Introduction to Statistics

Dr. Lauren Cappiello

Contents

For the Instructor

Thanks for checking out my Introduction to Statistics text! Sections are designed to be short, easy-to-read introductions to each concept. Some of the more conceptual sections do not have section exercises, but I am working on adding exercises wherever it seems appropriate. The topics and course ordering reflect the department syllabus for Introduction to Statistics at Sacramento State. I am sure there are topics we've left out, but there are only so many things one can cover in 15 weeks.

Each Chapter is designed to take approximately two weeks of class time. In an ideal world, I would cover all eight chapters in a 15 week semester. However, with assessment, activities, student questions, holidays, etc., I usually get through the first six or seven. Sometimes I introduce regression at the end of Chapter 2. Rarely do I get to ANOVA. Despite the time constraints, I am working on including additional topics.

This text is a work in progress and gets updated every semester that I each Introduction to Statistics (which is very nearly every semester) and sometimes during winter and summer breaks. Currently, I am working on

- overhauling the entire thing to remove some of the examples borrowed from OpenIntro (another great resource) from when this was just the typed version of my course notes.
- adding section exercises and additional topics.

Please feel free to reach out to me with any questions, comments, or concerns by emailing me at cappiello@csus.edu

Chapter 1

Introduction to Data

1.1 Module Overview

What is statistics? There are two ways to think about this:

1. Facts and data, organized or summarized in such a way that they provide useful information about something.
2. The science of analyzing, organizing, and summarizing data.

As a field, Statistics provides tools for scientists, practitioners, and laypeople to better understand data. You may find yourself using knowledge from this course in a research lab, while reading a research report, or even while watching the news!

Chapter Learning Objectives/Outcomes

After completing Chapter 1, you will:

1. Understand basic statistical terminology.
2. Produce data using sampling and experimental design techniques.
3. Organize and visualize data using techniques for exploratory data analysis.
4. Identify the shape of a data set.
5. Understand and interpret graphical displays.

This chapter's outcomes correspond to course outcomes (1) organize, summarize, and interpret data in tabular, graphical, and pictorial formats and (2) organize and interpret bivariate data and learn simple linear regression and correlation.

1.2 Statistics Terminology

There are two ways to think about statistics:

1. **Descriptive statistics** are methods for *describing* information.

For example, 66% of eligible voters voted in the 2020 presidential election (the highest turnout since 1900!).

2. **Inferential statistics** are methods for *drawing inference* (making decisions about something we are uncertain about).

For example, a poll suggests that 75% of voters will select a Candidate A. People haven't voted yet, so we don't know what will happen, but we could reasonably conclude that Candidate A will win the election.

Data is factual information. We collect data from a **population**, the collection of all individuals or items a researcher is interested in.

- Collecting data from an entire population is called a **census**.
  - This is complicated and expensive! There's a reason the United States only does a census every 10 years.
- We can also take a **sample**, a subset of the population we get data from.
  - If you think of the population as a pie, the sample is a small slice. Whether it's a pumpkin pie, a cherry pie, or a savory pie, the small slice will tell you that. We don't need to eat the entire pie to learn a lot about it!

Data are often organized in what we call a **data matrix**. If you've ever seen data in a spreadsheet, that's a data matrix!

Age

Gender

Smoker

Marital Status

Person 1

45

Male

yes

married

Person 2

23

Female

no

single

Person 3

36

Other

no

married

Person 4

29

Female

no

single

Each row (horizontal) represents one **observation** (also called **observational units**, **cases**, or **subjects**). These are the individuals or items in the sample.

Each column (vertical) represents a **variable**, the characteristic or thing being measured. Think of variables as measurements that can *vary* from one observation to the next.

There are two types of variable:

Numeric or **quantitative** variables take *numeric* values AND it is sensible to do math with those values.

Discrete numeric variables take numeric values with jumps. Typically, this means they can only take whole number values. These are often counts of something - for example, counting the number of pets you have.

Continuous numeric variables take values "between the jumps." Typically, this means they can take decimal values.

Categorical or **qualitative** variables take values that are *categories*.

The "Does it make sense?" Test

- Sometimes, categories can be represented by numbers. Ask yourself if it makes sense to do math with those numbers. If it doesn't make sense, it's probably a categorical variable. (Ex: zip codes)
- If you're unsure whether a variable is discrete or continuous, pick a number with some decimal places - like 1.83 - and ask yourself if that value makes sense. If it doesn't, it's probably discrete. (Ex: number of siblings)

Section Exercises

The following table shows part of the data matrix from a Stat 1 course survey.

Age

Year in college	**What is your major?**	**Units this semester**
1	19	
Sophomore	Health Sciences	15
2	19	
Sophomore	Business	15
3	19	
Sophomore	Undecided	14
⋮	⋮	⋮
29	21	
Junior	Business	15

1. What does each row of the data matrix represent?

2. What does each column of the data matrix represent?
3. Indicate whether each variable is discrete numeric, continuous numeric, or categorical.

Dig Deeper

Read the article, Here's Why an Accurate Census Count Is So Important from the New York Times. (If you can't access the article, try a Google search for "why an accurate census count is important.") Take a moment to write down your thoughts on the relationship between how we collect data (for example - the questions asked in the census) and the power data has over people's lives. As researchers, scientists, and consumers of media, what are some reasons this is important to think about?

1.3 Sampling and Design

1.3.1 Statistical Sampling

How do we get samples? We want a sample that represents our population. **Representative samples** reflect the relevant characteristics of our population.

In general, we get representative samples by selecting our samples *at random* and with an adequate sample size.

A non-representative sample is said to be **biased**. For example, if we used a sample of chihuahuas to represent all dogs, we probably wouldn't get very good information; that sample would be *biased*.

These can be a result of **convenience sampling**, choosing a sample based on ease.

In our daily lives, common sources of biases are **anecdotal evidence** and **availability bias**. Anecdotal evidence is data based on personal experience or observation. Typically this consists of only one or two observations and is NOT representative of the population.

> *Example*: anecdotal evidence. A friend tells you their grandpa smoked a pack of cigarettes a day and lived to be 100. Does this mean that cigarettes will help you live to 100? no!

Availability bias is your brain's tendency to think that examples of things that come readily to mind are more representative than is actually the case.

> *Example*: availability bias. Shark attacks. Shark attacks are actually extremely uncommon, but the media tends to report on extreme anecdotes, making us more prone to this kind of bias!

We avoid bias by taking random samples. One type of random sample is a **simple random sample**. We can think of this as "raffle sampling," like drawing names out of a hat. Each case (or each possible sample) has an equal chance of being selected. Knowing that A is selected doesn't tell us anything about whether B is selected. Instead of literally drawing from a hat, we usually use a **random number generator** from a computer.

R: Random Number Generation

To generate a random whole number using R, we can use the `sample` command. We use the `sample` command like `sample(minimum:maximum, size = n)`, replacing `minimum` with the minimum value (often the number 1), `maximum` with the maximum value, and `n` with the sample size.

The following command takes a random sample of size 1 from the values 1 through 10 (1, 2, 3, 4, 5, 6, 7, 8, 9, 10):

```
sample(1:10, size = 1)
```

```
## [1] 9
```

1.3.2 Experimental Design

When we do research, we have two options:

Conduct an **experiment**, where researchers assign treatments to cases.

Treatments are experimental conditions.

In an experiment, cases may also be called **experimental units** (items or individuals on which the experiment is performed).

Conduct an **observational study**, where no conditions are assigned. These are often done for ethical reasons, like examining the impacts of smoking cigarettes.

Experiments allow us to infer causation. Observational studies do not.

Experimental design principles:

Control: two or more treatments are compared.

Randomization: experimental units are assigned to treatment groups (usually and preferably at random).

Replication: a large enough sample size is used to test each treatment many times (on many different experimental units).

Blocking: if variables other than treatment are likely to have an impact on study outcome, we use blocks.

For example, I might separate patients in a medical study into "high risk" and "low risk" blocks. I would randomly assign all of the high risk patients to a treatment and then randomly assign all of the low risk patients to a treatment. This helps ensure an even distribution of high/low risk patients in each treatment group.

An experiment without blocking has a completely randomized design; an experiment with blocking has a randomized block design.

In an experimental setting, we talk about

Response variable: the characteristic of the experimental outcome being measured or observed.

Factor: a variable whose impact on the response variable is of interest in the experiment.

Levels: the possible values of a factor.

Treatments: experimental conditions (based on combinations of factor levels).

In human subjects research, we do a little extra work:

If subjects do not know what treatment group they are in, the study is called blind.

We use a placebo (fake treatment) to achieve this.

If neither the subjects nor the researchers who interact with them know the treatment group, it is called double blind.

This helps avoid bias caused by placebo effect, doctor's expectations for outcome, etc.!

Section Exercises

1. A study published in 2009 sought to examine whether supplementing with chia seeds contributed to weight loss. Researchers recruited 76 individuals and randomly assigned them into either a treatment group or a control group. The treatment group was given a set quantity of daily chia seeds; the control group was given a placebo. At the end of the 12-week study, they found no difference in average weight lost between the treatment and control group.

    a. Is this an observational study or an experiment? Explain.

    b. Identify the (i) cases and (ii) response variable.

1.4 Frequency Distributions

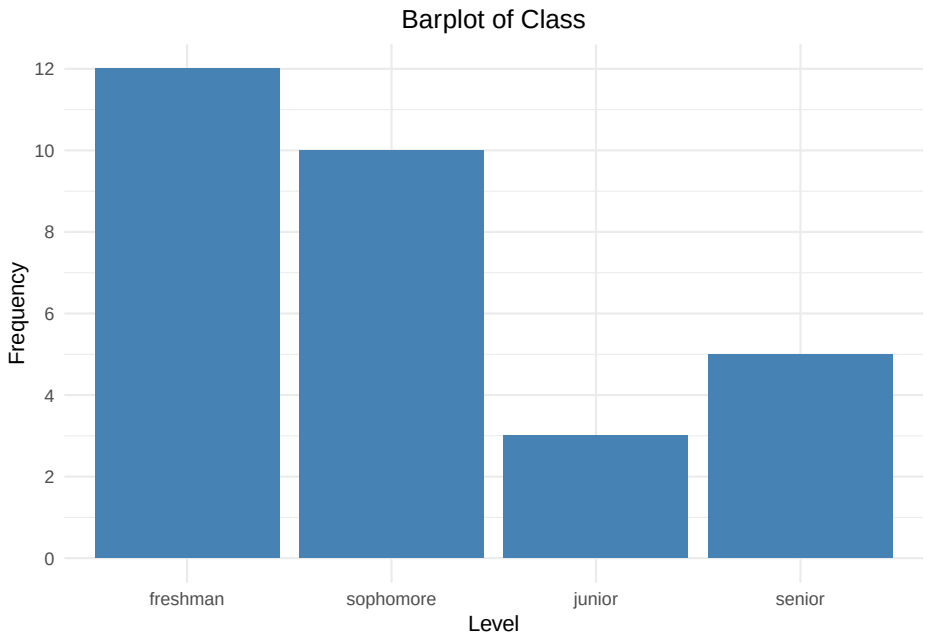
1.4.1 Qualitative Variables

Frequency (**count**): the number of times a particular value occurs.

A **frequency distribution** lists each distinct value with its frequency.

Class	Frequency
freshman	12
sophomore	10
junior	3
senior	5

A **bar plot** is a graphical representation of a frequency distribution. Each bar's height is based on the frequency of the corresponding category.



The bar plot above shows the class level breakdown for students in an Introductory Statistics course. Take a moment to notice how the bars match up with the frequency distribution above.

Relative frequency is the ratio of the frequency to the total number of observations.

$$\text{relative frequency} = \frac{\text{frequency}}{\text{number of observations}}$$

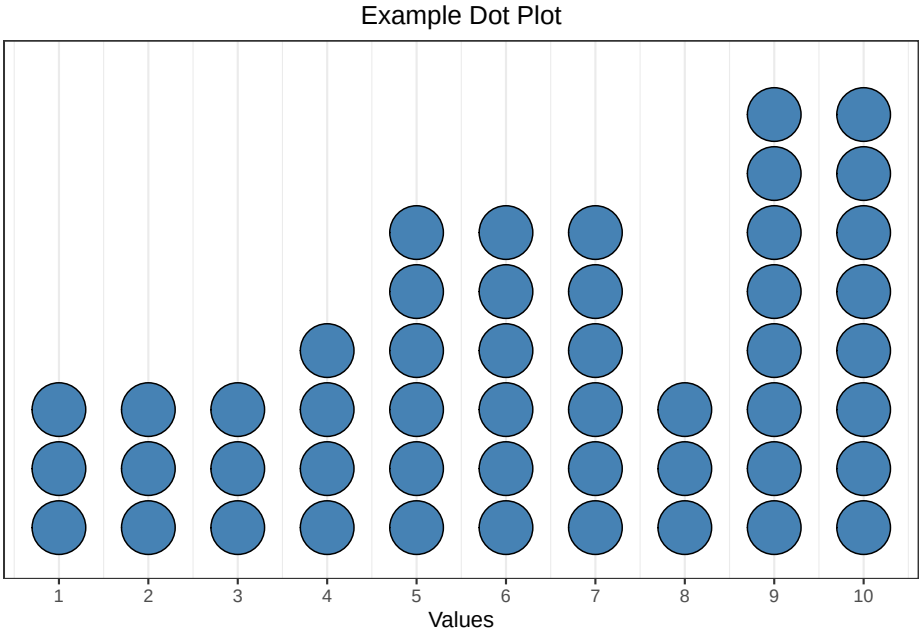
This is also called the **proportion**. The **percentage** can be obtained by multiplying the proportion by 100.

A **relative frequency distribution** lists each distinct value with its relative frequency.

Class	Frequency	Relative Frequency	Percent
freshman	12	$12/30 = 0.4$	40%
sophomore	10	$10/30 \approx 0.3333$	33.33%
junior	3	$3/30 = 0.1$	10%
senior	5	$5/30 \approx 0.1667$	16.67%

1.4.2 Quantitative Variables

We can also apply this concept to numeric data. A **dot plot** is one graphical representation of this. A dot plot shows a number line with dots drawn above the line. Each dot represents a single point.



For example, the dot plot above shows a sample where the value 1 appears three times, the value 5 appears six times, etc.

We would also like to be able to visualize larger, more complex data sets. This is hard to do using a do plot! Instead, we can do this using **bins**, which group numeric data into equal-width consecutive intervals.

Example: A random sample of weights (in lbs) from 12 cats:

6.2 11.6 7.2 17.1 15.1 8.4 7.7 13.9 21.0 5.5 9.1 7.3

The **minimum** (smallest value) is 5.5 and the **maximum** (largest value) is 21. There are lots of ways to break these into "bins," but what about...

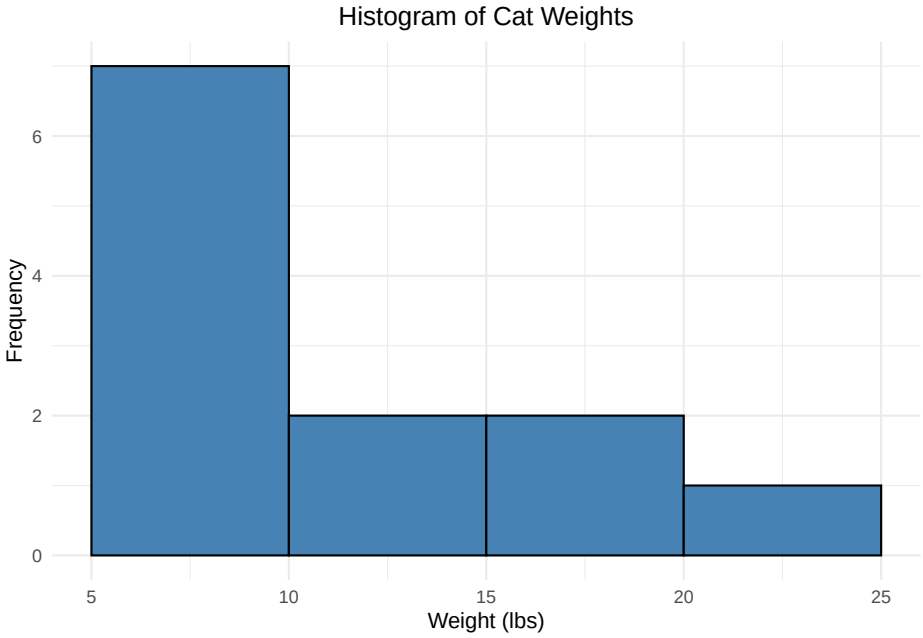
- 5 - 10
- 10 - 15
- 15 - 20
- 20 - 25

Each bin has an equal width of 5, but if we had a cat with a weight of exactly 15 lbs, would we use the second or third bin?? It's unclear. To make this clear, we need there to be no overlap. Instead, we could use:

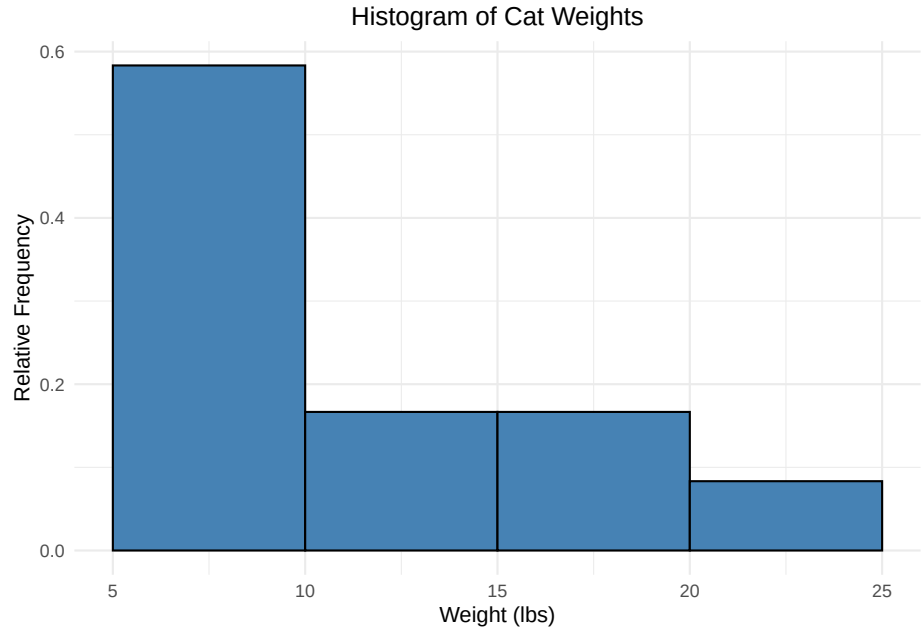
Weight	Count
5 - <10	7
10 - <15	2
15 - <20	2
20 - <25	1

Now, a cat with a weight of 15.0 lbs would be placed in the third bin (but not the second).

We will visualize this using a **histogram**, which is a lot like a bar plot but for numeric data:



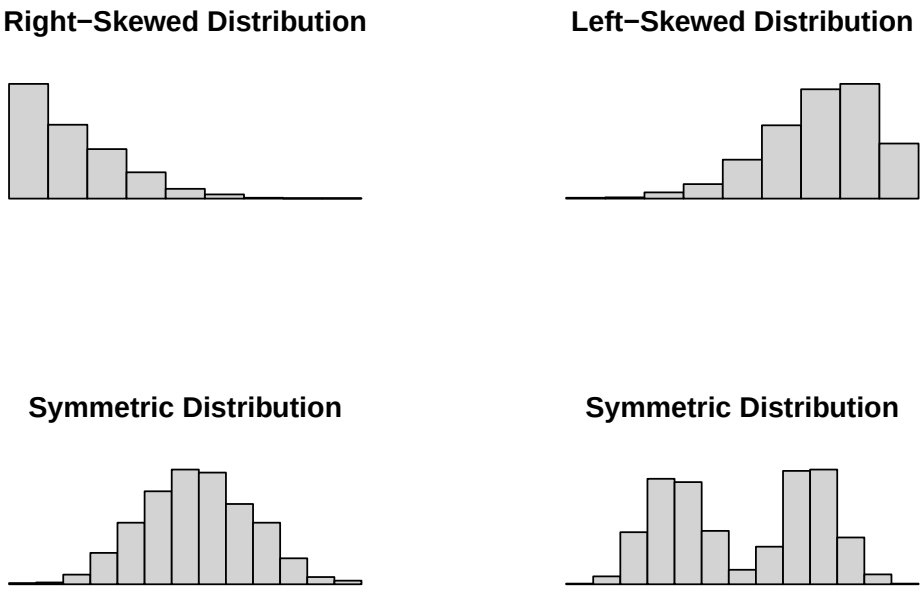
This is what we call a **frequency histogram** because each bar height reflects the frequency of that bin. We can also create a **relative frequency histogram** which displays the relative frequency instead of the frequency:



Notice that these last two histograms look the same *except for the numbers on the*

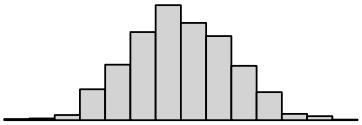
vertical axis! This gives us insight into the shape of the data **distribution**, literally how the values are distributed across the bins. The part of the distribution that "trails off" to one or both sides is called a **tail** of the distribution.

When a histogram trails off to one side, we say it is **skewed** (right-skewed if it trails off to the right, left-skewed if it trails off to the left). Data sets with roughly equal tails are **symmetric**.

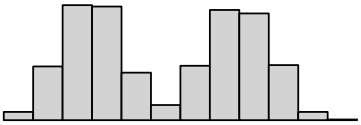


We can also use a histogram to identify **modes**. For numeric data, especially continuous variables, we think of modes as *prominent peaks*.

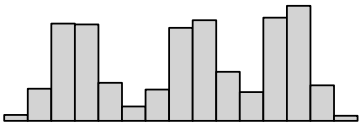
Unimodal



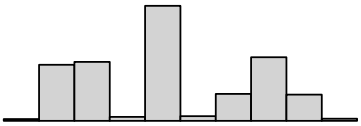
Bimodal



Multimodal



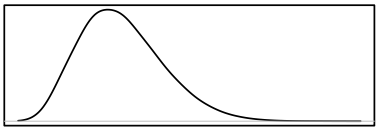
Multimodal



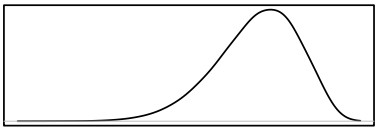
- **Unimodal**: one prominent peak.
- **Bimodal**: two prominent peaks.
- **Multimodal**: three or more prominent peaks.

