STAT 145 Course Pack

UWL Department of Mathematics & Statistics

1.1 Introduction to Data

Scientists seek to answer questions using rigorous methods and careful observations. These observations — collected from the likes of field notes, surveys, and experiments — form the backbone of a statistical investigation and are called ***data****. Statistics is the study of how best to collect, analyze, and draw conclusions from data. In this first chapter, we focus on both the properties of data and on the collection of data. We will introduce the building blocks of statistical thinking: observations, variables, data frames, and the types of questions that statistics can help us answer.*

Key Concepts

- Recognize that a data set consists of cases and variables
- Identify variables as either categorical or quantitative
- Determine explanatory and response variables where appropriate
- Determine whether a study is observational or experimental

Why we collect data

All statistical analyses start with a question or an idea. This question or idea is posed about a *population*, in particular a population *parameter*. The main purpose of collecting data is to learn about this unknown population parameter. We often cannot measure data for every individual in a population because it would cost too much money or take too long, so a *sample*, or subset of the population, is collected instead. Using the information in the sample, a *statistic* is calculated and used to estimate the corresponding population parameter. This process of using a sample statistic to learn about a population parameter is called *statistical inference*. One of the main goals of this class is to introduce you to the procedures for making statistical inferences in a variety of settings.

A *population* is a well-defined collection of *all* individuals or objects of interest.

A *parameter* is a number that describes a population and is almost always unknown.

A *sample* is a subset of the population on which we measure data.

A *statistic* is a number that describes a sample and is almost always known.

Inference is the process of drawing conclusions about a *population parameter* from a *sample statistic*.

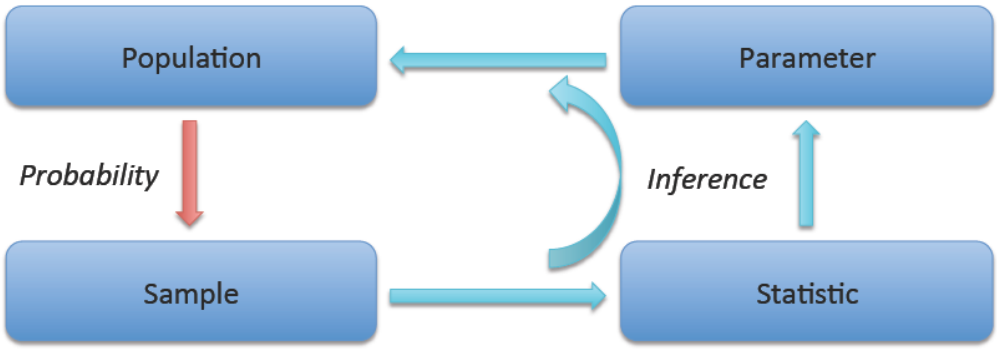


Figure 1: The cycle of statistical inference.

Data Basics

After data is gathered, it needs a place to be stored. In statistics, this is called a *data set*, or *data frame*. Data sets can be as simple as an Excel spreadsheet or as complicated as a major database. It all depends on the complexity of the question.

A *data set*, or *data frame* is used to store the set of measurements taken on individual units.

Structure: Cases and variables

In general, when data is stored in a table, each row of the table represents an *observation*, or a *case*, in the dataset. Each column of the table represents a *variable* in the dataset.

Observations, or *Cases*, are the subjects/objects that we obtain information about. These are most generally known as *experimental units*.

A *Variable* is a characteristic that is measured and recorded for each case.

Note: Many datasets use the first column as an identifier. This would not be considered a variable.

Class Example 1.1.1 *Distinguishing between cases and variables*

Let's get to know your classmastes. Find two people near you and ask them the following questions:

a) What is your first name?
b) What is your major?
c) How tall are you?
d) How many credit hours are you taking this semester?
e) How often do you drink caffeine (never, once a week, few times a week, everyday)?

Name	Major	Height	Credit Hours	Caffeine

How many cases are there?

How many variables are there?

Types of Variables

Before we can know how to answer a question, we need to think about the kinds of variables we have recorded. Variables that split cases into groups, called *levels*, are called *categorical*, or *qualitative*, variables. Variables that are numerical quantities are called *quantitative* variables.

Categorical, or *Qualitative* variables sort each case into exactly one of two or more non-numerical, non-overlapping categories.

Quantitative variables take on numerical values for each case, often by measuring or counting. Numerical operations like adding and averaging make sense for quantitative variables.

Categorical variables can further be subdivided:

- *Nominal variables* have levels with no natural ordering. The variable *major* from Example 1.1 is nominal, there is no reason to rank one major over another.
- *Ordinal variables* have levels with a meaningful order. The variable *caffeine* has a natural ordering.

Quantitative variables can further be subdivided:

- *Continuous variables* can take any value within a range, including decimals. The variable *height* is continous.
- *Discrete variables* can only take counting values (typically whole numbers: 0, 1, 2, ...). The variable *credit hours* is discrete.

Note: In most of this course, we will treat categorical variables as nominal (unordered) and quantitative variables as continuous. The distinction between nominal and ordinal, and between discrete and continuous becomes more important in advanced courses when special methods are used to take advantage of the ordering.

The ability to identify whether a variable is categorical or quantitative is important because different types of variables are summarized in different ways.

Class Example 1.1.2 *Computer ownership at UWL*

UWL would like to estimate the proportion of students who own a computer. The university asks 500 students.

a) Identify the cases and variable(s) recorded.

b) What is the sample?

c) What is the population?

d) Is the variable categorical or quantitative?

e) Identify an unmeasured quantitative variable of interest.

f) Is the result of this study a statistic or parameter?

Explanatory and Response Variables

In this class, we will discuss two ways in which data are used to answer questions. The first is looking only at a single variable, and the second is to determine if there is a relationship between two variables. When two variables show some connection or pattern with one another, they are called *associated* variables (or related variables). If two variables are not associated, there is no evident relationship between them. The two variables are said to be *independent*.

Associated variables show some connection or pattern with one another.

Independent variables have no evident relationship between them.

In studies that look for an association, many times we want to know if one variable explains the behavior of another variable. In this kind of study, the *explanatory* variable would explain the behavior in the *response* variable.

An *explanatory* variable is one that helps us understand or predict the value of another variable.

The *response* variable is the variable that we are trying to understand.

Class Example 1.1.3 *Explanatory and response variables*

For the following research questions, identify the explanatory variable and the response variable.

a) Does eating yogurt cause people to lose weight?

b) Does meditation help reduce stress?

c) Is hyperactivity in children affected by sugar consumption?

Class Example 1.1.4 *2023 Hollywood Movies*

Here is a small part of a dataset that includes information on all 136 movies to come out of Hollywood in 2023. "Rating" is audience rating on a 100-point scale, "Budget" is budget in millions of dollars, and "Opening" is opening weekend gross, in millions of dollars.

Title	Lead Studio	Rating	Genre	Budget	Opening
Oppenheimer	Atlas	93	Biography	$100	$82
Barbie	Warner Bros	88	Comedy	$145	$162
Spider-Man: Across the Spider-Verse	Sony	95	Action	$100	$120.5
Guardians of the Galaxy Vol. 3	Disney	82	Sci-fi	$250	$118
The Holdovers	Focus	97	Drama	$13	$2.7
Killers of the Flower Moon	Paramount	93	Drama	$200	$23

a) What are the cases in this dataset?

b) What are the variables in this dataset? For each variable, identify whether it is categorical or quantitative.

c) Name a question we might ask about this dataset that is about

- a single variable

- a relationship between two of the variables

Observational Studies and Experiments

In an *experiment* the researcher actively controls one or more of the explanatory variables.

In an *observational study* the researcher simply observes the values as they naturally exist.

Two variables are *associated* when the values of one tend to move with the values of the other.

Two variables are *causally associated* when changing one of them actually changes the other.

There are two primary types of data collection: experiments and observational studies. Understanding the difference is critical, because the type of study determines what conclusions we can draw.

If the researcher is only observing what happened, the study is an *observational study*. In statistics, the word observational does not mean that the researcher is watching the participant do every part of the study. Instead, observational refers to the fact that the researcher takes a dataset and makes observations about the data in the set. If the researcher is actively involved with assigning different values of the explanatory variable to the participants, then the study is an *experiment*. In general, observational studies can provide evidence of a naturally occurring *association* between variables, but they cannot by themselves show a *causal* connection. This is because in observational studies, researchers cannot control for all the other variables that might be influencing the outcome.

Warning: Association does not imply causation.

Class Example 1.1.5 *Experimental or Observational Studies*

Determine whether each of the following is an experiment or an observational study:

a) A pharmaceutical company randomly assigns 200 patients to receive either a new drug or a placebo and measures their cholesterol levels after 6 months.

b) A university surveys its graduates to determine whether those who participated in study-abroad programs earn higher salaries five years after graduation.

c) A teacher assigns students in her morning class to use a new learning app and students in her afternoon class to study using traditional methods, then compares exam scores.

Summary

This chapter introduced you to the world of data. Here are the key ideas:

- *Data* are observations collected about the world around us. *Statistics* is the science of collecting, analyzing, and drawing conclusions from data.
- A *data set*, or *data frame*, organizes data so that each row represents an *observation* and each column represents a *variable*.
- Variables come in two fundamental types: *quantitative* (continuous or discrete) and *categorical* (nominal or ordinal).
- The *explanatory* variable is the variable we think might influence the *response* variable.
- *Experiments* randomly assign subjects to treatments and can establish causation. *Observational* studies merely observe what happens and can identify associations, but generally cannot establish causation.
- Association does not imply causation — this is one of the most important principles in statistics.

Many of the ideas from this chapter will be revisited as we move on to doing end-to-end data analyses. In the next chapter, you will learn about how we can design studies to collect the data we need to make conclusions with the desired scope of inference.

1.2 Data Collection and Study Design

Before digging into the details of working with data, we pause to think about how data come to be. If data are to be used to draw broad conclusions, then it is important to understand who or what the data represent. One important aspect is ***sampling*** *— knowing how the observational units were selected from a larger group allows us to generalize back to the population from which the data were drawn. Additionally, by understanding the structure of the study, we can separate causal relationships from mere associations. A good question to ask before analyzing any dataset is:* ***How were these observations collected?*** *You will learn a lot about the data by understanding its source.*

Key Concepts

- Distinguish between a sample and a population
- Recognize when it is appropriate to use sample data to make inferences about the population
- Critically examine the way a sample is selected, identifying possible sources of bias
- Recognize that random sampling is a powerful way to avoid sampling bias
- Identify other potential sources of bias that may arise in studies on humans
- Distinguish between an observational study and a randomized experiment
- Recognize that only randomized experiments can lead to claims of causation
- Distinguish between a randomized comparative experiment and a matched pairs experiment
- Identify potential confounding variables in a study

Motivation

A state in the upper Midwest is currently considering a bill that would provide amnesty to underage persons who call for help from the authorities with an overly intoxicated person. One of the state senators who represents a district that contains a university is hesitant to initially support the bill, so she has commissioned a study to learn about how the residents of her district feel about the bill. The senator will select from one of the three study options.

Study 1: "Would you support a bill that prevents authorities from giving a ticket to an underage person who has asked for help with an overly intoxicated person?" Participants would be recruited by purchasing ads on the local radio stations.

Study 2: "Would you support a bill that has the potential to allow underage people to drink with no fear of retribution?" Participants would be a random sample of all taxpayers in the district gathered by door-to-door interviews.

Study 3: "Would you support a bill that helps reduce the chances of undergraduate death due to alcohol poisoning by guaranteeing that underage people who request emergency medical assistance for someone who is very intoxicated will not receive disciplinary sanctions from the university?" Participants would be a random sample of all property owners in the district contacted to fill out a survey via email.

a) Rank the three studies based on the question they ask. Briefly justify your ranking.

b) Rank the three studies based on the sampling method. Briefly justify your ranking.

c) If you could mix the questions and sampling methods, which combination do you think would be the best?

Populations and Samples

A *population* includes **all** individuals or objects of interest.

A *census* occurs when every case in the population is measured.

The *sampling frame* is a list of the names of all subjects or objects in the population.

Data are collected from a *sample*, which is a subset of the population.

Statistical inference is the process of drawing conclusions about a population using data from a sample.

The first step in conducting research is to identify the question to be investigated. A clearly stated research question helps identify what subjects or cases should be studied and what variables are important, or the *population*. It is also important to consider how data are collected so that they are reliable and help achieve the research goals.

One way we could find an answer is to look at every case by taking a *census*. However, most of the time this is not feasible. Taking a census could take months and is often very costly. Sometimes items are destroyed in the sampling process, such as measuring the tensile strength of a wire or the length of life of a light bulb or battery. Collecting a census in these cases would be foolish. So, most of the time, researchers take a *sample* from the population and use what they find from the sample to describe the population

In order to make sure that all units in the population are included, a *sampling frame* is compiled and used to select the subjects for the sample. This process of using a sample to describe the population is called *statistical inference*, and the accuracy of the inference can be greatly impacted by the quality of the sample and how well the sample reflects the characteristics of the population.

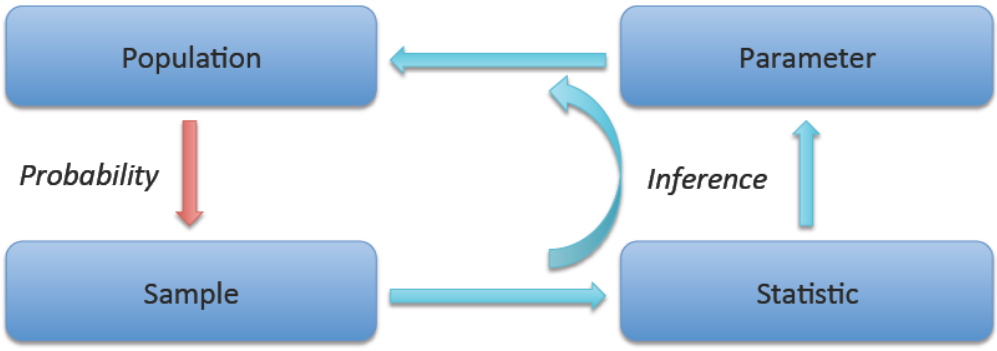


Figure 2: The cycle of statistical inference.

Class Example 1.2.1 *Population Identification*

For the following research questions, identify the target population and what represents an individual case

a) What is the average mercury content in swordfish in the Atlantic Ocean?

b) Over the last five years, what is the average time to complete a degree for UW–La Crosse undergrads?

c) Does a new drug reduce the number of deaths in patients with severe heart disease?

Parameters and Statistics

In most statistical analyses, the research question boils down to understanding a numerical summary — perhaps a quantity you already know (like the average) or one you will learn about in this course.

A numerical summary can be calculated on either the sample, called a *sample statistic*, or the entire population, called a *population parameter*. However, measuring every unit in the population is usually impossible. So we calculate the summary statistic from a sample and use it to estimate the corresponding population value.

A *parameter* is a number that describes a population and is almost always unknown.

A *statistic* is a number that describes a sample and is almost always known.

Class Example 1.2.2 *Parameter and Statistic Identification*

For the following research questions, identify the population parameter and the sample statistic.

a) A sample of 60 swordfish have an average of 0.97 ppm

b) An average of 4.37 years is reported from 25 UWL students as the time to complete their degree.

c) After 1000 patients started taking a new drug for severe heart disease, 28% were reported to have died within 5 years.

Sampling from a Population

We might try to estimate the time to graduation for UW–La Crosse undergraduates by collecting a sample of graduates. All graduates in the last five years represent the *population*, and graduates who are selected for review are collectively called the *sample*. In general, we always seek to *randomly* select a sample from a population. The most basic type of random selection is equivalent to how raffles are conducted. For example, we could write each graduate's name on a raffle ticket and draw 10 tickets. The selected names would represent a random sample of 10 graduates.

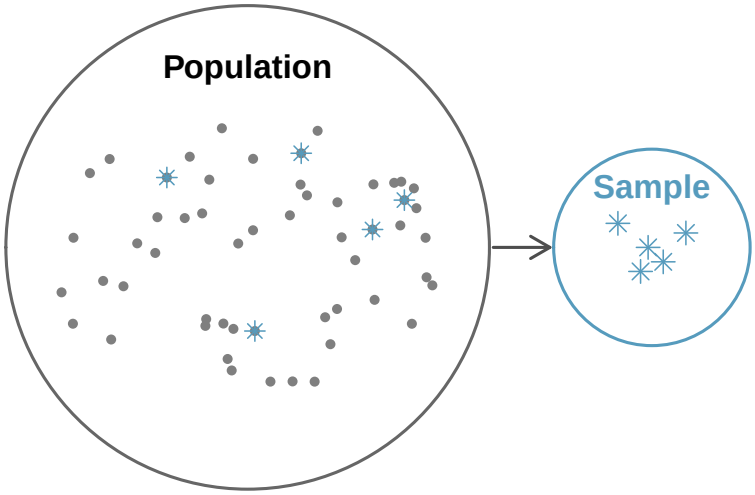


Figure 3: Sampling from a population. Individuals are randomly selected from the population (large circle) to be included in the sample (small circle).

If we ask a student who happens to be majoring in nutrition to select several graduates for the study, they might pick a disproportionate number of graduates from health-related fields. When selecting samples by hand, we run the risk of picking a *biased* sample, even if our bias is unintentional.

Bias occurs when the results from the sample differ from the population.

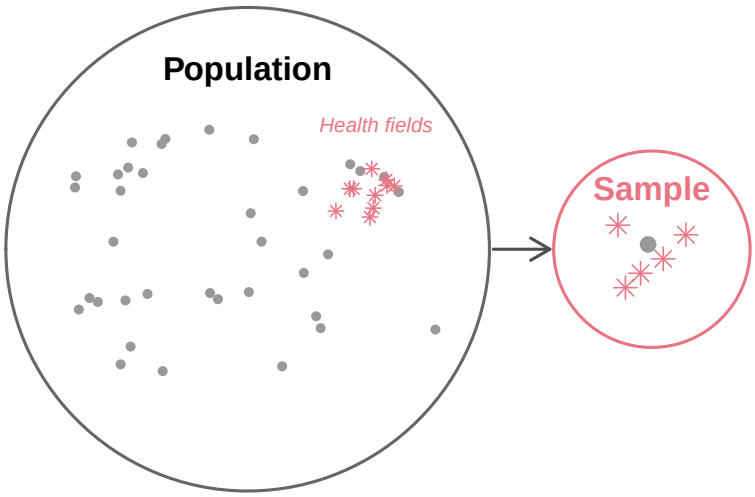


Figure 4: A biased sample. A nutrition major asked to select graduates might inadvertently pick a disproportionate number from health-related fields.

If someone is permitted to pick and choose exactly which individuals are included in the sample, it is entirely possible that the sample will overrepresent that person's interests — even if the bias is unintentional. Sampling randomly helps address this problem.

The act of taking a simple random sample helps minimize bias. However, bias can crop up in other ways. For example, asking sensitive questions may result in *response bias* because participants do not feel comfortable answering the question truthfully. If a lot of people do not respond to a survey, it will suffer from *non-response bias*. Finally, if the population from which the sample is drawn excludes groups of people, then the sample will suffer from *undercoverage*.

Response bias is the tendency for people to not answer questions honestly.

Non-response bias occurs when the response rate is very low.

Undercoverage occurs when part of the population is not considered for inclusion in the sample.

Class Example 1.2.3 *Wording bias*

A survey is to be conducted using a random sample of citizens in a town, asking them if they support raising taxes to increase funding for the public school.

a) Write the question in a way that is likely to bias the results toward more *yes* answers.

b) Write the question in a way that is likely to bias the results toward more *no* answers.

Sampling Methods

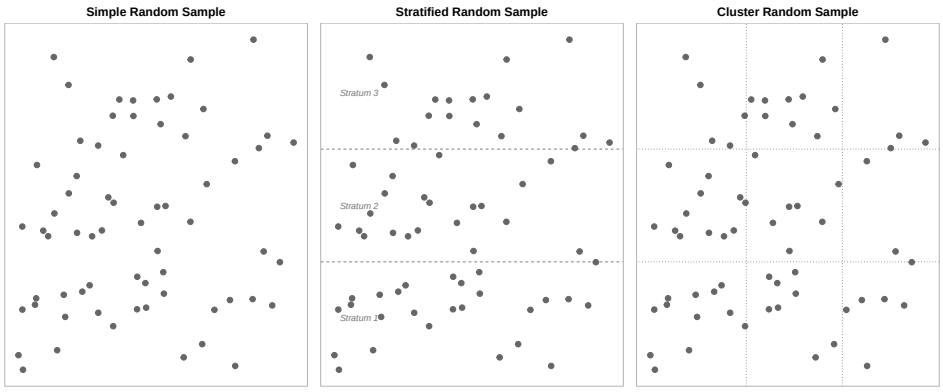
Good Ways to Obtain a Sample

There are a number of appropriate ways to randomly select your sample. The most basic type of sampling is the *simple random sample* (*SRS*). Taking a SRS avoids sampling bias and tends to produce a good representative sample of the population. An SRS is always valid, but it is not always the most practical choice: we may have no complete list of the population to draw from, reaching individuals scattered across a wide area may be too costly, or we may want to be sure that small subgroups are represented. The methods below handle those situations — and every one of them still relies on randomization. *Stratified random sampling* occurs when the population is divided into homogeneous, non-overlapping groups (called *strata*) and a simple random sample is selected from each group. *Cluster random sampling* occurs when

A *simple random sample* is obtained when every sample of a given size is equally likely to be the one selected.

Stratified random sampling occurs when the population is divided into homogeneous, non-overlapping groups (called *strata*) and a simple random sample is selected from each group.

the population is divided into many non-overlapping groups (called *clusters*), a simple random sample of clusters is obtained, and data is collected from every case in the cluster.



Cluster random sampling divides the population into many non-overlapping groups (called *clusters*), randomly selects a subset of clusters, and collects data from every case in the cluster.

Figure 5: Three different ways we can collect a random sample. Population shown as dots — circle sampled individuals during lecture.

Class Example 1.2.4 *Picking a good sampling method*

Which type of sampling method is most appropriate in the following scenarios?

a) Researchers want to test a new math textbook for third graders. They only have the budget to try the textbook in 10 elementary schools.

b) Processor chips are continuously manufactured at a production plant. To ensure that they meet size specifications so that they will fit inside an iPhone, manufacturers want to test 100 chips throughout the day.

c) A player's agent wants to know the average salary for each position of major league baseball.

Table 1: Sampling method summary

Method	How it works	Best used when...
SRS	Every case has an equal chance of selection	Population is relatively homogeneous and accessible
Stratified	Divide into similar groups (strata), then SRS within each	Known subgroups differ on the outcome of interest
Cluster	Randomly select entire groups (clusters)	Clusters are internally diverse but similar to each other; travel/cost is a concern
Multistage	Randomly select clusters, then SRS within each	Same as cluster, but clusters are large

Random sampling caution: Random is not the same as haphazard. We cannot obtain a random sample by haphazardly picking a sample on our own. We must use a formal random sampling method such as technology or drawing names out of a hat.

Class Activity: Sampling from Gettysburg Address Worksheet

Bad Ways to Take a Sample

Without a random sample, we may encounter *sampling bias* which occurs when the sample does not accurately represent the population of interest. One common sampling technique that will nearly always be biased is *voluntary response sampling*. With this kind of sampling, a researcher asks people to volunteer to be a part of a study. Another common type of sampling that produces biased results is the *convenience sample*. In a convenience sample, a researcher simply uses the most convenient group available as the sample.

Sampling bias occurs when the sample does not accurately represent the population of interest.

Voluntary response sampling occurs when a researcher asks people to volunteer to be participants instead of randomly selecting possible participants.

Convenience sampling happens when a researcher just uses the most convenient group as a sample.

Class Example 1.2.5 *Driving with a pet on your lap*

Over 30,000 people participated in an online poll on cnn.com conducted in April 2012 asking "Have you ever driven with a pet on your lap?" We see that 34% of the particpants answered yes and 66% answered no.

a) What is the sample?

b) Is the variable in this study quantitative or categorical?

c) Can we conclude that 34% of all drivers have driven with a pet on their lap?

d) Explain why it is not appropriate to generalize these results to all drivers, or even to all drivers who visit cnn.com. What is the problem with this method of data collection?

Experiments

In an *experiment* the researcher actively controls one or more of the explanatory variables.

In a *randomized experiment* the value of the explanatory variable for each unit is determined randomly, before the response variable is measured.

Well-designed *experiments* incorporate four key principles:

1. **Controlling.** Researchers assign treatments to cases and do their best to *control* any other differences between the groups. For example, when patients take a drug in pill form, some patients might take the pill with a sip of water while others drink a full glass. To control for the effect of water consumption, a doctor might instruct every patient to drink a 12-ounce glass of water with the pill.
2. **Randomization.** Researchers randomly assign subjects to treatment groups to account for variables that cannot be controlled. For example, some patients may be more susceptible to a disease than others due to their dietary habits. Randomly assigning patients to treatment or control groups helps even out such differences across groups.
3. **Replication.** The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response. In a single study, we **replicate** by collecting a sufficiently large sample. At a minimum, we want multiple subjects per treatment group. Replication also refers to repeating an entire study to verify earlier findings. (The *replication crisis* refers to the ongoing problem in several scientific disciplines where past findings have failed to be replicated in new studies.)
4. **Blocking.** Researchers sometimes know or suspect that variables other than the treatment influence the response. They may first group individuals based on this variable into *blocks* and then randomize cases within each block to the treatment groups. For instance, in studying the effect of a drug on heart attacks, researchers might first split patients into low-risk and high-risk blocks, then randomly assign half the patients from each block to the treatment and half to the control group.

To demonstrate the differences between some of the more commonly used experimental designs, consider the following study. A running shoe company wants to know if the type of shoe impacts the speed of runners. One aspect of the shoe that they are considering is the heel-toe drop, or the difference in height between the heel of the shoe to the toe of the shoe. Minimalist shoes reduce the heel-toe drop to 0 – 4 mm, whereas a more traditional shoe has an 8 – 12 mm heel-toe drop. This company wants to know if a lower heel-toe drop will cause runners to run more quickly, so they recruit 300 marathon runners of varying ability. For the study, the runner will run 5k, and their time will be used to establish their pace.

Randomized comparative design

In a *randomized comparative design*, we take all the participants in the study and randomize them to a different treatment group.

In a *randomized comparative* design, we randomly assign cases to different treatment groups and then compare results on the response variable(s).

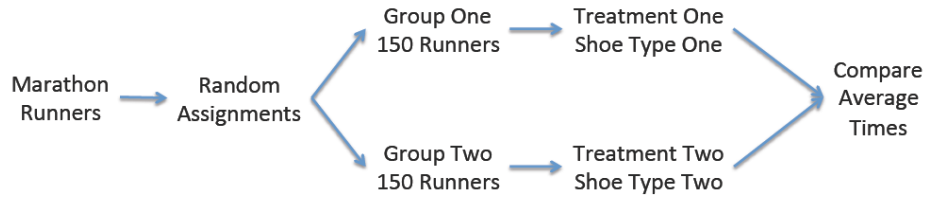


Figure 6: Randomized comparitive design.

Block design

In a *block design*, we first group participants by some logical grouping that we think might have an impact on the response in addition to the variable we are studying.

In a *block* design, cases are first split into groups (called *blocks*), and within each group the value of the explanatory variable for each unit is determined randomly.

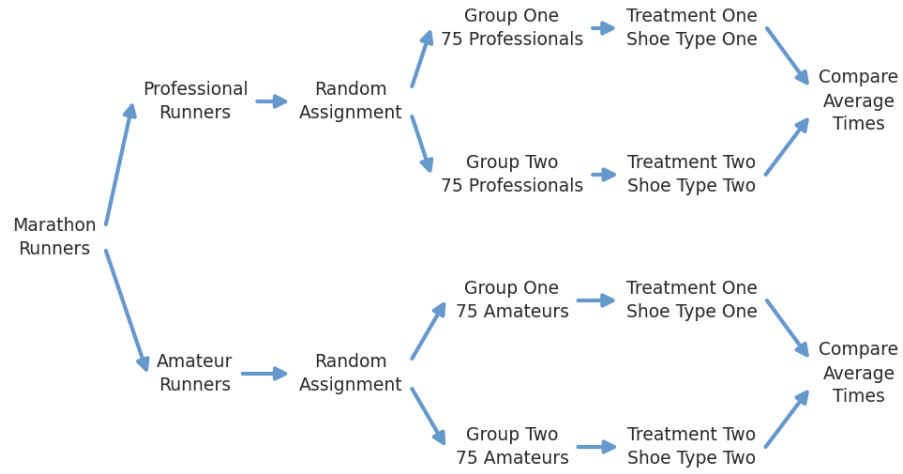


Figure 7: Block design.

Class Example 1.2.6 *Picking the blocking factor*

Why might it make sense to group the runners into professionals vs amateurs?

Matched pairs design

In a *matched pairs design*, each participant is assigned to each of two treatment settings, responses are measured for each treatment, and the differences in responses are computed for each participant. Because two measurements are obtained for each individual, the measurements are *dependent* upon each other. The two treatments often come in the form of "Before vs. After" or "Method A vs. Method B." In the "Method A vs. Method B" type of pairing, it is important to randomize the order of the treatments to avoid a confounding *order effect*.

In a *matched pairs* design, cases get both treatments (or cases are paired) and we examine differences in the response between the two treatments.

Class Example 1.2.7 *Describing the matched pairs running shoe experiment*

Thinking about the matched pairs design for the experiment about running shoes

a) What are the variables in the study? Are they categorical or quantitative?

b) What are the explanatory and response variables?

c) Explain how the treatments would be assigned.

d) Why might a matched pairs design be beneficial in this experiment?

Reducing Bias in Human Experiments

For an experiment, it is important to be able to determine if the treatment actually is making a difference, or if we are seeing results that would have happened regardless of the treatment. The purpose of a *control group* is to establish what we would expect to happen regardless of the treatment. Often times, in order to make sure that people adhere to the study guidelines, we use a *placebo* that is designed to look like the treatment.

The placebo makes it much easier to conduct *single-blinded* experiments or *double-blinded* experiments. Not all randomized experiments make use of control groups, placebos, or blinding; however, in studies involving humans, these techniques are often used to help make effective comparisons between groups.

The *control group* does not receive a treatment.

A *placebo* is a fake treatment that is designed to look like the treatment but has no therapeutic value.

In a *single-blind* study, participants do not know if they are in the treatment or control group.

In a *double-blind* study, neither the participants nor the people interacting with the participants knows if they are in the treatment or control group.

Class Example 1.2.8 *Depression and Prozac*

Researchers at a prominent Midwestern university would like to determine if Prozac is effective in reducing depression as determined by patient responses to a questionnaire. The researchers recruited 50 individuals suffering from depression. Describe a completely randomized double-blind experiment using a placebo to test this association is a causal association.

Confounding Variables

A *confounding variable* is a variable that is associated with both the explanatory variable and the response variable. Because of this dual association, a confounding variable makes it impossible to determine whether the explanatory variable truly caused the observed response.

A *confounding variable* is a third variable that is associated with both the explanatory and response variable, but not necessarily of interest in the study.

Consider a classic example: ice cream sales and drowning rates both increase during summer months. If we looked only at the association between ice cream sales and drownings, we might (absurdly) conclude that ice cream causes drowning. The confounding variable is temperature (or season) — warm weather drives both increased ice cream consumption and increased swimming, which leads to more drowning incidents.

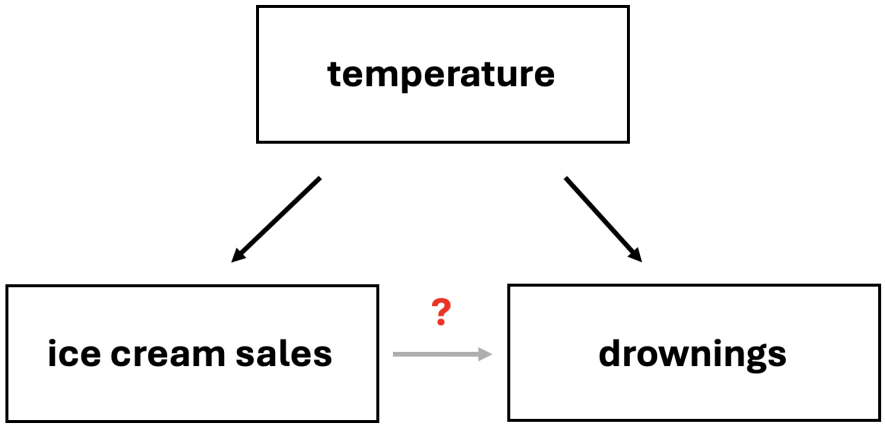


Figure 8: Confounding diagram.

Summary

In this chapter, we explored how data are collected and why the method of collection matters for the conclusions we can draw.

Key concepts:

- The *population* is the entire group of interest; a *sample* is a subset from which we collect data. A *census* collects data on every member of the population but is usually impractical.
- Good sampling methods include *simple random sampling* (*SRS*), *stratified sampling*, and *cluster sampling*. Each has advantages depending on the population's structure and practical constraints.
- *Bias* can enter through non-random sampling (selection bias), non-response, or convenience sampling.
- In an *experiment*, researchers actively assign treatments to subjects. In an *observational study*, researchers simply observe without intervening.
- *Confounding variables* are associated with both the explanatory and response variables and can create misleading associations.
- *Blinding* (single and double) and placebos help reduce bias in human experiments.

1.3 Exploring Categorical Data

*This chapter focuses on **categorical data** — variables whose values are categories or labels rather than numbers. We begin with visualizations for a single categorical variable (bar charts, pie charts, waffle charts), then move to tools for exploring the relationship between two categorical variables (contingency tables and stacked bar charts). We close with principles for creating effective, honest data visualizations.*

Key Concepts

- Display information from a categorical variable in a table or a graph
- Use information about a categorical variable to find and report a proportion, using correct notation
- Display information about a relationship between two categorical variables in a contingency table
- Use a contingency table to find proportions

Technology for Chapter

StatLens Use the One Categorical Variable explorer for bar charts and pie charts. Use the Two Categorical Variables explorer for contingency tables and stacked bar charts.

jamovi Use Analyses>Exploration>Descriptives for bar charts for one categorical variable. Use Analyses>Frequencies>Independent Samples for two categorical variables.

One Categorical Variable

When we want to summarize one categorical variable, we use a *proportion* which tells us what fraction of the sample is in each group. In this course, we will use p to represent the population proportion, and $\hat{p}$ (p-hat) to represent the sample proportion. The sample proportion is given by

The *proportion* (or *relative frequency*) in a category is found by dividing the number of cases in that category by the total number of cases.

$$\hat{p} = \frac{x}{n}$$

where x is the number of cases in a given category and n is the total number of cases.

Numerical displays

A single categorical variable is often summarized using a *frequency table* to display the number of observations that fall into each category. An alternative to the frequency table is the *relative frequency table* which displays the proportion of the sample that is in each category.

A *frequency table* is a table that displays the count for each category.

A *relative frequency table* is a table that displays the proportion of the sample in each category.

Class Example 1.3.1 *Hair Color*

The table below gives the frequency of hair color for a random sample of 54 UWL students. Find the proportions of students with each hair color and use them fill in the relative frequency column.

	Frequency	**Relative Frequency**
Brown	30	
Blonde	14	
Red	2	
Black	5	
Colorful	3	
Total	54	

Frequency table for hair color.

Graphical summaries

The standard graphical summaries for a categorical variable are the *bar chart* and the *pie chart*. A bar chart is the most common way to display the distribution of a single categorical variable. Each category gets its own bar, and the height (or length) of the bar represents either the count or the proportion of observations in that category. A pie chart divides a circle into slices, where each slice represents a category and the size of the slice corresponds to the proportion of observations in that category.

A *bar chart* includes one bar for each category, and the height of the bar is the frequency the category appears.

In a *pie chart*, the proportions correspond to the areas of sectors of a circle.

Class Example 1.3.2 *Hair Color revisited*

Sketch a bar chart and a pie chart for the hair color of random sample of 54 UWL students.

Pie charts can give a quick high-level overview, especially when a variable has only a few levels or when each level represents a simple fraction (one-half, one-quarter, etc.). However, bar charts are almost always easier to read because our eyes are better at comparing lengths than angles or areas.

Two Categorical Varaibles

A *contingency table* summarizes the data for two categorical variables. The rows represent the levels of one variables, the columns represent the levels of the other.

Sometimes our goal is to examine whether there is a relationship between two categorical variables. To investigate these relationships, we use a *contingency table*, or *two-way table*. In a two-way table, the rows represent the groups of the first category, and the columns represent the groups of the second category. Each cell of the table contains the number of cases that are in both the row and the column category.

Class Example 1.3.3 *Relationship status and class level*

169 college students were asked about their relationship status and class year (lower=freshman/sophomore, upper=junior/senior). The results are given in the table.

	Upper	**Lower**	**Total**
In a relationship	32	10	42
It's complicated	12	7	19
Single	63	45	108
Total	**107**	**62**	**169**

Contingency table for relationship status and class year.

a) What proportion of students in this sample are in a relationship?

b) What proportion of lower class students in this sample are in a relationship?

c) What proportion of upper class students in this sample are in a relationship?

d) Using $\hat{p}_U$ to represent the proportion of upper class students in a relationship and $\hat{p}_L$ to represent the proportion of lower class students in a relationship, find the difference in proportions $\hat{p}_U - \hat{p}_L$

e) What proportion of the people who are in a relationship in this sample are upper class students?

f) What proportion of the people in this sample are upper class students and in a relationship?

g) Answers c), e) and f) are all ways of representing upper class students in a relationship. Why are they different numbers?

