STAT 216 Coursepack

Fall 2022
Montana State University

Melinda Yager
Jade Schmidt
Stacey Hancock

This resource was developed by Melinda Yager, Jade Schmidt, and Stacey Hancock in 2021 to accompany the online textbook: Carnegie, N., Hancock, S., Meyer, E., Schmidt, J., and Yager, M. (2021). *Montana State Introductory Statistics with R*. Montana State University. https://mtstateintrostats.github.io/IntroStatTextbook/.

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 class period, 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 and labs in this coursepack will be completed during class time. Parts of each lab will be turned in on Gradescope. To aid in your understanding, read through the introduction for each activity before attending class each day.

STAT 216 is a 3-credit in-person course. In our experience, it takes six to nine hours per week outside of class to achieve a good grade in this class. By "good" we mean at least a C because a grade of D 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 day's in-class activity
  - 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 or weekly lab with your classmates and instructor, taking detailed notes on your answers to each question in the activity.
- *After class meeting*:
  - Complete any parts of the activity you did not complete in class.
  - Review the activity solutions in the Math and Stat Center, and take notes on key points.
  - Finish watching any remaining assigned videos or readings for the week.
  - Complete the week's homework assignment.

Fall 2023 Calendar of In-Class Activities

This calendar only lists the in-class activities, RStudio labs and exams each week. For required readings as well as due dates for assignments, refer to the calendar at:
https://mtstateintrostats.github.io/Syllabus/#Course_calendar

Week	**Day**	**Date**	**Activity**	
1	W	8/23	Intro to Data	
1	F	8/25	Data Lecture and Lab	
2	M	8/28	Study Design Lecture - Complete Out of Class Activity 2	
2	W	8/30	American Indian Address	
2	F	9/1	Week 2 Lab	
3	M	9/4	(*No class*)	
3	W	9/6	Categorical/Quantitative EDA Lecture - Complete Out of Class Activity 3	
3	F	9/8	Week 3 Lab	
4	M	9/11	Regression Lecture - Complete Out of Class Activity 4	
4	W	9/13	Movie Profits	
4	F	9/15	Week 4 Lab	
5	M	9/18	Exam 1 Review	
5	W	9/20	Group Midterm Exam 1	
5	F	9/21	Midterm Exam 1	
6	M	9/26	Hypothesis Testing Lecture	
6	W	9/28	Helper Hinderer — Simulation HT	
6	F	9/30	Week 6 Lab	
7	M	10/3	Theory-based Testing Lecture	
7	W	10/5	Handedness of Male Boxers — Theory	
7	F	10/7	Week 7 Lab	
8	M	10/10	Two Proportion Simulation Lecture	
8	W	10/12	Good Samaritan — Simulation HT	CI
8	F	10/14	Week 8 Lab	
9	M	10/17	Two Proportion Theory Lecture	
9	W	10/19	Helmet Use and Head Injuries — Theory HT	CI
9	F	10/21	Week 9 Lab	
10	M	10/24	Exam 2 Review	
10	W	10/26	Group Midterm Exam 2	
10	F	10/28	Midterm Exam 2	
11	M	10/31	Paired Inference Lecture	
11	W	11/2	Color Interference	
11	F	11/4	Week 11 Lab	
12	M	11/7	Weather Patterns and Snowfall	
12	W	11/9	Week 12 Lab	
12	F	11/11	(*No class*)	
13	M	11/14	Regression Inference Lecture	
13	W	11/16	Golf Driving Distances	
13	F	11/18	Week 13 Lab	
Holiday	M–F	11/21–11/25	**No Class — Fall Break**	
14	M	11/28	Probability	Relative Ris
14	W	11/30	Relative Risk	
14	F	12/2	Week 14 Lab	

Week	Day	Date	Activity	
15	M	12/5	Final Exam Review	
15	W	12/7	Final Group Exam Part 1	
15	F	12/9	Final Group Exam Part 2	
Finals	12/13	6 - 7:50 pm	Common Final Exam See www.montana.edu/registrar/Schedules.html	

MODULE 1

Inference for a Single Categorical Variable: Simulation-based Methods

1.1 Module 6 Reading Guide: Categorical Inference

Section 5.1 (Foundations of inference: Hypothesis tests)

Please note that Theory-based inference 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:

Confidence interval:

Margin of error:

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

$SD(\widehat{p}) =$

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. 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.

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

9. 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)?

Section 5.3 (Inference for one proportion)

You may skip Section 5.3.4, which will be covered next week.

Videos

- 5.3SimInf
- Bootstrapping

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.

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.

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.

Vocabulary

Point estimate:

Test statistic:

Null value:

Null distribution:

One-sided hypothesis test:

Two-sided hypothesis test:

Bootstrapping:

Bootstrapped resample:

Bootstrapped statistic:

Notes

Which hypothesis must we assume is true in order to simulate a null distribution?

Explain the differences between a one-sided and two-sided hypothesis test.

How will the research questions differ?

How will the notation in the alternative hypothesis differ?

How does the p-value calculation differ?

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

Should the default in research be a one-sided or two-sided hypothesis test? Explain why.

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?

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, using the example in 5.3.1.

4. Write the null and alternative hypotheses in notation, using the example in 5.3.1.

5. To simulate the null distribution, we would not be able to use coins. 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 in 5.3.1?

8. What was the p-value of the test from the example in 5.3.1?

9. At the 5% significance level, what decision would you make based on the p-value above?

10. What conclusion should the researcher make?

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

12. How does the alternative hypothesis change, both in words and in notation, when the example changes to a two-sided hypothesis test in 5.3.2?

13. Explain how the p-value calculation changes between the example in 5.3.1 (one-sided hypothesis test) and the example in 5.3.2 (two-sided hypothesis test).

14. Why does doubling the p-value from the one-sided hypothesis test (your answer to question 8) not match the two-sided p-value calculated in Figure 5.12?

15. 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?

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

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

1.2 Lecture Notes Week 6: Inference for one categorical variable using Simulation-based methods

1.2.1 Hypothesis Testing

Purpose of a hypothesis test:

- Use data collected on a ______________________________ to give information about the _______________________
- Determines ____________________________ of ________________________________ of an effect

General steps of a hypothesis test

1. Write a research question and hypotheses.
2. Collect data and calculate a summary statistic.
3. Model a sampling distribution which assumes the null hypothesis is true.
4. Calculate a p-value.
5. Draw conclusions based on a p-value.

Hypotheses

- Two possible outcomes:

- Always written about the ___________________________ (population)

Null hypothesis

- Skeptical perspective, no difference, no effect, random chance
- What the researcher hopes is _______________________.

Notation:

Alternative hypothesis

- New perspective, a chance, a difference, an effect
- What the researcher hopes is ________________________.

Notation:

Simulation vs. Theory-based Methods

Simulation-based method

Creation of the null distribution

- Simulate many samples assuming

- Find the proportion of ____________________________ at least as extreme as the observed sample __________________

- The null distribution estimates the sample to sample ______________________________ expected in the population

Theory-based method

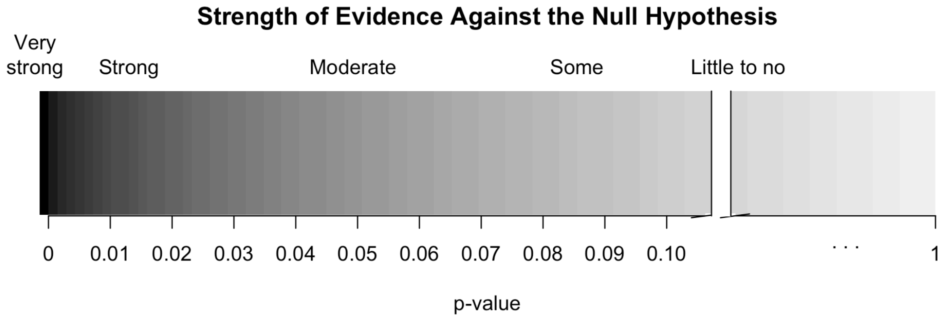
- Use a mathematical model to determine a distribution under the null hypothesis
- Compare the observed sample statistic to the model to calculate a probability
- *Theory-based methods will be discussed next week*

P-value

- What does the p-value measure?

