STAT 216 Coursepack

Spring 2021
Montana State University

Melinda Yager, Jade Schmidt, Dr. Stacey Hancock

Contents

Preface

This coursepack accompanies the textbook for STAT 216: Introduction to Statistics at Montana State University, which can be found at https://mtstateintrostats.github.io/IntroStatTextbook/. The syllabus for the course (including the course calendar), data sets, and links to D2L Brightspace, Gradescope, and the MSU RStudio server can be found on the course webpage: https://math.montana.edu/courses/s216/. Videos assigned in the course calendar and other notes and review materials are linked in D2L.

Each of the activities in this workbook is designed to target specific learning outcomes of the course, giving you practice with important statistical concepts in a group setting with instructor guidance. In addition to the in-class activities for the course, the coursepack includes reading guides to aid in taking notes while you complete the required readings and videos. Bring this workbook with you to class each week, and take notes in the workbook as you would your own notes. A well-written completed workbook will provide an optimal study guide for exams!

The activities in this coursepack are broken into three sections: pre-class, in-class, and after class. Read through the introduction for each activity and complete the pre-class questions before attending class each week. In class, you will work through the in-class section with your group and instructor. After class, you will complete the out-of-class part of the activity.

STAT 216 is a 3-credit blended course. Rather than meeting for a total of 150 structured minutes in class per week, students meet with their instructor and cohort of classmates for 50 minutes of class per week. The other 100 minutes typically spent in class are instead spent outside of class watching instructor video lectures, reading the textbook, working through case studies, and participating in online discussion with your classmates. This structure serves two purposes: (1) enhance the safety of our community during the COVID-19 pandemic, and (2) provide additional flexibility for students to create their own schedule and make their own decisions on how they learn best.

In our experience, it takes six to nine hours per week outside of class to achieve a good grade in STAT 216 – this means six to nine hours outside of the 150 minutes of time set aside for learning course material each week. By "good" we mean at least a C because a grade of C- or below does not count toward fulfilling degree requirements. Many of you set your goals higher than just getting a C, and we fully support that. You need roughly nine hours per week to review past activities, read feedback on previous assignments, complete current assignments, and prepare for the next day's class. A typical week in the life of a STAT 216 student looks like:

- *Prior to class meeting*:
    - Read assigned sections of the textbook, using the provided reading guides to take notes on the material.
    - Watch assigned videos on that week's content, pausing to take notes and answer video quiz questions.
    - Read through the introduction to the week's in-class activity and complete the pre-class questions.
    - Read through the week's homework assignment and note any questions you may have on the content.
- *During class meeting*:
    - Work through the in-class activity with your classmates and instructor, taking detailed notes on your answers to each question in the activity.
- *After class meeting*:
    - Complete the out-of-class part of the activity, plus any additional parts of the activity you did not complete in class.
    - Review the posted activity solutions and wrap-up videos, and take notes on key points.
    - Finish watching any remaining assigned videos or readings for the week.
    - Read through the week's case study and post case study discussion posts on D2L.
    - Complete the week's homework assignment.

Spring 2021 Calendar of In-Class Activities

This calendar only lists the in-class activities and exams each week. For required readings and videos, due dates for assignments and case study posts, and project deadlines, refer to your cohort's calendar at: https://mtstateintrostats.github.io/Syllabus/#Course_calendars

Week	Activity	Day	Date	Activity
1	1	M	1/11	Martian Alphabet
1	1	W	1/13	Martian Alphabet
1	1	F	1/15	Martian Alphabet
2	2	M	1/18	Study Design (*No class*)
2	2	W	1/20	Study Design
2	2	F	1/22	Study Design
3	3	M	1/25	What's the Probability?
3	3	W	1/27	What's the Probability?
3	3	F	1/29	What's the Probability?
4	4	M	2/1	IMDb Movie Reviews
4	4	W	2/3	IMDb Movie Reviews
4	4	F	2/5	IMDb Movie Reviews
5	4	M	2/8	Movie Profits
5	5	W	2/10	Movie Profits
5	5	F	2/12	Movie Profits
6	6	M	2/15	Handedness of Male Boxers — Testing (*No class*)
6	6	W	2/17	Handedness of Male Boxers — Testing
6	6	F	2/19	Handedness of Male Boxers — Testing
7	—	M–F	2/22–2/26	**Exam 1**
8	7	M	3/1	Handedness of Male Boxers — Estimation
8	7	W	3/3	Handedness of Male Boxers — Estimation
8	7	F	3/5	Handedness of Male Boxers — Estimation
9	8	M	3/8	Winter Sports Helmet Use and Head Injuries — Testing
9	8	W	3/10	Winter Sports Helmet Use and Head Injuries — Testing
9	8	F	3/12	Winter Sports Helmet Use and Head Injuries — Testing
10	9	M	3/15	Winter Sports Helmet Use and Head Injuries — Estimation
10	9	W	3/17	Winter Sports Helmet Use and Head Injuries — Estimation
10	9	F	3/19	Winter Sports Helmet Use and Head Injuries — Estimation
11	10	M	3/22	COVID-19 and Air Pollution
11	10	W	3/24	COVID-19 and Air Pollution
11	10	F	3/26	COVID-19 and Air Pollution
12	11	M	3/29	Weather Patterns and Record Snowfall
12	11	W	3/31	Weather Patterns and Record Snowfall
12	11	F	4/2	Weather Patterns and Record Snowfall (*No class*)
13	12	M	4/5	Moneyball
13	12	W	4/7	Moneyball
13	12	F	4/9	Moneyball
14	—	M–F	4/12–4/16	**Exam 2**
15	—	M–F	4/19–4/23	Looking Ahead
Finals	—	M–F	4/26–4/30	Group Project Presentations

Basics of Data

1.1 Reading Guide: Basics of Data

Sections 1.1 (Case study) and 1.2 (Data basics)

Videos

- Stat 216 Course_Tour
- Instructor bio
- 1.2.1_1.2.2
- 1.2.3_1.2.4_1.2.5

Vocabulary

Data:

Summary statistic:

Case/Observational unit:

Variable:

    Quantitative variable:

    Discrete variables:

        Examples of discrete variables using the `County` data:

    Continuous variables:

        Examples of continuous variables using the `County` data:

Example of a number which is NOT a numerical variable:

Categorical variable:

    Ordinal variable:

Example of an ordinal variable using the `County` data:

Nominal variable:

Examples of nominal variables using the `County` data:

Note: Ordinal and nominal variables will be treated the same in this course. We recommend taking more statistics courses in the future to learn better methods of analysis for ordinal variables.

Data frame:

Scatterplot:

Each point represents:

Positive association:

Negative association:

Association or Dependent variables:

Independent variables:

Explanatory variable:

Response variable:

Observational study:

Experiment:

Placebo:

Notes

Big Idea: Variability is inevitable! We would not expect to get *exactly* 50 heads in 100 coin flips. The statistical question then is whether any differences found in data are due to random variability, or if something else is going on.

The larger the difference, the **less we believe the difference was due to chance.**

In a data frame, rows correspond to _____

and columns correspond to _____.

How many types of variables are discussed? Explain the differences between them and give an example of each.

True or False: A pair of variables can be both associated AND independent.

True or False: Given a pair of variables, one will always be the explanatory variable and one the response variable.

True or False: If a study does have an explanatory and a response variable, that means changes in the explanatory variable must **cause** changes in the response variable.

True or False: Observational studies can show a naturally occurring association between variables.

Example (Section 1.1 - Case study: Using stents to prevent strokes)

1. What is the principle question the researchers hope to answer? (We call this the **research question**.)

2. When creating two groups to compare, do the groups have to be the same size (same number of people in each)?

3. What are the cases or observational units in this study?

4. Is there a clear explanatory and response variable? If so, name the variable in each role and determine the type of variable (discrete, continuous, nominal, or ordinal).

5. What is the purpose of the control group?

6. Is this an example of an observational study or an experiment? How do you know?

7. Consider Tables 1.1 and 1.2. Which table is more helpful in answering the research question? Justify your answer.

8. Describe in words what is shown in Figure 1.1. Specifically, compare the proportion of patients who had a stroke between the treatment and control groups after 30 days as well as after 365 days.

9. Given the notion that the larger the difference between the two groups (for a given sample size), the less believable it is that the difference was due to chance, which measurement period (30 days or 365 days) provide stronger evidence that there is an association between stents and strokes, or that the differences are not due to random chance?

10. This study reported finding evidence that stents *increase* the risk of stroke. Does this conclusion apply to all patients and all stents?

11. This study reported finding evidence that stents *increase* the risk of stroke. This conclusion implies a causal link between stents and an increased risk of stroke. Is that conclusion valid? Justify your answer.

1.2 Activity: Martian Alphabet

1.2.1 Learning outcomes

- Describe the statistical investigation process.
- Identify observational units, variables, and variable types in a statistical study.

1.2.2 Terminology review

Statistics is the study of how best to collect, analyze, and draw conclusions from data. Today in class you will be introduced to the following terms:

- Observational units or cases
- Variables: categorical or quantitative
- Proportions
- Graphs: frequency bar plot and relative frequency bar plot
- Distribution

For more on these concepts, read Sections 1.2 and 2.1 in the textbook.

1.2.3 Can you read "Martian"?

How well can humans distinguish one "Martian" letter from another? In this week's activity, we'll find out. When shown the two Martian letters, Kiki and Bumba, write down whether you think Bumba is on the left or on the right.

1. Were you correct or incorrect in identifying Bumba?

Steps of the statistical investigation process

Step 1: The first step of any statistical investigation is to *ask a research question*. In this study the research question is: Can we as a class read Martian? (We will refine this later on!).

Step 2: To answer any research question, we must *design a study and collect data*. For our question, the study consists of each student being presented with two Martian letters and asking which was Bumba. Your responses will become our observed data that we will explore.

Observational units or **cases** are the subjects data are collected on. In a spreadsheet of the data set, each row will represent a single observational unit.

2. What are the observational units in this study?

3. How many students are in class today? This is the **sample size**.

A **variable** is information collected or measured on each observational unit or case. Each column in a data set will represent a different variable. Today we are only measuring one variable on each observational unit.

4. Identify the variable we are collecting on each observational unit in this study, i.e., what are we measuring on each student? *Hint*: Your answer to question 1 is the outcome for the variable measured on one observational unit.

We will look at two types of variables: **quantitative** and **categorical** (see Figure 1.1).

Quantitative variables are numerical measurements that can be discrete (whole, non-negative numbers) or continuous (any value within an interval). The number of pets one owns would be a discrete variable as you can not have a partial pet. GPA would be a continuous variable ranging from 0 to 4.0.

The outcome of a categorical variable is a group or category such as eye color, state of residency, or whether or not a student lives on campus. Categorical variables with a natural ordering are considered ordinal variables while those without a natural ordering are considered nominal variables. All categorical variables will be treated as nominal for analysis in this course.
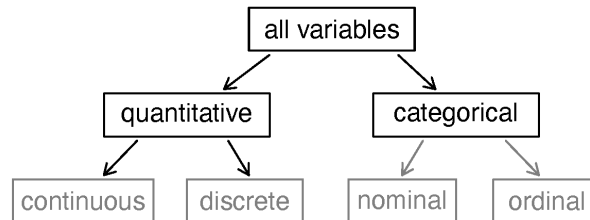
Figure 1.1: Types of variables.

5. Is the variable identified in question 4 categorical or quantitative?

Step 3: Once we have collected data, the next step is to *summarize and visualize the data.*

6. How many people in your class were correct in identifying Bumba? Using the class size from question 3, calculate the proportion of students who correctly identified Bumba.

$$\text{proportion} = \frac{\text{number of students who correctly identified Bumba}}{\text{total number of students}}$$

The proportion in question 6 is called a **summary statistic**—a single value that summarizes the data set. It is important to note that a variable is different than a summary statistic. A *variable* is measured on a *single observational unit* while a summary statistic is calculated from a group of observational units. For example, the variable "whether or not a student lives on campus" can be measured on each individual student. In a class of 50 students we can calculate the proportion of students who live on campus, the summary statistic. Look back and make sure you wrote the variable in question 4 as a variable, NOT a summary statistic.

Looking at the data set and the summary statistic is only one way to display the data. We will also want to create a visualization or picture of the data. A **frequency bar plot** is used to display categorical data as a count or frequency. Since our variable has two levels or outcomes, correct or incorrect, we will create two bars—one for each level.

7. Plot the observed class data using a frequency bar plot. Be sure to add a scale to the *y*-axis.
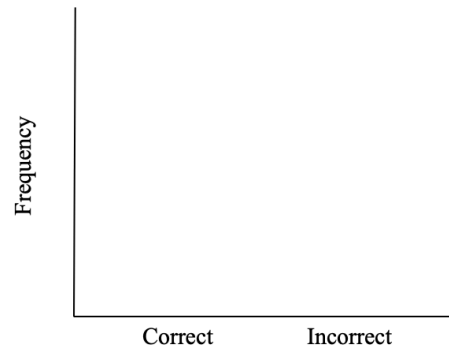
We can also visualize the data as a proportion in a **relative frequency bar plot**. Relative frequency is the proportion calculated for each level of the categorical variable.

8. Plot the observed class data using a relative frequency bar plot. Be sure to add a scale to the y-axis.
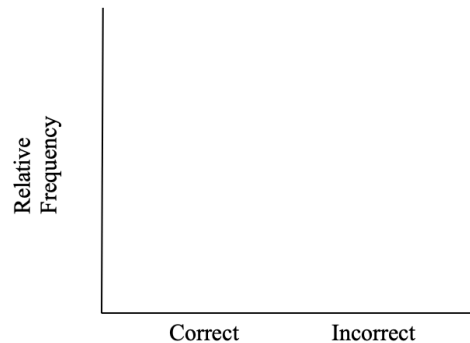
Step 4: The next step is to *use statistical analysis methods to draw inferences from the data.* To answer the research question, we will simulate what *could* have happened in our class given random chance, repeat many times to understand the expected *variability* between different "randomly guessing" classes, then compare our class's observed data to the simulation. This gives us an estimate of how often (or the probability of) the class's result would occur if students were all merely guessing, allowing us to determine if the data provides evidence that we as a class can in fact read Martian.

9. If humans really don't know Martian and are just guessing which is Bumba, what are the chances of getting it right?

How could we use a coin to simulate each student "just guessing" which Martian letter is Bumba?

How could we use coins to simulate the entire class "just guessing" which Martian letter is Bumba?

How many people in your class would you expect to choose Bumba correctly just by chance? Explain your reasoning.

10. Each student will flip a coin one time to simulate your "guess" under the assumption that we can't read Martian. Let Heads = correct, Tails = incorrect. What was the result of your one simulation?

What was the result from your class's simulation? What proportion of students "guessed" correctly in the simulation?

11. If students really don't know Martian and are just guessing which is Bumba, which seems more unusual: the result from your class's **simulation** or the observed proportion of students in your class that were correct (this is your summary statistic from question 6)? Explain your reasoning.

12. While your observed class data is likely far different from the simulated "just-guessing" class, comparing our class data to a single simulation does not provide enough information. The differences seen could just be due to the randomness of that set of coin flips! Let's simulate another class. Each student should flip their coin again. What was the result from your class's second simulation? What proportion of students "guessed" correctly in the second simulation? Create a plot to compare the two simulated results with the observed class result.

13. We still only have a couple of simulations to compare our class data to. It would be much better to be able to see how our class compared to hundreds or thousands of "just-guessing" classes. Since we don't want to flip coins all class period, your instructor will use a computer simulation to get 1000 trials. Fill in the following blanks to describe how we would create a simulation of random guessing with 1000 trials (repetitions).

Probability of correct guesses: _____

Sample size: _____

Number of repetitions: _____

14. Sketch the distribution displayed by your instructor here. Label each axis appropriately.

15. Is your class particularly good or bad at Martian? Use the plot in question 14 to explain your answer.

16. Is it *possible* that we could see our class results just by chance if everyone was just guessing? Explain your reasoning.

17. Is it *likely* that we could see our class results just by chance if everyone was just guessing? Explain your reasoning.

Step 5: The next step in the statistical investigation process is to *communicate the results and answer the research question.*

18. Does this activity provide strong evidence that students were not just guessing at random? If so, what do you think is going on here? Can we as a class read Martian?[1]

Step 6: The final step of any statistical investigation is to *revisit and look ahead.*

19. Can you think of any limitations of our study? Can you think of a new topic that might be of interest based on the results of our study?

---

[1]Reference for "Martian alphabet" is a TED talk given by Vilayanur Ramachandran in 2007. The synesthesia part begins at roughly 17:30 minutes: `http://www.ted.com/talks/vilayanur_ramachandran_on_your_mind`.

1.2.4 Out-of-class activity

Since this class is taught in a blended format, we are only in class one day per week. During class we will complete the in-class activity from the course pack. Outside of class, students will read from the textbook, watch course videos, and complete an out-of-class activity on the other two days of class. To become familiar with the course outline, read through the syllabus, https://mtstateintrostats.github.io/Syllabus/, your day specific cohort calendar, and watch the Stat 216 Course Tour on D2L before answering the following questions.

1. When are the case study discussion posts due on D2L each week?

2. For your cohort, what day and time are the weekly assignments due on Gradescope?

3. For your cohort, when is Exam 1? Exam 2?

1.2.4.1 Introduction to `R`

In Stat 216 we will use the statistical package `R` to analyze data through the IDE (integrated development environment) RStudio. Though it is possible to download `R` and RStudio on your own computer, we will use this program through the MSU RStudio server: https://rstudio.math.montana.edu/.

Read through the preliminaries chapter in the textbook and watch the video "Starting with `R`" before completing the following questions.

The RStudio workflow operates best by the use of "Projects". You should create a separate project for each activity or assignment in this course that requires the use of `R`. To get started with this activity, follow these steps:

- Log onto the RStudio server using your NetID and password: https://rstudio.math.montana.edu/.
- In the top right corner, you will see a dropdown menu next to "Project" that currently says "(None)". Click on this menu and choose "New Project". (Alternatively, you can click the "File" menu in the top left and select "New Project".)
    - A "New Project Wizard" window should pop up: click "New Directory", then click "New Project".
    - Give your project directory a name (e.g., Activity1). *Do not use spaces or other characters in the name.*
    - Click "Browse" and choose a location where you would like to save your project (you can create a new folder if desired). Note that this location is on your server account, not on your computer.
    - Leave all other boxes unchecked, and click "Create Project". (Now, if you click on the home icon in the top right, you will see your RStudio account, and the project should be listed under "Projects".)
- Download the Martian Alphabet `R` script file from D2L.
- Click "Upload" in the "Files" tab in the bottom right window of RStudio. Click "Choose File", and navigate to the folder where the Martian Alphabet `R` script file is saved. Then click "Open"; then click "Ok".
- You should see the uploaded file appear in the list of files. Click on the filename to open the file.

In the Martian Alphabet `R` script file, highlight the lines of code that starts with `library` and click "Run". This will load the **package** (or library) `catstats` needed for this activity; each package is a collection of `R` functions. We review a few of these packages here.

- Throughout the semester we will use the package `tidyverse` to allow us to use chaining (see Section 1.7 in the textbook for more on this symbol `%>%`.) Contained in `tidyverse` is the package `ggplot2`, used to create graphs in RStudio.
- The package `mosaic` contains the `favstats()` function to find summary statistics for quantitative variables.

- We will use the package `catstats`, starting in Chapter 5 (and in this activity), to create simulations for statistical inference.

These packages are already installed in the RStudio server, but you need to use the `library()` function to call the package into your R environment. We will only use the package `catstats` for this activity.

The # sign is not part of the R code. It is used by these authors to add comments to the R code and explain what each call is telling the program to do. R will ignore everything after a # sign when executing the code.

In the Martian Alphabet R script file for the `one_proportion_test()` function arguments, enter your class size (Q3 from the in-class activity) for `sample_size` and the number of students who were correct in identifying Bumba (Q6 from the in-class activity) for `as_extreme_as` argument. Highlight lines 3 – 8 and click run.

4. Is the distribution created from this code similar to what you saw in class in Q14?

1.2.5 Take-home messages

1. In this course we will learn how to evaluate a claim by comparing observed results (classes' "guesses" when asked to identify Bumba) to a distribution of many simulated results under an assumption like "blind guessing."

2. Blind guessing between two outcomes will be correct only about half the time. We can simulate data using a computer program to fit the assumption of blind guessing.

3. Unusual observed results will make us doubt the assumptions used to create the simulated distribution. A large number of correct "guesses" is evidence that a person was not just blindly guessing.

1.2.6 Additional notes

Use this space to summarize your thoughts and take additional notes on this week's activity and material covered, and to write down the names and contact information of your teammates.

Study Design

2.1 Reading Guide: Sampling, Experimental Design, and Scope of Inference

Section 1.3 (Sampling principles and strategies)

Videos

- 1.3

Vocabulary

(Target) Population:

Sample:

Anecdotal evidence:

Bias:

     Selection bias:

     Non-response bias:

     Response bias:

Convenience sample:

Simple Random Sample:

Non-response rate:

Representative:

Notes

Ideally, how should we sample cases from our target population? Using what sampling method?

Notes on types of sampling bias

- Someone must first be *chosen* to be in a study and refuse to participate in order to have **non-response bias**.

- There must be a valid reason for someone to lie or be untruthful to justify saying **response bias** is present. Yes, anyone could lie at any time to any question. Response bias is when those lies are predictable and systematic based on outside influences.

True or False: Convenience sampling tends to result in non-response bias.

True or False: Volunteer sampling tends to result in response bias.

True or False: Random sampling helps to resolve selection bias, but has no impact on non-response or response bias.

Sections 1.4 (Observational studies), 1.5 (Experiments), and 1.6 (Scope of inference)

Videos

- 1.4to1.6

Reminders from Section 1.2

Explanatory variable: The variable researchers think *may be* affecting the other variable. What the researchers control/assign in an experiment. If comparing groups, the explanatory variable puts the observational units into groups.

Response variable: The variable researchers think *may be* influenced by the other variable. This variable is always observed, never controlled or assigned.

Vocabulary

Observational study:

Observational data:

Prospective study:

Retrospective study:

Confounding variable:

Experiment:

    Randomized experiment:

    Blocking:

    Treatment group:

    Control group:

    Placebo:

    Placebo effect:

    Blinding:

Scope of inference:

Generalizability:

Causation:

Notes

What are the four principles of a well-designed randomized experiment?

Fill in the appropriate scope of inference for each study design.

	Study Type	
Selection of Cases	Randomized experiment	Observational study
Random sample (and no other sampling bias)		
Non-random sample (or other sampling bias)		

True or False: Observational studies can show an association between two variables, but cannot determine a causal relationship.

True or False: In order for an experiment to be valid, a placebo must be used.

True or False: If random sampling of the target population is used, and no other types of bias are suspected, results from the sample can be generalized to the entire target population.

True or False: If random sampling of the target population is used, and no other types of bias are suspected, results from the sample can be inferred as a causal relationship between the explanatory and response variables.

2.2 Activity: Study Design

2.2.1 Learning outcomes

- Explain the purpose of random sampling and its effect on scope of inference.
- Explain the purpose of random assignment and its effect on scope of inference.
- Identify whether a study design is observational or an experiment.
- Identify confounding variables in observational studies and explain why they are confounding.
- Identify the types of bias present in a study.

2.2.2 Terminology review

In this week's activity, we will examine different types of sampling bias and study designs, confounding variables, and how to determine the scope of inference for a study. Some terms covered in this activity are:

- Population
- Sample
- Parameter
- Statistic
- Selection bias
- Response bias
- Non-response bias
- Scope of inference
- Explanatory variable
- Response variable
- Confounding variable
- Experiment
- Observational study

To review these concepts, see Sections 1.3 through 1.6 in the textbook.

2.2.3 Types of sampling bias. Complete Q1 before class.

There are two parts to study design: how the sample was selected and how the study was conducted. First, we will look at sampling and types of bias (selection, non-response, or response).

In these next questions, identify the target population, the sample selected, the variable, and the type of bias present.

1. To determine if the proportion of out-of-state undergraduate students at Montana State University has increased in the last 10 years, a statistics instructor sent an email survey to 500 randomly selected current undergraduate students. One of the questions on the survey asked whether they had in-state or out-of-state residency. She only received 378 responses.

   Target population:

   Sample:

   Variable:

   Type(s) of bias:

2. Recently, a survey was conducted to assess current presidential approval in the United States. A random sample of 6395 US adults was taken. One of the questions asked in the survey was, "Do you agree or disagree with the President on many or nearly all of the top issues facing the country today?" Of those surveyed, 42% said they did agree.

   Target population:

   Sample:

   Variable:

   Type(s) of bias:

3. A television station is interested in predicting whether or not a local referendum to legalize marijuana for adult use will pass. It asks its viewers to phone in and indicate whether they are in favor or opposed to the referendum. Of the 2241 viewers who phoned in, forty-five percent were opposed to legalizing marijuana.

   Target population:

   Sample:

   Variable:

   Type(s) of bias:

4. To gauge the interest in a new swimming pool, a local organization stood outside of the Bogart Pool in Bozeman, MT, during open hours. One of the questions they asked was, "Since the Bogart Pool is in such bad repair, don't you agree that the city should fund a new pool?"

   Target population:

   Sample:

   Variable:

   Type(s) of bias:

2.2.4 Study design

The two main study designs we will cover are **observational studies** and **experiments**. Both the sampling method and the study design will help to determine the **scope of inference** for a study: To whom can we generalize, and can we conclude causation or only association? Remember that only in a randomized experiment can we conclude a **causal** (cause and effect) relationship between the explanatory and response variable.

Scope of Inference: If evidence of an association is found in our sample, what can be concluded?

Selection of cases	Study Type	
	Randomized experiment	Observational study
Random sample (and no other sampling bias)	Causal relationship, **and** can generalize results to population.	Cannot conclude causal relationship, **but** can generalize results to population.
No random sample (or other sampling bias)	Causal relationship, **but** cannot generalize results to a population.	Cannot conclude causal relationship, **and** cannot generalize results to a population.