Finally, we can also "smooth out" these histograms and use a smooth curve to examine the shape of the distribution. Below are the smooth curve versions of the distributions shown in the four histograms used to demonstrate skew and symmetry.

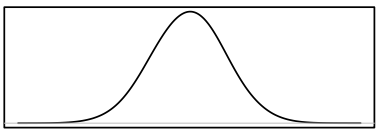
Right−Skewed Distribution



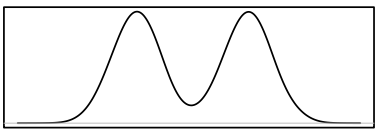
Left−Skewed Distribution



Symmetric Distribution



Symmetric Distribution



Chapter 2

Descriptive Measures

2.1 Chapter Overview

In the previous chapter, we thought about descriptive statistics using tables and graphs. Next, we summarize data by computing numbers. Some of these numbers you may already be familiar with, such as averages and percentiles. Numbers used to describe data are called *descriptive measures*.

Chapter Learning Objectives/Outcomes

After completing Chapter 2, you will be able to:

1. Calculate and interpret measures of center.
2. Calculate and interpret measures of variation.
3. Find and interpret measures of position.
4. Summarize data using boxplots.

This chapter's outcomes correspond to course outcomes (1) organize, summarize, and interpret data in tabular, graphical, and pictorial formats, (2) organize and interpret bivariate data and learn simple linear regression and correlation, and (6) apply statistical inference techniques of parameter estimation such as point estimation and confidence interval estimation.

2.2 Measures of Central Tendency

One research question we might ask is : what values are most common or most likely?

Mode: the most commonly occurring value. We can use this for numeric variables, but typically we use the mode when talking about categorical data.

Mean: this is what we usually think of as the "average." Denoted $\bar{x}$. Add up all of the values and divide by the number of observations (n):

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n}$$

where x_i denotes the ith observation and $\sum_{i=1}^{n}$ is the sum of all observations from 1 through n. This is called *summation notation*.

Median: the middle number when the data are ordered from smallest to largest.

- If there are an odd number of observations, this will be the number in the middle:
  {1, 3, **7**, 9, 9} has median 7
- If there are an even number of observations, there will be two numbers in the middle. The median will be their average.
  {1, 2, **4, 7**, 9, 9} has median $\frac{4+7}{2} = 5.5$

The mean is sensitive to extreme values and skew. The median is not!

x: 1, 3, 7, 9, 9

y: 1, 3, 7, 9, 45

Median

median = 7

median = 7

Mean

$\bar{x} = \frac{29}{5} = 5.8$

$\bar{y} = \frac{65}{5} = 13$

Notice how changing that 9 out for a 45 changes the *mean* a lot! But the *median* is 7 for both x and y.

Because the median is not affected by extreme observations or skew, we say it is a **resistant measure** or that it is **robust**.

Which measure should we use?

- Mean: symmetric, numeric data
- Median: skewed, numeric data
- Mode: categorical data

Note: If the mean and median are roughly equal, it is reasonable to assume the distribution is roughly symmetric.

2.3 Measures of Variability

How much do the data vary?

Should we care? Yes! The more variable the data, the harder it is to be confident in our measures of center!

If you live in a place with extremely variable weather, it is going to be much harder to be confident in how to dress for tomorrow's weather... but if you live in a place where the weather is always the same, it's much easier to be confident in what you plan to wear.

We want to think about how far observations are from the measure of center.

One easy way to think about variability is the **range** of the data:

$$\text{range} = \text{maximum} - \text{minimum}$$

This is quick and convenient, but it is *extremely* sensitive to outliers! It also takes into account only two of the observations - we would prefer a measure of variability that takes into account *all* the observations.

Deviation is the distance of an observation from the mean: $x - \bar{x}$. If we want to think about how far - on average - a typical observation is from the center, our intuition might be to take the average deviance... but it turns out that summing up the deviances will *always* result in 0! Conceptually, this is because the stuff below the mean (negative numbers) and the stuff above the mean (positive numbers) end up canceling each other out until we end up at 0. (If you are interested, *Appendix A: Average Deviance* has a mathematical proof of this using some relatively straightforward algebra.)

One way to deal with this is to make all of the numbers positive, which we accomplish by squaring the deviance.

	Deviance	Squared Deviance
x	$x - \bar{x}$	$(x - \bar{x})^2$
2	-1.2	1.44
5	1.8	3.24
3	-0.2	0.04
4	0.8	0.64
2	-1.2	1.44
$\bar{x} = 3.2$	Total = 0	Total = 6.8

Variance (denoted s^2) is the average squared distance from the mean:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

where n is the sample size. Notice that we divide by $n-1$ and NOT by n. There are some mathematical reasons why we do this, but the short version is that it'll be a better estimate when we talk about inference.

Finally, we come to **standard deviation** (denoted s).

$$s = \sqrt{s^2}$$

The standard deviation is the square root of the variance. We say that a "typical" observation is within about one standard deviation of the mean (between $\bar{x} - s$ and $\bar{x} + s$).

We will think about one more measure of variability, the interquartile range, in the next section.

2.4 Measures of Position

The **interquartile range** (**IQR**) represents the middle 50% of the data.

Recall that the *median* cut the data in half: 50% of the data is below and 50% is above the median. This is also called the **50th percentile**. The p**th percentile** is the value for which p% of the data is below it.

To get the middle 50%, we will split the data into four parts:

1	2	3	4
25%	25%	25%	25%

The 25th and 75th percentiles, along with the median, divide the data into four parts. We call these three measurements the **quartiles**:

- **Q1**, the first quartile, is the median of the lower 50% of the data.
- **Q2**, the second quartile, is the median.
- **Q3**, the third quartile, is the median of the upper 50% of the data.

  Example: Consider {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}

  - Cutting the data in half: {1, 2, 3, 4, 5 | 6, 7, 8, 9, 10}, the median (Q2) is $\frac{5+6}{2} = 5.5$.
  - Q1 is the median of {1, 2, 3, 4, 5}, or 3
  - Q3 is the median of {6, 7, 8, 9, 10}, or 8

Note: this is a "quick and dirty" way of finding quartiles. A computer will give a more exact result.

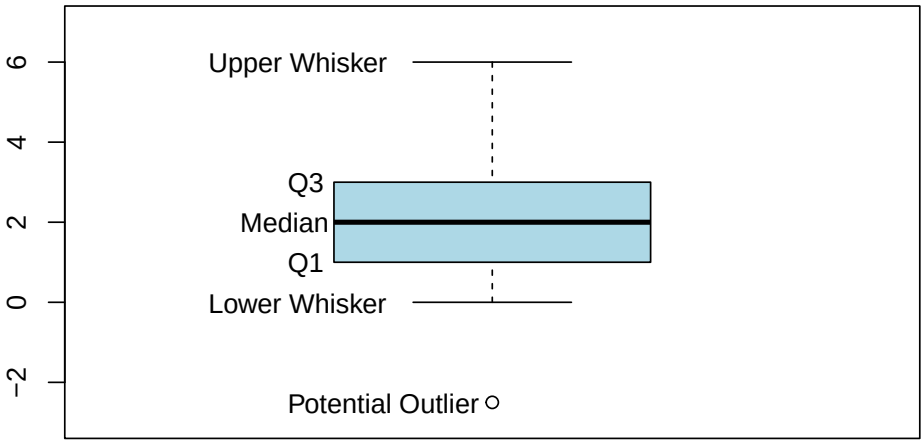
Then the interquartile range is

$$\text{IQR} = \text{Q3} - \text{Q1}$$

This is another measure of variability and is resistant to extreme values. In general, we prefer the mean and standard deviation when the data are symmetric and we prefer the median and IQR when the data are skewed.

2.4.1 Box Plots

Our measures of position are the foundation for constructing what we call a box plot, which summarizes the data with 5 statistics plus extreme observations:



Drawing a box plot:

1. Draw the vertical axis to include all possible values in the data.
2. Draw a horizontal line at the median, at Q1, and at Q3. Use these to form a box.
3. Draw the **whiskers**. The whiskers' upper limit is $Q3 + 1.5 \times IQR$ and the lower limit is $Q1 - 1.5 \times IQR$. The actual whiskers are then drawn *at the next closest data points within the limits.*
4. Any points outside the whisker limits are included as individual points. These are **potential outliers**.

(Potential) outliers can help us...

- examine skew (outliers in the negative direction suggest left skew; outliers in the positive direction suggest right skew).
- identify issues with data collection or entry, especially if the value of the outliers doesn't make sense.

As with most things in this text, we won't draw a lot of boxplots by hand. However, understanding how they are drawn will help us understand how to interpret them!

2.5 Descriptive Measures for Populations

So far, we've thought about calculating various descriptive statistics from a sample, but our long-term goal is to estimate descriptive information about a population. At the population level, these values are called **parameters**.

When we find a measure of center, spread, or position, we use a sample to calculate a single value. These single values are called **point estimates** and they are used to *estimate* the corresponding population parameter. For example, we use $\bar{x}$ to estimate the population mean, denoted μ (Greek letter "mu") and s to estimate the population standard deviation, denoted σ (Greek letter "sigma").

Point Estimate	Parameter
sample mean: $\bar{x}$	population mean: μ
sample standard deviation: s	population standard deviation: σ

. . . and so on and so forth. For each quantity we calculate from a sample (point estimate), there is some corresponding unknown population level value (parameter) that we wish to estimate.

We will discuss this in more detail when we discuss Random Variables and Statistical Inference.

Chapter 3

Probability Concepts

3.1 Chapter Overview

In previous chapters, we discussed ways to describe variables and the relationships between them. From here, we want to start asking inferential statistics questions like "If my sample mean is 10, how likely is it that the population mean is actually 11?" Probability is going to start us on this path.

Probability theory is the science of uncertainty and it is really interesting! But it can also be quite challenging. I try to frame probability around things most of us can do at home: flipping a coin, rolling a die, drawing from a deck of cards. You certainly don't need any of these things to get through this chapter, but you may find it helpful to have a coin/die/deck of cards on hand as you read through the examples.

Take your time running practice problems and going through the examples, using a tactile approach like sorting through your deck of cards whenever it seems helpful.

Chapter Learning Objectives/Outcomes

1. Find and interpret probabilities for equally likely events.
2. Find and interpret probabilities for events that are not equally likely.
3. Find and interpret joint and marginal probabilities.
4. Find and interpret conditional probabilities.
5. Use the multiplication rule and independence to calculate probabilities.

This chapter's outcomes correspond to course outcome (3) understand the basic rules of probability.

3.2 Experiments, Sample Spaces, and Events

Probability is the science of uncertainty. When we run an experiment, we are unsure of what the outcome will be. Because of this uncertainty, we say an experiment is a **random process**.

The probability of an event is the proportion of times is would occur if the experiment were run infinitely many times. For a collection of *equally likely events*, this looks like:

$$\text{probability of event} = \frac{\text{number of ways event can occur}}{\text{number of possible (unique) outcomes}}$$

An **event** is some specified possible outcome (or collection of outcomes) we are interested in observing.

> *Example*: If you want to roll a 6 on a six-sided die, there are six possible outcomes $\{1, 2, 3, 4, 5, 6\}$. In general, we assume that each die face is equally likely to appear on a single roll of the die, that is, that the die is *fair*. So the probability of rolling a 6 is
>
> $$\frac{\text{number of ways to roll a 6}}{\text{number of possible rolls}} = \frac{1}{6}$$
>
> *Example*: We can extend this to a collection of events, say the probability of rolling a 5 or a 6:
>
> $$\frac{\text{number of ways to roll a 5 or 6}}{\text{number of possible rolls}} = \frac{2}{6}$$

The collection of all possible outcomes is called a **sample space**, denoted S. For the six-sided die, $S = \{1, 2, 3, 4, 5, 6\}$.

To simplify our writing, we use **probability notation**:

- Events are assigned capital letters.
- $P(A)$ denotes the probability of event A.
- Sometimes we will also shorten simple events to just a number. For example, $P(1)$ might represent "the probability of rolling a 1."

We can estimate probabilities from a sample using a frequency distribution.

> *Example*: Consider the following frequncy distribution from section 1.6

Class	Frequency
freshman	12
sophomore	10
junior	3
senior	5

If a student is selected *at random* (meaning each student is equally likely to be selected), the probability of selecting a sophomore is

$$\text{probability of sophomore} = \frac{\text{number of ways to select a sophomore}}{\text{total number of students}} = \frac{10}{30} \approx 0.3333$$

The probability of selecting a *junior or a senior* is

$$\frac{\text{number of ways to select a junior or senior}}{\text{total number of students}} = \frac{3+5}{30} = \frac{8}{30} \approx 0.2667$$

Using probability notation, we might let A be the event we selected a junior and B be the event we selected a senior. Then

$$P(A \text{ or } B) = 0.2667$$

3.3 Probability Distributions

Two outcomes are **disjoint** or **mutually exclusive** if they cannot both happen (at the same time). Think back to how we developed bins for histograms - the bins need to be nonoverlapping - this is the same idea!

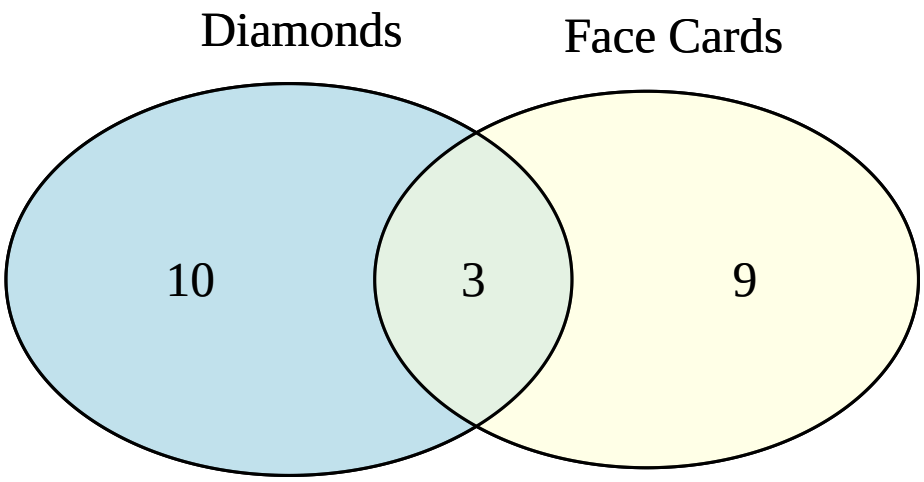
Example: If I roll a six-sided die one time, rolling a 5 and rolling a 6 are disjoint. I can get a 5 *or* a 6, but not both on the same roll.

Example: If I select a student, they can be a freshman *or* a sophomore, but that student cannot be both a freshman and a sophomore at the same time.

The outcome must be one event or the other (it cannot be both at the same time).

3.3.1 Venn Diagrams

Venn Diagrams show events as circles. The circles overlap where events share common outcomes.



When a Venn Diagram has *no overlap* the events are mutually exclusive. This Venn Diagram shows the event "Draw a Diamond" and the event "Draw a Face Card." There are 13 diamonds and 12 face cards in a deck. In this case, the events are *not* mutually exclusive: it's possible to draw both a diamond and a face card at the same time: the Jack of Diamonds, Queen of Diamonds, and King of Diamonds.

For quick reference, an image of a full 52-card deck is linked below. The "face cards" are the J, Q, and K. Each row represents a "suit." From top to bottom, the suits are clubs, spades, hearts, and diamonds. Cards can be either red (hearts and diamonds) or black (spades and clubs).

Click here for a graphic of a standard 52 card deck.

On your own: Consider events

- A: "Draw a spade"
- B: "Draw a queen"
- C: "Draw a red"

Which of these events are mutually exclusive?

3.3.2 Probability Axioms

A **probability distribution** lists all possible disjoint outcomes (think: all possible values of a variable) and their associated probabilities. This can be in the form of a table

Roll of a six-sided die	1	2	3	4	5	6
Probability	1/6	1/6	1/6	1/6	1/6	1/6

(note that we could visualize this with a bar plot!) or an equation, which we will discuss in a later chapter.

The **probability axioms** are requirements for a valid probability distribution. They are:

1. All listed outcomes must be disjoint.
2. Each probability must be between 0 and 1.
3. The probabilities must sum to 1.

Note that #2 is true for ALL probabilities. If you ever calculate a probability and get a negative number or a number greater than 1, you know something went wrong!

Example: Use the probability axioms to check whether the following tables are probability distributions.

A)

X	{1 or 2}	{3 or 4}	{5 or 6}
P(X)	1/3	1/3	1/3

Each axiom is satisfied, so this is a valid probability distribution.

B)

Y	{1 or 2}	{2 or 3}	{3 or 4}	{5 or 6}
P(Y)	1/3	1/3	1/3	-1/3

In this case, the outcomes are not disjoint and one of the probabilities is negative, so this is *not* a valid probability distribution.

3.3.3 Exercises

1. Use the probability axioms to determine whether each of the following is a valid probability distribution:
   A.

x	0	1	2	3
P(x)	0.1	0.2	0.1	0.3

   B.

x	0 or 1	1 or 2	3 or 4	5 or 6
P(x)	0.1	0.2	0.4	0.3

2. Determine whether the following events are mutually exclusive (disjoint).
   a. Your friend studies in the library. You study at home.
   b. You and your study group all earn As on an exam.
   c. You stay out until 3 am. You go to bed at 9 pm.
3. In a group of 24 people, 11 have cats and 13 have dogs. Four of them have both cats and dogs. Sketch a Venn Diagram for these events.

3.4 Rules of Probability

Consider a six-sided die.

$$P(\text{roll a 1 or 2}) = \frac{\text{2 ways}}{\text{6 outcomes}} = \frac{1}{3}.$$

Notice that we get the same result by taking

$$P(\text{roll a 1}) + P(\text{roll a 2}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

It turns out this is widely applicable!

3.4.1 Addition Rules

Addition Rule for Disjoint Outcomes

If A_1 and A_2 are disjoint outcomes, then the probability that one of them occurs is

$$P(A_1 \text{ or } A_2) = P(A_1) + P(A_2).$$

This can also be extended to more than two disjoint outcomes:

$$P(A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_k) = P(A_1) + P(A_2) + \cdots + P(A_k)$$

for k disjoint outcomes.

Now consider a deck of cards. Let A be the event that a card drawn is a diamond and let B be the event it is a face card. (Check back to 3.2 for the Venn Diagram of these events.)

- A: 2♢ 3♢ 4♢ 5♢ 6♢ 7♢ 8♢ 9♢ 10♢ J♢ Q♢ K♢ A♢
- B: J♡ Q♡ K♡ J♣ Q♣ K♣ J♢ Q♢ K♢ J♠ Q♠ K♠

The collection of cards that are diamonds or face cards (or both) is

$A\diamondsuit$ $2\diamondsuit$ $3\diamondsuit$ $4\diamondsuit$ $5\diamondsuit$ $6\diamondsuit$ $7\diamondsuit$ $8\diamondsuit$ $9\diamondsuit$ $10\diamondsuit$ $J\diamondsuit$ $Q\diamondsuit$ $K\diamondsuit$ $J\clubsuit$ $Q\clubsuit$ $K\clubsuit$ $J\heartsuit$ $Q\heartsuit$ $K\heartsuit$ $J\spadesuit$ $Q\spadesuit$ $K\spadesuit$

Looking at these cards, I can see that there are 22 of them, so

$$P(A \text{ or } B) = \frac{22}{52}$$

However, if I try to apply the addition rule for disjoint outcomes, $P(A) = \frac{13}{52}$ and $P(B) = \frac{12}{52}$ and I would get $\frac{13+15}{52} = \frac{25}{52}$, which isn't what we want!

What happened? When I tried to add these, I *double counted* the Jack of Diamonds, Queen of Diamonds, and King of Diamonds (the cards that are in both A and B). To deal with that, I need to subtract off the double count $\frac{13}{52} + \frac{12}{52} - \frac{3}{52}$.

General Addition Rule

For any two events A and B, the probability that *at least* one will occur is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

Notice that when we say "or," we include the situations where A is true, B is true, and the situation where are both A and B are true. This is an *inclusive or*. Basically, if I said "Do you like cats or dogs?" and you said "Yes." because you like cats *and* dogs, that would be a perfectly valid response. I recommend using the inclusive or with your friends any time you want to get out of making a decision.

Also notice that the general addition rule applies to *any* two events, even disjoint events. This is because, for disjoint events, $P(A \text{ and } B) = 0$; it's impossible for both to occur at the same time!

3.4.2 Complements

The **complement** of an event is all of the outcomes in the sample space that are *not* in the event. For an event A, we denote its complement by A^c.

> *Example*: For a single roll of a six-sided die, the sample space is all possible rolls: 1, 2, 3, 4, 5, or 6. If the event A is rolling a 1 or a 2, then the complement of this event, denoted A^c, is rolling a 3, 4, 5, or 6.
>
> We could also write this in probability notation: $S = \{1, 2, 3, 4, 5, 6\}$ and if $A = \{1, 2\}$, then $A^c = \{3, 4, 5, 6\}$.

Property:

$$P(A \text{ or } A^c) = 1$$

Using the addition rule,

$$P(A \text{ or } A^c) = P(A) + P(A^c) = 1.$$

(Make sure you can convince yourself that A and A^c are *always* disjoint.) This is especially useful written as

$$P(A) = 1 - P(A^c).$$

Example: Consider rolling 2 six-sided dice and taking their sum. The event of interest is a sum less than 12. Find

1. A^c
2. $P(A^c)$
3. $P(A)$

If $A =$ (sum less than 12), then $A^c =$ (sum greater than or equal to 12). Take a moment to notice that there is only one way to get a sum greater than or equal to 12: rolling two 6s.

The chart below shows the rolls of Die 1 as columns and the rolls for Die 2 as rows. The numbers in the middle are the sums. Note that there are 36 possible ways to roll 2 dice.

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Even without the chart, by noting that there's only one way to get a sum greater than or equal to 12, we can quickly find $P(A^C)$:

$$P(A^c) = \frac{1}{36}$$

But trying to count all of the ways to get A would take a long time! Instead, we can use

$$P(A) = 1 - P(A^c) = 1 - \frac{1}{36} = \frac{35}{36}$$

3.5 Conditional Probability

A **contingency table** is a way to summarize **bivariate data**, or data from two variables.

Smallpox in Boston (1726)

Inoculated

yes

no

total

Result

lived

238

5136

5374

died

6

855

850

total

244

5980

6224

5136 is the count of people who lived AND were not inoculated.

6224 is the total number of observations.

244 is the total number of people who were inoculated.

5374 is the total number of people who lived.