  - Probability of observing the sample ____________________________ or more ________________________ assuming the ____________ hypothesis is _______________.

- How much evidence does the p-value provide against the null hypothesis?



- The ___________________________the p-value, the _____________________________ the evidence against the null hypothesis.

- Write a conclusion based on the p-value.

  - Answers the ________________________ question.

  - Amount of _________________________ in support of the _________________________ hypothesis.

- Decision: can we reject or fail to reject the null hypothesis?

  - Significance level: cut-off of “small” vs “large” p-value

    - p-value $\leq \alpha$

      - Strong enough evidence against the null hypothesis

      - Decision:

      - Results are __________________________________ significant.

    - p-value $> \alpha$

      - Not enough evidence against the null hypothesis

      - Decision:

      - Results are not _______________________________ significant.

One proportion test

- Reminder: review summary measures and plots discussed in the Week 3 material and Chapter 4 of the textbook.
- The summary measure for a single categorical variable is a _____________________.

Notation:

- Population proportion:
- Sample proportion:

Parameter of Interest:

- Include:
  - Reference of the population (true, long-run, population, all)
  - Summary measure
  - Context
    - Observational units/cases
    - Response variable (and explanatory variable if present)
      - If the response variable is categorical, define a 'success' in context

π :

Hypothesis testing

Conditions:

- Independence:

Null hypothesis assumes "no effect", "no difference", "nothing interesting happening", etc.

Always of form: "parameter" = null value

H_0 :

H_A :

- Research question determines the alternative hypothesis.

Example: A 2007 study published in the Behavioral Ecology and Sociobiology journal was titled "Why do blue-eyed men prefer blue-eyed women?" In this study, conducted in Norway, 114 volunteer heterosexual blue-eyed males rated the attractiveness of 120 pictures of females. The researchers recorded which eye-color (blue, green, or brown) was rated the highest, on average. In the sample, 51 of the volunteers rated the blue-eyed women the most attractive. Do blue-eyed heterosexual men tend to find blue-eyed women the most attractive?

Parameter of interest:

Write the null and alternative hypotheses for the blue-eyed study:

In words:

H_0 :

H_A :

In notation:

H_0 :

H_A :

Statistic:

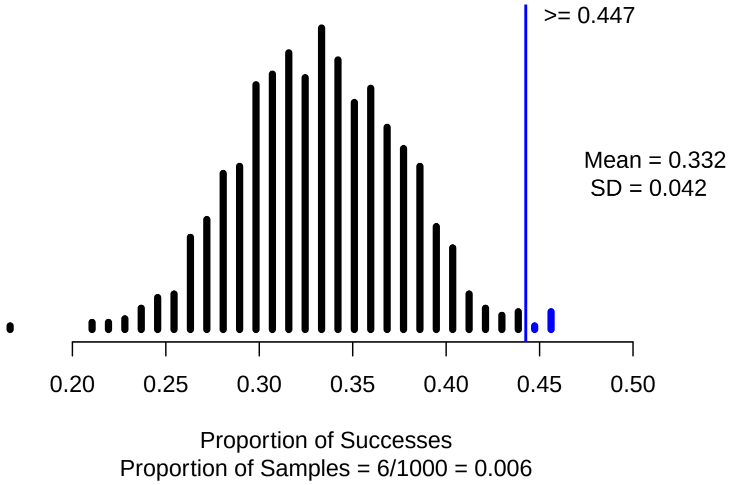
Is the independence condition met to analyze these data using a simulation-based approach?

Simulation-based method

- Simulate many samples assuming $H_0 : \pi = \pi_0$
  - Create a spinner with that represents the null value
  - Spin the spinner n times
  - Calculate and plot the simulated sample proportion from each simulation
  - Repeat 1000 times (simulations) to create the null distribution

- Find the proportion of simulations at least as extreme as $\hat{p}$

```
set.seed(216)
one_proportion_test(probability_success = 0.333, # Null hypothesis value
          sample_size = 114, # Enter sample size
          number_repetitions = 1000, # Enter number of simulations
          as_extreme_as = 0.447, # Observed statistic
          direction = "greater", # Specify direction of alternative hypothesis
          summary_measure = "proportion") # Reporting proportion or number of successes?
```



Explain why the null distribution is centered at the value of approximately 0.333:

Interpretation of the p-value:

- Statement about probability or proportion of samples
- Statistic (summary measure and value)
- Direction of the alternative
- Null hypothesis (in context)

Conclusion:

- Amount of evidence
- Parameter of interest
- Direction of the alternative hypothesis

Generalization:

- Can the results of the study be generalized to the target population?

Confidence interval

statistic ± margin of error

Vocabulary:

- Point estimate:
- Margin of error:

Purpose of a confidence interval

- To give an ______________________________ ____________________________ for the parameter of interest
- Determines how _____________________ an effect is

Sampling distribution

- Ideally, we would take many samples of the same ________________ from the same population to create a sampling distribution
- But only have 1 sample, so we will _________________________ with _________________________ from the one sample.
- Need to estimate the sampling distribution to see the _________________________ in the sample

Simulation-based methods

Bootstrap distribution:

- Write the response variable values on cards
- Sample with replacement n times (bootstrapping)
- Calculate and plot the simulated difference in sample means from each simulation
- Repeat 1000 times (simulations) to create the bootstrap distribution
- Find the cut-offs for the middle X% (confidence level) in a bootstrap distribution.

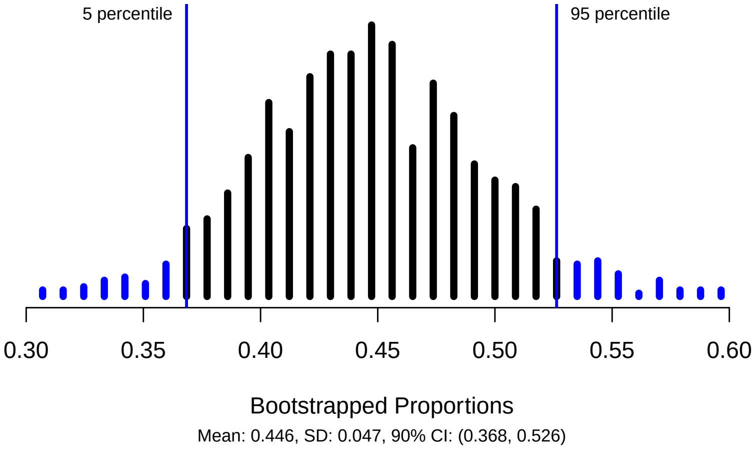
What is bootstrapping?

- Assume the "population" is many, many copies of the original sample.
- Randomly sample with replacement from the original sample n times.

Let's revisit the blue-eyed male study to estimate the *proportion of ALL heterosexual blue-eyed males who tend to find blue-eyed women the most attractive* by creating a 90% confidence interval.