Inferences to population can be made

Can only generalize to those similar to the sample due to potential sampling bias

Can draw cause-and-effect conclusions

Can only discuss association due to potential confounding variables

For the next exercises, identify the explanatory variable, the response variable, the study design (observational study or experiment), and the scope of inference (using the above chart).

5. The pharmaceutical company Moderna Therapeutics, working in conjunction with the National Institutes of Health, conducted Phase 3 clinical trials towards a vaccine for COVID-19 last fall. US clinical research sites enrolled 30,000 volunteers without COVID-19 to participate. Participants were randomly assigned to receive either the candidate vaccine or a saline placebo. They were then followed to assess whether or not they developed COVID-19. The trial was double-blind, so neither the investigators nor the participants knew who was assigned to which group.

   Explanatory variable:

   Response variable:

   Study design:

   What is the scope of inference for this study?

6. In another study, a local health department randomly selected 1000 US adults without COVID-19 to participate in a health survey. Each participant was assessed at the beginning of the study and then followed for one year. They were interested to see which participants elected to receive a vaccination for COVID-19 and whether any participants developed COVID-19.

   Explanatory variable:

   Response variable:

   Study design:

   What is the scope of inference for this study?

A **confounding variable** is a variable that is *both*

   1. associated with the explanatory variable, *and*
   2. associated with the response variable.

When both these conditions are met, if we observe an association between the explanatory variable and the response variable in the data, we cannot be sure if this association is due to the explanatory variable or the confounding variable—the explanatory and confounding variables are "confounded."
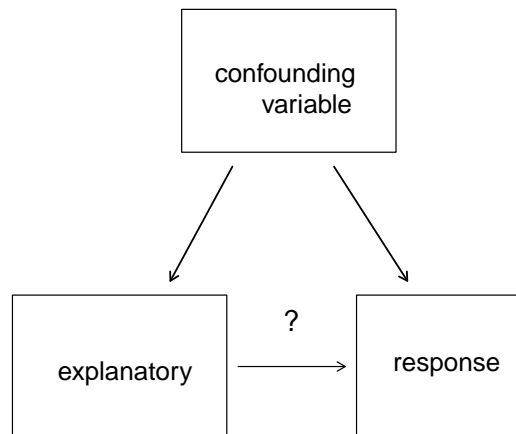
7. For each of the studies in questions 5 and 6, determine whether confounding variables could be an issue. If so, identify a potential confounding variable and explain how it meets the definition of a confounding variable.

2.2.5 Out-of-class activity

In the in-class activity, we looked at sampling methods and study design. Here are a few more questions to review that material.

1. The Bozeman school district is interested in surveying parents of students about their opinions on returning to in-person classes following the COVID-19 pandemic. They divided the school district into 10 divisions based on location and randomly surveyed 20 households within each division. Explain why selection bias would be present in this study design.

2. A study published in 2007 by Christopher Johnson, professor of music education and music therapy at the University of Kansas, revealed that students in elementary schools with superior music education programs scored around 20 percent higher in math scores on standardized tests, compared to schools with low-quality music programs. Explain how school budget could be a potential confounding variable. Be sure to address how the confounding variable is related to both the explanatory and response variable.

3. What is the purpose of random selection of a sample from the population?

4. What is the purpose of random assignment of the cases in a study to the explanatory variable groups?

2.2.6 Take-home messages

1. If the sample is selected using a random and non-biased method of selection (i.e., a random sample of the target population with no response or non-response bias), then the results of the study can be generalized to the target population. When using biased methods of selection, the results only apply to the sample selected or similar observational units.

2. The study design determines if we can draw causal inferences or not. If an association is detected, a randomized experiment allows us to conclude that there is a causal (cause-and-effect) relationship between the explanatory and response variable. Observational studies have potential confounding variables within the study that prevent us from inferring a causal relationship between the variables.

3. Confounding variables are variables not included in the study that are related to both the explanatory and the response variables. When there are potential confounding variables in the study we cannot draw causal inferences.

2.2.7 Additional notes

Use this space to summarize your thoughts and take additional notes on this week's activity and material covered.

Exploring Categorical Data

3.1 Reading Guide: Introduction to `R`, Categorical Data, and Probability

Section 1.7 (Data in `R`)

Videos

- Starting_with_R

Notes

`R` is case sensitive, meaning it reads `data` differently from `Data`. If you get an error message, check that your capitalization is correct.

`R` does not like spaces or special characters This means the column and row headers in the data set should not have spaces, periods, commas, etc. Instead of titling the variable `column header`, use `column_header` or `ColumnHeader`.

Tidy data: Data frames with

      1 row per _____,

      1 column per _____.

We highly recommend completing Tutorial 1 at the end of Chapter 1 (all four lessons) to give you practice with R/RStudio AND to help reflect on the content of Chapter 1: basics of data, sampling, study design, and scope of inference. These tutorials have some content questions and some places for you to practice using R online with some guidance.

      ___ indicate spots you need to type in functions, data sets, or variable names.

      There are Hint and Solution buttons on the R code box to help you.

We would not expect you to know the coding right now, especially for things like mutations or creating new variables in the data set. But seeing some initial coding for these more difficult functions will only make you more comfortable using the functions needed for this course!

Functions

State what these introductory functions do in `R`:

`glimpse(data_set_name)`

`head(data_set_name)`

```
data_set_name$variable_name
```

```
%>%
```

```
<-
```

Section 2.1 (Exploring categorical data)

Videos

- 2.1
- MosaicPlots

Vocabulary

Frequency table:

Relative frequency table:

Contingency or two-way table:

Unconditional proportion:

Conditional proportion:

    Row proportions:

    Column proportions:

Statistic:

    Sample proportion:

        Notation:

Parameter:

    Population proportion:

        Notation:

Bar plot:

Segmented bar plot:

Simpson's Paradox:

Notes

In a contingency table, which variable (explanatory or response) generally will make the columns of the table? Which variable will make the rows of the table?

In a segmented bar plot, the bars represent the levels of which variable? The segments represent the levels of which variable?

What type of plot(s) are appropriate to display a single categorical variable?

What type of plot(s) are appropriate to display two categorical variables?

What is the difference between a standardized segmented bar plot and a mosaic plot?

True or false: Pie charts are generally highly recommended ways to graphically display categorical data.

True or false: Two categorical variables are associated if the conditional proportions of a particular outcome (typically of the response variable) differ across levels of the other variable (typically the explanatory variable).

True or false: When a segmented bar plot has segments that sum to 1 (or 100%), the segment heights correspond to the proportions conditioned on the **segment**.

Review of Simpson's Paradox

Based on the segmented bar plot in Figure 2.6, which race of defendant was more likely to have the death penalty invoked?

Based on the segmented bar plot in Figure 2.7 and Table 2.9, which race of defendant was more likely to have the death penalty invoked when the victim was Caucasian?

Based on the segmented bar plot in Figure 2.7 and Table 2.9, which race of defendant was more likely to have the death penalty invoked when the victim was African American?

The direction of the relationship between the _____ and _____ variables is **reversed** when accounting for a _____ variable.

Section 2.2 (Probability with tables)

Videos

- 2.2

Vocabulary

Random process:

Probability:

Hypothetical two-way table:

Unconditional probability:

      Notation:

Conditional probability:

      Notation:

Event:

      Notation:

Complement:

      Notation:

Sensitivity:

Specificity:

Prevalence:

Notes

Method for creating a hypothetical two-way table:

1. Start with

2. Fill in the column or row totals using

3. Fill in the interior cells using

4. Add/Subtract to fill in the row/column totals not filled in at step 2.

     To find unconditional probabilities from the table,

     To find conditional probabilities from the table,

Example: Baby Jeff

1. Let D be the event a child has CPK. What does D^C represent?

2. Let T be the event a child tests positive for CPK. What does T^C represent?

3. Write each of the following values in proper probability notation:

    a. $1/10000 = 0.0001 = P($)

    b. $100\% = 1.0 = P($)

    c. $99.98\% = 0.9998 = P($)

4. Write out the steps for creating the hypothetical two-way table in section 2.2.4 of your textbook, then copy the table below.

First,

Next,

After that,

Finally,

Hypothetical two-way table:

	Test Positive	Test Negative	Total
Has CPK			
Does not have CPK			
Total			100,000

5. What is the probability that a child who had a positive test result actually does have CPK? What probability notation should be used for this value?

6. Explain how the probability in #5 was calculated.

3.2 Activity: What's the probability?

3.2.1 Learning outcomes

- Recognize and simulate probabilities as long-run frequencies.
- Construct two-way tables to evaluate conditional probabilities.
- Identify and create appropriate summary statistics and plots given a data set or research question involving categorical variables.
- Plots for a single categorical variable: bar plot.
- Plots for association between two categorical variables: segmented bar plot, mosaic plot.

3.2.2 Terminology review

In this week's in-class activity, we will cover two-way tables and probability. In the out-of-class activity, we will review summary measures and plots for categorical variables. Some terms covered in this activity are:

- Proportions
- Bar plots
- Segmented bar plots
- Probability
- Conditional probability
- Two-way tables

To review these concepts, see Sections 2.1 and 2.2 in the textbook.

3.2.3 "Current" Population Survey: 1985

The Current Population Survey (CPS) in 1985 is a survey sponsored by the Census Bureau and the Bureau of Labor Statistics to track labor force statistics for the United States population. The following table describes the variables in the data set:

Variable	Description
educ	Number of years of education
south	Whether lives in southern region of the US: S = lives in south, NS = does not live in south
sex	Sex: M = male, F = female
exper	Number of years of work experience (inferred from age and education)
union	Whether union member: Union or Not
wage	Wage (dollars per hour)
age	Age (years)
race	Race: W = white, NW = not white
sector	Sector of the economy: clerical, const (construction), management, manufacturing, professional, sales, service, other
married	Marital status: Married or Single

Vocabulary review. Complete Q1–Q4 before class.

1. What are the observational units?

2. Which variables are categorical?

3. What type(s) of plot could be used to display the proportion of individuals in each sector of the economy?

4. What type(s) of plot could be used to display the association between whether an individual is a union member and the sector of the economy in which they work?

3.2.4 Probability

5. Since the early 1980s, the rapid antigen detection test (RADT) of group A *streptococci* has been used to detect strep throat. A recent study of the accuracy of this test shows that the **sensitivity**, the probability of a positive RADT given the person has strep throat, is 86% in children, while the **specificity**, the probability of a negative RADT given the person does not have strep throat, is 92% in children. The **prevalence**, the probability of having group A strep, is 37% in children.

Let A = the event the child has strep throat, and B = the event the child has a positive RADT.

    a. Identify what each numerical value given in the problem represents in probability notation.

       $0.86 =$

       $0.92 =$

       $0.37 =$

    b. Create a hypothetical two-way table to represent the situation. Recall that in a two-way table, the explanatory variable should be your column headers (similar to the x-axis in a segmented bar graph!) while the response variable becomes the row headers.

			Total
Total			100,000

    c. Find $P(A \text{ and } B)$. What does this probability represent in the context of the problem?

    d. Find the probability that a child with a positive RADT actually has strep throat. What is the notation used for this probability?

    e. What is the probability that a child does not have strep given that they have a positive RADT? What is the notation used for this probability?

6. In a computer store, 30% of the computers in stock are laptops and 70% are desktops. Five percent of the laptops are on sale, while 10% of the desktops are on sale.

Let L = the event the computer is a laptop, and S = the event the computer is on sale.

a. Identify what each numerical value given in the problem represents in probability notation.

$0.30 =$

$0.70 =$

$0.05 =$

$0.10 =$

b. Create a hypothetical two-way table to represent the situation.

		Total
Total		100,000

c. Calculate the probability that a randomly selected computer will be a desktop, given that the computer is on sale. What is the notation used for this probability?

d. Find $P(S^C|L^C)$. What does this probability represent in context of the problem?

e. What is the probability a randomly selected computer is both a laptop and on sale? Give the appropriate probability notation.

3.2.5 Out-of-class activity

For this part of the activity we will focus on using RStudio and the provided `R` script file to create graphs and calculate proportions from each group.

3.2.5.1 Nightlight use and myopia

In a study reported in Nature (1999, Vol. 399, pp. 113-114), a survey of 479 children found that those who had slept with a nightlight or in a fully lit room before the age of 2 had a higher incidence of nearsightedness (myopia) later in childhood.

In this study, there are two variables studied: `Light`: level of light in room at night (no light, nightlight, full light) and `Sight`: level of myopia developed later in childhood (high myopia, myopia, no myopia).

1. Which variable is the explanatory variable? Which is the response variable?

An important part of understanding data is to create visual pictures of what the data represent. In this activity, we will create graphical representations of categorical data.

R code

Throughout these activities, we will often include the `R` code you would use in order to produce output or plots. These "code chunks" appear in gray. In the code chunk below, we demonstrate how to read the data set into `R` using the `read.csv()` function.

```
# This will read in the data set
myopia <- read.csv("https://math.montana.edu/courses/s216/data/ChildrenLightSight.csv")
```

Download and open the provided `R` script file for Activity 3 to answer the following questions. Highlight and run lines 1–5. These lines of code read in the data set and name the data set `myopia`. The `library()` function tells `R` which packages will be needed.

Displaying a single categorical variable

If we wanted to know how many children in our data set were in each level of myopia, we would create a frequency bar plot of the variable `Sight`. Enter the variable name, `Sight`, for xx into the `ggplot` code in line 10 in the `R` script file to create a bar plot. Highlight and run lines 9–15. Notice this is a **frequency** bar plot plotting counts (the number of children in each level of sight is displayed on the *y*-axis).

```
myopia %>% # Data set piped into...
ggplot(aes(y = xx)) +    # This specifies the variable
  geom_bar(stat = "count") +  # Tell it to make a bar plot
  labs(title = "Frequency Bar Plot of Level of Myopia",  # Give your plot a title
       x = "Frequency",   # Label the x axis
       y = "Level of Myopia")  + # Label the y axis
  coord_flip()  # Turn the bars so they are vertical
```

2. Sketch the bar plot created here. Be sure to label the axes.

3. Using the bar chart created, estimate how many children have some level of myopia.

We could also choose to display the data as a proportion in a **relative frequency** bar plot. To find the relative frequency, divide the count in each level of myopia by the sample size. These are sample proportions. Notice that in this code we told R to create a bar plot with proportions.

```r
myopia %>% # Data set piped into...
ggplot(aes(x = Sight)) +    # This specifies the variable
  geom_bar(aes(y = ..prop.., group = 1)) +  # Tell it to make a bar plot with proportions
  labs(title = "Relative Frequency Bar Plot of Level of Myopia",  # Give your plot a title
       x = "Level of Myopia",    # Label the x axis
       y = "Relative Frequency")  # Label the y axis
```


Relative Frequency Bar Plot of Level of Myopia
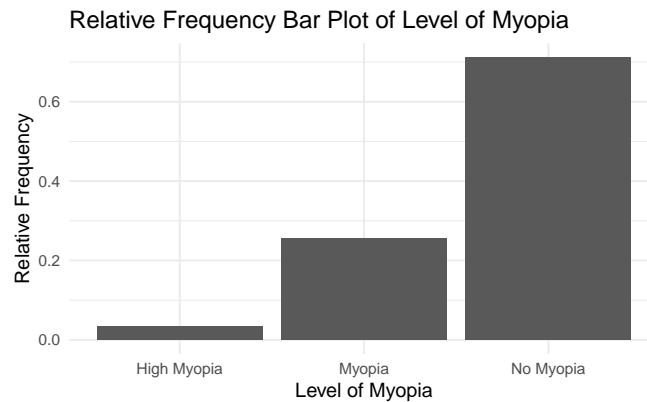
4. Which features in the relative frequency bar plot are the same as the frequency bar plot? Which are different?

Displaying two categorical variables

To examine the differences in level of myopia for the level of light, we would create a segmented bar plot of `Light` segmented by `Sight`. To create the segmented bar plot enter the variable name, `Light` (explanatory variable) for xx and the variable name, `Sight` (response variable) for yy in the R script file in line 28. Highlight and run lines 27–33.

```
myopia %>% # Data set piped into...
ggplot(aes(x = xx, fill = yy)) +    # This specifies the variables
  geom_bar(stat = "count", position = "fill") +  # Tell it to make a stacked bar plot
  labs(title = "Segmented Bar Plot of Night Light Use by Level of Myopia",
       # Make sure to title your plot
       x = "Level of Light",   # Label the x axis
       y = "") +  # Remove y axis label
    scale_fill_grey()  # Make figure black and white
```

5. Sketch the segmented bar plot created here. Be sure to label the axes.

6. From the segmented bar plot, estimate the proportion of no myopia for those that used a nightlight.

7. Which level of light has the highest proportion of `No Myopia`?

3.2.6   Take-home messages

1. Bar charts can be used to graphically display a single categorical variable either as counts or proportions. Segmented bar charts and mosaic plots are used to display two categorical variables.

2. Segmented bar charts always have a scale from 0 - 100%. The bars represent the outcomes of the explanatory variable. Each bar is segmented by the response variable. If the heights of each segment are the same for each bar there is no association between variables.

3. Mosaic plots are similar to segmented bar charts but the widths of the bars also show the number of observations within each outcome.

4. Conditional probabilities are calculated dependent on a second variable. In probability notation, the variable following | is the variable on which we are conditioning. The denominator used to calculate the probability will be the total for the variable on which we are conditioning.

3.2.7 Additional notes

Use this space to summarize your thoughts and take additional notes on this week's activity and material covered.

Exploring Quantitative Data

4.1 Reading Guide: Quantitative Data

Section 2.3 (Exploring quantitative data)

Type of Plots

Scatterplot:

Dot plot:

Histogram:

Density plot:

Box plot:

Vocabulary

Four characteristics of a scatterplot:

Form:

Strength:

Direction:

Unusual observations or outliers:

Data density:

Tail:

Skew:

Symmetric:

Modality:

Distribution (of a variable):

> Four characteristics of the distribution of one quantitative variable:
>
> Center:
>
> Variability:
>
> Shape:
>
> Outliers:

Point estimate:

Deviation:

Five number summary:

X^{th} percentile:

`e.g. if the value 6 is at the 10th percentile, then 10% of the data have values 6 or below.`
Interquartile range (IQR):

Robust statistics:

Notes

What type of plot(s) are appropriate for displaying one quantitative variable?

What type of plot(s) are appropriate for displaying two quantitative variables?

What type of plot(s) are appropriate for displaying one quantitative variable and one categorical variable?

What are the two ways to measure the 'center' of a distribution? Which one is considered robust to skew/outliers?

What are the three ways to measure the 'variability' of a distribution? Which one is considered robust to skew/outliers?

How are variance and standard deviation related?

Fill in the following table with the appropriate notation.

Summary Measure	Parameter	Statistic
Mean		
Variance		
Standard deviation		

How are outliers denoted on a box plot? How can you mathematically determine if a data set has outliers?

Section 2.4 (`R`: Exploratory data analysis) and Section 2.5 (Chapter 2 review)

Section 2.4 presents four tutorials on analyzing quantitative data in `R`. We recommend you complete all four.

Notes

Statistics summarize _____ .

Parameters summarize _____.

Fill in the following table with the appropriate notation for each summary measure.

Summary measure	Statistic	Parameter
Sample size		
Proportion (used to summarize one categorical variable)		
Mean (used to summarize one quantitative variable)		
Correlation (used to summarize two quantitative variables)		
Regression line slope (used to summarize two quantitative variables)		

Look at the table of vocabulary terms. If there are any you do not know, be sure to review the appropriate section of your text.

Data visualization summary

Fill in the following table to help associate type of plot for each of several scenarios.

	Appropriate plot(s)
One categorical variable (categorical response, no explanatory)	
One quantitative variable (quantitative response, no explanatory)	
Two categorical variables (categorical response, categorical explanatory)	
One of each (quantitative response, categorical explanatory)	
Two quantitative variables (quantitative response, quantitative explanatory)	

Decision tree for determining an appropriate plot given a number of variables and their types from Chapter 2 review:
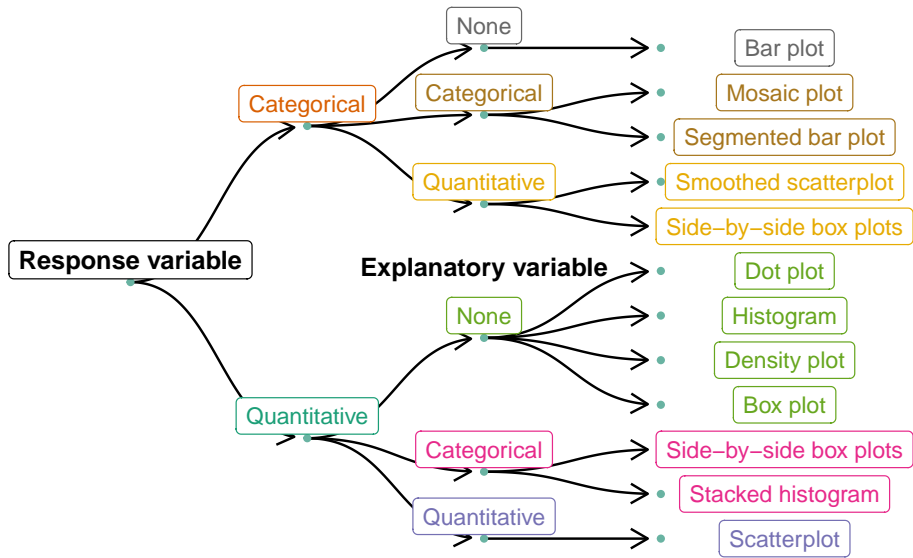
4.2 Activity: IMDb Movie Reviews

4.2.1 Learning objectives

- Identify and create appropriate summary statistics and plots given a data set or research question for quantitative data.

- Interpret the following summary statistics in context: median, lower quartile, upper quartile, standard deviation, interquartile range.

- Given a plot or set of plots, describe and compare the distribution(s) of a single quantitative variable (center, variability, shape, outliers).

4.2.2 Terminology review

In this week's activity, we will review summary measures and plots for quantitative variables. Some terms covered in this activity are:

- Two measures of center: mean, median

- Two measures of spread (variability): standard deviation, interquartile range (IQR)

- Types of graphs: box plots, dot plots, histograms

To review these concepts, see Section 2.3 in the textbook.

4.2.3 Movies released in 2016

A data set was collected on movies released in 2016. Here is a list of some of the variables collected on these movies.

Variable	Description
budget_mil	Amount of money (in US $ millions) budgeted for the production of the movie
revenue_mil	Amount of money (in US $ millions) the movie made after release
duration	Length of the movie (in minutes)
content_rating	Rating of the movie (G, PG, PG-13, R, Not Rated)
imdb_score	IMDb user rating score from 1 to 10
genres	Categories the movie falls into (e.g., Action, Drama, etc.)
movie_facebook_likes	Number of likes a movie receives on Facebook

Vocabulary review. Complete Q1 - 3 before class.

1. What are the observational units in this data set?

2. Which of the above listed variables are categorical?

3. Which of the above listed variables are quantitative?

Summarizing a single quantitative variable

The `favstats()` function from the `mosaic` package gives the summary statistics for a quantitative variable. Here we have the summary statistics for the variable `imdb_score`.

```r
# Read in data set
movies <- read.csv("https://math.montana.edu/courses/s216/data/Movies2016.csv")
movies %>% # Data set piped into...
  summarise(favstats(imdb_score)) # Apply favstats function to imdb_score
```

```
#>   min   Q1 median  Q3 max     mean       sd  n missing
#> 1 3.4 5.65    6.4 7.1 8.2 6.309783 1.086689 92       0
```

4. Give the values for the two measures of center.

5. Calculate the interquartile range (IQR = Q3 - Q1).

6. Report the value of the standard deviation and interpret this value in context of the problem.

Displaying a single quantitative variable

7. What are the three types of plots used to plot a single quantitative variable?

To create a histogram of the IMDb scores, enter the variable name, `imdb_score` in the provided R script file for xx at line 12, highlight and run lines 1 - 16. Visually, this shows us the range of IMDb scores for Movies released in 2016.

Notice that the **bin width** is 0.5. For example the first bin consists of the number of movies in the data set with an IMDb score of 3.25 to 3.75. It is important to note that a movie with a IMDb score on the boundary of a bin will fall into the bin above it; for example, 4.76 would be counted in the bin 4.75–5.25.

```
movies %>% # Data set piped into...
ggplot(aes(x = xx)) +    # Name variable to plot
  geom_histogram(binwidth = 0.5) +  # Create histogram with specified binwidth
  labs(title = "Histogram of IMDb Score of Movies in 2016", # Title for plot
       x = "IMDb Score", # Label for x axis
       y = "Frequency") # Label for y axis
```

8. Sketch the histogram created here.

9. Which range of IMDb scores have the highest frequency?

10. What is the shape of the distribution of IMDb scores?

11. Which five summary statistics are used in creating a box plot? *Hint*: Together they are called the **five-number summary** of the variable.

12. Using the code below we see that the three smallest IMDb scores in the data set are 3.4, 3.5, and 3.7 and the three largest IMDb scores are 8.0, 8.1, and 8.2:

```
movies %>% # Data set pipes into...
  select(imdb_score) %>% # Select imdb_score variable
  slice_min(imdb_score, n = 3)  # Show 3 smallest values

#>   imdb_score
#> 1        3.4
#> 2        3.5
#> 3        3.7
```

```
movies %>% # Data set pipes into...
  select(imdb_score) %>% # Select imdb_score variable
  slice_max(imdb_score, n = 3)  # Show 3 largest values

#>   imdb_score
#> 1        8.2
#> 2        8.1
#> 3        8.0
```

Using the summary statistics above, and the smallest and largest values of the variable to check for outliers, sketch a box plot of IMDb Score. Be sure to label the axes.

