PUBH 7405 -Block 1

LAND ACKNOWLEDGEMENT

The School of Public Health at the University of Minnesota Twin Cities is built within the traditional homelands of the Dakota people. Minnesota comes from the Dakota name for this region, Mni Sóta Maḳoce, which loosely translates to the land where the waters reflect the skies.

It is important to acknowledge the peoples on whose land we live, learn, and work as we seek to improve and strengthen our relations with our tribal nations. We also acknowledge that words are not enough. We must ensure that our institution provides support, resources, and programs that increase access to all aspects of higher education for our American Indian students, staff, faculty, and community members.

What is Biostatistics?

- Statistics applied to biomedical problems
- Decision making in the face of uncertainty or variability
- Design and analysis of experiments; detective work in observational studies (in epidemiology, outcomes research, etc.)
- Attempt to remove bias or find alternative explanations to those posited by researchers with vested interests
- Experimental design, measurement, description, statistical graphics, data analysis, inference

1.2 Types of Data Analysis and Inference

- Description: what happened to *past* patients

- Inference from specific (a sample) to general (a population)
  - Hypothesis testing: test a hypothesis about population or long-run effects
  - Estimation: approximate a population or long term average quantity
  - Prediction: predict the responses of other patients *like yours* based on analysis of patterns of responses in your patients

1.3 Types of Measurements by Their Role in the Study

- Response variable (clinical endpoint, final lab measurements, etc.)
- Independent variable (predictor or descriptor variable) — something measured when a patient begins to be studied, before the response; often not controllable by investigator, e.g. sex, weight, height, smoking history
- Adjustment variable (confounder) — a variable not of major interest but one needing accounting for because it explains an apparent effect of a variable of major interest or because it describes heterogeneity in severity of risk factors across patients
- Experimental variable, e.g. the treatment or dose to which a patient is randomized; this is an independent variable under the control of the researcher

Table 1.1: *Common alternatives for describing independent and response variables*

Response variable	Independent variable
Outcome variable	Exposure variable
Dependent variable	Predictor variable
y-variables	x-variable
Case-control group	Risk factor
	Explanatory variable

1.4 Types of Measurements According to Coding

- Binary: yes/no, present/absent

- Categorical (nominal, polytomous, discrete): more than 2 values that are not necessarily in special order

- Ordinal: a categorical variable whose possible values are in a special order, e.g., by severity of symptom or disease; spacing between categories is not assumed to be useful

- Count: a discrete variable that (in theory) has no upper limit, e.g. the number of ER visits in a day, the number of traffic accidents in a month

- Continuous: a numeric variable having many possible values representing an underlying spectrum

- Continuous variables have the most statistical information (assuming the raw values are used in the data analysis) and are usually the easiest to standardize across hospitals

- Turning continuous variables into categories by using intervals of values is arbitrary and requires more patients to yield the same statistical information (precision or power)

- Errors are not reduced by categorization unless that's the only way to get a subject to answer the question (e.g., income)

1.5 Random Variables

- A potential measurement X
- X might mean a blood pressure that will be measured on a randomly chosen US resident
- Once the subject is chosen and the measurement is made, we have a sample value of this variable
- Statistics often uses X to denote a potentially observed value from some population and x for an already-observed value (i.e., a constant)

Distributions

The *distribution* of a random variable X is a profile of its variability and other tendencies. Depending on the type of X, a distribution is characterized by the following.

- Binary variable: the probability of "yes" or "present" (for a population) or the proportion of same (for a sample).

- k-Category categorical (polytomous, multinomial) variable: the probability that a randomly chosen person in the population will be from category $i, i = 1, \ldots, k$. For a sample, use k proportions or percents.

- Continuous variable: any of the following 4 sets of statistics
  - probability density: value of x is on the x-axis, and the relative likelihood of observing a value "close" to x is on the y-axis. For a sample this yields a histogram.

- cumulative probability distribution: the y-axis contains the probability of observing $X \leq x$. This is a function that is always rising or staying flat, never decreasing. For a sample it corresponds to a cumulative histogram[a]

- all of the *quantiles* or *percentiles* of X

- all of the *moments* of X (mean, variance, skewness, kurtosis, . . .)

- If the distribution is characterized by one of the above four sets of numbers, the other three sets can be derived from this set

- Knowing the distribution we can make intelligent guesses about future observations from the same series, although unless the distribution really consists of a single point there is a lot of uncertainty in predicting an individual new patient's response. It is less difficult to predict the average response of a group of patients once the distribution is known.

- At the least, a distribution tells you what proportion of patients you would expect to see whose measurement falls in a given interval.

Random Variable

Formally speaking, a **random variable** is a real-valued function on the sample space S and maps elements of S, ω, to real numbers.

$$\begin{array}{ccc} S & \xrightarrow{X} & \mathbb{R} \\ \omega & \longmapsto & x = X(\omega) \end{array}$$

Ex 1. Let X be the number of heads in 3 tosses of a coin. Sample space $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. Then

$$X(HHH) = 3, \quad X(HHT) = 2, \quad X(HTH) = 2, \quad X(HTT) = 1,$$
$$X(THH) = 2, \quad X(THT) = 1, \quad X(TTH) = 1, \quad X(TTT) = 0$$

Ex 2. Let Y be the number of tosses required to get a head. $S = \{H, TH, TTH, TTTH, TTTTH, \ldots\}$ Then

$$Y(H) = 1, \; Y(TH) = 2, \; Y(TTH) = 3, \; Y(TTTH) = 4, \ldots$$

Probability Mass Function (pmf)

The *probability mass function* (pmf) of a random variable X is a function $p(x)$ that maps each possible value x_i to the corresponding probability $P(X = x_i)$.

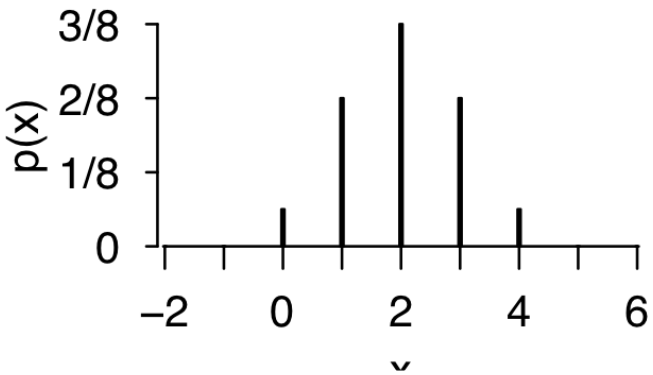
- A pmf $p(x)$ must satisfy $0 \leq p(x) \leq 1$ and $\sum_x p(x) = 1$.

Example (coin tossing on the previous slide)

Possible Values of X	0	1	2	3	4
Probabilities	1/16	4/16	6/16	4/16	1/16

The pmf of X is

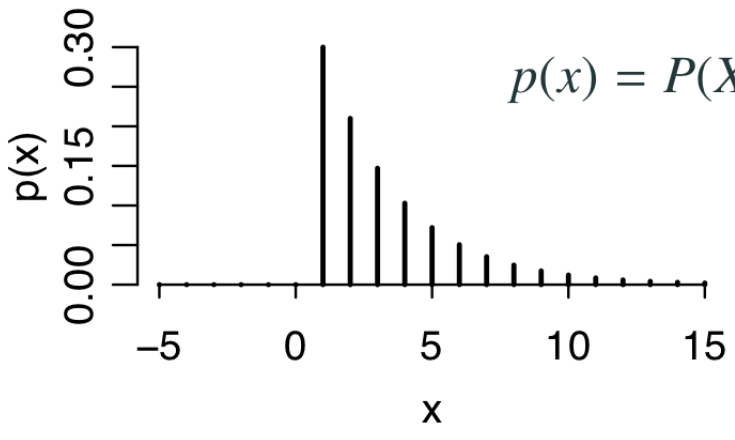
$$p(x) = \begin{cases} 1/16 & \text{if } x = 0 \text{ or } 4 \\ 4/16 & \text{if } x = 1 \text{ or } 3 \\ 6/16 & \text{if } x = 2 \\ 0 & \text{if } x \neq 0, 1, 2, 3, 4 \end{cases}$$



Example: Geometric Distribution

Let X be the number of tosses required to obtain the first heads, when tossing a coin with a probability of p to land heads.

The pmf of X is



$$
\begin{aligned}
p(x) = P(X = x) &= P(\overbrace{T \ldots T}^{x-1 \text{ tails}}\, H) \quad \text{by indep.} \\
&= P(T)(T) \cdots P(T)P(H) \\
&= \underbrace{(1-p)(1-p)\cdots(1-p)}_{x-1 \text{ copies}}\, p \\
&= (1-p)^{x-1} p,
\end{aligned}
$$

if x is a positive integer and $p(x) = 0$ if not.

- We say X has a **geometric distribution** since the pmf is a geometric sequence
- Does $\sum_{x=1}^{\infty} p(x) = 1$?