This is basically a two-variable frequency distribution. And, like a frequency distribution, we can convert to proportions (relative frequencies) by dividing each count (each number) by the total number of observations:

Inoculated

yes

no

total

Result

lived

0.0382

0.8252

0.8634

died

0.0010

0.1356

0.1366

total

0.0392

0.9608

1.0000

0.8252 is the proportion of people who lived AND were not inoculated.

1.000 is the proportion of total number of observations. Think of this as 100% of the observations.

0.0392 is the proportion of people who were inoculated.

0.8634 is the proportion of people who lived.

The row and column totals are **marginal probabilities**. The probability of two events together (A and B) is a **joint probability**.

What can we learn about the result of smallpox if we already know something about inoculation status? For example, given that a person is inoculated, what is the probability of death? To figure this out, we restrict our attention to the 244 inoculated cases. Of these, 6 died. So the probability is 6/244.

This is called **conditional probability**, the probability of some event A if we know that event B occurred (or is true):

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

where the symbol | is read as "given."

For death given inoculation,

$$P(\text{death}|\text{inoculation}) = \frac{P(\text{death and inoculation})}{P(\text{death})} = \frac{0.0010}{0.0392} = 0.0255.$$

Notie that we could also write this as

$$P(\text{death}|\text{inoculation}) = \frac{P(\text{death and inoculation})}{P(\text{death})} = \frac{6/6224}{244/6224} = \frac{6}{244},$$

which is what we found when using the table to restrict our attention to only the inoculated cases.

If knowing whether event B occurs tells us nothing about event A, the events are **independent**. For example, if we know that the first flip of a (fair) coin came up heads, that doesn't tell us anything about what will happen next time we flip that coin.

We can test for independence by checking if $P(A|B) = P(A)$.

3.5.1 Multiplication Rules

Multiplication Rule for Independent Processes

If A and B are independent events, then

$$P(A \text{ and } B) = P(A)P(B).$$

We can extend this to more than two events:

$$P(A \text{ and } B \text{ and } C \text{ and } \dots) = P(A)P(B)P(C)\dots.$$

Note that if $P(A \text{ and } B) \neq P(A)P(B)$, then A and B are *not* independent.

Example: Find the probability of rolling a 6 on your first roll of a die and a 6 on your second roll.

Let $A =$ (rolling a 6 on first roll) and $B =$ (rolling a 6 on second roll). For each roll, the probabiltiy of getting a 6 is 1/6, so $P(A) = \frac{1}{6}$ and $P(B) = \frac{1}{6}$.

Then, because each roll is independent of any other rolls,

$$P(A \text{ or } B) = P(A)P(B) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

General Multiplication Rule

If A and B are any two events, then

$$P(A \text{ and } B) = P(A|B)P(B).$$

Notice that this is just the conditional probability formula, rewritten in terms of $P(A \text{ and } B)$!

Example: Suppose we know that 38.4% of US households have dogs and that among those with dogs, 23.1% have cats. Find the probability that a US household has both dogs and cats.

Let $C =$ (household has cats) and $D =$ (household has dogs). We know from the problem statement that $P(D) = 0.384$.

The other piece tells us something about the probability of having cats *among those with dogs*. This means that we *know* that these people have dogs. That is, *given* a household has dogs, the probabiltiy of cats is 23.1%. In probability notation, $P(C|D) = 0.231$. Then

$$P(C \text{ and } D) = P(C|D)P(D) = 0.231 \times 0.384 = 0.0887$$

or the probability that a US household has both cats and dogs is 0.0887.

Chapter 4

Random Variables

4.1 Chapter Overview

In previous chapters, we introduced the idea of variables and examined their distributions. We also began our discussion on probability theory. Now, we extend these concepts into what are called random variables. We will introduce the concept of random variables in general and will discuss a specific type of distribution - the binomial distribution. Then we will discuss a continuous probability distribution, the normal distribution. The normal distribution will provide a foundation for much of the inference we will complete throughout the rest of this course.

Chapter Learning Objectives/Outcomes

1. Discuss discrete random variables using key terminology.
2. Express cumulative probabilities using probability notation.
3. Calculate the expected value and standard deviation of a discrete random variable.
4. Calculate binomial probabilities.
5. Convert normal distributions to standard normal distributions.
6. Calculate probabilities for a normal distribution using area under the curve.
7. Approximate binomial probabilities using the normal curve.

This chapter's outcomes correspond to course outcomes (4) use the binomial distribution as a model for discrete variables and (5) use the normal distribution as a model for continuous variables.

4.2 Discrete Random Variables

A **random variable** is a quantitative variable whose values are based on chance. By "chance," we mean that you can't *know* the outcome before it occurs.

A **discrete random variable** is a random variable whose possible values can be listed.

Notation:

- x,y,z (lower case letters) denote variables.
- X, Y, Z (upper case letters) denote *random* variables.

In contrast to events, where we usually used letters toward the start of the alphabet, (random) variables are typically denoted by letters from the end of the alphabet.

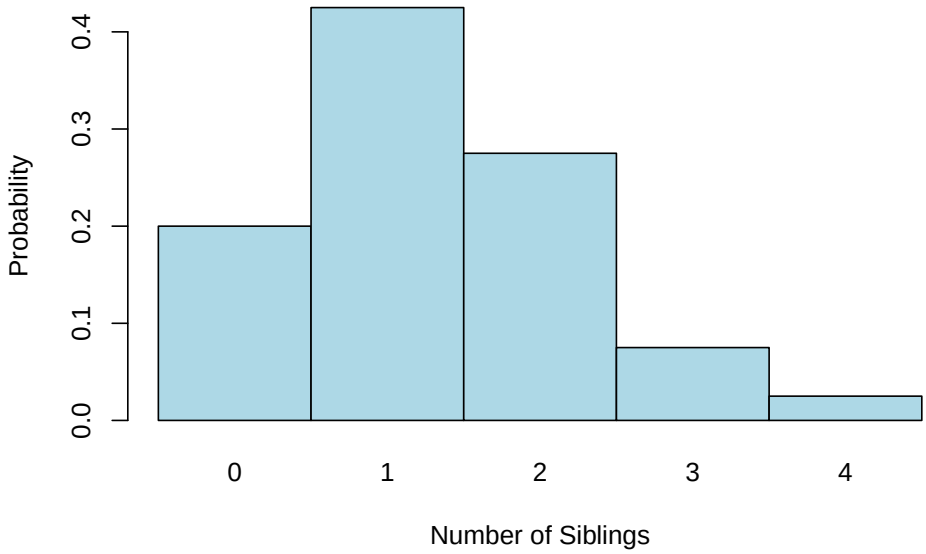
- $\{X = x\}$ denotes the event that the random variable X equals x.
- $P(X = x)$ denotes the probability that the random variable X equals x.

Recall: a probability distribution is a list of all possible values and their corresponding probabilities. (See Section 3.3 for a refresher.) A **probability histogram** is a histogram where the heights of the bars correspond to the probability of each value. For discrete random variables, each "bin" is one of the listed values.

Example:

Number of Siblings, x	0	1	2	3	4
Probability, $P(X = x)$	0.200	0.425	0.275	0.075	0.025

(Assume for the sake of the example that no one has more than 4 siblings.)



Interpretation: in a large number of independent observations of a random variable X, the proportion of times each possible value occurs will approximate the probability distribution of X.

4.2.1 The Mean and Standard Deviation

Mean of a Discrete Random Variable

The mean of a discrete random variable X is denoted μ_X. If it's clear which random variable we're talking about, we can drop the subscript and write μ.

$$\mu_X = \Sigma x P(X = x)$$

where Σ denotes "the sum over all values of x":

$$\Sigma x P(X = x) = x_1 P(X = x_1) + x_2 P(X = x_2) + \cdots + x_n P(X = x_n).$$

The mean of a random variable is also called the **expected value** or **expectation**. Recall that measures of center are meant to identify the most common or most likely, thus the value we can *expect* to see (most often).

Example: for the Siblings distribution,

$$\mu = 0(0.200) + 1(0.425) + 2(0.275) + 3(0.075) + 4(0.025) = 1.3$$

Make sure you understand how we used the formula for μ and the probability distribution to come up with this number.

Interpretation: in a large number of independent observations of a random variable X, the mean of those observations will approximately equal μ.

The larger the number of observations, the closer their average tends to be to μ. This is known as the **law of large numbers**.

> *Example*: Suppose I took a random sample of 10 people and asked how many siblings they have.
>
> $$2, 2, 2, 2, 1, 0, 3, 1, 2, 0$$
>
> In my random sample of 10, $\bar{x} = 2$, which is a reasonable estimate but not that close to the true mean $\mu = 1.3$.

- A random sample of 30 gave me a mean of $\bar{x} = 1.53$.
- A random sample of 100 gave me a mean of $\bar{x} = 1.47$.
- A random sample of 1000 gave me a mean of $\bar{x} = 1.307$.

We use concepts related to the law of large numbers as a foundation for statistical inference, but note that - although very large samples are nice to have - it's not necessary to take enormous samples all the time. Often, we can come to interesting conclusions with fewer than 30 observations!

Standard Deviation of a Discrete Random Variable

The variance of a discrete random variable X is denoted σ_X^2 (or σ^2 if it's clear which variable we're talking about).

$$\sigma_X^2 = \Sigma[(x - \mu_X)^2 P(X = x)]$$

OR

$$\sigma_X^2 = \Sigma[x^2 P(X = x)] - \mu_X^2$$

These formulas are *exactly* equivalent and you may use whichever you wish, but note that the second may be a little easier to work with.

As before, the standard deviation is the square root of the variance:

$$\sigma = \sqrt{\sigma^2}$$

> *Example*: Calculate the standard deviation of the Siblings variable.
>
> In general, a table is the best way to keep track of a variance calculation:

x	$P(X = x)$	$xP(X = x)$	x^2	$x^2 P(X = x)$
0	0.200	0	0	0
1	0.425	0.425	1	0.425
2	0.275	0.550	4	1.100
3	0.075	0.225	9	0.675
4	0.025	0.100	16	0.400
		$\mu = 1.3$		Total = 2.6

Then the variance is

$$\sigma^2 = 2.6 - 1.3^2 = 0.9$$

and the standard deviation is

$$\sigma = \sqrt{0.9} = 0.9539.$$

4.3 The Binomial Distribution

Think back to replication in an experiment. Each replication is what we call a **trial**. We will consider a setting where each trial has two possible outcomes.

> For example, suppose you want to know if a coin is fair (both sides equally likely). You might flip the coin 100 times (thus running 100 trials). Each trial is a flip of the coin with two possible outcomes: heads or tails.

The product of the first k positive integers $(1, 2, 3, \dots)$ is called **k-factorial**, denoted $k!$:

$$k! = k \times (k-1) \times \cdots \times 3 \times 2 \times 1$$

We define $0! = 1$.

Example: $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

If n is a positive integer $(1, 2, 3, \dots)$ and x is a nonnegative integer $(0, 1, 2, \dots)$ with $x \leq n$, the **binomial coefficient** is

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Example:

$$\binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{(2 \times 1)(3 \times 2 \times 1)}$$

Sometimes, we may want to simplify a binomial coefficient *before* taking all of the factorials. Why? Well,

$$20! = 2432902008176640000$$

Most calculators will not print this number. Instead, you'll get an error or a rounded version printed using scientific notation. Neither will help you accurately calculate the binomial coefficient.

Example:

$$\binom{20}{17} = \frac{20 \times 19 \times 18 \times 17 \times 16 \times \cdots \times 3 \times 2 \times 1}{(17 \times 16 \times \cdots \times 3 \times 2 \times 1)(3 \times 2 \times 1)}$$

but notice that I can rewrite 20! as $20 \times 19 \times 18 \times 17!$, so

$$\binom{20}{17} = \frac{20 \times 19 \times 18 \times 17!}{17!(3 \times 2 \times 1)} = \frac{20 \times 19 \times 18}{3 \times 2 \times 1} = \frac{6840}{6} = 1140$$

Bernoulli trials are repeated trials of an experiment that satisfy 1. Each trial has two possible outcomes: success and failure. 2. Trials are independent. 3. The probability of success (the **success probability**) p remains the same from one trial to the next:

$$P(X = \text{success}) = p$$

The **binomial distribution** is the probability distribution for the number of successes in a sequence of Bernoulli trials.

Fact: in n Bernoulli trials, the number of outcomes that contain exactly x successes equals the binomial coefficient $\binom{n}{x}$.

Binomial Probability Formula

Let x denote the total number of successes in n Bernoulli trials with success probability p. The probability distribution of the random variable X is given by

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, 2, \ldots, n$$

The random variable X is called a **binomial random variable** and is said to have the **binomial distribution**. Because n and p fully define this distribution, they are called the distribution's **parameters**.

To find a binomial probability formula:

Check assumptions.

Exactly n trials to be performed.

Two possible outcomes for each trial.

Trials are independent (each trial does not impact the result of the next)

Success probability p remains the same from trial to trial.

Identify a "success." Generally, this is whichever of the two possible outcomes we are most interested in.

Determine the success probability p.

Determine n, the number of trials.

Plug n and p into the binomial distribution formula.

We can also use the binomial probability formula to calculate probabilities like $P(X \leq x)$. Notice that we can rewrite this uisng concepts from the previous chapter

$$P(X \leq k) = P(X = k \text{ or } X = k-1 \text{ or } \ldots \text{ or } X = 2 \text{ or } X = 1 \text{ or } X = 0)$$

Since X is a discrete random variable, each possible value is *disjoint*. We can use this!

$$P(X \leq k) = P(X = k) + P(X = k-1) + \cdots + P(X = 2) + P(X = 1) + P(X = 0)$$

Example: $P(X \leq 3) = P(X = 3) + P(X = 2) + P(X = 1) + P(X = 0)$

We can also extend this concept to work with probabilities like $P(a < X \leq b)$.

Example: $P(2 < X \leq 5)$

First, notice that if $2 < X \leq 5$, then X can be 3, 4, or 5:

$$P(2 < X \leq 5) = P(X = 3) + P(X = 4) + P(X = 5)$$

Note: if going from $2 < X \leq 5$ to "X can be 3, 4, or 5" doesn't make sense to you, start by writing out the sample space. Suppose $n = 10$. Then the sample space for the binomial distribution is

$$S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

Then I can check any number in this sample space by plugging it in for X. So for 1, I can check $2 < 1 \leq 5$. Obviously this is not true, so we won't include 1. Checking the number 2, I get $2 < 2 \leq 5$. Since $2 < 2$ is NOT true, we don't include 2. Etc.

4.3.1 Mean and Variance

The shape of a binomial distribution is determined by the success probability:

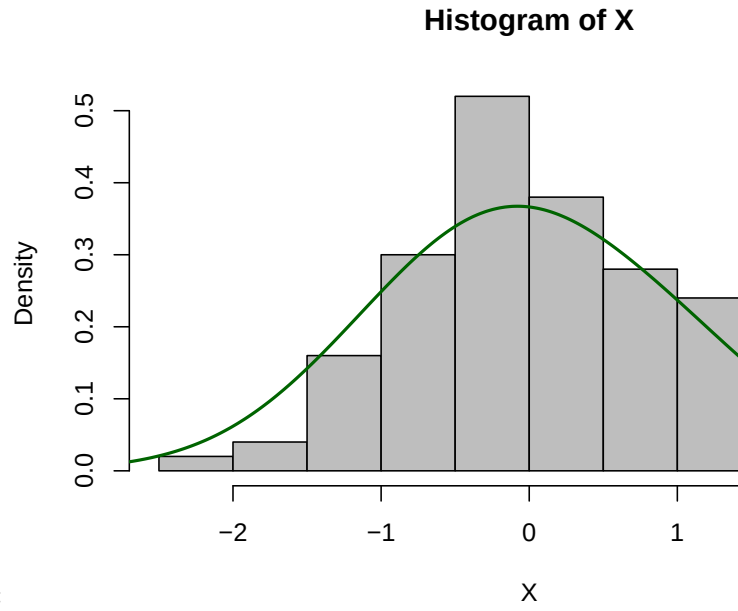
- If $p \approx 0.5$, the distribution is approximately symmetric.
- If $p < 0.5$, the distribution is right-skewed.
- If $p > 0.5$, the distribution is left-skewed.

The mean of a binomial distribution is $\mu = np$. The variance is $\sigma^2 = np(1-p)$.

4.4 The Normal Distribution

If we can represent a discrete variable with a probability histogram, what can we do with a continuous variable?

We represent the shape of a continuous variable using a **density curve**. This is

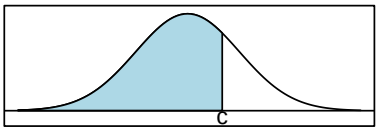


like a histogram, but with a smooth curve:

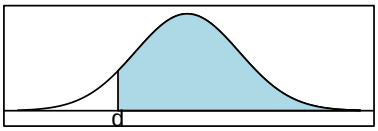
Properties:

1. The curve is always above the horizontal axis (because probabilities are always nonnegative).
2. The total area under the curve equals 1.

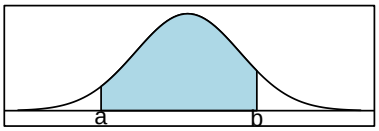
For a variable with a density curve, the proportion of all possible observations that lie within a specified range equals the corresponding area under the density curve.



P(X < c)



P(X > d)



P(a < X < b)

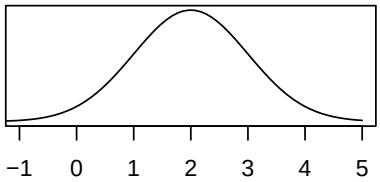
A **normal curve** is a special type of density curve that has a "bell-shaped" distribution. In fact, all of the density curves I've shown so far have been normal curves! We say that a variable is **normally distributed** or has a **normal distribution** if its distribution has the shape of a normal curve.

Why "normal?" Because it's very common! Lots of things are more common around the average and less common as you get farther from the average: height, amount of sleep people get each night, standardized test scores, etc. In practice, these things aren't *exactly* normally distributed... instead, they're **approximately normally distributed** (and that's ok).

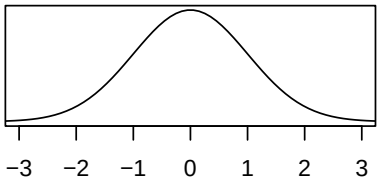
Normal distributions...

- are fully determined by parameters mean μ and standard deviation σ.
- are symmetric and centered at μ.
- have spreads that depend on σ.

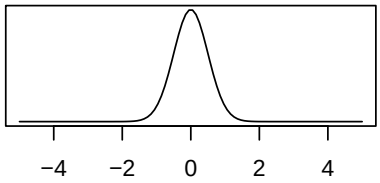
Pay close attention to the horizontal axis and how spread out the densities are in each of the following plots:



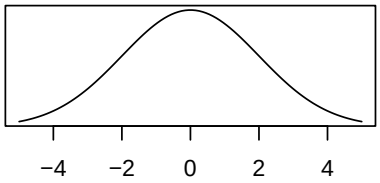
Normal(mu = 2, sigma = 1)



Normal(mu = 0, sigma = 1)



Normal(mu = 0, sigma = 0.5)



Normal(mu = 0, sigma = 2)

Notice that the bottom left plot comes to a sharper peak, while the bottom right has a gentler slope. This is what we mean by "spread": the density on the bottom right is the most spread out.

To check whether a variable is (approximately) normally distributed,

1. Check the histogram to see if it is symmetric and bell-shaped.
2. Estimate the parameters: μ using $\bar{x}$ and σ using s.

4.4.1 The Standard Normal Distribution

In order to make normal distributions easier to work with, we will **standardize** them. A **standard normal distribution** is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. We standardize a variable using

$$z = \frac{x - \mu}{\sigma}.$$

This is also called a **z-score**. Standardizing using this formula will *always* result in a variable with mean 0 and standard deviation 1 (even if it's not normal!). If X is approximately normal, then the standardized variable Z will have a standard normal distribution.

Note: when we z-score a variable, we preserve the area under the curve properties! If X is Normal(μ, σ), then

$$P(X < c) = P\left(Z < \frac{c - \mu}{\sigma}\right) = P(Z < z).$$

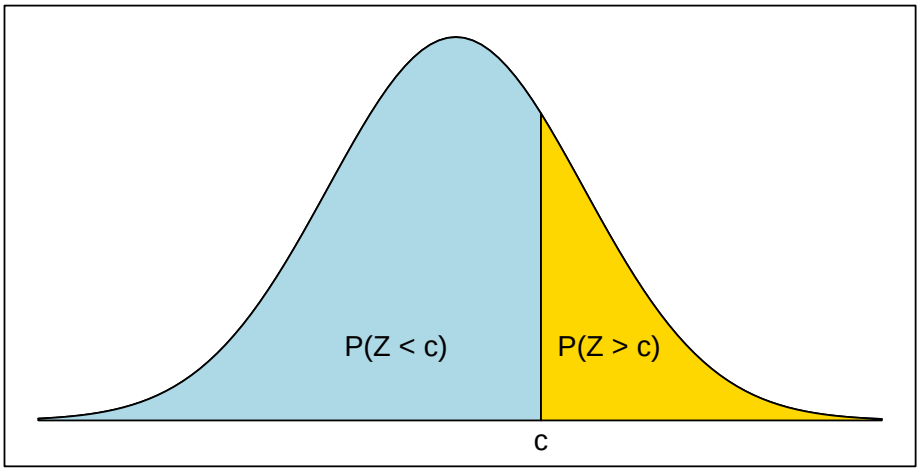
4.5 Area Under the Standard Normal Curve

Properties:

1. Total area under the curve is 1.
2. The curve extends infinitely in both directions, never touching the horizontal axis.
3. Symmetric about 0.
4. Almost all of the area under the curve is between -3 and 3.

We will think about area under the standard normal curve in terms of **cumulative probabilities** or probabilities of the form $P(Z < z)$.