Displaying a single categorical and single quantitative variable

The box plot of movie budgets (in millions) by content rating is plotted using the code below. Enter the variable `budget_mil` for yy and the variable `content_rating` for xx at line 31, highlight and run code lines 29 - 35. This plot helps to compare the budget for different levels of content rating.

```
movies %>%  # Data set piped into...
  filter(content_rating != "Not Rated") %>% # Remove Not Rated movies
  ggplot(aes(y = yy, x = xx))+  # Identify variables
  geom_boxplot()+  # Tell it to make a box plot
  labs(title = "Side by side box plot of budget by content rating",  # Title
       x = "Content Rating",    # x-axis label
       y = "Budget (in Millions)")  # y-axis label
```

13. Sketch the box plots created using the `R` code.

14. Answer the following questions about the box plots created.

a. Which content rating has the highest center?

b. Which content rating has the largest spread?

c. Which content rating has the most skewed distribution?

d. Fifty percent of movies in 2016 with a PG-13 content rating fall below what value? What is the name of this value?

e. What is the value for the third quartile (Q3) for the PG-13 rating? Interpret this value in context.

4.2.4 Out-of-class activity

For a little more practice using `Rstudio` to create graphs of quantitative variables we will look at some other variables in the `Movies` data set. Download and open the provided `R` script file, highlight and run the first 8 lines of code.

To use the `favstats()` function in the mosaic package with two variables, we will enter the variables as a formula, response~explanatory.

```
movies %>% # Data set piped into...
  summarise(favstats(imdb_score~content_rating)) # Apply favstats function to imdb_score
```

```
#>   content_rating min    Q1 median    Q3 max      mean        sd  n missing
#> 1      Not Rated 3.7 4.700    5.7 6.700 7.7 5.700000 2.8284271  2       0
#> 2             PG 3.4 6.150    6.8 7.225 7.8 6.425000 1.2757351 12       0
#> 3          PG-13 4.0 5.800    6.5 7.100 8.2 6.367391 0.9477586 46       0
#> 4              R 3.5 5.375    6.3 7.050 8.1 6.221875 1.1335740 32       0
```

Using the provided `R` script file, we will create side-by-side histograms of IMDb by movie content rating. Enter the variable name, `imdb_score` for yy and `content_rating` for xx at line 44, highlight and run lines 39 - 48.

```
movies %>%  # Data set piped into...
  filter(content_rating != "Not Rated") %>% # Remove Not Rated movies
  ggplot(aes(y = yy, x = xx))+  # Identify variables
  geom_boxplot()+  # Tell it to make a box plot
  labs(title = "Side by side box plot of budget by content rating",  # Title
       x = "Content Rating",     # x-axis label
       y = "Budget (in Millions)")  # y-axis label
```

1. Using the provided `favstats` output and the side-by-side box plots, interpret the value of quartile 1 for the R content rating.

2. Which content rating has the highest center?

3. Which variable is the explanatory variable? Response variable?

4.2.5 Take-home messages

1. Histograms, box plots, and dot plots can all be used to graphically display quantitative variables. When we have a single categorical variable and a single quantitative variable we will display the data in side-by-side plots.

2. The box plot is created using the five number summary: minimum value, quartile 1, median, quartile 3, and maximum value. Values in the data set that are less than $Q_1 - 1.5*IQR$ and greater than $Q_3 + 1.5*IQR$ are considering outliers and are graphically represented by a dot outside of the whiskers on the box plot.

3. When comparing distributions of quantitative variables we look at the shape, center, spread, and for outliers. There are two measures of center: mean and the median and two measures of spread: standard deviation and the interquartile range, IQR = Q1- Q3.

4. The median and IQR are robust measures that are not affected by the presence of outliers.

4.2.6   Additional notes

Use this space to summarize your thoughts and take additional notes on this week's activity and material covered.

Exploring Multivariable Data

5.1 Reading Guide: Quantitative Data

Section 3.1 (Fitting a line, residuals, and correlation)

Videos

- Chapter3

Reminders from Section 2.3

Scatterplot: displays two quantitative variables; one dot = two measurements (x, y) on one observational unit.

Four characteristics of a scatterplot:

- *Form*: pattern of the dots plotted. Is the trend generally linear (you can fit a straight line to the data) or non-linear?

- *Strength*: how closely do the points follow a trend? Very closely (strong)? No pattern (weak)?

- *Direction*: as the x values increase, do the y-values tend to increase (positive) or decrease (negative)?

- Unusual observations or *outliers*: points that do not fit the overall pattern of the data.

Vocabulary

Residual:

    Formula:

Residual plot:

Correlation:

Notes

General equation of a linear regression for a *population*: $y = \beta_0 + \beta_1 x + \epsilon$, where

    x represents

    y represents

β_0 represents

β_1 represents

ϵ represents

General equation of a linear regression model from *sample* data: $\hat{y} = b_0 + b_1 x$, where

x represents

$\hat{y}$ represents

b_0 represents

b_1 represents

Fill in the following table with the appropriate notation for each summary measure.

Summary Measure	Parameter	Statistic
Correlation		
Slope		
y-intercept		

Fill in the blanks below to define some of the properties of correlation:

The value of correlation must be between _____. (Includes the endpoints of the interval)

The sign of correlation gives the _____ of the linear relationship.

The magnitude of correlation gives the _____ of the linear relationship.

True or false: A scatterplot that shows random scatter would be considered non-linear.

True or false: If the correlation between two quantitative variables is equal to zero, then the two variables are not associated.

True or false: To calculate a predicted y-value from a given x-value, just look at the scatterplot and estimate the y-value.

True or false: A positive residual indicates the data point is above the regression line.

Example: Brushtail possums

1. What are the observational units?

2. Look at the scatterplot in Figure 3.5.

   a) What is the explanatory variable? The response variable? What type is each?

   b) What is the form of the scatterplot?

   c) What is the direction of the scatterplot?

   d) What is the strength of the scatterplot?

   e) Are there any outliers on the scatterplot?

3. Write the equation of the regression line, in context (do not use x and y, use variable names instead).

4. Calculate the predicted head length for a possum with a 76.0 cm total length.

5. One of the possums in the data set has a total length of 76.0 cm and a head length of 85.1 cm. Calculate the residual for this possum. Does this possum lie above or below the regression line?

Section 3.2 (Least squares regression)

You may skip the special topic Sections 3.2.3.1 and 3.2.6.

Videos

- Chapter3

Vocabulary

Least squares criterion:

Least squares line:

lm() R function: name_of_model <- lm(response ~ explanatory, data = data_set_name)

slope:

y-intercept:

Extrapolation:

      Assumes the pattern seen in the data extends beyond the data collected!

Coefficient of determination:

      s_y^2 (or SST) represents

      s_{RES}^2 (or SSE) represents

Notes

Two methods for determining the best line:

    1.

    2.

Notation for the coefficient of determination:

Formulas for calculating the coefficient of determination:

True or false: A correlation between two quantitative variables implies a causal relationship exists between the variables.

True or false: The slope of the line tells us how much to expect the y variable to increase or decrease when the x variable increases by 1 unit.

True or false: The coefficient of determination is just the square of the correlation.

Example: Elmhurst College

    1. What are the observational units?

    2. Look at the scatterplot in Figure 3.13.

       a) What is the explanatory variable? The response variable?

       b) What is the form of the scatterplot?

c) What is the direction of the scatterplot?

d) What is the strength of the scatterplot?

e) Are there any outliers on the scatterplot?

3. Write the equation of the regression line, in context (do not use x and y, use variable names instead).

4. Interpret the slope of the line, in the context of the problem. Remember that both family income and gift aid from the university are measured in $1000s.

5. Interpret the y-intercept of the line, in the context of the problem. Remember that both family income and gift aid from the university are measured in $1000s.

6. Is your interpretation in question 5 an example of extrapolation?

7. Give and interpret, in context, the value of the coefficient of determination.

Section 3.3 (Outliers in linear regression)

Videos

- Chapter3

Vocabulary

Outlier:

Leverage:

Influential:

Notes

Investigate, but do not remove, outliers. Unless you find there was an actual error in the data collection, ignoring outliers can make models poor predictors!

True or false: All high leverage outliers are influential.

True or false: An outlier is considered high leverage if it is extreme in its x-value.

Section 3.4 (R: Correlation and regression) and Section 3.5 (Chapter 3 review)

Videos

- Chapter3

Section 3.4 presents five tutorials on analyzing two quantitative variables in `R`. We recommend you complete all five.

Notes

Statistics summarize

Parameters summarize

What are the two ways to calculate the coefficient of determination?

What is the formula for calculating a residual?

Determine whether each of the following statements about the correlation coefficient are true or false:

1. The correlation coefficient must be a positive number.

2. Stronger linear relationships are indicated by correlation coefficients far from 0.

3. The correlation coefficient is a robust statistic.

4. When two variables are highly correlated, that indicates a causal relationship exists between the variables.

5. The sign of the correlation coefficient will be the same as the sign of the regression line slope, though the values are typically different.

Fill in the blanks to correctly interpret:

- Slope:

  For every _____, we expect _____ to increase (if slope is _____) or decrease (if slope is _____) by the absolute value of the _____.

- y-intercept:

  If _____, we predict the _____ to equal _____.

Look at the table of vocabulary terms. If there are any you do not know, be sure to review the appropriate section of your text.

Section 4.1 (Gapminder world)

Videos

- Chapter4

Reminder from Section 3.1

Use color and a legend to add a third variable to a scatterplot. E.g., Color the dots to represent different levels of a categorical variable or shading to represent different values of a quantitative variable.

Vocabulary

Interaction:

Aesthetic:

Notes

If the response and one predictor are quantitative and the other predictor categorical, we fit a regression line for each level of the categorical predictor.

- Parallel slopes would indicate that that the two predictors _____ in explaining the response.

- Non-parallel slopes would indicate that the two predictors _____ in explaining the response.

True or false: Scatterplots can only display two variables at a time.

Section 4.2 (Simpson's Paradox, revisited)

Videos

- Chapter4

Reminder from Section 2.1

Simpson's Paradox: when the relationship between the explanatory and response variable is reversed when looking at the relationship within different levels of a confounding variable.

Notes

True or false: Simpson's Paradox can only occur when the explanatory, response, and confounding variables are all categorical.

Example: SAT scores

1. What are the observational units?

2. Look at the scatterplot in Figure 4.5.

   a) What is the explanatory variable? The response variable?

   b) What is the form of the scatterplot?

   c) What is the direction of the scatterplot?

   d) What is the strength of the scatterplot?

   e) Are there any outliers on the scatterplot?

3. What would need to be done to the study design in order to eliminate the confounding variable: percent of eligible students taking the SAT?

4. What features of the scatterplots in Figure 4.6 demonstrate that the percent of eligible students taking the SAT is a confounding variable?

5. How does Figure 4.7 demonstrate Simpson's Paradox?

Section 4.4 (Chapter 4 review)

Section 4.3 discusses multiple regression and presents five tutorials on analyzing multiple variables in R. This section is a special topic, meaning you are not required to read or complete these tutorials.

Videos

- Chapter4

Notes

To determine if the relationship between two quantitative variables differs across levels of a categorical variable, you should compare

Simpson's Paradox:

5.2 Activity: Movie Profits

5.2.1 Learning objectives

- Identify and create appropriate summary statistics and plots given a data set with two quantitative variables.

- Use scatterplots to assess the relationship between two quantitative variables.

- Find the estimated line of regression using summary statistics and R linear model (`lm()`) output.

- Interpret the slope coefficient in context of the problem.

- Calculate and interpret R^2, the coefficient of determination, in context of the problem.

- Find the correlation coefficient from R output or from R^2 and the sign of the slope.

5.2.2 Terminology review

In this week's activity, we will review summary measures and plots for two quantitative variables. Some terms covered in this activity are:

- Scatterplot

- Correlation (r or R)

- Coefficient of determination (r-squared or R^2)

- Least-squares line of regression

- Slope and y-intercept

- Residuals

To review these concepts, see Chapter 3 in the textbook.

5.2.3 Movies released in 2016

We will revisit the data set used last week collected on Movies released in 2016. Here is a reminder of the variables collected on these movies.

Variable	Description
budget_mil	Amount of money (in US $ millions) budgeted for the production of the movie
revenue_mil	Amount of money (in US $ millions) the movie made after release
duration	Length of the movie (in minutes)
content_rating	Rating of the movie (G, PG, PG-13, R, Not Rated)
imdb_score	IMDb user rating score from 1 to 10
genres	Categories the movie falls into (e.g., Action, Drama, etc.)
movie_facebook_likes	Number of likes a movie receives on Facebook

Vocabulary review. Complete Q1–Q5 before class.

Note: You will need to use the provided R script file for Activity 5 to complete question 3.

1. What type of plot should be used to display the relationship between budget_mil and revenue_mil?

2. What three summary statistics could be used to describe the relationship between two quantitative variables?

We will look at the relationship between budget and revenue for movies released in 2016. Enter the variable name `budget_mil` for xx and `revenue_mil` for yy at line 7 in the `R` script file to create the scatterplot. (Note: both variables are measured in "millions of dollars", or $MM.) Highlight and run lines 1–12.

```
movies %>% # Data set pipes into...
ggplot(aes(x = xx, y = yy))+  # Specify variables
  geom_point() +  # Add scatterplot of points
  labs(x = "Budget in Millions ($)",  # Label x-axis
       y = "Revenue in Millions ($)",  # Label y-axis
       title = "Revenue vs. Budget") + # Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```

3. Sketch the scatterplot created from the code.

4. Assess the four features of the scatterplot that describe this relationship. Describe each feature using a complete sentence!

- Form (linear, non-linear)

- Direction (positive, negative)

- Strength

- Unusual observations or outliers

5. Does there appear to be an association between budget and revenue? Explain.

Correlation

Correlation measures the strength and the direction of the linear relationship between two quantitative variables. The closer the value of correlation to $+1$ or -1, the stronger the linear relationship. Values close to zero indicate a very weak linear relationship between the two variables. The following output shows a correlation matrix between several pairs of quantitative variables.

```
movies %>%  # Data set pipes into
  select(c("budget_mil", "revenue_mil",
           "duration", "imdb_score",
           "movie_facebook_likes")) %>%
  cor(use="pairwise.complete.obs") %>%
  round(3)
```

```
#>                      budget_mil revenue_mil duration imdb_score
#> budget_mil                1.000       0.686    0.463      0.292
#> revenue_mil               0.686       1.000    0.227      0.398
#> duration                  0.463       0.227    1.000      0.261
#> imdb_score                0.292       0.398    0.261      1.000
#> movie_facebook_likes      0.678       0.723    0.438      0.309
#>                      movie_facebook_likes
#> budget_mil                          0.678
#> revenue_mil                         0.723
#> duration                            0.438
#> imdb_score                          0.309
#> movie_facebook_likes                1.000
```

6. Using the output above, which two variables have the *strongest* correlation? What is the value of this correlation?

7. What is the value of correlation between budget and revenue?

8. Based on the value of correlation found in question 7, what would the sign of the slope be? Positive or negative? Explain.

9. Does your answer to question 8 match the direction you choose in question 4?

10. Explain why the correlation values on the diagonal are equal to 1.

Slope

The linear model function in R (`lm()`) gives us the summary for the least squares regression line. The estimate for (`Intercept`) is the y-intercept for the line of least squares, and the estimate for `budget_mil` (the x-variable name) is the value of b_1, the slope.

```
# Fit linear model: y ~ x
revenueLM <- lm(revenue_mil ~ budget_mil, data=movies)
summary(revenueLM)$coefficients # Display coefficient summary
```

```
#>             Estimate Std. Error  t value    Pr(>|t|)
#> (Intercept) 9.1693054  9.0175499 1.016829 3.119606e-01
#> budget_mil  0.9460001  0.1056786 8.951670 4.339561e-14
```

11. Write out the least squares line using the summary statistics provided above. Use proper statistical notation.

You may remember from middle and high school that slope $= \frac{\text{rise}}{\text{run}}$.

Using b_1 to represent slope, we can write that as the fraction $\frac{b_1}{1}$.

Therefore, the slope predicts how much the line will *rise* for each *run* of $+1$. In other words, as the x variable increases by 1 unit, the y variable is predicted to change (increase/decrease) by the value of slope.

12. Interpret the value of slope in context of the problem.

13. Using the least squares line from question 11, predict the revenue for a movie with a budget of 165 $MM.

Residuals

The model we are using assumes the relationship between the two variables follows a straight line. The residuals are the errors, or the variability in the response that hasn't been modeled by the line (model).

$$\text{Data} = \text{Model} + \text{Residual}$$

$$\implies \text{Residual} = \text{Data} - \text{Model}$$

$$e_i = y_i - \hat{y}_i$$

14. The movie *Independence Day: Resurgence* had a budget of 165 $MM and revenue of 102.315 $MM. Find the residual for this movie.

15. Did the line of regression overestimate or underestimate the revenue for this movie?

5.2.4 Out-of-class activity

Coefficient of determination (squared correlation)

The third summary measure used to explain the linear relationship between two quantitative variables is the coefficient of determination (r^2). The coefficient of determination, r^2, can also be used to describe the strength of the linear relationship between two quantitative variables. The value of r^2 (a value between 0 and 1) represents the **proportion of variation in the response that is explained by the least squares line with the explanatory variable**. There are two ways to calculate the coefficient of determination:

Square the correlation coefficient: $r^2 = (r)^2$

Use the variances of the response and the residuals: $r^2 = \dfrac{s_y^2 - s_{RES}^2}{s_y^2} = \dfrac{SST - SSE}{SST}$

1. Use the correlation, r, found in question 7 of the in-class activity, to calculate the coefficient of determination between budget and revenue, r^2.

2. The variance of the response variable, revenue in \$MM, is about $s_{revenue}^2 = 8024.261$ \$MM^2 and the variability in the residuals is about $s_{RES}^2 = 4244.832$ \$MM^2. Use these values to calculate the coefficient of determination. Verify that your answers to 1 and 2 are the same.

3. Write a sentence interpreting the coefficient of determination in context of the problem.

Multivariable plots

What if we wanted to see if the relationship between movie budget and revenue differs if we add another variable into the picture? The following plot visualizes three variables, creating a **multivariable** plot.

```
movies %>% # Data set pipes into...
  filter(content_rating != "Not Rated") %>% # Remove Not Rated movies
  ggplot(aes(x = budget_mil, y = revenue_mil, color = content_rating)) +  # Specify variables
  geom_point(aes(shape = content_rating), size = 3) +  # Add scatterplot of points
  labs(x = "Budget in Millions ($)",  # Label x-axis
       y = "Revenue in Millions ($)",  # Label y-axis
       color = "Content Rating",  # Label legend
       title = "Revenue vs. Budget") + # Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE, lwd = 2) + # Add regression lines
  scale_color_grey() # Make black and white
```

Revenue vs. Budget
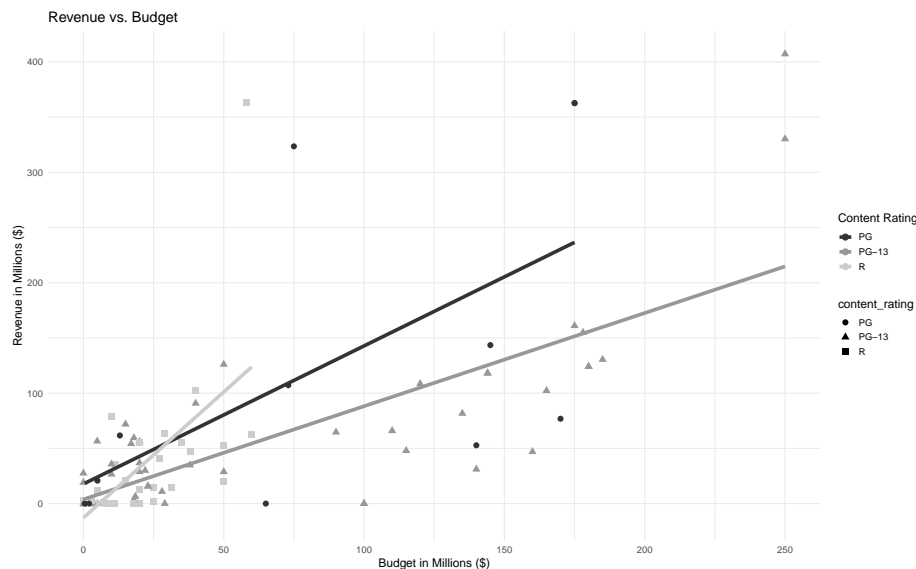
4. Identify the three variables plotted in this graph.

5. Does the *relationship* between movie budget and revenue differ among the different content ratings? Explain.

5.2.5 Take-home messages

1. Two quantitative variables are graphically displayed in a scatterplot. The explanatory variable is on the x-axis and the response variable is on the y-axis. When describing the relationship between two quantitative variables we look at the form (linear or non-linear), direction (positive or negative), strength, and for the presence of outliers.

2. There are three summary statistics used to summarize the relationship between two quantitative variables: correlation (r), slope of the regression line (b_1), and the coefficient of determination (r^2).

3. The sign of correlation and the sign of the slope will always be the same. The closer the value of correlation is to -1 or $+1$, the stronger the relationship between the explanatory and the response variable.

4. The coefficient of determination multiplied by 100 $(r^2 \times 100)$ measures the percent of variation in the response variable that is explained by the relationship with the explanatory variable. The closer the value of the coefficient of determination is to 100%, the stronger the relationship.

5. We can use the line of regression to predict values of the response variable for values of the explanatory variable. Do not use values of the explanatory variable that are outside of the range of values in the data set to predict values of the response variable (reflect on why this is true.). This is called **extrapolation**.

5.2.6 Additional notes

Use this space to summarize your thoughts and take additional notes on this week's activity and material covered.

Inference for a Single Categorical Variable: Hypothesis Testing

6.1 Reading Guide: Categorical Inference

Section 5.1 (Foundations of inference: Hypothesis tests)

You may skip Section 5.1.4. This section will be covered next week.

Videos

- 5.1

Vocabulary

Statistical inference:

Hypothesis test:

> Also called a 'significance test'.

Simulation-based method:

Theory-based method:

Central Limit Theorem:

Sampling distribution:

Standard deviation of a statistic:

Standard error of a statistic:

Null hypothesis (H_0):

Alternative hypothesis (H_A):

P-value:

Point estimate:

Test statistic:

Decision:

Significance level (α):

Statistically significant:

Notes

What 'theory' is behind the theory-based methods of analysis?

Consider the US judicial system:

What is the null hypothesis?

What is the alternative hypothesis?

The jury is presented with evidence.

- If the evidence is strong (beyond a reasonable doubt), the jury will find the defendant:

- If the evidence is not strong (not beyond a reasonable doubt), the jury will find the defendant:

To create a simulation, which hypothesis (null or alternative) do we assume is true?

More on p-values:

Lower the p-value:

Interpretations require:

General steps of a hypothesis test:

Conclusions should include:

Decision:

If p-value $\leq \alpha$, the decision is to:

If p-value $> \alpha$, the decision is to:

True or False: If the p-value is above 0.10, that means the null hypothesis is true.

True or False: When conducting a simulation-based hypothesis test, the null hypothesis is assumed to be true to create the simulation.

Formulas

$SD(\hat{p}) =$

Example: Martian alphabet

1. What is the sample statistic presented in this example? What notation would be used to represent this value?

2. What are the two possible explanations for how these data could have occurred?

3. Of the two explanations, which is the null and which is the alternative hypothesis?

4. How could coins be used to create a simulation of what should happen if everyone in the class was just guessing?

5. How can we use the simulation to determine which of the two possibilities is more believable?

6. What decision should be made at an $\alpha = 0.05$ significance level? Justify your answer.

7. Are the results in this example statistically significant? Justify your answer.

Section 5.2 (The normal distribution)

Videos

- 5.2

Vocabulary

Normal distribution (Also known as: normal curve, normal model, Gaussian distribution):

Notation:

Standard normal distribution:

    Notation:

Z-score:

Xth percentile:

68-95-99.7 rule:

Notes

Interpretation of a Z-score:

True or False: The more unusual observation will be the observation with the largest Z-score.

Approximately what percent of a normal distribution is in the interval

    (mean – standard deviation, mean + standard deviation):

    (mean – 2×(standard deviation), mean + 2×(standard deviation)):

    (mean – 3×(standard deviation), mean + 3×(standard deviation)):

Formulas

Z =

R coding

Calculating normal probabilities   When using the `pnorm()` R function, you will need to enter values for the arguments `mean`, `sd`, and `q` to match the question.

```
pnorm(mean = mu, sd = sigma, q = x, lower.tail = TRUE)
```

This function will return the proportion of the N(`mu`,`sigma`) distribution which is *below* the value `x`.

Example: `pnorm(mean = 5, sd = 2, q = 3, lower.tail = TRUE)` will give us the proportion of a N(5,2) distribution which is below 3, which equals 0.159:

```
pnorm(mean = 5, sd = 2, q = 3, lower.tail = TRUE)
#> [1] 0.1586553
```

Changing to `lower.tail = FALSE` will give the proportion of the distribution which is *above* the value x.

```
pnorm(mean = 5, sd = 2, q = 3, lower.tail = FALSE)
#> [1] 0.8413447
```

**Displaying normal probabilities**   When using the `normTail()` R function, you will need to enter values for the arguments m, s, and L (or U) to match the question.

```
normTail(m = mu, s = sigma, L = x)
```

This function (in the `openintro` package) will plot a N(mu, sigma) distribution and shade the area that is below the value x.

Example: `normTail(m = 5, s = 2, L = 3)` creates the plot pictured below.
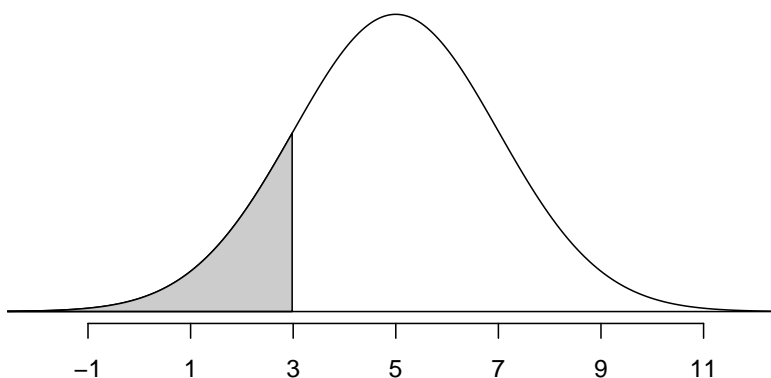
Changing L to U will shade the area *above* x.

Example: `normTail(m = 5, s = 2, U = 3)` plots a N(5,2) distribution with the area above 3 shaded.

Calculating normal percentiles   When using the `qnorm()` R function, you will need to enter values for the arguments mean, sd, and p to match the question.