Expected Value = Expectation = Mean

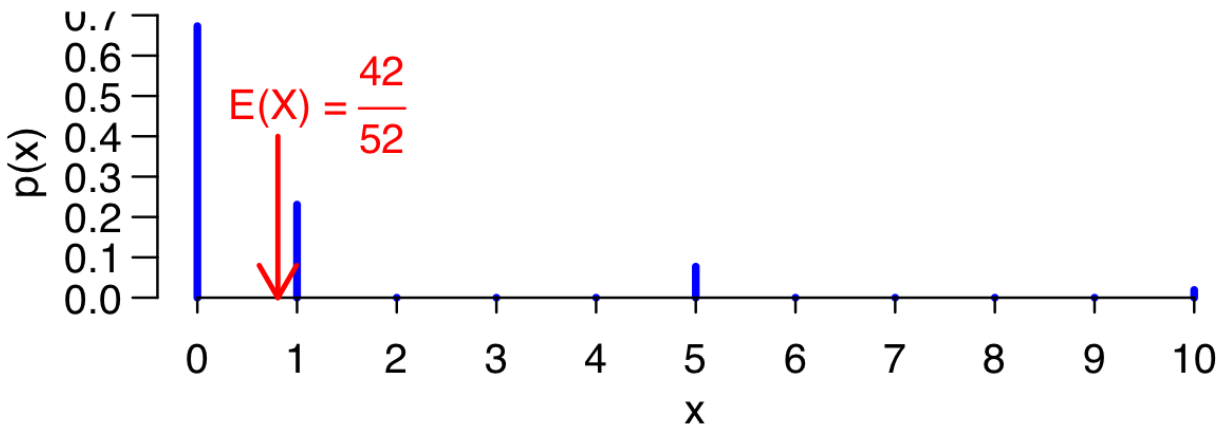
Let X be a discrete random variable with pmf $p(x)$. The **expected value** or the **expectation** or the **mean** of X, denoted by $\mathrm{E}[X]$, or μ is a *weighted average* of the possible values of X, where the weights are the probabilities of those values.

$$\begin{aligned}\mu &= \mathrm{E}[X] \\ &= \sum_{\text{all } x} xP(X = x) \\ &= \sum_{\text{all } x} xp(x)\end{aligned}$$

Example: Card Game — Expected Value

$$p(x) = \begin{cases} 35/52 & \text{if } x = 0 \\ 12/52 & \text{if } x = 1 \\ 4/52 & \text{if } x = 5 \\ 1/52 & \text{if } x = 10 \\ 0 & \text{if } x \neq 0, 1, 5, 10 \end{cases}$$

$$\mathrm{E}[X] = \sum_x xp(x) = 0 \times \frac{35}{52} + 1 \times \frac{12}{52} + 5 \times \frac{4}{52} + 10 \times \frac{1}{52} = \frac{42}{52} \approx 0.81$$



Expected Value of a Function of a Random Variable

In addition to the expected value of a random variable X itself, we might be also interested in the *expected value of a function of a random variable* $h(X)$, e.g.,

- the net profit from the card game $h(X) = 0.9X - 0.5$
- the net profit from the card game $h(X) = X - 0.02X^2 - 0.5$ with a new tax rule

Definition: If the pmf of X is $p_X(x)$, the expected value of $h(X)$ is

$$\mathrm{E}[h(X)] = \sum_x h(x)p_X(x).$$

Example 4 (Card Game w/ Tax)

One's expected net profit from the game is

Reward x	pmf $p(x)$	Net Profit $h(x) = 0.9x - 0.5$
1	$12/52$	$0.9 \cdot 1 - 0.5 = 0.4$
5	$4/52$	$0.9 \cdot 5 - 0.5 = 4.0$
10	$1/52$	$0.9 \cdot 10 - 0.5 = 8.5$
0	$35/52$	$0.9 \cdot 0 - 0.5 = -0.5$

$$\begin{aligned}
\mathrm{E}[h(X)] &= \sum_x h(x)p(x) \\
&= 0.4 \times \frac{12}{52} + 4.0 \times \frac{4}{52} + 8.5 \times \frac{1}{52} + (-0.5) \times \frac{35}{52} \\
&= \frac{11.8}{52} \approx 0.227
\end{aligned}$$

Variance of a Random Variable

One measure of spread of a random variable (or its probability distribution) is the *variance*.

The **variance** of a random variable X, denoted as σ_X^2 or $V(X)$ is defined as the *average squared distance from the mean*.

$$\text{Var}(X) = \sigma^2 = \text{"sigma squared"} = \text{E}\left[(X-\mu)^2\right]$$

Variance is in squared units.
Square root of the variance is the *standard deviation (SD).*

$$\text{SD}(X) = \sigma = \sqrt{\text{Var}(X)}$$

Example (Card Game)

Recall for the card game reward X:

$$\text{pmf: } \begin{array}{c|cccc} x & 0 & 1 & 5 & 10 \\ \hline p(x) & \frac{35}{52} & \frac{12}{52} & \frac{4}{52} & \frac{1}{52} \end{array}, \quad \text{and mean} = \mu = \mathrm{E}(X) = \frac{42}{52}.$$

Its variance is hence,

$$\begin{aligned} \mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2] &= \mathrm{E}\left[\left(X - \frac{42}{52}\right)^2\right] = \sum_x \left(x - \frac{42}{52}\right)^2 p(x) \\ &= \left(0 - \frac{42}{52}\right)^2 \cdot \frac{35}{52} + \left(1 - \frac{42}{52}\right)^2 \cdot \frac{12}{52} + \left(5 - \frac{42}{52}\right)^2 \cdot \frac{4}{52} + \left(10 - \frac{42}{52}\right)^2 \cdot \frac{1}{52} \\ &= \frac{9260}{52^2} \approx 3.42 \end{aligned}$$

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{9260}{52^2}} \approx \sqrt{3.42} \approx 1.85.$$

Observe the computation of the variance can be awkward if the expected value μ is not an integer.

A Shortcut Formula for Calculating Variance

$$\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2] = \mathrm{E}(X^2) - \mu^2$$

Proof.

$$\begin{aligned}
\mathrm{E}[(X-\mu)^2] &= \sum\nolimits_x (x-\mu)^2\, p(x) \\
&= \sum\nolimits_x (x^2 - 2\mu x + \mu^2)\, p(x) \\
&= \underbrace{\sum\nolimits_x x^2\, p(x)}_{=\mathrm{E}(X^2)} - 2\mu \underbrace{\sum\nolimits_x x\, p(x)}_{=\mu} + \mu^2 \underbrace{\sum\nolimits_x p(x)}_{=1} \\
&= \mathrm{E}(X^2) - 2\mu^2 + \mu^2 = \mathrm{E}(X^2) - \mu^2
\end{aligned}$$

Example (Card Game)

x	0	1	5	10
$p(x)$	$35/52$	$12/52$	$4/52$	$1/52$

Let's calculate the variance again using the shortcut formula $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2$. First we calculate $\mathrm{E}[X^2]$

$$\mathrm{E}[X^2] = 0^2 \cdot \frac{35}{52} + 1^2 \cdot \frac{12}{52} + 5^2 \cdot \frac{4}{52} + 10^2 \cdot \frac{1}{52} = \frac{212}{52}$$

and the variance is hence

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2 = \frac{212}{52} - \left(\frac{42}{52}\right)^2 = \frac{9260}{52^2}$$

which resembles our previous calculation.

Linear Transformation of a Random Variable

Linear transformation of a random variable $h(X) = aX + b$ is also a function of interest, e.g.,

- The net profit $h(X) = X - 0.1X - 0.5 = 0.9X - 0.5$ from the Card Game w/ tax

For $Y = aX + b$, we can show that

$$\mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b, \quad \text{and} \quad \mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$$

Before we get to the proofs.
Let's review properties of summation.

Review: Summation Notation and Its Properties

In the following, a is a fixed constant.

$$\sum_{i=1}^{n} a = (\underbrace{a + a + \cdots + a}_{n \text{ copies}}) = na$$

$$\begin{aligned}
\sum_{i=1}^{n} (ax_i) &= ax_1 + ax_2 + \cdots + ax_n \\
&= a(x_1 + x_2 + \cdots + x_n) \\
&= a \sum_{i=1}^{n} x_i
\end{aligned}$$

$$\begin{aligned}
\sum_{i=1}^{n} (x_i + y_i) &= (x_1 + y_1) + (x_2 + y_2) + \cdots + (x_n + y_n) \\
&= (x_1 + x_2 + \cdots + x_n) + (y_1 + y_2 + \cdots + y_n) \\
&= \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i
\end{aligned}$$

Proof of $\mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b$

We prove it for the case that X is discrete with pmf $p(x)$.\ This relation is also true when X is continuous.

$$
\begin{aligned}
&\mathrm{E}(aX + b) \\
&= \sum_x (ax + b)p(x) && \text{(definition of } \mathrm{E}(aX + b)\text{)} \\
&= \sum_x (axp(x) + bp(x)) \\
&= \sum_x axp(x) + \sum_x bp(x) && \text{(since } \sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i\text{)} \\
&= a\underbrace{\sum_x xp(x)}_{=\mathrm{E}(X)} + b\underbrace{\sum_x p(x)}_{=1} && \text{(since } \sum_{i=1}^{n}(ax_i) = a\sum_{i=1}^{n} x_i\text{)} \\
&= aE(X) + b
\end{aligned}
$$

Proof of $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$

Recall $\mathrm{Var}(Y)$ is the expected value of $[Y - \mathrm{E}(Y)]^2$.

For $Y = aX + b$, we have proved that $\mathrm{E}(Y) = E(aX + b) = a\mu + b$, where $\mu = \mathrm{E}(X)$ and hence

$$[Y - \mathrm{E}(Y)]^2 = [(aX + b) - \mathrm{E}(aX + b)]^2 = [aX + b - (a\mu + b)]^2 = a^2(X - \mu)^2.$$

Taking expected value of the above we get

$$\begin{array}{ccc} \mathrm{E}[Y - \mathrm{E}(Y)]^2 & = & E[a^2(X - \mu)^2] \\ \| & & \|^* \\ \mathrm{Var}(Y) & & a^2\,\mathrm{E}[(X - \mu)^2] \\ \| & & \| \\ \mathrm{Var}(aX + b) & & a^2\,\mathrm{Var}(X) \end{array}$$

in which the step $\mathrm{E}[a^2(X - \mu)^2] = a^2\,\mathrm{E}[(X - \mathrm{E}(X))^2]$ is justified using $\mathrm{E}[cW + d] = c\,\mathrm{E}[W] + d$ we just proved with $c = a^2$, $W = (X - E(X))^2$, and $d = 0$.