Contingency tables can be displayed graphically using an extension of the bar chart.
There are two common variants: a stacked bar chart and a side-by-side bar chart.

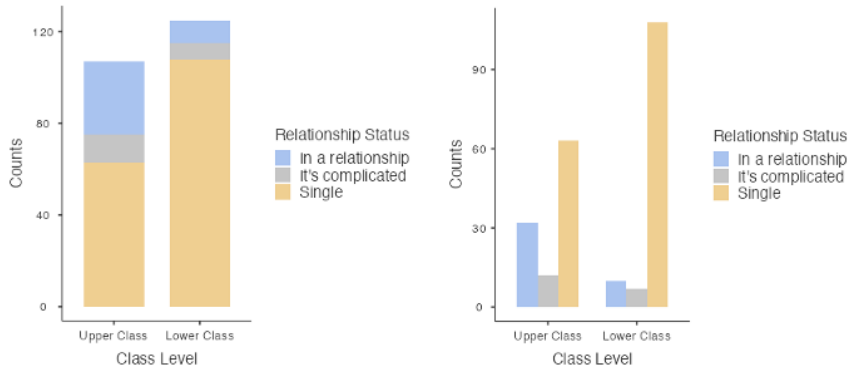


Figure 9: Bar plots for relationship status and class level.

Class Example 1.3.4 *Handedness and occupation*
In a study of handedness in occupations, 10 of 118 psychiatrists were left-handed, 26 of 148 architects were left-handed, 5 of 132 orthopedic surgeons were left-handed, and 16 of 105 lawyers were left-handed.

a) Make a two-way table of this relationship.

b) What proportion of the people in the sample are left-handed?

c) What proportion of left-handed people are architects?

d) Is your answer in part c) the same or different than the proportion of architects who are left-handed?

Effective Data Visualization

Graphs can powerfully communicate ideas directly and quickly. However, there are times when a visualization conveys a message that is inaccurate or misleading. The following are guiding principles for creating clear, honest, and effective visualizations.

- *Keep it simple*: Colors should be used purposefully and can be used to draw attention
  - Colors used only for decoration can be distracting and can confuse readers
- *Tell a story*: Include annotations (text, lines or labels that provide context beyond the raw data)
- *Order matters*: The arrangement of categories in a bar chart can help or hinder understanding
  - Alphabetical order is rarely the most informative choice
- *Make labels readable*
- *Select meaningful colors*: Default or rainbow color schemes are not always the best choice

Summary

This chapter introduced tools for exploring categorical data.

-For a single categorical variable, *bar charts* are the primary visualization, with *pie charts* reserved for simple cases.

-For two categorical variables, *contingency tables* organize counts, and stacked and side-by-side bar charts provide visual comparisons.

-Effective visualizations follow principles of simplicity, purposeful use of color, meaningful ordering, and accessibility.

1.4 *Exploring Quantitative Data*

*In the previous chapter we explored categorical data. Now we turn to **quantitative data** — variables that take numerical values where arithmetic makes sense (heights, incomes, interest rates). This chapter teaches you to both see and* measure *distributions. We introduce dot plots and histograms alongside the mean and median, box plots alongside the IQR, and the standard deviation alongside histogram shapes. By the end, you will be able to describe any distribution in terms of its shape, center, spread, and unusual features.*

Key Concepts

- Understand the difference between displays of categorical and quantitative data
- Use a dotplot or histogram to describe the shape of a distribution
- Calculate the mean and median of a data set, using appropriate notation
- Locate the mean and median approximately on a dotplot or histogram
- Explain how outliers and skewness affect the values for the mean and median
- Recognize the uses and meaning of the standard deviation
- Interpret a five number summary
- Use the range, the interquartile range, and the standard deviation to measure spread
- Describe the advantages and disadvantages of the different measures of spread
- Identify outliers in a dataset based on the IQR method
- Use a boxplot to describe data for a single quantitative variable
- Use the empirical rule to estimate the middle 68%, 95%, or 99.7% for bell-shaped distributions
- Use technology to compute summary statistics for a quantitative variable

Technology for Chapter

StatLens Use the One Quantitative Variable to visualize distributions as histograms, dot plots, or box plots, and compute means, medians, standard deviations, and quartiles for any dataset.

jamovi Use Analyses>Exploration>Descriptives to visualize distributions as histograms or box plots, and compute means, medians, standard deviations, and quartiles for any dataset.

Dot Plots and Histograms

Dot Plots

A *dot plot* is the simplest display for a single quantitative variable. Each observation is represented by a dot placed above its value on a number line. When observations share the same value (or are rounded to the same value), the dots stack vertically.

Consider the interest rates on 50 loans from a lending platform. In a dot plot, each of the 50 loans is a single dot. The red triangle marks the mean.

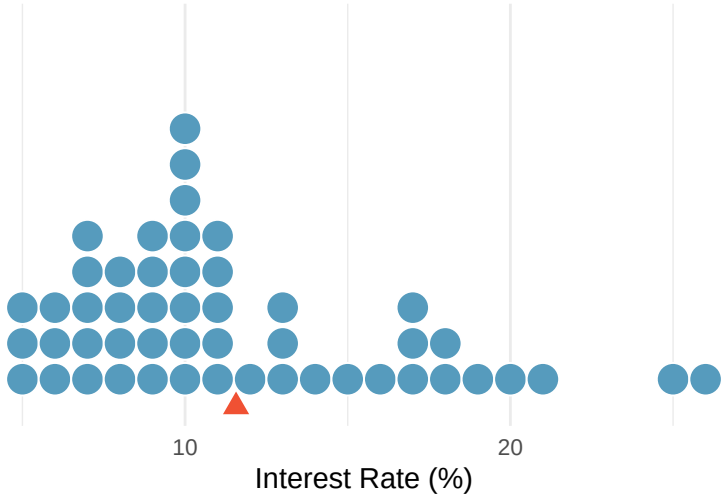


Figure 10: A dot plot of interest rates for 50 loans. Most values cluster between 5% and 15%, with a few high rates pulling the distribution to the right. The red triangle marks the mean (11.57%).

Dot plots are most useful for small datasets (up to about 50–100 observations) where you want to see every individual value. For larger datasets, histograms provide a better summary.

Histograms

A *histogram* groups quantitative data into bins (intervals) and displays the count or proportion of observations in each bin as a bar. Unlike bar charts for categorical data, histogram bars touch each other because the bins represent a continuous range of values.

To build a histogram, we:

1. Choose a bin width
2. Count how many observations fall in each bin
3. Draw bars whose heights equal the counts

Histograms are especially useful for identifying the *shape* of a distribution. One key characteristic is symmetry. When we say a distribution is *symmetric* it means that the left-side of a distribution is very similar to the right-side of the distribution. We say a distribution is *skewed left* when the tail of the distribution is longer on the left. Similarly, we say that a distribution is *skewed right* when the tail of the distribution is longer on the right.

A distribution is considered *symmetric* if we can fold the plot over a vertical center line and both sides match closely.

A distribution is considered *skewed left* if the longer tail of the distribution is to the left.

A distribution is considered *skewed right* if the longer tail of the distribution is to the right.

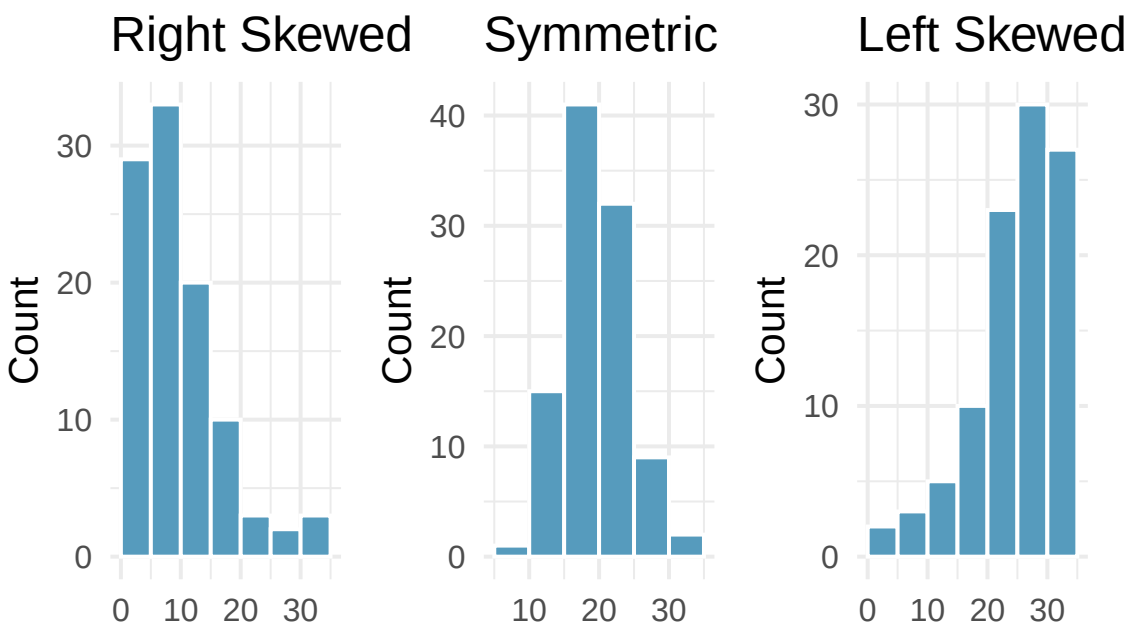


Figure 11: Three histograms depiciting the different shapes: skewed right, symmetric, and skewed left

Class Example 1.4.1 *Shape of the interest rates*

Below is the histogram for the interest rate dot plot. Describe the shape.

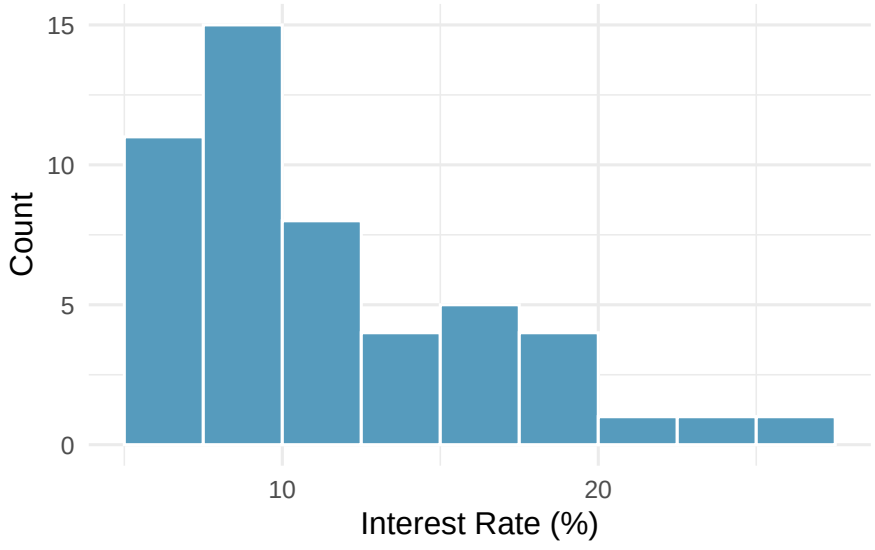


Figure 12: A histogram of interest rates with bin width 2.5 percentage points.

Measures of Center: Mean and Median

The *mean* of a data set is the sum of all its values divided by the number of observations.

The *median* of a dataset is the middle entry of an ordered list of data. If there are an even number of entries, then we take the average of the middle two values.

One type of numerical summaries you will see for a dataset is called measures of *center*. Measures of center provide us with a single number that best describes a quantitative variable's distribution. Essentially, they give us a guess of what a typical value would look like. There are two measures of center we will use in this class: the *mean* and *median*.

The mean

The *mean* of a sample tells us the average. For a sample of observations $x_1, x_2, \ldots, x_n$, the mean is given by

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

where n is the number of observations.

Notation:
μ: population mean
$\overline{x}$: sample mean

We will use μ (pronounced "mew") to represent the population mean and $\overline{x}$ (read "x-bar") to represent the sample mean. We often cannot measure μ directly because we rarely have access to every member of the population. Instead, we estimate μ using the sample mean $\bar{x}$.

The median

The *median*, denoted m, divides the sample into two groups of equal size. We find the median by following two steps:

1. Sort the sample values.
2. Find the middle value:
   - If n is odd, the median is the middle value in the sorted list.
   - If n is even, the median is the average (midpoint) of the middle two values in the ordered list.

The dot plot below shows both the mean and the median marked on the interest rate data. Notice that the median (9.93%) is lower than the mean (11.57%) — the few loans with very high interest rates pull the mean to the right, but the median stays anchored in the middle of the data.

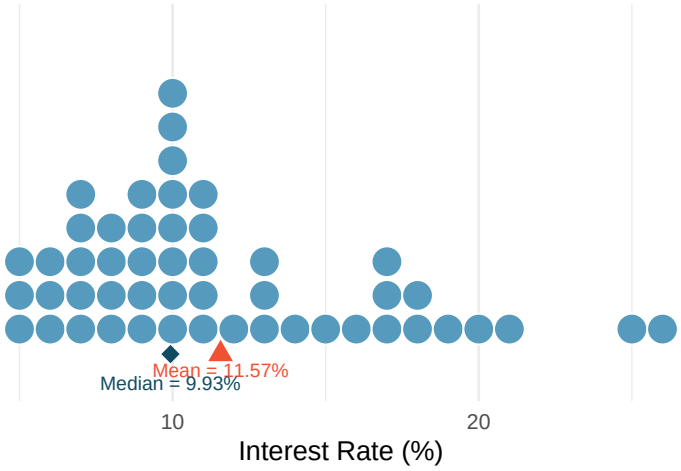


Figure 13: A dot plot of interest rates with the mean (red triangle, 11.57%) and median (blue diamond, 9.93%) marked. The mean is pulled to the right by the high-interest loans.

Class Example 1.4.2 *Ants on a sandwich*

The number of ants climbing on a piece of a peanut butter sandwich left on the ground near an anthill for a few minutes was measured 7 different times, and the results are: 43, 59, 22, 25, 36, 47, 19

a) Calculate the mean number of ants.

b) Calculate the median number of ants.

c) Suppose one of the sandwich bits was extremely appealing to the ants, and instead of 59, the number of ants for that bit was actually 159, giving the 7 values as: 43, 159, 22, 25, 36, 47, 19. Compute the mean and median for this new dataset.

d) Which statistic (mean or median) is impacted most by the outlier?

How skewness affects the mean and median

In a symmetric distribution, the mean and the median are approximately equal (mean $\approx$ median). In a right-skew distribution, the mean is pulled towards the right tail and is larger than the median (mean $>$ median). In a left-skewed distribution, the mean is pulled towards the left tail and is smaller than the median (mean $<$ median).

When to use the mean vs. the median

As we saw in the last example, the median is not impacted by extreme data values (outliers) like the mean is. We say that the median is resistant. Although the mean is a more commonly used measure of center, the median may be preferred in cases with extreme values or a high degree of skew.

If the concern is the total or aggregate, the mean is the best. If the concern is a typical observation, the median is preferred, especially when the distribution is skewed or contains extreme values.

Measures of Spread

A measure of *spread* tells us how much variability exists in the data.

Knowing the center of a distribution is important, but it tells only part of the story. Two distributions can have the same center but look very different if one is tightly clustered and the other is widely dispersed. We need measures of *spread* (also called variability or dispersion).

The range

The *range* is the difference between the maximum and minimum of the data.

The simplest measure of spread is the *range*: the difference between the maximum and minimum values.

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

While easy to compute, the range has a major weakness: it depends entirely on the two most extreme observations. A single extreme value can dramatically increase the range, giving a misleading impression of the overall variability. For this reason, we usually prefer the IQR or standard deviation.

The interquartile range (IQR)

The *IQR* (Interquartile range) is the difference between the 1st and 3rd quartiles of the data.

The *interquartile range* measures the spread of the middle 50% of the data. To find the IQR, we need to first estimate Q_1 and Q_3, the first and third quartiles. The first quartile, Q_1, is the value which 25% of the data falls below; and the third quartile, Q_3, is the value which 75% of the data falls below. To find Q_1, we need to take the median of all the values below the median. To find Q_3, we need to take the median of all the values above the median. The IQR is given by

$$\text{IQR} = Q_3 - Q_1.$$

The five-number summary

A way to combine the information from the range, the IQR, and a measure of center is through the *five-number summary*. The five-number summary consists of the minimum, Q_1, the median, Q_3, and the maximum of the data set.

Class Example 1.4.3 *Ants revisited*

In the previous example, we found that for the 7 different times, the results were 19, 22, 25, 36, 43, 47, 59 (sorted in ascending order)

a) Find the five number summary for this data.

b) Find the IQR for this data

c) Find the range for this data

d) Find the range and IQR if the value 59 was actually 159?

e) Which statistic (range or IQR) is impacted most by the extreme value?

Variance and standard deviation

The **standard deviation** is the most commonly used measure of spread. The standard deviation gives an estimate for the average distance each observation is from the mean.

The *standard deviation* of a variable roughly measures the typical distance of a data value from the mean.
A *deviation* is the difference between between an observation and the mean
The *variance* is the average of the squared deviations.

We start with the concept of a **deviation**: the distance of an observation from the mean.

$$\text{deviation of } x_i = x_i - \bar{x}$$

Some deviations are positive (observations above the mean) and some are negative (observations below the mean). If we simply averaged the deviations, the positives and negatives would cancel out, giving us zero. To avoid this, we *square* each deviation before averaging. The average of the squared deviations, using $n-1$ in the denominator is the *variance* and is given by

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

The *standard deviation* is the square root of the variance:

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

Notation:
σ: population standard deviation
s: sample standard deviation

In this course, we will use σ^2 (pronounced "sigma squared") to represent the population variance and s^2 to represent the sample variance. We will use σ (pronounced "sigma") to represent the population standard deviation and s to represent the sample standard deviation. When the standard deviation is large, it means there is a lot of variability amongst the observations. When the standard deviation is small, it means there is a lot of consistency amongst the observations.

Looking back at the interest rate data with shaded bands marking one and two standard deviations from the mean. The standard deviation of 5.05 percentage points gives us a sense of how spread out the data are around the mean of 11.57%.

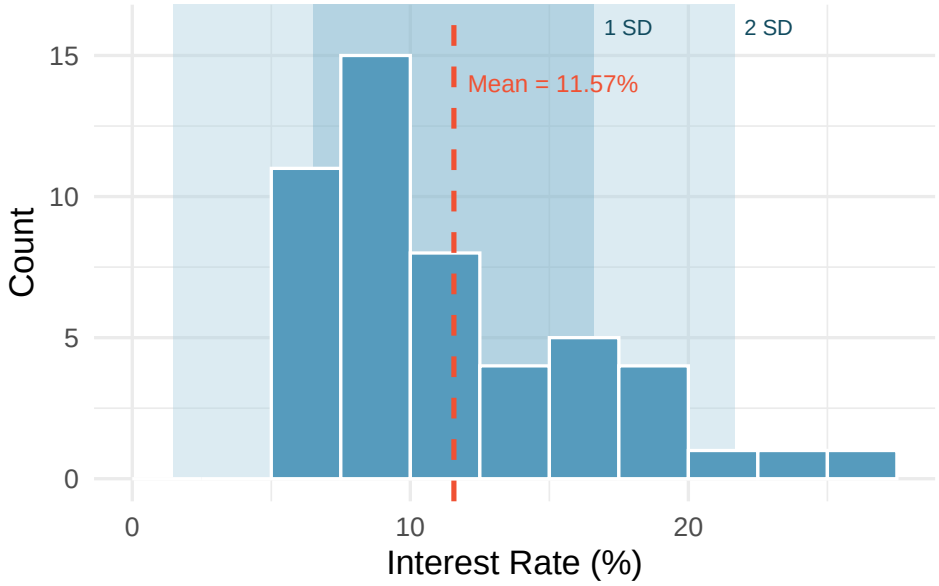


Figure 14: A histogram of interest rates with shaded regions showing one standard deviation (dark) and two standard deviations (light) from the mean.

The standard deviation tells us that interest rates typically deviate about 5 percentage points from the mean (11.57%).

Class Example 1.4.4 *Ants revisited*

a) Find the standard deviation of the ants on a sandwich data.

b) Find the standard deviation of the data if the value 59 was actually 159.

Outliers and Robust Statistics

Identifying outliers

An *outlier* is a data point that is unusual given the rest of the data.

An *outlier* is an observation that appears extreme relative to the rest of the data. A common rule for identifying outliers is the **1.5 IQR rule**: A data value, x_0 is considered to be an outlier if

$$x_0 < Q_1 - 1.5(IQR) \quad \text{or} \quad x_0 > Q_3 + 1.5(IQR).$$

Examining data for outliers serves several purposes:

- Identifying strong skew in the distribution
- Identifying possible data collection or data entry errors
- Providing insight into interesting or unusual properties of the data

Robust statistics

We say that a statistic is *robust*, or *resistant*, if its is relatively unaffected by extreme values.

We call a statistic **robust** (or **resistant**) if extreme observations have little effect on its value.

Class Example 1.4.5 *Resistant statistics*

Look back at your answers to the questions about ants on a sandwich.

a) Which measures of center (mean, median) are resistant?

b) Which measures of spread (IQR, range, standard deviation) are resistant?

Robust vs. non-robust statistics.

Robust (resistant to outliers)	**Not robust (sensitive to outliers)**
Median	Mean
IQR	Standard deviation
	Range

When a distribution is skewed or contains outliers, robust statistics (median and IQR) often give a more representative summary of the "typical" observation. When a distribution is roughly symmetric with no extreme outliers, the mean and standard deviation work well and carry additional mathematical advantages.

Box plots

The *boxplot* is a graphical display of the *five-number summary* which consists of the minimum,Q_1, median,Q_3, and the maximum of a dataset.

So far in this class, we have used histograms and dot plots to represent data graphically. There is another very commonly used graphical summary called the *boxplot*. The boxplot gives us a graphical display of the *five-number summary* which consists of the minimum,Q_1, median,Q_3, and the maximum of a dataset.

Constructing a box plot

The components of a box plot are:

1. *The box*: Spans from Q_1 (the 25th percentile) to Q_3 (the 75th percentile). This box contains the middle 50% of the data. The length of the box is the *interquartile range* (IQR = $Q_3 - Q_1$).
2. *The median line*: A line inside the box marks the median (the 50th percentile), which divides the data in half.
3. *The whiskers*: Lines extending from the box toward the smallest and largest observations that are *not* outliers. Specifically, the whiskers extend to the most extreme data points within $1.5 \times$ IQR of the box.
4. *Outlier points*: Any observations beyond $1.5 \times$ IQR from the edges of the box are plotted individually as dots. These are potential outliers.

The following gives a general sketch of a boxplot:

A box plot of interest rates shows: the left edge of the box at about 8%, the median line at about 10%, and the right edge of the box at about 14%. Two dots appear beyond the right whisker at about 25% and 26%.

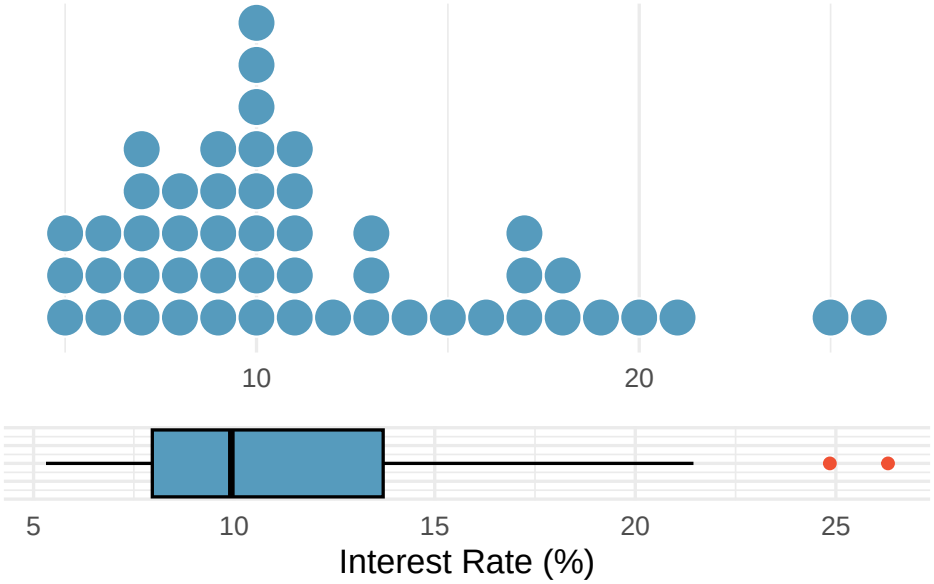


Figure 15: A dot plot and box plot of the same interest rate data, shown one above the other for comparison. The box plot summarizes the distribution with five numbers plus outliers.

Class Example 1.4.6 *Population of U.S. states*

The following data represent the populations of 12 of the 50 U.S. states, in millions of people, in ascending order.

0.506	0.830	1.315	1.813	2.333	2.953
3.524	4.198	5.504	6.207	8.685	22.472

a) Identify the five-number summary for the state populations.

b) Draw a boxplot for the data. If there are any outliers, provide a justification as to why they are outliers and identify them on your boxplot.

Class Example 1.4.7 *Mammal Gestation*

The figure below gives a boxplot of the gestation (in days) for 54 mammal species.

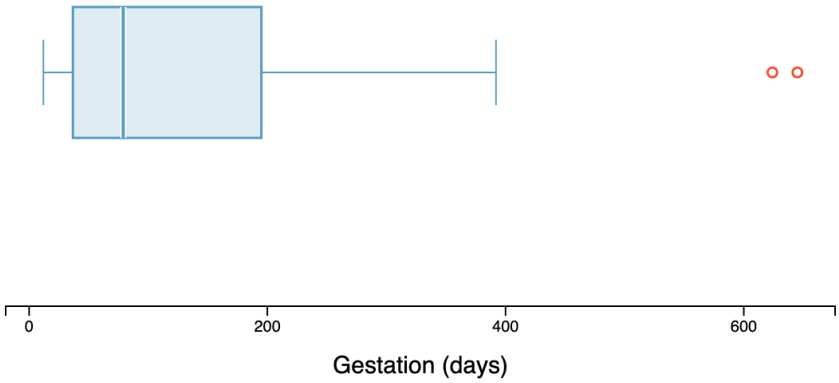


Figure 16: Box plot of gestation (in days).

a) Are there any outliers? If so, how many and estimate the values of the outliers.

b) Estimate the median, IQR and range of the data.

c) How would you describe the shape of this distribution?

d) Do you expect the mean to be less than, approximately equal to, or greater than the median? Explain.

Describing a distribution

A complete description of a distribution should address three things: *shape*, *center*, and *spread*.

A good description of a distribution should always:

1. Name the *shape*: symmetric or skewed (and direction)
2. Give the *center*: report the mean and/or median with units
3. Give the *spread*: report the standard deviation and/or IQR with units
4. Note any *unusual features*: outliers, gaps, clusters
5. Use *context*: refer to the actual variable and its units, not just abstract numbers

A way to describe a distribution is as follows: "The distribution of commute times is right skewed, with a median of 22 minutes and an IQR of 15 minutes. A few commuters have commute times exceeding 90 minutes."

The Empirical Rule (68-95-99.7 rule)

For distributions that are approximately *bell-shaped*, the standard deviation provides a useful ruler. Statisticians call this particular bell shape the *normal distribution* — it comes up so often in statistics that it has its own name. We'll study it formally later, but for now, just know that "bell-shaped" and "normal" mean the same thing.

The empirical rule.

For a roughly bell-shaped distribution:

- About **68%** of the data fall within 1 standard deviation of the mean:

$$(\bar{x} - s,\ \bar{x} + s)$$

- About **95%** of the data fall within 2 standard deviations of the mean:

$$(\bar{x} - 2s,\ \bar{x} + 2s)$$

- About **99.7%** of the data fall within 3 standard deviations of the mean:

$$(\bar{x} - 3s,\ \bar{x} + 3s)$$

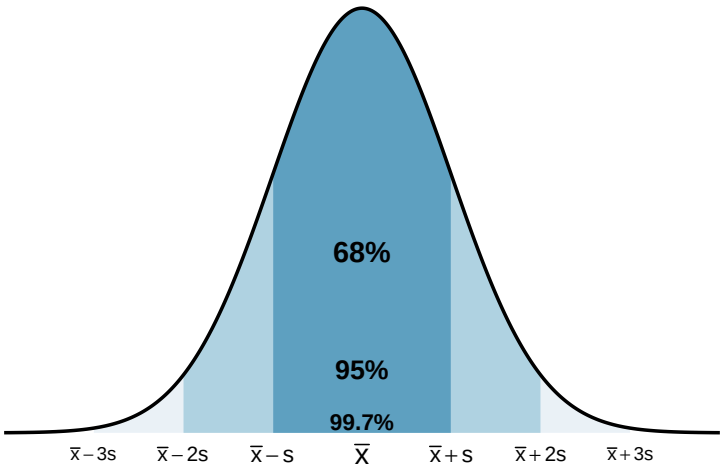


Figure 17: A bell-shaped distribution with shaded regions showing 68%, 95%, and 99.7% of the data within 1, 2, and 3 standard deviations of the mean, respectively.

Class Example 1.4.8 *SAT scores*

The SAT scores (standardized test scores) are approximately bell-shaped with a mean of 1150 and a standard deviation of 150. Use the empirical rule to estimate the range of scores for the middle 95% of students.

Summary

This chapter introduced the tools for exploring *quantitative* data — both visual and numerical.

- *Dot plots* show individual values for small datasets, *histograms* reveal the shape and density of a distribution, and *box plots* give a compact *five-number summary*.
- Distribution shape is described using symmetry (*left skewed*, *symmetric*, *right skewed*).
- The *mean* and *median* measure center, with the mean being the balancing point and the median being the middle value.
- The *standard deviation* and *IQR* measure spread, with the standard deviation roughly representing the typical deviation from the mean and the IQR representing the range of the middle 50%.
- The *median* and *IQR* are *robust* to outliers, while the mean and standard deviation are sensitive to extreme values.
- The *empirical rule* connects the standard deviation to the distribution shape for bell-shaped data: about 68%, 95%, and 99.7% of observations fall within 1, 2, and 3 standard deviations of the mean.

1.5 Relationships Between Variables

*The preceding chapters focused on exploring one variable at a time. But the most interesting statistical questions involve **relationships between variables**. Does higher education lead to higher income? Is a drug more effective for one group than another? This chapter introduces tools for exploring relationships: side-by-side box plots and faceted histograms for comparing a quantitative variable across groups, and scatterplots for examining the association between two quantitative variables. We close with the critical distinction between association and causation.*

Key Concepts

- Use a side-by-side graph to visualize a relationship between quantitative and categorical variables
- Examine a relationship between quantitative and categorical variables using comparative summary statistics

Technology for Chapter

StatLens Use the Quantitative by Group to compare distributions across groups with side-by-side box plots, histograms, and density overlays. Use the Two Quantitative Variables for scatterplots of two quantitative variables.

jamovi Use Analyses>Exploration>Descriptives, using the "Split by" input, to visualize distributions as histograms or box plots, and compute means, medians, standard deviations, and quartiles for any dataset. Use Analyses>Exploration>Scatterplot for scatterplots of two quantitative variables.

Comparing Groups

We use *side-by-side* plots to visualize the relationship between a quantitative variable and a categorical variable.

Some of the most interesting investigations involve comparing a quantitative variable across groups defined by a categorical variable. Sometimes when investigating the relationship between categorical and quantitative variables, it is helpful to plot them *side-by-side*.

Class Example 1.5.1 *Average household incomes broken down by U.S. geographical region*

The figure below shows a side-by-side boxplot of average U.S. household income broken down by geographical region of the U.S. Each case is one of the 50 states; the two variables are *region* (categorical) and *average household income* (quantitative).

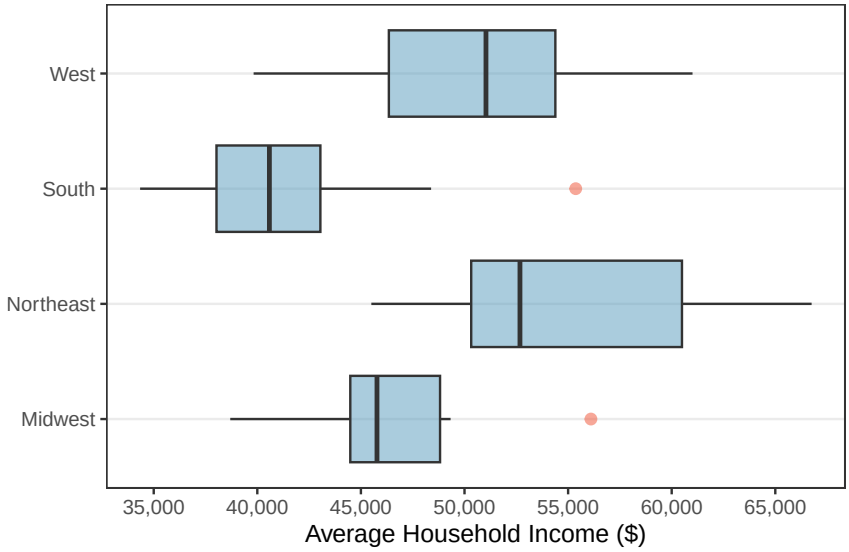


Figure 18: A side-by-side box plot of average household income split by geographic region in the US.

a) Use these side-by-side boxplots to discuss how the average U.S. household income differs by region.

b) Are there any outliers? If so, how many and estimate the values of the outliers.

c) For which region(s) would it be more appropriate to use the median instead of the mean to describe the average U.S. household income? Justify your answer.

Scatterplots

A *scatterplot* displays the relationship between two quantitative variables. Each observation is represented as a point, with one variable on the horizontal axis and the other on the vertical axis. When looking at a scatterplot, we should consider four different aspects of the plot:

A *scatterplot* is a graph of the relationship between two quantitative variables. If there are explanatory and response variables, we put the explanatory variable on the horizontal (x) axis and the response on the vertical (y) axis.

1. *Direction*: Is the association *positive* (both variables tend to increase together) or *negative* (one increases as the other decreases)?
2. *Form*: Is the relationship approximately *linear* (following a straight-line pattern) or *nonlinear* (curved)?
3. *Strength*: Is the association *strong* (points closely follow a pattern), *moderate*, or *weak* (points are loosely scattered)?
4. *Unusual observations*: Are there any *outliers* — points that deviate markedly from the overall pattern?

Example 5.2 Describing Associations

Consider a dataset of 50 loans, where we plot total income on the horizontal axis and loan amount on the vertical axis:

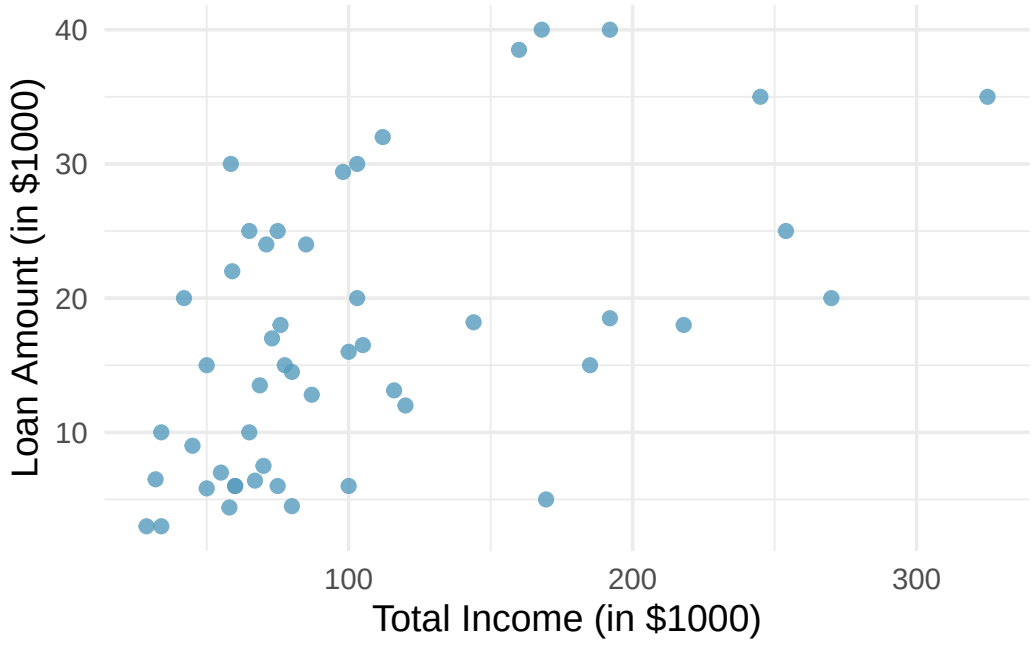


Figure 19: A scatterplot of loan amount versus total income for 50 loans. There is a moderate positive association.

Describe the association seen in the scatterplot in Figure 2.

Associations vs. causation

We have explored the association between variables, but an association is not the same as causation. When two variables are associated, it means they tend to vary together. That does not necessarily mean one variable *causes* the other. There may be a *confounding variable* that is related to both, creating the illusion of a direct relationship.

A *confounding variable* is a third variable that is associated with both the explanatory and response variable, but not necessarily of interest in the study.

