Consumer Reports Statistician Interview Test

Ivan E. Perez

May 11, 2021

1 Problem A: Washing Machine Consistency

Redacted

1.1 Experimental Design and Considerations

Redacted

A baseline would need to be established for both clean and soiled clothing using reflective spectrophotometer.[1] The reflectance for new white clothes is defined to be 1 and the reflectance of soiled clothes is 0.[2] A standard load of white clothes that have been uniformly soiled with water soluble black dye (e.g., Nigrosin WS aka Acid black 2) would be washed in a washing machine with a constant mass of detergent for a complete cycle. The recovered clothes would have their reflectance measured. The reflectance of the recovered clothes is the response variable, $X = \{n \in \mathbb{R} | 0 \leq n \leq 1\}$.

To test consistency between washing machines for a set of detergents we begin by labeling the set of washing machines A, B, C, D, E. The detergents are labeled by their brand, Tide, Seventh, Candado, and Shout. We seek to determine whether there is statistically significant source of variation from the random selection of Washing Machines, separate from the variation arising from the selection of detergents. The results of a mock trial are shown in Table 1.

1.2 Describing the Statistical Model

Redacted

The *statistical model*, is an additive model that takes two categorical factors Washing Machine, α, and Detergent, β, Tide, Seventh, Candado, and Shout. The response variable is the Reflectance, $X_{ij}, i = A, \ldots, E, j = \text{Tide}, \ldots, \text{Shout}$, is composed of the true average response for all levels, μ, with factors α and β, plus an unexplained error is ϵ_{ij}. We take $\mathbb{E}[\epsilon_{ij}] = 0, \text{Var}[\epsilon_{ij}] = \sigma^2, \sigma > 0$. We assume that there is no interaction between predictors, and therefore the only

[1] 11 page summary on reflective spectrophotometry, http://www-odp.tamu.edu/publications/tnotes/tn26/CHAP7.PDF

[2] Reflectance is a measurement of the intensity of a reflected color.

Table 1: Pilot Study of Reflectance of washed clothes with four Detergents in five randomly selected Washing Machines

Detergent	Machine	Reflectance	Mach. A Broken Reflectance
Tide	A	0.75	**0.00**
Tide	B	0.75	0.75
Tide	C	0.74	0.74
Seventh	A	0.22	**0.00**
Seventh	C	0.41	0.41
Seventh	D	0.33	0.33
Candado	B	0.10	0.10
Candado	D	0.05	0.05
Candado	E	0.20	0.20
Shout	A	0.93	**0.00**
Shout	E	0.64	0.64
Shout	B	0.77	0.77

sources of variation are assumed to be from the random selection of Washing Machines, and the Detergent.

$$X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \tag{1}$$

The *statistical test* is a Two-Way ANOVA. The Two-Way ANOVA tests whether there is a statistically significant difference in the population means for each grouping (i.e., by Washing Machine, Detergent) versus the total sample population. From the toy data above, the resulting Two-Way ANOVA is shown in Table 2. The Hypotheses are:

H_{0A} : the means of population grouped by Washing Machine are equal.

H_{1A} : the means of the population grouped by Washing Machine are not equal.

H_{0B} : the means of the population grouped by Detergent are equal.

H_{1B} : the means of the populations grouped by Detergent are not equal.

From the ANOVA table we can conclude that at the $\alpha = 0.05$ significance level, the population means grouped by Detergent are statistically different from one another. Therefore H_{0B} can be rejected. However, the means grouped by Washing Machine are not statistically significant different. Therefore we fail to reject H_{0A}.

Example 1.1 Broken Washing Machine: Suppose that Washing Machine A, despite being tested prior to the pilot study was defective and did not wash clothing regardless of what detergent was used. The observed reflectance would be 0.00 for all trials that used Washing Machine A. The results is shown in the third column of Table 1. The associated Two-Way ANOVA is shown in Table 3.

Table 2: Two-Way ANOVA of the effect of Detergent and Washing Machine Selection of Reflectance

Source of Variation	SS	df	MSS	F	Pr(>F)
Machine	0.01217	4	0.00304	0.203	0.92418
Detergents	0.69356	3	0.23119	15.422	0.01155
Residuals	0.05996	4	0.01499		

Table 3: Two-Way ANOVA of the effect of Detergent and Washing Machine Selection on Reflectance as a result of Washing Machine A being broken.