Example (Card Game w/ Tax)

For the Card Game, recall the mean and variance of the reward X are

$$\mathrm{E}(X) = \frac{42}{52}, \quad \mathrm{Var}(X) = \frac{9620}{52^2}$$

The mean and variance of the net profit with tax $h(X) = 0.9X - 0.5$ are

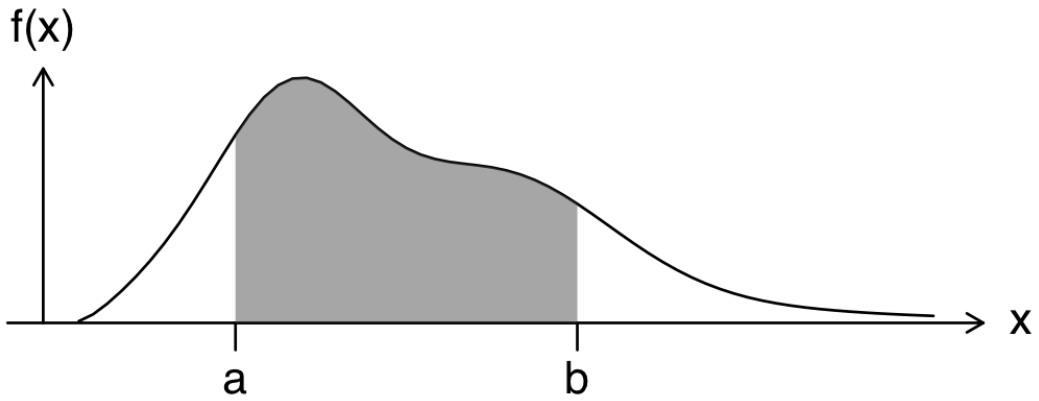
$$\mathrm{E}(0.9X - 0.5) = 0.9\,\mathrm{E}(X) - 0.5 = 0.9 \times \frac{42}{52} - 0.5 = \frac{11.8}{52}$$

$$\mathrm{Var}(0.9X - 0.5) = 0.9^2\,\mathrm{Var}(X) = 0.9^2 \times \frac{9620}{52^2} = \frac{7792.2}{52}$$

Continuous Random Variables

A random variable X is said to have a *continuous distribution* if there exists a non-negative function f such that

$$P(a < X \leq b) = \int_a^b f(x)\,dx, \quad \text{for all } -\infty \leq a < b \leq \infty.$$



Here f is called the *probability density function* (*pdf*), the *density curve*, or the *density* of X.

Conditions of pdf

A pdf $f(x)$ can be of any imaginable shape but must satisfy the following:

- It must be *nonnegative*

$$f(x) \geq 0 \text{ for all } x$$

- The total area under the pdf must be 1

$$\int_{-\infty}^{\infty} f(x)\,dx = P(-\infty < X \leq \infty) = 1$$

Interpretation of a pdf

Suppose f is the pdf of X. If f is continuous at a point x, then for small δ

$$P\left(x - \frac{\delta}{2} < X \leq x + \frac{\delta}{2}\right) = \int_{x-\delta/2}^{x+\delta/2} f(u)\,du = \delta f(x).$$

- Is the pdf f of a random variable always ≤ 1?
  No, the pdf $f(x)$ itself is not a probability.
  It's the area underneath $f(x)$ that represents the probability.
- For any continuous random variable X

$$P(X = x) = \int_{x}^{x} f(u)du = 0$$

- What percentage of men are 6-feet tall exactly?
  Those that are 6.00001 or 5.99999 feet tall don't count.
- A pdf $f(x)$ may not be continuous

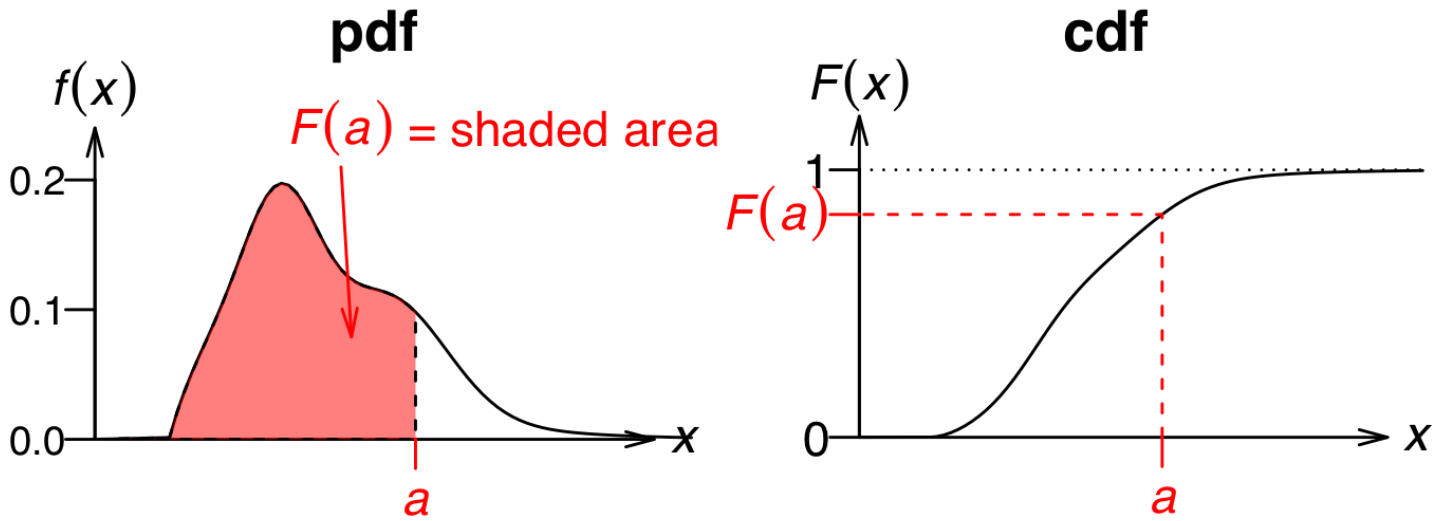
Cumulative Distribution Function (cdf)

For any random variable X, its *cumulative distribution function* (*cdf*) is the function defined by

$$F(x) = F_X(x) = P(X \leq x).$$

One get the cdf of a random variable from its pdf by **integration**:

$$F(x) = \int_{-\infty}^{x} f(u)\,du$$



Obtaining the PDF from the CDF

The PDF can be obtained from the cdf by differentiation.

$$f(x) = \frac{d}{dx}F(x).$$

Example 1

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{3}x^2 & \text{if } 0 \le x \le 1 \\ \frac{1}{3} + \frac{2}{3}(x-1) & \text{if } 1 \le x \le 2 \\ 1 & \text{if } x > 2 \end{cases} \Rightarrow \frac{d}{dx}F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{2}{3}x & \text{if } 0 \le x \le 1 \\ \frac{2}{3} & \text{if } 1 \le x \le 2 \\ 0 & \text{if } x > 2 \end{cases}$$

Observe $\frac{d}{dx}F(x)$ is exactly the pdf $f(x)$.

Example 2. For the cdf of the battery life distribution

$$F(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 - e^{-2t} & \text{for } t \ge 0 \end{cases} \Rightarrow \frac{d}{dx}F(x) = \begin{cases} 0 & \text{if } t < 0 \\ 2e^{-2t} & \text{for } t \ge 0 \end{cases}$$

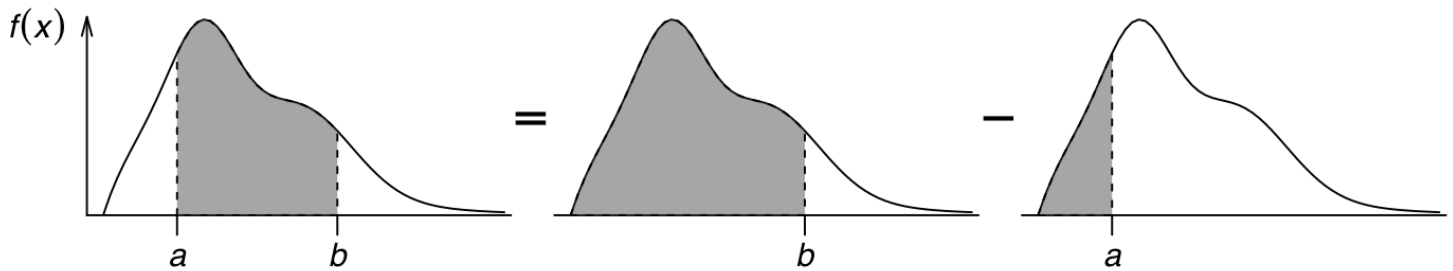
Using the cdf to Compute Probabilities

Let X be a continuous rv with pdf $f(x)$ and cdf $F(x)$. Then for any number a,

$$P(X > a) = 1 - F(a)$$

and for any two numbers a and b with $a < b$,

$$P(a \le X \le b) = F(b) - F(a)$$



Recall in **Example 2**, we computed $P(0.5 < T < 1)$ by integrating the pdf. We can also compute using the cdf, $F(t) = 1 - e^{-2t}$, $t > 0$.

$$P(0.5 < T < 1) = F(1) - F(0.5) = (1 - e^{-2}) - (1 - e^{-1}) = e^{-1} - e^{-2}$$

which agrees with our prior calculation.