Example 5.3 GPAs and Breakfast

A study finds that students who eat breakfast tend to have higher GPAs than students who skip breakfast. Can we conclude that eating breakfast causes higher GPAs?

Summary

This chapter introduced tools for exploring relationships between variables.

- For comparing a *quantitative variable across groups* defined by a categorical variable, side-by-side box plots provide a compact comparison of centers and spreads, faceted histograms reveal full distributional shapes, and ridge plots overlay density curves for compact multi-group comparison.
- For *two quantitative variables*, scatterplots reveal the direction, form, strength, and unusual features of an association.
- We closed with the fundamental principle that *association does not imply causation* — confounding variables can create misleading associations, and establishing causation generally requires a randomized experiment.

2.1 Sampling Variability

So far in this course, you have learned how to collect data, visualize it, and compute summary statistics. But statistics is about more than describing the data you have — it is about using data to learn about a larger population. The central challenge is this: ***every sample is different****. If you took another sample from the same population, you would get different results. This chapter introduces the foundational ideas that make statistical inference possible: the distinction between parameters and statistics, the concept of a sampling distribution, and the standard error as a measure of how much sample statistics vary. These ideas are the bridge from describing data to drawing conclusions about the world.*

Key Concepts

- Distinguish between a population parameter and a sample statistic, recognizing that a parameter is fixed while a statistic varies from sample to sample
- Compute a point estimate for a parameter using an appropriate statistic from a sample
- Recognize that a sampling distribution shows how sample statistics tend to vary
- Recognize that statistics from random samples tend to be centered at the population parameter
- Estimate the standard error of a statistic from its sampling distribution
- Explain how sample size affects a sampling distribution

Parameters vs. Statistics

In Unit 1, *statistical inference* was defined as the process of using a *statistic* calculated from a sample to estimate a *parameter*, or unknown value, from the population. The reason this cycle works is that when we gather data properly (i.e. use a random sampling technique), the statistics behave in a predictable way. Sample statistics are rarely, if ever, the exact population parameter, but they are usually close. If we only have one sample and we don't know the value of the population parameter, the sample statistic is our best estimate of the true population parameter (called a *point estimate* since it's a single number.)

Statistical inference is using the information in a sample to draw conclusions about the entire population.

A *parameter* is a number that describes some aspect of the population.

A *statistic* is a number that is computed from the data in a sample.

We use the sample statistic as a *point estimate* for the population

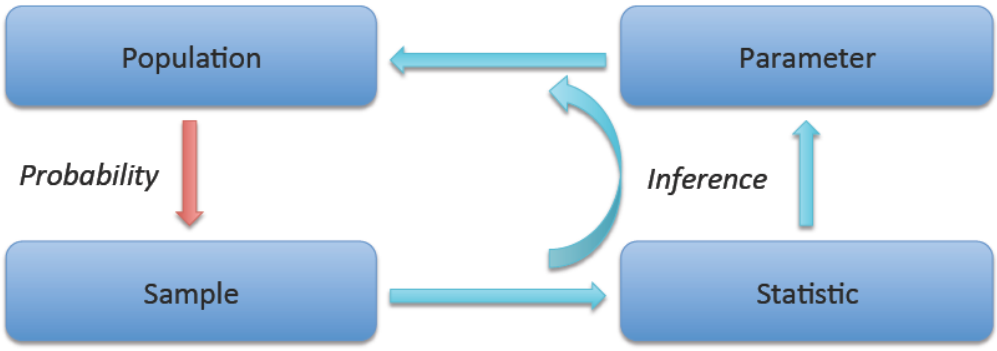


Figure 20: The cycle of statistical inference.

We use specific notation to distinguish parameters from statistics:

Quantity	**Parameter (population)**	**Statistic (sample)**
Proportion	p	$\hat{p}$
Mean	μ	$\bar{x}$
Standard deviation	σ	s
Difference in proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$
Difference in means	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$

Class Example 2.1.1 *Parameter vs. Statistic*

For each of the following, determine whether the quantity described is a parameter or a statistic and give the proper notation.

a) Average height of players on the 2024 Brazil World Cup team, using data from all 23 players on the roster.

b) Average daily high temperature in La Crosse in June, based off of randomly selecting 100 June days over the last 20 years.

c) Proportion of households in the U.S. who are renting their living accommodations using data from the 2020 Census.

d) Proportion of UWL students who live on campus, using data from 30 students.

Sampling Distributions

The *sampling distribution* shows how a statistic will vary from sample to sample.

To understand how far a statistic might be from its parameter, we need to understand what happens when we take many samples from the same population. A sampling distribution is the distribution of sample statistics computed for different samples of the same size drawn from the same population. The sampling distribution shows us how the sample statistic varies from sample to sample and gives us a way to see how close we can expect statistics to be to the true parameter value.

Class Example 2.1.2 *Illustrating sampling distributions*

Go to the Sampling Distribution Lab in StatLens. We will use the Right-Skew population shape. We do not have access to the population, but we wish to estimate the (then unknown) population mean.

a) What is the popuation mean (μ)?

b) Use StatLens to select a sample size of 10. What was the sample mean ($\overline{x}$) of this data set? Is it close to the popuation mean?

c) Keep taking samples of size 10, one by one. Visualize how the sample means of these samples change. Even though the sample means change from sample to sample, where do they appear to center?

d) Now generate 1000 samples at once to get a fuller picture of the sampling distribution of the sample mean. Where is the distribution centered? What range do the sample means lie within?

e) Reset the plot and now consider drawing samples of size 100 and exploring the sample mean. Begin with drawing samples one by one. Where do the sample means tend to center? Do they deviate much from this center?

f) Now generate 1000 samples at once. Where is the sampling distribution centered? What range do the sample means lie within? How does this range compare to the range from part d)?

g) What accounts for the differences in variability from part d) to part f)?

Facts about sampling distributions

For many common statistics, if the sample size is large enough and representative of the population (i.e. a random sample), the sampling distribution has three important features:

1. *Center*: The sampling distribution is centered at the population parameter. On average, the sample statistic equals the parameter.
2. *Spread*: The sampling distribution has a measureable spread that decreases as the sample size increases. Larger samples give more precise estimates.
3. *Shape*: For many statistics (e.g. $\overline{x}$ and $\hat{p}$), the sampling distributionis approximately bell-shaped (normal) when the sample size is large enough.

Sampling distributions tell us useful information about which sample statistics are likely to occur. Sample statistics that fall near the center of the sampling distribution are more likely to occur than those that fall far from the center.

The *sampling distribution* is fundamentally different from the *data distribution*:

- The data distribution shows how individual observations vary.
- The sampling distribution shows how a statistic (like the mean or proportion) varies from sample to sample.

Variability of sample statistics

As the sample size *increases* the variability of a statistic will *decrease*.

When gathering data, there is a tradeoff between how large a sample is and the accuracy of the statistic calculated from the sample. This happens because with a larger sample size, we get a more complete picture of what the population looks like. So, as the sample size increases, we are more likely to see sample statistics that are closer to the true value of the population parameter (i.e. the variability of the sample statistics tends to decrease). We saw this in the last example where we explored the sampling distribution for the mean in StatLens.

Two factors control how much a sample statistic bounces around from sample to sample:

1. *Sample size* (n): Larger samples produce statistics that are closer to the parameter. This is intuitive — a sample of 1,000 tells you more about the population than a sample of 10.

2. *Population variability*: If the population is very homogeneous (everyone is similar), then any sample will look like any other sample, and the statistic won't vary much. If the population is very heterogeneous (lots of individual differences), then different samples can look quite different, and the statistic will vary more.

Class Example 2.1.3 *Average pH levels in Florida lakes*

The figure below shows two sampling distributions for the average pH level in Florida lakes for two different sample sizes.

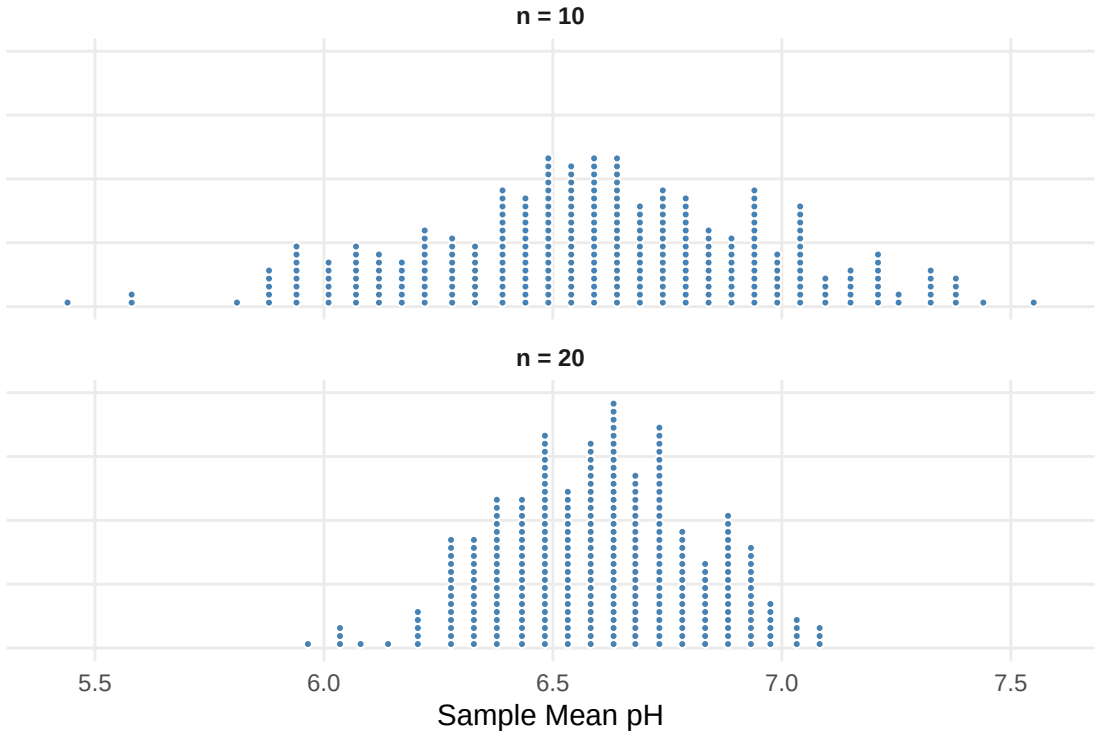


Figure 21: Two sampling distributions for average pH levels in Florida lakes.

a) The difference between these two dotplots is that one comes from samples where $n = 10$ and the other comes from samples where $n = 20$. Identify which dotplot corresponds to each sample size.

- $n = 10$ Top Bottom
- $n = 20$ Top Bottom

b) What does each dot in the dotplots represent?

c) For which sample size are we more likely to find an average pH greater than 7?

- Circle one: $n = 10$ $n = 20$

Standard Error

In a sampling distribution the *standard error* is the standard deviation of the statistic.

We need a single number that quantifies how much a statistic varies from sample to sample. That number is the *standard error*. It measures the typical distance between a sample statistic and the population parameter. The standard error of a statistic is very important because it gives us a sense of how certain we are that our statistic is accurate, so it is important that we can easily estimate the standard deviation of the sampling distribution. When we use a computer to generate a sampling distribution, it is very easy to have the computer calculate the standard deviation of the sampling distribution. However, there is also a graphical technique we can use to estimate the standard deviation of a distribution.

An *inflection point* is a point on a distribution where the up-cup becomes the down-cup.

Using some ideas from calculus, we can say the first standard deviation on either side of the mean occurs at the *inflection points*. Essentially, when tracing along the distribution from left to right, when your pen stops making an up-cup (a cup that can hold water) and starts making a down-cup (cannot hold water), that occurs approximately one standard deviation from the mean. The same occurs on the other side of the mean.

Class Example 2.1.4 *Estimating standard errors from a dot plot*

Refer back to the Figure in the previous example. Use the method above to estimate the standard error for each data set.

Do not confuse the standard error with the standard deviation of the data.

- The *standard deviation* (s or σ) measures how much individual observations vary around the mean.
- The *standard error* (SE) measures how much a sample statistic varies around the population parameter from sample to sample.

Why the standard error matters

The standard error is the engine of statistical inference. Here is why:

- *Confidence intervals* are built by going a certain number of standard errors above and below the point estimate. A rough 95% confidence interval is:

$$\text{statistic} \pm 2 \times SE$$

- *Hypothesis tests* compare an observed statistic to what we would expect under the null hypothesis. The comparison is made in units of standard errors: "Is the observed result more than 2 standard errors from the null value?"
- *Sample size planning* uses the standard error to determine how many observations are needed to achieve a desired level of precision.

In the coming chapters, we will use simulation to *estimate* the standard error. In later chapters (Unit 3), we will use mathematical formulas to *calculate* it directly.

Words of caution

The statistics calculated from a sample are only reliable if we collect our data *correctly* and have a *representative* sample. When we do not use proper techniques to gather the data, then the statistics we calculate will be biased.

Summary

- A *population parameter* is a fixed but unknown numerical summary of the population. A *sample statistic* (point estimate) is computed from data and serves as our best guess of the parameter.
- Different samples from the same population yield different statistics. This is *sampling variability* — it is not a mistake, but a fundamental feature of working with samples.
- The *sampling distribution* of a statistic describes how the statistic varies across all possible samples of a given size. It is typically centered at the parameter, and its spread decreases as sample size increases.
- The *standard error* is the standard deviation of the sampling distribution. It measures the precision of the statistic as an estimate of the parameter.
- The standard error is not the same as the standard deviation of the data. The standard deviation measures variability of individual observations; the standard error measures variability of a statistic.
- Larger samples produce smaller standard errors, meaning more precise estimates.

2.2 *Randomization Tests*

Statistical inference is primarily concerned with understanding and quantifying the uncertainty of parameter estimates. While the equations and details change depending on the setting, the foundations for inference are the same throughout all of statistics. In this chapter, we introduce the ***hypothesis testing framework*** *— a formal method for evaluating competing claims about a population using data. We learn how to build* ***randomization distributions*** *by shuffling data under the assumption that there is no effect, and we use these distributions to compute* ***p-values*** *that quantify the strength of the evidence.*

Key Concepts

- Recognize when and why statistical tests are needed
- Specify null and alternative hypotheses based on a question of interest, defining relevant parameters
- Interpret a p-value as the probability of results as extreme as the observed results happening by random chance, if the null hypothesis is true
- Estimate a p-value from a randomization distribution
- Comparatively quantify strength of evidence using p-values
- Distinguish between one-tailed and two-tailed tests in estimating p-values
- Make a formal decision in a hypothesis test by comparing a p-value to a given discernibility level
- State the conclusion to a hypothesis test in context
- Interpret Type I and Type II errors in hypothesis tests
- Recognize a discernibility level as measuring the tolerable chance of making a Type I error
- Recognize that statistical discernibility is not always the same as practical significance

Motivation: Cereal example

Consider the following scenario: A sample of 80 four-to six-year olds participated in a study in which they were asked to taste two cereals, both called "Oat Bits" (in random order). The cereal in each box was exactly the same, but the boxes differed in the fact that one of the boxes had pictures of a star mascot and the other did not.

Figure 22: Boxes for the cartoon marketing cereal study.

a) If the cartoons make no difference, because after all the cereal is the same in each box, how many of the 80 children do you expect to pick the cartoon box? What sample proportion is this?

b) Suppose that in this study, 52 of the children chose the box with cartoons. What sample proportion is this? Does this lead you to believe that there was a preference toward the cartoon box, or would you expect this sample proportion to occur by chance even if there was no preference? (We will use StatLens to explore the sampling distribution for the proportion to determine what values are likely to occur.)

c) Using the simulation from part b), what sample proportions do we expect to see if there is no preference toward the cartoons?

The hypothesis testing framework

There is almost always variability in data — one dataset will not be identical to a second dataset even if they are both collected from the same population using the same methods. However, quantifying the variability in the data is neither obvious nor easy. Answering the question "*how* different is one dataset from another?" is not trivial.

The question at the heart of hypothesis testing is: *Could the pattern we observe in the data have arisen from chance alone, or is there something real going on?*

Hypothesis testing

A *hypothesis test* is a statistical technique used to evaluate competing claims using data. The process involves:

A *hypothesis test*, or *statistical test* uses data from a sample to assess a claim about a population.

1. *State the hypotheses*: Define a null hypothesis (H_0) and an alternative hypothesis (H_A).
2. *Collect data*: Gather evidence that bears on the hypotheses.
3. *Assess the evidence*: Determine how likely the observed data would be if the null hypothesis were true.
4. *Draw a conclusion*: Decide whether the evidence is strong enough to reject the null hypothesis.

If the null hypothesis and the data notably disagree, then we reject the null hypothesis in favor of the alternative hypothesis. There are many nuances to hypothesis testing, and we will discuss these ideas throughout this chapter and those that follow.

Null and alternative hypotheses

Usually, a *null hypothesis*, notated H_0, is a claim that there really is "no effect" or "no difference", or represents the status quo.

An *alternative hypothesis*, notated H_A, is the claim for which we seek discernible evidence, and is established by observing evidence that contradicts the null hypothesis and supports the alternative hypothesis.

In the first step of a hypothesis test, we must state the hypotheses. The hypotheses consist of the *null hypothesis* and the *alternative hypothesis*. The null hypothesis, denoted H_0, often represents a skeptical perspective or a claim of "no difference" or "no effect." It assumes that any observed pattern in the data is due to chance alone. The alternative hypothesis, denoted H_A, represents the claim under investigation, typically that there is a real difference, effect, or relationship. The alternative hypotheisis is usually the reason the research was conducted in the first place.

Statistical tests, and in particular hypothesis tests, behave in much of the same way as the US justice system. The US court system considers two possible claims about a defendant: they are either innocent or guilty. We can set up these claims in a hypothesis framework. The skeptical perspective (null hypothesis) is that the person is innocent until evidence is presented that convinces the jury the person is guilty (alternative hypothesis).

$$H_0 : \text{The defendant is innocent.}$$
$$H_A : \text{The defendant is guilty.}$$

Caution

Notice that if a jury finds a defendant *not guilty*, this does not necessarily mean the jury is confident in the person's innocence. They are simply not convinced of the alternative. This is also the case with hypothesis testing: *even if we fail to reject the null hypothesis, we do not accept the null hypothesis as truth.* Failing to find evidence in favor of the alternative is not the same as finding evidence that the null hypothesis is true.

Sets of common hypotheses

The first set of statistical hypotheses that we will investigate in this class are for population parameters and differences in population parameters. The hypotheses have the following forms:

1. Single Proportion (p_0 is the proportion claimed or status quo value):
   - Null hypothesis:
   $$H_0 : p = p_0$$
   - Alternative hypothesis (pick one):
   $$H_A : p \neq p_0$$
   $$H_A : p > p_0$$
   $$H_A : p < p_0$$
2. Single Mean (μ_0 is the average claimed or status quo value):
   - Null hypothesis:
   $$H_0 : \mu = \mu_0$$
   - Alternative hypothesis (pick one):
   $$H_A : \mu \neq \mu_0$$
   $$H_A : \mu > \mu_0$$
   $$H_A : \mu < \mu_0$$

3. Difference in Proportions (p_1 and p_2 are the population proportions for each group):
   - Null hypothesis:
   $$H_0 : p_1 - p_2 = 0$$
   - Alternative hypothesis (pick one):
   $$H_A : p_1 - p_2 \neq 0$$
   $$H_A : p_1 - p_2 > 0$$
   $$H_A : p_1 - p_2 < 0$$
4. Difference in Means (μ_1 and μ_2 are the population averages for each group):
   - Null hypothesis:
   $$H_0 : \mu_1 - \mu_2 = 0$$
   - Alternative hypothesis (pick one):
   $$H_A : \mu_1 - \mu_2 \neq 0$$
   $$H_A : \mu_1 - \mu_2 > 0$$
   $$H_A : \mu_1 - \mu_2 < 0$$

Alternative hypotheses containing "$>$" or "$<$" are called *one-sided* alternatives and alternative hypotheses containing "$\neq$" are referred to as *two-sided* alternatives. One-sided alternatives specify a particular direction for the alternative, while two-sided alternatives do not.

Class Example 2.2.1 *Defining statistical hypotheses*

For each of the following scenarios, state the null and alternative hypotheses with the relevant parameter(s).

a) We are testing to see if the average income is different in eastern states than western states.

b) We are testing to see if the average income is greater in eastern states than western states.

c) We are testing to see if a particular coin is weighted (i.e., not fair).

d) We are testing to see if the average rent in La Crosse is less than $850.

e) We are testing to see if children have a preference for the cereal box with cartoons (above).

Case study: Opportunity cost

How rational and consistent is the behavior of the typical American college student? Here, we explore whether reminding students that money not spent now can be spent later causes them to be thriftier.

The participants were 150 students, all given a scenario about purchasing a video on sale for $14.99. Half (the control group) were given these options:

A. Buy this entertaining video.

B. Not buy this entertaining video.

The other half (the treatment group) saw a slightly modified option (B):

A. Buy this entertaining video.

B. Not buy this entertaining video. Keep the $14.99 for other purchases.

In this study, we define a "success" as a student who chooses *not* to buy the video and want to compare the proportion of students who do not buy the video in the control group compared to the treatment group.

Class Example 2.2.2 *Opportunity cost*

Using the opportunity cost case study, answer the following questions:

a) Is this an observational study or an experiment? How does the type of study impact what can be inferred from the results?

b) State the null and alternative hypotheses for this study.

The results are summarized below:

	Buy video	**Not buy video**	**Total**
Control	56	19	75
Treatment	41	34	75
Total	97	53	150

Class Example 2.2.3 *Opportunity cost, revisited*

Compute the difference in the sample proportions of students who choose not to buy the video.

Building a randomization distribution

What would it mean if the null hypothesis were true? It would mean that each student would make the decision on whether to buy the video without regard to the change in the second option, (B). The 97 purchases and 53 non-purchases were determined by factors other than the verbage of the options, and the difference we observed was just bad luck in how the students happened to be assigned to the two groups.

Suppose the students' decision were indpendent of the phrasing of the choices. Then, if we conducted the experiment again with a different random assignment of choices to the students, differences in the proportions of the choices would be based only on random fluction in decisions. We can perform this *randomization*, which simulates what would have happened if the students' decisions had been independent of the phrasing but we had distributed the phrases differently.

The card-shuffling metaphor

Think of the data as 150 cards:

- 97 red cards (buy video)
- 53 white cards (not buy video)

In the actual study, 75 cards were dealt to the 'control' pile and 75 to the 'treatment' pile. Under the null hypothesis, the color of each card (whether to buy the video) has nothing to do with which pile it lands in.

To simulate what the null hypothesis look like, we:

1. *Thoroughly shuffle* all 150 cards together.
2. *Deal* 75 into a 'control' pile and 75 into a 'treatment' pile
3. *Count* the number of white (not buy video) in each pile and compute the difference in proportions.

This gives us one simulated difference *under the null hypothesis*, a difference that arose purely from the random shuffle, not from the phrasing.

Many simulations

A *randomization distribution* approximates a sampling distribution from a population where the null hypothesis is true.

One simulation is not enough. We need to repeat the shuffle-and-deal process many times to build up a picture of what chance alone looks like. Using a computer, we can do this hundreds or thousands of times creating a *randomization distribution* showing how likely different statistics are to occur if H_0 is true.

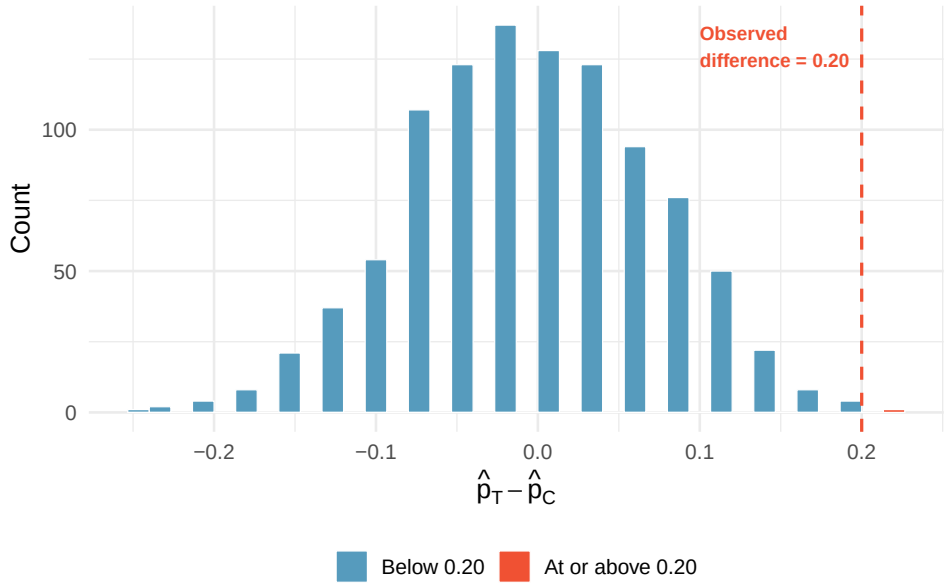


Figure 23: A histogram of 1,000 simulated differences produced under the null hypothesis. Differences at least as large as the observed 0.20 are highlighted in red.

Under the null hypothesis, we would observe a difference of at least +20% about 0.6% of the time. That is very rare! We conclude the data provide strong evidence there is a treatment effect: reminding students that they could spend money later lowers their chance of making a purchase. Because this is an experiment (with random assignment), we can make a causal claim.

P-values

The p-*value* of the sample data in a statistical test is the probability of obtaining a sample statistic as extreme as (or more extreme than) the observed sample when the null hypothesis is *true*.

When we claim that we have evidence of our alternative hypothesis, the underlying question we are asking is

> How unusual is it to see a sample statistics as extreme as the sample statistic we observed, *if* H_0 *is true*?

In statistics, we use the p-value to answer this question. The p-value is the proportion of the randomization distribution that is as extreme (or more extreme than) over observed statistic. Since the randomization distribution is built assuming the null hypothesis is true, the p-value is also defined as the probability of observing data at least as favorable to the alternative hypothesis as our current dataset, if the null hypothesis were true. We typically use a summary statistic of the data, such as a difference in the proportion, to compute the p-value.

If the summary statistic lies in the tails of the distribution, we have strong evidence against H_0 in support of H_A. The smaller the p-value, the more evidence we have to reject the null hypothesis.

Randomization Distributions and p-values

To find the p-value using a randomization distribution, we:

- *For a one-sided alternative hypothesis:* Find the proportion of randomization samples that equal or exceed the original statistic in the direction (tail) indicated by the alternative hypothesis.
- *For a two-sided alternative hypothesis:* Find the proportion of randomization samples in the smaller tail at or beyond the original statistic and then double the proportion to account for the other tail.

In order to correctly estimate the p-value from a randomization distribution, we need to keep our alternative hypothesis in mind!

Class Example 2.2.4 *Estimating a p-value from a randomization distribution*

The figure below shows a randomization distribution for testing $H_0 : p = 0.3$ vs $H_A : p < 0.3$. In each case, use the distribution to decide which value is closer to the p-value for the observed sample proportion. There are 1000 randomizations given in the figure below.

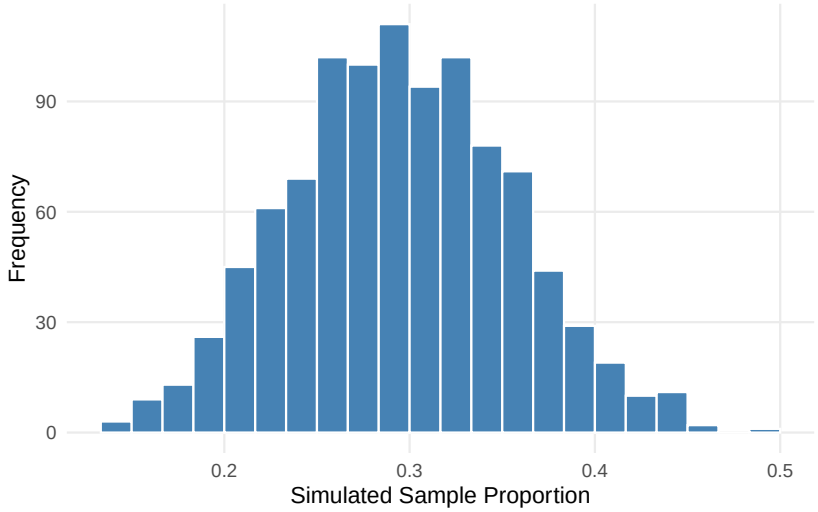


Figure 24: Randomization distribution with 1000 randomizations

a) Why is the distribution centered at 0.3?

b) The p-value for $\hat{p} = 0.15$ is closest to: 0.04 or 0.40?

c) Which proportion, $\hat{p} = 0.20$ or $\hat{p} = 0.10$, provides the strongest evidence to suggest that H_0 is not true? Explain.

Class Example 2.2.5 *Estimating a p-value from a randomization distribution...continued*

The figure below shows a randomization distribution for testing $H_0 : \mu_1 - \mu_2 = 0$ vs $H_A : \mu_1 - \mu_2 \neq 0$. The statistic used for each sample is $D = \overline{x}_1 - \overline{x}_2$. In each case, use the distribution to decide which value is closer to the p-value for the observed sample difference in averages. There are 1000 randomizations given in the figure below.

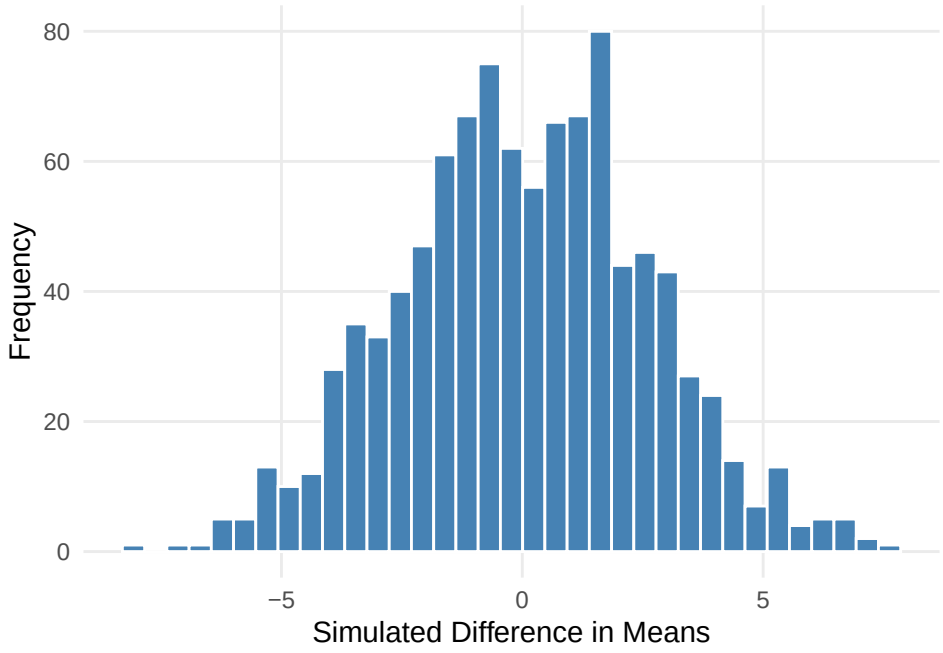


Figure 25: Randomization distribution with 1000 randomizations

a) The p-value for $D = \overline{x}_1 - \overline{x}_2 = -2.9$ is closest to: 0.01 or 0.25?

b) The p-value for $D = \overline{x}_1 - \overline{x}_2 = 1.2$ is closest to: 0.30 or 0.60?

c) Of the two values of D above, which provides strongest evidence to suggest that H_0 is not true? Explain.

Class Example 2.2.6 *Migraine Acupuncture Trial*

Can acupuncture relieve migraine pain? Allais et al. (2011) conducted a randomized controlled study where 89 individuals who suffered from migraines were randomly assigned to receive either acupuncture (treatment, $n = 43$) or placebo acupuncture (control, $n = 46$). The placebo acupuncture involved placing needles at non-acupuncture-point locations. Researchers recorded whether each patient was pain-free 24 hours after treatment. The full data are in the data set *Migraine Acupuncture Trial* on StatLens.

a) What are the null and alternative hypotheses for this study?

b) What was the sample statistic for these data?

c) What is the probability that the sample results (or those more extreme) occurred by random chance, assuming that the acupuncture had no effect?

d) How would your answer to c) change if the alternative hypothesis was two-sided?

Class Example 2.2.7 *Calorie Counts*

Is the average amount of kilocalories from a major U.S. fast food chain menu item more than 525 kilocalories? Calorie counts for 515 menu items from eight major fast food chains were collected. The full data are in the data set *Fast Food Calories* on StatLens.

a) State the null and alternative hypotheses

b) What is the sample statistic?

c) To create the randomization distribution, what do we have to assume?

d) What is the p-value for this study?

Class Example 2.2.8 *Cereal study revisited*

Recall the movtivating cereal study. We were trying to determine if children had a preference for the cereal in the box with the cartoon. In the study 52 out of 80 children chose the box with cartoons. Use StatLens to find the p-value for this scenario.

Interpreting the p-value

- A *small p-value* (e.g., below 0.05) means that the observed result would be rare if the null hypothesis were true. This is evidence *against* the null hypothesis.
- A *large p-value* (e.g., above 0.10) means that the observed result is not particularly unusual under the null hypothesis. The data do not provide strong evidence against the null hypothesis.
- The p-value is *not* the probability that the null hypothesis is true. It is the probability of the observed data (or more extreme data) *given that* the null hypothesis is true.

Caution

A common mistake is to interpret the p-value as the probability that the null hypothesis is true (e.g., "there is a 2% chance that sex has no effect on promotions"). This is incorrect. The p-value tells us how *surprising* the data would be *if* the null hypothesis were true, not how likely the null hypothesis is to be true.

Statistical discernibility

Statistical tests begin under the assumption that the null hypothesis,H_0, is true. In order to reject H_0, sufficient evidence to its contrary must be collected – just like in order to convict a defendant, sufficient evidence must be presented to the judge and/or jury. What exactly is sufficient evidence? In our courtroom example, evidence must be provided so that decisions can be made "beyond a reasonable doubt." In statistical tests, the data results must be unrealistic assuming H_0 is true. This means that we have found evidence to support H_A.

Decision framework

When we make a formal decision, we have two options:

- Reject H_0
- Do not reject H_0

For hypothesis tests, there are two conclusions:

- We *reject* H_0 when the sample statistic is extreme assuming H_0 is true.
- We *do not reject* H_0 when the sample statistic is not too extreme assuming H_0 is true.

If we reject the null hypothesis, we say our finding is *statistically discernible*.

If we reject the null hypothesis, we say our finding is *statistically discernible*. Not rejecting the null hypothesis is *not* the same thing as accepting the null hypothesis. The null hypothesis is *never accepted* because the hypothesis test procedure is not designed to *rule in* H_0, but only to rule it out. That is, evidence is never collected with the idea of favoring or proving the null hypothesis. Sample data are collected for the purpose of looking for evidence *against* H_0. To arrive at the correct conclusion, the p-value must be small enough to be considered unusual.

What "statistically discernible" does and does not mean

"Statistically discernible" means the p-value fell below the chosen discernibility level. It does *not* necessarily mean the effect is large or practically important. A very large sample can detect a tiny difference as "statistically discernible" even if the difference is too small to matter in practice.

Conversely, "not statistically discernible" does not mean there is no effect. It may simply mean the sample was too small to detect the effect.

How small is small enough?

We discussed how small p-values mean that a statistic is unlikely if H_0 is true. Now we need to think about some rules to determine when our p-value is small enough to say something is unusual assuming the null hypothesis is true.

We use the *discernibility level*, α, to decide the cutoff for our p-value to be discernible.

The *discernibility level*, notated α, is the point at which a p-value is considered unusual. Some commonly used discernibility levels are $\alpha = 0.05$,$\alpha = 0.01$, or $\alpha = 0.10$. Any p-value that is smaller than the discernibility level is considered unusual. So, to use a p-value to determine whether or not the results are discernible, you should use the following guidelines:

Common values for $\alpha : 0.05, 0.01, 0.10$

We reject H_0 when p-value $< \alpha$

We do not reject H_0 when p-value $\geq \alpha$

- When the p-value $< \alpha$, reject H_0.
- When the p-value $\geq \alpha$, do not reject H_0.

Class Example 2.2.9 *Red wine and weight loss*

Resveratrol, an ingredient in red wine and grapes, has been shown to promote weight loss in rodents. A recent study (*Science Daily*, 2010) investigates whether the same phenomenon holds true in primates. A sample of six lemurs had their resting metabolic rate, body mass gain, food intake, and locomotor activity measured for one week prior to resveratrol supplementation and then the four indicates were measured again after treatment. For each of the following, use the p-value to make the appropriate conclusion (reject or do not reject) for the hypothesis tests. Use $\alpha = 0.05$ for each of your decisions.

a) In a test to see if mean resting metabolic rate is higher after treatment (p-value of 0.013).

b) In a test to see if mean body mass gain is lower after treatment (p-value of 0.007).

c) In a test to see if mean locomotor activity is affected by treatment (p-value of 0.980).

Interpreting your results

After you have made a decision to reject H_0 or not to reject H_0, you need to interpret your conclusion in the context of the study. We always write our interpretation in the context of H_A.

A framework for interpreting your decision is as follows (fill in the blanks with the context of the study):

> At $\alpha =$ (discernibility level), we (do/do not) have discernible evidence that (alternative hypothesis in words).