We will use the fact that the total area under the curve is 1 to find probabilities like $P(Z > c)$:



Total area = 1

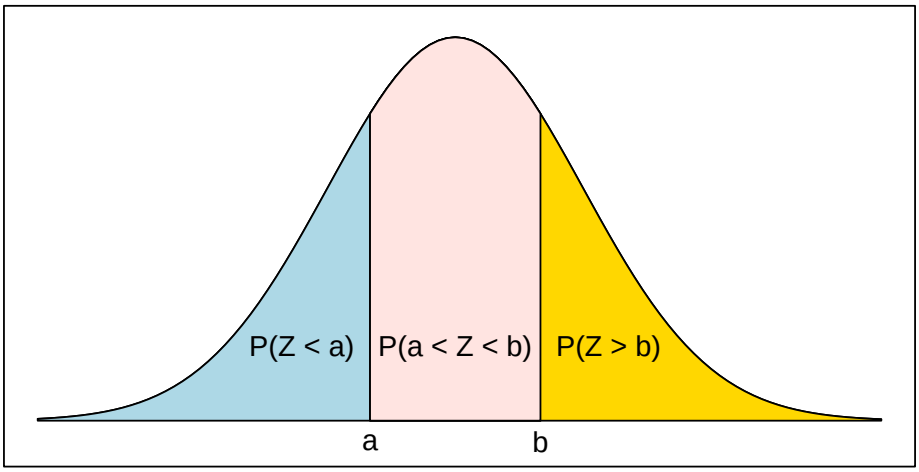
Using the graphic to help visualize, we can see that

$$1 = P(Z < c) + P(Z > c)$$

which we can then rewrite as

$$P(Z > c) = 1 - P(Z < c).$$

We can also use this concept to find $P(a < Z < b)$.



Notice that

$$1 = P(Z < a) + P(a < Z < b) + P(Z > b),$$

which we can rewrite as

$$P(a < Z < b) = 1 - P(Z > b) - P(Z < a)$$

and since we just found that $P(Z > b) = 1 - P(Z < b)$, we can replace $1 - P(Z > b)$ with $P(Z < b)$, and get

$$P(a < Z < b) = P(Z < b) - P(Z < a).$$

Key Cumulative Probability Concepts

- $P(Z > c) = 1 - P(Z < c)$
- $P(a < Z < b) = P(Z < b) - P(Z < a)$

A final note, because the normal distribution is symmetric, $P(X < \mu) = P(X > \mu) = 0.5$. Notice this also implies that, when a distribution is symmetric (and unimodal), the mean and median are the same!

Now that we can get all of our probabilities written as *cumulative* probabilities, we're ready to use software to find the area under the curve!

Finding Area Under the Curve: R

We will use statistical software called `R` to find areas under the curve. R is an incredibly powerful statistical programming language, but we're going to keep it simple. $P(Z < z)$ is found using the command 'pnorm(z).' To find $P(Z < 1)$, I would type `pnorm(1)`. That entry and `R` output look like this:

```
pnorm(1)
```

```
## [1] 0.8413447
```

so $P(Z < 1) = 0.8413447$. Since we are only going to use R for a few simple commands, we will run it completely online at the website rdrr.io/snippets (bookmark this website!)

For now, you can run R right here in the course notes! This is exactly what you will see on the rdrr.io website. Type in your command and click the green "Run" button. Try finding $P(Z < 2)$.

Make sure you are able to run the command and get $P(Z < 2) = 0.9772499$. (If it prints out "Sorry, something went wrong. All I know is:" just press the "Run" button again.)

We can also find a z-score given a specified area/probability. The notation z_α (z-alpha) is the z-score corresponding to a right-tail area of α. That is,

$$P(Z > z_\alpha) = \alpha$$

We can find z_α using the command `qnorm(p, lower.tail=FALSE)`. To find $P(Z > z_\alpha) = 0.1$, I would type

```
qnorm(0.1, lower.tail=FALSE)
```

```
## [1] 1.281552
```

so if $P(Z > z_\alpha) = 0.1$, then $z_\alpha = 1.281552$. (If you wanted to consider $P(Z < z) = p$, you would replace "FALSE" with "TRUE.")

A quick note about `R`: `R` will print very large numbers and numbers close to 0 using *scientific notation*. However, `R`'s scientific notation may not look the way you're used to! Check out the `R` output for $P(Z < -5)$:

```
pnorm(-5)
```

```
## [1] 2.866516e-07
```

When you see `e-07`, that means $\times 10^{-7}\ldots$ so $P(Z < -5) = 2.8665 \times 10^{-7} \approx 0.00000029$.

Finding Area Under the Curve: Applets

Another option for finding probabilities and z-scores associated with the normal curve is to use an online applet. The Rossman and Chance Normal Probability Calculator is my preferred applet. It's relatively straightforward to use and would be difficult to demonstrate in these course notes! We will demonstrate this applet in class. I recommend you bookmark any websites you use to find probabilities!

You can also find the area under a normal distribution using a Normal Distribution Table. These are outdated and not used anywhere but the statistics

classroom. As a result, I do not teach them. However, if you wish to use the table instead of R, there is a short tutorial here.

4.6 Working with Normally Distributed Variables

4.6.1 Normal Distribution Probabilities

Using z-scores and area under the standard normal curve, we can find probabilities for any normal distribution problem!

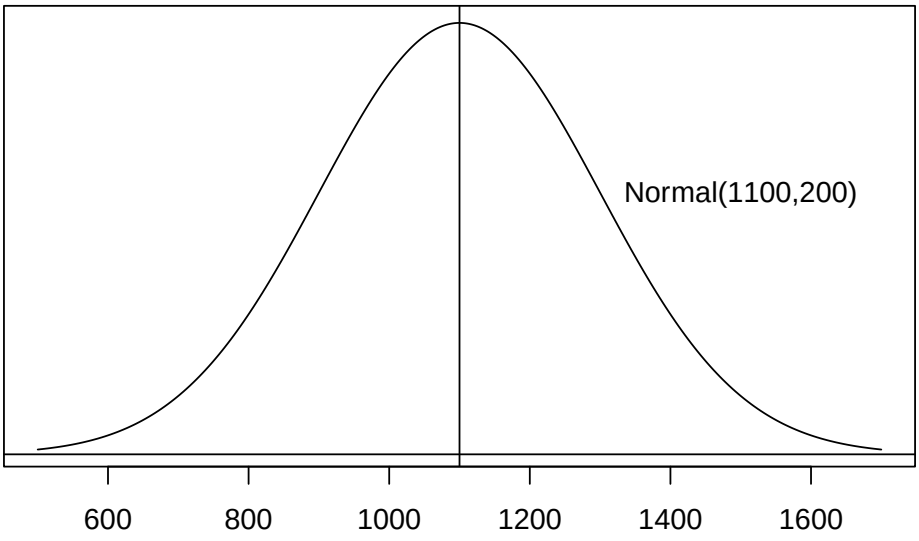
Determining Normal Distribution Probabilities

1. Sketch the normal curve for the variable.
2. Shade the region of interest and mark its delimiting x-value(s).
3. Find the z-score(s) for the value(s).
4. Use the `pnorm` command in `R` to find the associated area.

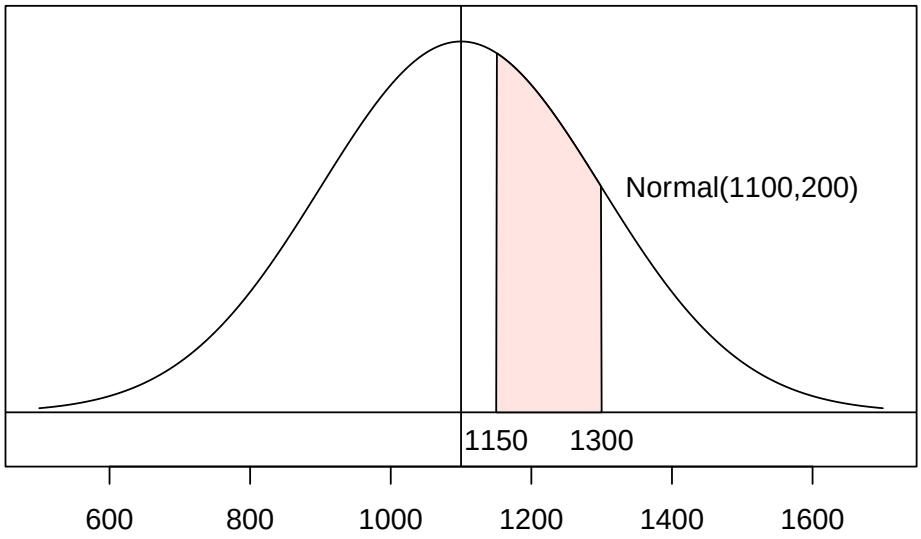
Example: Find the proportion of SAT-takers who score between 1150 and 1300. Assume that SAT scores are approximately normally distributed with mean $\mu = 1100$ and standard deviation $\sigma = 200$.

First, let's figure out what we want to calculate. Using area under the curve concepts, the proportion of test-takers who score *between* 1150 and 1300 will be $P(1150 < X < 1300)$.

1. Sketch:



2. Shade and label:



3. Calculate z-scores:

$$x = 1150 \rightarrow z = \frac{1150 - 1100}{200} = 0.25$$

and

$$x = 1300 \rightarrow z = \frac{1300 - 1100}{200} = 1.$$

4. Use `R` with `pnorm` to find $P(Z < 0.25)$ and $P(Z < 1)$:

```
pnorm(0.25)
```

```
## [1] 0.5987063
```

```
pnorm(1)
```

```
## [1] 0.8413447
```

Note that

$$P(1150 < X < 1300) = P\left(\frac{1150 - 1100}{200} < Z < \frac{1300 - 1100}{200}\right) = P(0.25 < Z < 1)$$

and, using cumulative probability concepts,

$$P(0.25 < Z < 1) = P(Z < 1) - P(Z < 0.25).$$

Using `R`, we found $P(Z < 0.25) \approx 0.5987$ and $P(Z < 1) \approx 0.8413$, so

$$P(Z < 1) - P(Z < 0.25) \approx 0.8413 - 0.5987 = 0.2426.$$

That is, approximately 26.26% of test-takers score between 1150 and 1300 on the SAT.

4.6.2 Empirical Rule for Variables

For any (approximately) normally distributed variable,

1. Approximately 68% of all possible observations lie within one standard deviation of the mean: $\mu \pm \sigma$.
2. Approximately 95% of all possible observations lie within two standard deviations of the mean: $\mu \pm 2\sigma$.
3. Approximately 99.7% of all possible observations lie within three standard deviations of the mean: $\mu \pm 3\sigma$.

Given some data, you can check if approximately 68% of the data falls within $\bar{x} \pm s$, 95% within $\bar{x} \pm 2s$, and 99.7% within $\bar{x} \pm 3s$ to examine whether the data follow the empirical rule.

Note that a z-score tells us how many standard deviations an observation is from the mean. A positive z-score $z > 0$ is *above* the mean; a negative z-score $z < 0$ is *below* the mean.

> *Example*: $z = -0.23$ is 0.23 standard deviations below the mean.

4.6.3 Percentiles

We can also find the *observation* associated with a percentage/proportion.

The wth **percentile** p_w is the observation that is higher than w% of all observations

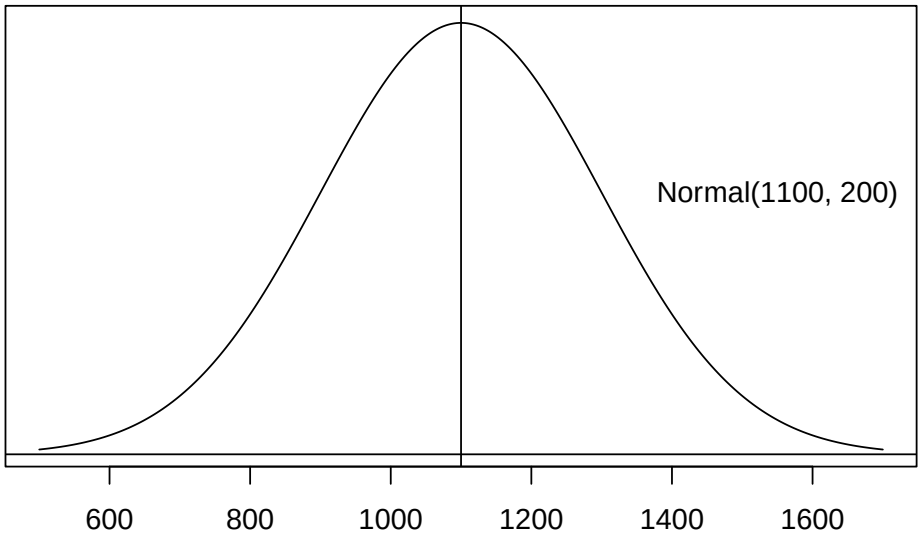
$$P(X < p_w) = w$$

Finding a Percentile

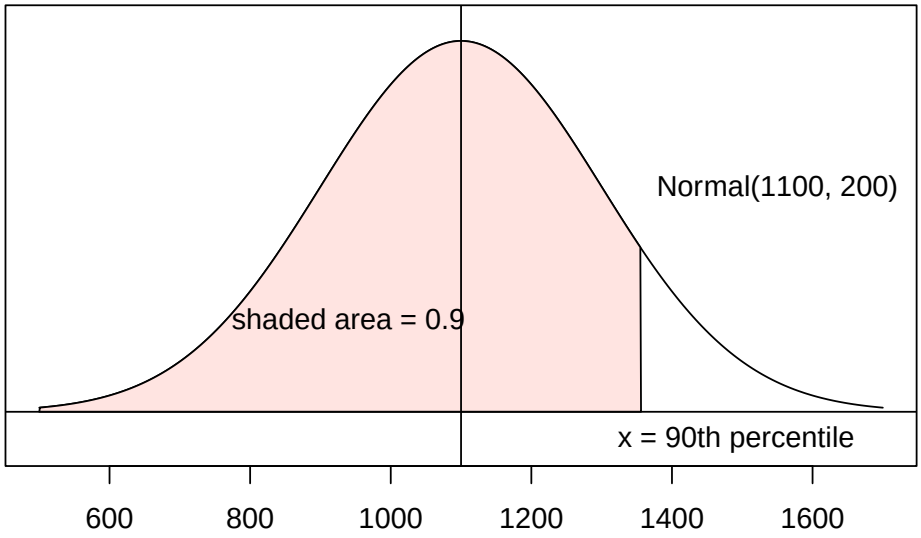
1. Sketch the normal curve for the variable.
2. Shade the region of interest and label the area.
3. Use the applet to determine the z-score for the area.
4. Find the x-value using z, μ, and σ.

Note that if $z = \frac{x-\mu}{\sigma}$, then $x = \mu + z\sigma$.

> *Example*: Find the 90th percentile for SAT scores.
>
> From the previous example, we know that SAT scores are approximately Normal($\mu = 1100$, $\sigma = 200$). 1. Sketch the normal curve.



2. Shade the region of interest and label the area.



3. Use `R` with `qnorm` to determine the z-score for the area:

```
qnorm(0.9)
```

```
## [1] 1.281552
```

Find the x-value using $z \approx 1.2816$, $\mu = 1100$, and $\sigma = 200$:

$$x = 1100 + 1.2816(200) = 1356.32$$

so 90% of SAT test-takers score below 1356.32.

Chapter 5

Introduction to Inference

5.1 Chapter Overview

This chapter will bridge the gap between our discussion on the normal distribution and our first forays into statistical inference. As it turns out, much of the statistical inference we will use relies on the normal distribution and the t-distribution, which we will introduce in this chapter. We begin our study of statistical inference by learning about confidence intervals.

Chapter Learning Objectives/Outcomes

1. Find the distribution of a sample mean.
2. Estimate probabilities for a sample mean.
3. Calculate and interpret confidence intervals for a population mean.
4. Use the standard normal and t-distributions to find critical values.

This chapter's outcomes correspond to course outcome (6) apply statistical inference techniques of parameter estimation such as point estimation and confidence interval estimation and (7) apply techniques of testing various statistical hypotheses concerning population parameters.

5.2 Sampling Distributions

5.2.1 Sampling Error

We want to use a sample to learn something about a population, but no sample is perfect! **Sampling error** is the error resulting from using a sample to estimate a population characteristic.

If we use a sample mean $\bar{x}$ to estimate μ, chances are that $\bar{x} \neq \mu$ (they might be close but... they might not be!). We will consider

- How close *is* $\bar{x}$ to μ?
- What if we took many samples and calculated $\bar{x}$ many times?
  - How would that relate to μ?
  - What would be the distribution of these values?

The distribution of a statistic (across all possible samples of size n) is called the **sampling distribution**. We will focus primarily on the distribution of the sample mean.

For a variable x and given a sample size n, the distribution of $\bar{x}$ is called the **sampling distribution of the sample mean** or the **distribution of $\bar{x}$**.

> *Example*: Suppose our population is the five starting players on a particular basketball team. We are interested in their heights (measures in inches). The full population data is

Player	A	B	C	D	E
Height	76	78	79	81	86

> The population mean is $\mu = 80$. Consider all possible samples of size $n = 2$:

Sample	A,B	A,C	A,D	A,E	B,C	B,D	B,E	C,D	C,E	D,E
$\bar{x}$	77	77.5	78.5	81.0	78.5	79.5	82.0	80.0	82.5	83.5

> There are 10 possible samples of size 2. Of these samples, 10% have means exactly equal to μ (for a *random* sample of size 2, you'd have a 10% chance to find $\bar{x} = \mu$... and a 90% chance not to!).

In general, the larger the sample size, the smaller the sampling error tends to be in estimating μ using $\bar{x}$.

In practice, we have one sample and μ is unknown. We also have limited resources to collect data, so it may not be feasible to collect a very large sample.

The mean of the distribution of $\bar{x}$ is $\mu_{\bar{X}} = \mu$ and the standard deviation is $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. We refer to the standard deviation of a sampling distribution as **standard error**. (Note that this standard error formula is built for very large populations, so it will not work well for our basketball players. This is okay! We usually work with populations so large that we treat them as "infinite.")

> *Example*: The mean living space for a detatched single family home in the United States is 1742 ft^2 with a standard deviation of 568 square

feet. (Does that mean seem huge to anyone else??) For samples of 25 homes, determine the mean and standard error of $\bar{x}$.

Using our formulae:

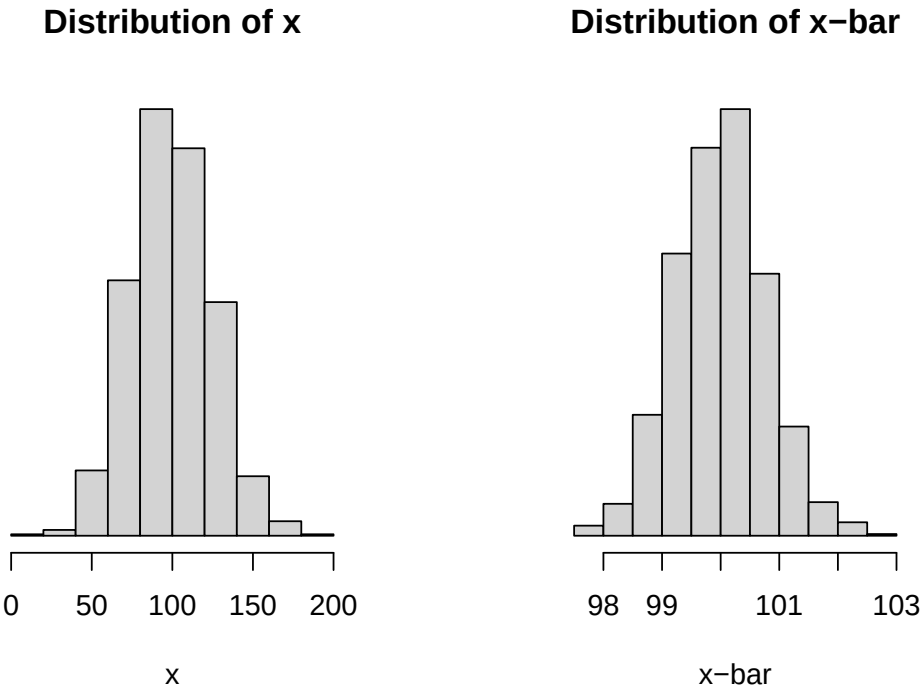
$$\mu_{\bar{X}} = \mu = 1742$$

and

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{568}{\sqrt{25}} = 113.6.$$

5.2.2 The Sampling Distribution of $\bar{X}$

First, we consider the setting where X is Normal(μ, σ). The plots below show (A) a random sample of 1000 from a Normal(100, 25) distribution and (B) the approximate sampling distribution of $\bar{X}$ when X is Normal(100, 25).



Notice how the x-axis changes from one plot to the next.

In fact, if X is Normal(μ, σ), then $\bar{X}$ is Normal($\mu_{\bar{X}} = \mu$, $\sigma_{\bar{X}} = \sigma/\sqrt{n}$).

Central Limit Theorem

For relatively large sample sizes, the random variable $\bar{X}$ is approximately normally distributed *regardless of the distribution of* X:

$$\bar{X} \text{ is Normal}(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}} = \sigma/\sqrt{n}).$$

Notes

- This approximation improves with increasing sample size.
- In general, "relatively large" means sample sizes $n \geq 30$.

5.3 Developing Confidence Intervals

Recall: A **point estimate** is a single-value estimate of a population parameter. We say that a statistic is an **unbiased estimator** if the mean of its distribution is equal to the population parameter. Otherwise, it is a **biased estimator**.

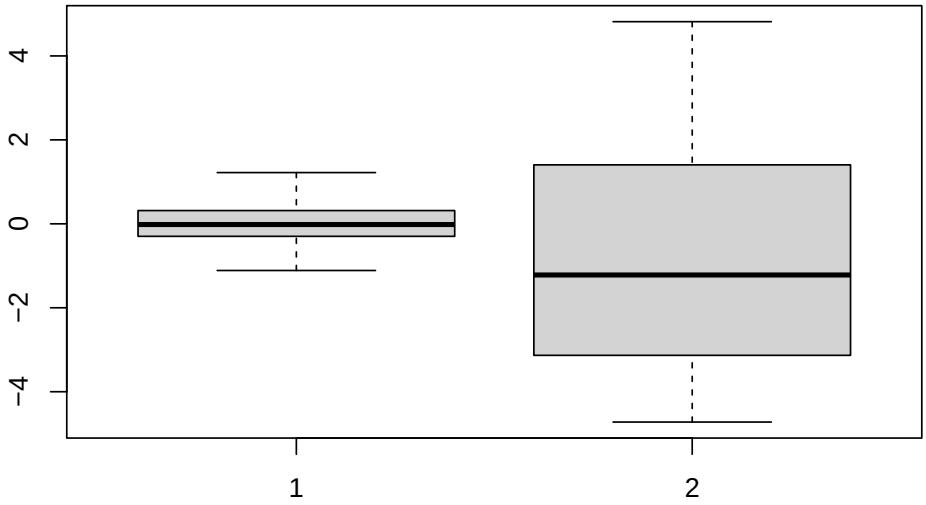
Comment Remember how our formula for standard deviation, the "mean squared deviance" divides by $n-1$ instead of n? We do this so that s is an *unbiased* estimate of σ.

Ideally, we want estimates that are unbiased with small standard error. For example, a sample mean (unbiased) with a large sample size (results in smaller standard error).