```
qnorm(mean = mu, sd = sigma, p = x, lower.tail = TRUE)
```

This function will return the value on the N(mu, sigma) distribution which has x area of the distribution *below* it.

Example: `qnorm(mean = 5, sd = 2, p = 0.159, lower.tail = TRUE)` will give us the value on a N(5,2) distribution which has 0.159 (15.9%) of the distribution below it, which equals 3 (from the R output above).

Changing to `lower.tail = FALSE` will give the value which has x area of the distribution *above* it.

We would recommend you work through each of the examples in Section 5.2.4 using R.

Section 5.3 (Inference for one proportion)

You may skip Section 5.3.2 and stop before the "Confidence interval for π" sub-section in Section 5.3.3. These sections will be covered next week.

Videos

- 5.3
- OnePropTheory

Reminders from previous sections

n = sample size

$\hat{p}$ = sample proportion

π = population proportion

General steps of a hypothesis test:

1. Frame the research question in terms of hypotheses.

2. Collect and summarize data using a test statistic.

3. Assume the null hypothesis is true, and simulate or mathematically model a null distribution for the test statistic.

4. Compare the observed test statistic to the null distribution to calculate a p-value.

5. Make a conclusion based on the p-value and write the conclusion in context.

Parameter: a value summarizing a variable(s) for a population.

Statistic: a value summarizing a variable(s) for a sample.

Sampling distribution: plot of statistics from 1000s of samples of the same size taken from the same population.

Standard deviation of a statistic: the variability of statistics from 1000s of samples; how far, on average, each statistic is from the true value of the parameter.

Standard error of a statistic: estimated standard deviation of a statistic.

Hypothesis test: a process to determine how strong the evidence of an effect is.

Also called a 'significance test'.

Simulation-based method: Simulate lots of samples of size n under assumption of the null hypothesis, then find the proportion of the simulations that are at least as extreme as the observed sample statistic.

Theory-based method: Develop a mathematical model for the sampling distribution of the statistic under the null hypothesis and use the model to calculate the probability of the observed sample statistic (or one more extreme) occurring.

Null hypothesis (H_0): the skeptical perspective; no difference; no change; no effect; random chance; what the researcher hopes to prove is **wrong**.

Alternative hypothesis (H_A): the new perspective; a difference/increase/decrease; an effect; not random chance; what the researcher hopes to prove is **correct**.

P-value: probability of seeing the observed sample data, or something more extreme, assuming the null hypothesis is true.

$\implies$ Lower the p-value the stronger the evidence AGAINST the null hypothesis and FOR the alternative hypothesis.

Decision: a determination of whether to 'reject' or 'fail to reject' a null hypothesis based on a p-value and a pre-set level of significance.

Significance level (α): a threshold used to determine if a p-value provides enough evidence to reject the null hypothesis or not.

Common levels of α include 0.01, 0.05, and 0.10.

Statistically significant: results are considered statistically significant if the p-value is below the significance level.

Central Limit Theorem: For large sample sizes, the sampling distribution of a sample proportion (or mean) will be approximately normal (bell-shaped and symmetric).

Vocabulary

Point estimate:

Test statistic:

Null value:

Null distribution:

Standardized statistic:

Notes

Conditions for the Central Limit Theorem to apply (for the sampling distribution of $\hat{p}$ to be approximately normal)

Independence:

Checked by:

Success-failure condition:

Checked by:

Formulas

$SD(\hat{p}) =$

Null standard error of the sample proportion:

$SE_0(\hat{p}) =$

Standardized statistic/standardized sample proportion:

$Z =$

Example: Organ donations

1. What is the sample statistic presented in this example? What notation would be used to represent this value?

2. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

3. Write the null and alternative hypotheses in words.

4. Write the null and alternative hypotheses in notation.

5. To simulate the null distribution, we would not be able to use coins. Why or why not?

6. How could we use cards to simulate 1 sample which assumes the null hypothesis is true? How many blue cards — to represent what? How many red cards — to represent what? How many times would we draw a card and replace it back in the deck? What would you record once you completed the draw-with-replacement process?

7. How can we calculate a p-value from the simulated null distribution for this example?

8. What was the p-value of the test?

9. At the 5% significance level, what decision would you make?

10. What conclusion should the researcher make?

11. Are the results in this example statistically significant? Justify your answer.

12. Are the conditions met to use theoretical methods to analyze these data?

Example: Payday loans

1. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

2. Write the null and alternative hypotheses in words.

3. Write the null and alternative hypotheses in notation.

4. Are the conditions met to use theoretical methods to analyze these data?

5. Calculate the null standard error of the sample proportion.

6. What is the sample statistic presented in this example? What notation would be used to represent this value?

7. Calculate the standardized sample proportion.

8. How can we calculate a p-value from the normal distribution for this example?

9. What was the p-value of the test?

10. At the 5% significance level, what decision would you make?

11. What conclusion should the researcher make?

12. Are the results in this example statistically significant? Justify your answer.

6.2 Activity: Handedness of Male Boxers — Testing

6.2.1 Learning objectives

- Identify the two possible explanations (one assuming the null hypothesis, and one assuming the alternative hypothesis) for a relationship seen in sample data.

- Given a research question involving a single categorical variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Describe and perform a simulation-based hypothesis test for a single proportion.

- Interpret and evaluate a p-value for a simulation-based hypothesis test for a single proportion.

6.2.2 Terminology review

In this week's in-class activity, we will introduce simulation-based hypothesis testing for a single categorical variable. Some terms covered in this activity are:

- Parameter of interest

- Null hypothesis

- Alternative hypothesis

- Simulation

- Null distribution

- p-value

To review these concepts, see Chapter 5 in your textbook, focusing on Sections 5.1 through 5.3.

6.2.3 Steps of the statistical investigation process

We will work through a six-step process to complete a hypothesis test for a single proportion, first introduced in the Martian Alphabet Activity in week 1.

- **Ask a research question** that can be addressed by collecting data. What are the researchers trying to show?

- **Design a study and collect data**. This step involves selecting the people or objects to be studied and how to gather relevant data on them.

- **Summarize and visualize the data**. Calculate summary statistics and create graphical plots that best represent the research question.

- **Use statistical analysis methods to draw inferences from the data**. Choose a statistical inference method appropriate for the data and identify the p-value and/or confidence interval after checking assumptions. In this study, we will focus on using randomization to generate a simulated p-value.

- **Communicate the results and answer the research question**. Using the p-value and confidence interval from the analysis, determine whether the data provide statistical evidence against the null hypothesis. Write a conclusion that addresses the research question.

- **Revisit and look forward** to point out limitations of the study and suggest new studies that could be performed to build on the findings of the study.

6.2.4 Handedness of male boxers

Left-handedness is a trait that is found in about 10% of the population. Past studies have shown that left-handed men are over-represented among professional boxers. The fighting claim states that left-handed men have an advantage in competition. In this random sample of 500 male professional boxers, we want to see if there is an over-prevalence of left-handed fighters.

Summary statistics review. Complete Q1–Q4 before class.

1. What are the observational units?

2. What variable are we measuring on each observational unit? Is it categorical or quantitative?

3. What type of plot would be appropriate to visually display these data?

4. Write out in context the statistic we will calculate to summarize the data.

Ask a research question

5. Identify the research question for this study.

Design a study and collect data

Before using statistical inference methods, we must check that the cases are independent. The sample observations are independent if the outcome of one observation does not influence the outcome of another. One way this condition is met is if data come from a simple random sample of the target population.

6. Are the cases independent? Justify your answer.

Summarize and visualize the data

```
 # Read in data set
handedness <- read.csv("https://math.montana.edu/courses/s216/data/Male_boxers_sample.csv")
handedness %>% count(Stance)  # Count number in each Stance category
```

```
#>        Stance    n
#> 1  left-handed   81
#> 2 right-handed  419
```

7. Using the output above, calculate the appropriate summary statistic to represent the research question. Use appropriate notation.

8. What type of plot should be used to represent these data? Sketch this plot.

Use statistical analysis methods to draw inferences from the data

When performing a hypothesis test, we must first identify the null hypothesis. The null hypothesis is written about the parameter of interest, or the value that summarizes the variable in the population. *For example, in the Martian Alphabet Activity, the parameter of interest is the true proportion of statistic students who would correctly identify Bumba.*

9. Write out the parameter of interest for this study.

10. Using the parameter of interest in question 9, write out the null hypothesis in words. That is, what do we assume to be true about the parameter of interest when we perform our simulation?

The notation used for a population proportion (or probability, or true proportion) is π. Since this summarizes a population, it is a parameter. When writing the **null hypothesis** in notation, we set the parameter equal to the null value, $H_0 : \pi = \pi_0$.

11. Write the null hypothesis in notation using the null value of 0.1 in place of π_0 in the equation given above.

The **alternative hypothesis** is the claim to be tested and the direction of the claim (less than, greater than, or not equal to) is based on the research question.

12. Based on the research question from question 5, are we testing that the parameter is greater than 0.1, less than 0.1 or different than 0.1?

13. Write out the alternative hypothesis in words.

14. Write out the alternative hypothesis in notation.

Remember that when utilizing a hypothesis test, we are evaluating two competing possibilities. For this study the **two possibilities** are either...

- The true proportion of male boxers who are left-handed is 0.1 and our results just occurred by random chance; or,
- The true proportion of male boxers who are left-handed is greater than 0.1 and our results reflect this.

Notice that these two competing possibilities represent the null and alternative hypotheses.

We will now simulate a **null distribution** of sample proportions. The null distribution is created under the assumption the null hypothesis is true. In this case, we assume the true proportion of male boxers who are left-handed is 0.1, so we will create 1000 (or more) different simulations of 500 boxers under this assumption.

Let's think about how to use red and blue cards to create one simulation of 500 boxers under the assumption the null hypothesis is true. Suppose blue cards represents left-handed boxers and red cards represents right-handed boxers.

15. How many cards total do we need? How many blue ones? How many red ones?

16. Next, we would mix the cards together and draw 1 card, write down if it's red or blue, and replace the card. How many times would we need to repeat this process to simulate one sample?

17. Once we have one simulated sample, what would we calculate and plot on the null distribution? *Hint*: What statistic are we calculating from the data?

We will use the computer to simulate a null distribution of 1000 different samples of 500 male boxers, plotting the proportion who are left-handed in each sample, based on the assumption that the true proportion of male boxers who are left-handed is 0.1 (or that the null hypothesis is true).

To use the computer simulation, we will need to enter the

- assumed "probability of success" (π_0),
- "sample size" (the number of observational units or cases in the sample),
- "number of repetitions" (the number of samples to be generated),
- "as extreme as" (the observed statistic), and
- the "direction" (matches the direction of the alternative hypothesis).

18. What values should be entered for each of the following into the one proportion test to create 1000 simulations?

- Probability of success:

- Sample size:

- Number of repetitions:

- As extreme as:

- Direction (`"greater"`, `"less"`, or `"two-sided"`):

We will use the `one_proportion_test()` function in R (in the `catstats` package) to simulate the null distribution of sample proportions and compute a p-value. Using the provided R script file, fill in the values/words for each `xx` with your answers from question 18 in the one proportion test to create a null distribution with 1000 simulations. Then highlight and run lines 1–14.

```
one_proportion_test(probability_success = xx, # Null hypothesis value
        sample_size = xx, # Enter sample size
        number_repetitions = 1000, # Enter number of simulations
        as_extreme_as = xx, # Observed statistic
        direction = "xx", # Specify direction of alternative hypothesis
        report_value = "proportion") # Reporting proportion or number of successes?
```

19. Sketch the null distribution created from the R code here.

20. Around what value is the null distribution centered? Why does that make sense?

21. Circle the observed statistic (value from question 7) on the distribution you drew in question 19. Where does this statistic fall in the null distribution: Is it near the center of the distribution (near 0.1) or in one of the tails of the distribution?

22. Is the observed statistic likely to happen or unlikely to happen if the true proportion of male boxers who are left-handed is 0.1? Explain your answer using the plot.

23. Using the simulation, what is the proportion of simulated samples that generated a sample proportion at the observed statistic or greater, if the true proportion of male boxers who are left-handed is 0.1? *Hint*: Look under the simulation.

The value in question 23 is the **p-value**. The smaller the p-value, the more evidence we have against the null hypothesis. Explain why this makes sense?

24. Using the following guidelines for the strength of evidence, how much evidence do the data provide against the null hypothesis? (Circle one of the five descriptions.)

Strength of Evidence Against the Null

Very strong	Strong	Moderate	Some	Little to no

1%   5%   10%
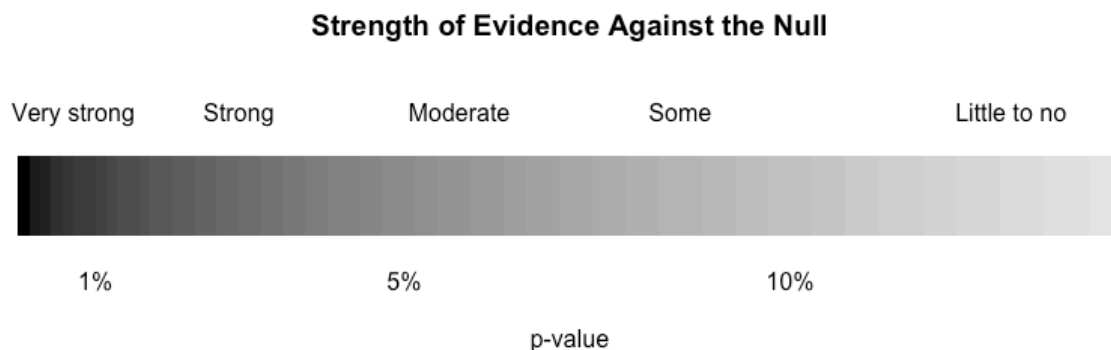
p-value

25. What does the p-value measure?: Interpret the p-value in context of the problem.

Communicate the results and answer the research question

When we write a conclusion we answer the research question by stating how much evidence there is for the alternative hypothesis.

26. Write a conclusion in context of the study.

27. Write a paragraph summarizing the results as if you were writing a press release. Be sure to describe:

- Summary statistic
- P-value and interpretation
- Conclusion (written to answer the research question)
- Generalization — to what group do the results apply?

Revisit and look forward

28. Suggest a new research question that you might investigate, building on what you learned in this study.

6.2.5 Out-of-class activity

The in-class activity covered simulation-based methods for hypothesis tests involving a single categorical variable. The remaining questions cover theory-based methods for testing a single categorical variable. Use Section 5.3.3 in the textbook and the OnePropTheory video to complete the following questions.

The sampling distribution of a single proportion — how that proportion varies from sample to sample — can be mathematically modeled using the normal distribution if certain conditions are met.

Conditions for the sampling distribution of $\hat{p}$ to follow an approximate normal distribution:

- **Independence**: The sample's observations are independent, e.g., are from a simple random sample. (*Remember*: This also must be true to use simulation methods!)

- **Success-failure condition**: We *expect* to see at least 10 successes and 10 failures in the sample, $n\pi \geq 10$ and $n(1 - \pi) \geq 10$.

1. We already verified that the independence condition is satisfied in question 6, since the independence condition is required for both simulation-based and theory-based methods. Is the success-failure condition met to model the data with the normal distribution? Show your work to support your answer. Hint: We don't know the true value of the parameter, π, so we use the null value, π_0, to check the success-failure condition.

To calculate the standardized statistic we use the general formula

$$Z = \frac{\text{point estimate} - \text{null value}}{SE_0(\text{point estimate})}.$$

For a single categorical variable the standardized sample proportion is calculated using

$$Z = \frac{\hat{p} - \pi_0}{SE_0(\hat{p})},$$

where the standard error is calculated using the null value:

$$SE_0(\hat{p}) = \sqrt{\frac{\pi_0(1 - \pi_0)}{n}}$$

.

2. Calculate the null standard error of the sample proportion.

3. Calculate the standardized sample proportion.

The standardized statistic is used as a ruler to measure how far the sample statistic is from the null value. Essentially, we are converting the sample proportion into a measure of standard errors to compare to the standard normal distribution.
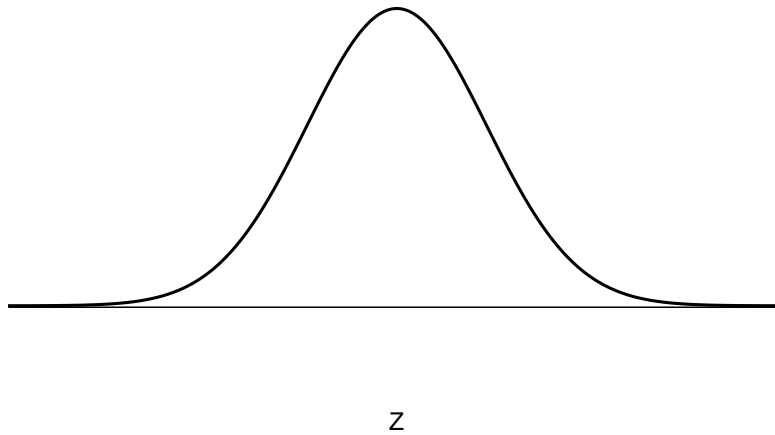

Z

Figure 6.1: A standard normal curve.

4. Using the 68-95-99.7 rule in Section 5.2.5 to guide you, fill in values on the x-axis of the standard normal distribution displayed in Figure 6.1, and also mark the value of the standardized statistic calculated in question 3.

The standardized statistic measures the *number of standard errors the sample statistic is from the null value.*

5. Interpret the standardized sample proportion from question 3 in context of the problem.

We will use the `pnorm()` function in R to find the p-value. Use the provided R script file and enter the value of the standardized statistic calculated in question 3 at `xx` in line 18; highlight and run lines 18–20. Notice that in line 20 it says `lower.tail = FALSE`. R will calculate the p-value *greater* than the value of the standardized statistic.

Notes:

- Use `lower.tail = TRUE` when doing a left-sided test.
- Use `lower.tail = FALSE` when doing a right-sided test.
- To find a two-sided p-value, use a left-sided test for negative Z or a right-sided test for positive Z, then multiply the value found by 2 to get the p-value.

```
pnorm(xx, # Enter value of standardized statistic
      m=0, s=1 # Using the standard normal mean = 0, sd = 1
      lower.tail=FALSE) # Gives a p-value greater than the standardized statistic
```

6. Report the p-value obtained from the R output.

7. Is the p-value found in question 6 similar to the p-value found using the simulation test? Explain why you would expect this to be true.

6.2.6 Take-home messages

1. In a hypothesis test we have two competing hypotheses, the null hypothesis and the alternative hypothesis. The null hypothesis represents either a skeptical perspective or a perspective of no difference or no effect. The alternative hypothesis represents a new perspective such as the possibility that there has been a change or that there is a treatment effect in an experiment.

2. In a simulation-based test, we create a distribution of possible simulated statistics for our sample if the null hypothesis is true. Then we see if the calculated observed statistic from the data is likely or unlikely to occur when compared to the null distribution.

3. The p-value is the probability of the observed statistic occurring or more extreme if the null hypothesis is true. The farther in the tail of the distribution the observed statistic is, the smaller the probability is (smaller the p-value!). The **smaller** the p-value, the **more** evidence the statistic provides **against** the null hypothesis. (Think carefully about why this makes sense!)

4. A **decision** is a statement about strength of evidence against the null hypothesis: reject the null if the p-value is below a pre-set significance level, and fail to reject the null if the p-value is above a pre-set significance level. When writing a **conclusion** to a hypothesis test, on the other hand, we are answering the research question. Thus, a conclusion is a statement about strength of evidence *for the alternative hypothesis*. Use the guidelines for the strength of evidence throughout this course to assess the evidence against the null hypothesis.

5. To create one simulated sample on the null distribution for a sample proportion, spin a spinner with probability equal to π_0 (the null value), n times or draw with replacement n times from a deck of cards created to reflect π_0 as the probability of success. Calculate and plot the proportion of successes from the simulated sample.

6.2.7 Additional notes

Use this space to summarize your thoughts and take additional notes on this week's activity and material covered.

ACTIVITY 7

Inference for a Single Categorical Variable: Confidence Intervals

---

7.1 Reading Guide: Categorical Inference

Section 5.1.4 (Foundations of inference: Confidence intervals)
Videos

- 5.1

Vocabulary

Confidence interval:

Margin of error:

Formulas

General form of a theory-based confidence interval:

Margin of error:

Example: Martian Alphabet

1. What is the sample statistic presented in this example? What notation would be used to represent this value?

2. Interpret the 95% confidence interval provided in the textbook.

3. The formula for the interval is $34/38 \pm (2 \times 0.08) = 0.89 \pm 0.16$. Calculating that, you should get (0.73, 1.05). Why was the interval shown in the textbook (0.73, 1) instead of (0.73, 1.05)?

Sections 5.3.2 and 5.3.3 (One proportion: Bootstrap confidence intervals and Theory-based inferential methods)

In Section 5.3.3, read only the sub-section on "Confidence interval for π". The other sections were covered last week.

Videos

- 5.3
- OnePropTheory

Reminders from previous sections

n = sample size

$\hat{p}$ = sample proportion

π = population proportion

Confidence interval: a process to determine how large an effect is; a range of plausible values for the parameter; also called 'estimation'.

Margin of error: the value that is added to and subtracted from the sample statistic to create a confidence interval; half the width of a confidence interval.

Parameter: a value summarizing a variable(s) for a population.

Statistic: a value summarizing a variable(s) for a sample.

Sampling distribution: plot of statistics from 1000s of samples of the same size taken from the same population.

Standard deviation of a statistic: the variability of statistics from 1000s of samples; how far, on average, each statistic is from the true value of the parameter.

Standard error of a statistic: estimated standard deviation of a statistic.

Central Limit Theorem: For large sample sizes, the sampling distribution of a sample proportion (or mean) will be approximately normal (bell-shaped and symmetric).

Vocabulary

Point estimate:

Test statistic:

Bootstrapping:

Bootstrapped resample:

Bootstrapped statistic:

Confidence level:

Notes

Purpose of bootstrapping:

How is bootstrapping used?

If we want to find a 90% confidence interval, what percentiles of the bootstrap distribution would we need?

Conditions for the Central Limit Theorem to apply (for the sampling distribution of $\hat{p}$ to be approximately normal)

    Independence:

        Checked by:

    Success-failure condition:

        Checked by:

How can we determine the value of $z^\star$ to use as the multiplier in a confidence interval?

    In R, use `qnorm(mean = __, sd = __, p = __)`.

Select one answer in each set of parentheses: The higher the confidence level, the (larger/smaller) the multiplier, meaning the confidence interval will be (wider/narrower).

Formulas

$SD(\hat{p}) =$

Standard error of the sample proportion when we do not assume the null hypothesis is true:

$SE(\hat{p}) =$

Theory-based confidence interval for a sample proportion:

Margin of error of a confidence interval for a sample proportion:

Example: Organ donations

1. What is the sample statistic presented in this example? What notation would be used to represent this value?

2. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

3. How could we use cards to simulate **one** bootstrapped resample? How many blue cards — to represent what? How many red cards — to represent what? How many times would we draw a card and replace it back in the deck? What would you record once you completed the draw-with-replacement process?

4. Interpret the 95% confidence interval provided in the textbook.

5. Are the results in this example statistically significant? Justify your answer.

6. Are the conditions met to use theoretical methods to analyze these data?

Example: Payday loans

1. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

2. Are the conditions met to use theoretical methods to analyze these data?

3. What is the sample statistic presented in this example? What notation would be used to represent this value?

4. Calculate the standard error of the sample proportion when we do not assume the null hypothesis is true.

5. Calculate the margin of error for a 95% confidence interval for π using 1.96 as the multiplier.

6. Calculate a 95% confidence interval for π using your margin of error calculated above.

7. Interpret the 95% confidence interval provided in the textbook.

8. Are the results in this example statistically significant? Justify your answer.

7.2   Activity: Handedness of Male Boxers — Estimation

7.2.1   Learning objectives

- Use bootstrapping to find a confidence interval for a single proportion.

- Interpret a confidence interval for a single proportion.

7.2.2   Terminology review

In this week's in-class activity, we will introduce simulation-based confidence intervals for a single proportion. Some terms covered in this activity are:

- Parameter of interest

- Bootstrapping

- Confidence interval

To review these concepts, see Chapter 5 in your textbook, focusing on Sections 5.1 through 5.3.

7.2.3   Handedness of male boxers

Last week, we found very strong evidence that the true proportion of male boxers who are left-handed is greater than the general population, 0.1. But what *is* the true proportion of male boxers who are left-handed? We will use this same study to estimate this parameter of interest.

A **point estimate** (our observed statistic) provides a single plausible value for a parameter. However, a point estimate is rarely perfect; usually there is some error in the estimate. In addition to supplying a point estimate of a parameter, a next logical step would be to provide a plausible *range* of values for the parameter. This plausible range of values for the population parameter is called an **interval estimate** or **confidence interval**.

As a reminder: left-handedness is a trait that is found in about 10% of the general population. Past studies have shown that left-handed men are over-represented among professional boxers. The fighting claim states that left-handed men have an advantage in competition. Using the data from this random sample of 500 male boxers, we want to estimate, with a given level of confidence, the true proportion of male boxers who are left-handed.

Recall from the last activity that in the sample of 500 male boxers, 81 were left-handed.

Activity intro. Complete Q1–Q4 before class.

1. What is the value of the point estimate?

2. If we took another random sample of 500 male boxers, would you get the exact same point estimate? Explain why or why not.

In this week's activity, we will use bootstrapping to find a 95% confidence interval for π, the true proportion of male boxers who are left-handed. See Section 5.3.2 to review bootstrapping.

3. In your own words, explain the bootstrapping process.

4. Write the conclusion to your test from question 26 in Activity 6.

Use statistical analysis methods to draw inferences from the data

5. Write out the parameter of interest for this study in words.

To use the computer simulation to create a bootstrap distribution, we will need to enter the

- "sample size" (the number of observational units or cases in the sample),
- "number of successes" (the number of cases that are left-handed),
- "number of repetitions" (the number of samples to be generated), and
- the "confidence level" (which level of confidence are we using to create the confidence interval).

6. What values should be entered for each of the following into the simulation to create the bootstrap distribution of sample proportions to find a 95% confidence interval?

- Sample size:

- Number of successes:

- Number of repetitions:

- Confidence level (as a decimal):

We will use the `one_proportion_bootstrap_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of sample proportions and calculate a confidence interval. Using the provided R script file, fill in the values/words for each `xx` with your answers from question 6 in the one proportion bootstrap confidence interval (CI) code to create a bootstrap distribution with 1000 simulations. Then highlight and run lines 1–11.

