldberger, Arthur S. (1972). "Structural equation methods in the social sciences," *Econometrica: Journal of the Econometric Society*, pp. 979–1001.

Good, I. J. (1983). "The Philosophy of Exploratory Data Analysis," *Philosophy of Science*, Vol. 50, pp. 283–295.

Gorman, Mark and Stephen Swenson (2013). "Building believers: How to expand the use of predictive analytics in claims," SAS, URL: https://www.the-digital-insurer.com/wp-content/uploads/2014/10/265-wp-59831.pdf, [Retrieved on July 23, 2020].

Greenwood, Major (1926). "The errors of sampling of the survivorship tables," in *Reports on Public Health and Statistical Subjects*, Vol. 33. London: Her Majesty's Stationary Office.

Halperin, Max, Herman O Hartley, and Paul G Hoel (1965). "Recommended standards for statistical symbols and notation: Copss Committee on Symbols and Notation," *The American Statistician*, Vol. 19, pp. 12–14.

Hardy, Mary R. (2006). *An Introduction to Risk Measures for Actuarial Applications*. Society of Actuaries, URL: https://www.soa.org/globalassets/assets/files/edu/c-25-07.pdf, [Retrieved on August 6, 2020].

Hartman, Brian M and Chris Groendyke (2013). "Model Selection and Averaging in Financial Risk Management," *North American Actuarial Journal*, Vol. 17, pp. 216–228.

Hartman, Brian M and Matthew J Heaton (2011). "Accounting for Regime and Parameter Uncertainty in Regime-Switching Models," *Insurance: Mathematics and Economics*, Vol. 49, pp. 429–437.

Hashem, Ibrahim Abaker Targio, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan (2015). "The rise of "big data" on cloud computing: Review and open research issues," *Information Systems*, Vol. 47, pp. 98 – 115.

Hastie, Trevor, Robert Tibshirani, and Jerome H Friedman (2009). *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, Vol. 2. Springer.

Hastings, WK (1970). “Monte Carlo sampling methods using Markov chains and their applications,” *Biometrika*, Vol. 57, pp. 97–109.

Haueter, Niels Viggo (2017). “A History of UK Insurance,”Technical report.

Heckman, Philip E and Glenn G Meyers (1983). “The calculation of aggregate loss distributions from claim severity and claim count distributions,” in *Proceedings of the Casualty Actuarial Society*, Vol. 70, pp. 49–66.

Hettmansperger, T. P. (1984). *Statistical Inference Based on Ranks*. Wiley.

Hoerl, Arthur E and Robert W. Kennard (1970). “Ridge regression: Biased estimation for nonorthogonal problems,” *Technometrics*, Vol. 12, pp. 55–67.

Hofert, Marius, Ivan Kojadinovic, Martin Mächler, and Jun Yan (2018). *Elements of Copula Modeling with R*. Springer.

Hogg, Robert V, Elliot A Tanis, and Dale L Zimmerman (2015). *Probability and Statistical Inference, 9th Edition*. Pearson, New York.

Hougaard, P (2000). *Analysis of Multivariate Survival Data*. Springer New York.

Hox, Joop J. and Hennie R. Boeije (2005). “Data collection, primary versus secondary,” in *Encyclopedia of social measurement*. Elsevier, pp. 593 – 599.

Huang, Yifan and Shengwang Meng (2020). “A Bayesian nonparametric model and its application in insurance loss prediction,” *Insurance: Mathematics and Economics*, Vol. 93, pp. 84–94.

Inmon, W.H. and Dan Linstedt (2014). *Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault*, Cambridge, MA. Morgan Kaufmann.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013). *An Introduction to Statistical Learning*, Vol. 112. Springer.

Janert, Philipp K. (2010). *Data Analysis with Open Source Tools*, Sebastopol, CA. O’Reilly Media.

Joe, Harry (2014). *Dependence Modeling with Copulas*. CRC Press.

Kaas, Rob, Marc Goovaerts, Jan Dhaene, and Michel Denuit (2008). *Modern Actuarial Risk Theory: using R*, Vol. 128. Springer Science & Business Media.