Source of Variation	SS	df	MSS	F	Pr(>F)
Machine	0.76751	4	0.19188	15.829	0.01840
Detergents	0.43256	3	0.14419	11.895	0.01017
Residuals	0.04849	4	0.012123		

From the ANOVA we see that now both tests show that population means grouped by both Washing Machine, and Detergent are statistically different at the $\alpha = 0.05$ significance level.

1.3 Software Recommendation

Redacted

I used R, with the package "Car" for the Two-Way ANOVA. Given the size of the study, this can be also done in MS Excel. Why? because it was quicker in R than in Excel.

1.4 Discarding Data Ex-post

Redacted

Recalling my Defective Machine example, suppose we simply omit the data for Machine A. The data is shown in Table X. We can try doing a Two-Way ANOVA with this subset of data.

Table 4: Experimental Trial for Clothing with Broken Machine A

Detergent	Machine	Reflectance, y
Tide	B	0.75
Tide	C	0.74
Seventh	C	0.41
Seventh	D	0.33
Candado	B	0.10
Candado	D	0.05
Candado	E	0.20
Shout	E	0.64
Shout	B	0.77

Table 5: Two-Way ANOVA of Effect of Washing Machine Selection and Detergent Selection on Reflectance, omitting data from Washing Machine A

Source of Variation	SS	df	MSS	F	Pr(>F)
Machine	0.01009	3	0.00336	0.5065	0.71631
Detergents	0.46777	3	0.15592	23.4856	0.04112
Residuals	0.01328	2	0.00664		

We see here that while for this data set the F statistic is still statistically significant for the source of variation between detergents. We see that with a reduced data set, the degrees of freedom are smaller, leading to a higher threshold for the F statistic, this effect is reflected in the p-values despite having larger but similar computed F statistics as in Table 2.

2 Problem B: Analysis of Verbatims

Redacted

2.1 Data Analysis for Consumers and Product Testers

Redacted

2.1.1 Parsing Data

We are given an $n \times 2$ array, where the first element of each row contains a manufacturer, Bosch, Kenmore, Electrolux, or GE. The second element of each row, contains a short phrase in the form of a string, reflecting their experience with that product.

We assume that the manufacturer has been spelled correctly such that there are only four possible categorical values that the first element of each row can take. Our objective is to identify common themes in each brand. To clean the data, we can:

1. Convert uppercase letters to lowercase using a method like "lower()".
2. Convert the string to a list where each word is an item in the list.
3. Filter out "Stop Words"[3]
4. Group each list of remaining words by brand yielding 4 data frames.
5. Iterate through each filtered verbatim list and counting the number of appearances of each word.
6. Display the frequency of each word by brand, as shown in Figure 1.

[3]connecting words, articles and prepositions irrelevant to identifying brand themes.

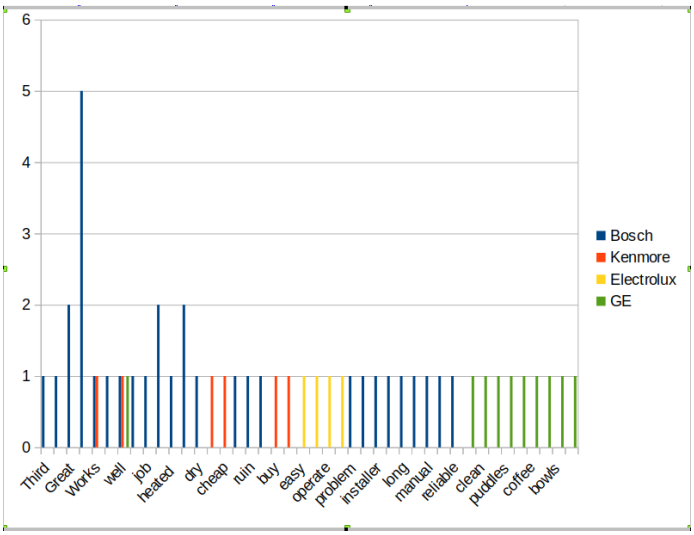

Figure 1: Frequency of each word from Verbatim Analysis grouped by Brand

2.1.2 Issues with Simple Verbatim Analysis

We can see that this set up has many limitations. The most important factor is determining the sentiment of the verbatim, (i.e., classifying the verbatim as positive or negative). This can be addressed through the "NaiveBayesClassifier()" object in nltk, but I would have to look into its practical application[4].