```
one_proportion_bootstrap_CI(sample_size = xx, # Sample size
                    number_successes = xx, # Observed number of successes
                    number_repetitions = 1000, # Number of bootstrap samples to use
                    confidence_level = 0.95) # Confidence level as a decimal
```

7. Sketch the bootstrap distribution created below.

8. What is the value at the center of this bootstrap distribution? Why does this make sense?

9. Explain why the two vertical lines are at the 2.5th percentile and the 97.5th percentile.

10. Report the 95% bootstrapped confidence interval for π. Use interval notation: (lower value, upper value).

11. Interpret the 95% confidence interval in context.

Communicate the results and answer the research question

12. Does the confidence interval confirm your conclusion from activity 6? Explain your answer.

Effect of confidence level

13. Suppose instead of finding a 95% confidence interval, we found a 90% confidence interval. Would you expect the 90% confidence interval to be narrower or wider? Explain your answer.

14. The following `R` code produced the bootstrap distribution with 1000 simulations that follows. Circle the value that changed in the code.

```
one_proportion_bootstrap_CI(sample_size = 500, # Sample size
                number_successes = 81, # Observed number of successes
                number_repetitions = 1000, # Number of bootstrap samples to use
                confidence_level = 0.90) # Confidence level as a decimal
```


Bootstrapped values of the proportion
90% CI: (0.134, 0.188)
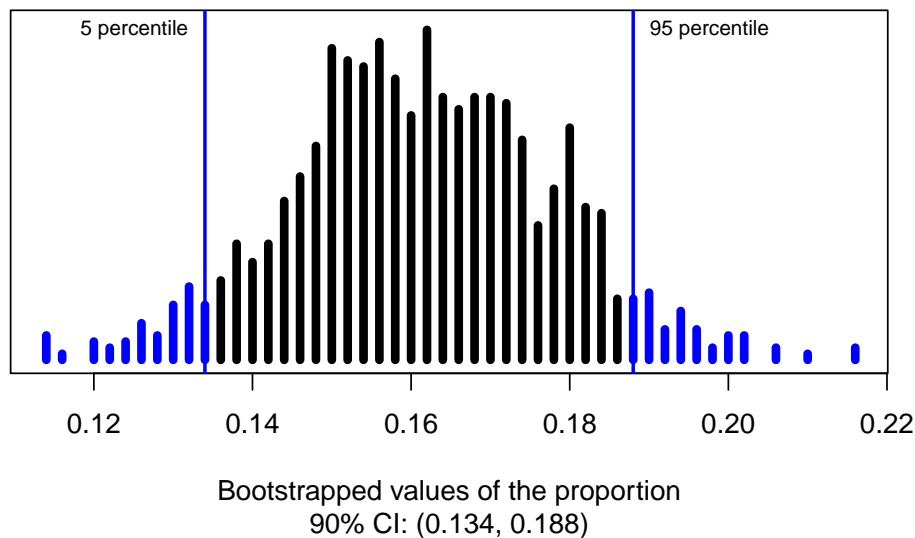
15. Report both the 95% confidence interval (question 10) and the 90% confidence interval (question 14). Is the 90% confidence interval narrower or wider than the 95% confidence interval?

7.2.4 What does *confidence* mean?

In the interpretation of a 95% confidence interval, we say that we are 95% confident that the parameter is within the confidence interval. Why are we able to make that claim? What does it mean to say "we are 95% confident"?

16. Go to this website, http://www.rossmanchance.com/ISIapplets.html and choose 'Simulating Confidence Intervals'. In the input on the left-hand side of the screen enter 0.1 for π, 500 for n, and 100 for 'Intervals'. Click 'sample'.

  a) In the graph on the bottom right, click on a green dot. Write down the confidence interval for this sample given on the graph on the left. Does this confidence interval contain the null value of 0.1?

  b) Now click on a red dot. Write down the confidence interval for this sample. Does this confidence interval contain the null value of 0.1.?

  c) How many intervals out of 100 contain π, the null value of 0.1? *Hint*: This is given to the left of the graph of green and red intervals.

17. Click on 'sample' nine more times. Write down the 'Running Total' for the proportion of intervals that contain π.

18. Interpret the level of confidence. *Hint*: What proportion of samples would we expect to give a confidence interval that contains the parameter of interest?

Revisit and look forward

19. Suggest a new research question that you might investigate, building on what you learned in this study.

7.2.5 Out-of-class activity

The in-class activity covered simulation-based methods for confidence intervals involving a single categorical variable. The remaining questions cover theory-based methods for estimating a single proportion. Use Section 5.3.3 in the textbook and the OnePropTheory video to complete the following questions.

Recall that to use theory-based methods we must check the conditions to approximate the sampling distribution with the normal distribution. From the previous activity, we saw that independence was satisfied as the researchers took a random sample.

To check the success-failure condition to use theory-based methods for confidence intervals, we use $\hat{p}$ in the calculations since we are not assuming a value for π. That is, check that we have at least 10 successes and 10 failures in our **sample**: $n\hat{p} \geq 0$ and $n(1 - \hat{p}) \geq 10$.

1. Verify that the success-failure condition is met to use theory based methods to find a 95% confidence interval.

To calculate a theory-based 95% confidence interval for π, we will first find the **standard error** of $\hat{p}$ by plugging in the value of $\hat{p}$ for π in $SD(\hat{p})$:

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

Note that we do not include a "0" subscript, since we are not assuming a null hypothesis.

2. Calculate the standard error of the sample proportion to find a 95% confidence interval.

To find the confidence interval, we will add and subtract the **margin of error** to the point estimate:

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

$$\hat{p} \pm z^* SE(\hat{p})$$

The z^* multiplier is the percentile of a standard normal distribution that corresponds to our confidence level. If our confidence level is 95%, we find the Z values that encompass the middle 95% of the standard normal distribution. If 95% of the standard normal distribution should be in the middle, that leaves 5% in the tails, or 2.5% in each tail. The `qnorm()` function in `R` will tell us the z^* value for the desired percentile (in this case, 95% + 2.5% = 97.5% percentile).

```
qnorm(0.975) # Multiplier for 95% confidence interval
```

```
#> [1] 1.959964
```

3. Using the multiplier of $z^* = 1.96$, calculate the 95% confidence interval for the true proportion of male boxers who are left-handed.

4. Verify that the simulation-based confidence interval found using bootstrapping is similar to the confidence interval calculating using theory-based methods.

7.2.6 Take-home messages

1. The goal in a hypothesis test is to assess the strength of evidence for an effect, while the goal in creating a confidence interval is to determine how large the effect is. A **confidence interval** is a range of *plausible* values for the parameter of interest.

2. A confidence interval is built around the point estimate or observed calculated statistic from the sample. This means that the sample statistic is always the center of the confidence interval. A confidence interval includes a measure of sample to sample variability represented by the **margin of error**.

3. In simulation-based methods (bootstrapping), a simulated distribution of possible sample statistics is created showing the possible sample to sample variability. Then we find the middle percent of the distribution around the sample statistic using the percentile method to give the range of values for the confidence interval. This shows us that we are $X\%$ confident that the parameter is within these values, where X represents the level of confidence.

4. In theory-based methods, we add and subtract a margin of error to the sample statistic. The margin of error is calculated using a multiplier that corresponds to the level of confidence times the variability (standard error) of the statistic.

5. When the null value is within the confidence interval, it is a plausible value for the parameter of interest; thus, we would find a larger p-value for a hypothesis test of that null value. Conversely, if the null value is NOT within the confidence interval, we would find a small p-value for the hypothesis test and strong evidence against this null hypothesis.

6. To create one simulated sample on the bootstrap distribution for a sample proportion, label n cards with the original responses. Draw with replacement n times. Calculate and plot the resampled proportion of successes.

7.2.7 Additional notes

Use this space to summarize your thoughts and take additional notes on this week's activity and material covered.

Inference for Two Categorical Variables: Hypothesis Testing

8.1 Reading Guide: Hypothesis Testing for a Difference in Proportions

Sections 5.4.1 and 5.4.2 (Simulation tests for a difference in proportions; Two-sided hypotheses)

Videos

- 5.4
- TwoPropSim

Reminders from previous sections

n = sample size

$\hat{p}$ = sample proportion

π = population proportion

General steps of a hypothesis test:

1. Frame the research question in terms of hypotheses.

2. Collect and summarize data using a test statistic.

3. Assume the null hypothesis is true, and simulate or mathematically model a null distribution for the test statistic.

4. Compare the observed test (standardized) statistic to the null distribution to calculate a p-value.

5. Make a conclusion based on the p-value and write the conclusion in context.

Parameter: a value summarizing a variable(s) for a population.

Statistic: a value summarizing a variable(s) for a sample.

Sampling distribution: plot of statistics from 1000s of samples of the same size taken from the same population.

Standard deviation of a statistic: the variability of statistics from 1000s of samples; how far, on average, each statistic is from the true value of the parameter.

Standard error of a statistic: estimated standard deviation of a statistic.

Hypothesis test: a process to determine how strong the evidence of an effect is.

Also called a 'significance test'.

Simulation-based method: Simulate lots of samples of size n under assumption of the null hypothesis, then find the proportion of the simulations that are at least as extreme as the observed sample statistic.

Theory-based method: Develop a mathematical model for the sampling distribution of the statistic under the null hypothesis and use the model to calculate the probability of the observed sample statistic (or one more extreme) occurring.

Null hypothesis (H_0): the skeptical perspective; no difference; no change; no effect; random chance; what the researcher hopes to prove is **wrong**.

Alternative hypothesis (H_A): the new perspective; a difference/increase/decrease; an effect; not random chance; what the researcher hopes to prove is **correct**.

Null value: the value of the parameter when we assume the null hypothesis is true (labeled as $parameter_0$).

Null distribution: the simulated or modeled distribution of statistics (sampling distribution) we would expect to occur if the null hypothesis is true.

P-value: probability of seeing the observed sample data, or something more extreme, assuming the null hypothesis is true.

$\implies$ Lower the p-value the stronger the evidence AGAINST the null hypothesis and FOR the alternative hypothesis.

Decision: a determination of whether to 'reject' or 'fail to reject' a null hypothesis based on a p-value and a pre-set level of significance.

Significance level (α): a threshold used to determine if a p-value provides enough evidence to reject the null hypothesis or not.

Common levels of α include 0.01, 0.05, and 0.10.

Statistically significant: results are considered statistically significant if the p-value is below the significance level.

Vocabulary

Randomization test:

Relative risk:

One-sided hypothesis test:

Two-sided hypothesis test:

Notes

In a randomization test involving two categorical variables, how many cards will you need and how will the cards be labeled?

Why, in the randomization test, are the cards all shuffled together and randomly dealt into two new groups?

After shuffling, how many cards are dealt into each pile?

Interpreting relative risk $(RR = \frac{\hat{p_1}}{\hat{p_2}})$

    The proportion of success in group 1 is _____ times the proportion of success in group 2.

    The proportion of success in group 1 is _____ % higher/lower than in group 2.

Write the null hypothesis in notation for a test of relative risk.

How does the p-value in a two-sided test compare to the p-value in a one-sided test?

Formulas

Relative risk =

Notation

Sample size of group 1:

Sample size of group 2:

Sample proportion of group 1:

Sample proportion of group 2:

Population proportion of group 1:

Population proportion of group 2:

Example: Gender discrimination

1. What is the research question?

2. What are the observational units?

3. What type of study design was used? Justify your answer.

4. What is the appropriate scope of inference for these data?

5. What is the sample statistic presented in this example? What notation would be used to represent this value?

6. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

7. Write the null and the alternative hypotheses in words.

8. Write the null and the alternative hypotheses in notation.

9. How could we use cards to simulate **ome** sample *which assumes the null hypothesis is true*? How many blue cards — to represent what? How many red cards — to represent what? What would we do with the cards? What would you record once you have a simulated sample?

10. How can we calculate a p-value from the simulated null distribution for this example?

11. What was the p-value of the test?

12. At the 5% significance level, what decision would you make?

13. What conclusion should the researcher make?

14. Are the results in this example statistically significant? Justify your answer.

Example: Opportunity cost

1. What is the research question?

2. What are the observational units?

3. What type of study design was used? Justify your answer.

4. What is the appropriate scope of inference for these data?

5. What is the sample statistic presented in this example? What notation would be used to represent this value?

6. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

7. Write the null and the alternative hypotheses in words.

8. Write the null and the alternative hypotheses in notation.

9. How could we use cards to simulate **one** sample *which assumes the null hypothesis is true*? How many blue cards — to represent what? How many red cards — to represent what? What would we do with the cards? What would you record once you have a simulated sample?

10. How can we calculate a p-value from the simulated null distribution for this example?

11. What was the p-value of the test?

12. Interpret the p-value in the context of the problem.

13. At the 5% significance level, what decision would you make?

14. What conclusion should the researcher make?

15. Are the results in this example statistically significant? Justify your answer.

Example: CPR and blood thinner

1. What is the research question?

2. What are the observational units?

3. What type of study design was used? Justify your answer.

4. What is the appropriate scope of inference for these data?

5. What is the sample difference in proportions presented in this example? What notation would be used to represent this value?

6. What is the sample relative risk? Interpret the value in the context of the study.

7. What is the parameter (using a difference in proportion) representing in the context of this problem? What notation would be used to represent this parameter?

8. Write the null and the alternative hypotheses in words.

9. Write the null and the alternative hypotheses in notation.

10. How could we use cards to simulate **one** sample *which assumes the null hypothesis is true*? How many blue cards — to represent what? How many red cards — to represent what? What would we do with the cards? What would you record once you have a simulated sample?

11. How can we calculate a p-value from the simulated null distribution for this example?

12. What was the p-value of the test?

13. Interpret the p-value in the context of the problem.

14. At the 5% significance level, what decision would you make?

15. What conclusion should the researcher make?

16. Are the results in this example statistically significant? Justify your answer.

Section 5.4.4 (Theory-based methods for a difference in proportions)

You may skip the sub-section on "Confidence Interval for $\pi_1 - \pi_2$". This section will be covered next week.

Videos

- 5.4

Reminders from previous sections

Sample size of group 1: n_1

Sample size of group 2: n_2

Sample proportion of group 1: $\hat{p}_1$

Sample proportion of group 2: $\hat{p}_2$

Population proportion of group 1: π_1

Population proportion of group 2: π_2

Test statistic/Point estimate: other names for a statistic from a sample; the point estimate is our best guess for the parameter of interest.

Central Limit Theorem: For large sample sizes, the sampling distribution of a sample proportion (or mean) will be approximately normal (bell-shaped and symmetric).

Notes

Conditions for the CLT to apply for a difference in proportions

Independence:

Checked by:

Success-failure condition:

Checked by:

Formulas

$SD(\hat{p}_1 - \hat{p}_2) =$

Null standard error of the difference in sample proportions: $SE_0(\hat{p}_1 - \hat{p}_2) =$

Standardized statistic/standardized difference in sample proportions: $Z =$

Notation

Overall (pooled) proportion of successes:

Example: CPR and blood thinner

1. What are the observational units?

2. What type of study design was used? Justify your answer.

3. What is the appropriate scope of inference for these data?

4. What is the sample difference in proportions presented in this example? What notation would be used to represent this value?

5. What is the parameter (using a difference in proportion) representing in the context of this problem? What notation would be used to represent this parameter?

6. Write the null and the alternative hypotheses in words.

7. Write the null and the alternative hypotheses in notation.

8. Is it valid to use theory-based methods to analyze these data?

9. Calculate the pooled or overall proportion of successes. What notation would be used to represent this value?

10. Calculate the null standard error of the difference in sample proportions.

11. Calculate the standardized statistic

12. Interpret the standardized statistic in the context of the problem.

Note: a p-value, p-value interpretation, decision, and conclusion for this example can be found in the Reading Guide solutions for Sections 5.4.1–5.4.3.

Section 5.5 (Errors, power, and practical importance)

Videos

- 5.5
- Errors_Power

Reminders from previous sections

Decision: a determination of whether to reject or fail to reject a null hypothesis based on a p-value and a pre-set level of significance.

- If p-value $\leq \alpha$, then reject H_0.

- If p-value $> \alpha$, then fail to reject H_0.

Significance level (α): a threshold used to determine if a p-value provides enough evidence to reject the null hypothesis or not.

  Common levels of α include 0.01, 0.05, and 0.10.

Statistically significant: results are considered statistically significant if the p-value is below the significance level.

Vocabulary

Type 1 error:

Type 2 error:

Confirmation bias:

Power:

Practical importance:

Notes

Fill in the following table with whether the decision was correct or not, and if not, what type of error was made.

	Test conclusion (based on data)	
Truth (unknown)	Reject null hyp.	Fail to reject null hyp.
H_0 is true		
H_A is true (H_0 is false)		

How are the significance level and type I error rate related?

How are the significance level and type II error rate related?

After collecting data, a researcher decides to change from a two-sided test to a one-sided test. Why is this a bad idea?

1. It _____ (increases/decreases) the chance of a type I error.

2. This can result in _____.

How are power and type I error rate related?

How are power and type II error rate related?

How can we increase the power of a test?

1. _____ (Increase/Decrease) the significance level

2. _____ (Increase/Decrease) the sample size

3. Change from a _____ (one/two)-sided to a _____ (one/two)-sided test

4. Have a _____ (larger/smaller) standard deviation of the statistic

5. Have the alternative parameter value _____ (closer/farther) from the null value

Results are likely to be statistically significant (but may not be practically important) if the sample size is _____(large/small).

Results are unlikely to be statistically significant (but may be practically important) if the sample size is _____(large/small).

Examples:

1. In the Gender Discrimination study in the textbook and presented as an example in Reading Guide 5.4.1–5.4.2,

    a. What was the p-value of the test?

    b. At the 5% significance level, what decision would you make?

    c. What type of error might have occurred in these data?

    d. Interpret that error in the context of the problem.

2. In the Opportunity Cost study in the textbook and presented as an example in the reading guide for sections 5.4.1–5.4.2,

   a. What was the p-value of the test?

   b. At the 5% significance level, what decision would you make?

   c. What type of error might have occurred in these data?

   d. Interpret that error in the context of the problem.

3. In the CPR and Blood Thinners study in the textbook and presented as an example in the reading guide for sections 5.4.1–5.4.2,

   a. What was the p-value of the test?

   b. At the 5% significance level, what decision would you make?

   c. What type of error might have occurred in these data?

   d. Interpret that error in the context of the problem.

8.2 Activity: Winter Sports Helmet Use and Head Injuries — Testing

8.2.1 Learning objectives

- Given a research question involving two categorical variables, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Assess the conditions to use the normal distribution model for a difference in proportions.

- Calculate the Z test statistic for a difference in proportions.

- Find, interpret, and evaluate the p-value for a theory-based hypothesis test for a difference in proportions.

8.2.2 Terminology review

In this week's in-class activity, we will use theory-based methods to analyze two categorical variables. Some terms covered in this activity are:

- Conditional proportion

- Z test

- z^* multiplier

- Null hypothesis

- Alternative hypothesis

- Test statistic

- Standard normal distribution

- Independence and success-failure conditions

- Type 1 and Type 2 errors

- Decision of a hypothesis test

To review these concepts, see Chapter 5 in your textbook.

8.2.3 Helmet use and head injuries

In "Helmet Use and Risk of Head Injuries in Alpine Skiers and Snowboarders" by Sullheim et. al., in the *Journal of the American Medical Association*, Vol. 295, No. 8 (2006), we can see the summary results from a random sample of 3562 skiers and snowboarders involved in accidents in the two-way table below. Is there evidence that safety helmet use is associated with a reduced risk of head injury for skiers and snowboarders?

	Helmet Use	No Helmet Use	Total
Head Injury	96	480	576
No Head Injury	656	2330	2986
Total	752	2810	3562

These counts can be found in `R` by using the `count()` function:

```r
# Read data set in
injury <- read.csv("https://math.montana.edu/courses/s216/data/HeadInjuries.csv")
injury <- # Write over original data with the following
  injury %>% # Pipe data set into
  mutate(Helmet <- factor(Helmet),
         Outcome <- factor(Outcome)) # Convert to factors

injury %>% group_by(Helmet) %>% count(Outcome)
```

```
#> # A tibble: 4 x 3
#> # Groups:   Helmet [2]
#>   Helmet Outcome           n
#>   <chr>  <chr>         <int>
#> 1 No     Head Injury     480
#> 2 No     No Head Injury 2330
#> 3 Yes    Head Injury      96
#> 4 Yes    No Head Injury  656
```

Vocabulary review. Complete Q1–Q4 before class.

1. What is the name of the explanatory variable in the `R` output? What are its categories?

2. What is the response variable in the `R` output? What are its categories?

3. Fill in the blanks with one answer from each set of parentheses: This is an

   _____ (experiment/observational study) because

   _____ (helmet use/head injury) _____ (was/was not)

   randomly _____ (assigned/selected).

4. Put an X in the box that represents the appropriate scope of inference for this study.

		Study Type	
		Randomized Experiment	Observational Study
Selection of Cases	Random Sample		
	No Random Sample		

Ask a research question

The research question as stated above is: Is there evidence that safety helmet use is associated with a reduced risk of head injury for skiers and snowboarders? In order to set up our hypotheses, we need to express this research question in terms of parameters. Remember, we define the parameter for a single categorical variable as the true proportion of observational units that are labeled as a "success" in the response variable.

5. Write the two parameters of interest for this study. Let 1 = skier/snowboarder wore helmet, 2 = skier/snowboarder did not wear helmet.

π_1 —

π_2 —

When comparing two groups, we assume the two parameters are equal in the null hypothesis—there is no association between the variables.

6. Write the null hypothesis out in words using your answers to question 5.

7. Based on the research question, fill in the appropriate sign for the alternative hypothesis ($<$, $>$, or $\neq$):

$H_A : \pi_1 - \pi_2$ _____ 0

Summarize and visualize the data

8. Using the two-way table above, calculate the conditional proportion of helmet-wearing skiers/snowboarders that sustained a head injury.

9. Using the two-way table above, calculate the conditional proportion of non-helmet-wearing skiers/snowboarders that sustained a head injury.

Segmented Bar Plot of Helmet Use by Injury
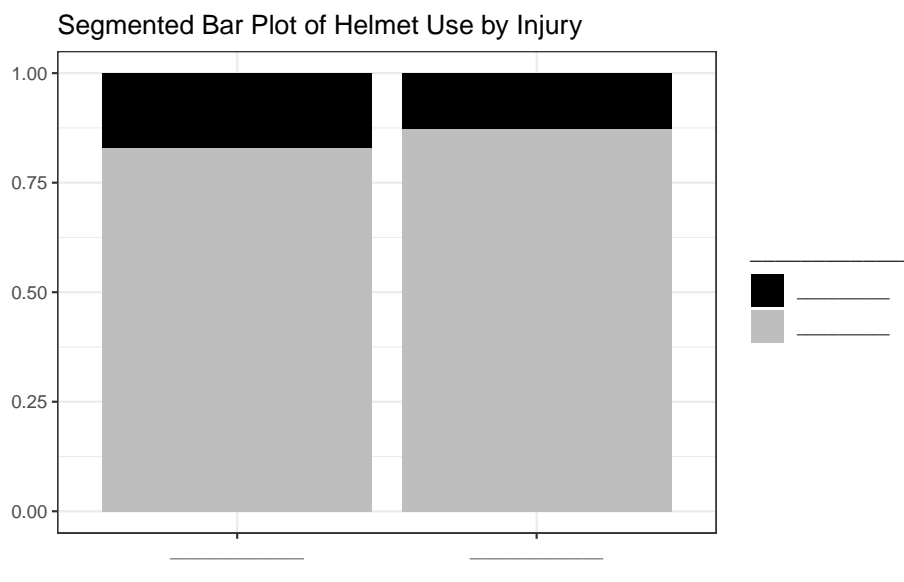
10. Fill in the blanks on the segmented bar plot on the previous page with the appropriate variable names and categories to complete the segmented bar plot comparing the proportion of head injuries between those who wear helmets and those who do not wear helmets. *Hint*: Use the conditional proportions from questions 8 and 9.

11. Based on the segmented bar plot, Does there appear to be an association between helmet use and head injury? Explain using the plot.

12. Calculate the summary statistic for this study. Use helmet use (`Yes`) minus no helmet use (`No`) as the order of subtraction.

13. What is the notation used for the value calculated in question 12?

Use statistical analysis methods to draw inferences from the data

To test the null hypothesis, we could use simulation-based methods as we did with a single categorical variable in class. In this in-class activity, we will focus on theory-based methods. Like with a single proportion, the sampling distribution of a difference in sample proportions can be mathematically modeled using the normal distribution if certain conditions are met.

Conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to follow an approximate normal distribution:

- **Independence**: The data are independent within and between the two groups. (*Remember*: This also must be true to use simulation methods!)

- **Success-failure condition**: The success-failure condition holds for each group. Under the null hypothesis, the proportions π_1 and π_2 are equal, so we check the success-failure condition with our best estimate of these values under H_0, the pooled proportion from the two samples,

$$\hat{p}_{pool} = \frac{\text{number of "successes"}}{\text{number of cases}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

We then check that all four of the following inequalities hold:

$$\hat{p}_{pool} \times n_1 \geq 10, \qquad (1 - \hat{p}_{pool}) \times n_1 \geq 10,$$
$$\hat{p}_{pool} \times n_2 \geq 10, \qquad (1 - \hat{p}_{pool}) \times n_2 \geq 10$$

14. Is the independence condition met? Explain your answer.

15. Is the success-failure condition met for each group? Show your work to verify your answer.

To calculate the standardized statistic we use:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE},$$

where the null standard error is calculated using the pooled proportion of successes:

$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}_{pool}(1 - \hat{p}_{pool})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

16. Calculate $SE_0(\hat{p}_1 - \hat{p}_2)$.

17. Calculate the standardized statistic.

We will use the `pnorm()` function in `R` to find the p-value. Use the provided `R` script file and enter the value of the standardized statistic found in question 17 at `xx` in line 27; highlight and run lines 27–29.