Point estimates are useful, but they only give us so much information. The variability of an estimate is also important!

> *Example* Think about estimating what tomorrow's weather will be like. If it's May in Sacramento, the average high temperature is 82 degrees Fahrenheit, but it's not uncommon to have highs anywhere from 75 to 90! Since the highs are so *variable*, it's hard to be confident using 82 to predict tomorrow's weather.
>
> On the flip side, think about July in Phoenix. The average high is 106 degrees Fahrenheit. In Phoenix, it's uncommon to have a July day with a high below 100. Since the highs are *not variable*, you could feel pretty confident using 105 to predict tomorrow's weather.

Take a look at these two boxplots:



Both samples are size $n = 100$ and have $\bar{x} = 0$, which would be our point estimate for μ... but Variable 1 has a standard deviation of $\sigma = 0.5$ and Variable 2 has standard deviation $\sigma = 5$. As a result, we can be more confident in our estimate of the population mean for Variable 1 than for Variable 2.

We want to formalize this idea of confidence in our estimates. A **confidence interval** is an interval of numbers based on the point estimate of the parameter. Say we want to be 95% confident about a statement. In Statistics, this means that we have arrived at our statement using a method that will give us a correct statement 95% of the time.

Assume we are taking a sample from a normal distribution with mean μ and standard deviation σ. We will assume the value of σ is known to us. Then $\bar{X}$ is Normal$(\mu, \sigma/\sqrt{n})$. If we standardize $\bar{X}$, we get

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

We want some interval (a, b). We will start by considering $a < Z < b$, so $a < Z$ and $Z < b$ (or $b > Z$). Then

$$\begin{aligned} Z &< b \\ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} &< b \\ \bar{X} - \mu &< b\sigma/\sqrt{n} \\ \bar{X} - b\sigma/\sqrt{n} &< \mu \end{aligned}$$

and

$$\begin{aligned} a &< Z \\ a &< \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \\ a\sigma/\sqrt{n} &< \bar{X} - \mu \\ \mu &< \bar{X} - a\sigma/\sqrt{n} \end{aligned}$$

putting these together,

$$\bar{X} - b\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} - a\frac{\sigma}{\sqrt{n}}.$$

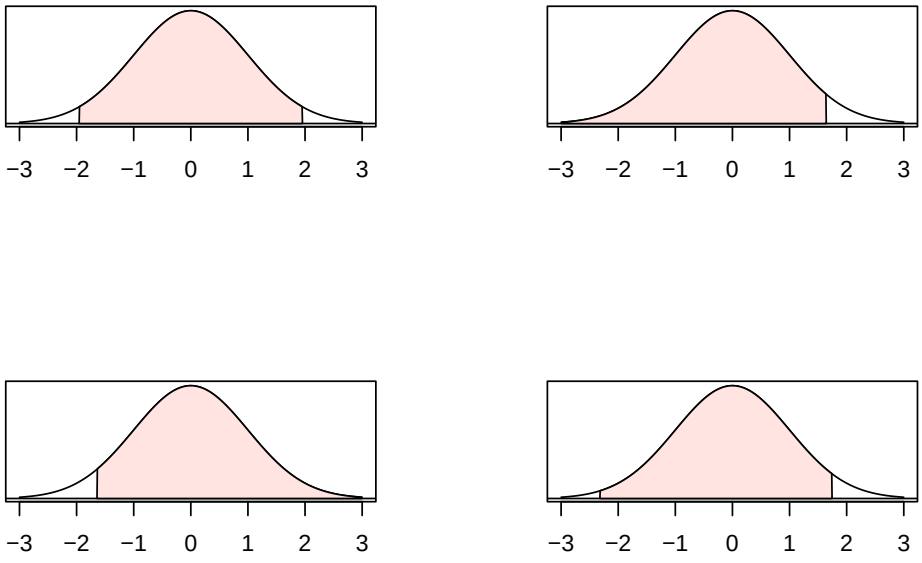
If we want to be 95% confident, then we want $P(a < Z < b) = 0.95$:

$$P\left(\bar{X} - b\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} - a\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

To calculate the 95% confidence interval, we need to find a and b such that $P(a < Z < b) = 0.95$.

We want this interval to be as narrow (small) as possible. Why? Narrower intervals are more informative. If I say I'm 95% confidence that tomorrow's high will be between -100 and 200 degrees Fahrenheit, that's a useless interval. If I change it to between 70 and 100, that's a little better. Changing it to between 85 and 90 is even better. This is what we mean by more informative.

It turns out that, with a symmetric distribution like the normal distribution, the way to make a confidence interval as narrow as possible is to take advantage of this symmetry. Each of the plots below show a shaded area of 0.95. The narrowest interval (along the horizontal axis) is the first interval, which is shaded on $(-1.96 < z < 1.96)$.



The confidence interval, then, is

$$\left(\bar{x} - z_* \frac{\sigma}{\sqrt{n}}, \bar{x} + z_* \frac{\sigma}{\sqrt{n}}\right)$$

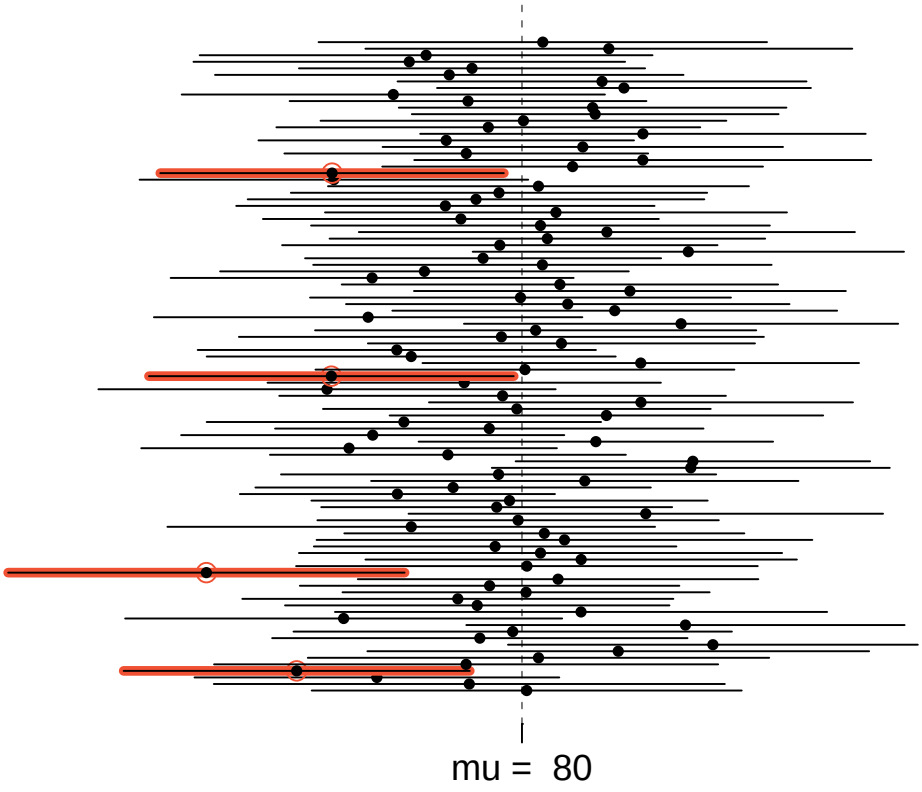
where $z_* = 1.96$. The midpoint of this interval is $\bar{x}$. The value of

$$z_* \frac{\sigma}{\sqrt{n}}$$

is called the **margin of error**.

5.3.1 Interpreting a Confidence Interval

To interpret a confidence interval, we need to think back to our definition of probability as "the proportion of times is would occur if the experiment were run infinitely many times." In the confidence interval case, if an experiment is run infinitely many times, the true value of μ will be contained in 95% of the intervals.



The graphic above shows 95% confidence intervals for 100 samples of size $n = 60$ drawn from a population with mean $\mu = 80$ and standard deviation $\sigma = 25$.

Each sample's confidence interval is represented by a horizontal line. The dot in the middle of each is the sample mean. When a confidence interval does *not* capture the population mean μ, the line is printed in red. Based on this concept of repeated sampling, we would expect about 95% of these intervals to capture μ. In fact, 96 of the 100 intervals capture μ.

Finally, when you interpret a confidence interval, it is important to do so in the context of the problem.

> *Example* The preferred keyboard height for typists is approximately normally distributed with $\sigma = 2.0$. A sample of size $n = 31$, resulted in a mean prefered keyboard height of $80cm$. Find and interpret a 95% confidence interval for keyboard height.
>
> The interval is
>
> $$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 80.0 \pm 1.96 \times \frac{2.0}{\sqrt{31}} = 80.0 \pm 0.70 = (79.3, 80.7).$$
>
> Interpretation:We can be 95% confident that the mean preferred keyboard height for typists is between 79.3cm and 80.7cm.

Notice that I kept the interpretation simple! That's okay - just be sure you are *also* able to explain what it means to be 95% confident (using the concept of repeated sampling).

Common mistakes:

- It is NOT accurate to say that "the probability that μ is in the confidence interval is 0.95." The parameter μ is some fixed quantity and it's either in the interval or it isn't.
- We are NOT "95% confident that $\bar{x}$ is in the interval." The value $\bar{x}$ is some known quantity and it's always in the interval.

5.3.2 Exercises

1. Suppose I took a random sample of 50 Sac State students and asked about their SAT scores and found a mean score of 1112. Prior experience with SAT scores in the CSU system suggests that SAT scores are well-approximated by a normal distribution with standard deviation known to be 50.
   a. Find a 95% confidence interval for Sac State SAT scores.
   b. Interpret your interval in the context of the problem.
   c. What is the width of your interval? If you want a narrower interval, what could you do?

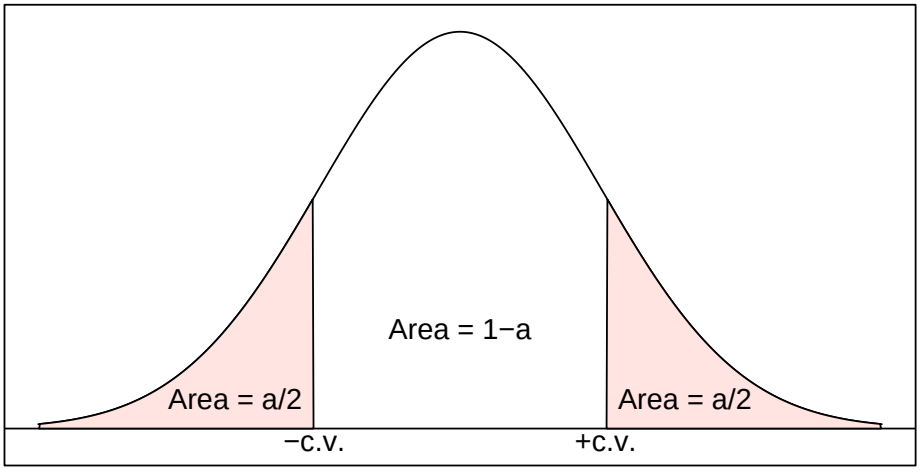
5.4 Other Levels of Confidence

While the 95% confidence interval is common in research, there's nothing inherently special about it. You could calculate a 90%, a 99%, or - if you're feeling spicy - something like a 43.8% confidence interval. These numbers are called the **confidence level** and they represent the proportion of times that the parameter will fall in the interval (if we took many samples).

The 100(1-α)% confidence interval for μ is given by

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2}$ is the z-score associated with the $[1-(\alpha/2)]$th percentile of the standard normal distribution. The value $z_{\alpha/2}$ is called the **critical value** ("c.v." on the plot, below).



We can find critical values in `R` using the same command we used to find percentiles: `qnorm(p)`. We want a 100(1-α)% confidence interval, so we need to quickly solve for α and divide by 2. For example, for a 98% interval,

$$100(1-\alpha) = 98 \implies \alpha = 0.02$$

Then $\alpha/2 = 0.01$ and

```
qnorm(0.01)
```

```
## [1] -2.326348
```

So the critical value is $z_{\alpha/2} = 2.326$. Notice that I dropped the negative sign here. That's because our formula uses $\pm z_{\alpha/2}$, so the sign doesn't matter. I'll always ignore that negative for critical values. As long as you write your interval as (smaller number, bigger number), it's all good.

Common Critical Values

Confidence Level	α	Critical Value, $z_{\alpha/2}$
90%	0.10	1.645
95%	0.05	1.96
98%	0.02	2.326
99%	0.01	2.575

5.4.1 Breaking Down a Confidence Interval

Consider

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

The key values are

- $\bar{x}$, the sample mean
- σ, the population standard deviation
- n, the sample size
- $z_{\alpha/2}$, the critical value

$$P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$$

The value of interest is μ, the (unknown) population mean; the confidence interval gives us a reasonable range of values for μ.

In addition, the formula includes

- The standard error, $\frac{\sigma}{\sqrt{n}}$
- The margin of error, $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

5.4.2 Confidence Level, Precision, and Sample Size

If we can be 99% confident (or even higher), why do we tend to "settle" for 95%?? Take a look at the common critical values (above) and the confidence interval formula

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

What will higher levels of confidence do to this interval? Think back to the intuitive interval width explanation with the weather. Mathematically, the same thing will happen: the interval will get wider! And remember, a narrow interval is a more informative interval. There is a trade off here between interval width and confidence. In general, the scientific community has settled on 95% as a compromise between the two, but different fields may use different levels of confidence.

There is one other thing we can control in the confidence interval: the sample size n. One strategy is to specify the confidence level and the maximum acceptable

interval width and use these to determine sample size. We know that

$$\text{interval width} \geq 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

(Note: I use $\geq$ because $2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is the *maximum* interval width - we would still be happy if this value turned out to be smaller!) Letting interval width equal w, we can solve for n:

$$n \geq \left(2z_{\alpha/2}\frac{\sigma}{w}\right)^2$$

Alternately, we may specify a maximum margin of error m instead:

$$n \geq \left(z_{\alpha/2}\frac{\sigma}{m}\right)^2$$

Once we've done this calculation, we need a whole number for n. Since $n \geq$ something, we will *always round up.*

> *Example* Suppose we want a 95% confidence interval for the mean of a normally distributed population with standard deviation $\sigma = 10$. It is important for our margin of error to be no more than 2. What sample size do we need?
>
> Using the formula for sample size with a desired margin of error, I can plug in $z_{0.05/2} = 1.96$, $m = 2$ and $\sigma = 10$:
>
> $$n = \left(1.96 \times \frac{10}{2}\right)^2 = 96.04$$
>
> . So (rounding up!) I need a sample size of *at least 97.*

A few comments:

- As desired width/margin of error decreases, n will increase.
- As σ increases, n will also increase. (More population variability will necessitate a larger sample size.)
- As confidence level increases, n will also increase.

5.4.3 Exercises

1. In the previous section, you worked with a random sample of 50 Sac State students with mean SAT score 1112. Prior experience with SAT scores in the CSU system suggests that SAT scores are well-approximated by a normal distribution with standard deviation known to be 50. Calculate a
   a. 98% confidence interval.
   b. 90% confidence interval.
   c. Interpret each interval in the context of the problem. Comment on how the intervals change as you change the confidence level.
   d. Find the sample size required for a 98% confidence interval with maximum margin of error 10.

5.5 Confidence Intervals, σ Unknown

In practice, the value of σ is almost never known... but we know that we can estimate σ using s. Can we plug in s for σ? Sometimes!

Remember the Central Limit Theorem (Section 5.1)? For samples of size $n \geq 30$, $\bar{X}$ will be approximately normal even if X isn't. In this case, we can plug in s for σ:

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}.$$

That setting is pretty straightforward! Now we need to consider the setting where $n < 30$, which will require a bit of additional work.

5.5.1 The T-Distribution

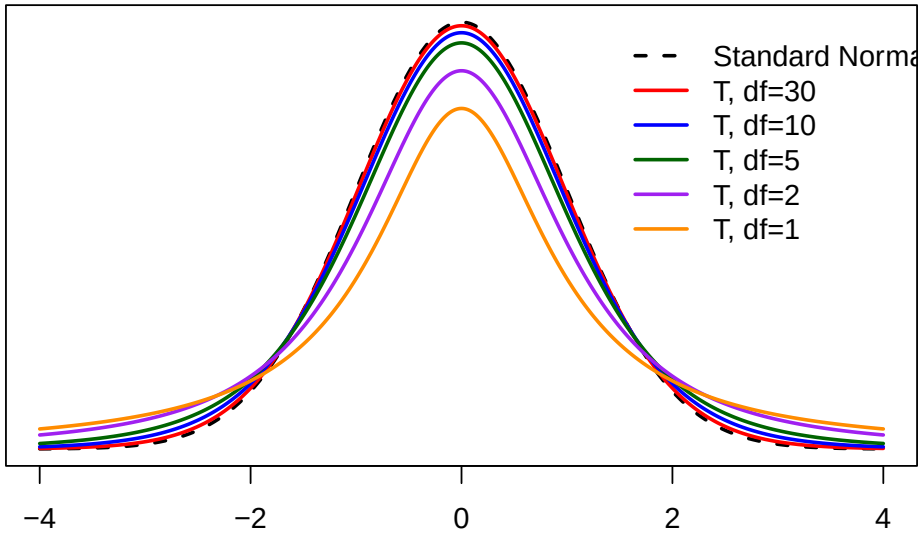
Enter: the t-distribution. If

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution (for X normal or $n \geq 30$), the slightly modified

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has what we call the **t-distribution** with $n - 1$ **degrees of freedom** (even when $n < 30$!). The only thing we need to know about degrees of freedom is that $df = n - 1$ is the t-distribution's only parameter.



The t-distribution is symmetric and always centered at 0. When $n \geq 30$, the t-distribution is approximately equivalent to the standard normal distribution. For smaller sample sizes, the t-distribution has more area in the tails (and therefore less area in the center of the distribution).

For a sample of size $n < 30$, we plug in s for σ and use a t critical value (instead of a z critical value):

$$\bar{x} \pm t_{df,\,\alpha/2}\frac{s}{\sqrt{n}}.$$

The t critical value is found through

$$P(T_{df} > t_{df,\,\alpha/2}) = \alpha/2$$

where T_{df} is the t-distribution with $df = n - 1$ degrees of freedom.

To find a t critical value, we will again use `R`, now with the command `qt(p, df)`. (Notice that this is similar to the command for the standard normal distribution, but instead of "norm" for normal it has "t" for the t-distribution.) For example, for a 98% interval with a sample size of 15,

$$100(1-\alpha) = 98 \implies \alpha = 0.02$$

Then $\alpha/2 = 0.01$ and $df = 15 - 1 = 14$.

```
qt(0.01, df=14)
```

```
## [1] -2.624494
```

which gives the t critical value $t_{14,\alpha/2} = 2.625$. Notice again that I am able to ignore the sign because our formula uses $\pm t_{df,\alpha/2}$.

As before, if you prefer you may use the applet, Rossman and Chance t Probability Calculator, instead of `R`. For this applet, enter the degrees of freedom $n - 1$ next to "df." Then check the top box under "t-value probability" and make sure the inequality is clicked to ">" . Enter the value of $\alpha/2$ for the probability. Click anywhere else on the page and the applet will automatically fill in the box under "t-value." This is your t critical value.

5.6 Confidence Intervals for a Proportion

Confidence intervals for a proportion are really similar to those for a mean! It turns out we can apply the Central Limit Theorem to the sampling distribution for a proportion. But wait - isn't our Central Limit Theorem only for means?

Think back to the binomial distribution (Section 4.3). A binomial experiment is made up of a series of Bernoulli trials, which result in 0s and 1s. If we add up these values, we get the number of successes x. If we take the mean of these

successes, we get the *proportion* of successes. In short, $\bar{x} = \hat{p}$ and we can work with the sampling distribution for a sample mean!

The mean of a Bernoulli random variable is $\mu = p$ and the standard deviation is $\sigma = \sqrt{p(1-p)}$. So if we apply the Central Limit Theorem, $\hat{p}$ is approximately normally distributed with mean

$$\mu_{\hat{p}} = p$$

and standard error

$$\sigma_{\hat{p}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$$

Each of the confidence intervals for a mean uses the same logic:

$$\text{estimate } \pm \text{ critical value } \times \text{ standard error}$$

Confidence intervals for a proportion will do the same. We do not know the true value of p for the standard error, so we will plug in $\hat{p}$.

The $100(1-\alpha)\%$ confidence interval for p is given by

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

To use this formula, we need to check that $n\hat{p} > 10$ and $n(1-\hat{p}) > 10$. Why? This relies on a normal approximation that does not work well if either of those quantities is less than or equal to 10. (This a topic which I have skipped, but the theory behind it is similar to the theory presented here for why we can use the Central Limit Theorem with proportions.)

5.7 Summary of Confidence Interval Settings

Setting 1: μ is target parameter, X is normal, σ known

- Critical value: $z_{\alpha/2}$
- Confidence interval:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Setting 2: μ is target parameter, $n \geq 30$, σ unknown

- Critical value: $z_{\alpha/2}$
- Confidence interval:

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

Setting 3: μ is target parameter, $n < 30$, σ unknown

- Critical value: $t_{df,\alpha/2}$

- Confidence interval:

$$\bar{x} \pm t_{df,\alpha/2} \frac{s}{\sqrt{n}}$$

Setting 4: p is target parameter, $n\hat{p} > 10$, $n(1-\hat{p}) > 10$

- Critical value: $z_{\alpha/2}$
- Confidence interval:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Chapter 6

Introduction to Hypothesis Testing

6.1 Chapter Overview

In this chapter, we will continue our discussion on statistical inference with a discussion on hypothesis testing. In hypothesis testing, we take a more active approach to our data by asking questions about population parameters and developing a framework to answer those questions. We will root this discussion in confidence intervals before learning about several other approaches to hypothesis testing.

Chapter Learning Outcomes/Objectives

Test one sample means using

1. confidence intervals.
2. the critical value approach.
3. the p-value approach.

This chapter's outcomes correspond to course outcomes (6) apply statistical inference techniques of parameter estimation such as point estimation and confidence interval estimation and (7) apply techniques of testing various statistical hypotheses concerning population parameters.

6.2 Logic of Hypothesis Testing

One of our goals with statistical inference is to make decisions or judgements about the value of a parameter. A confidence interval is a good starting point,

but we might also want to ask questions like

- Do cans of soda actually contain 12 oz?
- Is Medicine A better than Medicine B?

A **hypothesis** is a statement that something is true. A hypothesis test involves two (competing) hypotheses:

1. The **null hypothesis**, denoted H_0, is the hypothesis to be tested. This is the "default" assumption.
2. The **alternative hypothesis**, denoted H_A is the alternative to the null.