Properties of cdfs

- The cdf $F(x) = P(X \leq x)$ is a probability, and hence it must be *between 0 and 1*.

$$0 \leq F(x) \leq 1$$

- cdfs are always *non-decreasing*. For $a < b$

$$F(b) - F(a) = P(X \leq b) - P(X \leq a) = P(a < X \leq b) \geq 0$$

- The cdf of a continuous r.v. must be *continuous*. As $\delta \to 0$

$$F(x + \delta) - F(x) = \int_x^{x+\delta} f(u)du \to 0$$

Expected Values

Let X be a continuous random variable with density f_X, then the **expectation** of X is

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx.$$

Suppose $Y = g(X)$ is a function of X. The **expectation** of Y is

$$\mathrm{E}(Y) = \mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx.$$

Variance and Standard Deviation

The *variance* of a continuous r.v. X, with density $f(x)$, and mean μ, is denoted as $\text{Var}(X)$, σ_X^2, or simply σ^2, is defined as

$$\text{Var}(X) = \text{E}(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx., \quad \text{where } \mu = \text{E}(X).$$

The *standard deviation* (*SD*) is the square root of the variance,

$$\text{SD}(X) = \sigma = \sqrt{\text{Var}(X)}.$$

Properties of the Expected Value and Variance

Property 1. The shortcut formula to find the variance remains valid for continuous random variables.

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2.$$

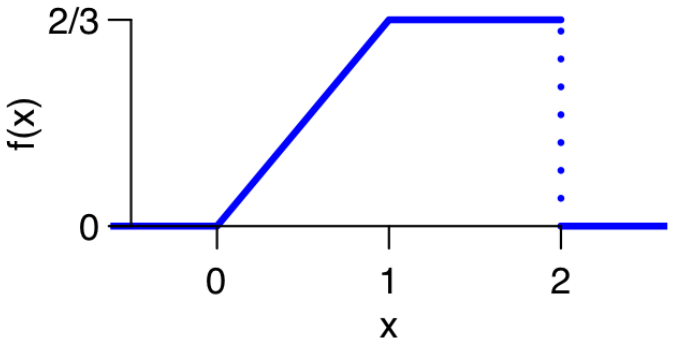
Property 2. For any constants a and b, the following identities are also valid for continuous r.v. X.

- $\mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b$
- $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$
- $\mathrm{SD}(aX + b) = |a|\mathrm{SD}(X)$

The proofs are similar to the ones for the discrete case, just replacing the summation $\sum$ with the integral $\int$, and hence are omitted.

Example (Mean, Variance, SD)

$$f(x) = \begin{cases} 2x/3 & \text{if } 0 \le x \le 1 \\ 2/3 & \text{if } 1 \le x \le 2 \\ 0 & \text{elsewhere} \end{cases}$$



$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x \frac{2x}{3} dx + \int_1^2 x \frac{2}{3} dx$$

$$= \frac{2x^3}{9}\Big|_0^1 + \frac{x^2}{3}\Big|_1^2 = \frac{2}{9} + \frac{4}{3} - \frac{1}{3} = \frac{11}{9}$$

$$\mathrm{E}(X^2) == \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^1 x^2 \frac{2x}{3} dx + \int_1^2 x^2 \frac{2}{3} dx$$

$$= \frac{x^4}{6}\Big|_0^1 + \frac{2x^3}{9}\Big|_1^2 = \frac{1}{6} + \frac{16}{9} - \frac{2}{9} = \frac{31}{18}$$

$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{31}{18} - (\frac{11}{9})^2 = \frac{37}{162}$ by the shortcut formula.

$\mathrm{SD}(X) = \sqrt{37/162} \approx 0.478.$

Expected Values of Functions of X & Y

For two random variable X, Y with

- a joint pmf $p(x, y)$, or
- a joint cdf $f(x, y)$,

the expected value of a function $g(X, Y)$ of X and Y is defined as

$$\mathrm{E}[g(X, Y)] = \begin{cases} \sum_{xy} g(x, y) p(x, y) & \text{for discrete case,} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y)\, dx\, dy & \text{in continuous case.} \end{cases}$$

Example (Gas Station)

Recall the joint pmf for the Gas Station Example in L09 is the table on the right. Suppose we are interested in

		Y (full-service)		
$p(x,y)$		0	1	2
X	0	0.10	0.04	0.02
self-	1	0.08	0.20	0.06
service	2	0.06	0.14	0.30

$$g(X,Y) = |X - Y|$$

$$= \text{the absolute diff in the \# of hoses in use}$$

$$\text{of the self-service and full-service islands.}$$

The expected value is

$$\begin{aligned}
\mathrm{E}\,|X - Y| &= \sum\nolimits_{xy} |x - y| p(x,y) \\
&= |0-0| \cdot 0.10 + |0-1| \cdot 0.04 + |0-2| \cdot 0.02 \\
&\quad + |1-0| \cdot 0.08 + |1-1| \cdot 0.20 + |1-2| \cdot 0.06 \\
&\quad + |2-0| \cdot 0.06 + |2-1| \cdot 0.14 + |2-2| \cdot 0.30 \\
&= 0.48
\end{aligned}$$

$\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$

If $g(X, Y) = aX + bY$ for two random variables X and Y and two constants a and b, we have

$$\mathrm{E}[g(X, Y)] = \mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$$

no matter X and Y are both discrete, both continuous, or one discrete and one continuous.

Proof. We will prove it for the case when X and Y are continuous with joint pdf $f(x, y)$. The proof for the discrete case is similar. By definition, the expected value of the function $g(X, Y) = aX + bY$ of X and Y is

$$\begin{aligned}\mathrm{E}(aX + bY) &= \iint (ax + by) f(x, y)\mathrm{d}x\mathrm{d}y \\ &= \underbrace{\iint ax f(x, y)\mathrm{d}x\mathrm{d}y}_{\text{Part I}} + \underbrace{\iint by f(x, y)\mathrm{d}x\mathrm{d}y}_{\text{Part II}}\end{aligned}$$

For Part I, we first integrate over y, and then over x.

$$\text{Part I} = \iint axf(x,y)\mathrm{d}x\mathrm{d}y = a\int\left(\int xf(x,y)dy\right)\mathrm{d}x$$
$$= a\int x\underbrace{\int f(x,y)dy}_{f_X(x)}\mathrm{d}x = a\underbrace{\int xf_X(x)\mathrm{d}x}_{\mathrm{E}(X)} = a\,\mathrm{E}(X)$$

For Part II, we first integrate over x, and then over y.

$$\text{Part II} = \iint byf(x,y)\mathrm{d}x\mathrm{d}y = b\int\left(\int yf(x,y)dx\right)\mathrm{d}y$$
$$= b\int y\underbrace{\int f(x,y)dx}_{f_Y(y)}\mathrm{d}y = b\underbrace{\int yf_Y(y)\mathrm{d}y}_{\mathrm{E}(Y)} = b\,\mathrm{E}(Y)$$

Putting Parts I & II together, we get

$$\mathrm{E}(aX+bY) = \mathrm{E}(aX) + \mathrm{E}(bY).$$

Expected Value for Linear Combination of Random Variables

The result $\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$ can be generalized to linear combinations of several random variables

$$\mathrm{E}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1\,\mathrm{E}(X_1) + a_2\,\mathrm{E}(X_2) + \cdots + a_n\,\mathrm{E}(X_n),$$

no matter the rv's are discrete or continuous, independent or not.

$\mathrm{E}[g(X)h(Y)] = \mathrm{E}[g(X)]\,\mathrm{E}[h(Y)]$ **if** X **&** Y **are independent**

When X and Y are **independent**, for any functions g and h,

$$\mathrm{E}[g(X)h(Y)] = \mathrm{E}[g(X)]\,\mathrm{E}[h(Y)].$$

In particular, $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$.

Proof. We prove the discrete case. The continuous case is similar. Using that $p(x, y) = p_X(x)p_Y(y)$ when X, Y are indep, one has

$$\begin{aligned}
\mathrm{E}[g(X)h(Y)] &= \sum_{xy} g(x)h(y)p(x,y) \\
&= \sum_x \sum_y g(x)h(y)p_X(x)p_Y(y) \quad (p(x,y) = p_X(x)p_Y(y) \text{ by indep.}) \\
&= \underbrace{\sum_x g(x)p_X(x)}_{\mathrm{E}[g(X)]} \underbrace{\sum_y h(y)p_Y(y)}_{\mathrm{E}[h(Y)]} = \mathrm{E}[g(X)]\,\mathrm{E}[h(Y)]
\end{aligned}$$

Covariance

The **covariance** of X and Y, denoted as $\mathrm{Cov}(X, Y)$ or σ_{XY}, is defined as

$$\mathrm{Cov}(X, Y) = \sigma_{XY} = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)],$$

in which $\mu_X = \mathrm{E}(X)$, $\mu_Y = \mathrm{E}(Y)$

- Covariance is a generalization of variance:

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X) = \mathrm{E}[(X - \mu_X)^2]$$

- Covariance can be positive or negative:

  - $\mathrm{Cov}(X, Y) > 0$ means positive association between X, Y
  - $\mathrm{Cov}(X, Y) < 0$ means negative association between X, Y

Shortcut Formula for Covariance

$$\boxed{\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)}$$

- Like the Shortcut Formula for Variance $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$.
- If X & Y are indep., then $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$, which implies $\mathrm{Cov}(X, Y) = 0$.
- However $\mathrm{Cov}(X, Y) = 0$ does **not** imply the independence of X and Y. In this case, we say X and Y are **uncorrelated.**
- Proof of the shortcut formula:

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathrm{E}[(X - \mu_X)(Y - \mu_Y)] \\
&= \mathrm{E}(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) \\
&= \mathrm{E}(XY) - \mu_X \underbrace{\mathrm{E}(Y)}_{=\mu_Y} - \mu_Y \underbrace{\mathrm{E}(X)}_{=\mu_X} + \mu_X \mu_Y \\
&= \mathrm{E}(XY) - \mu_X \mu_Y
\end{aligned}$$

Conditional Distributions of Discrete Random Variables

A conditional pmf of Y given $X = x$ is $p_{Y|X}(y \mid x) = \dfrac{P(x, y)}{p_X(x)}$ which satisfies

$$0 \leq p_{Y|X}(y \mid x) \leq 1 \quad \text{and} \quad \sum_y p_{Y|X}(y \mid x) = 1, \quad \text{for all } x.$$

A conditional pmf of X given $Y = y$ is $p_{X|Y}(x \mid y) = \dfrac{P(x, y)}{p_Y(y)}$ which satisfies

$$0 \leq p_{X|Y}(x \mid y) \leq 1 \quad \text{and} \quad \sum_x p_{X|Y}(x \mid y) = 1, \quad \text{for all } y.$$

Example

conditional pmf $p(y \mid x)$		Y = 0	1	2	Row Sum
	0	0.625	0.25	0.125	1
X	1	0.235	0.588	0.176	1
	2	0.12	0.18	0.60	1
marginal pmf	$p_Y(y)$	0.24	0.38	0.38	1

- Each row is a pmf for Y given some x value
- Observed the row sums of $p_{Y|X}(y \mid x)$ are all 1

Conditional Distributions of Continuous Random Variables

Given two continuous random variables with the joint pdf $f(x, y)$, the **conditional probability distribution of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$** is the function $f_{X|Y}$, and the **conditional probability distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$** is the function $f_{Y|X}$ are defined as

$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)}, \quad f_{Y|X}(y \mid x) = \frac{f(x, y)}{f_X(x)}$$

- Recall the definition of conditional probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

- Similarly, given the joint pdf of two continuous r.v.'s, the conditional pdf is the ratio of the joint pdf and marginal pdf,

$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)}, \quad f_{Y|X}(y \mid x) = \frac{f(x, y)}{f_X(x)}$$

Conditional $=$ Marginal, when Independent

What is the conditional distribution of Y given $X = x$ if X and Y are independent?

$$f_{Y|X}(y|X = x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{f_X(x)f_Y(y)}{f_X(x)} = f_Y(y).$$

i.e., *conditional pdf $Y|X$ is the marginal pdf of Y*.

In fact, the following three are equivalent definitions of the independence of X and Y

- $f(x, y) = f_X(x)f_Y(y)$ . . . . . (joint = product of marginal)
- $f_{Y|X}(y|X = x) = f_Y(y)$ . . . . . (conditional $Y|X$ = marginal of Y)
- $f_{X|Y}(x|Y = y) = f_X(x)$ . . . . . (conditional $X|Y$ = marginal of X)

All the things above apply to joint/conditional/marginal **pmf** for discrete $X, Y.$, too.

Conditional Expected Values

For two random variables X and Y, the *conditional mean* or *conditional expected value of* Y given $X = x$ is

$$\mu_{Y|X=X} = \mathrm{E}(Y \mid X = x) = \begin{cases} \sum_y y\, p_{Y|X}(y \mid x) & \text{if } X, Y \text{ are discrete} \\ \int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x) dy & \text{if } X, Y \text{ are continuous} \end{cases}$$

where $p_{Y|X}(y|x)$ and $f_{Y|X}(y|x)$ are the conditional pmf/pdf of Y given X.

Note: The conditional mean of Y given $X = x$ $\mathrm{E}(Y \mid X = x)$ is NOT a single value but a function of the x value given.

Example - Discrete Case

Recall the conditional pmf of Y given $X = x$ is as follows.

conditional pmf $p(y \mid x)$		Y = 0	1	2	Row Sum
	0	0.625	0.25	0.125	1
X	1	0.235	0.588	0.176	1
	2	0.12	0.18	0.60	1

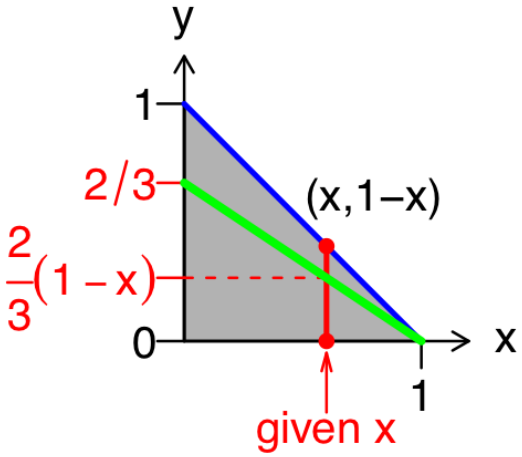
The conditional mean of Y given $X = x$ is

$$\mathrm{E}(Y \mid X = x) = \begin{cases} 0 \cdot 0.625 + 1 \cdot 0.25 + 2 \cdot 0.125 = 0.5 & \text{if } x = 0 \\ 0 \cdot 0.235 + 1 \cdot 0.588 + 2 \cdot 0.176 = 0.94 & \text{if } x = 1 \\ 0 \cdot 0.12 + 1 \cdot 0.18 + 2 \cdot 0.6 = 1.38 & \text{if } x = 2 \end{cases}$$

Example – Continuous Case

$$f_{Y|X}(y \mid x) = \frac{f(x,y)}{f_X(x)} = \frac{2y}{(1-x)^2}, \quad \text{for } 0 \le y \le 1-x\,.$$

$$\begin{aligned} \mathrm{E}(Y \mid X = x) &= \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) dy \\ &= \int_0^{1-x} \frac{y \times 2y}{(1-x)^2} dy \\ &= \frac{2y^3}{3(1-x)^2}\Big|_{y=0}^{y=1-x} = \frac{2}{3}(1-x). \end{aligned}$$



Principles of experimental design

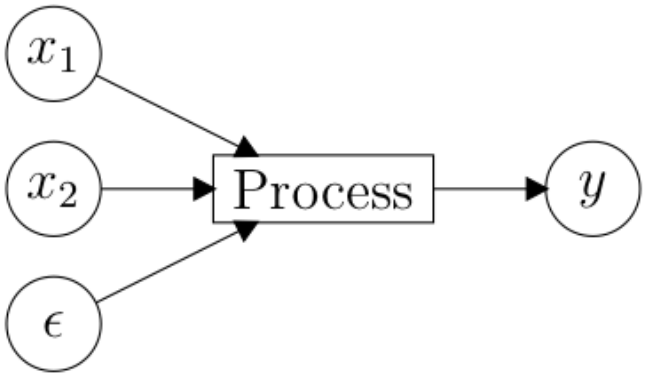
Much of our scientific knowledge about processes and systems is based on *induction*: reasoning from the specific to the general.

Example (survey): Do you favor increasing the gas tax for public transportation?

- Specific cases: 200 people called for a telephone survey
- Inferential goal: get information on the opinion of the **entire** city.

Example (Women's Health Initiative): Does hormone replacement improve health status in post-menopausal women?

- Specific cases: Health status monitored in 16,608 women over a 5-year period. Some took hormones, others did not.
- Inferential goal : Determine if hormones improve the health of women not in the study.



We are interested in how the *inputs* of a process affect an *output*. Input variables consist of

controllable factors x_1: measured and determined by scientist.

uncontrollable factors x_2: measured but not determined by scientist.

noise factors ϵ: unmeasured, uncontrolled factors, often called experimental variability or "error".

For any interesting process, there are inputs such that:

$$\text{variability in input} \rightarrow \text{variability in output}$$

If variability in an input factor x leads to variability in output y, we say x is a *source of variation*. In this class we will discuss methods of designing and analyzing experiments to determine important sources of variation.

Experiments and observational studies

Information on how *inputs* affect *output* can be gained from:

- Observational studies: Input and output variables are observed from a pre-existing population. It may be hard to say what is input and what is output.
- Controlled experiments: One or more input variables are controlled and manipulated by the experimenter to determine their *effect* on the output.

Example (Women's Health Initiative, WHI):

- Population: Healthy, post-menopausal women in the U.S.
- Input variables:
  1. estrogen treatment, yes/no
  2. demographic variables (age, race, diet, etc.)
  3. unmeasured variables (?)
- Output variables:
  1. coronary heart disease (eg. MI)
  2. invasive breast cancer
  3. other health related outcomes
- Scientific question: How does estrogen treatment affect health outcomes?

Observational Study:

1. **Observational population:** 93,676 women enlisted starting in 1991, tracked over eight years on average. Data consists of x= input variables, y=health outcomes, gathered concurrently on existing populations.
2. **Results:** good health/low rates of CHD generally associated with estrogen treatment.
3. **Conclusion:** Estrogen treatment is positively associated with health outcomes, such as prevalence of CHD.

Experimental Study (WHI randomized controlled trial):

1. Experimental population:

373,092 women determined to be eligible

$\hookrightarrow$ 18,845 provided consent to be in experiment

$\hookrightarrow$ 16,608 included in the experiment

16,608 women randomized to either $\begin{cases} x = 1 & \text{(estrogen treatment)} \\ x = 0 & \text{(no estrogen treatment)} \end{cases}$

Women were of different ages and were treated at different clinics. Women were *blocked* together by age and clinic, and then treatments were randomly assigned within each age×treatment *block*. This type of random allocation is called a *randomized block design*.

		age group		
		1 (50-59)	2 (60-69)	3 (70-79)
clinic	1	n_{11}	n_{12}	n_{13}
	2	n_{21}	n_{22}	n_{23}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$

$$\begin{aligned} n_{i,j} &= \#\text{ of women in study, in clinic } i \text{ and in age group } j \\ &= \#\text{ of women in } \textit{block } i, j \end{aligned}$$

Randomization scheme: For each block, 50% of the women in that block were randomly assigned to treatment ($x = 1$) and the remaining assigned to control ($x = 0$).