```
pnorm(xx, # Enter value of standardized statistic
      m=0, s=1 # Using the standard normal mean = 0, sd = 1
      lower.tail=TRUE) # Gives a p-value less than the standardized statistic
```

18. Report the p-value from the `R` output.

19. Interpret the p-value in context of the study.

20. How much evidence does the p-value provide against the null hypothesis? *Hint*: Refer to the guidelines given in Activity 6.

21. Write a conclusion to the test.

Table 8.3: Four different possible scenarios for hypothesis test decisions.

		Test conclusion	
		Fail to reject H_0	Reject H_0
	H_0 true	Good decision	Type 1 Error
Truth	H_A true	Type 2 Error	Good decision

Types of errors

Hypothesis tests are not flawless. In a hypothesis test, there are two competing hypotheses: the null and alternative. We make a decision about which might be true, but we may choose incorrectly.

Shown in Table 8.3, a **Type 1 Error** happens when we reject the null hypothesis when H_0 is actually true. A **Type 2 Error** happens when we fail to reject the null hypothesis when the alternative is actually true.

22. Using a significance level of 0.05, based on the p-value found in question 18, what decision do you make in regards to the null hypothesis?

23. What type of error could we have made?

24. Write this error in context of the problem.

25. Write a paragraph summarizing the results of the study as if writing a press release. Be sure to describe:

- Summary statistic
- Test statistic and interpretation
- P-value and interpretation
- Conclusion (written to answer the research question)
- Scope of inference

8.2.4 Out-of-class activity

The remaining questions cover simulation-based methods for testing two categorical variables. Use Section 5.4.1 in the textbook and the TwoPropSim video to complete the following questions.

1. First, let's think about how one simulation would be created on the null distribution using cards.

   How many cards would you need?

   What would be written on each card?

2. Next, we would mix the cards together and shuffle into two piles. How many cards would be in each pile? What would each pile represent?

3. Once we have one simulated sample, what would we calculate and plot on the null distribution? *Hint*: What statistic are we calculating from the data?

To create the null distribution of differences in sample proportions, we will use the `two_proportion_test()` function in `R` (in the `catstats` package). We will need to enter the response variable name and the explanatory variable name for the formula, the data set name (identified above as `injury`), the outcome for the explanatory variable that is first in subtraction, number of repetitions, the outcome for the response variable that is a success (what the numerator counts when calculating a sample proportion), and the direction of the alternative hypothesis.

The response variable name is `Outcome` and the explanatory variable name is `Helmet`.

4. What inputs should be entered for each of the following to create the simulation?

   - First in subtraction (What is the outcome for the explanatory variable that is used as first in the order of subtraction? `"Yes"` or `"No"`):

   - Number of repetitions:

   - Response value numerator (What is the outcome for the response variable that is considered a success? `"Head Injury"` or `"No Head Injury"`):

   - As extreme as (enter the value for the sample difference in proportions):

   - Direction (`"greater"`, `"less"`, or `"two-sided"`):

Using the R script file for this activity, enter your answers for question 4 in place of the **xx**'s to produce the null distribution with 1000 simulations; highlight and run lines 1–12 and then 33–39.

```
two_proportion_test(formula = Outcome ~ Helmet, # response ~ explanatory
    data= injury, # Name of data set
    first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1
    number_repetitions = 1000, # Always use a minimum of 1000 repetitions
    response_value_numerator = "xx", # Define which outcome is a success
    as_extreme_as = xx, # Calculated observed statistic (difference in sample proportions)
    direction="xx") # Alternative hypothesis direction ("greater","less","two-sided")
```

5. Sketch the null distribution created here.

6. What value is the null distribution centered around? Explain why this makes sense.

7. What is the p-value? *Remember*: This is the value given at the bottom of the null distribution.

8. Is the p-value found in question 7 for the out-of-class activity similar to the p-value found using the theory-based test? Explain why you would expect this to be true.

8.2.5   Take-home messages

1. When comparing two groups, we are looking at the difference between two parameters. In the null hypothesis, we assume the two parameters are equal, or that there is no difference between the two proportions.

2. We use the same guidelines for the strength of evidence as we did in Activity 6.

3. The standardized statistic when the response variable is categorical is a Z-score and is compared to the standard normal distribution to find the p-value. To find the standardized statistic, we take the value of the statistic minus the null value, divided by the null standard error of the statistic. The standardized statistic measures the number of standard errors the statistic is from the null value.

4. If we make the decision to reject the null hypothesis (the p-value is less than the significance level), we could have a possible Type 1 error. A Type 1 error occurs when we reject a true null hypothesis (false positive).

5. If we make the decision to fail to reject the null hypothesis (the p-value is greater than the significance level), we could have a possible Type 2 error. A Type 2 error occurs when we fail to reject a false null hypothesis (false negative).

6. To create one simulated sample on the null distribution for a difference in sample proportions, label $n_1 + n_2$ cards with the response variable outcomes from the original data. Mix cards together and shuffle into two new groups of sizes n_1 and n_2, representing the explanatory variable groups. Calculate and plot the difference in proportion of successes.

8.2.6   Additional notes

Use this space to summarize your thoughts and take additional notes on this week's activity and material covered.

Inference for Two Categorical Variables: Confidence Intervals

9.1 Reading Guide: Confidence Intervals for a Difference in Proportions

Section 5.4.3 (Bootstrap confidence interval for a difference in proportions)

Videos

- 5.4

Reminders from previous sections

Sample size of group 1: n_1

Sample size of group 2: n_2

Sample proportion of group 1: $\hat{p}_1$

Sample proportion of group 2: $\hat{p}_2$

Population proportion of group 1: π_1

Population proportion of group 2: π_2

Parameter: a value summarizing a variable(s) for a population.

Statistic: a value summarizing a variable(s) for a sample.

Point estimate: another name for a statistic from a sample; our best guess for the parameter of interest.

Sampling distribution: plot of statistics from 1000s of samples of the same size taken from the same population.

Standard deviation of a statistic: the variability of statistics from 1000s of samples; how far, on average, each statistic is from the true value of the parameter.

Standard error of a statistic: estimated standard deviation of a statistic.

Confidence interval: a process to determine how large an effect is; a range of plausible values for the parameter. Also called 'estimation'.

Margin of error: the value that is added to and subtracted from the sample statistic to create a confidence interval; half the width of a confidence interval.

Bootstrapping: the process of drawing with replacement n times from the original sample.

Bootstrapped resample: a random sample of size n from the original sample, selected with replacement.

Bootstrapped statistic: the statistic recorded from the bootstrapped resample.

Confidence level: how confident we are that the confidence interval will capture the parameter.

Notes

To create a single bootstrap resample for two categorical variables, how many cards will you need and how will the cards be labeled?

What is done with the cards once they are labeled?

Interpretations of confidence level must include:

How do you determine if the results of a hypothesis test agree with a confidence interval?

How are the confidence level and the significance level related (for a two-sided test)?

Example: CPR and blood thinner

1. What is the research question?

2. What is the sample difference in proportions presented in this example? What notation would be used to represent this value?

3. What is the parameter (using a difference in proportion) representing in the context of this problem? What notation would be used to represent this parameter?

4. How could we use cards to simulate **one** bootstrap resample? How many blue cards — to represent what? How many red cards — to represent what? What would we do with the cards? What would you record once you have a simulated sample?

5. How can we calculate a 90% confidence interval from the bootstrap distribution for this example?

6. What was the 90% confidence interval?

7. Interpret the confidence *interval* in the context of the problem.

8. Interpret the confidence *level* in the context of the problem.

9. Does the conclusion of the hypothesis test match the confidence interval?

Section 5.4.4 (Theory-based methods for a difference in proportions)

In section 5.4.4, read only the sub-section on "Confidence interval for $\pi_1 - \pi_2$". The other sections were covered last week.

Videos

- 5.4

Reminders from previous sections

Central Limit Theorem: For large sample sizes, the sampling distribution of a sample proportion (or mean) will be approximately normal (bell-shaped and symmetric).

Notes

Conditions for the CLT to apply for two categorical variables

Independence:

Checked by:

Success-failure condition:

Checked by:

Formulas

$SD(\hat{p_1} - \hat{p_2}) =$

Standard error of the difference in sample proportions when we do not assume the null hypothesis is true: $SE(\hat{p_1} - \hat{p_2}) =$

Theory-based confidence interval for a difference in proportions:

Margin of error of a confidence interval for a difference in proportions:

Example: CPR and blood thinner

1. What is the sample difference in proportions presented in this example? What notation would be used to represent this value?

2. What is the parameter (using a difference in proportion) representing in the context of this problem? What notation would be used to represent this parameter?

3. Calculate the standard error of the difference in sample proportions without assuming a null hypothesis.

4. Calculate the 90% confidence interval using $z^* = 1.65$ as the multiplier.

Note: A confidence interval interpretation and confidence level interpretation for this example can be found in the Reading Guide solutions for Sections 5.4.1–5.4.3.

9.2 Activity: Winter Sports Helmet Use and Head Injuries — Estimation

9.2.1 Learning objectives

- Assess the conditions to use the normal distribution model for a difference in proportions.
- Create and interpret a theory-based confidence interval for a difference in proportions.

9.2.2 Terminology review

In this week's activity, we will use theory-based methods to estimate the difference in two proportions. Some terms covered in this activity are:

- Standard normal distribution
- Independence and success-failure conditions

To review these concepts, see Chapter 5 in your textbook.

9.2.3 Helmet use and head injuries

In Activity 8, we found very strong evidence that helmet use is associated with a reduction in head injury for skiers and snowboarders. In this in-class activity we will estimate the parameter of interest, the difference in true proportion of skiers and snowboarders with head injuries for those who wore helmets and those that did not, using theory-based methods.

The summary results from a random sample of 3562 skiers and snowboarders involved in accidents is displayed in the two-way table below.

	Helmet Use	No Helmet Use	Total
Head Injury	96	480	576
No Head Injury	656	2330	2986
Total	752	2810	3562

Vocabulary review. Complete Q1–Q3 before class.

1. Report the point estimate for this study. Use 'Helmet Use' minus 'No Helmet Use' as the order of subtraction.

2. Write the hypothesis test conclusion for the study from Activity 8.

3. Based on the results from Activity 8, do you expect the 95% confidence interval for the true difference in proportion with head injuries between the two groups to contain the null value of zero? Explain your answer.

Use statistical analysis methods to draw inferences from the data

In this activity we will focus on theory-based methods to calculate a confidence interval. Like with a single proportion, the sampling distribution of a difference in proportions can be mathematically modeled using the normal distribution if certain conditions are met.

Conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to follow an approximate normal distribution:

- **Independence**: The data are independent within and between the two groups. (*Remember*: This also must be true to use simulation methods!)

- **Success-failure condition**: The success-failure condition holds for each group. Since we are not assuming a null hypothesis, we do not use the pooled sample proportion to check this condition as we did in Activity 8. Instead, we use the individual sample proportions $\hat{p}_1$ and $\hat{p}_2$. Equivalently, we check that all cells in the table have at least 10 observations.

4. In Activity 8, we saw that the independence condition was met since the study used a random sample. Is the success-failure condition to find the theory-based confidence interval met for each group? Explain your answer.

To find a confidence interval for the difference in proportions we will add and subtract the margin of error from the point estimate to find the two endpoints.

$$\hat{p}_1 - \hat{p}_2 \pm z^* SE(\hat{p}_1 - \hat{p}_2), \text{ where}$$

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

Note that the formula changes when calculating the variability around the statistic in order to calculate a confidence interval from the formula used in Activity 8! Here, we use the sample proportions for each group to calculate the standard error for the difference in proportions since we are not assuming that the true difference is zero.

5. Calculate the standard error for a difference in proportions to create a 95% confidence interval.

6. Interpret the value calculated in question 5 in context of the problem.

The z^* multiplier is the percentile of a standard normal distribution that corresponds to our confidence level. If our confidence level is 95%, we find the Z values that encompass the middle 95% of the standard normal distribution. If 95% of the standard normal distribution should be in the middle, that leaves 5% in the tails, or 2.5% in each tail. The `qnorm()` function in R will tell us the z^* value for the desired percentile (in this case, 95% + 2.5% = 97.5% percentile).

```
qnorm(0.975) # Multiplier for 95% confidence interval
```

```
#> [1] 1.959964
```

7. Sketch a graph of the standard normal distribution and use the graph to explain how the R code above is used to find the z^* multiplier.

8. Using the multiplier of $z^* = 1.96$ and the standard error found in question 5, calculate the margin of error for a 95% confidence interval.

9. Calculate the 95% confidence interval for the difference in true proportion of head injuries for those that used helmets minus those who did not.

10. Interpret the confidence interval found in question 9 in context of the problem.

9.2.4   Effect of sample size

Suppose in another sample of skiers and snowboarders involved in accidents we saw these results:

	Helmet Use	No Helmet Use	Total
Head Injury	135	674	809
No Head Injury	921	3270	4191
Total	1056	3944	5000

11. Calculate the margin of error for a 95% confidence interval using a multiplier of $z^* = 1.96$ for this new sample. Is the margin of error larger or smaller than the margin of error for the original study?

12. Calculate the 95% confidence interval for this new study using the margin of error from question 11.

13. Is the confidence interval calculated in question 12 with the larger sample size wider or smaller than the confidence interval in question 9? Why?

9.2.5   Relative risk

Another summary statistic that can be calculated for two categorical variables is the relative risk. The relative risk is calculated as the ratio of the conditional proportions:

$$\text{relative risk} = \frac{\hat{p}_1}{\hat{p}_2}.$$

14. Calculate the relative risk of head injury for those who wore helmets compared to those who did not.

15. Interpret the relative risk in context of the problem.

9.2.6 Out-of-class activity

The remaining questions cover simulation-based methods for creating a bootstrap distribution of differences in sample proportions to find a confidence interval. Use Section 5.4.3 in the textbook and the TwoPropSim video to complete the following questions.

We will use the `two_proportion_bootstrap_CI()` function in `R` (in the `catstats` package) to simulate the bootstrap distribution of differences in sample proportions and calculate a confidence interval. We will need to enter the response variable name and the explanatory variable name for the formula, the data set name (identified above as `injury`), the outcome for the explanatory variable that is first in subtraction, number of repetitions, the outcome for the response variable that is a success (what the numerator counts when calculating a sample proportion), and the confidence level as a decimal.

The response variable name is `Outcome` and the explanatory variable name is `Helmet`.

1. What values should be entered for each of the following into the simulation to create a 95% confidence interval?

   - First in subtraction (What is the outcome for the explanatory variable that is used as first in the order of subtraction? `"Yes"` or `"No"`):

   - Response value numerator (What is the outcome for the response variable that is considered a success? `"Head Injury"` or `"No Head Injury"`):

   - Number of repetitions:

   - Confidence level (entered as a decimal):

Using the `R` script file for this activity, enter your answers for question 1 in place of the `xx`'s to produce the bootstrap distribution with 1000 simulations; highlight and run lines 1–22.

```
two_proportion_bootstrap_CI(formula = Outcome ~ Helmet,
      data=injury, # Name of data set
      first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1
      response_value_numerator = "xx", # Define which outcome is a success
      number_repetitions = 1000, # Always use a minimum of 1000 repetitions
      confidence_level = 0.95) # Enter the level of confidence as a decimal
```

2. Report the bootstrap 95% confidence interval.

3. What percentile of the bootstrap distribution does the upper value of the confidence interval represent?

4. Is the bootstrap 95% confidence interval similar to what was found using theory-based methods? Why did you expect this to be true?

9.2.7   Take-home messages

1. Simulation-based methods and theory-based methods should give the same results for a study *if the validity conditions are met*. For both methods, observational units need to be independent. To use theory-based methods, additionally, the success-failure condition must be met. Check the validity conditions for each type of test to determine if theory-based methods can be used.

2. When calculating the standard error for the difference in sample proportions when doing a hypothesis test, we use the pooled proportion of successes, the best estimate for calculating the variability *under the assumption the null hypothesis is true*. For a confidence interval, we are not assuming a null hypothesis, so we use the values of the two conditional proportions to calculate the standard error. Make note of the difference in these two formulas.

3. In addition to estimating the difference in proportions for two categorical variables we can also find the relative risk, the ratio of conditional proportions.

4. Increasing sample size will result in less sample-to-sample variability in statistics, which will result in a smaller standard error, and thus a narrower confidence interval.

5. To create one simulated sample on the bootstrap distribution for a difference in sample proportions, label $n_1 + n_2$ cards with the outcomes for the original responses. Keep groups separate and randomly draw with replacement n_1 times from group 1 and n_2 times from group 2. Calculate and plot the resampled difference in the proportion of successes.

9.2.8   Additional notes

Use this space to summarize your thoughts and take additional notes on this week's activity and material covered.

Inference for a Quantitative Response with Paired Samples

10.1 Reading Guide: Inference for a Single Mean or Paired Mean Difference

Section 6.1 (Inference for one mean)

Videos

- 6.1
- OneMeanTheory

Reminders from previous sections

n = sample size

$\overline{x}$ = sample mean

s = sample standard deviation

μ = population mean

σ = population standard deviation

General steps of a hypothesis test:

1. Frame the research question in terms of hypotheses.

2. Collect and summarize data using a test statistic.

3. Assume the null hypothesis is true, and simulate or mathematically model a null distribution for the test statistic.

4. Compare the observed test statistic to the null distribution to calculate a p-value.

5. Make a conclusion based on the p-value and write the conclusion in context.

Parameter: a value summarizing a variable(s) for a population.

Statistic: a value summarizing a variable(s) for a sample.

Sampling distribution: plot of statistics from 1000s of samples of the same size taken from the same population.

Standard deviation of a statistic: the variability of statistics from 1000s of samples; how far, on average, each statistic is from the true value of the parameter.

Standard error of a statistic: estimated standard deviation of a statistic.

Hypothesis test: a process to determine how strong the evidence of an effect is. Also called a 'significance test'.

Simulation-based method: Simulate lots of samples of size n under assumption of the null hypothesis, then find the proportion of the simulations that are at least as extreme as the observed sample statistic.

Theory-based method: Develop a mathematical model for the sampling distribution of the statistic under the null hypothesis and use the model to calculate the probability of the observed sample statistic (or one more extreme) occurring.

Null hypothesis (H_0): the skeptical perspective; no difference; no change; no effect; random chance; what the researcher hopes to prove is **wrong**.

Alternative hypothesis (H_A): the new perspective; a difference/increase/decrease; an effect; not random chance; what the researcher hopes to prove is **correct**.

Null value: the value of the parameter when we assume the null hypothesis is true (labeled as $parameter_0$).

Null distribution: the simulated or modeled distribution of statistics (sampling distribution) we would expect to occur if the null hypothesis is true.

P-value: probability of seeing the observed sample data, or something more extreme, assuming the null hypothesis is true.

$\implies$ Lower the p-value the stronger the evidence AGAINST the null hypothesis and FOR the alternative hypothesis.

Decision: a determination of whether to reject or fail to reject a null hypothesis based on a p-value and a pre-set level of significance.

- If p-value $\leq \alpha$, then reject H_0.

- If p-value $> \alpha$, then fail to reject H_0.

Significance level (α): a threshold used to determine if a p-value provides enough evidence to reject the null hypothesis or not.

   Common levels of α include 0.01, 0.05, and 0.10.

Statistically significant: results are considered statistically significant if the p-value is below the significance level.

Confidence interval: a process to determine how large an effect is; a range of plausible values for the parameter. Also called 'estimation'.

Margin of error: the value that is added to and subtracted from the sample statistic to create a confidence interval; half the width of a confidence interval.

Bootstrapping: the process of drawing with replacement n times from the original sample.

Bootstrapped resample: a random sample of size n from the original sample, selected with replacement.

Bootstrapped statistic: the statistic recorded from the bootstrapped resample.

Confidence level: how confident we are that the confidence interval will capture the parameter.

Bootstrap $X\%$ confidence interval: $((\frac{(1-X)}{2})^{th}$ percentile, $(X + (\frac{(1-X)}{2})^{th}$ percentile) of a bootstrap distribution.

Central Limit Theorem: For large sample sizes, the sampling distribution of a sample mean (or proportion) will be approximately normal (bell-shaped and symmetric).

Vocabulary

t-distribution:

- The variability in the t-distribution depends on the sample size (used to calculate degrees of freedom — df for short).
- The larger df, the closer the t distribution is to the standard normal distribution.

Degrees of freedom (df):

T-score:

Notes

To create a bootstrap distribution test, how many cards will you need and how will the cards be labeled?

What do you do with the cards after labeling them?

After resampling, what value will be plotted on the bootstrap distribution?

True or false: Bootstrapping can only be used if the sample size is small.

Why do we use a t-distribution rather than the normal distribution when analyzing quantitative data?

How do we calculate degrees of freedom for the t-distribution?

Conditions to use the CLT for means:

    Independence:

        Checked by:

    Normality:

        Checked by:

Formulas

$SE(\overline{x}) =$

$T =$

Confidence interval for a mean:

Notation

μ_0 represents

Example: Edinburgh rentals

1. What are the observational units?

2. What are the sample statistics presented in this example? What notation would be used to represent each value?

3. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

4. How could we use cards to simulate **one** bootstrap resample *which does not assume the null hypothesis is true*? How many cards? What is written on the cards? What would we do with the cards? What would you record once you have a simulated sample?

5. After 1000 resamples are generated, where is the resulting bootstrap distribution centered? Why does that make sense?

6. Based on Figure 6.3, give the confidence interval for each of the following confidence levels.

   90% confidence interval =

   95% confidence interval =

   99% confidence interval =

7. Interpret your 99% confidence interval in the context of the problem.

8. Use Figure 6.4 to determine a 90% confidence interval for the true standard deviation for three bedroom flats in Edinburgh.

Example: Mercury content of dolphin muscle

1. What is the research question?

2. What are the observational units?

3. Can the results of this study be generalized to a larger population? Why or why not?

4. What are the sample statistics presented in this example? What notation would be used to represent each value?

5. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

6. Are the independence and normality conditions satisfied?

7. Calculate the standard error of the sample mean.

8. What distribution should be referenced to find the multiplier for a 95% confidence interval?

9. Using $t^\star = 2.10$, calculate a 95% confidence interval for μ.

10. Interpret the interval calculated in the context of the problem.

Example: Cherry Blossom Race

1. What is the research question?

2. What are the observational units?

3. Can the results of this study be generalized to a larger population? Why or why not?

4. What are the sample statistics presented in this example? What notation would be used to represent each value?

5. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

6. Are the independence and normality conditions satisfied?

7. Write the null and the alternative hypotheses in words.

8. Write the null and the alternative hypotheses in notation.

9. Calculate the standard error of the sample mean.

10. Calculate the T-score (the standardized statistic for the sample mean).

11. What distribution should the T-score be compared to in order to calculate a p-value?

12. What was the p-value of the test?

13. Interpret the p-value in the context of the problem.

14. At the 5% significance level, what decision would you make? What type of error might that be?

15. What conclusion should the researcher make?

16. Are the results in this example statistically significant? Justify your answer.

Section 6.2 (Inference for paired mean difference)

Videos

- 6.2

Vocabulary

Paired data:

    Paired with repeated measures:

    Paired with matching:

Notes

For each of the following scenarios, determine if the two sets of observations are paired or independent.

1. To test whether the IQ is related to genetics, researchers measured the IQ of two biological parents and the IQ of their first-born child. The average parent IQ was compared to the IQ of the first born child.

2. Hoping to see how exercise is related to heart rates, researchers asked a group of 30 volunteers to do either bicycle kicks or jumping jacks for 30 seconds. Volunteer's heart rate was measured at the end of 30 seconds, then the volunteers sat for a 5 minute rest period. At the end of the rest period, the volunteer performed the other activity and their heart rate was measured again. Which activity was done first was randomly assigned.

3. Researchers hoping to look into the effectiveness of blended learning gathered two random samples of 50 8th graders (one at Belgrade Middle School which has 5 full-day instruction currently, the other from Chief Joseph Middle School which utilizes a 2-day on, 3-day off blended learning structure). All 8th graders were given the same lessons and same homework, then asked to take the same end-of-unit test.

Conditions to use the CLT for paired mean difference:

    Independence:

        Checked by:

    Normality:

        Checked by:

Formulas

$SE(\overline{x_d}) =$

$T =$

Confidence interval for a paired mean difference:

Notation

$\overline{x_d} =$

$s_d =$

$\mu_d =$

$\sigma_d =$

Example: Tires

1. What are the observational units?

2. Why should we treat these data as paired rather than two independent samples?

3. What are the sample statistics presented in this example? What notation would be used to represent each value?

4. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

5. Write the null and alternative hypotheses in appropriate notation.

6. How could we use cards to simulate **one** bootstrap resample *which assumes the null hypothesis is true*? How many cards? What is written on the cards? What would we do with the cards? What would you record once you have a simulated sample?

7. After 1000 resamples are generated, where is the resulting null distribution centered? Why does that make sense?

8. What was the p-value of the test? Interpret this p-value in the context of the problem.

9. Write a conclusion in the context of the problem.

Example: College textbook prices

1. What is the research question?

2. What are the observational units?

3. Why should we treat these data as paired rather than two independent samples?

4. What are the sample statistics presented in this example? What notation would be used to represent each value?

5. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

6. How could we use cards to simulate **one** bootstrap resample *which does not assume the null hypothesis is true*? How many cards? What is written on the cards? What would we do with the cards? What would you record once you have a simulated sample?

7. After 1000 resamples are generated, where is the resulting bootstrap distribution centered? Why does that make sense?

8. Give the 95% confidence interval for μ_d.

9. Interpret your 95% confidence interval in the context of the problem.

10. Are the independence and normality conditions satisfied?

11. Write the null and the alternative hypotheses in words.

12. Calculate the standard error of the sample mean difference.

13. Calculate the T-score (the standardized statistic for the sample mean difference).

14. What distribution should the T-score be compared to in order to calculate a p-value?

15. What was the p-value of the test?

16. At the 5% significance level, what decision would you make? What type of error might that be?

17. What conclusion should the researcher make?

18. Are the results in this example statistically significant? Justify your answer.

19. Using $t^\star = 2.00$, calculate a 95% confidence interval for μ_d.

20. Interpret the interval calculated in the context of the problem.

10.2 Activity: COVID-19 and Air Pollution

10.2.1 Learning outcomes

- Given a research question involving paired differences, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Describe and perform a simulation-based hypothesis test for a paired mean difference.

- Interpret and evaluate a p-value for a simulation-based hypothesis test for a paired mean difference.

- Use bootstrapping to find a confidence interval for a paired mean difference.

- Interpret a confidence interval for a paired mean difference.

- Use a confidence interval to determine the conclusion of a hypothesis test.

10.2.2 Terminology review

In this week's activity, we will analyze paired quantitative data using simulation-based methods. Some terms covered in this activity are:

- Mean difference

- Paired data

- Independent groups

- Shifted bootstrap (null) distribution

To review these concepts, see Section 6.2 in the textbook.

10.2.3 COVID-19 and air pollution

In June 2020, the social distancing efforts and stay-at-home directives to help combat the spread of COVID-19 appeared to help 'flatten the curve' across the United States, albeit at a high cost to many individuals and businesses. The impact of these measures, though, goes far beyond the infection and death rates from the disease. You may have seen images comparing air quality in large international cities like Rome, Milan, Wuhan, and New Delhi such as the one pictured in Figure 10.1, which seem to indicate, perhaps unsurprisingly, that fewer people driving and factories being shut down have reduced air pollutants.

Have high population-density US cities seen the same improved air quality conditions? To study this question, data were gathered from the US Environmental Protection Agency (EPA) AirData website which records the ozone (O3) and fine particulate matter (PM2.5) values for cities across the US. These measures are used to calculate an air quality index (AQI) score for each city each day of the year. Thirty-three of the most densely populated US cities were selected and the AQI score recorded for April 20, 2020 as well as the five-year median AQI score for April 20th (2015–2019). Note that higher AQI scores indicate worse air quality. A box plot of the differences in AQI scores for the 33 cities and a table of summary statistics are shown on the next page.

Figure 10.1: The India Gate in New Delhi, India.

Boxplot of the Differences in AQI Scores

Difference in air quality score
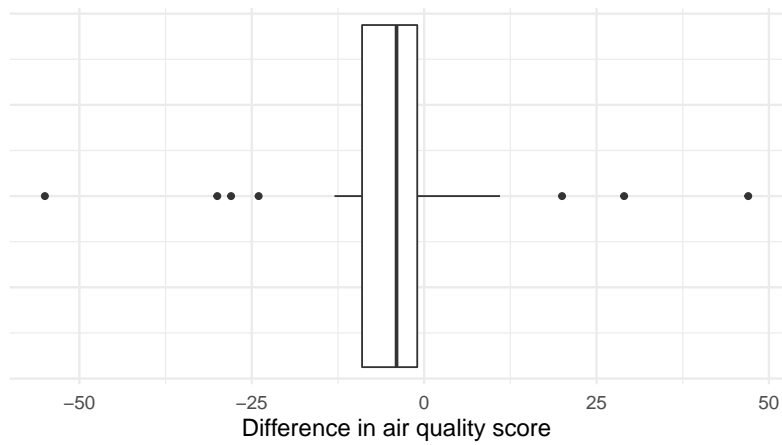
Table 10.1: Summary statistics for current AQI scores, median AQI scores from 2015–2019, and the differences in AQI scores.

	Mean	Standard deviation	Sample size
Current	$\bar{x}_1 = 47.394$	$s_1 = 14.107$	$n_1 = 33$
5 Year Median	$\bar{x}_2 = 51.545$	$s_2 = 17.447$	$n_2 = 33$
Differences	$\bar{x}_d = -4.152$	$s_d = 17.096$	$n_d = 33$

Vocabulary review. Complete Q1–Q5 before class.

1. What is the sample size?

2. Identify the variables in this study. What role (explanatory or response) do each have?

3. Are the differences in AQI scores independent for each case (US city)? Explain.

4. Why is this treated as a paired study design and not two independent samples?

5. Is this an experiment or observational study? Justify your answer.

Ask a research question

6. What are the two competing possibilities to run a hypothesis test for this study?

7. Write the null hypothesis in words.

8. What is the research question?

9. Write the alternative hypothesis in notation.

Summarize and visualize the data

10. Report the summary statistic of interest for the data.

11. What notation is used for the value in question 10?

Use statistical inferential methods to draw inferences from the data

Hypothesis test To simulate the null distribution of paired sample mean differences we will use a bootstrapping method. Recall that the null distribution must be created under the assumption that the null hypothesis is true. Therefore, before bootstrapping, we will need to *shift* each data point by the difference $\mu_0 - \bar{x}_d$. This will ensure that the mean of the shifted data is μ_0 (rather than the mean of the original data, $\bar{x}_d$), and that the simulated null distribution will be centered at the null value.

12. Calculate the difference $\mu_0 - \bar{x}_d$. Will we need to shift the data up or down?

13. We will use the `paired_test()` function in R (in the `catstats` package) to simulate the shifted bootstrap (null) distribution of sample mean differences and compute a p-value. Use the provided R script file and enter the calculated value from question 12 for `xx` to simulate the null distribution and enter the summary statistic from question 10 for `yy` to find the p-value. Highlight and run lines 1–21.