Note that the subscript 0 is "nought" (pronounced "not"). A **hypothesis test** helps us decide whether the null hypothesis should be rejected in favor of the alternative.

> *Example*: Cans of soda are labeled with "12 FL OZ." Is this accurate?
>
> The default, or uninteresting, assumption is that cans of soda contain 12 oz.
>
> - H_0: the mean volume of soda in a can is 12 oz.
> - H_A: the mean volume of soda in a can is NOT 12 oz.

We can write these hypotheses in words (as above) or in statistical notation. The null specifies a single value of μ

- H_0: $\mu = \mu_0$

or of p

- H_0: $p = p_0$

We call μ_0 (or p_0) the **null value**. When we run a hypothesis test, μ_0 (or p_0) will be replaced by some number. For the soda can example, the null value is 12. We would write $H_0 : \mu = 12$.

The alternative specifies a *range* of possible values for μ:

- H_A: $\mu \neq \mu_0$. "The true mean is different from the null value."

or p:

- H_A: $p \neq p_0$. "The true proportion is different from the null value."

The Logic of Hypothesis Testing

Take a random sample from the population. If the data area consistent with the null hypothesis, do not reject the null hypothesis. If the data are inconsistent with the null hypothesis *and* supportive of the alternative hypothesis, reject the null in favor of the alternative.

> *Example*: One way to think about the logic of hypothesis testing is by comparing it to the U.S. court system. In a jury trial, jurors

are told to assume the defendant is "innocent until proven guilty." Innocence is the default assumption, so

- H_0: the defendant is innocent.
- H_A: the defendant is guilty.

Like in hypothesis testing, it is not the jury's job to decide if the defendant is innocent. That should be their default assumption. They are only there to decide if the defendant is guilty or if there is not enough evidence to override that default assumption. The *burden of proof* lies on the alternative hypothesis.

Notice the careful language in the logic of hypothesis testing: we either reject, or fail to reject, the null hypothesis. We never "accept" a null hypothesis.

6.2.1 Decision Errors

- A **Type I Error** is rejecting the null when it is true. (Null is true, but we conclude null is false.)
- A **Type II Error** is not rejecting the null when it is false. (Null is false, but we do not conclude it is false.)

H_0 is

True

False

Decision

Do not reject H_0

Correct decision

Type II Error

Reject H_0

Type I Error

Correct decision

Example: In our jury trial,

- H_0: the defendant is innocent.
- H_A: the defendant is guilty.

A Type I error is concluding guilt when the defendant is innocent. A Type II error is failing to convict when the person is guilty.

How likely are we to make errors? Well, $P(\text{Type I Error}) = \alpha$, the **significance level**. (Yes, this is the same α we saw in confidence intervals!) For Type II error,

$P(\text{Type II Error}) = \beta$. This is related to the sample size calculation from the previous chapter, but is otherwise something we don't have time to cover.

We would like both α and β to be small but, like many other things in statistics, there's a trade off! For a fixed sample size,

- If we decrease α, then β will increase.
- If we increase α, then β will decrease.

In practice, we set α (as we did in confidence intervals). We can improve β by increasing sample size. Since resources are finite (we can't get enormous sample sizes all the time), we will need to consider the consequences of each type of error.

> *Example* We could think about assessing consequences through the jury trial example. Consider two possible charges:
>
> 1. Defendant is accused of stealing a loaf of bread. If found guilty, they may face some jail time and will have a criminal record.
> 2. Defendant is accused of murder. If found guilty, they will have a felony and may spend decades in prison.
>
> Since these are moral questions, I will let you consider the consequences of each type of error. However, keep in mind that we do make scientific decisions that have lasting impacts on people's lives.

Hypothesis Test Conclusions

- If the null hypothesis is rejected, we say the result is **statistically significant**. We can interpret this result with:
  - At the α level of significance, the data provide sufficient evidence to support the alternative hypothesis.
- If the null hypothesis is *not* rejected, we say the result is **not statistically significant**. We can interpret this result with:
  - At the α level of significance, the data do *not* provide sufficient evidence to support the alternative hypothesis.

Notice that these conclusions are framed in terms of the alternative hypothesis, which is either supported or not supported. We will *never* conclude the null hypothesis. Finally, when we write these types of conclusions, we will write them in the context of the problem.

6.3 Confidence Interval Approach to Hypothesis Testing

I frame this section in terms of confidence intervals for μ, but the same logic/procedure applies to confidence intervals for p.

We can use a confidence interval to help us weigh the evidence against the null hypothesis. A confidence interval gives us a range of *plausible* values for μ. If the null value is in the interval, then μ_0 is a plausible value for μ. If the null value is *not* in the interval, then μ_0 is *not* a plausible value for μ.

1. State null and alternative hypotheses.
2. Decide on significance level α. Check assumptions (decide which confidence interval setting to use).
3. Find the critical value.
4. Compute confidence interval.
5. If the null value is *not* in the confidence interval, reject the null hypothesis. Otherwise, do not reject.
6. Interpret results in the context of the problem.

Example: Is the average mercury level in dolphin muslces different from $2.5\mu g/g$? Test at the 0.05 level of significance. A random sample of 19 dolphins resulted in a mean of $4.4\mu g/g$ and a standard deviation of $2.3\mu g/g$.

1. $H_0 : \mu = 2.5$ and $H_A : \mu \neq 2.5$.
2. Significance level is $\alpha = 0.05$. The value of σ is unknown and $n = 19 < 30$, so we are in setting 3.
3. For setting 3, the critical value is $t_{df,\alpha/2}$. Here, $df = n - 1 = 18$ and $\alpha/2 = 0.025$:

```
qt(0.025, 18)
```

```
## [1] -2.100922
```

4. The confidence interval is

$$\bar{x} \pm t_{df,\alpha/2} \frac{s}{\sqrt{n}} \tag{6.1}$$

$$4.4 \pm 2.101 \frac{2.3}{\sqrt{19}} \tag{6.2}$$

$$4.4 \pm 1.109 \tag{6.3}$$

or $(3.29, 5.51)$.

5. Since the null value, 2.5, is not in the interval, it is *not* a plausible value for μ (at the 95% level of confidence). Therefore we reject the null hypothesis.
6. At the 0.05 level of significance, the data provide sufficient evidence to conclude that the true mean mercury level in dolphin muscles is *greater than* $2.5\mu g/g$.

Note: The alternative hypothesis is "not equal to," but we conclude "greater than" because all of the plausible values in the confidence interval are greater than the null value.

6.4 Critical Value Approach to Hypothesis Testing

We learned about critical values when we discussed confidence intervals. Now, we want to use these values directly in a hypothesis test. We will compare these values to a value based on the data, called a **test statistic**.

Idea: the null is our "default assumption." If the null is true, how likely are we to observe a sample that looks like the one we have? If our sample is very inconsistent with the null hypothesis, we want to reject the null hypothesis.

6.4.1 Test statistics

Test statistics are similar to z- and t-scores:

$$\text{test statistic} = \frac{\text{point estimate} - \text{null value}}{\text{standard error}}.$$

In fact, they serve a similar function in converting a variable $\bar{X}$ into a distribution we can work with easily.

- **Setting 1**: μ is target parameter, X is normal, σ known

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- **Setting 2**: μ is target parameter, $n \geq 30$, σ unknown

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- **Setting 3**: μ is target parameter, $n < 30$, σ unknown

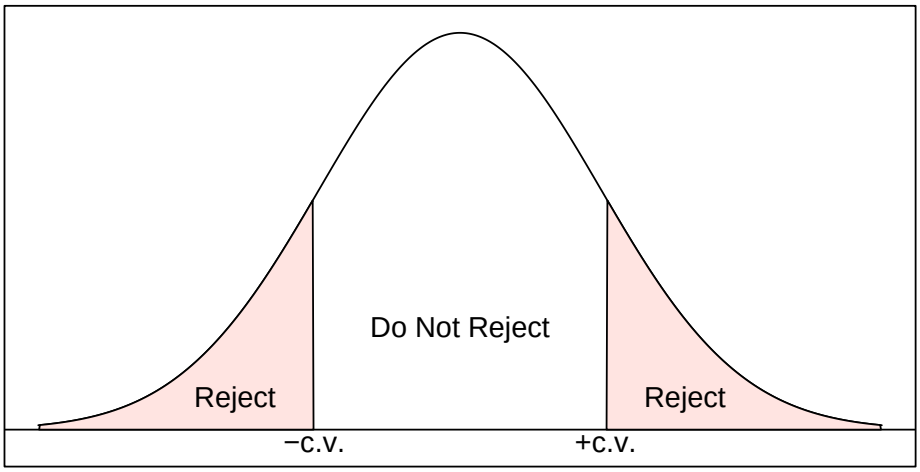
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- **Setting 4**: p is target parameter, $np > 10$, $n(1-p) > 10$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

  Notice that we plug in p_0 for the standard error! This is different from how we dealt with the standard error when calculating confidence intervals.

The set of values for the test statistic that cause us to reject H_0 is the **rejection region**. The remaining values are the **nonrejection region**. The value that separates these is the critical value!



Steps:

1. State the null and alternative hypotheses.
2. Determine the significance level α. Check assumptions (decide which setting to use).
3. Compute the value of the test statistic.
4. Determine the critical values.
5. If the test statistic is in the rejection region, reject the null hypothesis. Otherwise, do not reject.
6. Interpret results.

   Example: Is the average mercury level in dolphin muslces different from $2.5\mu g/g$? Test at the 0.05 level of significance. A random sample of 19 dolphins resulted in a mean of $4.4\mu g/g$ and a standard deviation of $2.3\mu g/g$.

   1. $H_0 : \mu = 2.5$ and $H_A : \mu \neq 2.5$.
   2. Significance level is $\alpha = 0.05$. The value of σ is unknown and $n = 19 < 30$, so we are in setting 3.
   3. The test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{6.4}$$

$$= \frac{4.4 - 2.5}{2.3/\sqrt{19}} \tag{6.5}$$

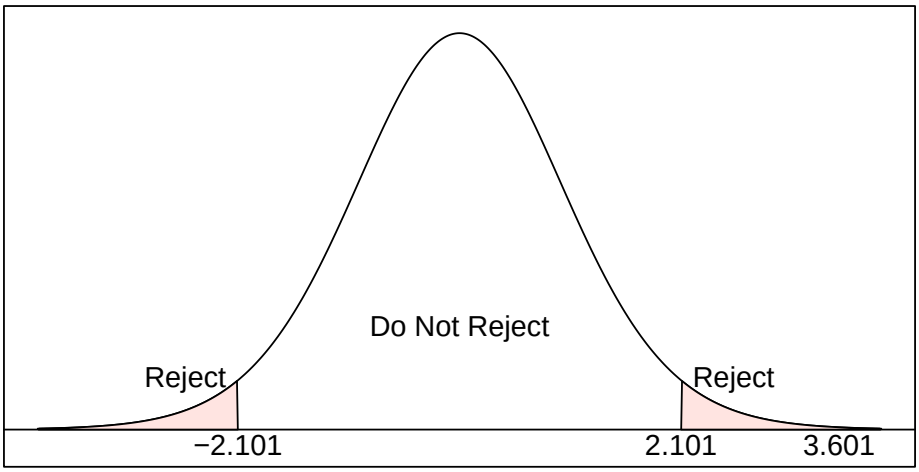
$$= 3.601 \tag{6.6}$$

   4. The critical value is $t_{df,\alpha/2}$. Here, $df = n - 1 = 18$ and $\alpha/2 = 0.025$:

```
qt(0.025, 18)
```

```
## [1] -2.100922
```

5. The test statistic is in the rejection region, so we will reject the null hypothesis:



6. At the 0.05 level of significance, the data provide sufficient evidence to conclude that the true mean mercury level in dolphin muscles is greater than $2.5\mu g/g$.

Notice that this is the same conclusion we came to when we used the confidence interval approach. These approaches are exactly equivalent!

6.5 P-Value Approach to Hypothesis Testing

If the null hypothesis is true, what is the probability of getting a random sample that is as inconsistent with the null hypothesis as the random sample we got? This probability is called the **p-value**.

Example: Is the average mercury level in dolphin muslces different from $2.5\mu g/g$? Test at the 0.05 level of significance. A random sample of 19 dolphins resulted in a mean of $4.4\mu g/g$ and a standard deviation of $2.3\mu g/g$.

Probability of a sample *as inconsistent* as our sample is $P(t_{df}$ is as extreme as the test statistic). Consider

$$P(t_{18} > 3.6) = 0.001$$

but we want to think about the probability of being "as extreme" in *either direction* (either tail), so

$$\text{p-value} = 2P(t_{18} > 3.6) = 0.002$$

If p-value $< \alpha$, reject the null hypothesis. Otherwise, do not reject.

6.5.1 P-Values

- **Setting 1**: μ is target parameter, X is normal, σ known

  $$2P(Z > |z|)$$

  where z is the test statistic.

- **Setting 2**: μ is target parameter, $n \geq 30$, σ unknown

  $$2P(Z > |z|)$$

  where z is the test statistic.

- **Setting 3**: μ is target parameter, $n < 30$, σ unknown

  $$2P(t_{df} > |t|)$$

  where t is the test statistic.

- **Setting 4**: p is target parameter, $np > 10$, $n(1-p) > 10$

  $$2P(Z > |z|)$$

  where z is the test statistic.

Note: $|a|$ is the "absolute value" of a. The absolute value takes a number and throws away the sign, so $|2| = 2$ and $|-3| = 3$.

Steps:

1. State the null and alternative hypotheses.
2. Determine the significance level α. Check assumptions (decide which setting to use).
3. Compute the value of the test statistic.
4. Determine the p-value.
5. If p-value $< \alpha$, reject the null hypothesis. Otherwise, do not reject.
6. Interpret results.

We often use p-values instead of the critical value approach because they are meaningful on their own (they have a direct interpretation).

Example: For the dolphins,

1. $H_0 : \mu = 2.5$ and $H_A : \mu \neq 2.5$.
2. Significance level is $\alpha = 0.05$. The value of σ is unknown and $n = 19 < 30$, so we are in setting 3.
3. The test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{6.7}$$

$$= \frac{4.4 - 2.5}{2.3/\sqrt{19}} \tag{6.8}$$

$$= 3.601 \tag{6.9}$$

4. The p-value is

$$2P(t_{df} > |t|) - 2P(t_{18} > 3.601) = 0.002$$

5. Since p-value $= 0.002 < \alpha = 0.05$, reject the null hypothesis.
6. At the 0.05 level of significance, the data provide sufficient evidence to conclude that the true mean mercury level in dolphin muscles is greater than $2.5\mu g/g$.

As before, this is the same conclusion we came to when we used the confidence interval and critical value approaches. All of these approaches are exactly equivalent.

Chapter 7

Inference: Comparing Parameters

7.1 Chapter Overview

In this chapter, we extend the concepts from Chapter 6 to answer questions like "is there a difference between these means?" We will also consider hypothesis tests for whether a sample represents the population or closely matches a particular distribution.

Chapter Learning Outcomes/Objectives

1. Test paired data and two sample means using
   a. confidence intervals.
   b. the critical value approach.
   c. the p-value approach.
2. Interpret an ANOVA.
3. Use the Bonferroni correction to conduct multiple comparisons.

7.2 Hypothesis Tests for Two Proportions

Sometimes, we might like to *compare* two proportions. We do this by looking at their *difference*: $p_1 - p_2$. This is going to be fairly similar to the tests we used for a single proportion. Let n_1 be the sample size for the first group and p_1 the proportion for the first group. Similarly, let n_2 be the sample size for the second group and p_2 the proportion for the second group.

Conditions:

1. Independence within and between groups (generally satisfied if the data are from random samples or a randomized experiment).
2. We need $n_1 p_1 > 10$ and $n_1(1-p_1) > 10$ **and** $n_2 p_2 > 10$ and $n_2(1-p_2) > 10$

If these conditions are satisfied, the standard error is

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

and we can calculate confidence intervals and perform hypothesis tests on $p_1 - p_2$.

7.2.1 Confidence Intervals for Two Proportions

A $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is

$$\hat{p_1} - \hat{p_2} \pm z_{\alpha/2} \times \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

7.2.2 Critical Values, Test Statistics, and P-Values

Often, we are interested in checking whether $p_1 = p_2$, which results in a null hypothesis of $H_0 : p_1 - p_2 = 0$ (where the null value is zero). In this case, we use a *pooled proportion* to estimate p in the standard error.

This pooled proportion is calculated as

$$\hat{p}_{\text{pooled}} = \frac{\text{total number of successes}}{\text{total number of cases}} = \frac{\hat{p_1} n_1 + \hat{p_2} n_2}{n_1 + n_2}$$

which makes the standard error in this case

$$\text{Standard Error} = \sqrt{\frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_1} + \frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_2}}$$

The critical value is $z_{\alpha/2}$. The test statistic is

$$z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_1} + \frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_2}}}$$

and the p-value is

$$2P(Z > |z|)$$

where z is the test statistic.

Steps:

1. State the null and alternative hypotheses.
2. Determine the significance level α. Check assumptions, $n_1 p_1 > 10$ and $n_1(1-p_1) > 10$ **and** $n_2 p_2 > 10$ and $n_2(1-p_2) > 10$.
3. Compute the value of the test statistic.
4. Determine the critical value or p-value.
5. For the *critical value approach*: If the test statistic is in the rejection region, reject the null hypothesis. For the *p-value approach*: If p-value $< \alpha$, reject the null hypothesis. Otherwise, do not reject.
6. Interpret results.

7.3 Hypothesis Tests for Two Means

What if we wanted to compare two means? We begin by discussing paired samples. This will feel very familiar, since it's essentially the same as hypothesis testing for a single mean. Then we will move on to independent samples, which will require a couple of adjustments.

7.3.1 Paired Samples

Sometimes there is a special correspondence between two sets of observations. We say that two sets of observations are **paired** if each observation has a natural connection with exactly one observation in the other data set. Consider the following data from 30 students given a pre- and post-test on a course concept:

Student	Pre-Test	Post-Test
1	52	70
2	71	98
3	13	65
...	...	...
30	48	81

The natural connection between "pre-test" and "post-test" is the student who took each test! Often, paired data will involve similar measures taken on the *same item or individual*. We *pair* these data because we want to compare two means, but we also want to account for the pairing.

Why? Consider: If a student got a 13% on the pre-test, I would love to see them get a 60% on the post-test - that's a huge improvement! But if a student got an 82% on the pre-test, I would *not* like to see them get a 60% on the post-test. Pairing the data lets us account for this connection.

So what do we do with paired data? Fortunately, this part is easy! We start by taking the difference between the two sets of observations. In the pre- and

post-test example, I will take the pre-test score and subtract the post-test score:

Student	Pre-Test	Post-Test	**Difference**
1	52	70	**18**
2	71	98	**27**
3	13	65	**52**
...	...	...	...
30	48	81	**33**

Then, we do a test of a *single mean* on the differences where

- $H_0 : \mu_{\text{d}} = 0$
- $H_A : \mu_{\text{d}} \neq 0$

Note that the subscript "d" denotes "difference." We will use the exact same test(s) as in the previous sections:

- **Setting 1**: μ_{d} is target parameter, the differences are approximately normal, σ_{d} known

  $$z = \frac{\bar{x}_{\text{d}}}{\sigma_{\text{d}}/\sqrt{n_{\text{d}}}}$$

  and the p-value is

  $$2P(Z > |z|)$$

  where z is the test statistic.

- **Setting 2**: μ_{d} is target parameter, $n_{\text{d}} \geq 30$, σ_{d} unknown

  $$z = \frac{\bar{x}_{\text{d}}}{s_{\text{d}}/\sqrt{n_{\text{d}}}}$$

  and the p-value is

  $$2P(Z > |z|)$$

  where z is the test statistic.

- **Setting 3**: μ_{d} is target parameter, $n_{\text{d}} < 30$, σ_{d} unknown

  $$t = \frac{\bar{x}_{\text{d}}}{s_{\text{d}}/\sqrt{n_{\text{d}}}}$$

  and the p-value is

  $$2P(t_{df} > |t|)$$

  where t is the test statistic.

Here, n_{d} is the number of pairs.

Steps:

1. State the null and alternative hypotheses.

2. Determine the significance level α. Check assumptions (decide which setting to use).
3. Compute the value of the test statistic.
4. Determine the critical values or p-value.
5. For the *critical value approach*: If the test statistic is in the rejection region, reject the null hypothesis. For the *p-value approach*: If p-value $< \alpha$, reject the null hypothesis. Otherwise, do not reject.
6. Interpret results.

7.3.2 Independent Samples

In **independent samples**, the sample from one population does not impact the sample from the other population. In short, we take two *separate samples* and compare them.

- $H_0 : \mu_1 = \mu_2 \quad \rightarrow \quad H_0 : \mu_1 - \mu_2 = 0$
- $H_A : \mu_1 \neq \mu_2 \quad \rightarrow \quad H_A : \mu_1 - \mu_2 \neq 0$

If we use $\bar{x}$ to estimate μ, intuitively we might use $\bar{x}_1 - \bar{x}_2$ to estimate $\mu_1 - \mu_2$. To do this, we need to know something about the sampling distribution of $\bar{x}_1 - \bar{x}_2$.

Consider: if X_1 is Normal(μ_1, σ_1) and X_2 is Normal(μ_2,σ_2) with σ_1 and σ_2 are known, then for independent samples of size n_1 and n_2,

- $\bar{X}_1 - \bar{X}_2$ is Normal($\mu_{\bar{X}_1 - \bar{X}_2}$, $\sigma_{\bar{X}_1 - \bar{X}_2}$).
- $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$
- $\sigma_{\bar{X}_1 - \bar{X}_2} = \sigma_1 - \sigma_2$

so then

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1/n_1 - \sigma_2/n_2}}$$

has a standard normal distribution. But, as we mentioned earlier, we rarely work in that setting where the population standard deviation is known. Instead, we will use s_1 and s_2 to estimate σ_1 and σ_2. For independent samples of size n_1 and n_2,

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1/n_1 - s_2/n_2}}$$

has a t-distribution with degrees of freedom

$$\Delta = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

rounded *down* to the nearest whole number. (Note that Δ is the uppercase Greek letter, "delta.") If $n_1 = n_2$, this simplifies to

$$\Delta = (n-1)\left(\frac{(s_1^2 + s_2^2)^2}{s_1^4 + s_2^4}\right)$$

Tip: Generally, people do not calculate Δ by hand. Instead, we use a computer to do these kinds of tests.