Kaplan, Edward L. and Paul Meier (1958). “Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, Vol. 53, pp. 457–481.

Kendall, M. G (1945). "The treatment of ties in ranking problems," *Biometrika*, Vol. 33(3), pp. 239–251.

Kendall, Maurice G (1938). "A new measure of rank correlation," *Biometrika*, pp. 81–93.

Klugman, Stuart A., Harry H. Panjer, and Gordon E. Willmot (2012). *Loss Models: From Data to Decisions*. John Wiley & Sons.

Krämer, Nicole, Eike C Brechmann, Daniel Silvestrini, and Claudia Czado (2013). "Total loss estimation using copula-based regression models," *Insurance: Mathematics and Economics*, Vol. 53, pp. 829–839.

Kreer, Markus, Ayşe Kızılersü, Anthony W Thomas, and Alfredo D Egídio dos Reis (2015). "Goodness-of-fit tests and applications for left-truncated Weibull distributions to non-life insurance," *European Actuarial Journal*, Vol. 5, pp. 139–163.

Kremer, Erhard (1982). "IBNR-claims and the two-way model of ANOVA," *Scandinavian Actuarial Journal*, Vol. 1982, pp. 47–55.

——— (1984). "A class of autoregressive models for predicting the final claims amount," *Insurance: Mathematics and Economics*, Vol. 3, pp. 111–119.

Lee Rodgers, J and W. A Nicewander (1998). "Thirteen ways to look at the correlation coeffeicient," *The American Statistician*, Vol. 42, pp. 59–66.

Lemaire, Jean (1998). "Bonus-malus systems: the European and Asian approach to merit rating," *North American Actuarial Journal*, Vol. 2(1), pp. 26–38.

Lemaire, Jean and Hongmin Zi (1994). "A comparative analysis of 30 bonus-malus systems," *ASTIN Bulletin*, Vol. 24(2), pp. 287–309.

Levin, Bruce, James Reeds et al. (1977). "Compound multinomial likelihood functions are unimodal: Proof of a conjecture of IJ Good," *The Annals of Statistics*, Vol. 5, pp. 79–87.

Mack, Thomas (1991). "A simple parametric model for rating automobile insurance or estimating IBNR claims reserves," *ASTIN Bulletin: The Journal of the IAA*, Vol. 21, pp. 93–109.

——— (1993). "Distribution-free calculation of the standard error of chain ladder reserve estimates," *ASTIN Bulletin: The Journal of the IAA*, Vol. 23, pp. 213–225.

Mack, Thomas and Gary Venter (2000). "A comparison of stochastic models that reproduce chain ladder reserve estimates," *Insurance: mathematics and economics*, Vol. 26, pp. 101–107.

McCullagh, Peter and John A. Nelder (1989). *Generalized Linear Models, Second Edition*, Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, London.

McDonald, James B (1984). "Some generalized functions for the size distribution of income," *Econometrica: journal of the Econometric Society*, pp. 647–663.

McDonald, James B and Yexiao J Xu (1995). "A generalization of the beta distribution with applications," *Journal of Econometrics*, Vol. 66, pp. 133–152.

Metropolis, Nicholas, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller (1953). "Equation of state calculations by fast computing machines," *Journal of Chemical Physics*, Vol. 21, pp. 1087–1092.

Meyers, Glenn (1994). "Quantifying the Uncertainty in Claim Severity Estimates for an Excess Layer When Using the Single Parameter Pareto," in *Proceedings of the Casualty Actuarial Society*, Vol. 81, pp. 91–122.

Meyers, Glenn and Nathaniel Schenker (1983). "Parameter uncertainty in the collective risk model," *PCAS LXX*, Vol. 111, p. 15.

Mildenhall, Stephen J and John A Major (2022). *Pricing Insurance Risk: Theory and Practice*. John Wiley & Sons: New York, NY, United States of America.

Miles, Matthew, Michael Hberman, and Johnny Sdana (2014). *Qualitative Data Analysis: A Methods Sourcebook*, Thousand Oaks, CA. Sage, 3rd edition.