```
paired_test(data = Air$Difference,   # Vector of differences
                                      # or data set with column for each group
            shift = xx,    # Shift needed for bootstrap hypothesis test
            as_extreme_as = yy,   # Observed statistic
            direction = "less",   # Direction of alternative
            number_repetitions = 1000,   # Number of simulated samples for null distribution
            which_first = 1)   # Not needed when using calculated differences
```

14. Sketch the null distribution created in question 13 here.

15. Explain why the null distribution is centered at zero.

16. What proportion of samples are at or less than the observed sample mean difference in AQI scores for current scores minus 5 year median scores? What is the statistical term for this proportion?

17. Interpret the p-value in the context of the problem.

18. How much evidence does this provide for improved air quality in US cities?

19. If evidence was found for improved air quality in US cities, could we conclude that the stay-at-home directives *caused* the improvement in air quality? Explain.

Confidence interval We will use the `paired_bootstrap_CI()` function in `R` (in the `catstats` package) to simulate the bootstrap distribution of sample mean differences and calculate a confidence interval.

20. Write out the parameter of interest in context of the study.

21. Using the provided `R` script file, fill in the missing value at `xx` to find a 99% bootstrap confidence interval; highlight and run lines 24–27. Report the confidence interval in interval notation.

```
paired_bootstrap_CI(data = Air$Difference, # Enter vector of differences
                    number_repetitions = 1000, # Number of bootstrap samples for CI
                    confidence_level = xx,  # Confidence level in decimal form
                    which_first = 1)  # Not needed when entering vector of differences
```

Communicate the results and answer the research question

22. Interpret the 99% confidence interval in the context of the problem.

23. Do the results of your confidence interval and hypothesis test agree? What does each tell you about the null hypothesis?

24. Write a paragraph summarizes the results of this study as if you were describing the results to your roommate. Be sure to describe:

- Summary statistic

- P-value and interpretation

- Conclusion (written to answer the research question)

- Confidence interval and interpretation

- Scope of inference

Revisit and look forward

25. Would it be possible to design an experiment to determine if the changed human behavior due to the COVID-19 pandemic causes a decrease in air pollution? Explain.

10.2.4   Out-of-class activity

The remaining questions cover theory-based methods for testing and estimating a paired mean difference (or single mean). Use Section 6.2.3 in the textbook and the OneMeanTheory video to complete the following questions.

The sampling distribution for $\bar{x}$ based on a sample of size n from a population with a true mean μ and true standard deviation σ can be modeled using a normal distribution when certain conditions are met.

Conditions for the sampling distribution of $\bar{x}$ to follow an approximate normal distribution:

- **Independence**: The sample's observations are independent

- **Normality**: The data should be approximately normal or the sample size should be large.

  - $n < 30$: If the sample size n is less than 30 and there are no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.

  - $n \geq 30$: If the sample size n is at least 30 and there are no particularly extreme outliers, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

1. In the in-class activity, we verified that the independence condition was satisfied. Is the normality condition met to use the theory-based methods for analysis? Explain your answer.

To find the standardized statistic for the paired differences we will use the following formula:

$$T = \frac{\bar{x}_d}{SE(\bar{x}_d)},$$

where the standard error of the sample mean difference is:

$$SE(\bar{x}_d) = \frac{s_d}{\sqrt{n}}.$$

2. Calculate the standard error of the sample mean difference.

3. Calculate the standardized statistic.

Using the provided R script file, enter the T-score (for xx) into the pt() function using a df $= n_d-1 = 33-1 = 32$, and lower.tail = TRUE to find the p-value. Highlight and run line 31.

```
pt(xx, df=32, lower.tail=TRUE)
```

4. Is the p-value found using theory-based methods similar to the simulation p-value found in the in-class activity?

To calculate the 99% theory-based confidence interval for the paired mean difference, use the following formula:

$$\bar{x}_d \pm t^* SE(\bar{x}_d).$$

We will need to find the t^* multiplier using the function qt(). For a 99% confidence level, we are finding the t^* value at the 99.5th percentile with df $= n_d - 1 = 33 - 1 = 32$.

```
qt(0.995, df = 32, lower.tail=TRUE)
#> [1] 2.738481
```

5. Calculate the 99% confidence interval for the paired mean difference using theory-based methods.

6. Explain why the theory-based and simulation confidence intervals are not quite the same.

10.2.5 Take-home messages

1. The differences in a paired data set are treated like a single quantitative variable when performing a statistical analysis. Paired data (or paired samples) occur when pairs of measurements are collected. We are only interested in the population (and sample) of differences, and not in the original data.

2. When using bootstrapping to create a null distribution centered at the null value for both paired data and a single quantitative variable, we first need to shift the data by the difference $\mu_0 - \bar{x}_d$, and then sample with replacement from the shifted data.

3. When analyzing paired data, the summary statistic is the 'mean difference' NOT the 'difference in means'[1]. This terminology will be *very* important in interpretations.

4. To create one simulated sample on the null distribution for a sample mean or mean difference, shift the original data by adding $(\mu_0 - \bar{x})$ or $(0 - \bar{x}_d)$. Sample with replacement from the shifted data n times. Calculate and plot the sample mean or the sample mean difference.

5. To create one simulated sample on the bootstrap distribution for a sample mean or mean difference, label n cards with the original response values. Randomly draw with replacement n times. Calculate and plot the resampled mean or the resampled mean difference.

10.2.6 Additional notes

Use this space to summarize your thoughts and take additional notes on this week's activity and material covered.

---

[1]Technically, if we calculate the differences and then take the mean (mean difference), and we calculate the two means and then take the difference (difference in means), the value will be the same. However, the *sampling variability* of the two statistics will differ, as we will see in Activity 11.

ACTIVITY 11

Inference for a Quantitative Response with Independent Samples

11.1 Reading Guide: Inference for a Difference in Two Means

Section 6.3 (Inference for a difference in two means)

Videos

- 6.3
- TwoMeanRand

Reminders from previous sections

n_1 = sample size of group 1

n_2 = sample size of group 2

$\overline{x}$ = sample mean

s = sample standard deviation

μ = population mean

σ = population standard deviation

General steps of a hypothesis test:

1. Frame the research question in terms of hypotheses.

2. Collect and summarize data using a test statistic.

3. Assume the null hypothesis is true, and simulate or mathematically model a null distribution for the test statistic.

4. Compare the observed test statistic to the null distribution to calculate a p-value.

5. Make a conclusion based on the p-value and write the conclusion in context.

Parameter: a value summarizing a variable(s) for a population.

Statistic: a value summarizing a variable(s) for a sample.

Sampling distribution: plot of statistics from 1000s of samples of the same size taken from the same population.

Standard deviation of a statistic: the variability of statistics from 1000s of samples; how far, on average, each statistic is from the true value of the parameter.

Standard error of a statistic: estimated standard deviation of a statistic.

Hypothesis test: a process to determine how strong the evidence of an effect is. Also called a 'significance test'.

Simulation-based method: Simulate lots of samples of size n under assumption of the null hypothesis, then find the proportion of the simulations that are at least as extreme as the observed sample statistic.

Theory-based method: Develop a mathematical model for the sampling distribution of the statistic under the null hypothesis and use the model to calculate the probability of the observed sample statistic (or one more extreme) occurring.

Null hypothesis (H_0): the skeptical perspective; no difference; no change; no effect; random chance; what the researcher hopes to prove is **wrong**.

Alternative hypothesis (H_A): the new perspective; a difference/increase/decrease; an effect; not random chance; what the researcher hopes to prove is **correct**.

Null value: the value of the parameter when we assume the null hypothesis is true (labeled as $parameter_0$).

Null distribution: the simulated or modeled distribution of statistics (sampling distribution) we would expect to occur if the null hypothesis is true.

P-value: probability of seeing the observed sample data, or something more extreme, assuming the null hypothesis is true.

$\implies$ Lower the p-value the stronger the evidence AGAINST the null hypothesis and FOR the alternative hypothesis.

Decision: a determination of whether to reject or fail to reject a null hypothesis based on a p-value and a pre-set level of significance.

- If p-value $\leq \alpha$, then reject H_0.

- If p-value $> \alpha$, then fail to reject H_0.

Significance level (α): a threshold used to determine if a p-value provides enough evidence to reject the null hypothesis or not.

Common levels of α include 0.01, 0.05, and 0.10.

Statistically significant: results are considered statistically significant if the p-value is below the significance level.

Confidence interval: a process to determine how large an effect is; a range of plausible values for the parameter. Also called 'estimation'.

Margin of error: the value that is added to and subtracted from the sample statistic to create a confidence interval; half the width of a confidence interval.

Bootstrapping: the process of drawing with replacement n times from the original sample.

Bootstrapped resample: a random sample of size n from the original sample, selected with replacement.

Bootstrapped statistic: the statistic recorded from the bootstrapped resample.

Confidence level: how confident we are that the confidence interval will capture the parameter.

Bootstrap $X\%$ confidence interval: $((\frac{(1-X)}{2})^{th}$ percentile, $(X + (\frac{(1-X)}{2})^{th}$ percentile) of a bootstrap distribution.

Central Limit Theorem: For large sample sizes, the sampling distribution of a sample mean (or proportion) will be approximately normal (bell-shaped and symmetric).

t-distribution: A bell-shaped symmetric distribution, centered at 0, wider than the standard normal distribution.

- The variability in a t-distribution depends on the sample size (used to calculate degrees of freedom — df for short).
- The t-distribution gets closer to the standard normal distribution as df increases.

Degrees of freedom (df): describes the variability of the t-distribution.

T-score: the name for a standardized statistic which is compared to a t-distribution.

Notes

To create a **simulated null distribution** of differences in sample means,

1. How many cards will you need and how will the cards be labeled?

2. What do you do with the cards after labeling them?

3. After shuffling, what value will be plotted on the simulated null distribution?

To create a **bootstrap distribution** of differences in sample means,

1. How many cards will you need and how will the cards be labeled?

2. What do you do with the cards after labeling them?

3. After shuffling, what value will be plotted on the bootstrap distribution?

Conditions to use the CLT for a difference in two means:

Independence:

Checked by:

Normality:

Checked by:

In a two-sample t-test, how are the degrees of freedom determined?

True or false: A large p-value indicates that the null hypothesis is true.

Formulas

$SE(\overline{x_1} - \overline{x_2}) =$

$T =$

Confidence interval for a difference in means:

Notation

μ_1 represents

μ_2 represents

σ_1 represents

σ_2 represents

$\overline{x_1}$ represents

$\overline{x_2}$ represents

s_1 represents

s_2 represents

Example: Test scores

1. What are the observational units?

2. What are the sample statistics presented in this example? What notation would be used to represent each value?

3. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

4. What is the research question?

5. Write the null and alternative hypothesis in appropriate notation.

6. How could we use cards to simulate **one** sample *which assumes the null hypothesis is true*? How many cards? What is written on the cards? What would we do with the cards? What would you record once you have a simulated sample?

7. After 1000 shuffles are generated, where is the resulting simulated distribution centered? Why does that make sense?

8. How was the p-value for this test found? The proportion of simulated null samples at _____ or _____.

9. Interpret the p-value in the context of the problem.

10. From these data, can we conclude the exams are equally difficult?

11. What type of error may have occurred at the 5% significance level? Interpret that error in context.

Example: ESC and heart attacks

1. What is the research question?

2. What are the observational units?

3. What variables are recorded? Give the type (categorical or quantitative) and role (explanatory or response) of each.

4. What are the sample statistics presented in this example? What notation would be used to represent each value?

5. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

6. How could we use cards to simulate **one** bootstrap resample *which does not assume the null hypothesis is true*? How many cards? What is written on the cards? What would we do with the cards? What would you record once you have a simulated sample?

7. After 1000 resamples are generated, where is the resulting bootstrap distribution centered? Why does that make sense?

8. Does the 90% confidence interval provide evidence of a difference across the two treatments?

Example: NC births

1. What is the research question?

2. What are the observational units?

3. What variables will be analyzed? Give the type and role of each.

4. Can the results of this study be generalized to a larger population?

5. Are causal conclusions appropriate for these data?

6. Write the null and the alternative hypotheses in words.

7. Write the null and the alternative hypotheses in notation.

8. What are the sample statistics presented in this example? What notation would be used to represent each value?

9. Are the independence and normality conditions satisfied?

10. Calculate the standard error of the difference in sample means.

11. Calculate the T-score (the standardized statistic for the sample mean).

12. What distribution should the T-score be compared to in order to calculate a p-value?

13. What was the p-value of the test?

14. What conclusion should the researcher make?

15. Calculate a 95% confidence interval for the parameter of interest using `qt(0.975, df = 49) = 1.677` as the $t^\star$ value.

16. Interpret your interval in the context of the problem.

11.2 Activity: Weather Patterns and Record Snowfall

11.2.1 Learning objectives

- Given a research question involving one categorical explanatory variable and one quantitative response variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Describe and perform a simulation-based hypothesis test for a difference in means.

- Interpret and evaluate a p-value for a simulation-based hypothesis test for a difference in means.

- Use bootstrapping to find a confidence interval for a difference in means.

- Interpret a confidence interval for a difference in means.

- Use a confidence interval to determine the conclusion of a hypothesis test.

11.2.2 Terminology review

In this week's in-class activity, we will use simulation-based methods to analyze the association between one categorical explanatory variable and one quantitative response variable, where the groups formed by the categorical variable are independent. Some terms covered in this activity are:

- Independent groups

- Difference in means

To review these concepts, see Section 6.3 in the textbook.

11.2.3 Weather patterns and record snowfall

In the winter of 2018–2019, Bozeman had a record snowfall which resulted in the collapse of two flat-roofed buildings on the MSU campus. A writer for the *Washington Post* predicted the heavy snowfall for 2018–2019 due to the El Niño weather pattern that occurred in that season. A meteorologist in Montana wanted to see if the weather pattern really was associated with total snowfall. She obtained historical data from 44 years on the weather pattern (El Niño or La Niña) and snowfall (in inches) at the Billings Weather Station. Side-by-side boxplots and summary statistics for each group are shown on the following page.

Notice from the `R` code that the name of the data set is `Snow`.

```
# Read in data set
Snow <- read.csv("https://math.montana.edu/courses/s216/data/SnowfallByWeatherPattern.csv")

# Code categorical variables as factors
Snow <- # Write over original data with the following
  Snow %>% # Pipe data set into
  mutate(WeatherPattern = factor(WeatherPattern)) # Convert to factor
```

```
# Side-by-side box plots
Snow %>%
ggplot(aes(x = WeatherPattern, y = Snowfall)) +
    geom_boxplot() +
    labs(title = "Snowfall by weather pattern",
         x = "Weather pattern") +
    coord_flip()
```


Snowfall by weather pattern
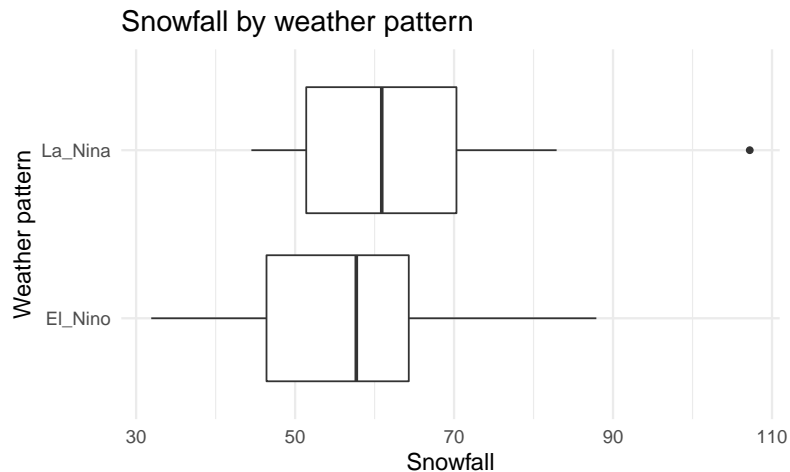
```
# Summary statistics
Snow %>%
    summarize(favstats(Snowfall ~ WeatherPattern))
```

```
#>  WeatherPattern  min   Q1 median   Q3   max     mean       sd  n missing
#> 1        El_Nino 31.9 46.4   57.7 64.3  87.9 56.23043 13.00823 23       0
#> 2        La_Nina 44.5 51.4   60.9 70.3 107.2 63.13333 15.48626 21       0
```

Quantitative variables review. Complete Q1–Q5 before class.

1. The two variables assessed in this study are the type of weather pattern and snowfall. Identify the role for each variable (explanatory or response).

2. Which group (El Niño or La Niña) has the highest center in the distributions of snowfall? Explain which measure of center you are using.

3. Using the side-by-side box plots, which group has the largest spread in snowfall? How did you make that choice?

4. Is this an experiment or an observational study? Justify your answer.

5. Is this a paired data set or two independent groups? Explain your reasoning.

Ask a research question

6. Write out the parameter of interest in context of the study. Use proper notation and be sure to define your subscripts. Use El Niño minus La Niña as the order of subtraction.

7. What are the two competing possibilities we will evaluate in this study?

8. Identify which of your answers in question 7 is the null hypothesis and which is the alternative hypothesis.

Summarize and visualize the data

9. Calculate the summary statistic of interest. Use El Niño minus La Niña as the order of subtraction. What is the appropriate notation for this statistic?

Use statistical inferential methods to draw inferences from the data

Hypothesis test   Remember that the null distribution is created based on the assumption the null hypothesis is true. In this study, the null hypothesis states that there is no association between the two variables. This means that the snowfall values observed in the data set would have been the same regardless of the weather pattern that year.

To demonstrate this simulation, your instructor will use cards to represent the sample.

10. How many cards will we start with?

11. What will we write on each card?

12. Next, we will mix the cards together and shuffle into two piles. How many cards will go into each pile? What should we label the piles?

13. What value is calculated from the cards and plotted on the null distribution? *Hint*: What statistic are we calculating from the data?

14. Once we create a null distribution of 1000 simulations, at what value do you expect the distribution to be centered? Explain your reasoning.

We will use the `two_mean_test()` function in R (in the `catstats` package) to simulate the null distribution of differences in sample means and compute a p-value.

15. When using the `two_mean_test()` function, we need to enter the name of the response variable, `Snowfall`, and the name of the explanatory variable, `WeatherPattern`, for the formula. The name of the data set as shown above is `Snow`. What values should be entered for each of the following to create 1000 simulated samples?

   - First in subtraction (What is the outcome for the explanatory variable that is used as first in the order of subtraction? `"El_Nino"` or `"La_Nina"`):

   - Number of repetitions:

   - As extreme as:

   - Direction (`"greater"`, `"less"`, or `"two-sided"`):

16. Simulate a null distribution and compute the p-value. Using the `R` script file for this activity, enter your answers for question 15 in place of the `xx`'s to produce the null distribution with 1000 simulations. Highlight and run lines 1–29.

```
two_mean_test(Snowfall ~ WeatherPattern, data = Snow,  # Variables and data
          first_in_subtraction = "xx", # First outcome in order of subtraction
          number_repetitions = 1000,  # Number of simulations
          as_extreme_as = xx,  # Observed statistic
          direction = "xx")  # Direction of alternative: "greater", "less", or "two-sided"
```

Sketch the null distribution created using the code above.

17. Report the p-value. Based off of this p-value, write a conclusion to the hypothesis test.

Confidence interval   We will use the `two_mean_bootstrap_CI()` function in `R` (in the `catstats` package) to simulate the bootstrap distribution of differences in sample means and calculate a confidence interval.

18. Using bootstrapping find a 95% confidence interval. Using the provided `R` script file, enter the variable names and data set name as in the `two_mean_test()` function, outcome name for the first in subtraction, number of repetitions, and the confidence level as a decimal. Highlight and run lines 32–35. Report the 95% confidence interval in interval notation.