Bootstrap distribution:

```
set.seed(216)
one_proportion_bootstrap_CI(sample_size = 114, # Sample size
                            number_successes = 51, # Observed number of successes
                            number_repetitions = 1000, # Number of bootstrap samples to use
                            confidence_level = 0.90) # Confidence level as a decimal
```



Confidence interval interpretation:

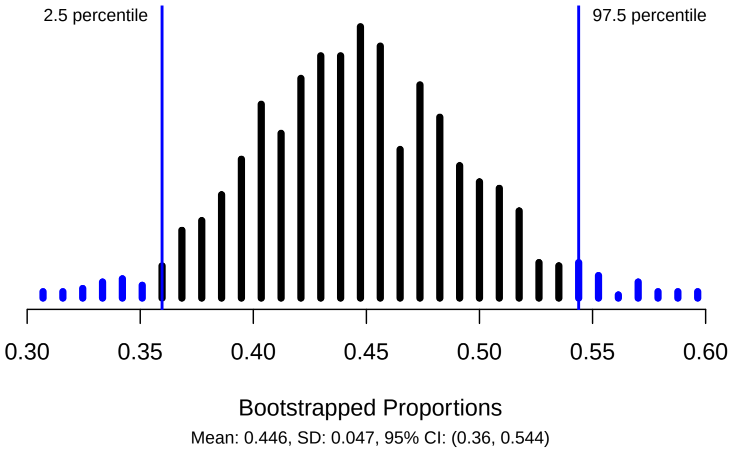
- How confident you are (e.g., 90%, 95%, 98%, 99%)
- Parameter of interest
- Calculated interval
- Order of subtraction when comparing two groups

Do the results of the confidence interval *match* the results based on the p-value?

How does changing the confidence level impact the width of the confidence interval?

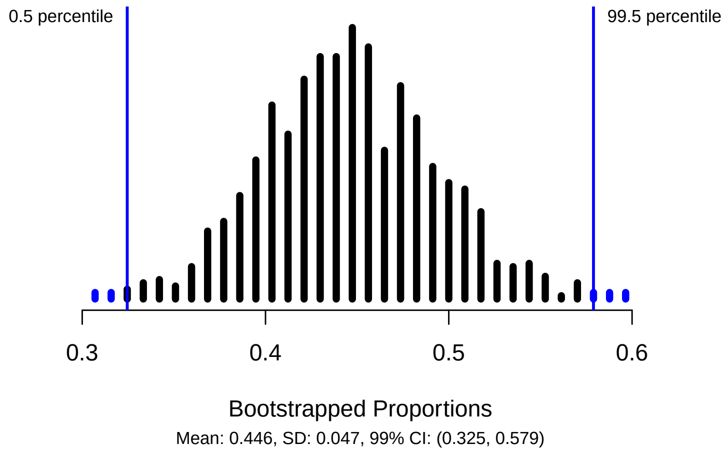
95% Confidence Interval:

```
set.seed(216)
one_proportion_bootstrap_CI(sample_size = 114, # Sample size
                            number_successes = 51, # Observed number of successes
                            number_repetitions = 1000, # Number of bootstrap samples to use
                            confidence_level = 0.95) # Confidence level as a decimal
```



99% Confidence Interval:

```
set.seed(216)
one_proportion_bootstrap_CI(sample_size = 114, # Sample size
                            number_successes = 51, # Observed number of successes
                            number_repetitions = 1000, # Number of bootstrap samples to use
                            confidence_level = 0.99) # Confidence level as a decimal
```



1.3 Out of Class Activity 6: Helperer-Hinderer — Simulation-based Hypothesis Test

1.3.1 Learning outcomes

- 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.

1.3.2 Terminology review

In today's 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

To review these concepts, see Chapters 9 & 14 in your textbook.

1.3.3 Steps of the statistical investigation process

We will work through a five-step process to complete a hypothesis test for a single proportion, first introduced in the 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.

1.3.4 Helper-Hinderer

A study by Hamblin, Wynn, and Bloom reported in Nature (Hamblin, Wynn, and Bloom 2007) was intended to check young kids' feelings about helpful and non-helpful behavior. Non-verbal infants ages 6 to 10 months were shown short videos with different shapes either helping or hindering the climber. As a class we will watch this short video to see how the experiment was run: https://youtu.be/anCaGBsBOxM. Researchers were hoping to assess: Are infants more likely to preferentially choose the helper toy over the hinderer toy? In the study, of the 16 infants age 6 to 10 months, 14 chose the *helper* toy and 2 chose the *hinderer* toy.

In this study, the **observational units are the infants ages 6 to 10 months**. The **variable measured on each observational unit (infant) is whether they chose the helper or the hinderer toy**. This is a categorical variable so we will be assessing the proportion of infants ages 6 to 10 months that choose the helper toy. Choosing the helper toy in this study will be considered a success.

Ask a research question

1. Identify the research question for this study. What are the researchers hoping to show?

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.

2. Are the cases independent? Justify your answer.

Summarize and visualize the data

The following code reads in the data set and gives the number of infants in each level of the variable, whether the infant chose the helper or the hinderer. Remember to visually display this data we can use either a frequency bar plot or a relative frequency bar plot.

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

```
#>     choice  n
#> 1   helper 14
#> 2 hinderer  2
```

$$\hat{p} = \frac{\text{number of successes}}{\text{total number of observational units}}$$

3. Using the `R` output and the formula given, calculate the summary statistic (sample proportion) to represent the research question. Recall that `choosing the helper toy` is a considered a success. Use appropriate notation.

4. Sketch a relative frequency bar plot of these data.

We cannot assess whether infants are more likely to choose the helper toy based on the statistic and plot alone. The next step is to analyze the data by using a hypothesis test to discover if there is evidence against the null hypothesis.

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 this study, the parameter of interest is the **true or population proportion of infants ages 6–10 months who will choose the helper toy**.

If the children are just randomly choosing the toy, we would expect half (0.5) of the infants to choose the helper toy. This is the null value for our study.

5. Using the parameter of interest given above, 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, $H_0 : \pi = \pi_0$.

6. Write the null hypothesis in notation using the null value of 0.5 in place of π_0 in the equation given on the previous page.

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.

7. Based on the research question from question 1, are we testing that the parameter is greater than 0.5, less than 0.5 or different than 0.5?

8. 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 infants who choose the helper is 0.5 and our results just occurred by random chance; or,
- The true proportion of infants who choose the helper is greater than 0.5 and our results reflect this.

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

We will now simulate a one sample of 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 infants who choose the helper is 0.5, so we will create 1000 (or more) different simulations of 16 infants under this assumption.

Let's think about how to use a coin to create one simulation of 16 infants under the assumption the null hypothesis is true. Let heads equal infant chose the helper toy and tails equal infant chose the hinderer toy.

9. How many times would you flip a coin to simulate the sample of infants?

10. Flip a coin 16 times recording the number of times the coin lands on heads. This represents one simulated sample of 16 infants randomly choosing the toy.

11. Is the value from question 10 closer to 0.5, the null value, or closer to the sample proportion, 0.875?

12. Report your simulated sample proportion to your instructor via the Google sheet provided on D2L. Sketch the distribution created by your class below.

13. Circle the observed statistic (value from question 3) on the distribution you drew in question 15. Where does this statistic fall in this distribution: Is it near the center of the distribution (near 0.5) or in one of the tails of the distribution?

In the next class, we will continue to assess the strength of evidence against the null hypothesis by using a computer to simulate 1000 samples when we assume the null hypothesis is true.

1.3.5 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. 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.

1.3.6 Additional notes

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

1.4 Activity 6: Helper-Hinderer (continued)

1.4.1 Learning outcomes

- 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.
- Explore what a p-value represents

1.4.2 Steps of the statistical investigation process

In today's activity we will continue with steps 4 and 5 in the statistical investigation process. We will continue to assess the Helper-Hinderer study from last class.

- **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.