2. **Results:** JAMA, July 17 2002. Also see the NLHBI press release. Women on treatment had a **higher incidence rate** of

   - CHD
   - breast cancer
   - stroke
   - pulmonary embolism

   and a **lower incidence rate** of

   - colorectal cancer
   - hip fracture

3. **Conclusion:** Estrogen treatment is not a viable preventative measure for CHD in the general population. That is, our *inductive inference* is

   (specific) higher CHD rate in treatment population than control

   suggests

   (general) if everyone in the population were treated, the incidence rate of CHD would increase.

Descriptive Statistics (Using Sample)

Categorical Variables

- Proportions of observations in each category
  Note: The mean of a binary variable coded 1/0 is the proportion of ones.

- For variables representing counts (e.g., number of comorbidities), the mean is a good summary measure (but not the median)

- Modal (most frequent) category

Continuous Variables

Denote the sample values as $x_1, x_2, \ldots, x_n$

Measures of Location

"Center" of a sample

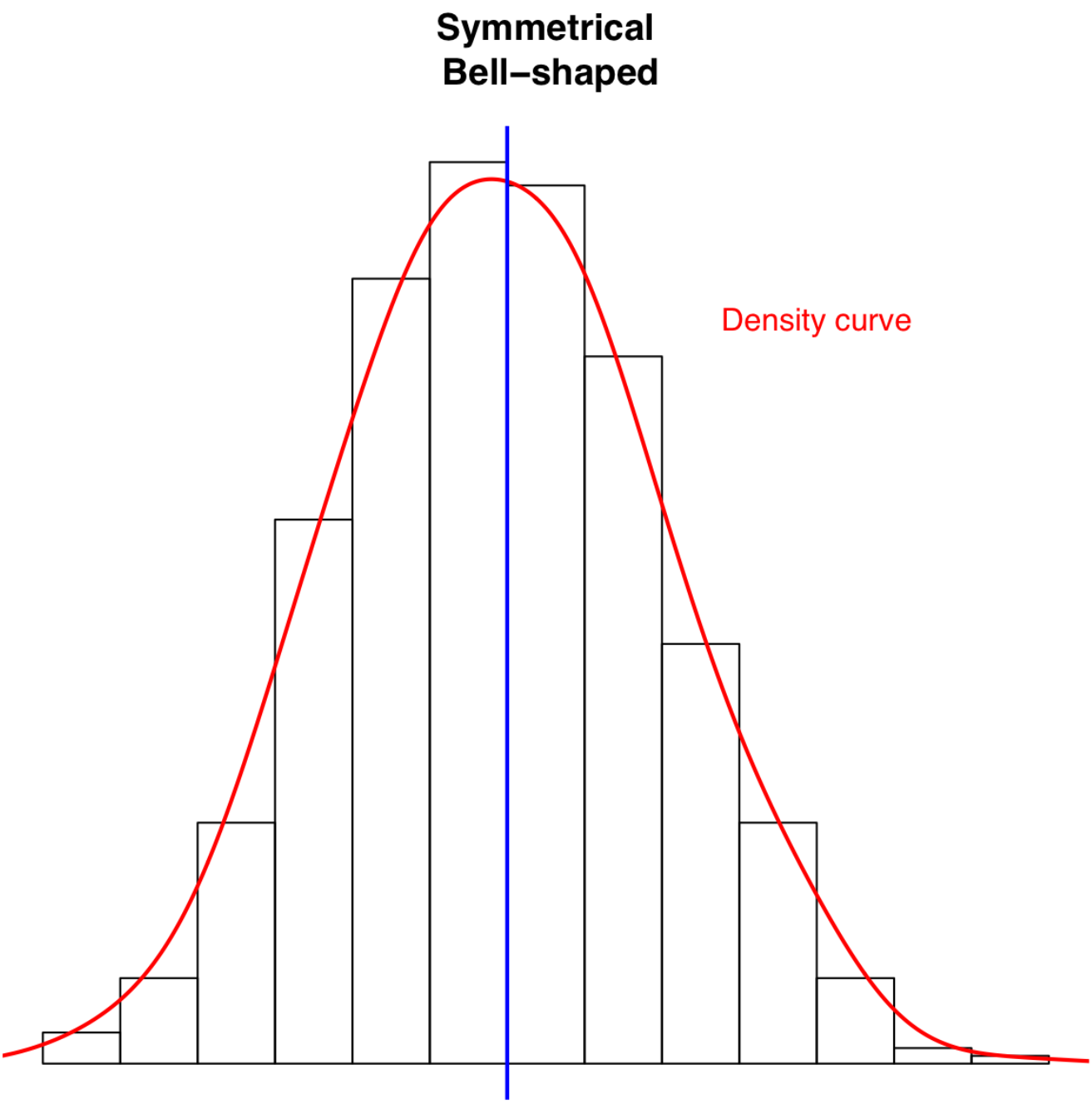
- Mean: arithmetic average

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

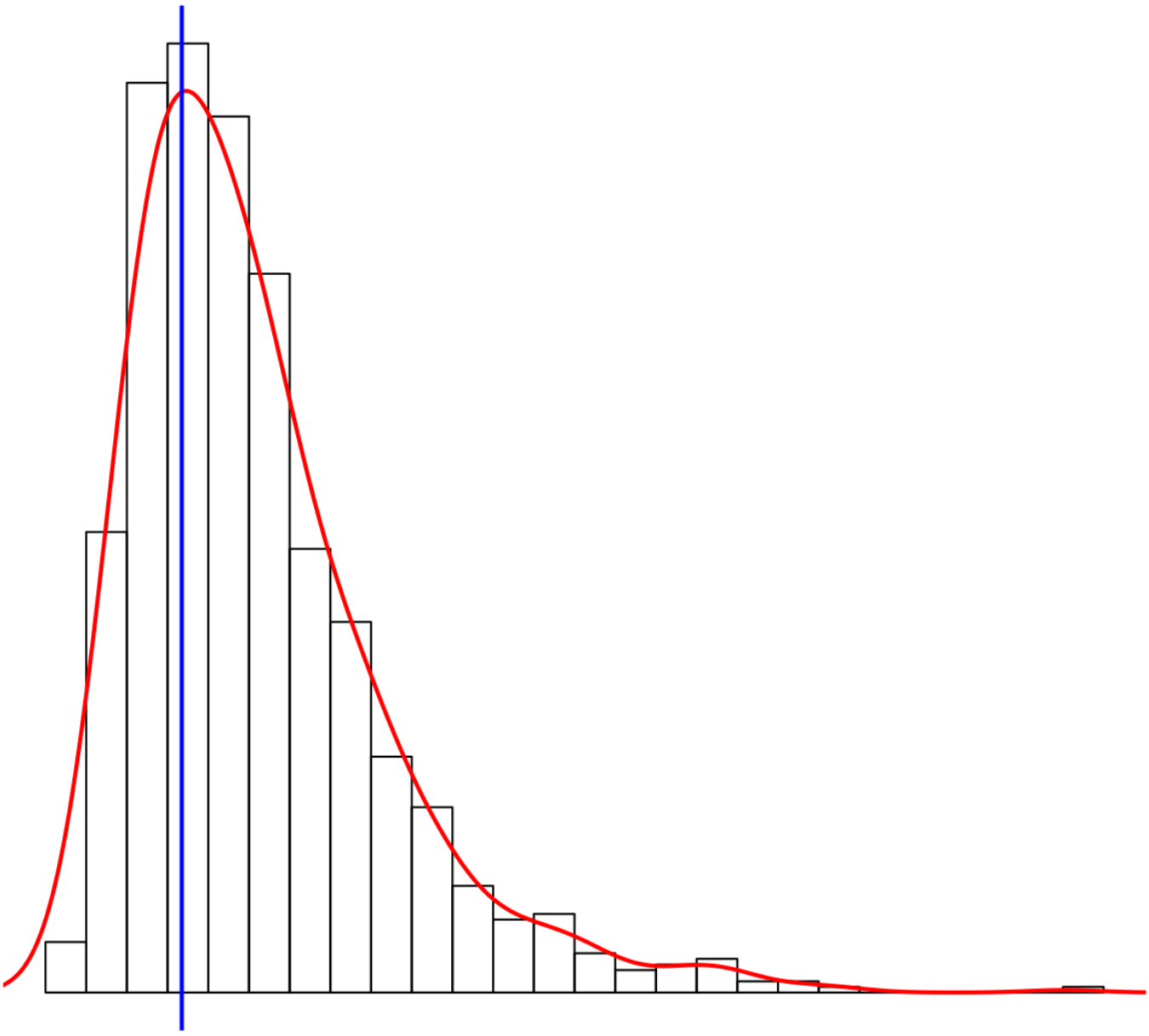
  Population mean μ is the long-run average (let $n \to \infty$ in computing $\bar{x}$)

  – center of mass of the data (balancing point)