```
two_mean_bootstrap_CI(RESPONSE ~ EXPLANATORY, data = DATASET,  # Variables and data
                    first_in_subtraction = "xx", # First value in order of subtraction
                    number_repetitions = 1000,  # Number of simulations
                    confidence_level = xx)
```

19. Interpret the interval you calculated in question 18.

Communicate the results and answer the research question

20. Write a paragraph summarizing the results of the study as if you are reporting the results to your supervisor. Be sure to describe:

- Summary statistic

- P-value and interpretation

- Conclusion (written to answer the research question)

- Confidence interval and interpretation

- Scope of inference

Revisit and look forward

21. Would the results from a theory-based test match the results we saw with the simulation? Explain why or why not.

22. If we had data on 45 La Niña years and 47 El Niño years and found a similar-valued summary statistic, what would happen to the p-value? The width of the confidence interval? The power of the test?

11.2.4 Out-of-class activity

The remaining questions cover inference for a difference in means using theory-based methods. Use Section 6.3.3 in the textbook and the TwoMeanTheory video to complete the following questions.

The sampling distribution for $\bar{x}_1 - \bar{x}_2$ can be modeled using a normal distribution when certain conditions are met.

Conditions for the sampling distribution of $\bar{x}_1 - \bar{x}_2$ to follow an approximate normal distribution:

- **Independence**: The sample's observations are independent

- **Normality**: Each sample should be approximately normal or have a large sample size. For *each* sample:

    - $n < 30$: If the sample size n is less than 30 and there are no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.

    - $n \leq 30$: If the sample size n is at least 30 and there are no particularly extreme outliers, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

1. In question 21 of the in-class activity, we noted that there were issues with the normality condition. Explain how that will affect the p-value and confidence interval found with theory-based methods.

To find the standardized statistic for the difference in means we will calculate:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)},$$

where the standard error of the difference in means is calculated using:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

2. Calculate the standard error of the difference in sample means.

3. Calculate the standardized statistic for the difference in sample eans.

Using the provided R script file, enter the T-score (for xx) into the pt() function using a df $= \text{minimum}(n_1 - 1, n_2 - 1) = 21 - 1 = 20$, and lower.tail = TRUE to find the p-value. Highlight and run line 39.

```
2*pt(xx, df=20, lower.tail=TRUE)
```

4. Explain why we multiplied by 2 in the code above.

5. Report the p-value from the R output.

6. Explain why the p-value found using theory-based methods differs from the p-value found using simulation methods in the in-class activity.

To calculate a theory-based 95% confidence interval for a difference in means, use the formula:

$$\bar{x}_1 - \bar{x}_2 \pm t^* SE(\bar{x}_1 - \bar{x}_2).$$

We will need to find the t^* multiplier using the function `qt()`. For a 95% confidence level, we are finding the t^* value at the 97.5th percentile with $\text{df} = \text{minimum}(n_1 - 1, n_2 - 1) = 20$.

```
qt(0.975, df = 20, lower.tail=TRUE)
#> [1] 2.085963
```

7. Calculate the 95% confidence interval using theory-based methods.

Note that since the normality condition was not met, neither the theory-based p-value nor the theory-based confidence interval are valid.

11.2.5 Take-home messages

1. This activity differs from Activity 10 because the responses are independent, not paired. These data are analyzed as a difference in means, not a mean difference.

2. Review the take-home messages in Activity 6 and apply them to this context.

3. To create one simulated sample on the null distribution for a difference in sample means, label cards with the response variable values from the original data. Mix cards together and shuffle into two new groups of sizes n_1 and n_2. Calculate and plot the difference in means.

4. To create one simulated sample on the bootstrap distribution for a difference in sample means, label $n_1 + n_2$ cards with the original response values. Keep groups separate and randomly draw with replacement n_1 times from group 1 and n_2 times from group 2. Calculate and plot the resampled difference in means.

11.2.6 Additional notes

Use this space to summarize your thoughts and take additional notes on this week's activity and material covered

Inference for Two Quantitative Variables

12.1 Reading Guide: Inference for Slope and Correlation

Sections 7.1 and 7.2 (Inference for regression and model conditions)

Videos

- 7.1and7.2
- RegressionSim

Reminders from previous sections

β_0: population y-intercept

β_1: population slope

ρ: population correlation

b_0: sample y-intercept

b_1: sample slope

r: sample correlation

Scatterplot: displays two quantitative variables; one dot = two measurements (x, y) on one observational unit.

Four characteristics of a scatterplot:

- *Form*: pattern of the dots plotted. Is the trend generally linear (you can fit a straight line to the data) or non-linear?

- *Strength*: how closely do the points follow a trend? Very closely (strong)? No pattern (weak)?

- *Direction*: as the x values increase, do the y-values tend to increase (positive) or decrease (negative)?

- Unusual observations or *outliers*: points that do not fit the overall pattern of the data.

Least squares regression line: $\hat{y} = b_0 + b_1 x$, where b_0 is the sample y-intercept (the estimate for the `(Intercept)` row in the `R` regression output), and b_1 is the sample slope (the estimate for the `x-variable_name` row in the `R`).

Sample slope interpretation: a 1 unit increase in the x variable is associated with a $|b_1|$ unit *predicted* increase/decrease in the y-variable.

General steps of a hypothesis test:

1. Frame the research question in terms of hypotheses.

2. Collect and summarize data using a test statistic.

3. Assume the null hypothesis is true, and simulate or mathematically model a null distribution for the test statistic.

4. Compare the observed test statistic to the null distribution to calculate a p-value.

5. Make a conclusion based on the p-value and write the conclusion in context.

Parameter: a value summarizing a variable(s) for a population.

Statistic: a value summarizing a variable(s) for a sample.

Sampling distribution: plot of statistics from 1000s of samples of the same size taken from the same population.

Standard deviation of a statistic: the variability of statistics from 1000s of samples; how far, on average, each statistic is from the true value of the parameter.

Standard error of a statistic: estimated standard deviation of a statistic.

Hypothesis test: a process to determine how strong the evidence of an effect is. Also called a 'significance test'.

Simulation-based method: Simulate lots of samples of size n under assumption of the null hypothesis, then find the proportion of the simulations that are at least as extreme as the observed sample statistic.

Theory-based method: Develop a mathematical model for the sampling distribution of the statistic under the null hypothesis and use the model to calculate the probability of the observed sample statistic (or one more extreme) occurring.

Null hypothesis (H_0): the skeptical perspective; no difference; no change; no effect; random chance; what the researcher hopes to prove is **wrong**.

Alternative hypothesis (H_A): the new perspective; a difference/increase/decrease; an effect; not random chance; what the researcher hopes to prove is **correct**.

Null value: the value of the parameter when we assume the null hypothesis is true (labeled as $parameter_0$).

Null distribution: the simulated or modeled distribution of statistics (sampling distribution) we would expect to occur if the null hypothesis is true.

P-value: probability of seeing the observed sample data, or something more extreme, assuming the null hypothesis is true.

$\implies$ Lower the p-value the stronger the evidence AGAINST the null hypothesis and FOR the alternative hypothesis.

Decision: a determination of whether to reject or fail to reject a null hypothesis based on a p-value and a pre-set level of significance.

- If p-value $\leq \alpha$, then reject H_0.

- If p-value $> \alpha$, then fail to reject H_0.

Significance level (α): a threshold used to determine if a p-value provides enough evidence to reject the null hypothesis or not.

Common levels of α include 0.01, 0.05, and 0.10.

Statistically significant: results are considered statistically significant if the p-value is below the significance level.

Confidence interval: a process to determine how large an effect is; a range of plausible values for the parameter. Also called 'estimation'.

Margin of error: the value that is added to and subtracted from the sample statistic to create a confidence interval; half the width of a confidence interval.

Bootstrapping: the process of drawing with replacement n times from the original sample.

Bootstrapped resample: a random sample of size n from the original sample, selected with replacement.

Bootstrapped statistic: the statistic recorded from the bootstrapped resample.

Confidence level: how confident we are that the confidence interval will capture the parameter.

Bootstrap $X\%$ confidence interval: $((\frac{(1-X)}{2})^{th}$ percentile, $(X + (\frac{(1-X)}{2})^{th}$ percentile) of a bootstrap distribution

t-distribution: A bell-shaped symmetric distribution, centered at 0, wider than the standard normal distribution.

- The variability in a t-distribution depends on the sample size (used to calculate degrees of freedom — df for short).
- The t-distribution gets closer to the standard normal distribution as df increases.

Degrees of freedom (df): describes the variability of the t-distribution.

T-score: the name for a standardized statistic which is compared to a t-distribution.

Notes

To create a **simulated null distribution** of sample slopes or sample correlations,

1. How many cards will you need and how will the cards be labeled?

2. What do you do with the cards after labeling them?

3. After shuffling, what value will be plotted on the simulated null distribution?

To create a **bootstrap distribution** of sample slopes or sample correlations,

1. How many cards will you need and how will the cards be labeled?

2. What do you do with the cards after labeling them?

3. After shuffling, what value will be plotted on the bootstrap distribution?

Conditions to use the CLT for testing slope (or correlation):

Linearity:

Checked by:

Independent observations:

Checked by:

Nearly normal residuals:

Checked by:

Constant or equal variance:

Checked by:

In a theory-based test of slope or correlation, how are the degrees of freedom determined?

Explain why testing for slope is equivalent to testing for correlation.

Where in the `R` output can $SE(b_1)$ be found?

Formulas

$T =$

Confidence interval:

Example: Crop yields

1. What are the observational units?

2. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

3. What is the research question?

4. Write the null and alternative hypotheses in appropriate notation.

5. How could we use cards to simulate **one** sample which assumes *the null hypothesis is true*? How many cards? What is written on the cards? What would we do with the cards? What would you record once you have a simulated sample?

6. After 1000 shuffles are generated, where is the resulting simulated distribution centered? Why does that make sense?

7. What are the sample statistics presented in this example? What notation would be used to represent each value?

8. Write the least squares regression line for these data in appropriate notation.

9. How was the p-value for this test found? The proportion of simulated null samples at _____ or _____.

10. Interpret the p-value in the context of the problem.

11. What conclusion can be drawn from these data?

12. How could we use cards to simulate **one** bootstrap resample *which does not assume the null hypothesis is true*? How many cards? What is written on the cards? What would we do with the cards? What would you record once you have a simulated sample?

13. Interpret the 95% confidence interval provided.

Example: Midterm elections and unemployment

1. What is the research question?

2. What are the observational units?

3. What variables will be analyzed? Give the type and role of each.

161

4. Can the results of this study be generalized to a larger population?

5. Are causal conclusions appropriate for these data?

6. Write the null and the alternative hypotheses in words.

7. Write the null and the alternative hypotheses in notation.

8. What are the sample statistics presented in this example? What notation would be used to represent each value?

9. Write the least squares regression line for these data in appropriate notation.

10. From the R output, what is the standard error of the slope estimate?

11. Calculate the T-score (the standardized statistic for the slope).

12. What distribution should the T-score be compared to in order to calculate a p-value?

13. What was the p-value of the test?

14. What conclusion should the researcher make?

15. Calculate a 95% confidence interval for the parameter of interest using `qt(0.975, df = 27) = 2.052` as the $t^\star$ value.

16. Interpret your interval in the context of the problem.

12.2 Activity: Moneyball

12.2.1 Learning outcomes

- Given a research question involving two quantitative variables, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Assess the conditions to use the normal distribution model for a slope or correlation.

- Find the T test statistic (T-score) for a slope or correlation based off of `lm()` output in `R`.

- Find, interpret, and evaluate the p-value for a theory-based hypothesis test for a slope or correlation.

- Create and interpret a theory-based confidence interval for a slope or correlation.

- Use a confidence interval to determine the conclusion of a hypothesis test.

12.2.2 Terminology review

In this week's in-class activity, we will use theory-based methods for hypothesis tests and confidence intervals for a linear regression slope or correlation. Some terms covered in this activity are:

- Correlation
- Slope
- Regression line

To review these concepts, see Chapters 3 and 7 in the textbook.

12.2.3 Moneyball

The goal of a Major League baseball team is to make the playoffs. In 2002, the manager of the Oakland A's, Billy Bean, with the help of Paul DePodesta began to use statistics to determine which players to choose for their season. Based on past data, DePodesta determined that to make it to the playoffs, the A's would need to win at least 95 games in the regular season. In order to win more games, they would need to score more runs than they allowed. The Oakland A's won 20 consecutive games and a total of 103 games for the season. The success of this use of sports analytics was portrayed by the 2011 movie, *Moneyball*.

In this study, we will see if there is evidence of a positive linear relationship between the difference in the number of runs scored minus the number of runs allowed and the number of wins for Major League baseball teams in the years before 2002.

Some of the variables collected in the data set `baseball` consist of the following:

Variable	Description
RA	Runs allowed
RS	Runs scored
OBP	On-base percentage
SLG	Slugging percentage
BA	Batting average
OOBP	Opponent's on-base percentage
OSLG	Opponent's slugging percentage
W	Number of wins in the season
RD	Difference of runs scored minus runs allowed

```
# Read in data set
baseball <- read.csv("https://math.montana.edu/courses/s216/data/baseball.csv")

baseball$RD <- baseball$RS - baseball$RA # Create variable run difference

baseball <- # Write over original data with the following
  baseball %>% # Pipe data set into
  subset(Year < 2002) # Select only years before 2002
```

Vocabulary review. Complete Q1–Q4 before class.

1. Explain why regression methods are appropriate to use to address the researchers' question. Make sure you clearly define the variables of interest in your explanation and their roles.

2. Use the provided R script file to create a scatterplot to examine the relationship between the difference in number of runs scored minus number of runs allowed and the number of wins by filling in the variable names (RD and W) for xx and yy in line 17. Highlight and run lines 1–22.

```
baseball %>% # Pipe data set into...
ggplot(aes(x = xx, y = yy))+  # Specify variables
  geom_point() +  # Add scatterplot of points
  labs(x = "Difference in number of runs",  # Label x-axis
       y = "Number of wins",  # Label y-axis
       title = "Scatterplot of Run Difference vs. Number of Wins") +
                # Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```

3. Sketch the plot created below. Based on your plot, does it appear that there is a relationship between run difference and wins? Note: RD should be on the x-axis.

4. Describe the features of the plot above, addressing all four characteristics of a scatterplot.

   If you indicated there are potential outliers, which points are they?

Conditions for the least squares line

When performing inference on a least squares line, the follow conditions are generally required:

- *Independent observations* (for both simulation-based and theory-based methods): individual data points must be independent.
    - Check this assumption by investigating the sampling method and determining if the observational units are related in any way.
- *Linearity* (for both simulation-based and theory-based methods): the data should follow a linear trend.
    - Check this assumption by examining the scatterplot of the two variables, and a scatterplot of the residuals (on the y-axis) versus the fitted values (on the x-axis). The pattern in the residual plot should display a horizontal line.
- *Constant variability* (for theory-based methods only): the variability of points around the least squares line remains roughly constant
    - Check this assumption by examining a scatterplot of the residuals (on the y-axis) versus the fitted values (on the x-axis). The variability in the residuals around zero should be approximately the same for all fitted values.
- *Nearly normal residuals* (for theory-based methods only: residuals must be nearly normal.
    - Check this assumption by examining a histogram of the residuals, which should appear approximately normal[1].

The scatterplot generated in question 2 and the residual plots shown below will be used to assess these conditions for approximating the data with the t-distribution.
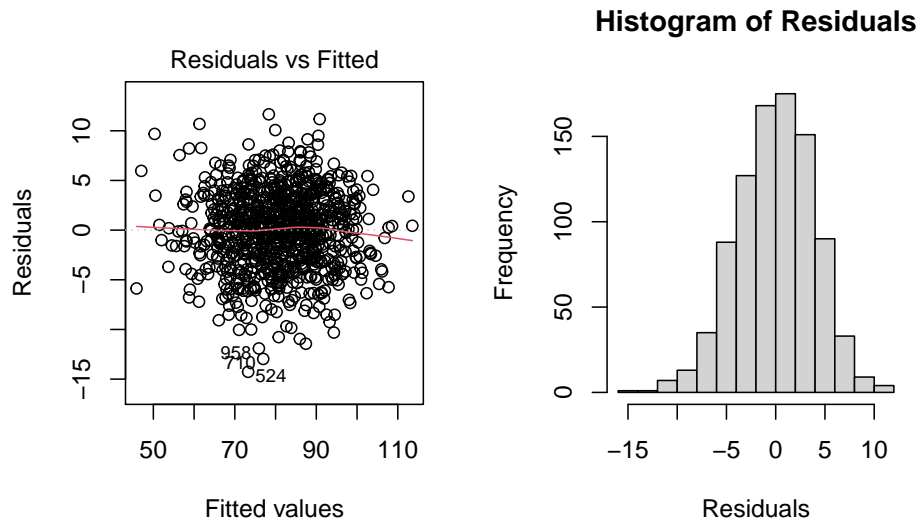
5. Are the conditions met to use the t-distribution to approximate the sampling distribution of the standardized statistic? Justify your answer.

---

[1]A better plot for checking the normality assumption is called a *normal quantile-quantile plot* (or QQ-plot). However, this type of plot will be covered in a future course

Ask a research question

6. Write out the null hypothesis in words.

7. Using the research question, write the alternative hypothesis in notation to test the slope.

Summarize and visualize the data

Using the provided `R` script file, enter the response variable name, `W`, into the `lm()` (linear model) function for `yy` and the explanatory variable name, `RD`, for `xx` in line 32 to get the linear model output. Highlight and run lines 32–33.

```
lm.baseball <- lm(yy~xx, data=baseball) # lm(response~explanatory)
round(summary(lm.baseball)$coefficients, 5)
```

8. Using the output from the evaluated `R` code above, write the equation of the regression line using appropriate statistical notation.

9. Interpret the estimated slope in context of the problem.

10. Using your estimated line of best fit, predict the number of wins for a run difference of 147. Show all work.

11. In 2002, the Oakland A's had a run difference of 147 and 103 wins. Calculate the residual associated with the observation (147, 103) using your estimated regression line from question 8.

Use statistical inferential methods to draw inferences from the data

Hypothesis test   To find the value of the standardized statistic to test the slope we will use,

$$T = \frac{\text{slope estimate}}{SE} = \frac{b_1}{SE(b_1)}.$$

We will use the linear model `R` output above to get the estimate for slope and the standard error of the slope.

12. Calculate the standardized statistic for slope. Identify where this calculated value is in the linear model `R` output.

13. Interpret the standardized statistic in context of the problem.

14. Using the linear model `R` output, report the p-value for the test of significance for slope.

15. Based on the p-value, how much evidence is there against the null hypothesis?

Confidence interval   Recall that a confidence interval is calculated by adding and subtracting the margin of error to the point estimate.

$$\text{point estimate} \pm t^* SE(\text{estimate}).$$

When the point estimate is a regression slope, this formula becomes

$$b_1 \pm t^* SE(b_1).$$

The t^* multiplier comes from a t-distribution with $n-2$ degrees of freedom. Recall for a 95% confidence interval, we use the 97.5% percentile (95% of the distribution is in the middle, leaving 2.5% in each tail). The sample size for this study is 902 so we will use the degrees of freedom 900 $(n-2)$.

```
qt(0.95+0.025, 900) # 95% t* multiplier
```

```
#> [1] 1.962603
```

16. Calculate the 95% confidence interval for the true slope.

17. Interpret the 95% confidence interval in context of the problem.

Communicate the results and answer the research question

18. Based on the p-value, write a conclusion in context of the problem.

19. Does the p-value agree with the 95% confidence interval? What does each tell you about the null hypothesis?

20. Write a paragraph summarizing the results of the study as if you are reporting these results in your local newspaper. Be sure to describe:

- Summary statistic
- Test statistic and interpretation
- P-value and interpretation
- Confidence interval and interpretation
- Conclusion (written to answer the research question)
- Scope of inference

Revisit and look forward

In order to see what team attributes would have the most impact on the number of runs scored, Beane and DePodesta created several prediction models. Using 2001 statistics, they looked at a prediction model using both the team's on-base percentage and slugging percentage to predict the number of runs scored. The following R code gives the estimates for the regression model with both of these variables included.

```
lm.score <- lm(RS ~ OBP + SLG, data=baseball)
round(summary(lm.score)$coefficients, 5)
#>             Estimate Std. Error   t value Pr(>|t|)
#> (Intercept) -804.6271   18.92079 -42.52608        0
#> OBP         2737.7680   90.68455  30.19001        0
#> SLG         1584.9086   42.15559  37.59665        0
```

21. Use the provided R output to write the linear regression model including all variables. *Hint*: The estimated line of regression is of the form:

$$\widehat{\text{runs scored}} = b_0 + b_1 \times OBP + b_2 \times SLG.$$

22. Using the fitted regression model above, predict the number of runs for the Oakland A's in 2002 if the team OBP is 0.339 and the team SLPG is 0.430.

12.2.4 Out-of-class activity

In the out-of-class activity, we will focus on using simulation-based methods for inference in regression. Use Section 7.1 in the textbook and the TwoQuantSim video to complete the following questions.

Simulation-based hypothesis test

First, let's think about how one simulation would be created on the null distribution using cards. First, we would write the values for the response variable, wins, on each card. Next, we would shuffle these y values while keeping the x values (explanatory variable) in the same order. Then, find the line of regression for the shuffled cards and calculate either the sample slope or sample correlation.

1. Once we have one simulated sample, what would we calculate and plot on the null distribution? *Hint*: What statistic are we calculating from the data?

We will use the `regression_test()` function in R (in the `catstats` package) to simulate the null distribution of sample slopes (or sample correlations) and compute a p-value. We will need to enter the response variable name and the explanatory variable name for the formula, the data set name (identified above as `baseball`), the statistic for the test (either slope or correlation), number of repetitions, the sample statistic (value of slope or correlation), and the direction of the alternative hypothesis.

The response variable name is `W` (wins) and the explanatory variable name is `RD` (run difference).

2. What inputs should be entered for each of the following to create the simulation to test regression slope?

- Direction (`"greater"`, `"less"`, or `"two-sided"`):

- Statistic (choose `"slope"` or `"correlation"`):

- As extreme as (enter the value for the sample slope):

- Number of repetitions:

Using the R script file for this activity, enter your answers for question 2 in place of the `xx`'s to produce the null distribution with 1000 simulations. Highlight and run lines 1–13 and then lines 44–49.

```
regression_test(W ~ RD, # response ~ explanatory
                data = baseball, # Name of data set
                direction = "xx", # Sign in alternative ("greater", "less", "two-sided")
                statistic = "xx", # "slope" or "correlation"
                as_extreme_as = x, # Observed slope or correlation
                number_repetitions = 1000) # Number of simulated samples for null distribution
```

3. Report the p-value from the R output. Is this value similar to what was seen with the theory-based methods?

Simulation-based confidence interval

We will use the `regression_bootstrap_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of sample slopes (or sample correlations) and calculate a confidence interval. Fill in the `xx`'s in the the provided R script file to find a 95% confidence interval. Highlight and run lines 52–56.

```
regression_bootstrap_CI(W ~ RD, # response ~ explanatory
  data = baseball, # Name of data set
  confidence_level = xx, # Confidence level as decimal
  statistic = "xx", # Slope or correlation
  number_repetitions = 1000) # Number of simulated samples for bootstrap distribution
```

4. Report the bootstrap 95% confidence interval in interval notation.

5. Is the bootstrap 95% confidence interval similar to what was found using theory-based methods? Why did you expect this to be true?

12.2.5 Take-home messages

1. The p-value for a test for correlation should be approximately the same as the p-value for the test of slope. In the simulation test, we just change the statistic type from slope to correlation and use the appropriate sample statistic value.

2. To check the validity conditions for using theory-based methods we must use the residual diagnostic plots to check for normality of residuals and constant variability, and the scatterplot to check for linearity.

3. To interpret a confidence interval for the slope, think about how to interpret the sample slope and use that information in the confidence interval for slope.

4. To create one simulated sample on the null distribution for a sample slope or sample correlation, hold the x values constant and shuffle the y values to new x values. Find the regression line for the shuffled data and plot the slope or the correlation for the shuffled data.

5. To create one simulated sample on the bootstrap distribution for a sample slope or sample correlation, label n cards with the original (response, explanatory) values. Randomly draw with replacement n times. Find the regression line for the resampled data and plot the resampled slope or correlation.

12.2.6 Additional notes

Use this space to summarize your thoughts and take additional notes on this week's activity and material covered.

References

Banerjee, S. 2018. "Linear Regression: Moneyball – Part 1." *Medium.* https://towardsdatascience.com/linear-regression-moneyball-part-1-b93b3b9f5b53.

"IMDb Movies Extensive Dataset." n.d. https://kaggle.com/stefanoleone992/imdb-extensive-dataset.

Johnson, C. M.., and J. E. Memmott. 2006. "Examination of Relationships Between Participation in School Music Programs of Differing Quality and Standardized Test Results." *Journal of Research in Music Education* 54 (4): 293–307. https://doi.org/10.1177/002242940605400403.

"Moneyball Dataset." n.d. https://kaggle.com/wduckett/moneyball-mlb-stats-19622012.

National Weather Service Corporate Image Web Team. n.d. "National Weather Service – NWS Billings." https://w2.weather.gov/climate/xmacis.php?wfo=byz.

O'Hair, A. 2017. "Video 4: Using the Model to Make Predictions." MIT: MIT Open Courseware. https://ocw.mit.edu/courses/sloan-school-of-management/15-071-the-analytics-edge-spring-2017/linear-regression/moneyball-the-power-of-sports-analytics/video-4-using-the-models-to-make-predictions/.

Quinn, G. E., C. H. Shin, M. G. Maguire, and R. A. Stone. 1999. "Myopia and Ambient Lighting at Night." *Nature* 399 (6732): 113–14. https://doi.org/10.1038/20094.

Ramachandran, V. 2007. "3 Clues to Understanding Your Brain." https://www.ted.com/talks/vs_ramachandran_3_clues_to_understanding_your_brain.

Richardson, T., and R. T. Gilman. 2019. "Left-Handedness Is Associated with Greater Fighting Success in Humans." *Scientific Reports* 9 (1): 15402. https://doi.org/10.1038/s41598-019-51975-3.

Stewart, E. H., B. Davis, B. L. Clemans-Taylor, B. Littenberg, C. A. Estrada, and R. M. Centor. 2014. "Rapid Antigen Group a Streptococcus Test to Diagnose Pharyngitis: A Systematic Review and Meta-Analysis" 9 (11). https://doi.org/10.1371/journal.pone.0111727.

Sulheim, S., A. Ekeland, I. Holme, and R. Bahr. 2017. "Helmet Use and Risk of Head Injuries in Alpine Skiers and Snowboarders: Changes After an Interval of One Decade" 51 (1): 44–50. https://doi.org/10.1136/bjsports-2015-095798.

US Environmental Protection Agency. n.d. "Air Data – Daily Air Quality Tracker." https://www.epa.gov/outdoor-air-quality-data/air-data-daily-air-quality-tracker.