Mirkin, Boris (2011). *Core Concepts in Data Analysis: Summarization, Correlation and Visualization*, London, UK. Springer.

Mitchell, Tom M. (1997). *Machine Learning*. McGraw-Hill.

NAIC Glossary (2018). "Glossary of Insurance Terms," National Association of Insurance Commissioners, URL: https://www.naic.org/consumer_glossary.htm, [Retrieved on Sept 11, 2018].

Nelson, Roger B. (1997). *An Introduction to Copulas*. Lecture Notes in Statistics 139.

Niehaus, Gregory and Scott Harrington (2003). *Risk Management and Insurance*, New York. McGraw Hill.

Norberg, Ragnar (1976). “A credibility theory for automobile bonus system,” *Scandinavian Actuarial Journal*, Vol. 2, pp. 92–107.

O’Donnell, Terence (1936). *History of Life Insurance in its Formative Years.* American Conservation Company: Chicago, IL, United States of America.

Oh, Rosy, Joseph H. T. Kim, and Jae Youn Ahn (2020a). “Designing a Bonus-Malus system reflecting the claim size under the dependent frequency-severity model.”

Oh, Rosy, Kyung Suk Lee, Sojung C. Park, and Jae Youn Ahn (2020b). “Double-counting problem of the bonus-malus system,” *Insurance: Mathematics and Economics*, Vol. 93, pp. 141–155.

Oh, Rosy, Peng Shi, and Jae Youn Ahn (2020c). “Bonus-malus premiums under the dependent frequency-severity modelling,” *Scandinavian Actuarial Journal*, Vol. 2020(3), pp. 172–195.

Ohlsson, Esbjörn and Björn Johansson (2010). *Non-life Insurance Pricing with Generalized Linear Models*, Vol. 21. Springer.

O’Leary, D. E. (2013). “Artificial Intelligence and Big Data,” *IEEE Intelligent Systems*, Vol. 28, pp. 96–99.

Olkin, Ingram, A John Petkau, and James V Zidek (1981). “A comparison of n estimators for the binomial distribution,” *Journal of the American Statistical Association*, Vol. 76, pp. 637–642.

Parsa, Rahul A and Stuart A Klugman (2011). “Copula regression,” *Variance: Advancing and Science of Risk*, Vol. 5, pp. 45–54.

Picard, Richard R. and Kenneth N. Berk (1990). “Data splitting,” *The American Statistician*, Vol. 44, pp. 140–147.

Pitrebois, Sandra, Michel Denuit, and Jean-François Walhin (2003). “Setting a bonus-malus scale in the presence of other rating factors: Taylor’s work revisited,” *ASTIN Bulletin*, Vol. 33(2), pp. 419–436.

Pries, Kim H. and Robert Dunnigan (2015). *Big Data Analytics: A Practical Guide for Managers*, Boca Raton, FL. CRC Press.

Quenouille, Maurice H (1949). “Approximate tests of correlation in time-series,” *Journal of the Royal Statistical Society. Series B*, Vol. 11, pp. 68–84.

Renshaw, A. and R. Verrall (1998). “A stochastic model underlying the chain-ladder technique,” *British Actuarial Journal*, Vol. 4/4, pp. 903–923.

Renshaw, Arthur E (1989). “Chain ladder and interactive modelling.(Claims reserving and GLIM)," *Journal of the Institute of Actuaries*, Vol. 116, pp. 559–587.

Robert, Christian P and George Casella (1999). *Monte Carlo Statistical Methods*. Springer: New York, NY, United States of America.

Ruppert, David, Matt P Wand, and Raymond J Carroll (2003). *Semiparametric Regression*, No. 12. Cambridge University Press.

Samuel, A. L. (1959). "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, Vol. 3, pp. 210–229.

Schmidt, R (2005). "Tail dependence," in Weron R Cizek P, Häardle W ed. *Statistical tools in finance and insurance.*, New York. Springer.

Schweizer, Berthold, Edward F Wolff et al. (1981). "On nonparametric measures of dependence for random variables," *The Annals of Statistics*, Vol. 9, pp. 879–885.