The second issue noticable from this subset of verbatims is that Bosch appears to have the highest count for "Great" which is misleading considering Bosch had 5 verbatims, while other brands had 1 or 2 verbatims each. expressing word frequency as rate (Word Count/Verbatim) may correct this bias.

The third issue is the inclusion of personal experiences by consumers and product testers. This makes assessment of the machines performance difficult, as these verbatims are not about the machine.

2.1.3 Recommendation for Consumers and Product Testers

While your personal experiences are valued and do help us understand how well a product works, they may misguide the algorithm towards adding importance to words from personal stories. We would appreciate consumers and product testers be specific about features or qualities that they enjoy about machine.

Product Testers, would greatly benefit greatly by systematically and categorically writing about features that each consumer would be concerned with. An example verbatim, shown in Table 7, could have a sentence concerning Build Quality, Noise, Cleaning Ability, Other categories.

[4]https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk

Table 6: Example Verbatim For Bosch Brand Dishwasher

Category	Verbatim
Build Quality	Reliable, beautiful brushed steel finish.
Noise	Very Quiet.
Cleaning Ability	Works well, great cleaning.
Other	No heating element, third rack is great.

2.2 Software Recommendation

Redacted

I would use Python with modules numpy, pandas, nltk to clean and parse the data. Once in an acceptable format. Statistical analyses can be done in R.

Why? unstructured data requires extensive manipulation with user defined functions, and requires modules that are more commonly found in Python. I find it hard to do these manipulations in R, and other languages.

3 Problem C: Analysis of Big Data

Given the nature of our work, Statisticians at Consumer Reports wear many hats. One of these hats is to support our data journalism effort. Several years ago, we purchased a large amount of insurance premium data to evaluate the effect of various factors on insurance premiums. We ultimately ended up with more than 4 billion data points, covering a wide range of age, gender, type of automobile, and many other factors affecting insurance premiums.

3.1 Data Validation

Redacted

3.1.1 Understanding The Data

The data frame has at least Age, Gender, and AutoType. Larson et al.(2017)[5] explains other factors to consider such as credit score, education level, family size, occupation, commute/driving habits, coverage amount (USD), and most important, ZIP Code. Something not considered in the study was discounts from bundled insurance. The table describing the data types for Age, Gender, AutoType, and ZIP code are shown below.

3.1.2 Sampling the Data - bootstrap method

Given that there are $\sim$ 4 billion entries, with $>$ 4 predictors. and assuming the constraint that I am on a laptop computer for my initial analysis. I would

[5]https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-methodology

Table 7: Vectoriztion of Entries

Column	Variable Type	Data Type (Original)	Data Type (Post-Validation)
Age	Numerical	integer	Age Range
Gender	Categorical	M/F	0,1
AutoType	Categorical?	String	Product ID
Zip code	Categorical	Integer	Zip Code Group
Monthly Premium	numerical	Float	No Change

bootstrap the numeric predictors (e.g., Age) and the response variable, Monthly Premium to get an estimate with a confidence interval of their means. To reduce the amount of data I am working with directly I take take a should take a sample.From my experience, my computer (a 2017 Lenovo E480) can handle data frames with $\sim 3,000,000$ entries. The caveat of taking a random sample is that by the central limit theorem, a random sample will take on a normal distribution which might not be the case from the population data frame. While not perfect, I can address the issue by ordering the data by Monthly insurance premium and taking a sequential sample of every 1000th entry.

3.1.3 Visualizing the Data

This reduced data frame would be most readily visualized by creating a 100 bin histogram of Monthly Premiums. A multimodal distribution will give us a hint that there are groups that need to be identified.