The Two-Sample T-Test

Assumptions:

- Simple random samples.
- Independent samples.
- Normal populations or large ($n \geq 30$) samples.

Steps for Critical Value Approach:

1. $H_0 : \mu_1 - \mu_2 = 0$ and $H_A : \mu_1 - \mu_2 \neq 0$
2. Check assumptions; select the significance level α.
3. Compute the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1/n_1 - s_2/n_2}}$$

   Note that we assume under the null hypothesis that $\mu_1 - \mu_2 = 0$, which is why we replace this quantity with 0 in the test statistic.
4. The critical value is $\pm t_{df,\alpha/2}$ with $df = \Delta$.
5. If the test statistic falls in the rejection region, reject the null hypothesis.
6. Interpret in the context of the problem.

Steps for P-Value Approach:

1. $H_0 : \mu_1 - \mu_2 = 0$ and $H_A : \mu_1 - \mu_2 \neq 0$
2. Check assumptions; select the significance level α.
3. Compute the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1/n_1 - s_2/n_2}}$$

   Note that we assume under the null hypothesis that $\mu_1 - \mu_2 = 0$, which is why we replace this quantity with 0 in the test statistic.
4. The p-value is $2P(t_{df} > |t|)$ with $df = \Delta$.
5. If p-value $< \alpha$, reject the null hypothesis.
6. Interpret in the context of the problem.

Notice that the only difference between the critical value and p-value approaches are steps 4 and 5.

> *Example*: Researchers wanted to detemine whether a dymanic or static approach would impact the time needed to complete neurosurgeries. The experiment resulted in the following data from simple random samples of patients:

Dynamic	Static
$\bar{x}_1 = 394.6$	$\bar{x}_2 = 468.3$

Dynamic	Static
$s_1 = 84.7$	$s_2 = 38.2$
$n_1 = 14$	$n_2 = 6$

Times are measured in minutes. Assume X_1 and X_2 are reasonably normal.

1. $H_0 : \mu_1 = \mu_2$ and $H_A : \mu_1 \neq \mu_2$
2. Let $\alpha = 0.05$ (this will be our default when a significance level is not given)
   - We are told these are simple random samples.
   - There's no reason that time for a neurosurgery with the dynamic system would impact time for the static system (or vice versa), so it's reasonable to assume these samples are independent.
   - We are told to assume that X_1 and X_2 are reasonably normal.
3. The test statistic is

$$t = \frac{394.6 - 468.3}{84.7^2/14 + 38.2^2/6} = -2.681$$

4. Then

$$df = \Delta = \frac{(84.7^2/14) + (38.2^2/6)^2}{\frac{(84.7^2/14)^2}{14-1} + \frac{(38.2^2/6)^2}{6-1}} = 17$$

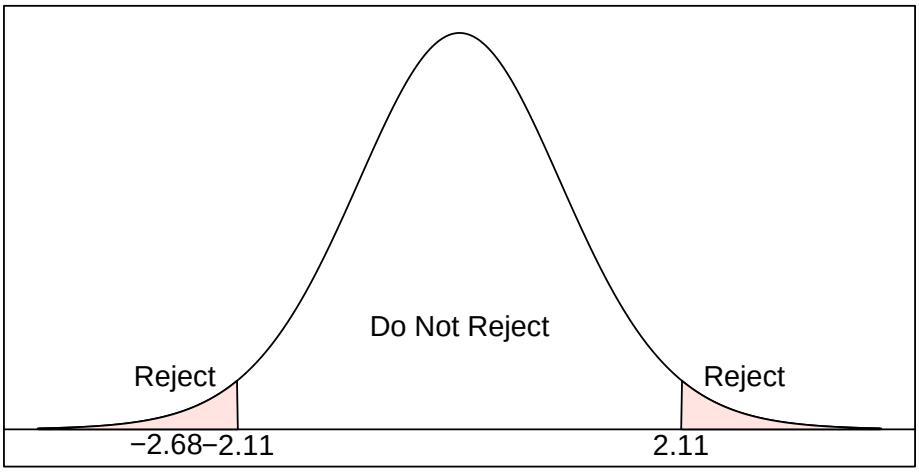
   when rounded down. The critical value is

$$t_{17,0.025} = 2.110$$

   and the p-value is

$$2P(t_{17} > |-2.681|) = 2(0.0079) = 0.0158$$

5. For the critical value approach,



Since the test statistic is in the rejection region, we reject the null hypothesis. For the p-value approach, since p-value $= 0.158 < \alpha = 0.05$, reject the null hypothesis.

6. At the 0.05 level of significance, the data provide sufficient evidence to conclude that the mean time for the dynamic system is less than the mean time for the static system.

We can also construct a $(1-\alpha)100\%$ **confidence interval** for the difference of the two population means:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{df,\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

which we interpret as we interpret other confidence intervals, including in our interpretation that we are now considering the *difference of two means**.

7.4 Analysis of Variance (ANOVA)

Now that we've examined tests for one and two means, it's natural to wonder about three or more means. For example, we might want to compare three different medications: treatment 1 (t_1), treatment 2 (t_2), and treatment 3 (t_3). Based on what we've learned so far, we might think to do pairwise comparisons, examining t_1 vs t_2, then t_2 vs t_3, then t_1 vs t_3. Unfortunately, this tends to increase our Type I error!

Think of it this way: if I set my confidence level to 95%, I'm setting my Type I error rate to $\alpha = 0.05$. In general terms, this means that about 1 out of every 20 times I run my experiment, I would make a type I error. If I went ahead and ran, say, 20 tests comparing two means, my *overall* Type I error rate is

going to increase - there's a pretty significant chance that at least one of those comparisons will results in a Type I error!

Instead, we will use a test that allows us to ask: "Are all these means the same?" This is called the **an**alysis **o**f **va**riance, or ANOVA.

- H_0: The mean outcome is the same across all groups.
- H_A: At least one mean differs from the rest.

In statistical notation, these hypotheses look like:

- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
- $H_A : \mu_i \neq \mu_j$ for at least one pair (i, j)

where k is the number of means being compared and the notation μ_i represents the mean for the ith group (i can take on any whole number value between 1 and k).

For ANOVA, we have three key conditions:

1. Observations are independent within and across groups.

Independence within groups is the way we've been thinking about independence already. We want to convince ourselves that for any particular group, the observations do not impact each other. For independence across groups, we want to convince ourselves that the groups do not impact each other. Note: if we have a simple random sample, this assumption is always satisfied.

2. Data within each group are approximately normal.

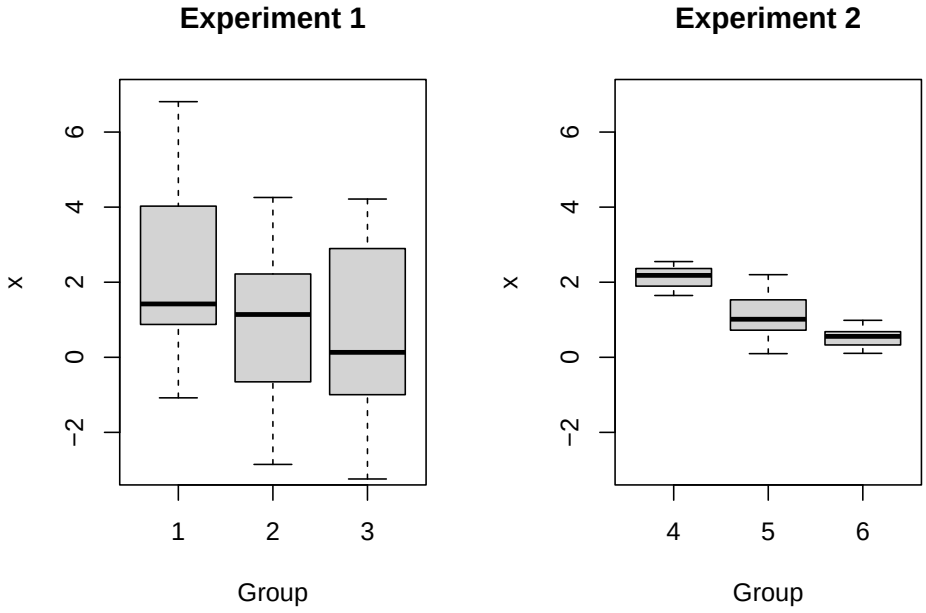
If you make a histogram of the data for each group, each histogram will look approximately bell-shaped.

3. Variability is approximately equal across groups.

Take the standard deviation for each group and check if they are approximately equal. A boxplot is an appropriate way to do this visually.

Why Variance?

You may have seen the name "analysis of variance" and wondered what the variance has to do with comparing many means. Consider the following boxplots:



Is there a difference in the means for Experiment 1? What about Experiment 2?

In fact, the means are $\mu_1 = \mu_4 = 2$, $\mu_2 = \mu_5 = 1$, and $\mu_3 = \mu_6 = 0.5$. But the variances for the Experiment 1 groups are much larger than for the Experiment 2 groups! The larger variances in Experiment 1 obscure any differences between the group means. It is for this reason that we analyze variance as part of our test for differences in means.

> Aside: Why can't we look at the data first and just test the two means that have the largest difference?
>
> When we look at the data *and then choose a test*, this inflates our Type I error rate! It's bad practice and not something we want to engage in as scientists.

In order to perform an ANOVA, we need to consider whether the sample means differ more than we would expect them to based on natural variation (remember that we expect random samples to produce slightly different sample statistics each time!). This type of variation is called **mean square between groups** or MSG. It has associated degrees of freedom $df_G = k - 1$ where k is the number of groups. Note that

$$MSG = \frac{SSG}{df_G}$$

where SSE is the **sum of squares group**. If the null hypothesis is true, variation in the sample means is due to chance. In this case, we would expect the MSG to be relatively small.

When I say "relatively small," I mean we need to compare this quantity to

something. We need some quantity that will give us an idea of how much variability to expect if the null hypothesis is true. This is the **mean square error** or MSE, which has degrees of freedom $df_E = n - k$. Again, we have the relationship that

$$MSE = \frac{SSE}{df_E}$$

where SSE is the **sum of squares error**. These calculations are very similar to the calculation for variance (and standard deviation)! (Note: we will not calculate these quantities by hand, but if you are interested in the mathematical details they are available in the OpenIntro Statistics textbook in the footnote on page 289.)

We compare these two quantities by examining their ratio:

$$F = \frac{MSG}{MSE}$$

This is the test statistic for the ANOVA.

7.4.1 The F-Distribution

The ***F*-test** relies on something called the F distribution. The F distribution has two parameters: $df_1 = df_G$ and $df_1 = df_E$. The F distribution always takes on positive values, so an *extreme* or *unusual* value for the F distribution will correspond to a large (positive) number.

When we run an ANOVA, we almost always use the p-value approach. If you are using `R` for your distributions, the command is `pf(F, df1, df2, lower.tail=FALSE)` where `F` is the test statistic.

> *Example:* Suppose I have a test with 100 observations and 5 groups. I find $MSG = 0.041$ and $MSE = 0.023$. Then
>
> $$df_G = k - 1 = 5 - 1 = 4$$
>
> and
>
> $$df_E = n - k = 100 - 5 = 95$$
>
> The test statistic is
>
> $$f = \frac{0.041}{0.023} = 1.7826$$
>
> To find the p-value using `R`, I would write the command

```
pf(1.7826, 4, 95, lower.tail=FALSE)
```

```
## [1] 0.1387132
```

> and find a p-value of 0.1387.

Here is a nice F-distribution applet. For this applet, $\nu_1 = df_1$ and $\nu_2 = df_2$. Plug in your F test statistic where it indicates "x =" and your p=value will appear in the red box next to "P(X>x)." When you enter your degrees of freedom, a visualization will appear similar to those in the Rossman and Chance applets we used previously.

The ANOVA Table

Generally, when we run an ANOVA, we create an ANOVA table (or we have software create one for us!). This table looks something like this

	df	Sum of Squares	Mean Squares	F Value	P-Value
group	df_G	SSG	MSG	F	p-value
error	df_E	SSE	MSE		

Example: chick weights

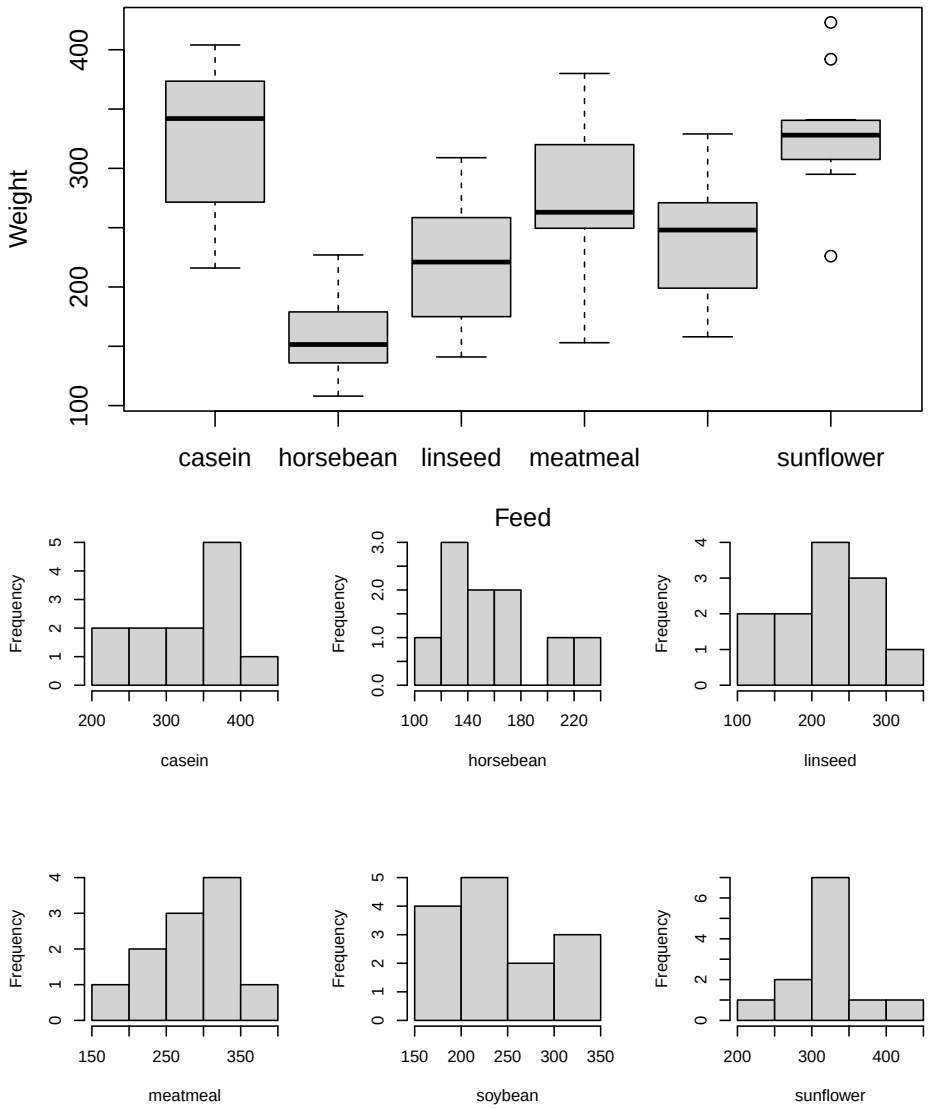
`R` has data on the weights of chicks fed six different feeds (diets). Assume these data are based on a random sample of chicks. There are $n = 71$ total observations and $k = 6$ different feeds. Let's assume we want to test with a 0.05 level of significance.

The ANOVA hypotheses are

- H_0: the mean weight is the same for all six feeds.
- H_A: at least one feed has a mean weight that differs.

The summaries for these data are

```
##         casein horsebean linseed meatmeal soybean sunflower
## n        12.00     10.00   12.00    11.00   14.00     12.00
## Mean    323.58    160.20  218.75   276.91  246.43    328.92
## Std Dev  64.43     38.63   52.24    64.90   54.13     48.84
```