Class Example 2.2.10 *Red wine and weight loss revisted*

For each of the following, provide an interpretation of the conclusion in the context of the original study.

a) In a test to see if mean resting metabolic rate is higher after treatment (p-value of 0.013).

b) In a test to see if mean body mass gain is lower after treatment (p-value of 0.007).

c) In a test to see if mean locomotor activity is affected by treatment (p-value of 0.980).

Class Example 2.2.11 *Migraine Acupuncture Trial revisited*

Can acupuncture relieve migraine pain? Using your decision from the Acupunture Trial example, interpret your results in the context of the problem.

Class Example 2.2.12 *Calorie count revisited*

Is the average amount of kilocalories from a major U.S. fast food chain menu item more than 525 kilocalories? Using your decision from the Calorie Count example, interpret your results in the context of the problem.

Class Example 2.2.13 *Cereal study revisited*

Recall the movtivating cereal study. Interpret the results of the study in the context of the problem.

Type I and Type II errors

A *Type I error* occurs when we reject H_0, but it is true (e.g., convict an innocent person).

A *Type II error* occurs when we do not reject H_0, but it is false (e.g., let a guilty person go free).

Every hypothesis test ends with one of two decisions: reject H_0 or fail to reject H_0. And the truth is one of two realities: H_0 is true or H_A is true. Combining these gives four scenarios:

	H_0 is true	**H_A is true**
Reject H_0	Type I error	Correct decision
Fail to reject H_0	Correct decision	Type II error

Four possible outcomes of a hypothesis test.

Two of the four outcomes are correct decisions:

- Rejecting H_0 when H_A is true: we correctly detected the real effect.
- Failing to reject H_0 when H_0 is true: we correctly avoided a false alarm.

The other two outcomes are errors:

- *Type I error* (upper-left): We rejected H_0 but it was actually true. False alarm.
- *Type II error* (lower-right): We failed to reject H_0 but H_A was actually true. Missed detection.

Class Example 2.2.14 *Red wine and weight loss revisted*

For each of the following, identify which kind of error could have occurred (Type I or Type II) using $\alpha = 0.05$.

a) In a test to see if mean resting metabolic rate is higher after treatment (p-value of 0.013).

b) In a test to see if mean body mass gain is lower after treatment (p-value of 0.007).

c) In a test to see if mean locomotor activity is affected by treatment (p-value of 0.980).

What can cause these errors?

The two types of errors that can be made both involve the hypothesis test leading to the wrong conclusion about the null hypothesis. This can happen because the process of random sampling does not give complete or perfect information about the population, and it is always possible to draw a rare sample just by chance. Recall that we have two options when making statistical inferences: *reject* H_0 or *do not reject* H_0. It is important to carefully consider what these decisions could mean.

- Decision: Reject H_0. The following outcomes are possible:
  1. You are making the correct decision; H_0 is not true.
  2. You are making a Type I error; the sample statistic was so far from the hypothesized value just by chance.
  3. Your sampling process or experiment was poor; the sample statistic is biased.
- Decision: Do not reject H_0
  1. You are making the correct decision; H_0 is true.
  2. You are making a Type II error; the sample statistic was close to the hypothesized value just by chance.
  3. Your sampling process or experiment was poor; the sample statistic is biased.

Balancing out the errors

The discernibility level, α, is the probability we make a Type I error.

The discernibility level, α, is the probability we make a Type I error. If we can choose the discernibility level and it gives the chance of an error being made, why not always make the discernibility level really, really low? This is because there are trade-offs between controlling for Type I and Type II errors, and balancing the risk of these errors is important.

The *power* of a test tells us how likely we are to reject H_0 when we should.

1. Setting α to be very small makes it difficult to make a Type I error, but it also makes it easier to make a Type II error. Conversely, setting α to be large increases the making a Type I error, but it also reduces the risk of making a Type II error.
2. Since a Type I error is usually thought to be more serious than a Type II error, we typically use a low $\alpha =$ Type I error probabilities like 0.05, 0.1, or 0.01
3. We can reduce the chance of a Type II error without impacting the risk of a Type I error by increasing the sample size.
4. Statistical *power* tells us how likely we are to reject H_0 when we should (so we want *higher* power!), and it is related to the risk of a Type II error by power = 1 - chance of Type II error.
5. Calculating the probability of Type II error and power are fairly complicated for most cases, so we won't discuss that in this class. But understanding the relationships between these errors is important.

The *power* of a test tells us how likely we are to reject H_0 when we should

Class Example 2.2.15 *Errors in a drug study*

A pharmaceutical company is researching a new anti-anxiety drug. Marketing a new drug is costly, and the company only makes a profit when the drug is effective for over 80% of patients. A random sample of 60 participants are placed on the drug and the number of participants that respond favorably to the drug will be recorded.

a) Give the null and alternative hypothesis that you would use for determining if the drug should be marketed.

b) Describe in words what a Type I error would mean in the context of this study.

c) Describe in words what a Type II error would mean in the context of this study.

d) Suppose the discernibility level is set at a = .05, but the researchers find that this produces a Type II error rate of 0.4 and this is too high for their liking. List two ways they could reduce the Type II error rate.

e) Given a Type II error rate of 0.4, what is the power of the test?

The randomization test procedure: a summary

We can summarize the randomization test procedure as follows:

1. *Frame the research question in terms of hypotheses.* The null hypothesis (H_0) usually represents no difference or no effect. The alternative hypothesis (H_A) represents the claim of interest.
2. *Collect data.* If the research question focuses on associations but not causation, use an observational study. If you want to establish a causal connection, use an experiment with random assignment.
3. *Model what would happen under the null hypothesis.* Shuffle (randomize) the data to break any association between the explanatory and response variables. Compute the test statistic for each shuffle. Repeat many times to build the *randomization distribution* (also called the null distribution).
4. *Compute the p-value.* Determine what proportion of the randomization distribution is at least as extreme as the observed test statistic.
5. *Draw a conclusion.* If the p-value is small (less than α), reject H_0 and conclude there is evidence for H_A. If the p-value is not small, fail to reject H_0. Write the conclusion in context, in plain language.

Summary

- A *hypothesis test* evaluates two competing claims: the null hypothesis (H_0), which represents no effect or no difference, and the alternative hypothesis (H_A), which represents the claim under investigation.
- A *randomization test* (or permutation test) assesses the null hypothesis by shuffling the data to simulate what would happen if there were truly no effect. The collection of simulated test statistics forms the *randomization distribution*.
- The *test statistic* is the summary value computed from the data (e.g., a difference in proportions) that we use to evaluate the hypotheses.
- The *p-value* is the probability of observing a test statistic at least as extreme as the one observed, assuming the null hypothesis is true. It quantifies the strength of the evidence against H_0.
- We reject H_0 when the p-value is less than the *discernibility level* α (commonly 0.05). This is called a *statistically discernible* result.
- A *one-sided test* looks for effects in a specified direction; a *two-sided test* looks for effects in either direction.
- Failing to reject H_0 is not the same as proving H_0 is true.
- Every hypothesis test ends in one of four outcomes, organized by the *decision table*: two correct decisions and two errors.
- A *Type I error* rejects a true H_0 (a false alarm); its probability is the discernibility level α, which we choose.
- A *Type II error* fails to reject H_0 when H_A is true (a missed detection); which we do not directly control.
- Statistical discernibility does not imply practical importance.

2.3 Bootstrap Confidence Intervals

*In the sampling variability chapter, we saw that sample statistics vary from sample to sample. This variability is not a nuisance, it's the key to quantifying uncertainty. In this chapter, we learn **bootstrapping**: a simulation-based method for constructing confidence intervals that requires no formulas and no assumptions about the shape of the population.*

Key Concepts

- Construct a confidence interval for a parameter based on a point estimate and a margin of error
- Use a confidence interval to recognize plausible values of the population parameter
- Construct a confidence interval for a parameter given a point estimate and an estimate of the standard error
- Interpret (in context) what a confidence interval says about a population parameter
- Recognize that a bootstrap distribution tends to be centered at the value of the original statistic
- Use technology to create a bootstrap distribution
- Estimate the standard error of a statistic from a bootstrap distribution
- Construct a 95% confidence interval for a parameter based on a sample statistic and the standard error from a bootstrap distribution
- Construct a confidence interval based on the percentiles of a bootstrap distribution
- Explain how the width of an interval is affected by the desired level of confidence and the sample size
- Recognize when it is appropriate to construct a bootstrap confidence interval using percentiles or the standard error

The confidence interval

An *interval estimate* gives a range of plausible values for a population parameter, commonly formed using *point estimate*$\pm$*margin of error*.

Margin of error is a number that reflects the precision of the sample statistic as a point estimate for the parameter.

From our discussion about sampling distributions, recall that sample statistics vary from sample to sample. Any time we give a point estimate for a population parameter, we want to describe that variability. This is commonly done using an *interval estimate* or a range of values that are plausible for the population parameter. One common form of an interval estimate is

$$\text{point estimate} \pm \text{margin of error}$$

The *margin of error* reflects the precision of our statistic as a point estimate, and it indicates how far out we should go to obtain a range of plausible values for the population parameter.

Class Example 2.3.1 *Adopting a child in the US*

A survey of 1000 American adults conducted in January 2013 stated that "44% say it's too hard to adopt a child in the US." The survey goes on to say that "The margin of error is $\pm$3 percentage points with a 95% level of confidence."

a) What is the relevant sample statistic? Give appropriate notation and the value of the statistic.

b) What population parameter are we estimating with this sample statistic?

c) Use the margin of error to give a confidence interval for the estimate. Is 0.42 a plausible value of the population proportion? Is 0.50 a plausible value?

d) If we took a sample of 100 instead, do you expect the margin of error to increase or decrease?

Typically in statistics, we refer to our interval estimates as *confidence intervals*. A confidence interval is an interval estimate along with a *confidence level* that tells us how successful our intervals would be at capturing the true (but unknown) parameter if we repeated our sampling thousands of times.

A *confidence interval* is a range of plausible values for the population parameter.

The success rate (proportion of all samples whose intervals contain the parameter) is known as the *confidence level*.

Class Example 2.3.2 *Interpreting confidence level*

For each of the following indicate the number of confidence intervals that should capture the true population parameter.

a) We repeat our sampling 30 times and calculate 95% confidence intervals each time.

b) We repeat our sampling 100 times and calculate 99% confidence intervals each time.

One of the more common confidence intervals is the 95% confidence interval, and we can estimate a 95% confidence interval using

$$\text{statistic} \pm 2 \cdot \text{standard error}.$$

A *95% confidence interval* is constructed using a procedure that, when applied repeatedly to many different sample, caputers the true parameter in approximately 95% of all intervals produced

Using the standard error, the *95% confidence interval* for a sampling distribution that is relatively symmetric and bell shaped is formed using *Statistic* $\pm 2 \cdot SE$.

Class Example 2.3.3 *Constructing confidence intervals*

For each of the following, use the information to construct a 95% confidence interval and give notation for the quantity being estimated.

a) $\overline{x} = 27$ with standard error 3.2.

Interval: Notation (circle one): p, μ, $p_1 - p_2$, $\mu_1 - \mu_2$

b) $\hat{p}_1 - \hat{p}_2 = 0.05$ with margin of error for 95% confidence of 0.02.

Interval: Notation (circle one): p, μ, $p_1 - p_2$, $\mu_1 - \mu_2$

Recap of precision

For a 95% confidence interval: $MOE = 2 \cdot SE$

Before learning about the interpretation of confidence intervals, it is important to highlight some differences between the measures of precision we have discussed thus far.

- The *margin of error* is the amount we add and subtract in a confidence interval.
- The *standard error* is the standard deviation of the sampling distribution of the statistic.
- The *sample standard deviation* represents the average amount a single observation in our sample is from the sample average.

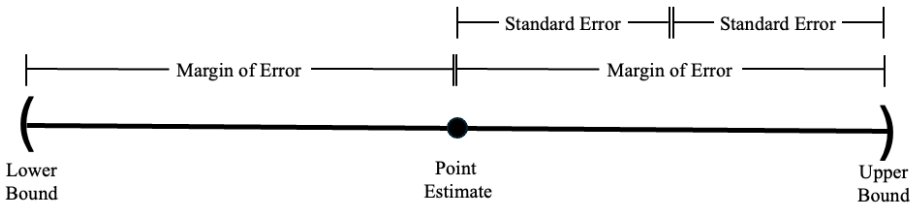


Figure 26: Diagram of the relationship between the 95% confidence interval, margin of error, and standard error.

Understanding confidence intervals

Confidence intervals appear all over the place. These intervals appear in research articles and academic publications, but they also appear in mainstream news sources as well. For example, during almost any election cycle, it is almost impossible to find any article about the upcoming vote that does not cite some kind of poll with an associated interval estimate.

Class Example 2.3.4 *Have you ever been arrested?*

According to a recent study of 7335 young people in the US, 30% had been arrested for a crime other than a traffic violation by the age of 23. Crimes included such things as vandalism, underage drinking, drunken driving, shoplifting, and drug possession.

a) Is the 30%, that the study reports, a parameter or statistic?

b) The margin of error is 0.01 for a 95% confidence interval. Use this information to give a range of plausible values for the parameter.

c) A correctional facility network claims that the true proportion of young people who had been arrested for a crime other than a traffic violation is less than 25%. Given the margin of error in part b), if we asked *all* young people in the US if they have ever been arrested, is it likely that the actual proportion is less than 25%?

Interpreting a confidence interval

We can interpret a 95% confidence interval by saying that we are 95% confident (or sure) that the parameter of interest is in the interval. It is important to note that the confidence interval only refers to the *population parameter*, it has nothing to do with individuals in the population or for the statistics calculated from other samples.

A framework for interpreting a confidence interval is as follows (fill in the blanks with the context of the study):

> We are (confidence level%) confident, that the (population parameter in words) is between (lower bound) and (upper bound).

The following are common misinterpretations of confidence intervals that you should *avoid*.

1. ~~A 95% confidence interval contains 95% of the data in the population~~: This statement is not what we mean by the 95%. The 95% indicates how confident that we are that we captured the population parameter (mean, proportion, etc.) and doesn't reflect how many individual cases are captured.
2. ~~We are 95% sure that the sample mean falls in the 95% confidence interval for the mean~~: We know the sample mean and it will always (100% of the time) fall in the confidence interval.
3. ~~The probability that the population parameter is in this particular 95% confidence interval is 0.95~~: Once an interval is constructed it either does or does not contain the population value.

Class Example 2.3.5 *BPA in soup cans*

To investigate BPA exposure from cans, 75 participants ate either canned soup or fresh soup for lunch for five days. On the fifth day, urinary BPA levels were measured. After a two-day break, the participants switched groups and repeated the process (matched pairs experiment). The difference in the BPA levels between the canned and fresh soup were measured for each participant, yielding a 95% confidence interval for the difference in means (canned minus fresh) of 19.6 to 25.5 micrograms/L.

a) Provide an interpretation of the confidence interval in the context of the study.

b) What is the margin of error for this confidence interval.

c) Does this inteval suggest that BPA levels differ for canned and fresh soup? Explain.

d) If the study had included 500 participants instead of 75, would you expect the confidence interval to be wider or narrower? Explain.

Class Example 2.3.6 *Biomass in tropical forests*

Using a sample of 4079 inventory plots, scientists give a 95% confidence interval of 9600 to 13600 tons for the mean amount of carbon per square kilometer in tropical forests.

a) What is wrong with this interpretation? "We believe that 95% of tropical forests will have an average of carbon per square kilometer that is between 9600 to 13600 tons."

b) Provide a correct interpretation of the confidence interval in the context of the study.

Class Example 2.3.7 *Tipping for pizza delivery*

Using a sample of 24 deliveries described in "Diary of a Pizza Girl", we find a 95% confidence interval for the mean tip given for a pizza delivery to be \$4.18 to \$5.90. Which of the following is a correct interpretation of this interval? Circle the correct interpretation(s).

a) I am 95% confident that all pizza delivery tips will be between \$4.18 and \$5.90.
b) 95% of all pizza delivery tips will be between \$4.18 and \$5.90.
c) I am 95% confident that the mean pizza delivery tip for this sample will be between \$4.18 and \$5.90.
d) I am 95% confident that the mean tip for all pizza deliveries in this area will be between \$4.18 and \$5.90.
e) I am 95% confident that the confidence interval for the mean pizza delivery tip will be between \$4.18 and \$5.90.

The big idea behind bootstrap sampling: resampling

A *point estimate* is a single numerical value used as a "best guess" or approximation to estimate an unknown population parameter.

In 2011, researchers collected a sample of 648 pennies in circulation and recorded how old each penny was (current year minus year minted). The question: what is the average age of all pennies currently in circulation in the United States?

The sample mean is $\overline{x} = 10.4$ years. This is our *point estimate* of the population mean μ. But how precise is this estimate? If we collected a different sample of 648 pennies, we'd get a different $\overline{x}$. We can't keep collecting new samples — we only have this one.

The *bootstrap distribution* is a collection of *bootstrap statistics*, each computed from a *bootstrap resample*. The boostrap resample is a sample of size n drawn with replacement from the original sample.

The bootstrap idea: Treat your sample as a stand-in for the population and resample from it. Each *bootstrap resample* is drawn with replacement from your original sample, at the same sample size n. The *bootstrap statistic* is computed from the boostrap resample and the collection of many bootstrap statistics forms the *bootstrap distribution*.

Bootstrapping a sample mean

Let's build a bootstrap confidence interval step by step.

Step 1: Collect data

Our sample of 648 pennies has a right-skewed distribution — many young pennies and a long tail of older ones:

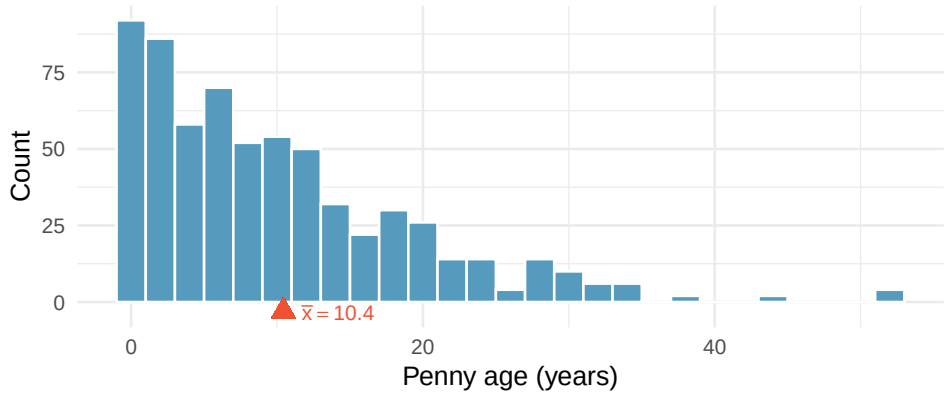


Figure 27: Distribution of ages for 648 pennies sampled from circulation. The distribution is right-skewed with a mean of about 10.4 years. The sample mean is marked with a red triangle. Open in StatLens

The sample mean is $\bar{x} = 10.4$ years. Notice the distribution is clearly not bell-shaped, it's skewed right. Bootstrapping doesn't care. It works regardless of the population shape.

Step 2: Resample with replacement

We draw a new sample of 648 values from the original 648, *with replacement*. Some pennies will appear more than once; others will be left out. Each such draw is one *bootstrap resample*, and the mean of that resample is a *bootstrap statistic* $\bar{x}^*$.

Step 3: Repeat many times

We repeat Step 2 a large number of times, typically $B = 1{,}000$ or more, and record the bootstrap mean $\bar{x}^*$ each time. This gives us a *bootstrap distribution*.

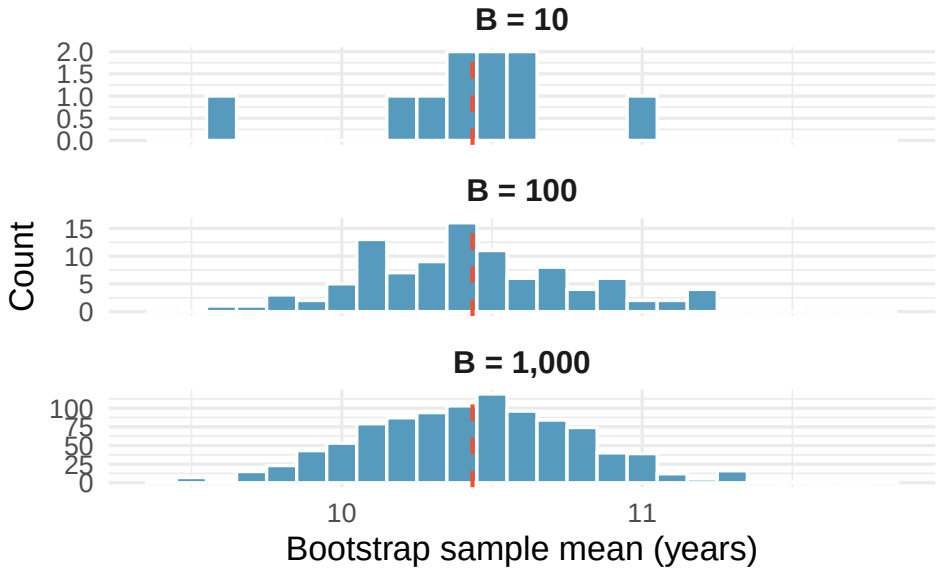


Figure 28: Building up the bootstrap distribution. With 10 resamples we see only a rough sketch; by 1,000 resamples the shape of the distribution is clear — and approximately bell-shaped (normal), even though the original data are skewed.

Notice two things: (1) the bootstrap distribution is centered near the original sample mean (dashed red line), and (2) even though the penny ages are right-skewed, the bootstrap distribution of the *mean* is approximately bell-shaped (what statisticians call a *normal* shape). This is the Central Limit Theorem at work, we'll explore it more in the next chapter.

Step 4: Find the confidence interval

For a 95% confidence interval, we find the middle 95% of the bootstrap distribution by taking the *2.5th and 97.5th percentiles*.

The *Pth percentile* is the value of a quantiative variable which is greater than P percent of the data

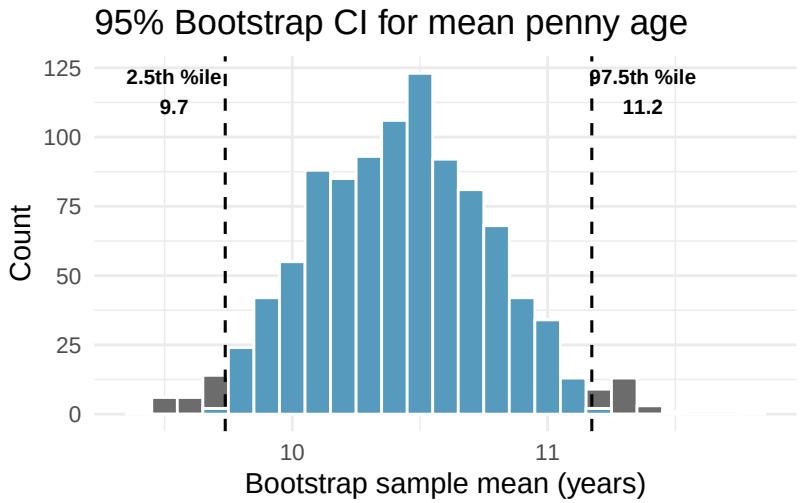


Figure 29: Bootstrap distribution of 1,000 sample means with the middle 95% shaded in blue. The 2.5th and 97.5th percentiles form the 95% confidence interval for the mean penny age.

The 95% bootstrap confidence interval for the mean penny age is approximately $(9.7, 11.2)$ years.

We can also illustrate the concept of a *confidence level* within the boostrapping framework. Imagine we could treat our 648 pennies as if they were the entire population (with true mean $\mu \approx 10.4$ years). We then draw 25 smaller random samples of 50 pennies each and build a bootstrap CI from each one. This lets us check whether each interval captures the "true" mean — something we can never do with real data, because we don't know the true parameter.

Figure 30: Twenty 95% confidence intervals from 20 different random samples of pennies. The true population mean is shown as a vertical dashed line. Approximately 95% of intervals capture the true mean; a few (shown as red and dotted) miss it.

Each horizontal line represents a 95% CI from a different sample. Most intervals (blue and solid) capture the true mean, but a few (red and dashed) miss it entirely. Over many samples, roughly 95% of the intervals succeed. This is the meaning of "95% confidence."

What "95% confident" means: If we repeated the entire study many times — collecting a new sample each time and computing a new 95% confidence interval — approximately 95% of those intervals would contain the true population parameter. Any single interval either contains the parameter or it doesn't; the 95% refers to the long-run success rate of the *procedure*.

Class Example 2.3.8 *Can these be a bootstrap sample?*

We have a random sample of 5 exam grades: 85, 72, 79, 97, 88. For the following state whether or not it is a possible bootstrap sample. If it is not possible, provide a correction to the sample.

a) 79, 79, 97, 85, 82

b) 85, 88, 97, 72

c) 85, 85, 85, 85, 85

The percentile method

The approach used in the pennies example is called the *percentile method* for bootstrap confidence intervals.

For a $(1-\alpha)\times 100\%$ confidence interval using the percentile method:

- **Lower bound** = the $(\alpha/2)\times 100$th percentile of the bootstrap distribution
- **Upper bound** = the $(1-\alpha/2)\times 100$th percentile of the bootstrap distribution

For a 95% CI ($\alpha = 0.05$): use the 2.5th and 97.5th percentiles. For a 90% CI ($\alpha = 0.10$): use the 5th and 95th percentiles. For a 99% CI ($\alpha = 0.01$): use the 0.5th and 99.5th percentiles.

Class Example 2.3.9 *Increasing confidence level*

If you increase the confidence level from 95% to 99%, what happens to the width of the confidence interval? Why?

Class Example 2.3.10 *Average penalty minutes in the NHL*

We have a random sample of the penalty minutes for $n = 26$ players from the 2018–2019 season for the NHL team the Ottawa Senators. Some percentiles from a bootstrap distribution of 1000 sample means are shown in the table.

	0.5%	**1.0%**	**2.0%**	**2.5%**	**5.0%**	**95.0%**	**97.5%**	**98.0%**	**99.0%**	**99.5%**
Percentile	13.8	14.6	15.4	15.8	17.2	33.1	35.1	35.3	36.4	38.0

Selected percentiles from bootstrap distribution

a) What percentiles of the bootstrap distribution would be needed to create a 90% confidence interval?

b) Create a 90% confidence interval for the average penalty minutes for a player on the Ottawa Senators in the 2018–2019 season using the percentile method.

c) What percentiles of the bootstrap distribution would be needed to create a 98% confidence interval?

d) Create a 98% confidence interval for the average penalty minutes for a player on the Ottawa Senators in the 2018–2019 season using the percentile method.

e) Give an interpretation for one of the intervals (90% or 98%).

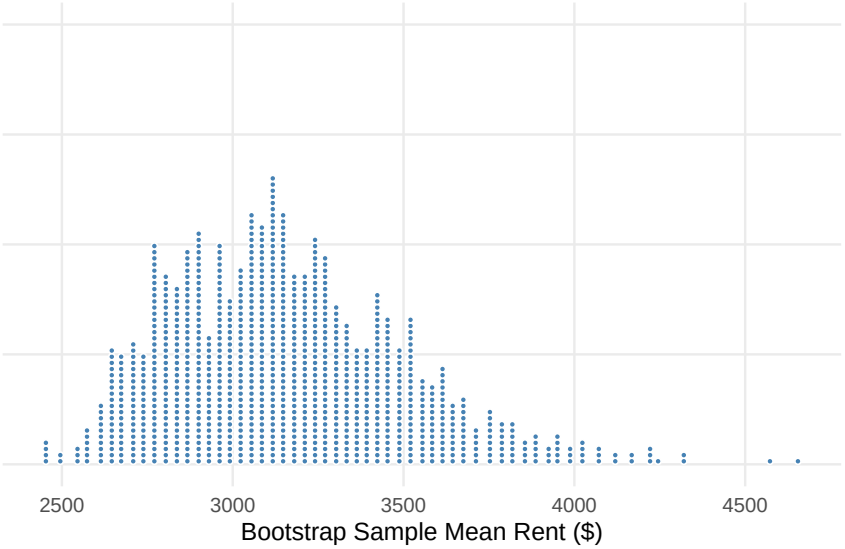
f) Using the 90% interval, is it plausible to say that the Ottawa Senators players were given more than 34 penalty minutes on average per player? Explain.

Class Example 2.3.11 *Manhattan apartment rent*

The New York City housing authority wants to know about the average rent for a one-bedroom apartment in Manhattan. One of their employees goes to Craigslist and records the monthly rent for a randomly selected sample of 8 listed one-bedroom apartments in Manhattan. Her sample is given below

3150 2275 3100 2650 2925 3305 2325 5495

Use the dotplot and the table below to answer the following questions.



	1.0%	**2.5%**	**5.0%**	**10.0%**	**90.0%**	**95.0%**	**97.5%**	**99.0%**
Percentile:	2595.25	2651.85	2720.40	2788.53	3546.80	3679.48	3796.40	3913.68

Selected percentiles from bootstrap distribution

a) Write down one possible bootstrap sample from the original sample.

b) Is it appropriate to use the bootstrap distribution to create a confidence interval for the average rent price in Manhattan? Justify your answer.

c) Regardless of how you answered b), create and interpret a 95% confidence interval for the average rent in Manhattan using the percentile method.

The standard error method

The percentile method reads the confidence interval directly off the two tails of the bootstrap distribution. A second approach, the *standard error method*, instead summarizes the bootstrap distribution by its spread. It is the bridge to the formula-based confidence intervals we develop with the normal and t-distributions in Unit 3.

The *standard deviation* of the bootstrap distribution provides an estimate of the *standard error* for our statistic.

The *bootstrap standard error* (SE) is the standard deviation of the bootstrap distribution. When the bootstrap distribution is roughly symmetric and bell-shaped, an approximate 95% confidence interval is

$$\text{point estimate} \pm 2 \times SE$$

The multiplier 2 comes from the normal model: about 95% of a bell-shaped distribution falls within two standard deviations of its center. (The more precise multiplier is 1.96, which we begin using in Unit 3.)

For the penny ages, the observed sample mean is 10.4 years and the bootstrap standard error, the standard deviation of the 1,000 bootstrap means, is 0.36 years. The standard error 95% confidence interval is

$$10.4 \pm 2(0.36) = (9.7,\ 11.2) \text{ years.}$$

This is essentially the same interval we found with the percentile method, as we should expect, because the bootstrap distribution of the mean penny age is nearly symmetric.

Two methods, usually one answer. For a symmetric, bell-shaped bootstrap distribution, the percentile method and the standard error method give almost identical

intervals. They can disagree when the bootstrap distribution is *skewed*, there the percentile method, which follows the actual shape of the distribution, is the safer choice. The standard error method matters because it connects directly to the formula-based confidence intervals of Unit 3, where the standard error comes from a formula instead of from resampling.

Class Example 2.3.12 *Skateboard Prices*

In a random sample of 20 skateboard prices on eBay, we find the average price is \$67.59. A bootstrap distribution gives a standard error of 10.9.

a) Find and interpret a 95% confidence interval.

b) Should a 90% confidence interval be wider or narrower than this interval?

c) If we took a sample of 50 prices instead of 20, would the 95% confidence interval be wider or narrower than this interval? Explain.

What affects the width of the CI

Three factors control how wide a confidence interval is:

1. *Sample size* (n): Larger samples → narrower intervals (more information about the population)
2. *Variability in the data*: More spread → wider intervals (less precision)
3. *Confidence level*: Higher confidence → wider intervals (more certainty requires a wider net)

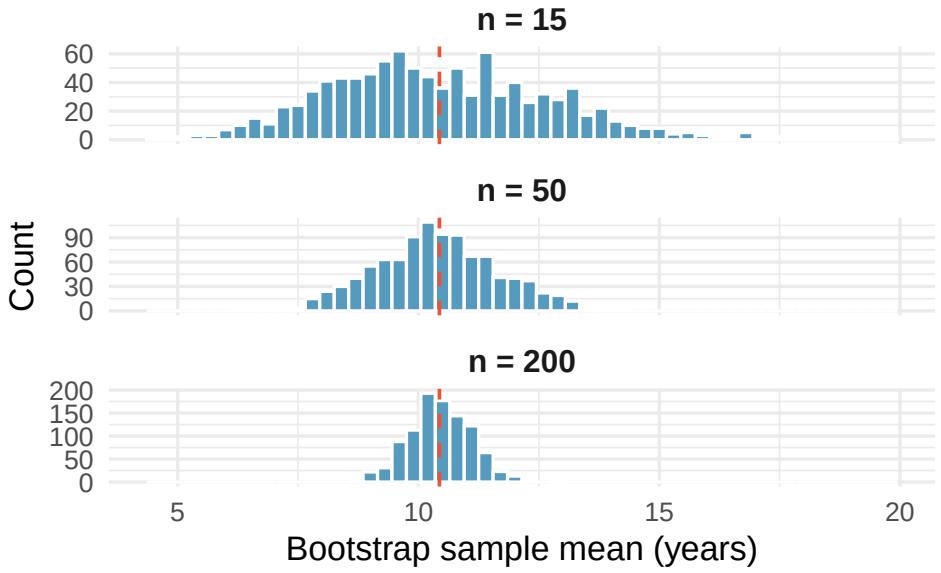


Figure 31: The effect of sample size on the bootstrap distribution. Larger samples produce narrower bootstrap distributions, leading to narrower confidence intervals. All three panels use the same population of penny ages.

As sample size increases, the bootstrap distribution becomes tighter around the sample mean. This is why large studies produce more precise estimates than small ones.

When does bootstrapping work?

Bootstrapping is remarkably flexible, but it does require some conditions to produce reliable intervals:

1. *Independence*: The observations in the sample should be independent of one another. This is usually satisfied when data come from a random sample or a randomized experiment.
2. *Representative sample*: The sample should be reasonably representative of the population. If the sample is biased (e.g., only volunteers), the bootstrap CI will reflect that bias.
3. *Sufficient sample size*: Very small samples (e.g., $n < 10$) may not contain enough information about the population's shape. With small n, the bootstrap distribution can be unreliable.

Bootstrapping does *not* require the population to follow a bell-shaped (normal) distribution. This is one of its greatest strengths compared to traditional formula-based methods, which often do require normality. The penny ages data are clearly right-skewed, yet the bootstrap procedure works perfectly well. The bootstrap lets the data "speak for themselves."

Summary

- The *bootstrap* constructs a confidence interval by resampling with replacement from the original sample.
- The *bootstrap distribution* approximates the sampling distribution of the statistic.
- The *percentile method* uses the middle $(1-\alpha) \times 100\%$ of the bootstrap distribution as the CI.
- The *standard error method* uses point estimate $\pm 2 \times SE$, where the SE is the standard deviation of the bootstrap distribution; it agrees with the percentile method for symmetric distributions and bridges to the formula-based intervals of Unit 3.
- A *95% confidence interval* means that if we repeated the study many times, about 95% of the resulting intervals would capture the true parameter.
- CI width depends on sample size, data variability, and confidence level.
- Bootstrap CIs work for any statistic — means, medians, proportions, differences, and more.
- Bootstrapping requires independent observations from a representative sample with sufficient sample size, but does *not* require the population to be bell-shaped (normal).

2.4 From Simulation to Theory

In the randomization tests and bootstrap confidence intervals chapters, we used simulation to answer two fundamental questions of inference: Is there an effect? (*randomization tests*) *and* How big is the parameter? (*bootstrap confidence intervals*). *Both methods relied on the computer to generate distributions — randomization distributions and bootstrap distributions — that helped us quantify uncertainty. In this chapter, we step back to see the bigger picture. We examine the shape, center, and spread of these simulated distributions and discover a remarkable pattern: under certain conditions, they all look approximately* ***normal*** (*bell-shaped*). *This observation, formalized as the* ***Central Limit Theorem****, is the bridge from the simulation-based methods of Unit 2 to the formula-based methods of Unit 3.*

Key Concepts

- Explain the conclusion of the Central Limit Theorem (CLT)
- Recognize when the CLT can be applied to a study
- Identify when each of the three inference procedures should be used

The Central Limit Theorem

The *Central Limit Theorem* (*CLT*) tells us that for any population distribution (no matter how skewed or strange the population is), if we repeatedly take new random samples from this distribution and calculate the mean or proportion each time, then:

The *Central Limit Theorem (CLT)* stats that when we take sufficiently large random samples from a population, the sampling distribution of many common statistics (such as $\widehat{p}$ or $\overline{x}$) is approximately *normal* (bell-shaped), regarless of the shape of the underlying population distribution

- As sample size increases, the sample average or proportion should get closer to the population average.
- As sample size increases, the averages or proportions will be less spread out.
- As sample size increases, the distribution of the sample averages or proportions will look more like a bell-shaped (normal) curve.

Conditions for the CLT

The bell-shaped approximation does not apply in all situations. Two conditions must hold:

- *Independent observations.* The observations in the sample must be independent of each other. This is typically guaranteed by random sampling from a population, or by random assignment in an experiment.
- *Large enough sample.* The sample size cannot be too small. What counts as "large enough" depends on the context:
  - For proportions: we generally need at least 10 expected successes and 10 expected failures.
  - For means: the rule of thumb depends on how skewed the populations is. We will generally use $n > 30$ for STAT 145.