1.4.3 Helper-Hinderer

A study by Hamblin, Wynn, and Bloom reported in Nature (Hamblin, Wynn, and Bloom 2007) was intended to check young kids' feelings about helpful and non-helpful behavior. Non-verbal infants ages 6 to 10 months were shown short videos with different shapes either helping or hindering the climber. As a class we will watch this short video to see how the experiment was run: https://youtu.be/anCaGBsBOxM. Researchers were hoping to assess: Are infants more likely to preferentially choose the helper toy over the hinderer toy? In the study, of the 16 infants age 6 to 10 months, 14 chose the *helper* toy and 2 chose the *hinderer* toy.

1. Report the sample proportion calculated in the out of class activity.

2. Write the alternative hypothesis in words in context of the problem. Remember the direction we are testing is dependent on the research question.

Today, we will use the computer to simulate a null distribution of 1000 different samples of 16 infants, plotting the proportion who chose the helper in each sample, based on the assumption that the true proportion of infants who choose the helper is 0.5 (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).

3. 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 3 in the one proportion test to create a null distribution with 1000 simulations. Then highlight and run lines 1–15.

```r
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
          summary_measure = "proportion") # Reporting proportion or number of successes?
```

4. Sketch the null distribution created from the `R` code here.

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

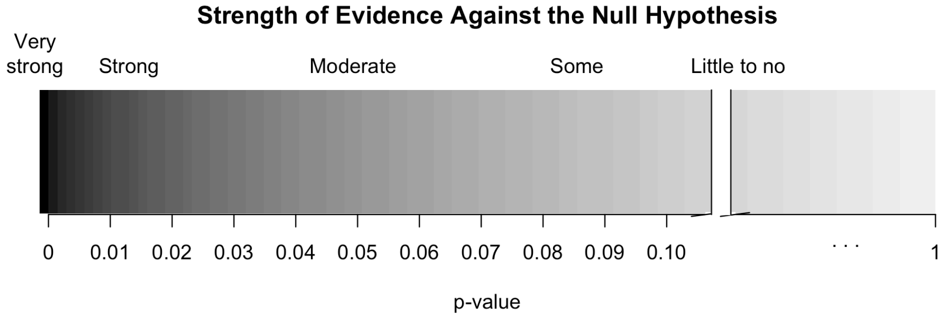
6. Circle the observed statistic (value from question 1) on the distribution you drew in question 4. Where does this statistic fall in the null distribution: Is it near the center of the distribution (near 0.5) or in one of the tails of the distribution?

7. Is the observed statistic likely to happen or unlikely to happen if the true proportion of infants who choose the helper is 0.5? Explain your answer using the plot.

8. 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 infants who choose the helper is 0.5? *Hint*: Look under the simulation.

The value in question 8 is the **p-value**. The smaller the p-value, the more evidence we have against the null hypothesis.

9. **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.)**



Interpret the p-value

The p-value measures the probability that we observe a sample proportion as extreme as what was seen in the data or more extreme (matching the direction of the Ha) IF the null hypothesis is true.

10. What did we assume to create the null distribution?

11. What value did we compare to the null distribution to find the p-value?

12. What direction did we count simulations from the statistic?

13. Fill in the blanks below to interpret the p-value.

We would observe a sample proportion of (value of the sample proportion)______________________________

or (greater, less, more extreme) ____

with a probability of (value of p-value) ____

IF we assume (H_0 in context) ____.

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.

14. Write a conclusion in context of the study. How much evidence does the data provide in support of the alternative hypothesis?

1.4.4 Take-home messages

1. The null distribution is created based on the assumption the null hypothesis is true. We compare the sample statistic to the distribution to find the likelihood of observing this statistic.
2. The p-value measures the probability of observing the sample statistic or more extreme (in direction of the alternative hypothesis) is the null hypothesis is true.

1.4.5 Additional notes

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

1.5 Week 6 Lab: Helper-Hinderer — Simulation-based Confidence Interval

1.5.1 Learning outcomes

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

1.5.2 Terminology review

In today's 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 Chapters 10 & 14 in your textbook.

1.5.3 Helper-Hinderer

In the last class, we found very strong evidence that the true proportion of infants who will choose the helper character is greater than 0.5. But what *is* the true proportion of infants who will choose the helper character? We will use this same study to estimate this parameter of interest by creating a confidence interval.

As a reminder: Do young children know the difference between helpful and unhelpful behavior? A study by Hamblin, Wynn, and Bloom reported in Nature (Hamblin, Wynn, and Bloom 2007) was intended to check young kids' feelings about helpful and non-helpful behavior. Non-verbal infants ages 6 to 10 months were shown short videos with different shapes either helping or hindering the climber. Researchers were hoping to assess: Are infants more likely to preferentially choose the helper toy over the hinderer toy? In the study, of the 16 infants age 6 to 10 months, 14 chose the *helper* toy and 2 chose the *hinderer* toy.

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**.

Activity intro

1. What is the value of the point estimate?

2. If we took another random sample of 16 infants, would we get the exact same point estimate? Explain why or why not.

In today's activity, we will use bootstrapping to find a 95% confidence interval for π, the parameter of interest.

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

Use statistical analysis methods to draw inferences from the data

4. Write out the parameter of interest for this study in words. *Hint: this is the same as in Activity 6A.*

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 choose the helper character),
- "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).

5. 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 5 in the one proportion bootstrap confidence interval (CI) code to create a bootstrap distribution with 1000 simulations. Then highlight and run lines 1–7.

```r
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
```

6. Sketch the bootstrap distribution created below.

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

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

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

10. **Interpret the 95% confidence interval in context.**

Communicate the results and answer the research question

11. **Is the value 0.5 (the null value) in the 95% confidence interval?**

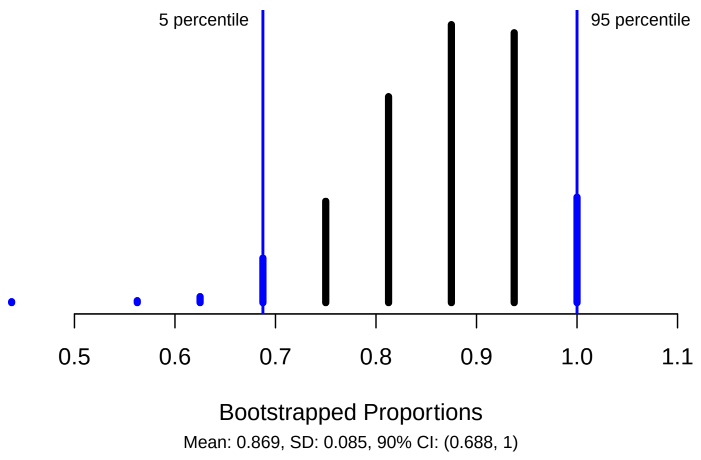
    Explain how this indicates that the p-value provides strong evidence against the null.

Effect of confidence level

12. 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.