The group sizes are relatively small, so it's difficult to determine how far from normality these data are based on the histograms. We may also run into some issues with constant variance. However, for the sake of the example, let's push ahead with the ANOVA! Since we usually use software to calculate ANOVAs, I've used `R` to create the following ANOVA table:

```
## Analysis of Variance Table
##
## Response: chickwts$weight
##               Df Sum Sq Mean Sq F value    Pr(>F)
```

```
## chickwts$feed  5 231129   46226  15.365 5.936e-10 ***
## Residuals     65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the table, we can confirm that $df_G = 6 - 1 = 5$ and $df_E = 71 - 6 = 65$. The F test statistic is

$$MSG/MSE = 46226/3009 = 15.365$$

Finally, the p-value is 5.936×10^{-10}. Clearly $5.936 \times 10^{-10} < \alpha = 0.05$, so we will reject the null hypothesis and conclude that at least one of the feed groups has a mean weight that differs.

7.4.2 Multiple Comparisons and Type I Error Rate

Let's return for a moment to our ANOVA hypotheses:

- H_0: The mean outcome is the same across all groups.
- H_A: At least one mean differs from the rest.

If we reject H_0 and conclude that "at least one mean differs from the rest," how do we determine which mean(s) differ? *If* we reject H_0, we will perform a series of two-sample t-tests. But wait! What about the Type I error? Isn't this exactly what we decided we couldn't do when we introduced ANOVA?

In order to avoid this increased Type I error rate, we run these **mulitple comparisons** with a modified significance level. There are several ways to do this, but the most common way is with the **Bonferroni correction**. Here, if we want to test at the $100(1-\alpha)$ level of significance, we run each of our pairwise comparisons with

$$\alpha^* = \alpha/K$$

where K is the number of comparisons being considered. For k groups, there are

$$K = \frac{k(k-1)}{2}$$

possible pairwise comparisons.

For these comparisons, we use a special pooled estimate of the standard deviation, s_{pooled} in place of s_1 and s_2:

$$\text{standard error} = \sqrt{\frac{s^2_{\text{pooled}}}{n_1} + \frac{s^2_{\text{pooled}}}{n_2}}$$

Other than changing α to α^* and the standard error to this new formula, the test is exactly the same as that discussed in the previous section. Note that

$$s_{\text{pooled}} = \sqrt{MSE}$$

and the degrees of freedom is df_E.

Example: chick weights

Let's extend our discussion on the chick weights to multiple comparisons. Since we were able to conclude that at least one feed has a weight that differs, we want to find out where the difference(s) lie!

We will test all possible pairwise comparisons. This will require $K = \frac{6(6-1)}{2} = 15$ tests. The pooled standard deviation is $s_{pooled} = \sqrt{3009} \approx 54.85$. Let's walk through the test of casein ($\bar{x}_1 = 323.58, n = 12$) vs horsebean ($\bar{x}_2 = 160.20, n = 10$):

- $H_0 : \mu_1 = \mu_2$
- $H_A : \mu_1 \neq \mu_2$

The estimated difference and standard error are

$$\bar{x}_1 - \bar{x}_2 = 323.58 - 160.20 = 163.38 \qquad SE = \sqrt{\frac{54.85^2}{11} + \frac{54.85^2}{9}} = 25.65$$

which results in a test statistic of $t = 6.37$ and a p-value of 1.11×10^{-8}. We then compare this to $\alpha^* = 0.05/15 = 0.0033$. Since the p-value of $1.11 \times 10^{-8} < \alpha^* = 0.0033$, we reject the null hypothesis and conclude there is a significant difference in mean chick weight between the casein and horsebean feeds.

In order to complete the pairwise comparisons, we would then run the remaining 14 tests. I will leave this as an optional exercise for the particularly motivated student.

Note: occasionally, we may reject H_0 in the ANOVA but may fail to find any statistically significant differences when performing multiple comparisons with the Bonferroni correction. This is ok! It just means we were unable to identify which specific groups differ.

Chapter 8

Regression and Correlation

8.1 Chapter Overview

We will extend our conversation on descriptive measures for quantitative variables to include the relationship between two variables.

Chapter Learning Outcomes/Objectives

1. Calculate and interpret a correlation coefficient.
2. Calculate and interpret a regression line.
3. Use a regression line to make predictions.

8.2 Linear Equations

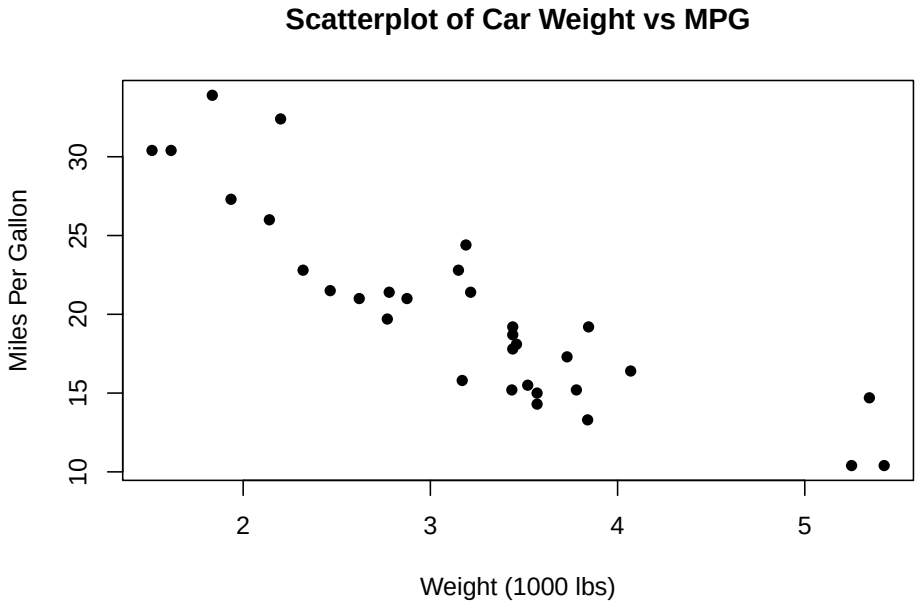
From your previous math classes, you should have a passing familiarity with linear equations like $y = mx + b$. In statistics, we write these as

$$y = b_0 + b_1 x$$

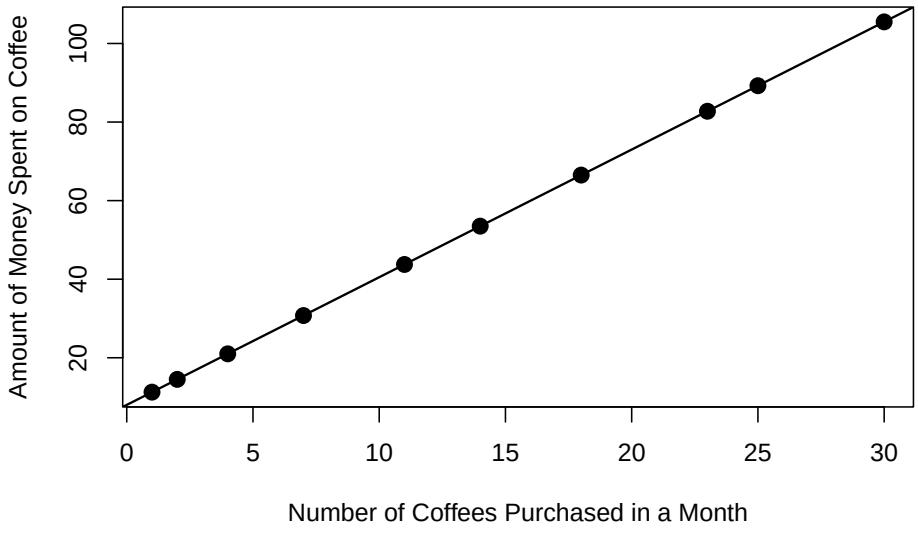
where b_0 and b_1 are constants, x is the independent variable, and y is the dependent variable. The graph of a linear function is always a (straight) line.

The **y-intercept** is b_0, the value the dependent variable takes when the independent variable $x = 0$. The **slope** is b_1, the change in y for a 1-unit change in x.

A **scatterplot** shows the relationship between two (numeric) variables.



At a glance, we can see that (in general) heavier cars have lower MPG. We call this type of data **bivariate data**. Now consider

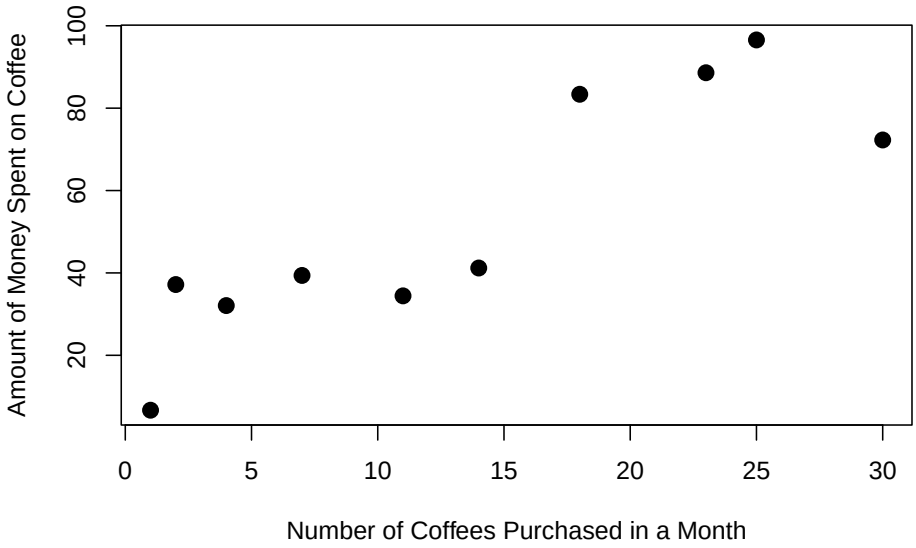


This relationship can be modeled perfectly with a straight line:

$$y = 8 + 3.25x$$

When we can do this - model a relationship perfectly - we know the exact value of y whenever we know the value of x. This is nice (we would love to be able to do this all the time!) but typically data is more complex than this.

Linear regression takes the idea of fitting a line and allows the relationship to be imperfect. Imagine in the previous scenario that you buy an \$8 pound of coffee each month and individual coffees cost \$3.25... but what if your pound of coffee didn't always cost \$8? Or your coffee drinks didn't always cost \$3.25? In this case, you might get a plot that looks something like this:



The linear regression line looks like

$$y = \beta_0 + \beta_1 x + \epsilon$$

- β is the Greek letter "beta."
- β_0 and β_1 are constants.
- Error (the fact that the points don't all line up perfectly) is represented by ϵ.

Think of this as the 2-dimensional version of a point estimate!

We estimate β_0 and β_1 using data and denote the estimated line by

$$\hat{y} = b_0 + b_1 x$$

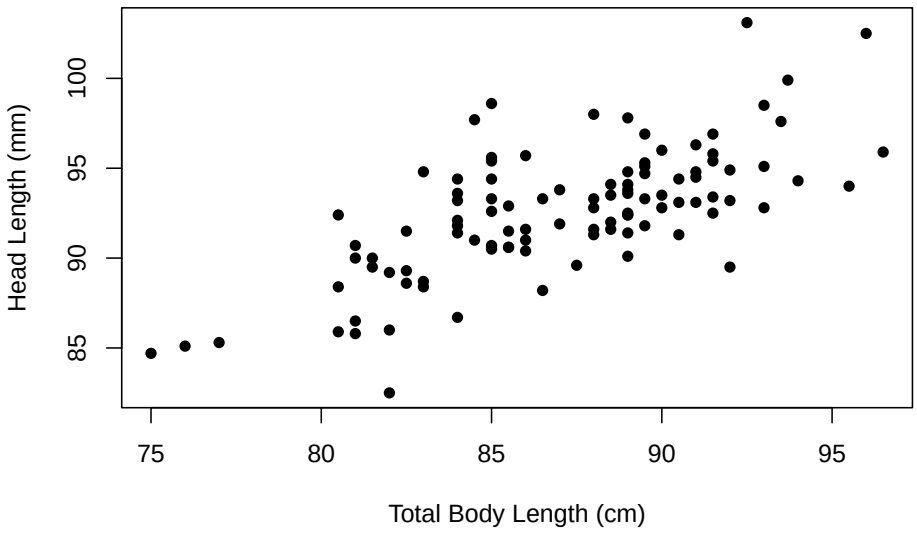
- $\hat{y}$, "y-hat," is the estimated value of y.
- b_0 is the estimate for β_0.
- b_1 is the estimate for β_1.

We drop the error term ϵ when we estimate the constants for a regression line; we assume that the mean error is 0, so *on average* we can ignore this error.

We use a regression line to make predictions about y using values of x.

- y is the **response variable**.
- x is the **predictor variable**.

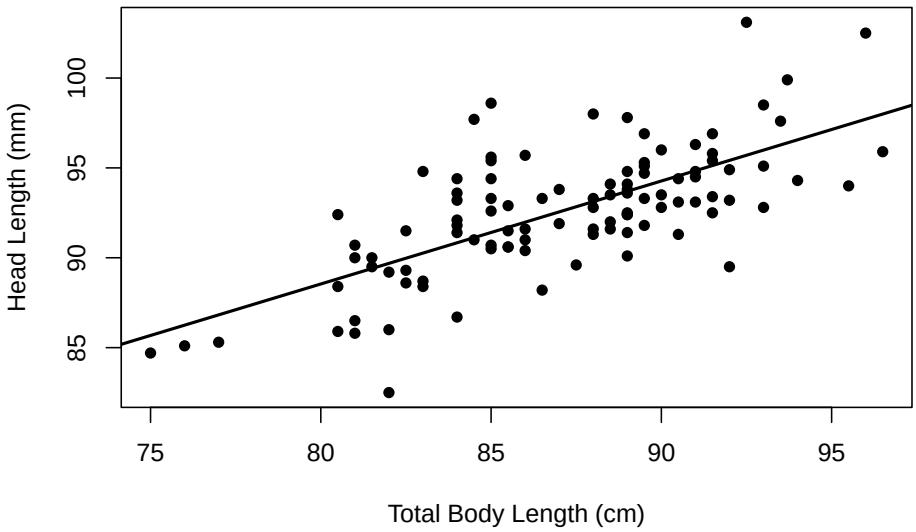
Example: (from OpenIntro Statistics 8.1.2) Researchers captured 104 brushtail possums and took a variety of body measurements on each before releasing them back into the wild. We consider two measurements for each possum: total body length and head length.



Clearly, the relationship isn't perfectly linear, but there does appear to be some kind of linear relationship (as body length increases, head length also increases). We want to try to use body length (x) to predict head length (y).

The regression model for these data is

$$\hat{y} = 42.7 + 0.57x$$



To predict the head length for a possum with a body length of 80cm, we just need to plug in 80 for body length (x):

$$\hat{y} = 42.7 + 0.57(80) = 88.3\text{mm}.$$

Note: because the regression line is built using the data's original units (cm for body length, mm for head length), the regression line will preserve those units. That means that when we plugged in a value in cm, the equation spit out a predicted value in mm.

8.3 Correlation

We've talked about the strength of linear relationships, but it would be nice to formalize this concept. The **correlation** between two variables describes the strength of their linear relationship. It always takes values between -1 and 1. We denote the correlation (or correlation coefficient) by R:

$$R = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x} \times \frac{y_i - \bar{y}}{s_y}\right)$$

where s_x and s_y are the respective standard deviations for x and y. The sample size n is the total number of (x, y) pairs.

Example: Consider

x	y
1	3
2	3

x	y
3	4
$\bar{x} = 2$	$\bar{y} = 3.333$
$s_x = 1$	$s_y = 0.577$

Like we did with variance/standard deviation, I recommend using a table to calculate the correlation between x and y:

$x - \bar{x}$	$\frac{x-\bar{x}}{s_x}$	$y - \bar{y}$	$\frac{y-\bar{y}}{s_y}$	$\frac{x-\bar{x}}{s_x} \times \frac{y-\bar{y}}{s_y}$
-1	-1	-0.333	-0.577	0.577
0	0	-0.333	-0.577	0.000
1	1	0.667	1.155	1.155
				sum = 1.732

So $R = \frac{1}{3-1}(1.732) = 0.866$

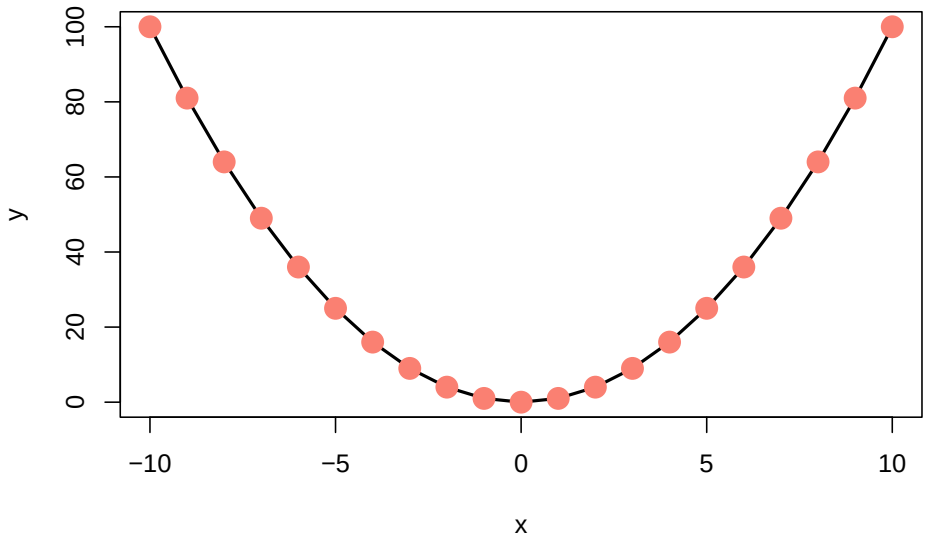
Correlations

- close to -1 suggest strong, negative linear relationships.
- close to +1 suggest strong, positive linear relationships.
- close to 0 have little-to-no linear relationship.

Note: the sign of the correlation will match the sign of the slope!

- If $R < 0$, there is a downward trend and $b_1 < 0$.
- If $R > 0$, there is an upward trend and $b_1 > 0$.
- If $R \approx 0$, there is no relationship and $b_1 \approx 0$.

A final note: correlations only represent *linear* trends. Consider the following scatterplot:



Obviously there's a strong relationship between x and y. In fact, there's a perfect relationship here: $y = x^2$. But the *correlation* between x and y is 0! This is one reason why it's important to examine the data both through visual and numeric measures.

8.4 Finding a Regression Line

Residuals are the leftover *stuff* (variation) in the data after accounting for model fit:

$$\text{data} = \text{prediction} + \text{residual}$$

Each observation has its own residual. The residual for an observation (x, y) is the difference between observed (y) and predicted $(\hat{y})$:

$$e = y - \hat{y}$$

We denote the residuals by e and find $\hat{y}$ by plugging x into the regression equation. If an observation lands above the regression line, $e > 0$. If below, $e < 0$.

When we estimate the parameters for the regression, our goal is to get each residual as close to 0 as possible. We might think to try minimizing

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

but that would just give us very large negative residuals. As with the variance,

we will use squares to shift the focus to magnitude:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{8.1}$$

$$= \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 \tag{8.2}$$

This will allow us to shrink the residuals toward 0: the values b_0 and b_1 that minimize this will make up our regression line.

This is a calculus-free course, so we'll skip the proof of the minimization part. The slope can be estimated as

$$b_1 = \frac{s_y}{s_x} \times R$$

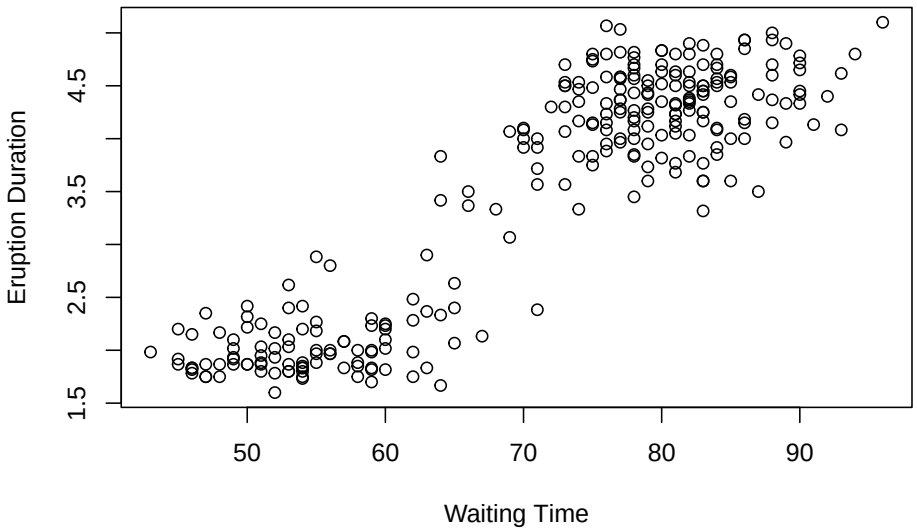
and the intercept as

$$b_0 = \bar{y} - b_1 \bar{x}$$

8.4.1 Coefficient of Determination

With the correlation and regression line in hand, we will add one last piece for considering the fit of a regression line. The **coefficient of determination**, R^2, is the square of the correlation coefficient. This value tells us how much of the variability around the regression line is accounted for by the regression. An easy way to interpret this value is to assign it a letter grade. For example, if $R^2 = 0.84$, the predictive capabilities of the regression line get a B.

> *Example*: Consider two measurements taken on the Old Faithful Geyser in Yellowstone National Park: `eruptions`, the length of each eruption and `waiting`, the time between eruptions. Each is measured in minutes.



There does appear to be some kind of linear relationship here, so we will see if we can use the wait time to predict the eruption duration. The sample statistics for these data are

	`waiting`	`eruptions`
mean	$\bar{x} = 70.90$	$\bar{y} = 3.49$
sd	$s_x = 13.60$	$s_y = 1.14$
		$R = 0.90$

Since we want to use wait time to predict eruption duration, wait time is x and eruption duration is y. Then

$$b_1 = \frac{1.14}{13.60} \times 0.90 \approx 0.076$$

and

$$b_0 = 3.49 - 0.076 \times 70.90 \approx -1.87$$

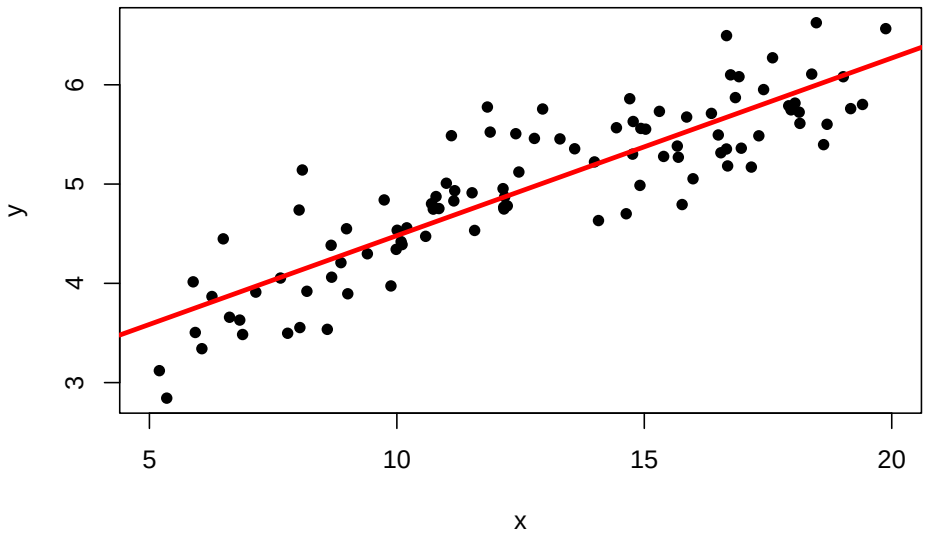
so the estimated regression line is

$$\hat{y} = -1.87 + 0.076x$$

To interpret b_1, the slope, we would say that for a one-minute increase in waiting time, we would predict a 0.076 minute increase in eruption duration. The intercept is a little bit trickier. Plugging in 0 for x, we get a predicted eruption duration of -1.87 minutes. There are two issues with this. First, a negative eruption duration doesn't make sense... but it also doesn't make sense to have a waiting time of 0 minutes.

It's important to stop and think about our predictions. Sometimes, the numbers don't make sense and it's easy to see that there's something wrong with the prediction. Other times, these issues are more insidious. Usually, all of these issues result from what we call *extrapolation*, applying a model estimate for values outside of the data's range for x. Our linear model is only an approximation, and we don't know anything about how the relationship outside of the scope of our data!

Consider the following data with the best fit line drawn on the scatterplot.



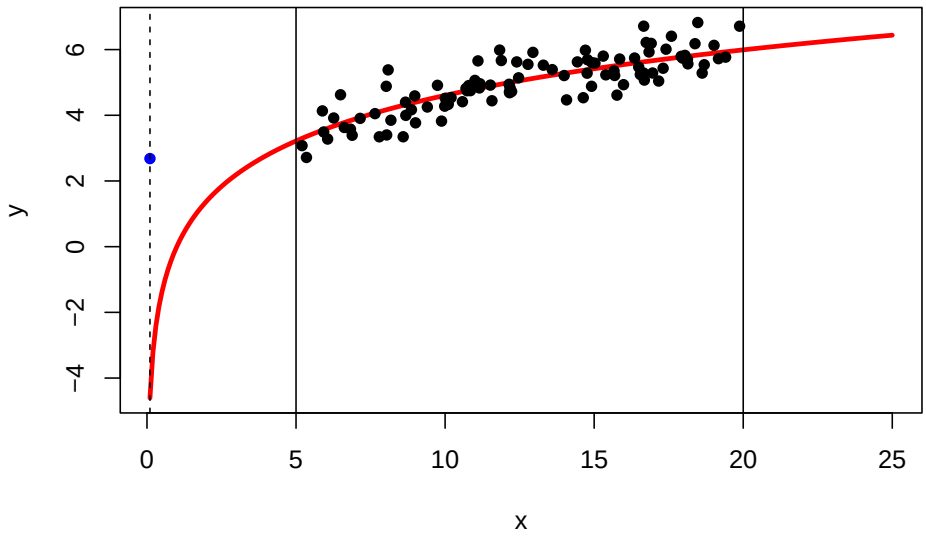
The best fit line is

$$\hat{y} = 2.69 + 0.179x$$

and the correlation is $R = 0.877$. Then the coefficient of determination is $R^2 = 0.767$ (think: a C grade), so the model has decent predictive capabilities. More precisely, the model accounts for 76.7% of the variability about the regression line. Now suppose we wanted to predict the value of y when $x = 0.1$:

$$\hat{y} = 2.66 + 0.181 \times 0.1 = 2.67$$

This seems like a perfectly reasonable number... But what if I told you that I generated the data using the model $y = 2\ln(x) + \text{random error}$? (If you're not familiar with the natural log, ln, don't worry about it! You won't need to use it.) The true (population) best-fit model would look like this:



The vertical lines at $x = 5$ and $x = 20$ show the bounds of our data. The blue dot at $x = 0.1$ is the predicted value $\hat{y}$ based on the linear model. The dashed horizontal line helps demonstrate just how far this estimate is from the true population value! This does *not* mean there's anything inherently wrong with our model. If it works well from $x = 5$ to $x = 20$, great, it's doing its job!

Appendix

Appendix A: Important Links and Additional Resources

Applets

- Normal Distribution Calculator
- Rossman and Chance Applets
- Simulating the Central Limit Theorem

Run R Online

- Run R Online
- RStudio Cloud

Appendix B: Average Deviance

The deviance of an observation from its mean is $x - \bar{x}$. We denote the deviation for the ith observation as $x_i - \bar{x}$. So the sum over all n deviances is

$$\begin{aligned}
\text{Sum of Deviances} &= \Sigma_{i=1}^{n}(x_i - \bar{x}) && (8.3)\\
&= (x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_{n-1} - \bar{x}) + (x_n - \bar{x}) && (8.4)\\
&= x_1 - \bar{x} + x_2 - \bar{x} + \cdots + x_{n-1} - \bar{x} + x_n - \bar{x} && (8.5)\\
&= x_1 + x_2 + \cdots + x_{n-1} + x_n - \bar{x} - \bar{x} - \cdots - \bar{x} - \bar{x} && (8.6)\\
&= (x_1 + x_2 + \cdots + x_{n-1} + x_n) - (\bar{x} + \bar{x} + \cdots + \bar{x} + \bar{x}) && (8.7)
\end{aligned}$$

where the first half is the sum over all of the x values and the term $(\bar{x})$ appears n times. So we can rewrite this as

$$\text{Sum of Deviances} = \Sigma_{i=1}^{n}(x_i) - n\bar{x}$$

Now notice that, because $\bar{x} = \frac{\Sigma_{i=1}^{n}(x_i)}{n}$, we can multiply both sides by n to get $n\bar{x} = \Sigma_{i=1}^{n}(x_i)$ and rewrite the sum over the deviances as

$$\text{Sum of Deviances} = n\bar{x} - n\bar{x} \tag{8.8}$$

$$= 0 \tag{8.9}$$