Class Example 2.4.1 *Visualizing the CLT*

Go to the Sampling Distribution Lab in StatLens. We will use the Right-Skew population shape. We wish to visualize the distribution of the sample mean.

a) Set the sample size to $n = 5$ and draw 1000 samples. What is the shape of the sampling distribution?

b) Set the sample size to $n = 30$ and draw 1000 samples. How does the shape of the sampling distribution change?

c) Set the sample size to $n = 100$ and draw 1000 samples. How does the shape of the sampling distribution change?

d) Which sample size would you use if you wanted the shape of the sampling distribution to be bell-shaped

e) Now, selected a Normal population shape. Set the sample size to $n = 5$ and draw 1000 samples. How does the shape of this sampling distribution compare to the sampling distribution shape from part (a)?

f) Set the sample size to $n = 30$ and draw 1000 samples. How does the shape of this sampling distribution change from part (e). How does it compare to the sampling distribution shape from part (b)?

Comparing randomization, bootstrap, and mathematical approaches

We now have three tools for statistical inference, each grounded in the same core ideas but using different mechanisms:

Table 2: Comparison of inference methods

	Randomization	Bootstrap	Mathematical model
Question answered	Is there an effect? (hypothesis test)	How big is the parameter? (confidence interval)	Both
How variability is generated	Shuffle labels to simulate H_0	Resample with replacement from sample	Use formulas based on the normal distribution
What it models	Variability due to random assignment	Variability due to random sampling	Variability predicted by mathematical theory
Center of distribution	Null hypothesis value	Sample statistic	Depends on context (null values for tests, point estimates for CIs)
Conditions required	Minimal	Minimal	Independence + large sample
Computational cost	Moderate (need many simulations)	Moderate (need many resamples)	Low (just plug into formula)

All three methods answer the same fundamental question: *How much does the statistic vary?* They just measure that variability in different ways.

- *Randomization* asks: "How much would the statistic vary if the null hypothesis were true?"
- *Bootstrap* asks: "How much would the statistic vary across different samples from this population?"
- *Mathematical models* ask: "How much does theory predict the statistic would vary?"

When conditions are met, all three give similar answers. The mathematical approach is faster and requires no simulation, which is why it is widely used. But the simulation approaches are more flexible and work even when conditions for the mathematical model are not met.

When is the normal approximation good enough?

The mathematical approach to inference (which we will develop in Unit 3) replaces simulation with the normal distribution. Instead of generating thousands of shuffles or resamples, we use a formula to compute the standard error and then rely on the bell curve to find p-values or construct confidence intervals.

But when can we trust this shortcut?

The success-failure condition for proportions

For a sample proportion $\hat{p}$, the sampling distribution is approximately normal when:

$$\text{Expected Successes: } np \geq 10 \quad \text{and} \quad \text{Expected Failures: } n(1-p) \geq 10$$

where n is the sample size and p is the proportion of interest. In practice, since p is unknown, we use $\hat{p}$ as a stand-in:

$$n\hat{p} \geq 10 \quad \text{and} \quad n(1-\hat{p}) \geq 10$$

This is called the **success-failure condition**.

The sample size condition for means

For a sample mean $\overline{x}$, the CLT tells us the sampling distribution is approximately normal when:

- The sample size is "large enough," and
- The observations are independent.

A common guideline is $n \geq 30$, but this is only a rough rule of thumb. If the population distribution is nearly symmetric, the normal approximation can work well even for smaller samples. If the population is highly skewed or has extreme outliers, a larger sample may be needed.

Population shape	**Minimum n for normal approximation**
Symmetric, no outliers	$n \geq 15$ or even smaller
Slightly skewed	$n \geq 30$
Highly skewed or heavy-tailed	$n \geq 60$ or more

When in doubt, use simulation (bootstrap) to check whether the normal approximation is reasonable.

Class Example 2.4.2 *Applying the CLT*

For each of the following indicate whether the CLT can be applied. Why or why not?

a) We are interested in approximating the sampling distribution of the sample mean with samples of size 10.

b) We are interested in approximating the sampling distribution of the sample mean with samples of size 100.

c) We are interested in approximating the sampling distribution of the sample proportion when $n = 62$ and $\hat{p} = 0.05$.

d) We are interested in approximating the sampling distribution of the sample proportion when $n = 1000$ and $\hat{p} = 0.5$.

e) We are interested in approximating the sampling distribution of the sample minimum with samples of size 100.

Summary

- Randomization distributions, bootstrap distributions, and theoretical sampling distributions all tend to be approximately *bell-shaped* (normal) under common conditions. This is the **Central Limit Theorem** (CLT).
- The CLT states that for sufficiently large samples with independent observations, the sampling distribution of many common statistics is approximately normal, regardless of the shape of the population.
- The *success-failure condition* ($n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$) is the practical check for whether the normal approximation is appropriate for proportions.
- The three inference approaches — randomization, bootstrap, and mathematical models — answer the same fundamental questions. When conditions are met, they produce similar results. Mathematical models are faster; simulation methods are more flexible.
- This chapter bridges Unit 2 (simulation-based inference) and Unit 3 (formula-based inference). The CLT is the theoretical justification for the formulas we will use in Unit 3.

3.1 Normal Distributions

*When we discussed randomization tests, we used simulation-based methods, randomization distributions and bootstrapping, to make inferences about population parameters. Those methods are powerful and flexible, but they share a limitation: every new problem requires running thousands of simulations. In this chapter, we introduce the **normal distribution**, a mathematical model that describes the bell-shaped pattern we keep seeing in our sampling distributions. When certain conditions are met, the normal distribution lets us calculate probabilities and construct confidence intervals using formulas instead of simulations. This is not a replacement for simulation-based thinking, it is a complement. The normal model provides the theoretical justification for why those simulations work the way they do.*

Key Concepts

- Estimate probabilities as areas under a density curve
- Recognize how the mean and standard deviation represent to the center and spread of a normal distribution
- Use the 68-95-99.7 rule to compute probabilities of intervals for normal distributions
- Use technology to compute probabilities of intervals for normal distributions
- Use technology to find endpoint(s) of intervals with a specified probability for normal distributions
- Convert in either direction between a general normal distribution, denoted $N(\mu, \sigma)$, and a standard normal distribution, denoted $N(0, 1)$

Density curves

A *density curve* is a graphical representation of the distribution for a *continuous* random variable (*i.e.* sample space with probabilities).

Although density curves come in many different shapes, there are two key characteristics that all density curves have:

- The total area under the curve equals 1 (which corresponds to 100% of the distribution).
- The area under the density curve within any interval is the proportion of the distribution in that interval and, thus, represents the probability that a randomly selected value from that distribution falls in that interval.

Normal distribution

The *normal distribution* is a symmetric, bell-shaped distribution described by two parameters: the mean, μ (which determines the center), and the standard deviation, σ (which determines the spread). We write the distribution shorthand as $N(\mu, \sigma)$.

It is not always easy to produce a bootstrap or randomization distribution to answer the statistical question of interest. However, in many circumstances, a theoretical model can be used in place of the randomization distribution. Most of the bootstrap and randomization distributions we have seen in this class have been symmetric and bell-shaped. This type of distribution shape occurs frequently in statistics and is also a characteristic of a family of continuous probability distributions called the *normal distributions*. The normal distribution is density curve that has been well-studied and has a number of very nice properties.

All normal distributions are centered at their mean, μ, and the standard deviation, σ, tells us how spread out the distribution is. We will use the notation $X \sim N(\mu, \sigma)$ to indicate that the random variable X follows a normal distribution with mean, μ, and standard deviation, σ.

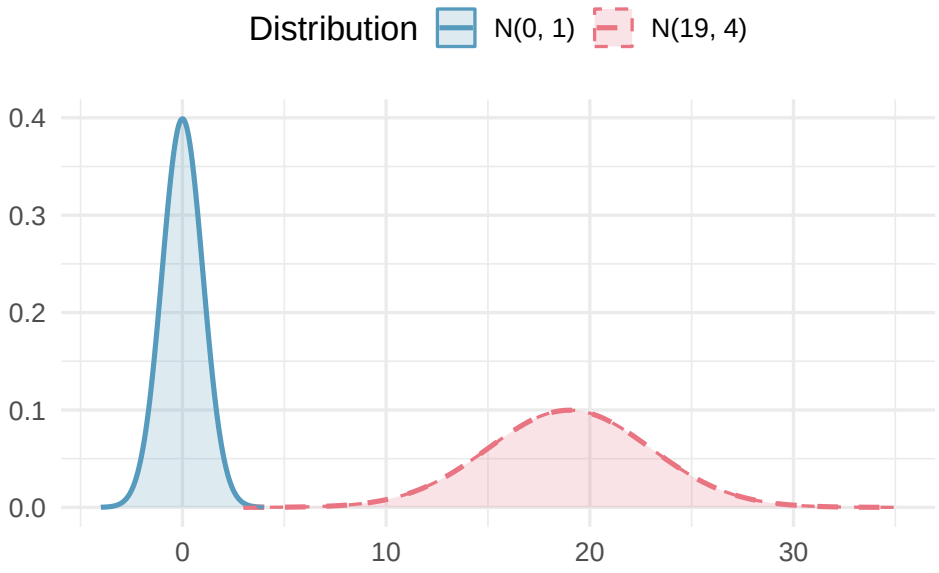


Figure 32: Two normal distributions with different parameters. The left curve has $\mu = 0$ and $\sigma = 1$; the right curve has $\mu = 19$ and $\sigma = 4$. Changing μ shifts the curve left or right; changing σ stretches or compresses it.

The *standard normal distribution* is denoted $N(0, 1)$.

The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ is called the *standard normal distribution*, written $N(0, 1)$. We use the letter Z to represent the standard normal distribution.

Z-scores and standardizing

The *standard score* or *z-score* indicates how many standard deviations a particular value is from the mean.

Often times, we want to compare two different observations relative to their populations. To do so, it is helpful to first standardize them and then compare the *standard score* or *z-score*. The z-score is a measure of relative standing and tells us how many standard deviations the observation is from the mean. We can standardize a random variable X by using the following formula:

$$Z = \frac{X - \mu}{\sigma}.$$

We can unstandardize Z by using the following formula:

$$X = \mu + Z\sigma.$$

A positive z-score indicates the observation is above the mean and a negative z-score indicates below the mean. Z-scores closer to 0 are more common and z-scores further away from 0 are more rare.

Class Example 3.1.1 *Standardizing to compare*

For each of the following, compare the results using their standardized score

a) Benny and Aaron are on the same track team. They have recently competed at a track and field day, and have each won their respective events. Benny won the high jump with a jump of 75 inches, and Aaron won the 100m dash with a time of 12 seconds. Let $J \sim N(65, 2)$ be the height of the jumps and $D \sim N(15, 1)$ be the times of the 100m dash. Who performed better Benny or Aaron?

b) A car manufacturer has a small car that gets 36 mpg and a truck that gets 22 mpg. Let $C \sim N(35, 4)$ be the distribution of mpgs for small cars and $T \sim N(19, 3)$ be the mpgs for trucks. Which vehicle performs better relative to the other vehicles in its class?

The 68-95-99.7 rule (Empirical rule)

One of the most useful facts about the normal distribution is how observations spread out around the mean. This relationship is captured by the *68-95-99.7 rule* (sometimes called the *empirical rule*). The *68-95-99.7 rule* gives for any normal distribution $N(\mu, \sigma)$:

- About **68%** of observations fall within **1 standard deviation** of the mean: between $\mu - \sigma$ and $\mu + \sigma$.
- About **95%** of observations fall within **2 standard deviations** of the mean: between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- About **99.7%** of observations fall within **3 standard deviations** of the mean: between $\mu - 3\sigma$ and $\mu + 3\sigma$.

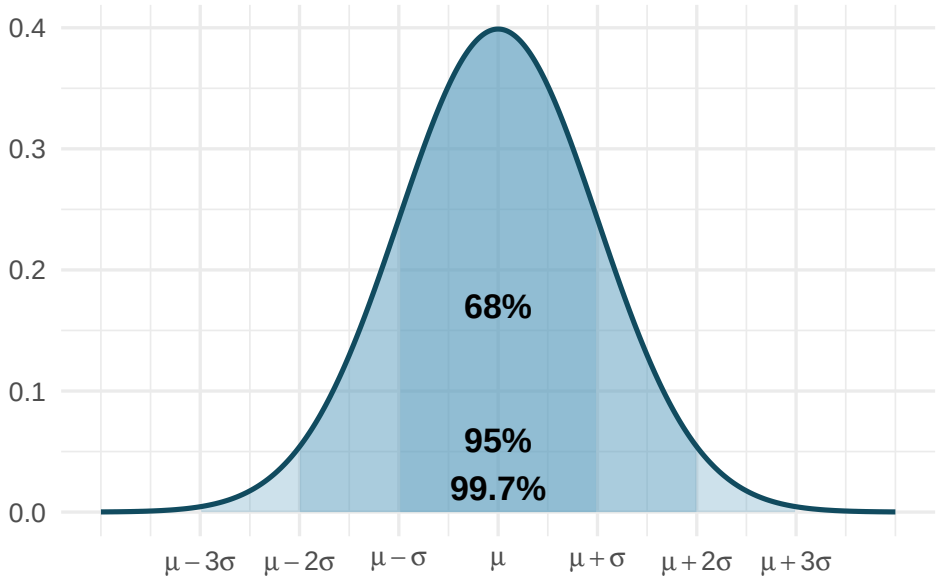


Figure 33: The 68-95-99.7 rule illustrated on a normal curve. The innermost shaded region (68%) extends from $\mu - \sigma$ to $\mu + \sigma$. The middle region (95%) extends from $\mu - 2\sigma$ to $\mu + 2\sigma$. The outermost region (99.7%) extends from $\mu - 3\sigma$ to $\mu + 3\sigma$.

Class Example 3.1.2 *SAT scores*

Let X be the SAT math scores for science and engineering majors, $X \sim N(609, 80)$. Using the 68-95-99.7 rule, answer the following questions.

a) What proportion of SAT scores should be between 529 and 689.

b) What is the probability that a randomly selected student with a science and engineering major has an SAT score of at least 769?

c) What proportion of SAT scores should be between 609 and 769?

Finding probabilities from the normal distribution

To find the probability that a normally distributed variable falls in a specified range, we must find the area under the density curve. You've already been doing this intuitively — in the randomization tests and bootstrap confidence intervals chapters, every time you counted "what fraction of simulated statistics were this extreme?" you were estimating a probability by looking at proportions in a distribution. The normal model just gives us a formula for the curve, so we can compute the area exactly instead of simulating it.

Strategy for normal probability problems. Always follow these steps:

1. *Draw a picture.* Sketch the normal curve, mark the mean, and shade the area you want to find.
2. *Find the Z-score.* Standardize the value(s) of interest using $Z = \frac{x-\mu}{\sigma}$.
3. *Look up the area.* Use a normal probability table, statistical software, or the StatLens Normal Distribution Explorer to find the area to the *left* of the Z-score.
4. *Adjust if needed.* If you need a right-tail area, subtract the left-tail area from 1. For an area between two values, subtract the smaller left-tail area from the larger.

Class Example 3.1.3 *Rainfall in Ithaca*

Let R be the yearly rainfall (in inches) in Ithaca, NY. Data from Cornell's Northeast Regional Climate Center indicates that $R \sim N(35.4, 4.2)$.

a) How much rain should Ithaca expect to get in the year that is the 99^{th} percentile of rainfall?

b) How much rain should Ithaca expect to get in the year that marks the top 20 percent of rainfall?

c) What should the IQR be for yearly rainfall in Ithaca? (*i.e.* the middle 50%)

Class Example 3.1.4 *Furniture assembly*

A company that markets build-it-yourself furniture sells a computer desk that is advertised with the claim "less than an hour to assemble." Let T be the time (in hours) it takes to assemble the furniture. Through post-purchase surveys, it appears that $T \sim N(1.29, 0.43)$.

a) What percentage of people should be able to assemble the desk in less than an hour?

b) What is the probability that a randomly selected person can assemble the desk in between 1 and 1.5 hours?

c) The company would like to change the advertising claim to read "less than _____ to assemble" so that 90% of people can assemble the desk in the time listed. What time should they use (as the cutoff for the bottom 90%)?

The normal approximation to sampling distributions

The *standard error (SE)* of a statistic is the standard deivation of its sampling distribution. It measures how much the statistic varies from sample to sample.

In the sampling distributions chapter, we used simulation to explore *sampling distributions*, the distributions of sample statistics computed from many repeated samples. A striking pattern emerged: regardless of whether we were looking at sample means, sample proportions, or differences in means or proportions, the sampling distributions tended to look *normal* centered at the parameter of interest due to the *Central Limit Thorem*. But how wide is the normal distribution? The spread of the sampling distibution is given by the *standard error*.

The standard error depends on the parameter of interest:

Parameter	**Point estimate**	**Standard error**
Population mean μ	$\bar{x}$	$SE = \frac{s}{\sqrt{n}}$
Population proportion p	$\hat{p}$	$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Difference in means $\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Difference in proportions $p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Standard error formulas for common parameters.

Putting it all together, for the sample mean, $\overline{x}$, and sample proportion $\hat{p}$:

- The sampling distribution of $\bar{x}$ is approximately $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.
- The sampling distribution of $\hat{p}$ is approximately $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

Confidence intervals using the normal model

The normal approximation also lets us build confidence intervals without bootstrapping.

General confidence interval formula. When the sampling distribution of a statistic is approximately normal with standard error SE, a confidence interval takes the form:

$$\text{point estimate} \pm z^* \times SE$$

where z^* is chosen based on the desired confidence level:

Confidence level	z^*
90%	1.645
95%	1.960
99%	2.575

Critical values z^* for common confidence levels.

The distance $z^* \times SE$ is called the *margin of error*. For a 95% confidence interval, $z^* = 1.96$ (often rounded to 2), giving a margin of error of approximately $2 \times SE$. This is directly connected to the 68-95-99.7 rule: 95% of sample statistics fall within about 2 standard errors of the true parameter value.

Class Example 3.1.5 *Stents and stroke risk*

A study examined whether stents reduce stroke risk. The observed difference in 30-day stroke rates (treatment minus control) was $\hat{p}_T - \hat{p}_C = 0.090$, with $SE = 0.028$. Construct a 95% confidence interval.

Summary

This chapter introduced the normal distribution, the most commonly encountered distribution in statistics. Key concepts include:

- The normal distribution is described by two parameters: the mean μ (center) and standard deviation σ (spread), written as $N(\mu, \sigma)$.
- The *68-95-99.7 rule* provides quick estimates: about 68%, 95%, and 99.7% of observations fall within 1, 2, and 3 standard deviations of the mean.
- A *Z-score* standardizes an observation by measuring how many standard deviations it lies from the mean: $Z = \frac{x-\mu}{\sigma}$.
- Normal probabilities are found by computing Z-scores and looking up areas under the standard normal curve using tables, software, or StatLens.
- The *standard error* measures how much a statistic varies from sample to sample and is the key ingredient in confidence intervals and hypothesis tests.

3.2 *Inference for Proportions*

In earlier chapters, we used simulation-based methods to draw conclusions about population proportions. In the randomization tests and bootstrap confidence intervals chapters, we built randomization distributions and bootstrap confidence intervals; in the normal approximation chapter, we learned that the normal distribution provides a mathematical model for these sampling distributions when certain conditions are met. In this chapter, we bring everything together for proportions. We develop normal-based confidence intervals and hypothesis tests for a single proportion and for the difference between two proportions. Throughout, we compare these formula-based results to the simulation-based results from earlier chapters, reinforcing the idea that these are two routes to the same destination.

Key Concepts

- Find the mean and standard error for a distribution of sample proportions and difference in sample proportions
- Identify when a normal distribution is an appropriate model for a distribution of sample proportions
- Recognize when a two-sample z-interval for proportions is the appropriate statistical inference procedure
- State the conditions for the one-sample z-test for a single proportion and two-sample z-test for a difference in population proportions
- Use a normal distribution, when appropriate, to compute a confidence interval for a population proportion and difference in proportions between two populations
- Use a normal distribution, when appropriate, to test a hypothesis about a population proportion and for the difference in population proportions

The sampling distribution of the sample proportion

For categorical data, the goal is often to learn about a population proportion from a sample proportion. We can use the CLT to conduct inference about sample proportions if we have a large enough sample size.

If a sample size is large enough, the CLT says that the distribution of all sample proportions ($\hat{p}$) will follow a normal distribution that is centered at the population proportion (p). We can state these conditions precisely:

- *Independence.* The observations are independent (e.g., from a simple random sample)
- *Success-failure condition.* We expect to see at least 10 successes and 10 failures: $np \geq 10$ and $n(1-p) \geq 10$.

To ensure that the sampling distribution of $\hat{p}$ is not too skewed for the normal approximation to work well, the *success-failure condition* requires that

- $np \geq 10$ and
- $n(1-p) \geq 10$.

We previously discussed estimating the standard error of a statistic using a bootstrap distribution in Unit 2. When we use the CLT, we can estimate the standard error for a sample proportion by

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

where p is the population proportion, and n is the sample size.

Class Example 3.2.1 *Use the CLT to estimate the standard error*

For each of the following, find $SE(\hat{p})$.

a) Samples of size $n = 100$ from a population with $p = 0.5$

b) Samples of size $n = 200$ from a population with $p = 0.5$

c) Samples of size $n = 500$ from a population with $p = 0.5$

d) What relationship did you notice from the answers from parts a) - c)?

So when the conditions are met, the sampling distribution is approximately

$$\hat{p} \sim N\left(p,\ \sqrt{\frac{p(1-p)}{n}}\right).$$

Class Example 3.2.2 *Describe the sampling distribution*

For each of the following questions, identify if you can use the CLT. If so, fully describe the distribution. If not, identify the smallest sample size you need in order to use the CLT.

a) Samples of size $n = 10$ from a population with $p = 0.4$.

$\hat{p} \sim N($_____ , _____$)$ or Smallest n: _____

b) Samples of size $n = 100$ from a population with $p = 0.6$.

$\hat{p} \sim N($_____ , _____$)$ or Smallest n: _____

c) Samples of size $n = 40$ from a population with $p = 0.9$.

$\hat{p} \sim N($_____ , _____$)$ or Smallest n: _____

Class Example 3.2.3 *Distributional Shape*

For each of the scenarios in Example 11.2, go to the Sampling Distribution Lab in StatLens to generate the Sampling Distribution for the Proportion using 3000 samples. What distributional shape is exhibited in each case? Do the cases that satisfy the "large sample" requirements look normal as the CLT suggest they shoul

a) Distributional shape:

b) Distributional shape:

c) Distributional shape:

Since we rarely know the true p:

- *For confidence intervals*, use $\hat{p}$ as the best guess: $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- *For hypothesis tests*, use p_0 (the value from H_0): $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$

Confidence intervals for a single proportion

We can only use the CLT for confidence intervals if - $n\hat{p} \geq 10$ and - $n(1-\hat{p}) \geq 10$

In the previous chapter, we defined the general confidence interval formula using the normal distribution:

$$\text{point estimate} \pm z^* \times SE.$$

Provided that the *success-failure condition* holds, we can substitute the point estimate with $\hat{p}$, our best guess at p, and $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, the spread of the sampling statistic. A confidence interval for a population proportion p can be constructed using

$$\hat{p} \pm z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where z^* is the critical value for the desired confidence level (e.g., $z^* = 1.96$ for 95%). Recall from the previous chapter, we stated common z^* values:

Confidence	z^*
90%	1.645
95%	1.960
99%	2.575

Commonly used z^* values.

We can also use technology, like StatLens or jamovi, or a table to get other z^* values.

Class Example 3.2.4 *Finding z^**

Find the z^* values for the following confidence intervals.

a) 92%:

b) 85%:

c) 80%:

Class Example 3.2.5 *Left-handed desks*

A university would like to estimate the proportion of its students that are left-handed to purchase new desks for a building renovation. They take a random sample of 300 students and find that 32 people in the sample are left-handed.

a) Using the CLT find a 95% confidence interval for the proportion of students that are left-handed.

b) Provide an interpretation for your interval in a).

c) The Vice Chancellor for facilities decides that 15% of the desks they order should be left-handed. Use the confidence interval in a) to argue whether or not this will be an adequate proportion of left-handed desks.

Class Example 3.2.6 *Adopting a child in the US revisited*

A survey of 1000 randomly selected American adults conducted in January 2013 stated that "44% say its too hard to adopt a child in the US."

a) Using the CLT find a 99% confidence interval for the proportion of American adults who think it is too difficult to adopt.

b) Provide an interpretation for your interval in a).

c) Use the confidence interval in a) to argue whether or not a majority of American adults believe that it is too difficult to adopt.

d) How would this confidence interval change if the 44% figure was based on a sample size of 2000 adults instead of 1000?

Determining Sample Size for Estimating a Proportion

In the confidence interval for the population proportion, we have a margin of error of

$$ME = z^* \sqrt{\frac{p(1-p)}{n}}.$$

If we solve this expression for n, we obtain

$$n = \left(\frac{z^*}{ME}\right)^2 \hat{p}(1-\hat{p}).$$

Unfortunately, we are trying to determine the sample size which means that we likely haven't taken a sample yet. This means we don't have a value for $\hat{p}$. There are two options for replacing this value in the equation for n above.

- If we have some idea of what the proportion might be perhaps from taking a preliminary small sample, we can use this value in place of $\hat{p}$.
- If we are not willing to make a reasonable guess, we use a value of 0.5. This is the conservative approach as it will make n the largest it can be with all other values fixed.

Often times the formula for n will not produce a whole number. In these cases, we always *round up* to the next whole number.

Class Example 3.2.7 *Sample size determination 1*

Consider the previous example on adopting a child. Using the 44% figure from the initial poll, how large of a sample size is needed to estimate the proportion of adults that say it is too hard to adopt a child in the US to within $\pm 2\%$ with 99% confidence.

Class Example 3.2.8 *Sample size determination 2*

A company produces plastic bottles. Because millions of bottles are produced each day, they are unable to check every bottle for defects. Instead, they check a sample of bottles each day to estimate the population proportion of defectives. If no preliminary sample is taken, how large a sample of bottles is needed to obtain a 95% confidence interval that estimates the population proportion within $\pm 4\%$?

Hypothesis test for a single proportion

Recall that provided the sample size is large enough, then according to the CLT,

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

We can use this fact to conduct a hypothesis test. In general, if the distribution is normal, then we can comput a p-value using a standard normal curve and a standardized test statisics of the form

$$x = \frac{\text{Sample Statistic} - \text{Null Value}}{SE}.$$

As with the hypothesis tests we discussed in Unit 2, there are a specific set of steps to conduct a hypothesis test using the CLT.

Step 1: Check conditions for the test

Check whether both $np_0 \geq 10$ and $n(1-p_0) \geq 10$ are true.

We can only use the CLT for a hypothesis test if

- $np_0 \geq 10$ and
- $n(1-p_0) \geq 10$

Step 2: State the hypotheses

For a single proportion, the null hypothesis always has the form

$$H_0\colon\ p = p_0$$

where p_0 is a number that represents the proportion claimed or *status quo* value. The alternative hypothesis has one of the following three forms:

$$H_A\colon\ p < p_0$$
$$H_A\colon\ p > p_0$$
$$H_A\colon\ p \neq p_0$$

Step 3: Calculate the test statistic

To conduct the hypothesis test using the CLT, we use a standardized test statistic of

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

which follows a $N(0,1)$ distribution.

Step 4: Find the p-value

To find the p-value, you must use technology to find what proportion of the standard normal distribution is more extreme than the test statistic (in the direction of the alternative hypothesis). If your alternative is two-sided, then you multiply the one-sided p-value by 2.

Step 5: State your decision and interpret in the context of the hypotheses

As with the randomization hypothesis test, you make your decision based upon the p-value from step 3.

- If p-value $< \alpha$
  - Decision: *reject* H_0,
  - Interpretation: We have discernible evidence to suggest that the alternative hypothesis is correct.
- If p-value $\geq \alpha$
  - Decision: *fail to reject* H_0
  - Interpretation: We do not have discernible evidence to suggest that the alternative hypothesis is correct.

You should *always* write your interpretation in the context of the original hypothesis.

How to draw a sketch to illustrate the p-value

Class Example 3.2.9 *Infected Trees*

A ponderosa pine forest in Colorado has a pine beetle infestation. The beetles bore into a tree and carry a fungus that ultimately kills the tree. The forest can be treated and not eliminated if the proportion of infested trees in the forest is less than 20%. In a random sample of 180 trees, it was found that 26 trees were infested. Based on this data and a 5% level of significance, should the forest be treated?

a) Define the relevant parameter(s) and state the null and alternative hypotheses. Be sure to check any relevant conditions.

b) What is the test statistic?

c) Find the p-value, and give a sketch to illustrate your p-value.

d) State your decision and interpret your findings in the context of the study

Class Example 3.2.10 *Cereal example from Unit 2 revisited*

Recall the motivating example in Unit 2 where researchers were interested in whether children showed a preference for the cereal box that had cartoons on it. In the sample of 80 children, 52 chose the box with cartoons. Conduct a hypothesis test at the 5% level (include all steps) to determine if children prefer cereal with cartoon marketing.

Class Example 3.2.11 *Plato's* Republic*: Syllable Patterns*

Prose rhythm is characterized as the occurrence of five-syllable sequences in long passages of text. This characterization may be used to assess the similarity among passages of text and sometimes the identity of authors. Syllables are often categorized as long or short. On analyzing Plato's *Republic*, researchers found that 26.1% of the five-syllable sequences contained two short and three long syllables (Wishart and Leach, *Computer Studies of the Humanities and Verbal Behavior*, Vol.3, pp. 90-99). Suppose that Greek archaeologists have found an ancient manuscript dating back to Plato's time (about 427-347 B.C.E.). A random sample of 317 five-syllable sequences from the newly discovered manuscript showed that 61 contained two short and three long syllables. Do the data indicate that the population proportion of this type of five-syllable sequence is different from the text of Plato's *Republic*? You may either use a 95% confidence interval or hypothesis test (use $\alpha = 0.05$) to support your conclusion.

Class Activity: Taste test and Vulcan Salute group activity

Difference of two proportions

Consider two samples of sizes n_1 and n_2 that come from populations with proportions p_1 and p_2, respectively. It is a property of sample proportions that the distribution of all differences in sample proportions ($\hat{p}_1$and $\hat{p}_2$) will be centered at the true population difference in proportions ($p_1 - p_2$). When the samples are *independent* of each other, the standard error of $\hat{p}_1 - \hat{p}_2$ is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

Using statistical notation, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ for "large" samples can be summarized as

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

As with single proportions, the population proportions (p_1 and p_2) are rarely ever known (which is why we are estimating them in the first place!), so they are estimated by the corresponding sample proportions $\hat{p}_1$ and $\hat{p}_2$. According to the Central Limit Theorem, when the sample sizes are large enough the sampling distribution of $\hat{p}_1 - \hat{p}_2$ follows a normal distribution.

Class Example 3.2.12 *Proportion of Elderly, Florida and Texas*

From the 2010 US Census, we learn that 17.6% of the residents of Florida are 65 years old or older while 10.5% of the residents of Texas are 65 years old or older. If independent, random samples of 300 Florida residents and 200 Texas residents are selected and the difference in the proportion of elderly residents ($p_{FL} - p_{TX}$) is calculated:

a) Describe the shape of the sampling distribution of differences in sample proportions.

b) What is the mean of the distribution of differences in sample proportions.

c) Find the standard error of the distribution of differences in sample proportions.

In order to be able to use the standard normal distribution to construct a CI for a difference in proportions, the following conditions must be satisfied.

- The two samples must be randomly selected.
- The two samples must be independent of each other.
- The success-failure condition must be met separately for each group; $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$, $n_2\hat{p}_2 \geq 10$, and $n_2(1-\hat{p}_2) \geq 10$ is a good guideline to follow.

Confidence intervals for a difference in proportions

Provided that the conditions are met, the difference between two unknown population proportions can be estimated with a confidence interval as

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where z^* is the appropriate critical value for the desired level of confidence from a $N(0,1)$ distribution.

Class Example 3.2.13 *Mobile Connections to Libraries*

In a random sample of 2,252 Americans age 16 and older, 11% of the 1,059 men and 16% of the 1,193 women said they have accessed library services via a mobile device.

a) What must be *assumed* in order to construct a confidence interval for the difference in proportions?

b) Find and interpret a 95% confidence interval for the difference in proportions accessing libraries via mobile devices, between men and women.

c) Does your interval in (b) suggest that there is a difference in the proportions of men and women accessing libraries via mobile devices? Which population has a higher proportion?

Class Example 3.2.14 *Seat belt use*

We are interested in comparing the proportions of high school and college students in Wisconsin who always wear seat belts. In a random sample of 712 high school students, 615 say they always use seat belts. In a random sample of 641 college students, 578 always use seat belts. Compute a 95% confidence interval for the difference in the population proportions of high school and college students in WI who report always wearing their seat belt. Does this interval suggest that the population proportion is higher for either group?

Hypothesis test for a difference in proportions

Two unknown population proportions can be compared using a hypothesis test by following the steps below.

Step 1: Check the conditions

- The two samples must be randomly selected,
- The two samples must be independent of each other,
- The success-failure condition for each group; $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$, $n_2\hat{p}_2 \geq 10$, and $n_2(1-\hat{p}_2) \geq 10$.

Step 2: State the hypotheses

$H_0 : p_1 - p_2 = 0$ is equivalent to writing $H_0 : p_1 = p_2$.

For a difference in averages, the null hypothesis always has the form

$$H_0\colon\ p_1 - p_2 = 0$$

The alternative hypothesis has one of the following three forms:

$$H_A\colon\ p_1 - p_2 < 0$$
$$H_A\colon\ p_1 - p_2 > 0$$
$$H_A\colon\ p_1 - p_2 \neq 0$$

Step 3: Calculate the test statistic

To conduct the hypothesis test, a standardized test statistic of

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

where

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}.$$

$\hat{p}$ represents the overall proportion of successes under the presumption that H_0 is true. Since $p_1 - p_2 = 0$ implies that $p_1 = p_2$, then the two independent random samples could be seen as coming from *one* sample of size $n_1 + n_2$. Furthermore, x_1 and x_2 denote the observed number of successes in samples 1 and 2, respectively. The sampling distribution of this test statistic is a standard normal distribution for large samples.

Step 4: Find the p-value

To determine the p-value, we find what proportion of the standard normal distribution is more extreme than the test statistic (in the direction of the alternative hypothesis). If your alternative is two-sided, then you multiply the one-sided p-value by 2.

Step 5: State your conclusion and interpret in the context of the hypotheses

As with the randomization hypothesis test, you make your decision based on the P-value from step 4.

- If p-value $< \alpha$
  - Conclusion: *reject* H_0,
  - Interpretation: There is discernible evidence to suggest that . . . (the alternative hypothesis is true).
- If p-value $\geq \alpha$
  - Conclusion: *fail to reject* H_0
  - Interpretation: There is not discernible evidence to suggest that . . . (the alternative hypothesis is true).

Always write the interpretation in the context of the original hypothesis.

Class Example 3.2.15 *Accuracy of Lie Detectors*

Participants in a study to evaluate the accuracy of lie detectors were divided into two groups, with one group reading true material and the other group reading false material, while connected to a lie detector. The two way table indicates whether the participants were lying or telling the truth and also whether the lie detector indicated they were lying or not. We are interested in determining if there is a difference in the proportion of times the lie detector says the person is lying, depending on whether the person is lying or telling the truth.

	Detector says lying	**Detector says not**	**Total**
Person lying	31	17	48
Person not lying	27	21	48
Total	58	38	96

a) Are the conditions met for the two-sample test for proportions? Verify your answer.

b) Find the three sample proportions for the proportion of times the lie detector says the person is lying (the proportion for the lying people, the proportion for the truthful people, and the pooled proportion)

c) Test to see if there is a difference in the proportion of times the lie detector says the person is lying, depending on whether the person is lying or telling the truth. Use a 5% level of significance and show all details of the test.

Class Example 3.2.16 *Tagging Penguins*

A study was conducted to see if tagging penguins with metal tags harms them. In the study, 100 penguins were randomly assigned to receive a metal tag or (as a control group) an electronic tag. One of the variables studied is survival rate ten years after the penguins were tagged. The scientists observed that 20% of the 50 metal tagged penguins survived while 36% of the 50 electronically tagged penguins survived.

a) Are the conditions met for using the normal distribution? Verify your answer.

b) Define the two population proportions p_1 and p_2.

c) Test at $\alpha = 0.05$ to see if the survival rate is lower for all metal tagged penguins than for all electronically tagged penguins. Do metal tags appear to reduce survival rate in penguins?

Class Activity: Orange Candies group activity

Summary

This chapter developed normal-based inference methods for proportions:

- *Success-failure condition.* The normal approximation for $\hat{p}$ requires at least 10 expected successes and 10 expected failures. For confidence intervals, check with $\hat{p}$; for hypothesis tests, check with p_0.
- *CI for one proportion*: $\hat{p} \pm z^* \times \sqrt{\hat{p}(1-\hat{p})/n}$. Use $\hat{p}$ in the SE.
- *Hypothesis test for one proportion*: $Z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$. Use p_0 in the SE.
- *CI for difference of proportions*: $(\hat{p}_1 - \hat{p}_2) \pm z^* \times SE$, where SE uses separate sample proportions.
- *Hypothesis test for difference of proportions*: $Z = \frac{\hat{p}_1-\hat{p}_2}{SE}$, where SE uses $SE_{\hat{p}_1-\hat{p}_2}$.
- *Confidence level* is controlled by z^*: higher confidence means wider intervals.