Shmueli, Galit (2010). "To explain or to predict?" *Statistical Science*, Vol. 25, pp. 289–310.

Sklar, M (1959). "Fonctions de repartition a N dimensions et leurs marges," *Publ. inst. statist. univ. Paris*, Vol. 8, pp. 229–231.

Snee, Ronald D. (1977). "Validation of regression models: methods and examples," *Technometrics*, Vol. 19, pp. 415–428.

Spearman, C (1904). "The proof and measurement of association between two things," *The American Journal of Psychology*, Vol. 15, pp. 72–101.

Stigler, Stephen M (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press.

Tan, Chong It (2016). "Optimal design of a bonus-malus system: linear relativities revisited," *Annals of Actuarial Science*, Vol. 10(1), pp. 52–64.

Tan, Chong It, Jackie Li, Johnny Siu-Hang Li, and Uditha Balasooriya (2015). "Optimal relativities and transition rules of a bonus-malus system," *Insurance: Mathematics and Economics*, Vol. 61, pp. 255–263.

Taylor, G. (2000). *Loss Reserving: An Actuarial Perspective*. Kluwer Academic Publishers.

Taylor, Gregory Clive (1986). *Claims Reserving in Non-life Insurance*. North Holland.

Tevet, Dan (2016). "Applying generalized linear models to insurance data," *Predictive Modeling Applications in Actuarial Science: Volume 2, Case Studies in Insurance*, p. 39.

The Organization for Economic Cooperation and Development (OECD) (2021). "OECD Insurance Statistics 2021," OECD iLibrary, URL: https://read.oecd-ilibrary.org/finance-and-investment/oecd-insurance-statistics-2021_841fa619-en#page1, [Retrieved on 1 August, 2022].

Tse, Yiu-Kuen (2009). *Nonlife Actuarial Models: Theory, Methods and Evaluation*. Cambridge University Press.

Tukey, John W. (1962). "The Future of Data Analysis," *The Annals of Mathematical Statistics*, Vol. 33, pp. 1–67.

de Valpine, Perry, Daniel Turek, Christopher J Paciorek, Clifford Anderson-Bergman, Duncan Temple Lang, and Rastislav Bodik (2017). "Programming with models: Writing statistical algorithms for general model structures with nimble," *Journal of Computational and Graphical Statistics*, Vol. 26, pp. 403–413.

Vats, Dootika, James M Flegal, and Galin L Jones (2019). "Multivariate output analysis for Markov chain Monte Carlo," *Biometrika*, Vol. 106, pp. 321–337.

Venter, Gary (1983). "Transformed beta and gamma distributions and aggregate losses," in *Proceedings of the Casualty Actuarial Society*, Vol. 70, pp. 289–308.

Venter, Gary G. (2002). "Tails of copulas," in *Proceedings of the Casualty Actuarial Society*, Vol. 89, pp. 68–113.

Venter, Gary G (2006). "Discussion of the mean square error of prediction in the chain ladder reserving method," *ASTIN Bulletin: The Journal of the IAA*, Vol. 36, pp. 566–571.

Wang, Ruodu and Ričardas Zitikis (2022). "An axiomatic foundation for the expected shortfall," *Management Science*, Vol. 67, pp. 1413–1429, URL: https://doi.org/10.1287/mnsc.2020.3617.

Werner, Geoff and Claudine Modlin (2016). *Basic Ratemaking, Fifth Edition*. Casualty Actuarial Society, URL: https://www.casact.org/library/studynotes/werner_modlin_ratemaking.pdf, [Retrieved on April 1, 2019].

Wolny-Dominiak, Alicja and Michal Trzesiok (2014). "Package 'insuranceData',"Technical report, The Comprehensive R Archive Network.

Wüthrich, Mario V. and Michael Merz (2008). *Stochastic claims reserving methods in insurance*, Vol. 435 of Wiley Finance. John Wiley & Sons.

——— (2015). *Stochastic Claims Reserving Manual: Advances in Dynamic Modeling*. SSRN.

Young, Virginia R (2014). “Premium principles,” *Wiley StatsRef: Statistics Reference Online*.