Given some distribution, we can now begin to identify groups. I can identify neighborhoods, and counties of similar "risk" levels as clusters of ZIP Codes by plotting ZIP Code against insurance premium. Despite ZIP Code being a categorical predictor, neighborhoods are typically in adjacent ZIP Codes. However, an appropriate clustering algorithm would have to be identified.

I could also visualize the data by plotting insurance premium against categorical predictor M/F, by encoding $M = 0$, and $F = 1$, but given that women pay $< 1\%$ more on car insurance[6], the plot would not yield much information. The caveat here is that there could be a unique interaction of predictors that together have a significant effect on Monthly Premium (e.g., Women in neighborhood containing 10463 pay significantly more than Men in neighborhood containing 10010).

A third visualization would be plotting Age against Monthly Premium. This would likely confirm that younger people have higher Monthly Premiums[7].

To compare categorical variables we could use contingency tables. This can offer some qualitative information as to possible reasons some ZIP Code groups would pay more for car insurance(e.g., maybe there is a neighborhood popular among younger drivers). With these preliminary statistics, we can move onto validating the data set.

[6] https://www.caranddriver.com/research/a31268333/which-gender-pays-more-for-car-insurance/

[7] https://wallethub.com/edu/ci/average-car-insurance-rates-by-age/69321/

3.1.4 Assessing Completeness

From my sampled data set, I can readily query each column and estimate the proportion of missing data. For a determined proportion of missing data, I would have to make an assess on how to impute missing values, or omit entries and acknowledge the bias in reporting.

3.1.5 Assessing Consistency

Each column has a possible range of values, or Nan for missing values. While imperfect, one can create an dictionary of unique values for each categorical column. For float and integer columns, we will have to establish reasonable ranges for each value.

Age: We can bound the driving age at 16 in the united states and cap it at 85, as it is rare to see people above that age driving freely. We can also group these into common insurance age groups $\{16\text{-}25, 25\text{-}45, 45\text{-}65, 65+\}$

Gender: Here I assume it is attributed to biological gender. Only Oregon, California, and Maine allow for a third option.[8]

AutoType: This will be the most difficult, as identical AutoTypes can have have misspellings, inconsistent formatting between insurance companies.

ZIP Code: This must be an 5 digit integer, some ZIP Codes are listed with a block number, that may have to be omitted.

Monthly Premium: This number must be a non-zero dollars and cents. Larson et al. (2017)[5] had ranges \$25 to \$400. For a new data set I would need to understand the extent of outliers. A Bonferroni Method may be used, but I do not know what is the best method for outlier detection.

3.1.6 Assessing Accuracy

Assessing the accuracy requires access to external sources to cross reference the data. A second data frame with insurance rates by ZIP Code would be beneficial for cross validation. The US Census Bureau has information on the ZIP Code level for median household income[9]. One could check that the mean premiums from the sampled data do not exceed a defined percentage of household income.

Other things to consider with accuracy is the age of the data. Without knowledge of when the insurance rate was agreed upon, some rates could be lower or higher depending on the business cycle, or pricing power of insurance companies.

3.1.7 Assessing Uniqueness

Assessing uniqueness of the data will not be feasible in the sampled data frame, as sampled values may be unique, but the parent data base could have duplicates. Without more information on unique identifiers (e.g., policy number) we would not be able to assess uniqueness based on predictors alone.

[8] https://www.compare.com/auto-insurance/resources/transgender-car-insurance
[9] https://www.census.gov/quickfacts/fact/table/US/PST045219

3.2 Ranking Insurance Data

Redacted

From Larson et. al (2017), we see that there are different insurance premiums for a given "state risk", more precisely zipcodes, which varies by insurance companies. The article also stated that the insurance premium is dependent on the driver profile. So I would reduce the space of insurance premiums by:

1. Restricting comparison/estimation set to the Zip Code.
2. Then restricting by Age Range, and Gender?
3. Then querying historical monthly premiums from each insurance provider.
4. For missing historical quotes, a linear model would have to be developed to give an board estimate of insurance quotes.
5. For a set of insurance quotes given a driver profile, they could be sorted by price to afford a final ranking.