3.3 Confidence Intervals for Means

In the bootstrap confidence intervals chapter, we used bootstrapping to construct confidence intervals: a flexible, simulation-based approach that works well in many settings. In the normal approximation chapter, we saw that sampling distributions are often well approximated by the normal distribution. For means, however, we encounter a complication: the population standard deviation σ is almost never known, and estimating it with the sample standard deviation s introduces extra uncertainty. To account for this, we use a slightly different bell-shaped distribution called the ***t-distribution****. In this chapter, we develop t-based confidence intervals for one mean, the difference between two independent means, and the mean of paired differences.*

Key Concepts

- Find the mean and standard error of sample means
- Distinguish between separate samples and paried data when comparing two means
- Recognize when a t-distribution is an appropriate model for a distribution of sample means, difference in means between two populations, and the mean of paired differences.
- Use a t-distribution, when appropriate, to compute a confidence interval for a population mean, difference in means between two populations, and the mean of paired differences.

The t-distribution

Earlier we used the normal distribution to model sampling distributions, and in the previous chapter we applied it to proportions, where the standard error formula $SE = \sqrt{\hat{p}(1-\hat{p})/n}$ depends only on the sample proportion and the sample size: both of which are known. For quantitative data, the goal is often to learn about a population average from a sample average. If sample size is large enough, the CLT says that the distribution of all sample averages $(\overline{x})$ will follow a normal distribution that is centered at the population average (μ). In Unit 2, we discussed estimating the standard error of a statistic using a bootstrap distribution. When we use the CLT, we can estimate the standard error for a sample average by

$$SE_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation, and n is the sample size.

We estimate the standard error of $\overline{x}$ with $\frac{s}{\sqrt{n}}$

For averages, we almost never know the population standard deviation (σ), so we estimate it by using the sample standard deviation (s). When we use s instead of σ, $\overline{x}$ no longer follows a normal distribution. Instead, $\overline{x}$ follows a t-distribution.

Back in the early 1900s, a man named William Sealy Gosset was working for the Guinness brewery. During his career at Guinness, he was using statistics to select the best yielding varieties of barley for brewing. At the time Guinness had forbidden any of its employees from publishing in scientific journals. So, Gosset published his findings, a new statistical distribution called the t-distribution under the name Student. The t-distribution is based on a new parameter called the degrees of freedom (d.f.).

The *t-distribution* is a bell-shaped, symmetric distribution that is similar to the normal distribution but has *heavier tails* (more probability in the extremes). The exact shape of the t-distribution depends on a parameter called the *degrees of freedom* (*df*).

- When df is small, the t-distribution has noticeably heavier tails than the normal (extreme values are more likely).
- As df increases, the t-distribution approaches the standard normal distribution.

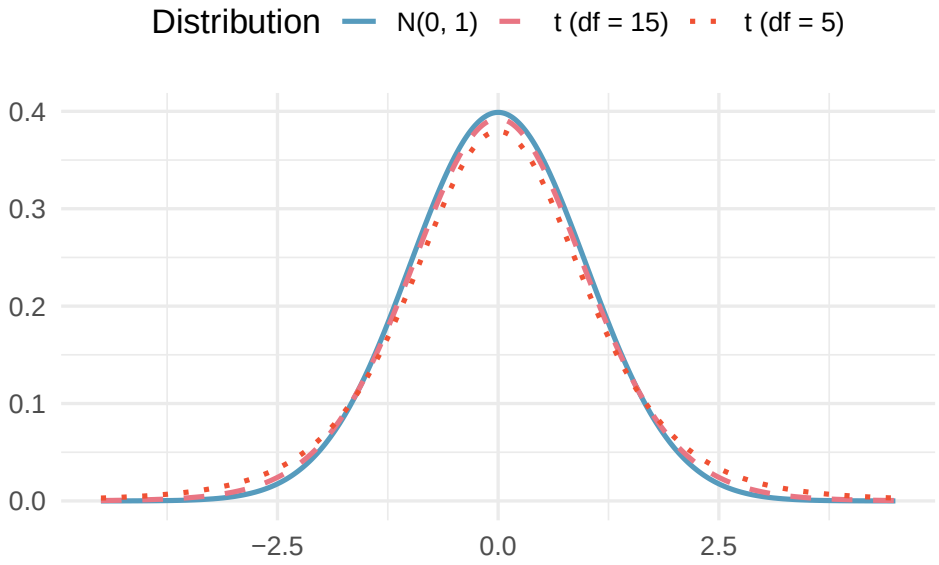


Figure 34: Three distributions plotted on the same axes: the standard normal distribution $N(0, 1)$, a t-distribution with $df = 5$, and a t-distribution with $df = 15$. All are centered at 0 and bell-shaped, but the t-distributions have heavier tails, especially when df is small.

What Gosset found is that provided we have met one of two conditions, then

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} \sim t_{n-1},$$

or t follows a t-distribution with $n - 1$ degrees of freedom. There are two possible ways you can justify using the t-distribution for averages.

- For a large enough sample size ($n \geq 30$), you can use the t-distribution
- If $n < 30$, then the t-distribution is only a good approximation if the population is approximately normal.

To use the t-distribution, you need either

- $n \geq 30$ or
- a population that is approximately normal

Class Example 3.3.1 *Describe the sampling distribution*

For each of the following questions, identify whether or not a t-distribution would be appropriate. If so, identify the number of degrees of freedom that you would use for the t-distribution.

a) Samples of size $n = 10$, from a population that is approximately normal.

b) Samples of size $n = 15$, from a population that is heavily skewed to the right.

c) Samples of size $n = 50$, from a population that is heavily skewed to the left.

Just as we use z^* for the normal distribution, we use t^*_{df} for the t-distribution. Because the t-distribution has heavier tails, t^* values are larger than the corresponding z^* values (especially for small df).

Selected critical values for 95% confidence intervals:

df	t^*_{df} **(95% CI)**	z^* **(for comparison)**
5	2.571	1.960
10	2.228	1.960
30	2.042	1.960
100	1.984	1.960
∞	1.960	1.960

Critical values t^*_{df} for a 95% confidence interval. As df increases, t^* approaches $z^* = 1.96$.

Generally, to find the appropriate t^* for the confidence interval for a population average, we need to use technology to find the corresponding endpoints for the confidence interval.

Confidence intervals for a single mean

We can only use the t-distribution for confidence intervals if

- $n \geq 30$ or
- the population is approximately normal

Provided that the sample size is large enough ($n \geq 30$) or the population is approximately normal, a confidence interval for a population average μcan be constructed using

$$\overline{x} \pm t^* \frac{s}{\sqrt{n}}$$

where $\overline{x}$ is the sample average and t^* is the appropriate critical value from a t_{n-1} distribution.

Class Example 3.3.2 *Triathlon Heart Rates*

The article "Cardiovascular and Thermal Response of Triathlon Performance" reports on a research study involving nine male triathletes. Maximum heart rate (beats/min) was recorded during performace of each of the three events. For swimming, the recorded values were 178, 182, 184, 184, 188, 190, 192, 196, 198.

a) What is the variable in this study? Is it categorical or quantitative?

b) What requirement is needed to make it appropriate to use the t confidence interval for this data?

c) Compute a 90% confidence interval for the mean heart rate of triathletes when swimming.

d) Provide an interpretation for your interval in c).

e) If we had calculated a 99% confidence interval instead of a 90% confidence interval, would it have been larger or smaller than the 90% confidence interval? Explain.

Class Example 3.3.3 *Dairy cows*

A study of 66 dairy cows found that the mean milk yield was 12.5 kg per milking with a standard deviation of 4.3 kg per milking.

a) Compute a 95% confidence interval for the average milk yield in the population. Make sure to check your conditions.

b) Provide an interpretation for your interval in a).

c) Is it plausible that the population mean milk yield is 11kg per milking?

d) If the mean and standard deviation were based off a sample fo 660 cows instead of 66, how would the 95% interval change?

Confidence intervals for a difference in means

We often want to learn about a difference in averages from two groups. We can use the CLT to conduct inference about differences in averages if we have a large enough sample size. Consider two groups n_1 and n_2 that come from populations with means μ_1 and μ_2 and standard deviations σ_1 and σ_2, respectively. If sample size is large enough, the CLT says that the distribution of all differences in sample averages $(\overline{x}_1 - \overline{x}_2)$ will follow a normal distribution that is centered at the true population difference in averages $(\mu_1 - \mu_2)$. When we use the CLT, we can estimate the standard error for a sample average by

$$SE_{\overline{x}_1 - \overline{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

As with single averages, we almost never know the population standard deviations (σ_1 and σ_2), so we estimate them by using the sample standard deviations (s_1 and s_2).

We estimate the standard error of $\overline{x}_1 - \overline{x}_2$ with $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

When we use s_1 and s_2 instead of σ_1 and σ_2, $\overline{x}_1 - \overline{x}_2$ no longer follows a normal distribution. The difference follows a t-distribution and the degrees of freedom are approximately equal to $n_1 + n_2 - 2$.

In order to be able to use this distribution for a difference in averages, you need to make sure that each group satisfies the conditions for an average. Recall from earlier in this chapter:

- For a large enough sample size ($n_1 \geq 30$ and $n_2 \geq 30$), you can use the t-distribution
- If $n < 30$, then the t-distribution is only a good approximation if the population is approximately normal.

Class Example 3.3.4 *Describe the sampling distribution*

For each of the following questions, identify whether or not a t-distribution would be appropriate.

a) Comparing averages from two samples of sizes $n_1 = 10$ and $n_2 = 15$, from populations that are approximately normal.

b) Comparing averages from two samples of sizes $n_1 = 15$ and $n_2 = 40$, where population 1 is heavily skewed right, and population 2 is approximately normal.

c) Comparing averages from two samples of sizes $n_1 = 50$ and $n_2 = 50$, from two populations that are heavily skewed to the left.

Remember, each sample must satisfy the necessary conditions for averages.

Provided that for each group the sample size is large enough ($n \geq 30$) or the population is approximately normal, a confidence interval for a difference in population averages $\mu_1 - \mu_2$ can be constructed using

$$(\overline{x}_1 - \overline{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $\overline{x}_1 - \overline{x}_2$ is the difference in sample averages and t^* is the appropriate confidence multiplier.

Class Example 3.3.5 *Highball vs tumbler*

Two common types of glassware at a bar are "highballs" (tall and thin) and "tumblers"(short and wide). Researchers randomly assigned participants either a "highball" glass or a "tumbler," each of which held 355 ml. Participants were asked to pour a shot (1.5 oz = 44.3 ml) into their glass. The summary statistics are given in the table below.

	n	$\overline{x}$	s
Highball	99	42.2 ml	16.2 ml
Tumbler	99	60.9 ml	17.9 ml

Summary statistics for amount poured

We want to know about the difference in the average amount of liquid poured into the two types of glasses.

a) What are the variables of interest in this study? Are they categorical or quantitative? Which is the explanatory variable and which is the response?

b) What is the parameter of interest?

c) Find and interpret a 95% confidence interval estimate for the difference in the average amount poured in the two different types of glasses. Use highball for group 1 and tumbler for group 2.

d) If you were the owner of a local bar, from a profit perspective, do you think there is any reason to prefer the highball to the tumbler glass?

Class Example 3.3.6 *Caffeine amounts*

A consumer agency wanted to estimate the difference in the amount of caffeine in two brands of coffee. The agency took a random sample of 15 one-pound jars of Brand I coffee, yielding a mean amount of caffeine of 80mg/jar with a standard deviation of 5mg. Another random sample of 12 jars of Brand II gave a eman of 77mg/jar with a standard deviation of 6mg. Find and interpret a 90% confidence interval for the difference in the population mean caffeine amounts. What does this interval tell us about the difference in caffiene for the two brands? Are there any conditions that must be satistified for the interval to be valid?

Confidence intervals for an average difference

When examining differences in averages, typically, we use one of two methods. One option is to use the tools we discussed for the difference in averages. However, if there is a reason that the groups are not separate groups, then there is a better technique to use. If there is some kind of natural pairing between subjects in a study, then we can use techniques based on the concept of *paired differences*. Most often, paired differences occur when the subjects of the study receive two different levels of a factor (most commonly a *control* and a *treatment*). However, another way this appears is in a study in which two participants have been paired based on a number of different characteristics.

Class Example 3.3.7 *Deciding between difference in averages and paired difference*

For each of the following, identify the cases for the study and decide if it would be more appropriate to use a difference in averages or paired difference.

a) In a medical study to learn about the difference in the efficacy of generic versus brand name blood pressure medications, 50 individuals under 30 years old and 50 indiviudals over 30 years old were randomly assigned to receive either a generic medication or a brand name medication.

b) Researchers want to learn about the average age differences between partners who have been together more than 10 years. They randomly select 50 sets of partners and ask them their age.

c) In a study on the average difference in calories between servings of strawberry and vanilla yogurt, a food scientist randomly selected 12 brands of yogurt and recorded the calories in their strawberry and vanilla yogurts.

d) The Office of Greek Life randomly selects 45 fraternity and sorority members and looks to see if there is a difference in the average GPAs for fraternities vs sororities.

In order to be able to use the t-distribution for a difference in averages, you need to be sure that the differences, $d_i = x_{1,i} - x_{2,i}$, for each pair satisfy the conditions for an average. Recall from earlier in this chapter:

1. For a large enough sample size ($n_d \geq 30$), you can use the t-distribution
2. If $n_d < 30$, then the t-distribution is only a good approximation if the population is approximately normal.

In this section, we will use a subscript d to indicate that we are working with differences.

We either need

- $n_d \geq 30$ or
- the population of differences to be approximately normal

Provided that the conditions are met, a confidence interval for the average difference in the population μ_d can be constructed using

$$\overline{x}_d \pm t^* \frac{s_d}{\sqrt{n_d}}$$

where $\overline{x}_d$ is average of the differences in the sample and t^* is the appropriate critical value from a t-distribution with $n_d - 1$ degrees of freedom.

Class Example 3.3.8 *Book prices*

You are given the task of determining if books differ in cost through Amazon.com or a local bookstore. The other student gathers data by selecting a random sample of 8 books from Amazon.com and a second independent random sample of 8 books from the local bookstore. The data are given below along with summary statistics.

Source									**Mean**	**SD**
Amazon	53	81	14	14	22	85	88	38	$\overline{x}_1 = 49.375$	$s_1 = 31.972$
Bookstore	45	76	43	64	21	94	47	25	$\overline{x}_2 = 51.875$	$s_2 = 24.868$

Summary statistics for book prices

a) Compute a 95% confidence interval based on how the other student collected data. What does this interval tell you about the difference in average price between Amazon and the local bookstore?

b) How might you have conducted this study instead? Why would a different study design be advantageous here?

c) Suppose that the data you collect from the different study design is given below. Based on the data that you collected, compute a 95% confidence interval for the difference in the cost of books for the two places? Does your conclusion differ from part a)? What accounted for this difference? (Hint: look at the SE for both cases)

Source								
Amazon	53	81	14	14	22	85	88	38
Bookstore	45	76	43	64	21	94	47	25

Summary statistics for book prices

Summary

This chapter developed formula-based confidence intervals for means using the t-distribution.

- The *t-distribution* accounts for the extra uncertainty from estimating σ with s. It has heavier tails than the normal distribution, especially for small degrees of freedom.
- *One-sample t CI*: $\bar{x} \pm t^*_{n-1} \times \frac{s}{\sqrt{n}}$. Requires independent observations and approximate normality (or large n).
- *Two-sample t CI*: $(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. Requires independence within and between groups, and approximate normality or large samples in each group.
- *Paired t CI*: Compute differences for each pair, then apply the one-sample t CI to the differences: $\bar{x}_d \pm t^*_{n_d-1} \times \frac{s_d}{\sqrt{n_d}}$.
- Always check conditions before using t-methods: independence and normality/sample size.

3.4 Hypothesis Tests for Means

In the confidence intervals for means chapter, we introduced the t-distribution as a mathematical model for the sampling distribution of means. In this chapter, we use the t-distribution to conduct hypothesis tests for a single mean, the difference between two independent means, and the mean of paired differences.

Key Concepts

- Distinguish between separate samples and paried data when comparing two means
- Recognize when a t-distribution is an appropriate model for a distribution of sample means, difference in means between two populations, and the mean of paired differences.
- Use a t-distribution, when appropriate, to test a hypothesis for a population mean, difference in means between two populations, and the mean of paired differences.

The t-test framework

As a reminder from testing a single and the difference of two proportions, every hypothesis test follows the same five-step structure whether we use simulation of a mathematical model:

1. *Check conditions for the test.* Verify that the mathematical model is appropriate.
2. *State the hypotheses.* Write out H_0 (null hypothesis) and H_A (alternative hypothesis).
3. *Compute the test statistic,* Calculate how far the observed data are from what H_0 predicts. For hypothesis tests for means, we will use the t-test statistic.
4. *Find the p-value.* Determine how unusual the test statistic is, assuming H_0 is true.
5. *Draw a conclusion.* Compare the p-value to the discernibility level, α, and state and interpret the conclusion in context.

Hypothesis test for a single mean

Remember, t_{n-1} means a t-distribution with $n-1$ degrees of freedom

Provided that the sample size is large enough ($n \geq 30$) or the population is approximately normal, then according to the CLT

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$

We can use this fact to conduct a hypothesis test. As with the hypothesis test we have discussed thus far, there are a specific set of steps to conduct a hypothesis test using the CLT.

Step 1: Check conditions for the test

Check whether either $n \geq 30$ or the population is approximately normal.

We can only use the t-distribution for confidence intervals if

- $n \geq 30$ or
- the population is approximately normal.

Step 2: State the hypotheses

For a single average, the null hypothesis always has the form

$$H_0 : \mu = \mu_0$$

where μ_0 is a number that represents the averages claimed or *status quo* value. The alternative hypothesis has one of the following three forms:

$$\begin{aligned} H_A&: \ \mu < \mu_0 \\ H_A&: \ \mu > \mu_0 \\ H_A&: \ \mu \neq \mu_0 \end{aligned}$$

Step 3: Calculate the test statistic

To conduct the hypothesis test using the CLT, we use a standardized test statistic of

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

which follows a t_{n-1} distribution.

Step 4: Find the p-value

To find the p-value, you must use technology to find what proportion of the t-distribution is more extreme than the test statistic (in the direction of the alternative hypothesis). If your alternative is two-sided, then you multiply the one-sided p-value by 2.

Step 5: State your decision and interpret in the context of the hypotheses

- If p-value $< \alpha$, we *reject* H_0 and have discernible evidence to suggest that H_A is correct.
- If p-value $\geq \alpha$, we *fail to reject* H_0 and do not have discernible evidence to suggest that the alternative hypothesis is correct.

You should *always* write your interpretation in the context of the original hypothesis.

How to draw a sketch to illustrate the p-value

Class Example 3.4.1 *How much is left?!*

Have you ever been frustrated because you could not get a container of some sort to release the last bit of its contents? The article "Shake, Rattle, and Squeeze: How Much is Left in That Container?" (*Consumer Reports*, May 2009: 8) reported on an investigation of this issue for various consumer products. In a random sample of five 6.0 ounce tubes of toothpaste from a particular brand and squeeze them until no more toothpaste will come out. Then each tube is cut open and the amount remaining is weighed. For this sample, the mean and standard deviation of the amount of toothpaste remaining (in ounces) was 0.502 and 0.102, respectively. Suppose the toothpaste manufacturer claims that it's packaging allows at least 90% of its contents to be released. Conduct a hypothesis test (using $\alpha = 0.05$) to test the manufacturer's claim. You may assume that the remaining amounts of toothpaste are approximately normally distributed.

a) Define the relevant parameter(s) and state the null and alternative hypotheses. Be sure to check any relevant conditions.

b) What is the test statistic.

c) Find the p-value and give a sketch to illustrate your p-value.

d) State your decision and interpret your findings in the context of the study.

Class Example 3.4.2 *Ski Patrol: Avalanches*

Snow avalanches can be a real problem for travelers in the western United States and Canada. A very common type of avalanche is called the slab avalanche. These have been studied extensively by David McClung, a professor of civil engineering at the University of British Columbia. Slab avalanches studied in Canada had an average thickness of 67 (in cm) (*Avalanche Handbook*, by David McClung and P. Schaerer.) The ski patrol at Vail, Colorado, is studying slab avalanches in their region. A random sample of avalanches in spring gave the following thicknesses (in cm).

59	**51**	**76**	**38**	**65**	**54**	**49**	**62**	**68**	**55**	**64**	**67**	**63**	**74**	**65**	**79**

Assuming that slab thickness has an approximately normal distribution, is there evidence that the slab thickness in the Vail region differs from that in Canada? You may either us a 95% confidence interval or a hypothesis test (use $\alpha = 0.05$) to support your conclusion.

Hypothesis test for a difference in means from independent samples

Each sample must satisfy the necessary conditions for averages.

Provided that for each group the sample size is large enough ($n \geq 30$) or the population is approximately normal, then according to the CLT

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

follows a t-distribution with d.f. equal to $n_1 - n_2 - 2$.

We can use this fact to conduct a hypothesis test.

Step 1: Check conditions for the test

Check that for *both* groups $n \geq 30$ or the population is approximately normal.

Step 2: State the hypotheses

For a difference in averages, the null hypothesis always has the form

$$H_0 : \mu_1 - \mu_2 = 0$$

The alternative hypothesis has one of the following three forms:

$$H_A\colon\ \mu_1 - \mu_2 < 0$$
$$H_A\colon\ \mu_1 - \mu_2 > 0$$
$$H_A\colon\ \mu_1 - \mu_2 \neq 0$$

Step 3: Calculate the test statistic

To conduct the hypothesis test using the CLT, we use a standardized test statistic of

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

which follows a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Step 4: Find the p-value and give a sketch to illustrate the p-value

To find the p-value, you must use technology to find what proportion of the standard normal distribution is more extreme than the test statistic (in the direction of the alternative hypothesis). If your alternative is two-sided, then you multiply the one-sided p-value by 2. The sketch that you give should be the same as the sketch for a single mean; however, you do not need to give the degrees of freedom.

Step 5: State your decision and interpret in the context of the hypotheses

- If p-value $< \alpha$ we *reject* H_0 and have discernible evidence to suggest that the alternative hypothesis is correct.
- If p-value $\geq \alpha$, we *fail to reject* H_0 and do not have discernible evidence to suggest that the alternative hypothesis is correct.

You should *always* write your interpretation in the context of the original hypothesis.

Class Example 3.4.3 *Tire lifetimes*

A large automobile manufacturing company is trying to decide whether to purchase Goodyear or Firestone tires for its new models. They believe that Firestone tires might have a longer lifetime. To help arrive at a decision, an experiment is conducted where tires are run until they wear out. A random sample of 55 Goodyear tires gave a mean lifetime of 40,900 miles with a standard deviation of 3,200 miles. A random sample of 40 Firestone tires gave a mean lifetime of 42,800 miles and a standard deviation of 2,700 miles. Using a 1% discernibility level, determine if Firestone tires have a larger population mean lifetime.

a) Define the relevant parameter(s) and state the null and alternative hypotheses. Be sure to check any relevant conditions.

b) What is the test statistic.

c) Find the p-value and give a sketch to illustrate your p-value.

d) State your decision and interpret your findings in the context of the study.

Class Example 3.4.4 *Caffeine amounts revisited*

The summary statistics from the amount of caffeine in coffee example are given in the table below.

	n	$\overline{x}$	s
Brand I	15	80mg	5mg
Brand II	12	77mg	6mg

Summary statistics for the amoung of caffeine in one-pound of coffee for two brands.

Conduct a hypothesis test to see if there is a difference between the average amount of caffiene for the two brands (use $\alpha = 0.05$). Do the findings from your hypothesis test agree with your confidence interval from Example 12.7?

Class Activity: Memorization Group Activity 2

Hypothesis test for an average difference

When data are paired ($d_i = x_{1,i} - x_{2,i}$), provided that the sample size is large enough so that $n_d \geq 30$ or the population is approximately normal, then according to the CLT

$$t = \frac{\overline{x}_d}{s_d/\sqrt{n_d}}$$

follows a t-distribution with $n_d - 1$ degrees of freedom.

We can use this fact to conduct a hypothesis test. As with the hypothesis test we have discussed thus far, there are a specific set of steps to conduct a hypothesis test using the CLT.

Step 1: Check conditions for the test

Check whether $n_d \geq 30$ or the population of differences is approximately normal.

Step 2: State the hypotheses

For a difference in averages, the null hypothesis always has the form

$$H_0 : \mu_d = 0$$

The alternative hypothesis has one of the following three forms:

$$\begin{aligned} H_A&: \ \mu_d < \mu_0 \\ H_A&: \ \mu_d > \mu_0 \\ H_A&: \ \mu_d \neq \mu_0 \end{aligned}$$

Step 3: Calculate the test statistic

To conduct the hypothesis test using the CLT, we use a standardized test statistic of

$$t = \frac{\overline{x}_d}{s_d/\sqrt{n_d}}$$

which follows a t-distribution with $n_d - 1$ degrees of freedom.

Step 4: Find the p-value and give a sketch to illustrate the p-value

To find the p-value, you must use technology to find what proportion of the t-distribution is more extreme than the test statistic (in the direction of the alternative hypothesis). If your alternative is two-sided, then you multiply the one-sided p-value by 2. The sketch is the same as for a single mean.

Step 5: State your decision and interpret in the context of the hypotheses

- If p-value $< \alpha$, we *reject* H_0 and have discernible evidence to suggest that the alternative hypothesis is correct.
- If p-value $\geq \alpha$, we *fail to reject* H_0 and do not have discernible evidence to suggest that the alternative hypothesis is correct.

You should *always* write your interpretation in the context of the original hypothesis.

One advantage of conducting a study using paired observations is that it can greatly reduce variation over independent samples and produce a much more powerful test (i.e. reject the null when you should be) and a more precise confidence interval estimate of the population mean difference. This reduction in variation is greatest when the underlying measurements vary greatly from pair to pair, but the differences do not.

Class Example 3.4.5 *Sales*

A company wants to know if attending a course on "how to be a successful salesperson" can increase the average sales of its employees. The company sent 6 of its salespeople to attend this course. The following gives the one-week sales of these employees before and after they attended the course. Determine if the course increases sales using a 5% significance level. What conditions/assumptions are necessary for the test to be valid?

Before	12	18	25	9	14	16
After	18	24	24	14	19	20

Sales of employees before and after they attended the course.

Practical vs. statistical discernibility

Statistical discernibility $\neq$ practical importance.

A result is *statistically discernible* if the p-value is below the chosen discernibility level α. But statistical discernibility only means the observed effect is unlikely to be due to chance alone. It says nothing about whether the effect is *large enough to matter in practice*.

With very large samples, even trivially small differences can be statistically discernible. With very small samples, even large differences may not be statistically discernible (due to low power).

Summary

This chapter presented hypothesis tests for means using the t-distribution:

- *One-sample t-test*: Tests whether a population mean μ equals a hypothesized value μ_0. Test statistic: $T = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ with $df = n - 1$.
- *Two-sample t-test*: Tests whether two population means are equal. Test statistic: $T = \frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{s_1^2/n_1+s_2^2/n_2}}$ with $df = n_1 + n_2 - 2$.
- *Paired t-test*: Tests whether the mean of paired differences equals zero. Compute differences first, then apply the one-sample t-test to the differences: $T = \frac{\bar{x}_{diff}}{s_{diff}/\sqrt{n_{diff}}}$ with $df = n_{diff} - 1$.
- All three tests require *independence* and *approximate normality or large sample size*.
- *Statistical discernibility* (small p-value) does not imply *practical importance* (large enough to matter).

4.1 Chi-Squared Goodness of Fit

In previous chapters, we have developed inferential methods for comparing proportions across two groups. But what if we want to assess whether a single categorical variable with multiple levels follows a particular distribution? For instance, we might want to know if a jury pool is racially representative of the community, or whether the days of the week on which accidents occur are equally likely. The chi-square goodness of fit test is designed for exactly this purpose. It compares a set of observed counts to a set of expected counts derived from a hypothesized distribution, using a single test statistic that summarizes all of the differences at once.

Key Concepts

- Test a hypothesis about a categorical variable using a chi-square goodness-of-fit test
- Recognize when a chi-square distribution is appropriate for testing a categorical variable

Testing more than two groups

Up until now, we have discussed testing for categorical variables when there are only two groups. However, there are many statistical questions that require us to test categorical variables with three or more groups. In this chapter, we will learn about two new tests that give us additional methods to learn about categorical variables.

In this chapter, we will be discussing a new hypothesis test called the χ^2(chi-square) goodness-of-fit test. The χ^2 goodness-of-fit test is used to test the relative frequency (or proportion) of observations for three or more groups.

The χ^2 goodness-of-fit test is used to test proportions when we have more than two groups in a single categorical variable.

Setting up the hypotheses

Class Example 4.1.1 *Motivating dice example*

Suppose you are interested in determining if a die is fair, so you roll it 54 times and get 8 ones, 7 twos, 13 threes, 11 fours, 9 fives, and 6 sixes. What null and alternative hypothesis would you wante to use to determine if the die is fair?

General Form of the Hypotheses

$$H_0\text{: } p_1 = p_{10}, p_2 = p_{20}, \dots, p_k = p_{k0}$$
$$H_A\text{: At least one } p_i \text{ is different}$$

There are k groups here an one restriction is that the null proportions must sum to 1 ($p_{10} + p_{20} + \cdots + p_{k0} = 1$). Note that the alternative only states that at least one proportion differs from the stated value. If we reject the null hypothesis, we do not know which one is different. We only know that at least one of them is.

Observed versus expected counts

As with all hypothesis tests, we assume that the null hypothesis is true when we conduct our test: we would expect the sample proportions to roughly match the population proportions. For the χ^2 goodness-of-fit test, we use this assumption to figure out the counts we *expect* to see in the sample.

The *expected count* for each category is found by multiplying the sample size by the hypothesized proportion for that category.

Class Example 4.1.2 *Motivating dice example continued*

If the null is true, and the die is fair, what would you expect for the number of ones, twos, threes, ect.?

If the null hypothesis specifies proportions $p_{10}, p_{20}, \ldots, p_{k0}$ for k categories, and the total sample size is n, then the expected count for category i is:

$$E_i = n \times p_{i0}$$

The χ^2 statistic

To build a test statistic, we first compute the standardized difference between each observed and expected count. For each category i:

$$Z_i = \frac{O_i - E_i}{\sqrt{E_i}}$$

where O_i is the observed count and E_i is the expected count. The denominator $\sqrt{E_i}$ serves as the standard error for the count.

We would like to use a single test statistic to determine if these four standardized differences are unusually far from zero as a group. We do this by squaring each standardized difference and adding them up:

$$X^2 = Z_1^2 + Z_2^2 + Z_3^2 + \cdots + Z_k^2$$

Squaring each standardized difference before adding them together does two things:

- Any standardized difference that is squared will be positive, so negative and positive deviations both contribute to the test statistic.
- Differences that are already unusual (e.g., a standardized difference of 2.5) become much larger after squaring, which gives them more weight.

The *chi-square test statistic* is the sum of the squared standardized differences between observed and expected counts:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where the sum is over all k categories, O_i is the observed count in category i, and E_i is the expected count in category i under the null hypothesis.

Notice that a larger test statistic indicates observed data that differs from what H_0 suggests. This should lead us to want to reject H_0. Because we always wish to reject the null for larger test statistics, we always think of this as a *right-tailed* test. Therefore, our p-value will be the area in the right-tail beyond our test statistic.

The χ^2 distribution

The chi-square test statistic X^2 follows a χ^2 **distribution** when the null hypothesis is true and all the expected counts are at least 5 or larger. The χ^2 distribution is characterized by a single parameter called the **degrees of freedom** (df), computed by $k - 1$, which influences the shape, center, and spread of the distribution.

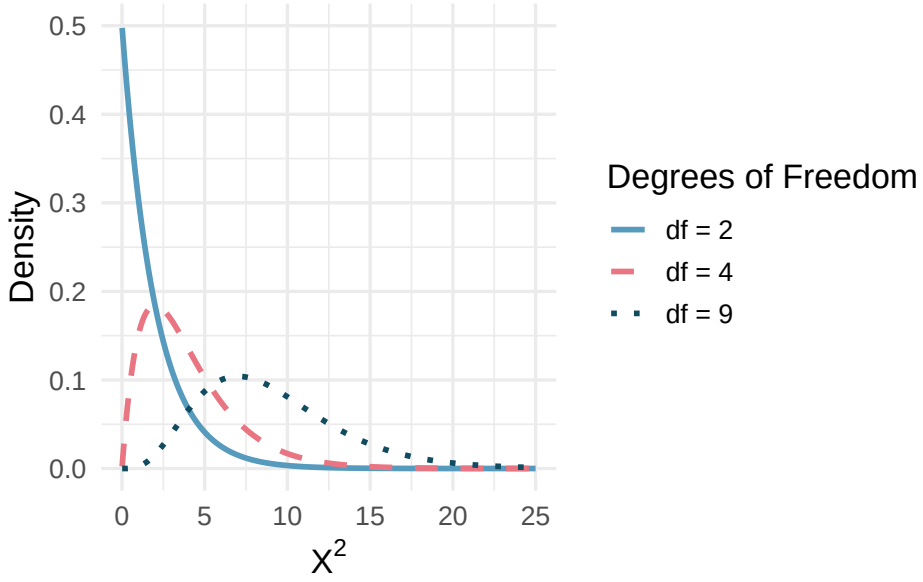


Figure 35: Three chi-square distributions with degrees of freedom 2, 4, and 9. As the degrees of freedom increase, the distribution becomes more symmetric, the center moves to the right, and the variability increases.

Key properties of the chi-square distribution:

- It is always non-negative (the chi-square statistic is a sum of squared values).
- It is right-skewed, especially for small degrees of freedom.
- As the degrees of freedom increase, the distribution becomes more symmetric and shifts to the right.
- The mean of the distribution equals the degrees of freedom.

Class Example 4.1.3 *Motivating dice example continued*

What are the test statistic and p-value for determineing if the die is fair? (Check conditions!) What is our conclusion?

Chi-square goodness of fit test

We can now summarize the complete procedure for a chi-square goodness of fit test.

Chi-square test for goodness of fit.

Suppose we want to evaluate whether the observed counts $O_1, O_2, \dots, O_k$ in k categories are consistent with a hypothesized distribution that specifies expected counts $E_1, E_2, \dots, E_k$.

Hypotheses:

$$H_0\text{: } p_1 = p_{10}, p_2 = p_{20}, \dots, p_k = p_{k0}$$
$$H_A\text{: At least one } p_i \text{ is different}$$

Test statistic:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

When the null hypothesis is true and the conditions are met, X^2 follows a χ^2 distribution with $df = k - 1$.

The p-value is found from the upper tail of the chi-square distribution.

Conditions:

- *Independence.* Each case that contributes a count must be independent of all other cases.
- *Sample size.* Each expected count must be at least 5.

Class Example 4.1.4 *Distribution of fish*

The Game and Fish Department stocked a lake with fish in the following proportions: 30% catfish, 15% bass, 40% bluegill, and 15% pike. Five years later it sampled the lake to see if the distribution of the fish had changed. It found that the 500 fish in the sample were distributed as follows: 117 catfish, 88 bass, 223 bluegill, and 72 pike. Determine if the proportions of the types of fish in the lake have changed in the past 5 years by using a 5% significance level.

Class Example 4.1.5 *Age discrimination on U.S. juries*

It is often said that older people are overrepresented on juries in the United States. The *UCLA Law Review* randomly selected cases in one county in California, and recorded the number of jurors in various age groups given in the table below. Do the following data suggest that the *UCLA Law Review* should expand the scope of their study to more counties (they would do this if they observed discrimination)? Use $\alpha = 0.05$.

	21-40	**41-50**	**51-60**	**61+**	**Total**
Countywide Proportion (%)	42.0	23.0	16.0	19.0	100.0
Grand Jurors (counts)	10	18	38	66	132
Expected (counts)					

Actual ages of jurors in California county

Summary

In this chapter, we developed the chi-square goodness of fit test, which assesses whether a sample of categorical data is consistent with a hypothesized distribution. The test uses the chi-square statistic, which measures how far the observed counts deviate from the expected counts. When the null hypothesis is true and the conditions are met, this statistic follows a chi-square distribution with $k - 1$ degrees of freedom, where k is the number of categories. The p-value is always computed from the upper tail of the chi-square distribution, since larger values indicate greater discrepancy between the observed and expected counts.

4.2 Chi-Squared Test of Independence

In the previous chapter, we used the chi-square statistic to test whether a single categorical variable follows a particular distribution. In this chapter, we extend the chi-square framework to two-way tables, where we have two categorical variables and want to assess whether they are associated. The key question becomes: is the distribution of one variable the same across all levels of the other variable, or does the distribution shift depending on the group? As with our other inference methods, we begin with a simulation-based approach (randomization) and then introduce the mathematical chi-square approximation.
Note that with two-way tables, there is not an obvious single parameter of interest. Instead, research questions focus on how the proportions of the response variable change (or not) across the different levels of the explanatory variable. Because there is no single population parameter to estimate, we focus on hypothesis testing (and not confidence intervals).