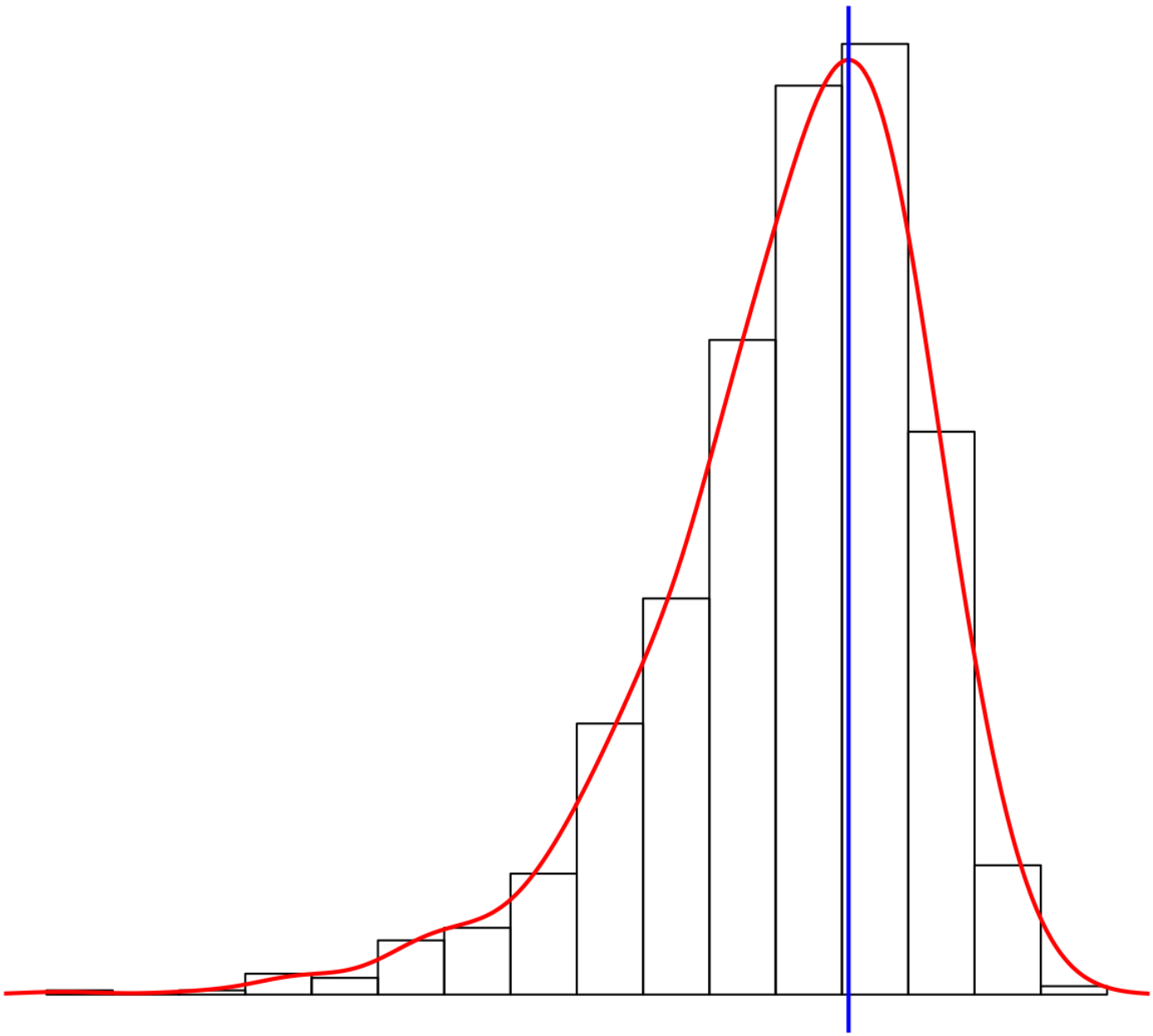
  – highly influenced by extreme values even if they are highly atypical


Symmetrical
Bell–shaped
Density curve

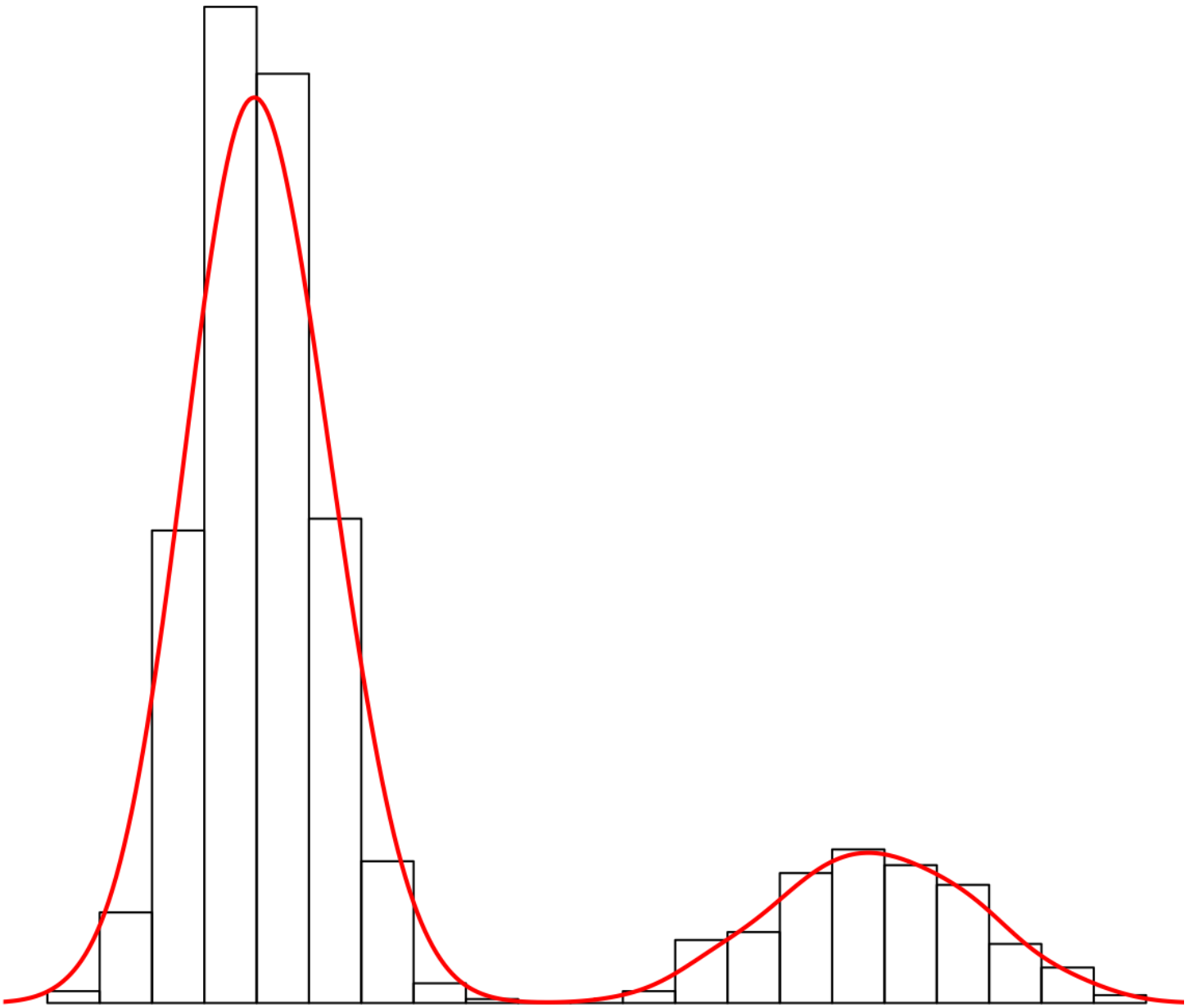
Positively Skewed
Skewed to the right



Negatively skewed
Skewed to the left



Bimodal



- Median: middle sorted value, i.e., value such that $\frac{1}{2}$ of the values are below it and above it
  - always descriptive
  - unaffected by extreme values
  - not a good measure of central tendency when there are heavy ties in the data
  - if there are heavy ties and the distribution is limited or well-behaved, the mean often performs better than the median (e.g., mean number of diseased fingers)

- Geometric mean: hard to interpret and effected by low outliers; better to use median

Example:

All of these sample summaries are easily obtained in R:

```
> yA<-c(11.4, 23.7, 17.9, 16.5, 21.1, 19.6)
> yB<-c(26.9, 26.6, 25.3, 28.5, 14.2, 24.3)


> mean(yA)
[1] 18.36667
> mean(yB)
[1] 24.3

> median(yA)
[1] 18.75
> median(yB)
[1] 25.95
```

trt A

trt B

$|\bar{y}_B - \bar{y}_A| = 5.93$

small y's to indicate based on actual data

Quantiles

Quantiles are general statistics that can be used to describe central tendency, spread, symmetry, heavy tailedness, and other quantities.

- Sample median: the 0.5 quantile or 50^{th} percentile
- Quartiles Q_1, Q_2, Q_3: 0.25 0.5 0.75 quantiles or $25^{th}, 50^{th}, 75^{th}$ percentiles
- Quintiles: by 0.2
- In general the pth sample quantile x_p is the value such that a fraction p of the observations fall below that value
- p^{th} population quantile: value x such that the probability that $X \leq x$ is p

Spread or Variability

- Interquartile range: Q_1 to Q_3
  Interval containing $\frac{1}{2}$ of the subjects
  Meaningful for any continuous distribution

- Other quantile intervals

- Variance (for symmetric distributions): averaged squared difference between a randomly chosen observation and the mean of all observations

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  The -1 is there to increase our estimate to compensate for our estimating the center of mass from the data instead of knowing the population mean.[b]

- Standard deviation: s — $\sqrt{}$ of variance
  - $\sqrt{}$ of average squared difference of an observation from the mean
  - can be defined in terms of proportion of sample population within $\pm$ 1 SD of the mean **if the population is normal**

- SD and variance are not useful for very asymmetric data, e.g. "the mean hospital cost was \$10000 $\pm$ \$15000"

- range: not recommended because range $\uparrow$ as $n \uparrow$ and is dominated by a single outlier

- coefficient of variation: not recommended (depends too much on how close the mean is to zero)

Easily calculated in R. For example:

```
> sd(yA)
[1] 4.234934
> sd(yB)
[1] 5.151699

> quantile(yA,prob=c(.25,.75))
   25%    75%
16.850 20.725
> quantile(yB,prob=c(.25,.75))
   25%    75%
24.550 26.825
```

Hypothesis Testing Via Randomization

$\mathbf{H}_0$: No difference in trt A and trt B means.