Key Concepts

- Test a hypothesis about an association between two categorical variable using a chi-square association test
- Recognize when a chi-square distribution is appropriate for testing an association between two categorical variables

Association Between Categorical Variables

The χ^2 association test is used to test the relationship between two categorical variables

In this chapter, we will be discussing a new hypothesis test called the χ^2 (chi-square) association test. The χ^2 association test is used to test about the relative frequencies (or proportions) of observations for two categorical variables.

Hypotheses

The hypotheses for the χ^2 association test tends to be described more generally than the other tests that we have seen so far. In fact, rather than using parameters in the hypotheses, we will typically express our hypotheses in words. The hypotheses that we want to test in this test are as follows:

H_0: *Variable 1* is not associated with *Variable 2*
H_A: *Variable 1* is associated with *Variable 2*

If we reject the null hypothesis here, we can only say that there is an association/dependency, but we cannot state exactly what that relationship is.

Class Example 4.2.1 *Traffic cameras and injuries*

Many cities have introduced traffic cameras at high-risk street intersections in an effort to reduce the severity of traffic accidents. Citizen rights groups have complained about the efficacy of these cameras. In fact, they doubt they have any impact whatsoever. A sample of 2938 reported accidents was collected including the type of accident (Death/Disabling Injury, Evident Injury, or Probable Injury) and camera status at the intersection (Before Camera, After Camera). What are the hypotheses to test the relationship between accident severity and presence of a traffic camera?

Expected counts in two-way tables

As with all hypothesis tests, we assume that the null hypothesis is true when we conduct our test.

For the χ^2 association test, if the null hypothesis is true, the two categorical variables are *independent*. We use this assumption to figure out the counts we *expect* to see in the sample. To find the expected count for (Variable 1 group i, Variable 2 group j):

The expected counts come from the probability rules. If A and B are *independent*, then $P(A \cap B) = P(A) * P(B)$

$$\begin{aligned} E_{ij} &= n \times p_{ij0} = n \times (p_{i0} \times p_{j0}) \\ &= n \times \left(\frac{\text{Row total}}{n} \times \frac{\text{Column total}}{n} \right) \\ &= \frac{\text{Row total} \times \text{Column total}}{n} \end{aligned}$$

Class Example 4.2.2 *Traffic cameras and injuries, continued*

The table below summarizes the traffic camera data. Use this data to compute a table of expected counts.

	Before Camera	**After Camera**
Death/Disabling	61	27
Evident	210	136
Probable	1659	845

Traffic camera status and accident severity

Test statistic

We want to see if the *observed* counts are far enough away from the *expected* counts

When we conduct the test, the goal is to see if the counts we *observed* are farther from the *expected* counts than we would reasonably expect to see due to random chance. We use the same test-statistic for the χ^2 test of independence that we used for the χ^2 goodness-of-fit test:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}.$$

Just like in the last section, a larger test statistic indicates observed data that differs from what H_0 suggests. This should lead us to want to reject H_0. Therefore, this is always considered a *right-tailed* test.

p-value for χ^2 test of independence

To use the χ^2 distribution, each expected count must be at least 5 or larger

In order for us to be able to use the χ^2 distribution, all of the expected counts need to be at least 5 or larger. If we have met this condition, then we can use a χ^2 distribution with $(r-1) \times (c-1)$ degrees of freedom where r is the number of groups in the first categorical variable (number of rows) and c is the number of groups in the second categorical variable (number of columns). Our p-value will be the area in the *right-tail* beyond our test statistic.

Class Example 4.2.3 *Traffic cameras and injuries, continued*

Find the χ^2 test statistic and p-value using the traffic camera data to test if there is an association between traffic cameras and injury severity. (Check conditions!) What is our conclusion?

Class Example 4.2.4 *Star Trek*

In Star Trek, the colored uniforms worn by the crew identify their work area. Among Trekkies (fans of the show), they have begun to notice a pattern - they believe there is an association with uniform color and if the character suffers a fatality on the show. The table below gives a breakdown of the various crew colors and the number of fatalities - does this give evidence to suggest they are related?

	Blue	**Gold**	**Red**
Fatality	7	9	24
Survival	129	64	263

Uniform color and fatality count

Class Activity: Test of Independence

Summary

In this chapter, we extended the chi-square framework to two-way tables, testing whether two categorical variables are independent. The randomization approach shuffles the data under the null hypothesis to build a distribution of the chi-square statistic, while the mathematical approach uses the chi-square distribution with $(R-1) \times (C-1)$ degrees of freedom. Both approaches require that observations are independent and that expected counts are sufficiently large (at least 5 in each cell).

The chi-square test of independence tells us *whether* two variables are related, but does not tell us *how* they are related or *which* groups differ. To understand the nature of the relationship, we must examine the individual cells and their contributions to the chi-square statistic.

4.3 Analysis of Variance

In earlier chapters, we developed methods to compare the mean of a quantitative outcome across two groups. An important aspect of those analyses was to look at the difference in sample means as an estimate for the difference in population means. When comparing more than two groups, simple subtraction will not fully capture the variation across three or more groups. As with two groups, the research question focuses on whether group membership is independent of the quantitative response variable. In this chapter, we focus on a new statistic that incorporates differences in means across more than two groups. Although the ideas are similar to the t-test, they have earned their own name: ***AN****alysis* ***O****f* ***VA****riance, or* ***ANOVA****.*

Key Concepts

- Use ANOVA to test for a difference in means among several groups
- Explain how variation between groups and variation within groups are relevant for testing a difference in means between multiple groups.
- Create confidence intervals for single means and differences of means after doing an ANOVA for means, recognizing the problem of multiplicity

Sometimes we want to compare means across many groups. We might initially think to do pairwise comparisons. For example, if there were three groups, we might compare the first mean with the second, then with the third, and finally compare the second and third means, a total of three comparisons. However, this strategy can be treacherous. If we have many groups and do many comparisons, it is likely that we will eventually find a difference just by chance, even if there is no difference in the populations. Instead, we should apply a holistic test to check whether there is evidence that at least one pair of groups is different, which is where ANOVA saves the day.

Consider the figure below showing two sets of three groups.

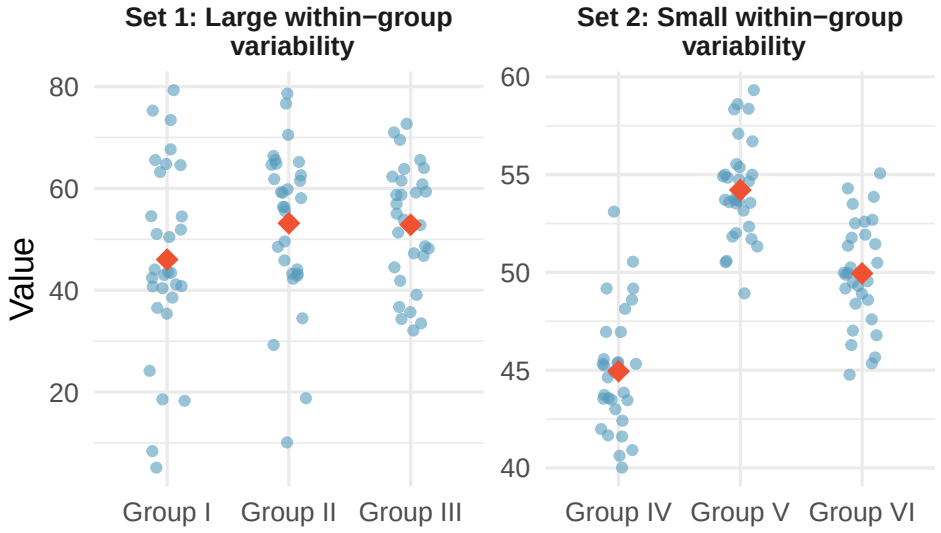


Figure 36: Side-by-side dot plots for two sets of three groups. In the first set (Groups I, II, III), the within-group variability is large relative to the between-group differences. In the second set (Groups IV, V, VI), the within-group variability is small and the between-group differences are clearly visible.

Compare Groups I, II, and III. Can you visually determine if the group differences are real? Now compare Groups IV, V, and VI.

Analysis of Variance

In the example above, any real difference in Groups I, II, and III is difficult to discern because the *spread within* each group was large compared to the *spread between* the means. On the other hand, Groups IV, V, and VI show noticeable differences because the *spread within* was samller relative to the variability between the means. We will use the spread/variability to decide which proveded stronger evidence.

Recall, the test statistic for a two-sample t-test is:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

In the numerator of the test statistic, we calculated the *difference* between the two sample means, or the spread between the two groups. In the denominator, we

calculated the standard error, which is a measure of spread within the two groups. How would this test statistic change if we wanted to test if there was a difference in means between three or more groups?

ANOVA is an extension of the two-sample t-test to more than two groups.

The Analysis of Variance (ANOVA) procedure can be regarded as an extension of the two-sample t-test to more than two samples. In fact, the test statistic is formed in much the same way. The test statistic for ANOVA is an F-statistic, and has the following form.

$$F = \frac{\text{Spread } \textit{between} \text{ group means}}{\text{Spread } \textit{within} \text{ groups}}$$

This is the same concept as our t-statistic from a two sample t-test; in fact, we can think of the difference in means (numerator of t) to be a measure of spread between the means.

Statement of hypotheses

The ANOVA hypothesis test is used to see if groups have significantly different averages. If we are considering k groups, then the hypotheses for the ANOVA test are

$$H_0\text{: } \mu_1 = \mu_2 = ... = \mu_k$$
$$H_A\text{: At least two } \mu_i\text{'s differ}$$

Class Example 4.3.1 *Quality control for lenses*

A company that produces glass lenses would like to know if the three presses they use are producing the same average thickness in lens. They hire an engineer to test to see if the machines should be calibrated. Write the hypothesis to test this using an ANOVA hypothesis test.

Finding the test statistic

The test statistic for an ANOVA hypothesis is an F-statistic with a numerator that measures variability between groups and a denominator that measures variability within groups.

SS stands from *sum of sqaures*
MS stands from *mean sqaures*
B indicates "between" and W indicates "within"

$$F = \frac{MSB}{MSW} = \frac{\frac{SSB}{df_B}}{\frac{SSW}{df_W}}$$

The equations for the sums of squares and df's are

$$SSB = n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + \cdots + n_k(\overline{x}_k - \overline{x})^2$$
$$SSW = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2$$
$$df_B = k - 1$$
$$df_W = n - k$$

where $\overline{x}_i$ is the sample mean for group i, s_i is the standard deviation for group i, and n_i is the sample size for group i. The values of n and $\overline{x}$ are from *all* data. n is the total sample size, or $n = n_1 + n_2 + \cdots + n_k$. $\overline{x}$ is the grand or overall mean. It is the mean of *all* data or can be computed from the individual group means by

$$\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + \cdots + n_k\overline{x}_k}{n}.$$

When the variation from groups to group (between) is large enough to overshadow the variation within the groups, we have a large *MSB* as compare to *MSW*, which produces a larger *F*. When the variation from group to group is large, we would want to conclude at least one mean differs and reject H_0. Therefore, the ANOVA is always considered a right-tailed test.

Finding the p*-value*

To find the p-value we use an F-distribution. The F distribution a right-skewed distribution that depends on *two* degrees of freedom, which are called the df numerator and df denominator. For an ANOVA the df numerator is $df_B = k - 1$ and the df denominator is $df_W = n - k$. We can use StatLens or some other technology to find the area in the right-tail beyond our F test statistic using the appropriate degrees of freedom.

Summarizing all quantities in an ANOVA table

Because there are various quantities calculated when doing an ANOVA test, they are nicely summarized in an ANOVA table. The general form of an ANOVA table is given below.

Source	**df**	**Sum Sq**	**Mean Sq**	**F value**	**p-value**
Between/Group	$k-1$	SSB	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$	p-value
Within/Error	$n-k$	SSW	$MSW = \frac{SSW}{n-k}$		
Total	$n-1$	SST			

General form of an ANOVA table

Note that $SST = SSB + SSE$ and

$df_{Total} = df_B + df_W = (k-1) + (n-k) = n - 1$

Class Example 4.3.2 *Quality control continued*

Fill in the remaining spots in the ANOVA table below.

Source	**df**	**SS**	**MS**	**F**	p**-value**
Group		10.30			
Error					
Total	14	34.40			

Quality control ANOVA table

Find the p-value for the F-statistic that tests to see whether or not average glass thickness differs for each press. What are the degrees of freedom for the numerator and denominator? State your decision and provide an interpretation in the context of the problem.

Conditions for ANOVA

There are three conditions to check before performing ANOVA:

- *Independence.* If the data are a simple random sample, this condition is generally satisfied. For experiments, carefully consider whether the data may be independent (e.g., no pairing).
- *Approximately normal.* As with one- and two-sample testing for means, the normality assumption is especially important when sample sizes are small. With large sample sizes within each group, we look mainly for extreme outliers.
- *Constant variance.* The variability in the groups should be about equal. This can be checked by examining side-by-side box plots or comparing the standard deviations across groups. A common rule of thumb is that the ratio of the largest to smallest group standard deviation should be less than 2.

Class Example 4.3.3 *Caffeine and finger tapping*

Does caffeine consumption have an effect on twitch movements? To investigate this question, the following experiment was conducted. 30 subjects were randomly assigned to one of three groups, with 10 subjects in each group. The subjects were then given a tablet containing 0mg, 100mg, or 200mg of caffeine. Two hours later, they were asked to tap their finger as quickly as possible for one minute. Summary statistics are provided in the table below.

	0 mg	**100 mg**	**200 mg**
$\overline{x}$	244.8	246.4	248.3
s	2.394	2.066	2.214

Summary statistics for caffeine and finger tapping experiment

a) What are the variables recorded in this study? Are they categorical or quantitative? Which of the variables is the independent variable? Dependent variable?

b) Based on the summary statistics above, do you think that caffeine has an effect on finger tapping?

c) Test to see whether caffeine consumption causees a difference in the average number of finger taps (use $\alpha = 0.05$)

Table 3: ANOVA summary table

Source	df	SS	MS	F	p-value
Group		61.4			
Error		134.1			
Total					

d) Is there a *practical* difference between the groups? Is this the same as a *statistical* difference?

Pairwise comparisons and inference after ANOVA

Although an ANOVA hypothesis test will tell you that there is a difference in at least one of the population means, it does not specify which averages are different. If H_0 is not rejected in an ANOVA, then there is no need to look for differences among population means. However, if H_0 is rejected, the next step is to see which means are actually different from the others. For a significant ANOVA test, there are a number of methods to compare the averages for the different groups.

When we compare these groups, we do what is called *pairwise comparisons* of the means. The phrase *pairwise comparisons* means conducting hypothesis tests between all the possible pairs of means to see which pairs are different. We can find the number of comparisons among k population means by using the formula

Pairwise comparisons is just a fancy way of saying consider all the possible pairs of means and conduct hypothesis tests to see if they are different.
Pairwise comparisons are also called *multiple comparisons* or *post-hoc tests*

$$\text{\# of comparisons} = \binom{k}{2} = \frac{k!}{2!(k-2)!} = \frac{k \times (k-1)}{2}$$

where k is the number of groups. This is read as "k choose 2" and you can use the `nCr` function on your calculator to do the calculation.

Class Example 4.3.4 *Caffeine and finger tapping continues*

How many pairwise comparisons would we have? Suppose that there had been four caffeine levels instead. How many pairwise comparisons would this make?

The multiple testing problem

The *family-wise error rate (FWER)* is the chance that we made a Type I error in at least one of our tests.

If you use a two-sample t-test to compare two averages at $\alpha = 0.05$, then there is a 5% chance that we will make a Type I error. However, as we conduct two-sample t-tests to compare more pairs of averages, then we don't want to consider the Type I error rate for only a single test. We want to know the *family-wise error rate* $(FWER)$, or the chance that we make a Type I error in at least one of the tests we conduct. To find the overall Type I error rate when comparing all pairwise comparisons, c, for k groups, we use the following formula:

$$FWER = 1 - (1 - \alpha)^c$$

Class Example 4.3.5 *Caffeine and finger tapping continues*

IF we conduct pairwise comparisons for the average number of finger taps for the 3 caffeiene levels at $\alpha = 0.05$, what is the FWER?

The table below shows how the FWER grows with the number of comparisons when $\alpha = 0.05$:

Number of groups (k)	**Number of comparisons** (c)	**FWER**
3	3	0.143
4	6	0.265
5	10	0.401
6	15	0.537
10	45	0.901

Relationship between number of comparisons and FWER

With 10 groups, there is a 90% chance of making at least one Type I error if we use unadjusted pairwise tests. This is the core of the multiple comparisons problem: we need procedures that control the FWER at an acceptable level (typically 0.05).

The Bonferroni correction

There are a number of ways to correct for the overall Type I error rate, but one of the easiest is the Bonferroni correction. When we do this correction, if we are conducting c tests and want the overall FWER to be at most α, we use a stricter discernibility level for each individual test.

$$\alpha^* = \frac{\alpha}{c}$$

Class Example 4.3.6 *Bonferroni correction for finger taps*

If we want the overall Type I error rate to be 0.05, what is the Bonferroni corrected α^* we should use when comparing the average number of finger taps for the 3 caffeine levels? What Bonferroni corrected α^* would we use if there were 4 caffeine levels instead of 3?

Class Example 4.3.7 *Restaurant tips*

In this example, we will be conducting many different tests from a dataset that contains information on the tipping patterns of patrons of the *First Crush* bistro in northern New York state. The data from 157 bills include the amount of the bill, size of the tip, percentage tip, number of customers in the group, whether or not a credit card was used, day of the week, and a coded identity of the server (A, B, or C). The first four variables are quantitative and the last three are categorical. For each test, you can use a hypothesis test or confidence interval (if appropriate). On each test, be sure to verify the necessary conditions.

a) Most restaurant patrons leave a tip between 15% and 20% of the bill, and some restaurant customers determine the percent tip to leave based on the quality of the service. Summary statistics for mean tip percentage left for each of the three servers and a partial ANOVA table are given in the tables below.

Server	**Sample size**	**Mean**	**Std. Dev.**
A	60	17.543	5.504
B	65	16.017	3.485
C	32	16.109	3.376

Mean tip percentage by server

Table 4: ANOVA table for mean tip percentage

Source	df	SS	MS	F	p-value
Group			41.57		
Error					
Total		3001			

i. Conduct the appropriate hypothesis test to see if there a difference in mean tip percentage between the three servers (use $\alpha = 0.05$).

ii. Is it appropriate to examine pairwise differences to see which of the servers differs in average tip percentage? Explain.

iii. Regardless of your answer in a.ii, if we want the overall Type I error rate to be 0.05, what is the Bonferroni corrected α we should use when comparing the average tip percentage for the servers?

b) The table below summarizes the number of bills by server. Conduct the appropriate hypothesis test to see if the servers serve equal numbers of tables.

Server	**A**	**B**	**C**	**Total**
Number of bills	60	65	32	157

Number of bills by server

Summary

In this chapter, we introduced the mathematical model for comparing means across two or more groups using ANOVA. The F-statistic measures the ratio of between-group variability (MSB) to within-group variability (MSW). When the null hypothesis is true and the conditions are satisfied, the F-statistic follows an F-distribution. A discernible result tells us that at least one group mean differs from the others, but does not identify *which* groups differ.

Performing many pairwise tests without correction inflates the family-wise error rate, making false positives likely. The Bonferroni correction divides the discernibility level by the number of comparisons, which is simple but conservative and allows us to determine *which* groups differ.

4.4 Correlation and Linear Regression

When we have two quantitative variables, we often want to understand the relationship between them. Does one variable increase as the other increases? Does it decrease? Or is there no apparent pattern? In this chapter, we introduce tools for visualizing and quantifying the strength and direction of the linear relationship between two quantitative variables and how to model the relationship. We introduce the correlation coefficient r, discuss its properties and limitations, and address the ever-important distinction between correlation and causation. We continue to build on the ideas of correlation to develop the least squares regression line via linear regression. We learn how to interpret the slope and intercept, use residuals to assess model fit, quantify the strength of the model with R^2, and perform inference on the slope to test whether a linear relationship exists in the population.

Key Concepts

- Use the linear correlation coefficient, r, to quantify the strength and direction of a linear relationship
- Identify when the correlation coefficient can capture the trend between two quantitative variables
- Recognize that correlation does not equal causation
- Use computer output to make predictions and interpret coefficients using a fitted simple linear model
- Construct a confidence interval for the slope in a regression model
- Test a hypothesis about the slope of a regression model
- Test for evidence of a linear association between two quantitative variables using a sample correlation
- Compute and interpret the value of R^2 for a regression model
- Check a scatterplot for obvious departures from a simple linear model

Scatterplots

A *scatterplot* is a graph of the relationship between two quantitative variables

In Unit 1, we saw that we can visualize the relationship between two quantitative variables with *scatterplots*.

When examining a scatterplot, we look for four key features:

- *Direction*: Is the relationship positive (upward trend), negative (downward trend), or neither?
- *Form*: Is the relationship linear (points cluster around a line) or nonlinear (points follow a curve)?
- *Strength*: How tightly do the points cluster around the trend? A strong relationship has little scatter; a weak relationship has lots of scatter.
- *Unusual observations*: Are there any outliers that deviate from the overall pattern?

Class Example 4.4.1 *Describing Relationships*

Researchers captured 104 brushtail possums in Australia and took body measurements before releasing them. Consider the relationship between total body length (cm) and head length (mm).

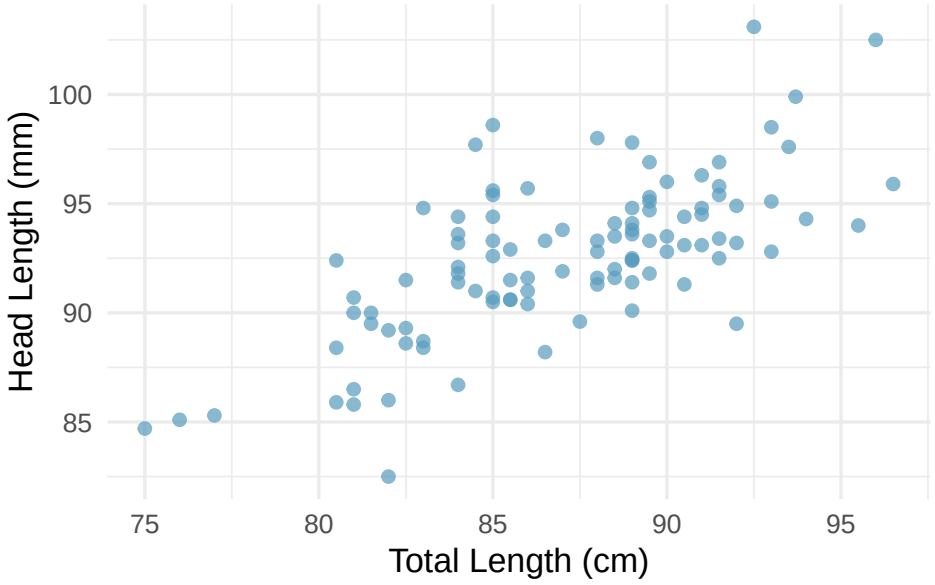


Figure 37: A scatterplot showing head length against total length for 104 brushtail possums. The relationship is moderately strong, positive, and linear.

Describe the relationship between total length and head length.

However, there are cases where fitting a straight line to the data is not helpful, even if there is a clear relationship between the variables.

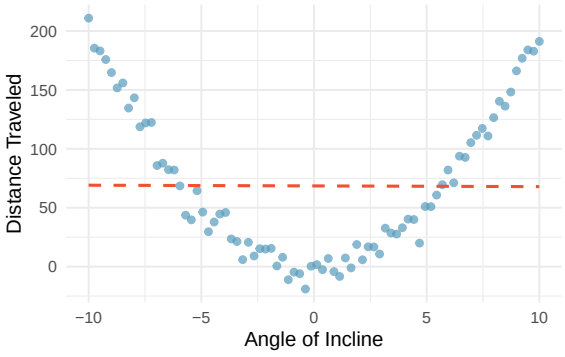


Figure 38: A scatterplot showing a clear quadratic (U-shaped) relationship between angle of incline and distance traveled. The best-fitting straight line is flat and does not capture the pattern at all.

When the relationship is nonlinear, fitting a straight line can be misleading. We must always examine scatterplots before computing correlations or fitting linear models.

The linear correlation coefficient

The *correlation coefficient*, denoted r, quantifies the strength and direction of a linear relationship.

The formula for the correlation coefficient is:

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

where $\bar{x}$ and $\bar{y}$ are the sample means, and s_x and s_y are the sample standard deviations for each variable. In practice, we always use a computer or calculator to compute r.

The correlation coefficient has several important properties:

1. *Range:* $-1 \leq r \leq 1$.
2. *Unitless:* Correlation has no units. Changing the units of x or y (e.g., from cm to inches) does not change r.
3. *Symmetric:* The correlation of x with y is the same as the correlation of y with x.
4. *Not affected by linear transformations:* Adding a constant to or multiplying a variable by a positive constant does not change r.
5. *Only measures linear relationships:* Correlation can be misleading for nonlinear relationships.

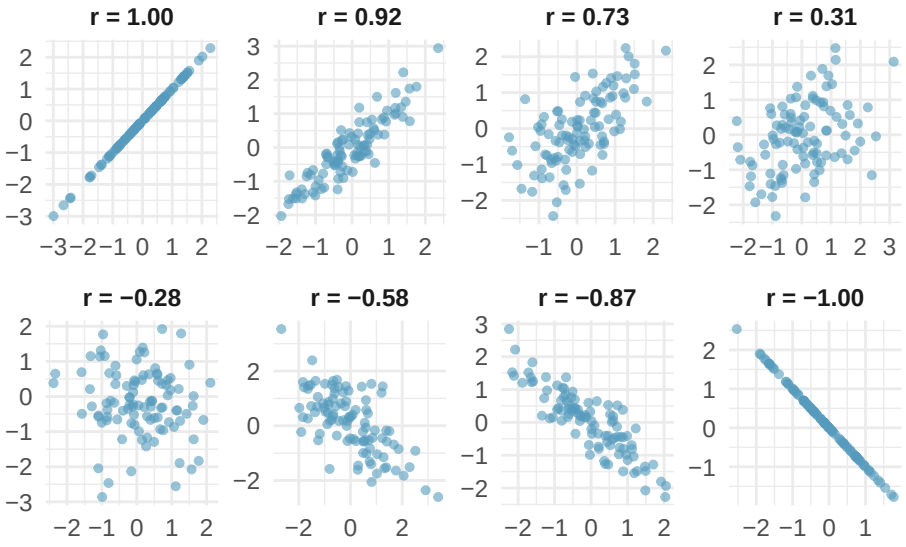


Figure 39: Eight scatterplots showing different correlation values. The first row shows positive relationships with r values of 1.00, 0.92, 0.73, and 0.31. The second row shows negative relationships with r values of -0.28, -0.58, -0.87, and -1.00.

Key patterns to notice:

- When $r = 1$ or $r = -1$, all points fall exactly on a straight line (a perfect linear relationship).
- When r is close to $+1$ or -1, the relationship is strong and linear.
- When r is close to 0, there is no linear relationship (but there could still be a nonlinear relationship!).
- Positive r means both variables tend to increase together; negative r means one tends to decrease as the other increases.

Important Note

Correlation does not imply causation. Just because two variables are correlated does not mean that one causes the other. There may be a *confounding variable* (also called a lurking variable) that is associated with both x and y, creating a correlation between them even though there is no direct causal link.

Class Example 4.4.2 *Firefighters and Damage*

Studies have found a positive correlation between the number of firefighters sent to a fire and the amount of damage the fire causes. Does this mean sending more firefighters causes more damage?

Linear Regression

Notation:

- β_0: population intercept
- β_1: population slope
- ε: random error
- $\hat{y}$: predicted value
- b_0: intercept estimate
- b_1: slope estimate

Linear regression is the statistical method for modeling a linear relationship between two variables, x and y. It is rare for all of the data to fall perfectly on a straight line. Instead, it's more common for the data to appear a a cloud of points. Thus, to have a population linear model, we must include an error term:

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

The values β_0 and β_1 represent the model's *population intercept* and *population slope*, respectively, and the error is represented by ε. If we collect data from a random sample from a target population, we can estimate β_0 and β_1 with b_0 and b_1, the *sample intercept* and *sample slope*, respectively:

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

The "hat" on y signifies that this is a predicted (estimated) value.

When we use x to predict y, we call x the **predictor** variable (or explanatory variable) and y the **outcome** variable (or response variable). The *slope*, b_1, represents the average predicted change in the response variable,y, given a one unit increase in the explanatory variable, x. The *intercept*, b_0, represents the predicted value of the response variable, y, if the explanatory variable, x is zero. The equation for the intercept and slope are given as

$$b_1 = r\frac{s_y}{s_x} \qquad b_0 = \overline{y} - b_1\overline{x}.$$

One thing to notes from this formula is that the *slope* and *correlation* should always have the same sign.

The *slope* and *correlation coefficient* should *always* have the same sign

Class Example 4.4.3 *Possums*

Researchers captured 104 brushtail possums in Australia and took body measurements. We consider total body length (cm) as the predictor to predict head length (mm). Some summary statistics and a scatterplot of the data are given below

Variable	Average	Std. Dev.
Body Length	87.09	4.31
Head Length	92.60	3.57
	$r = .691$	

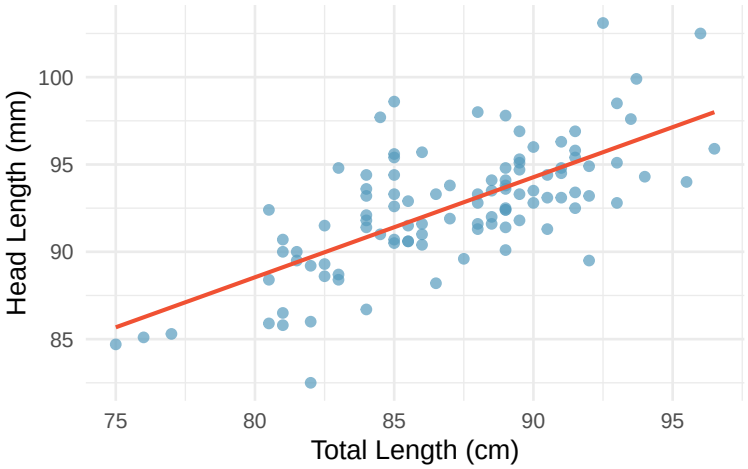


Figure 40: A scatterplot showing head length against total length for 104 brushtail possums, with a linear model fit to the data.

a) Looking at the plot, would you categorize the correlation as positive or negative? Strong, moderate, or weak?

b) Use the summaries to find estimates for the slope and intercept for the regression line. Record your answer in the form $\hat{y} = b_0 + b_1 x$.

Interpreting Slope and Intercept

For the regression line $\hat{y} = b_0 + b_1 x$:

- The slope, b_1, is interpreted as the predicted change in the response variable y for a one unit increase in the explanatory variable.
- The intercept, b_0, is interpreted as the predicted value of y when $x = 0$. In some situations it may not make sense to have $x = 0$ (a person's height, for example), and in these cases, we do not interpret b_0.

Residuals

A *residual* (or error) is the difference between an observed value of y and the predicted value of y at an observed value of X:

A *residual* is the difference between an observed values and a predicted value of a response

The *least squares line* is the equation that has the smallest sum of squared residuals

$$\text{Residual} = \text{Observed} - \text{Predicted} = y - \hat{y}$$

Visually, this is the vertical distance between and the regression line. We say that the line that fits the data best is the *least squares line*, which minimizes the sum of the squared residuals.

Regression Cautions

- While we can predict y by plugging an x in the least squares regression line, we should try to avoid predicting for x value outside the range of the data collected (called *extrapolation*). This is because the fit of the line may no longer hold for values outside the range of collected data.
- While we can find the best fitting line for *any* data set, we should first plot the data to ensure a linear fit is appropriate.
- Outliers can have strong influence on the regression line, just as we saw for correlation.

Class Example 4.4.4 *Possums, revisited*

Consider the previous example.

a) Predict the head length of a possum whose body length is 90mm. Is this prediction extrapolation?

b) One of the possums had a body length of 83mm and a head length of 88.4mm. Find the predicted value and residual for this possum.

c) Provide an interpretation of the slope and intercept (if reasonable) in the context of this problem.

Inference for Slopes

As with means and proportions, the regression coefficients,b_0 and b_1, are statistics and are estimates of the population intercept, β_0, and population slope, β_1. Most of the time, we will not be interested in inference for β_0, so the rest of the lecture on regression will focus on inference for the population slope β_1. To conduct a hypothesis test for the slope, we will be testing

$$H_0\text{: } \beta_1 = 0 \qquad H_A\text{: } \begin{array}{r} \beta_1 < 0 \\ \beta_1 > 0 \\ \beta_1 \neq 0 \end{array}$$

What do each of these alternative hypotheses represent in words?

As with all statistics, we need to report a standard error in order to determine if our findings are significant. This will be provided to you with software output, so you do not need to worry about calculating this on your own. However, you will need to know how to find the test statistic in order to find a p-value. For a hypothesis test about a slope, we use the test statistic

SE_{b_1} is the standard error of b_1

$$t = \frac{b_1 - 0}{\text{SE}_{b_1}}$$

where SE_{b_1} stands for standard error of b_1. To find a p-value, you should use a t-distribution with $n-2$ degrees of freedom. The remaining steps of the hypothesis test are the same as all other hypothesis tests.

The table below shows output that is typically given by software when conducting linear regression.

Coefficients	**Estimate**	**Std. Error**	t
(Intercept)	b_0	SE_{b_0}	t for b_0
Explanatory Variable	b_1	SE_{b_1}	t for b_1

Example output from software

We can also construct a confidence interval for the population slope, β_1, using the following with $n-2$ degrees of freedom.

$$b_1 \pm t^*_{n-2}\text{SE}_{b_1}$$

Class Example 4.4.5 *Possums continued*

The table below gives the output from results for a simple linear regression of possum head length. Use the output to conduct a hypothesis test to see if there is some linear relationship between the average body length and average head length of possums, using $\alpha = 0.05$. Compute a 95% confidence interval for β_1. Does the results of the confidence interval match the results of the hypothesis test?

Coefficients	**Estimate**	**Std. Error**	t
(Intercept)	42.71	5.17	8.257
Length	0.57	0.06	?

Output from results for a simple linear regression of possum head lengths with body length

Conditions for inference

For the mathematical inference to be valid, we check four conditions, often remembered by the *LINE* mnemonic:

- *L – Linear model* The data should show a linear trend. Check the scatterplot and residual plot.
- *I – Independent observations:* Be cautious with time series data or clustered observations.
- *N – Nearly normal residuals:* Residuals should be approximately normally distributed. This is less critical for large samples (Central Limit Theorem), but outliers are always a concern.
- *E – Equal variability:* The variability of points around the line should be roughly constant across all values of x. A deviation from this look like more of a fan or megaphone shape.

We can check conditions 1, 3, and 4 using a scatterplot of the data. Condition 2 can be checked using a histogram of residuals.

Coefficient of determination, R^2

R^2 gives the percentage of variation in the *response* variable that we can explain with the *explanatory* variable

There is another statistic from regression that is often used to compare two simple linear regressions models with each other. The *coefficient of determination*, notated R^2, gives the percentage (or proportion) of variation in the *response* variable that we can explain by the linear regression using the *explanatory* variable. Because we hope the explanatory variable and fitted model do a good job in explaining the response, a higher R^2 is desired.

When fitting a *simple* linear regression model, we can get the coefficient of determination by squaring the correlation, r.

A more general idea of how R^2 is computed will be discussed in the next chapter, but *if we are fitting a simple linear model*, we can find the coefficient of determination, R^2, by squaring the correlation coefficient, r. Therefore, $0 \leq R^2 \leq 1$. Values closer to 1 suggest a higher percentage of variation that *is* being explained. What if we are not interested in a *linear* model but some sort of curve instead? Then we know that the correlation r does not tell us about the adequacy of the fitted curve since it only describes a linear relationship. However, the coefficient of determination R^2 *can* tell us how well a curve fits. This makes it more general in purpose than correlation.