$\mathbf{H}_1$: A difference in treatment means exists.

A statistical hypothesis test evaluates the plausibility of H_0 in light of the data/sample.

To answer above, we need to compare

$$|\bar{y}_B - \bar{y}_A| = 5.93 \quad \text{observed difference}$$

To values of $|\bar{y}_B - \bar{y}_A|$ that could have been observed if H_0 were true.

Hypothetical values of $|\bar{y}_B - \bar{y}_A|$ under H_0 are called the null distribution.

$$\text{Let } g(Y_A, Y_B) = g\{(Y_{1,A}, \ldots, Y_{6,A}), (Y_{1,B}, \ldots, Y_{6,B})\} = \left|\bar{Y}_B - \bar{Y}_A\right|$$

This is a function of the outcome of an experiment called the test statistic.

Idea:

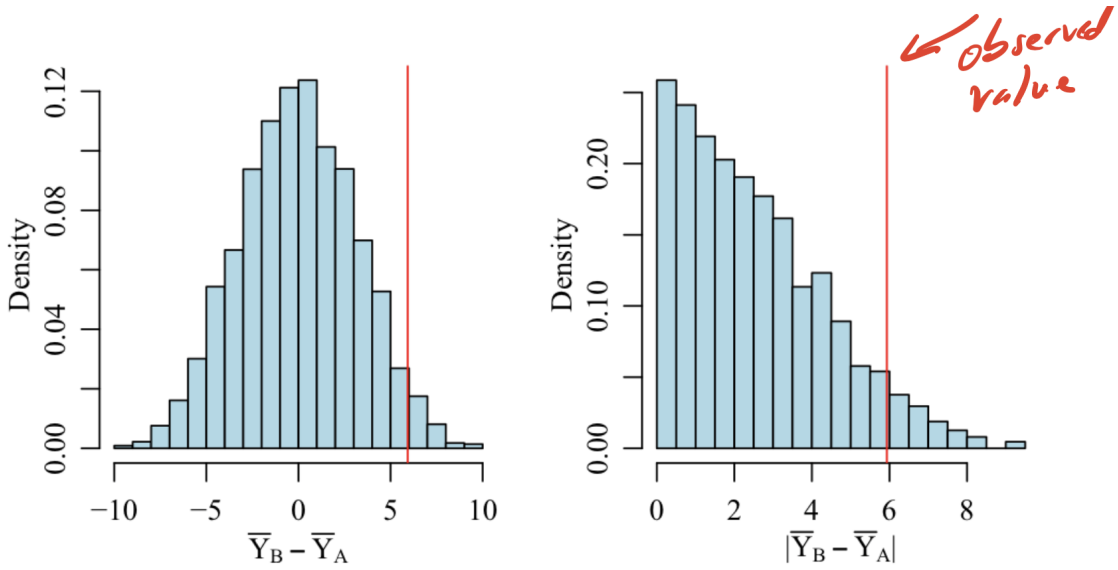
To consider what types of outcomes we would see in universes where H_0 is true, compute $g(Y_A, Y_B)$ under every possible assignment assuming H_0 is true.

Under randomization (H_0) to trt A or trt B, there are

$$\frac{12!}{6!\ 6!} = \binom{12}{6} = 924$$

equally likely ways treatments could have been assigned.

$$\Rightarrow \{g_1, g_2, \ldots, g_{924}\}$$



Randomization Null Distribution

Comparing Sample to Null Distribution

$$P\left(g(Y_A, Y_B) \geq 5.93 \mid H_0\right) = 0.056 \quad \text{(kind of unlikely)}$$

This calculation is called the p-value: "the prob., under H_0, of obtaining a result as or more extreme than the observed result."

Basic idea

$$\text{small p-value} \quad \rightarrow \quad \text{evidence against } H_0$$

$$\text{large p-value} \quad \rightarrow \quad \text{no evidence against } H_0$$

Approximating a Randomization Distribution

Enumerating all $\binom{n_A+n_B}{n_A}$ possible treatment assignments can be computationally costly.

Instead can do:

1) randomly simulate a treatment assignment from the population of possible treatment assignments
2) compute test statistic given the simulated assignment under H_0

Empirical Distribution of $\{\xi_1, \xi_2, \ldots, \xi_S\}$ approximates H_0

$$\frac{\#(gs \geq g_{\text{obs}})}{S} \approx P\left(g(Y_A, Y_B) \geq g_{\text{obs}} \mid H_0\right)$$

Here is some R-code:

```
y<-c(26.9,11.4,26.6,23.7,25.3,28.5,14.2,17.9,16.5,21.1,24.3,19.6)
x<-c("B", "A", "B", "A", "B", "B", "B", "A", "A", "A", "B", "A")

g.null<-real()
for(s in 1:10000)
{
  xsim<-sample(x)
  g.null[s]<- abs( mean(y[xsim=="B"]) - mean(y[xsim=="A"] ) )
}
```

Note:

Hypothesis testing is sensitive to the alternative hypothesis H_1.

For example, could use t-test (sensitive to mean differences)(will see next) or Kolmogorov-Smirnov (sensitive to dist. differences) test as two different test statistics. Can get different conclusions.

Basics of Decision Making

Decisions on Null Hypothesis	States of Nature	
	Null Hypothesis Is True	Null Hypothesis Is False
Fail to reject H_0	Correct decision Probability $= 1 - \alpha$	Type II error Probability $= \beta$
Reject H_0	Type I error Probability $= \alpha$ (α is called the significance level)	Correct decision Probability $= 1 - \beta$ ($1 - \beta$ is called the power of the test)

Decision procedure

1. Compute p-value
2. Reject H_0 if p-value is small - typical to set this as α

This procedure is called a level α test.
Controls pre-experimental probability of a Type I error or for a series of (independent) experiments, controls the type I error rate.

$$P(\text{Type I error} \mid H_0) = P(\text{Reject } H_0 \mid H_0) = P(\text{p-value} \leq \alpha \mid H_0) = \alpha$$

(Some hidden theory about the distribution of the p-value under H_0)

Single Experiment Interpretation

If you use a level α test when H_0 is true, the probability is α that you will erroneously reject H_0.

Many Experiments Interpretation

If level α tests are used in a large population of experiments, then H_0 will be declared false in $(100 \times \alpha)\%$ of them in which H_0 is true.

Note:

$$P(\text{Reject } H_0 \mid H_0 \text{ false}) = 1 - \beta = \text{power}$$

need to be more specific on what this means to calculate power.

[will see more later]

- α and β cannot be simultaneously minimized, so there is a trade-off between α and β.

- Usually, fix α and try to minimize β (or equivalently maximize $1-\beta$).

Trade-off Between Type-I and Type-II Errors

- Suppose we have only one data point z in hand and we know $z \sim N(\mu, 1)$.
- We want to test H_0: $\mu = 0$ against H_1: $\mu = 3$.
- A natural test is to reject H_0 if z is large, e.g., $z > c$ for some $c > 0$.
- $\alpha = P(z > c|\mu = 0) = 1 - \Phi(c)$, which is a decreasing function of c, where $\Phi(\cdot)$ is the distribution function of the standard normal.
- $\beta = P(z \leq c|\mu = 3) = P(z - 3 \leq c - 3) = \Phi(c - 3)$ which is an increasing function of c. [figure here]
- Fix $\alpha = 0.05$, then c is chosen such that $1 - \Phi(c) = 0.05$, i.e., $c = 1.645$.
- Now, $\pi = 1 - \Phi(1.645 - 3)$.

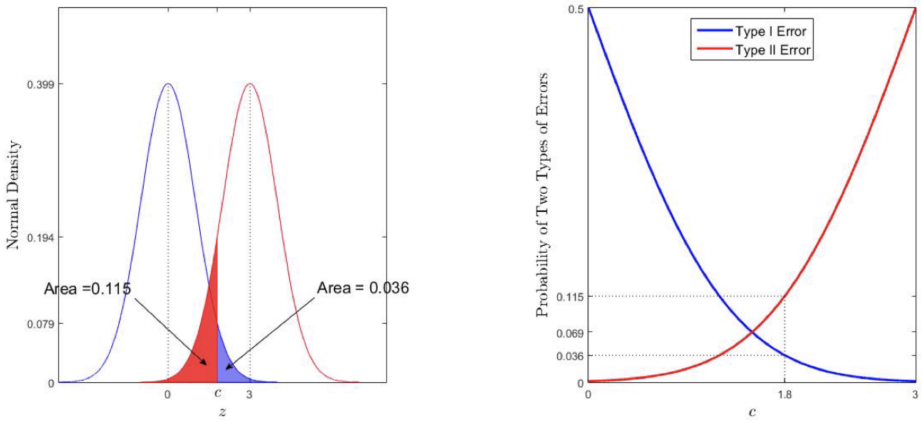


Figure: Trade-Off Between Type-I and Type-II Errors

- The left panel illustrates α and β when $c = 1.8$, and the right panel illustrates these probabilities as a function of c.