Class Example 4.4.6 *Investigating the coefficient of determination*

Match each of the following scatterplots below (each displays the fitted model as the dashed line/curve) to one of the following values of coefficient of determination: $R^2 = .98$, $R^2 = .80$, $R^2 = .37$, or $R^2 = .001$

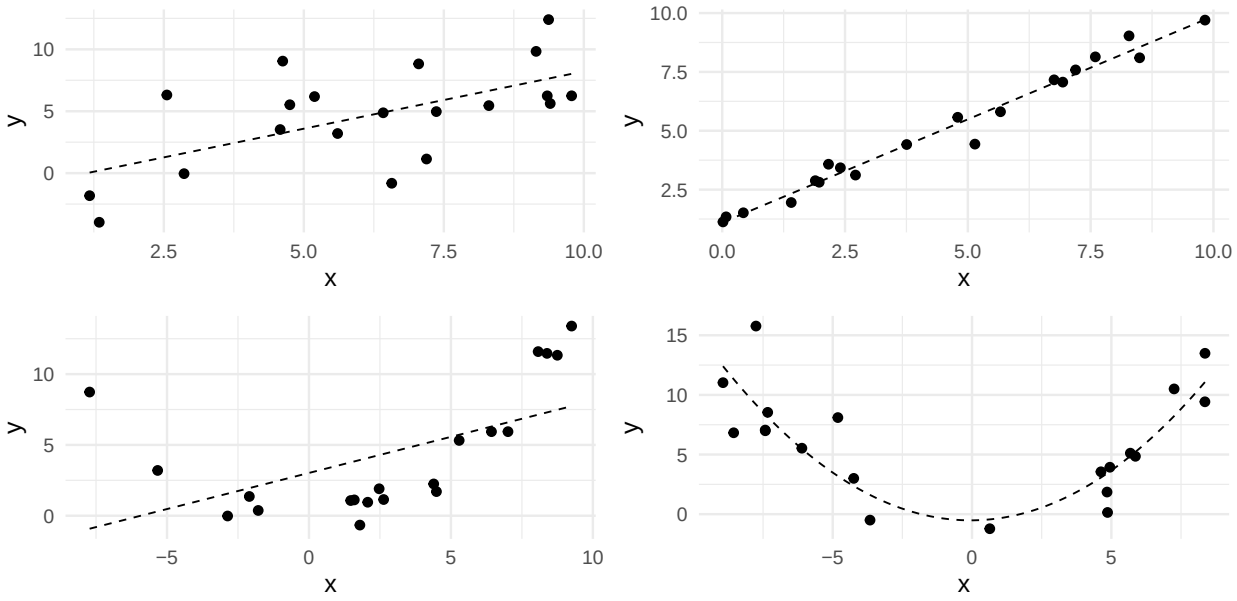


Figure 41: Four scatterplots with fitted regression model fit to the data.

Class Example 4.4.7 R^2 *for the Possum data*

If the correlation coefficient for the simple linear regression model relating the body length and head length of possums is 0.691, find and interpret the coefficient of determination, R^2.

Class Example 4.4.8 *Possum data, revisited*

In the previous examples, we use the variable *body length* to predict *head length*. Now, we are going to use *body length* to predict *skull width*. The following table give the output from a simple linear regression model of skull width with body length.

Coefficients	**Estimate**	**Std. Error**	t	p**-value (2-sided)**
(Intercept)	23.77	5.30	4.48	<0.0001
Length	0.38	0.06	?	<0.001
		$r = 0.526$		

Output from results for a simple linear regression of skull width with body length

a) Using the output above, write the regression equation in the form $\hat{y} = b_0 + b_1 x$. Use it to find the predicted and residual value for a possum with a length of 80.5mm and a skull width of 56mm.

b) Interpret the slope of the line in the context of the problem.

c) Conduct a hypothesis test to see if there is a significant positive relationship between length and skull width. Use $\alpha = 0.05$.

d) Find R^2 for this linear fit. Does body length explain head length or skull width better? Explain.

Summary

In this chapter, we introduced scatterplots as the primary tool for visualizing the relationship between two quantitative variables. We then defined the correlation coefficient r as a measure of the strength and direction of the linear relationship. Key properties of r include: it ranges from -1 to $+1$, it is unitless, it is symmetric, and it measures only linear relationships. We emphasized that correlation does not imply causation – confounding variables can create spurious associations. Finally, we noted that the sample correlation is subject to sampling variability, especially in small samples.

We developed the least squares regression line as a tool for modeling the linear relationship between two quantitative variables. The slope describes the predicted change in y for a one-unit increase in x, and the intercept describes the predicted value of y when $x = 0$. Residuals measure the difference between observed and predicted values, and residual plots help assess whether the linear model is appropriate. The coefficient of determination R^2 quantifies the proportion of variability in y explained by the model. For inference, we use either a randomization test or the t-distribution to test whether the slope is discernibly different from zero, and we can construct confidence intervals for the slope. The LINE conditions (linearity, independence, normality, equal variance) must be checked for the mathematical model to be valid.

4.5 Multiple Linear Regression

In the previous chapter, we used a single explanatory variable to predict the value of our response variable using simple linear regression. In this chapter, we extend the linear regression framework to include multiple explanatory variables. First, we connect the ideas of linear regression and ANOVA. Then, using the ANOVA for regression framework, we add additional explanatory variables into a regression model and can make decisions on which model best explains the response variable.

Key Concepts

- Use ANOVA to test the effectiveness of a simple linear model
- Compute and interpret the value of R^2 for a simple linear model using values in the ANOVA table.
- Find the standard deviation of the error term for a simple linear model
- Find the standard error of the slope for a simple linear model
- Use computer output to make predictions and interpret coefficients using a multiple regression model
- Use ANOVA to test the overall effectiveness of a multiple regression model
- Test the effectiveness of individual terms in a multiple regression model
- Compute and interpret the value of R^2 for a multiple regression model

Anova for Regression

Variability in the response variable can be partitioned into two components:

- $SS_{\text{Regression}}$
- SS_{Error}

When conducting a simple linear regression using the computer, often times, an ANOVA table is included in the output. The Sum of Squares column of the ANOVA table gives us a way to partition the variability of our response into two components, variability we can explain with our regression model, and variability that is unexplained. In the context of regression, this looks like

$$SS_{\text{Total}} = SS_{\text{Regression}} + SS_{\text{Error}}$$

The F-statistic

Another approach to determining if the linear model is effective (beyond the tests from the previous chapter) is to use the ANOVA table for regression, which includes an F test statistic. Recall that the F-distribution has two degrees of freedom. The numerator df for a regression model equals the number of explanaotory, x, variables in the model. For simple linear regression with one x, this is 1. The denominator df is $n-1-df_{\text{numerator}}$, so for simple linear regression this is $n-2$. Note that you can have multiple explanatory variables as predictors for a response variable and this topic will be discuss later in this chapter. If we take the sums of squares divided by the respective degrees of freedom, we get the mean squares (MS) and the F test statistic is

$$F = \frac{MS_{\text{Regression}}}{MS_{\text{Error}}}$$

The F test will allow us to test the hypotheses of

H_0: The model is ineffective
H_A: The model is effective

Understanding the definition of R^2 through ANOVA

In the previous chapter, the coefficient of determination, R^2, could be found by squaring the correlation coefficient, r (if considering a simple linear model). It can also be computed as

$$R^2 = \frac{SS_{\text{Regression}}}{SS_{\text{Total}}}$$

This is the reason R^2 is interpreted as the proportion of variability in our response that is explained by the model

Class Example 4.5.1 *Elmhurst College*

Elmhurst College conducted a study on the amount of gift aid given to students based on their family income (*The Chronicles of Higher Education*, 2012). 50 incoming freshman were randomly sampled and the amount of aid they were given along with their family's income were collected, both in thousands of dollars. A partially completed ANOVA table and other regression output are given below.

Table 5: Family income predicted gift aid ANOVA table

Source	df	SS	MS	F	p-value
Group		363.16			
Error		1097.92			
Total					

Table 6: Output from results for a simple linear regression of gift aid with family income

Coefficients	Estimate	Std. Error	t	p-value (2-sided)
(Intercept)	24.32	1.291	18.831	<0.0001
Income	−0.043	0.011	−3.985	0.0002

a) Complete the partial ANOVA table and use it to calculate R^2 and provide an interpretation in the context of the study.

b) Conduct an F test to determine if the linear model fit is appropriate.

c) How does the F test statistic relate to the t test statistic provided in the output?

Standard deviation of the error

In the previous chapter, we had to use the standard error of the slope to do inference for β_1. This quantity can be obtained using the ANOVA regression table. The error sums of squares from the table, SSE, is the sum of the squared residuals $\sum(y-\hat{y})^2$. We can get what is called the standard deviation of the error as

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE}$$

Then the standard error of the slope can be calculated as follows, where s_x is the standard deviation of the x-value.

$$SE_{b_1} = \frac{s_\varepsilon}{s_x\sqrt{n-1}}$$

Note that when data values are more scattered about the regression line, our residuals are larger. This produced a larges SSE, which means a larger s_ε. This in turn gives a larger SE_{b_1} which will produce wider CIs and a smaller test statistic (large p-values) when doing inference on the population slope.

Class Example 4.5.2 *Elmhurst College, continued*

Calculate the standard error of the slope using the formula above and verify it matches what is in the output. Note that $s_x = 63.21$.

Multiple Regression Model

In the previous chapter, we were using a simple linear regerssion model. This meant we fit a line to the sample, yielding $\hat{y} = b_0 + b_1 x$, and this estimated the true population line of $y = \beta_0 + \beta_1 x + \varepsilon$. However, there are many data sets with several quantitative variables that may serve as predictors for a response. In this case we use *multiple regression*. If we have k predictors, the multiple regression model looks like

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

which is estimated by

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k.$$

Note that this model still suggests a *linear* relationship between each x and the response. Therefore, we should look at plots of y against each x variable separately to make sure a linear trend is valid. Many of the topics we saw for the simple linear regression case like the F test, t tests for slop, and R^2 can also be used to investigate the adequacy of a multiple regression model.

ANOVA for Multiple Regression

Just as we saw earlier, we can divide the total variability into the part that is explained by the fitted model and the predictors ($SS_{\text{Regression}}$), and the part that is not explained (SS_{Error}).

$$SS_{\text{Total}} = SS_{\text{Regression}} + SS_{\text{Error}}$$

We can divide the sums of squares by their degrees of freedom to get the mean squares, and then get an F statistic as

$$F = \frac{MS_{\text{Regression}}}{MS_{\text{Error}}}$$

What are the degrees of freedom? As earlier mentioned, the numerator df for a regression model equals the number of x variables in the model, denoted as k. The denominator df is $n - k - 1$.

When we use this *overall* F test we are testing the following hypotheses

$$\begin{aligned} H_0&: \ \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_A&: \ \text{At least one} \beta_i \neq 0 \end{aligned}$$

Notice that the null says all "slope" parameters are 0. In this case they drop out of the model meaning that their corresponding x values were not useful predictors of y. The alternative hypothesis just states that at least one of the predictors is useful, but does not indicate which ones. If H_0 is rejected in favor of H_A, we can later try to determine which predictors are useful through individual t tests.

Coefficient of Determination, R^2

As before, we can find the coefficient of determination as

$$R^2 = \frac{SS_{\text{Regression}}}{SS_{\text{Total}}}$$

and interpret this as the percent of variability in y that can be explained by the fitted regression model. Note that you can compute the correlation, r, between y and each x_i separately, but that it cannot be used to explain the entire model as a whole.

Interpreting coefficients in multiple regression

The way we interpret the intercept and slopes in a multiple regression model is very similar to what we saw for the simple linear regression model in the previous chapter. We just need to make slight adjustments to account for the fact that we now have more than one x.

For the regression model $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$,

- The slopes $b_i (i = 1, 2, \ldots, k)$ represents the predicted change in the response y for a one unit increase in the predictor x_i, *holding all other predictors constant*.
- The intercept b_0 represents the predicted value of y when $x_1 = x_2 = \cdots = x_k = 0$. Similar to simple linear regression, in some situations it may not make sense to have all x-values be 0 and in these cases we do not interpret b_0.

Class Example 4.5.3 *Cherry Trees*

Timber yield is approximately equal to the volume of a tree, however, this value is difficult to measure without first cutting the tree down. Instead, other variables, such as height and diameter, may be used to predict a tree's volume and yield. Researchers wanting to understand the relationship between these variables for black cherry trees collected data from 31 such trees in the Allegheny National Forest, Pennsylvania. Height, is measured in feet, diameter in inches (at 54in above the ground), and volume in cubic feet (Hand, 1994). The multiple linear regression output obtained via Jamovi are provided below.

Model Fit Measures

Model	R	R^2
1	0.974	0.948

Note. Models estimated using sample size of N=31

Omnibus ANOVA Test

	Sum of Squares	df	Mean Square	F	p
diam	4782.974	1	4782.974	317.413	<.001
height	102.381	1	102.381	6.794	0.014
Residuals	421.921	28	15.069		

Note. Type 3 sum of squares

[3]

Model Coefficients - volume

Predictor	Estimate	SE	t	p
Intercept	−57.988	8.638	−6.713	<.001
diam	4.708	0.264	17.816	<.001
height	0.339	0.130	2.607	0.014

Figure 42: Jamovi output for multiple linear regression for the black cherry tree data.

a) Provide the estimated model equation relating height and diameter to the volume of the cherry trees.

b) Interpret each of the slope parameter estimates.

c) Give the coefficient of determination and interpret

d) Determine if at least one of height or diameter is a useful predictor of volume using an overall F test.

Testing individual terms in a model

If the overall F test reveals that at least one of the predictors is needed in the multiple regression model, the next step is to identify which variables are useful. This is done through separate/individual t tests. One important thing to note is that these tests assess the importance of the predictor in question *after the other predictors are in the model*. So the test is really asking "If all the other predictors are in the model, is this predictor useful in addition?" The hypothesis can be written as follows, where i ranges from 1 to k:

H_0: $\beta_i = 0$ (Predictor x_i is not needed when all other predictors are in the model)
H_A: $\beta_i \neq 0$ (Predictor x_i is useful in addition to all other predictors are in the model)

We have seen for simple linear regression that we have output that provides estimates of the coefficients along with the standard error of the slope, the t test statistic, and p-value. For multiple linear regression we are provided these same quantities for each predictor. So along with estimated slopes, we have SE_{b_i} and test statistic of

$$t = \frac{b_i}{SE_{b_i}}$$

which has a t distribution with $n - k - 1$ degrees of freedom.

Class Example 4.5.4 *Cherry Trees, continued*

Using the Jamovi output, determine separately which predictors are needed in the multiple regression model. For each provide the null and alternative hypothesis, test statistic, p-value, and a brief statement as to whether or not the variable is needed.

Multicollinearity

Multicollinearity refers to when two (or more) of the predictors are associated with each other.

In some situations a predictor may look strongly related to the response as evidenced by a scatterplot, yet not be needed in the model. Why would this happen? This is likely due to an association between the predictor variables (i.e. one x-variable associated to another x-variable). This is referred to as *multicollinearity*. Keep in mind that in a multiple regression, each coefficient's estimate and significance depend on the other explanatory variables included in the model. If these explanatory variables are associated, we may not need some of them, as they convey the same information about the predictor that is already captured by the associated predictor.

Class Example 4.5.5 *Cherry Trees, continued*

A scatterplot of the predictors height and diameter is provided below. Does this plot suggest multicollinearity is present? How does this impact the individual t tests done previously?

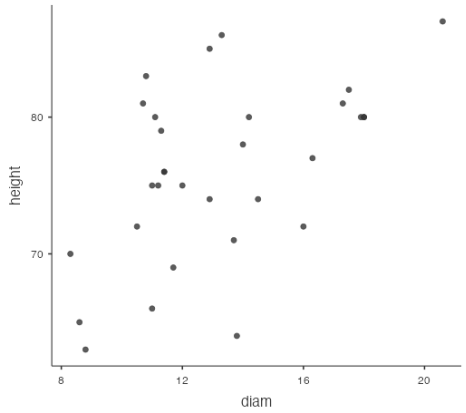


Figure 43: Scatterplot of the height and diameter of the cherry trees.

Summary

This chapter introduced the ANOVA-for-regression framework. The total sum of squares, SST, decomposes as $SST = SSR + SSE$, splitting the response's variability into a piece the model explains and a piece it does not. Organizing this decomposition alongside degrees of freedom, mean squares, and an F-statistic gives the ANOVA table for regression. The F-statistic tests the model as a whole; in simple regression it satisfies $F = t^2$ and gives the same p-value as the slope t-test, but the framework generalizes to multiple regression where the two views come apart. Along the way we reinterpreted $R^2 = SSR/SST$ as the fraction of total variability the model explains.

Multiple regression extends the simple regression model to two or more predictors, fit by least squares. The interpretation of each coefficient becomes "the mean change in y per one unit change in x_j holding the other predictors fixed", a qualifier that separates multiple regression's slope from the simple-regression slope that would absorb the other predictors' effects. Software reports coefficient estimates alongside a coefficient table, an ANOVA table with a model-level F-test, and the goodness-of-fit statistic R^2. Individual t-tests for each coefficient are conditional on the other predictors being in the model. Multicollinearity, the overlapping information between predictors, is a new concern that only arises with multiple predictors. The standard check is a scatterplot of each pair of predictors.

5.1 Probability Rules

Probability forms the foundation of statistics, and you are probably already aware of many of the ideas we will discuss in this chapter. However, formalizing these concepts is likely new. While this chapter provides the theoretical foundation for the inference methods you have already been using, it also offers practical tools for reasoning about uncertainty. We will define probability, develop rules for combining probabilities, and explore conditional probability and independence, all concepts that underpin the inference methods from Parts II and III.

Key Concepts

- Compute the probability of events if outcomes are equally likely
- Identify when a probability question is asking for A and B, A or B, not A, or A if/given B
- Use the complement, additive, multiplicative, and conditional rules to compute probabilities of events
- Recognize when two events are disjoint

Defining Probabilities

Statistics is built on probability, and probability gives us the language to describe uncertainty. We begin with a few simple examples that may feel familiar.

Class Example 5.1.1 *Roll a fair die*

A "die", the singular of "dice", is a cube with six faces numbered 1, 2, 3, 4, 5, and 6.

a) What is the chance of getting a 1 when rolling a fair die?

b) What is the chance of getting a 1 or 2 on the next roll?

c) What is the chance of not rolling a 2?

We use *probability* to describe and understand apparent randomness. Recall that a *population* is the collection of all individuals or items under consideration. An individual could refer to a person, a playing card, or whatever object we are interested in. We typically refer to a population when referencing sampling. However, when we talk about experiments, we use the phrase *sample space*. The *sample space* is the set of all possible non-overlapping outcomes for a random experiment, and is denoted Ω. *Events* can be simple (e.g., the outcome of a single flip of a coin) or complex combinations of outcomes (e.g., the results from 10 flips of a coin). We have three important rules for probability:

The *probability* of an event is the proportion of the times the event would occur if we observed the random process an infinite number of times. Probabilities are *always* between 0 and 1.

Sample space, Ω, is the set of all possible outcomes.

An *event* is a subset of the sample space, that is, a collection of one or more outcomes.

We use the notation $P()$ to mean *probability of*.

We use the notation $N()$ to mean *number of outcomes in* an event.

When outcomes are equally likely, the probability of event is the number of outcomes in the event divided by the total number of outcomes.

P.1) Any probability must be between 0 and 1 (inclusive)

P.2) When outcomes are equally likely, the probability of the event E is given by

$$P(E) = \frac{N(E)}{N(\Omega)}$$

P.3) The sum of the probabilities for all of the experimental outcomes must equal 1

Class Example 5.1.2 *Roll a fair die, continued*

Define the sample space for rolling a die.

We can visualize the sample space along with several events. Let A represent the event that the die roll results in a 1 or a 2, so $A = \{1, 2\}$. Let B represent the event that the die roll is a 4 or 6, so $B = \{4, 6\}$.

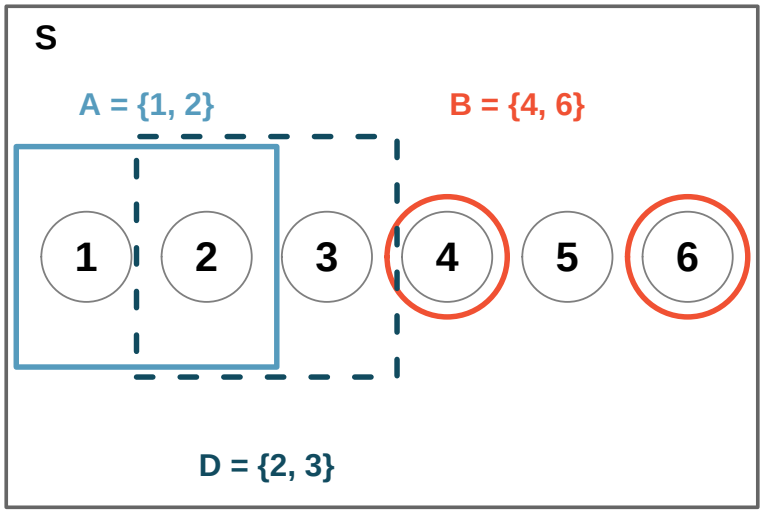


Figure 44: Three events for die outcomes. A and B are disjoint since they share no outcomes.

The complement rule

Let $D = \{2, 3\}$ represent the event that the outcome of a die roll is a 2 or 3. Then the *complement* of D, denoted D^c, represents all outcomes in the sample space that are *not* in D.

The *complement* of a condition, written "not A" or A^c, includes events in which the condition does not happen.

Class Example 5.1.3 *Roll a fair die, continued*

a) Write out the event for the complement of D and find its probability.

b) What is $P(D) + P(D^c)$?

P.4) P(not A) = $P(A^c)$ = Probability A does not occur is given by

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

Class Example 5.1.4 *Two rolls of a fair die*

Suppose a fair, six-sided die is rolled *twice*. Determine the following.

a) $N(\Omega)$

b) $N(A)$, where A is the event that the sum of the two rolls is 5

c) $P(B)$, where B is the event that the two rolls are the same

d) $P(C)$, where C is the event that the sum of the two rolls is even

e) $P(D)$, where D is the event that the two rolls are not the same

The addition rule

Many of the terms we used for describing the relationship between two categorical variables is also used when describing events matching two (or more) conditions. In addition to the ideas of *joint*, *marginal*, and *conditional* proportions, we will use terms like *intersection*, and *union* of conditions.

- P(A and B) = $P(A \cap B)$ = Probability conditions A and B both occur.
- P(A or B) = $P(A \cup B)$ = Probability either condition A or condition B (or both) occur.

Two outcomes or events are called *disjoint*, or *mututally exclusive*, if they cannot both happen at the same time. For instance, event A and event B in the die rolling example are disjoint since they cannot both occur at the same time.

The *joint* or *intersection* of two conditions, written "A and B" or $A \cap B$, includes events that satisfy *both* conditions.

The *union* of two conditions, written "A or B" or $A \cup B$, includes events that satisfy *either or both* conditions.

The *conditional occurrence* of condition A given the occurrence of condition B is written "A if B" or $A \mid B$.

Events are *disjoint* if they cannot occur at the same time.

P.5) If A_1 and A_2 are disjoint events, then:

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

More generally, if $A_1, A_2, \ldots, A_k$ are all mutually disjoint, then:

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

Class Example 5.1.5 *Roll a fair die, revisited*

We are interested in the probability of rolling a 1, 4, or 5.

a) Explain why the outcomes 1, 4, and 5 are disjoint.

b) Apply the addition rule to determine $P(1 \text{ or } 4 \text{ or } 5)$.

When events are *not* disjoint, we cannot simply add their probabilities, doing so would double-count the outcomes they share.

P.6) Additive Rule:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

For disjoint events $P(A \text{ or } B) = 0$. If we replace $P(A \text{ or } B)$ with 0 in the addition rule above, we see that for disjoint events

$$P(A \cup B) = P(A) + P(B)$$

Class Example 5.1.6 *Two rolls of a fair die, revisited*

We had the following events: A = sum of the two rolls is 5, B = the two rolls are the same, C = the sum of the two rolls is even.

a) Circle the pairs of events that are disjoint: A and B; A and C; B and C.

b) Find the probability of the union for all three pairs listed in part a).

Class Example 5.1.7 *Deck of Cards*

Consider a standard deck of 52 cards. In a standard deck, there are 2 colors: red and black. The red cards are split into two suits: diamonds and hearts. The black cards are split into clubs and spades. Each suit has 13 cards 2 through 10, jack, queen, king, ace. Let A be the event "the card is a diamond" and B "the card is a face card (jack, queen, or king)." Find the probability that a random card is a diamond or a face card.

Venn diagrams

Venn diagrams are visual displays that are useful for investigating *intersections*, *unions*, and *complements*. These diagrams are constructed by first drawing a rectangle, representing the sample space. Within this rectangle, overlapping circles are drawn for each of the conditions of interest.

Example 19.8 Types of classes

If 60% of college freshmen take a math class, 30% of college freshmen take a history class, and 12% of college freshmen take both a college math class and history class, what is the probability that a freshman selected at random is in a math or history class?

Example 19.9 La Crosse pets

Suppose 35% of households in La Crosse have a dog, 28% of households in La Crosse have a cat, and 52% of households in La Crosse have a dog or a cat. What is the probability that a La Crosse household selected at random has a dog and a cat?

Conditional Probabilities

The *conditional* occurrence of A given the occurrence of B is written "A if B" or "A given B" or A j B.

Conditional probability is used in situations where we find the probability of one event occurring, if/given we know another event has occurred. To find the conditional probability of an event A occurring if/given B occurs, we use

P.7) Conditional Probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Example 19.10 Types of classes, revisited

If you know a selected student is in a math class, what is the probability they are in a history class?

A *marginal probability* is a probability based on a single value, without regard to the other variables. A *joint probability* is the probability of two or more outcomes occurring together.

The multiplication rule

If we rearrange the formula in P.7 we can solve for $P(A \cap B)$ and this is know as the multiplication rule.

P.8) Multiplication Rule:

$$\begin{aligned} P(A \cap B) &= P(B|A) * P(A) \\ &= P(A|B) * P(B) \end{aligned}$$

Example 19.11 Baseball games

A Little League baseball team has a projected 80% chance of winning their first game in a tournament. If they win their first game, they are projected to win their second game with 60% probability. What is the probability that they win both their first and second game?

Events A and B are *independent* if knowing that one of the events has occurred does not change the probability of the other event occurring. Mathematically, events A and B are independent whenever

P.9) Independent events satisfy:

$$P(A|B) = P(A)$$
$$P(B|A) = P(B)$$

When events are independent, the multiplication rule becomes the more simplified form of

P.10) Multiplication Rule for Independent Events:

$$P(A \cap B) = P(A) \times P(B)$$

Class Example 5.1.8 *Flipping a fair coin*

Suppose we have a fair coin.

a) Suppose we flip the coin two times, are the flips independent of one another?

b) Suppose that the first three flips were HHT. What is the probability the 4^{th} flip is heads?

c) Suppose that the first three flips were TTT. What is the probability the 4^{th} flip is tails?

d) What is the probability of the event the 4 flips were HHTH?

e) What is the probability of the event the 4 flips were TTTT?

Example 19.13 Bin of marbles

A jar contains 20 blue marbles, 30 red marbles, and 50 yellow marbles. You select two marbles, one at a time, without looking. Find the following probabilities.

a) The probability that the second marble is red if the first marble was blue. You *have not* replaced the blue marble before drawing the second marble.

b) The probability that the second marble is red if the first marble was blue. You *have* replaced the blue marble before drawing the second marble.

c) The probability that the first marble is blue and the second marble is red. You *have not* replaced the blue marble before drawing the second marble.

d) The probability that the first marble is blue and the second marble is red. You *have* replaced the blue marble before drawing the second marble.

Class Example 5.1.9 *Relationship status and class year revisited*

Recall the following example from Unit 1. 169 college students were asked about relationship status and class year (lower=freshman/sophomore, upper=junior/senior). The results are given in the table below. For each of the following, assume that a student is chosen at random from this sample.

	Upper	**Lower**	**Total**
In a relationship	32	10	42
It's complicated	12	7	19
Single	63	45	108
Total	107	62	169

Contingency table for relationship status and class year.

a) What is the probability the student is in a relationship? Is this a joint, marginal, or conditional probability?

b) What is the probability the student is in a relationship given the student an upperclassmen? Is this a joint, marginal, or conditional probability?

c) What is the probability the student is an upperclassmen given the student is in a relationship? Is this a joint, marginal, or conditional probability?

d) What is the probability the student is a lowerclassmen and single? Is this a joint, marginal, or conditional probability?

Counting Rules

The *basic counting rule* provides the number of possible outcomes of a random experiment in which several actions are performed.

A *combination* is any unordered collection elements chosen from a collection of elements.

It is often useful to be able to count (enumerate) the number of possible outcomes of a random experiment. Understanding a couple of simple counting rules can help to determine probabilities of events.

C.1) *Basic Counting Rule:* If event A consists of r actions that are performed in a specific order, with m_1 possibilities for the first action, m_2 for the second, etc. Then the total outcomes in A is

$$N(A) = m_1 * m_2 * \cdots * m_r.$$

C.2) *Arranging items:* If event A has n items, then the number of ways we can arrange the items in A is

$$n! = n * (n-1) * (n-2) * \cdots * 2 * 1.$$

C.3) *Combinations:* If event A has n items, then the number of ways we can choose k items from A is

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

Class Example 5.1.10 *True/False Quiz*

Suppose that a professor gives her students a 10 question True/False quiz.

a) How many ways are there to answer the quiz?

b) How many ways are there to answer exactly 5 questions correctly?

Class Example 5.1.11 *Cola Taste Test*

Suppose that three glasses are filled with the same cola, and then labeled C, D, and P. A randomly selected person is asked to order the three glasses by preference (this person is unaware that all three glasses contain the same cola).

a) How many ways can the cola preferences be ranked?

b) What is the probability that C is ranked first?

c) What is the probability that C is ranked first and D is ranked last?

Summary

In this chapter, we developed the mathematical rules of probability:

- **Probability** is the long-run proportion of times an outcome occurs
- The **sample space** Ω is the set of all possible outcomes; an **event** is a subset of Ω.
- **Complement rule**: $P(A) = 1 - P(A^c)$.
- **Addition rule**: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$. For disjoint events, $P(A \text{ and } B) = 0$.
- **Conditional probability**: $P(A \mid B) = P(A \text{ and } B)/P(B)$.
- **Multiplication rule**: $P(A \text{ and } B) = P(A \mid B) \times P(B)$. For independent events, $P(A \text{ and } B) = P(A) \times P(B)$.
- Events are **independent** if knowing one occurred does not change the probability of the other.
- It is often useful to be able to count (enumerate) the number of possible outcomes

5.2 Binomial Distribution

How many of 400 randomly sampled adults are left-handed? How many of 20 randomly selected parts from an assembly line are defective? These "count of successes" questions arise constantly in statistics and are answered by the ***binomial distribution****. In this chapter, we identify the four conditions that define a binomial setting, develop the formula for computing binomial probabilities, and introduce the mean and standard deviation.*

Key Concepts

- Identify a binomial random variable
- Compute probabilities for a binomial random variable
- Compute the mean and/or standard deviation for a binomial random variable

The binomial setting

A *binomial random variable* measures the number of successes in a fixed number of trials.

Random variables often follow simple, predictable patterns. In these cases, calculating probabilities and the mean or standard deviation are much simpler and do not require a table to be created. One type of random variable with a simple pattern is called a "Binomial Random Variable." Suppose we call the occurrence of an event a *success*. "Success" does not necessarily indicate the outcome was positive (e.g., if we are investigating cancer in a sample of mice, success could be defined as the presence of cancer). A *binomial random variable* measures the number of successes in a fixed number of trials.

In order to be considered a binomial random variable, the following properties must hold:

1. *Fixed number of trials:* The number of trials n is fixed in advanced
2. *Binary outcomes*: Each trial results in one of two outcomes, which we label "success" and "failure."
3. *Independence*: The trials are independent, the outcome of one trial does not affect the outcome of any other trial.
4. *Constant probability*: The probability of success p is the same for every trial.

Class Example 5.2.1 *Is it binomial?*

For each of the following, determine if the random variable is binomial. If it is, determine n and p.

a) On average, 25% of La Crosse residents recycle. 500 residents of La Crosse were asked if they recycle. A is the number of 'yes' answers in the sample.

b) A box contains 50 marbles, 35 green and 15 white. Fifteen marbles are selected without replacement. B is the number of white marbles in the sample.

c) An American roulette wheel consists of 18 red, 18 black, and 2 green numbers. Suppose you observed red on 10 consecutive spins. J is the number of additional spins needed until you observe black.

d) D is the number of 5s in ten rolls of a fair die.

e) Customers arrive at Alice's with a rate of 5 per hour. E is the number of customers that enter Alice's between 2 A.M. and 4 A.M.

Computing binomial probabilites

A health insurance company found that 70% of the people they insure stay below their deductible in any given year. Consider a random sample of four individuals. Let's find the probability that exactly 1 of 4 randomly selected insured individuals exceeds the deductible (so 3 "successes" — staying below — and 1 "failure").

Consider the specific scenario where person A exceeds and persons B, C, D do not:

$$P(A = \text{exceed}, B = \text{not}, C = \text{not}, D = \text{not}) = (0.3)(0.7)(0.7)(0.7) = (0.7)^3(0.3)^1 = 0.103$$

But there are four such scenarios — any one of the four people could be the one who exceeds. Each scenario has the same probability. So:

$$P(\text{exactly 3 successes in 4 trials}) = 4 \times (0.7)^3(0.3)^1 = 0.412$$

The general pattern is:

$$P(X = k) = [\text{number of arrangements}] \times [\text{probability of one arrangement}]$$

The number of ways to arrange k successes in n trials is given by a *combination*:

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

For our example: $\binom{4}{3} = \frac{4!}{3!\cdot 1!} = \frac{24}{6\times 1} = 4$, confirming the four arrangements we identified.

For a binomial random variable X with n trials and probability of success p,

- $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ for k = 0, 1, 2, ..., n.
- Mean: $\mu = E[X] = np$
- Standard Deviation: $\sigma = sd(X) = \sqrt{np(1-p)}$

Class Example 5.2.2 *Testing ESP*

To test for ESP, we have 4 cards, each with different symbols. These cards are shuffled and one is chosen at random (symbol down). Each time, you guess the symbol on the card. Suppose that you repeat this test 8 times and you do not have ESP.

a) Let X be the number of times you guess correctly. Is X binomial? If so, what are n and p?

b) What is the expected value of X?

c) What is the standard deviation of X?

d) To be certified as having ESP, you need to correctly identify at least 6 cards of the 8 drawn. What is the probability you get certified?

Class Example 5.2.3 *Customer service*

Joan takes 10 phone calls per day from customers, and, on average, 30% of these callers are angry. If 7 or more of these callers in a given day are angry, Joan has a bad day.

a) Let C be the number of angry callers in a day. Is C binomial? If so, what are n and p?

b) How many angry callers will Joan speak to on average each day? What is the standard deviation for the number of angry callers?

c) What is the probability that Friday is a good day?

d) What is the probability that 3 of 5 days in her workweek are good?

Class Example 5.2.4 *Jordan Love*

In the 2023 NFL season, Jordan Love threw 579 passes. A fair-weather fan happened to only see a few games that season. The fan only saw Jordan Love throw 33 times, and of those 18 were completed (caught).

a) What would the fan estimate is Jordan Love's probability to complete a pass?

b) Based on the probability in part (a), how many passes would the fan expect Jordan Love to complete in the next 10 throws?

Summary

- The *binomial distribution* counts the number of successes in n independent trials, each with the same probability of success p.
- *Four conditions*: fixed n, binary outcomes, independence, constant p.
- *Binomial probability formula*: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- *Mean and standard deviation*: $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

5.3 Data Sources

The datasets used in this course pack come from the following sources. We gratefully acknowledge the original authors and data collectors.

Lock5Data R Package (GPL-2)

The following datasets are from the **Lock5Data** R package, which accompanies *Statistics: Unlocking the Power of Data* by Lock, Lock, Lock, Lock, and Lock (Wiley). The Lock5Data package is distributed under the GPL-2 license.

- **CAOSExam.csv** — Pre-test and post-test scores from a Comprehensive Assessment of Outcomes in Statistics (CAOS) exam. 10 students, 3 variables.
- **FloridaLakes.csv** — Water quality measurements from 53 Florida lakes. Original source: Lange, Royals, and Connor (1993), *Journal of Environmental Quality*.
- **PulseRates.csv** — Resting pulse rates for 10 students before and after exercise.
- **SleepStudy.csv** — Sleep habits and academic performance for 253 college students. Original source: Onyper, Thacher, Gilbert, and Gradess (2012), *Teaching of Psychology*.
- **USStates.csv** — State-level data (50 US states) compiled from US government sources including the Census Bureau and Bureau of Labor Statistics.

R Packages (GPL-2+)

- **Baumann.csv** — Reading comprehension scores from a study comparing three teaching methods (Basal, DRTA, Strat). 66 observations. From the **carData** R package (GPL-2+). Original source: Baumann and Jones (1992).
- **InsectSprays.csv** — Insect counts for six different spray treatments. 72 observations. Built-in R dataset (GPL-2/3). Original source: Beall (1942).
- **cherryblossom.csv** - Race results of the Cherry Blossom Run, which is an annual road race that takes place in Washington, DC. From the **cherryblossom** R package.

openintro

- **loan50.csv** - This dataset represents thousands of loans made through the Lending Club platform, which is a platform that allows individuals to lend to other individuals.
- **mammals.csv** - This dataset includes data for 39 species of mammals distributed over 13 orders.
- **migrane.csv** - Experiment involving acupuncture and sham acupuncture (as placebo) in the treatment of migraines.
- **fastfood.csv** - Nutrition amounts in 515 fast food items.
- **penny_ages.csv** - Sample of 648 pennies and their ages taken in 2004.
- **possum.csv** - Data representing possums in Australia and New Guinea.

Public Domain / Custom

- **HollywoodMovies2023.csv** — Opening weekend revenue and world gross for six 2023 Hollywood films. Created for this course pack from publicly available box office data (Box Office Mojo).
- **MastersWinners.csv** — Winning scores at the Masters Tournament (1934–2024). 87 observations. Compiled from public tournament records.

- **NBASalaries2022.csv** — NBA player salaries for the 2022–23 season. 467 observations. Synthetic data generated to match the published salary distribution.
- **Salaries.csv** — Monthly sales figures (in dollars) for 15 salespeople under three pay structures (Commission, Fixed Salary, Commission Plus Salary). Synthetic data created for this course pack to illustrate one-way ANOVA and post-hoc comparisons.