13. 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 = 16, # Sample size
                            number_successes = 14, # Observed number of successes
                            number_repetitions = 1000, # Number of bootstrap samples to use
                            confidence_level = 0.90) # Confidence level as a decimal
```



14. Report both the 95% confidence interval (question 9) and the 90% confidence interval (question 13). Is the 90% confidence interval narrower or wider than the 95% confidence interval?

15. Explain why the upper value of the confidence interval is truncated at 1.

16. Fill in the blanks below to write a paragraph summarizing the results of the study as if writing a press release. **Complete your group's paragraph on Gradescope.**

    Researchers were interested if infants observe social cues and would be more likely to choose the helper toy over the hinderer toy. In a sample of (sample size) ____________________infants, (number of successes) ______________________chose the helper toy. A simulation null distribution with 1000 simulations was created in RStudio. The p-value was found by calculating the proportion of simulations in the null distribution at the sample statistic of 0.875 and greater. This resulted in a p-value of (value of p-value)______________________. We would observe a sample proportion of (value of the sample proportion) ________________________________ or (greater, less, more extreme) _______________________________ with a probability of (value of p-value)____________________________ IF we assume (H_0 in context) ____. Based on this p-value, there is (very strong/little to no) ________________________________ evidence that the (sample/true)________________________________ proportion of infants age 6 to 10 months who will choose the helper toy is (greater than, less than, not equal to) _______________________________ 0.5. In addition, a 95% confidence interval was found for the parameter of interest. We are 95% confident that the (true/sample)_____________________________________ proportion of infants age 6 to 10 months who will choose the helper toy is between (lower value)________________________ and (upper value)_____________________________. The results of this study can be generalized to (all infants age 6 to 10 months/infants similar to those in this study)____ as the researchers (did/did not)________________________________ select a random sample.

1.5.4 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 X 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 this range, where X represents the level of confidence.

4. 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.

5. 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.

1.5.5 Additional notes

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

MODULE 2

Exam 2 Review

Use the provided data set from the Islands (ExamReviewData.csv) and the appropriate Exam 2 Review R script file to answer the following questions. Each adult (>21) islander was selected at random from all adult islanders. Variables and their descriptions are listed below. Music type (classical or heavy metal) was randomly assigned to the Islanders. Time to complete the puzzle cube was measured after listening to music for each Islander. Heart rate and blood glucose levels were both measured before and then after drinking a caffeinated beverage.

Variable	**Description**
`Island`	Name of Island that the Islander resides on
`City`	Name of City in which the Islander resides
`Population`	Population of the City
`Name`	Name of Islander
`Consent`	Whether the Islander consented to be in the study
`Gender`	Gender of Islander (M = male, F = Female)
`Age`	Age of Islander
`Married`	Marital status of Islander
`Smoking_Status`	Whether the Islander is a current smoker
`Children`	Whether the Islander has children
`weight_kg`	Weight measured in kg
`height_cm`	Height measured in cm
`respiratory_rate`	Breaths per minute
`Type_of_Music`	Music type (Classical or Heavy Medal) Islander was randomly assigned to listen to
`After_PuzzleCube`	Time to complete puzzle cube (minutes) after listening to assigned music
`Education_Level`	Highest level of education completed
`Balance_Test`	Time balanced measured in seconds with eyes closed
`Blood_Glucose_before`	Level of blood glucose (mg/dL) before consuming assigned drink
`Heart_Rate_before`	Heart rate (bpm) before consuming assigned drink
`Blood_Glucose_after`	Level of blood glucose (mg/dL) after consuming assigned drink
`Heart_Rate_after`	Heart rate (bpm) after consuming assigned drink
`Diff_Heart_Rate`	Difference in heart rate (bpm) for Before - After consuming assigned drink
`Diff_Blood_Glucose`	Difference in blood glucose (mg/dL) for Before - After consuming assigned drink

1. Use the appropriate Exam 2 Review R script file and analyze the following research question: The proportion of university graduates in the US is 42%. “Is there evidence that the proportion of university graduates in the Islands differs from the proportion in the US?”

a. Parameter of Interest:

b. Null Hypothesis:

Notation:

Words:

c. Alternative Hypothesis:

Notation:

Words:

d. Use the R script file to get the counts for each level of the variable. Fill in the following table with the success, failure, variable name, and counts using the values from the R output.

Variable	**Counts**
Success	
Failure	
Total	

e. Calculate the value of summary statistic to answer the research question. Give appropriate notation.

f. Interpret the value of the summary statistic in context of the problem:

g. Assess if the following conditions are met:
Independence (needed for both simulation and theory-based methods):

Success-Failure (must be met to use theory-based methods):

h. Use the provided R script file to find the simulation p-value to assess the research question. Report the p-value.

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

j. Write a conclusion to the research question based on the p-value.

k. Write a decision based on the p-value.

l. Use the provided R script file to find a 90% confidence interval.

m. Interpret the 90% confidence interval in context of the problem.

n. Regardless to your answer in part g, calculate the standardized statistic.

o. Interpret the value of the standardized statistic in context of the problem.

p. Use the provided R script file to find the theory-based p-value.

q. Use the provided R script file to find the appropriate z* multiplier and calculate the theory-based confidence interval.

r. Does the theory-based p-value and CI match those found using simulation methods? Explain why or why not.

s. To what group can the results be generalized?

2. Use the appropriate Exam 2 Review R script file and analyze the following research question: “Is there evidence that those with a higher education level are less likely to smoke?”

a. Parameter of Interest:

b. Null Hypothesis:

Notation:

Words:

c. Alternative Hypothesis:

Notation:

Words:

d. Use the R script file to get the counts for each level and combination of variables. Fill in the following table with the variable names, levels of each variable, and counts using the values from the R output.

	Explanatory Variable		
Response variable	Group 1	Group 2	Total
Success			
Failure			
Total			

d. Calculate the value of summary statistic to answer the research question. Give appropriate notation.

e. Interpret the value of the summary statistic in context of the problem:

g. Assess if the following conditions are met:
Independence (needed for both simulation and theory-based methods):

Success-Failure (must be met to use theory-based methods):

h. Use the provided R script file to find the simulation p-value to assess the research question. Report the p-value.

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

j. Write a conclusion to the research question based on the p-value.

k. Write a decision based on the p-value.

l. Use the provided R script file to find a 95% confidence interval.

m. Interpret the 95% confidence interval in context of the problem.

n. Regardless to your answer in part g, calculate the standardized statistic.

o. Interpret the value of the standardized statistic in context of the problem.

p. Use the provided R script file to find the theory-based p-value.

q. Use the provided R script file to find the appropriate z* multiplier and calculate the theory-based confidence interval.

r. Does the theory-based p-value and CI match those found using simulation methods? Explain why or why not.

s. What is the scope of inference for this study?

MODULE 3

Inference for a Quantitative Response with Paired Samples

3.1 Module 11 Reading Guide: Inference for a Single Mean or Paired Mean Difference

Chapter 17 (Inference for a single mean)

Videos

- 17.1
- 17.2
- 17.3Tests
- 17.4Intervals

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

Shifted bootstrap test:

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 shifted bootstrap distribution test,

How many cards will you need and how will the cards be labeled?

Why are the data values shifted prior to being written on the cards?

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 single mean:

Notation

μ_0 represents

Example from section 17.1: 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 17.3, give the confidence interval for the true mean 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.

Example from section 17.2: Sleep times of MSU students

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. Write the null and the alternative hypotheses in words.

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

8. How could we use cards to simulate **one** shifted bootstrap resample *which assumes the null hypothesis is true*? How many cards? What is written on the cards (be sure to include the amount and direction of the shift)? What would we do with the cards? What would you record once you have a simulated sample?

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

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

11. At the 5% significance level, what decision would you make? What type of error might that be?

12. What conclusion should the researcher make?

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

Example from section 17.3: 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 from section 17.3: 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.

Chapter 18 (Inference for paired mean difference)

Videos

- Paired_Data
- 18.1and18.2
- 18.3

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. Each volunteer's heart rate was measured at the end of 30 seconds, then the volunteer 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 had 5 full-day instruction at the time of the study, the other from Chief Joseph Middle School which utilized 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 from section 18.1: 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** shifted 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 from sections 18.2 and 18.3: UCLA 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.

3.2 Lecture Notes Week 11: Inference for a Single Mean & Paired Data

3.2.1 Single quantitative variable

- Reminder: review summary measures and plots discussed in the Week 3 material and Chapter 5 of the textbook.
- The summary measure for a single quantitative variable is the ____________________.

Notation:

- Population mean:
- Population standard deviation:
- Sample mean:
- Sample standard deviation:
- Sample size:

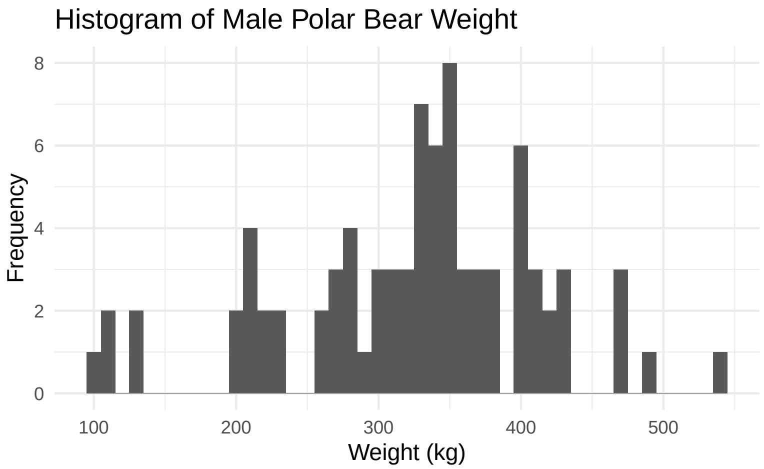
Example: What is the average weight of adult male polar bears? The weight was measured on a representative sample of 83 male polar bears from the Southern Beaufort Sea.

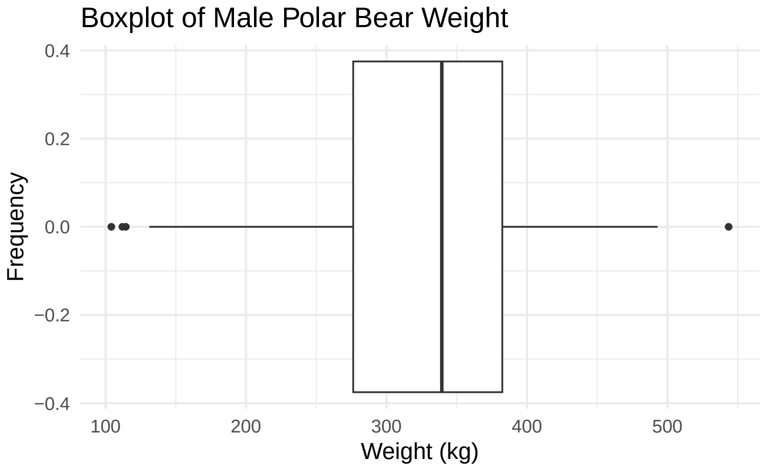
```r
pb <- read.csv("https://math.montana.edu/courses/s216/data/polarbear.csv")
```

Plots of the data:

```r
pb %>%
    ggplot(aes(x = Weight)) +   # Name variable to plot
    geom_histogram(binwidth = 10) +  # Create histogram with specified binwidth
    labs(title = "Histogram of Male Polar Bear Weight", # Title for plot
       x = "Weight (kg)", # Label for x axis
       y = "Frequency") # Label for y axis

pb %>% # Data set piped into...
ggplot(aes(x = Weight)) +   # Name variable to plot
  geom_boxplot() +  # Create boxplot
  labs(title = "Boxplot of Male Polar Bear Weight", # Title for plot
       x = "Weight (kg)", # Label for x axis
       y = "Frequency") # Label for y axis
```





Summary Statistics:

```
pb %>%
  summarise(favstats(Weight)) #Gives the summary statistics
#>     min    Q1 median     Q3   max     mean       sd  n missing
#> 1 104.1 276.3  339.4 382.45 543.6 324.5988 88.32615 83       0
```

Confidence interval

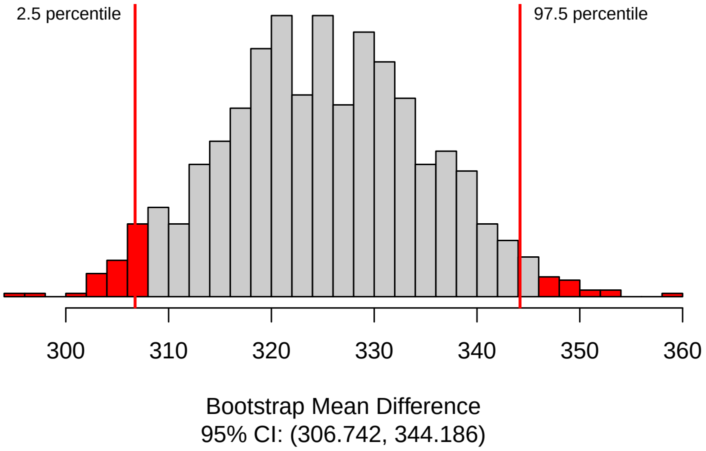
Simulation-based method

- Label cards with the values from the data set
- Sample with replacement (bootstrap) from the original sample n times
- Plot the simulated sample mean on the bootstrap distribution
- Repeat at least 1000 times (simulations)
- Find the cut-offs for the middle X% (confidence level) in a bootstrap distribution.
- ie. 95% CI = (2.5th percentile, 97.5th percentile)

Conditions for inference for a single mean:

- Independence:

```
set.seed(216)
paired_bootstrap_CI(data = pb$Weight, # Enter vector of differences
           number_repetitions = 1000, # Number of bootstrap samples for CI
           confidence_level = 0.95,  # Confidence level in decimal form
           which_first = 1)  # Not needed when entering vector of differences
```



The confidence interval estimates the ______________________________ of ______________________________________.

Confidence interval interpretation:

- How confident you are (e.g., 90%, 95%, 98%, 99%)
- Parameter of interest
- Calculated interval
- Order of subtraction when comparing two groups

Theory-based method

- Calculate the interval centered at the sample statistic

  statistic ± margin of error

Conditions for inference using theory-based methods:

- Independence:

- Large enough sample size:

T - distribution

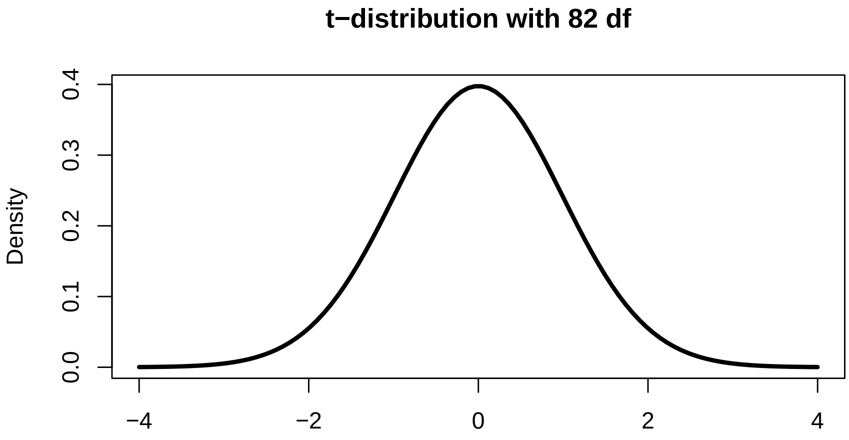
In the theoretical approach, we use the CLT to tell us that the distribution of sample means will be approximately normal, centered at the assumed true mean under H_0 and with standard deviation $\frac{\sigma}{\sqrt{n}}$.

$$\bar{x} \sim N(\mu_0, \frac{\sigma}{\sqrt{n}})$$

- Estimate the population standard deviation, σ, with the ____ standard deviation, ____________.

- For a single quantitative variable we use the ______ - distribution with ______________________ degrees of freedom to approximate the sampling distribution.

The t^* multiplier is the value at the given percentile of the t-distribution with $n - 1$ degrees of freedom.



To find the t^* multiplier for a 95% confidence interval:

```
qt(0.975, df = 82)
#> [1] 1.989319
```

Calculation of the confidence interval for the true mean weight of polar bears from the Southern Beaufort Sea:

Hypothesis testing

- Hypotheses are always written about the ______________________________. For a single mean we will use the notation ____________.

Null Hypothesis:

H_0 :

Alternative Hypothesis:

H_A :

- Direction of the alternative depends on the ______________________ ______________________.

Simulation-based method

- Simulate many samples assuming $H_0 : \mu = \mu_0$
  - Shift the data by the difference between μ_0 and $\bar{x}$
  - Sample with replacement n times from the shifted data
  - Plot the simulated shifted sample mean from each simulation
  - Repeat 1000 times (simulations) to create the null distribution
  - Find the proportion of simulations at least as extreme as $\bar{x}$

Example: Is there evidence that male polar bears weigh less than 370kg (previously recorded measure), on average? The weight was measured on a representative sample of 83 male polar bears from the Southern Beaufort Sea.

Hypotheses:

In notation:

H_0 :

H_A :

In words:

H_0 :

H_A :

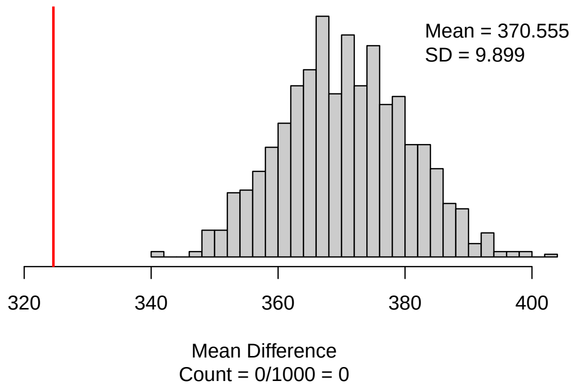
Reminder of summary statistics:

```
pb %>%
  summarise(favstats(Weight)) #Gives the summary statistics
#>     min    Q1 median     Q3   max     mean       sd  n missing
#> 1 104.1 276.3  339.4 382.45 543.6 324.5988 88.32615 83       0
```

Find the difference:

$\mu_0 - \bar{x} =$

```
set.seed(216)
paired_test(data = pb$Weight,   # Vector of differences
                                         # or data set with column for each group
            shift = 45.4,    # Shift needed for bootstrap hypothesis test
            as_extreme_as = 324.6,   # 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
```



Interpretation of the p-value:

- Statement about probability or proportion of samples
- Statistic (summary measure and value)
- Direction of the alternative
- Null hypothesis (in context)

Conclusion:

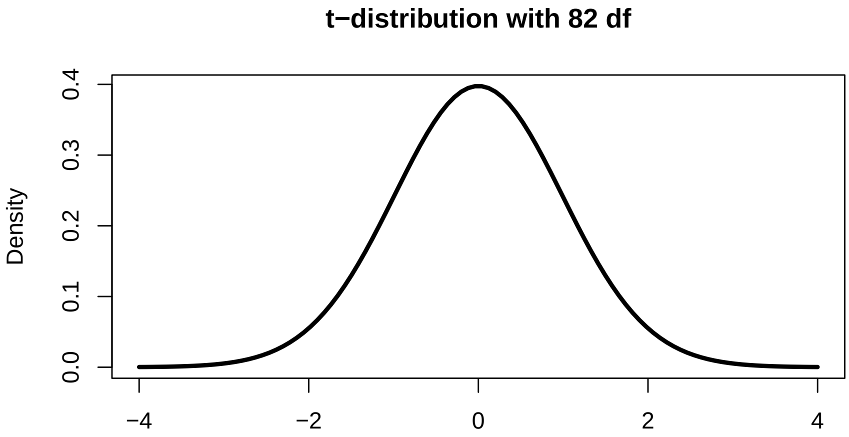
- Amount of evidence
- Parameter of interest
- Direction of the alternative hypothesis

Theory-based method

- Calculate the standardized statistic
- Find the area under the t-distribution with $n-1$ df at least as extreme as the standardized statistic

Standardized sample mean:

Calculate the standardized sample mean weight of adult male polar bears:



Interpret the standardized sample mean weight:

To find the theory-based p-value:

```
pt(-4.683, df=82, lower.tail=TRUE)
#> [1] 5.531605e-06
```

3.2.2 Paired vs. Independent Samples

Two groups are paired if an observational unit in one group is connected to an observational unit in another group

Data is paired if the samples are ______________________________

Examples:

- Change in test score from pre and post test
- Weight of college students before and after 1st year
- Change in blood pressure

Example: Three hundred registered voters were selected at random to participate in a study on attitudes about how well the president is performing. They were each asked to answer a short multiple-choice questionnaire and then they watched a 20-minute video that presented information about the job description of the president. After watching the video, the same 300 selected voters were asked to answer a follow-up multiple-choice questionnaire.

- Is this an example of a paired samples or independent samples study?

Thirty dogs were selected at random from those residing at the humane society last month. The 30 dogs were split at random into two groups. The first group of 15 dogs was trained to perform a certain task using a reward method. The second group of 15 dogs was trained to perform the same task using a reward-punishment method.

- Is this an example of a paired samples or independent samples study?

Fifty skiers volunteered to study how different waxes impacted their downhill race times. The participants were split into groups of two based on similar race times from the previous race. One of the two then had their skis treated with Wax A while the other was treated with Wax B. The downhill ski race times were then measured for each of the 25 volunteers who used Wax A as well as for each of the 25 volunteers who used Wax B.

- Is this an example of a paired samples or independent samples study?

For a paired experiment, we look at the difference between responses for each unit (pair), rather than just the average difference between treatment groups

- The summary measure for paired data is the ___________________________________.
- Mean difference: the average _______________________ in the __________________________ variable outcomes for observational units between ______________________________ variable groups

Parameter of Interest:

- Include:
  - Reference of the population (true, long-run, population, all)
  - Summary measure
  - Context
    - Observational units/cases
    - Response variable (and explanatory variable if present)
      - If the response variable is categorical, define a ‘success’ in context

μ_d :

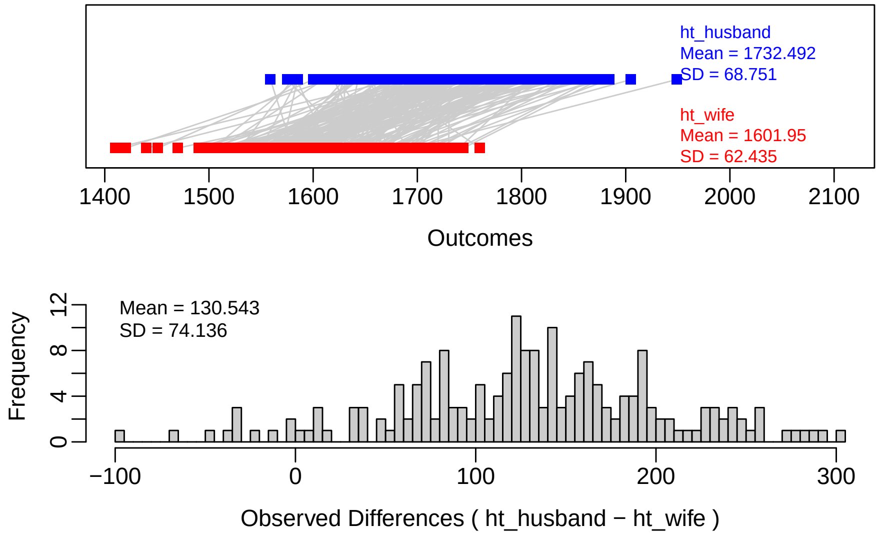
Notation for the Sample Statistics

- Sample mean of the differences:
- Sample standard deviation of the differences:

Example: Is there a difference in heights between husbands and wives? The heights were measured on the husband and wife in a random sample of 199 married couples from Great Britain.

Parameter of interest:

```
hw <-read.csv("data/husbands_wives_ht.csv")
paired_observed_plot(hw)
```



```
hw_diff <- hw %>%
  select(ht_husband, ht_wife) %>%
  mutate(ht_diff = ht_husband-ht_wife)
```

```
hw_diff %>%
    summarise(favstats(ht_husband))
#>     min   Q1 median   Q3  max     mean       sd   n missing
#> 1 1559 1691   1725 1774 1949 1732.492 68.75067 199       0
```

```
hw_diff %>%
    summarise(fav_stats(ht_wife))
#>     min   Q1 median   Q3  max    mean     sd   n missing
#> 1 1410 1560   1600 1650 1760 1601.95 62.435 199       0
```

```
hw_diff %>%
    summarise(fav_stats(ht_diff))
#>   min   Q1 median  Q3 max     mean       sd   n missing
#> 1 -96 83.5    131 179 303 130.5427 74.13608 199       0
```

Hypothesis testing

Null hypothesis assumes “no effect”, “no difference”, “nothing interesting happening”, etc.

- Treat the differences like a single mean
- Always of form: “parameter” = null value

H_0 :

H_A :

- Research question determines the alternative hypothesis.

Write the null and alternative for the height study:

In words:

H_0 :

H_A :

In notation:

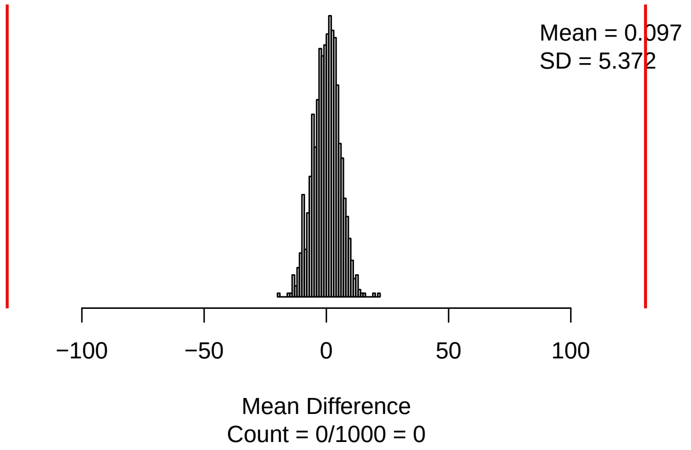
H_0 :

H_A :

Simulation-based method

Simulated null distribution:

```
set.seed(216)
paired_test(data = hw_diff$ht_diff,   # Vector of differences
                                        # or data set with column for each group
            shift = -130.543,   # Shift needed for bootstrap hypothesis test
            as_extreme_as = 130.543,  # Observed statistic
            direction = "two-sided",  # Direction of alternative
            number_repetitions = 1000,  # Number of simulated samples for null distribution
            which_first = 1)  # Not needed when using calculated differences
```



Interpret the p-value:

- Statement about probability or proportion of samples
- Statistic (summary measure and value)
- Direction of the alternative
- Null hypothesis (in context)

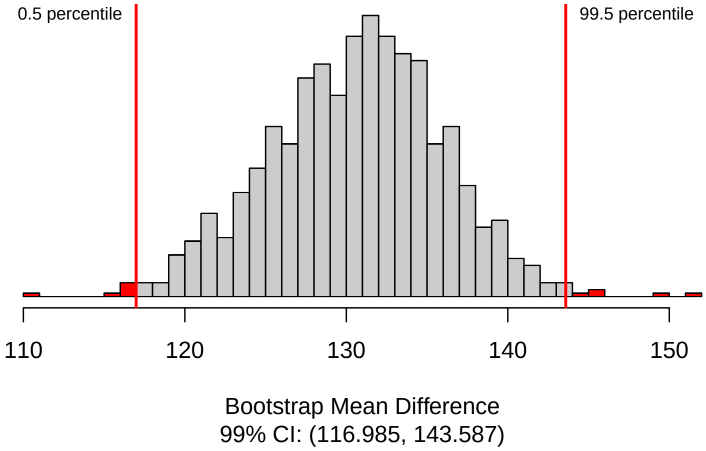
Conclusion:

- Amount of evidence
- Parameter of interest
- Direction of the alternative hypothesis

Confidence interval

Simulated bootstrap distribution:

```
set.seed(216)
paired_bootstrap_CI(data = hw_diff$ht_diff, # Enter vector of differences
            number_repetitions = 1000, # Number of bootstrap samples for CI
            confidence_level = 0.99,  # Confidence level in decimal form
            which_first = 1)  # Not needed when entering vector of differences
```



Interpret the 99% confidence interval:

- How confident you are (e.g., 90%, 95%, 98%, 99%)
- Parameter of interest
- Calculated interval
- Order of subtraction when comparing two groups

Theory-based method

```
hw_diff %>%
    summarise(fav_stats(ht_diff))
#>   min   Q1 median  Q3 max     mean       sd   n missing
#> 1 -96 83.5    131 179 303 130.5427 74.13608 199       0
```

Check the conditions to use theory-based methods:

Calculate the standardized sample mean difference in height:

Interpret the standardized statistic:

What theoretical distribution should we use to find the p-value using the value of the standardized statistic?

To find the p-value:

```
pt(24.84, df = 198, lower.tail=FALSE)*2
#> [1] 9.477617e-63
```

Calculate a 99% confidence interval:

- First need to find the $t*$ multiplier from the t-distribution with 198 df

```
qt(0.995, df=198, lower.tail = TRUE)
#> [1] 2.600887
```

Calculate the margin of error:

Calculate the theory-based confidence interval.

3.3 Out of Class Activity 11: Color Interference

3.3.1 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 theory-based hypothesis test for a paired mean difference.
- Interpret and evaluate a p-value for a theory-based hypothesis test for a paired mean difference.
- Use theory-based methods 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.

3.3.2 Terminology review

In today's activity, we will analyze paired quantitative data using theory-based methods. Some terms covered in this activity are:

- Paired data
- Mean difference
- Independent observational units
- Normality
- t-distribution
- Degrees of freedom
- T-score

To review these concepts, see Chapter 18 in the textbook.

3.3.3 Color Interference

The abstract of the article "Studies of interference in serial verbal reactions" in the *Journal of Experimental Psychology* (Stroop 1935) reads:

> In this study pairs of conflicting stimuli, both being inherent aspects of the same symbols, were presented simultaneously (a name of one color printed in the ink of another color—a word stimulus and a color stimulus). The difference in time for reading the words printed in colors and the same words printed in black is the measure of interference of color stimuli upon reading words. … The interference of conflicting color stimuli upon the time for reading 100 words (each word naming a color unlike the ink-color of its print) caused an increase of 2.3 seconds or 5.6% over the normal time for reading the same words printed in black.

The article reports on the results of a study in which seventy college undergraduates were given forms with 100 names of colors written in black ink, and the same 100 names of colors written in another color (i.e., the word purple written in green ink). The total time (in seconds) for reading the 100 words printed in black, and the total time (in seconds) for reading the 100 words printed in different colors were recorded for each subject. The order in which the forms (black or color) were given was randomized to the subjects. Does printing the name of colors in a different color increase the time it takes to read the words? Use color - black as the order of subtraction.

Identify the scenario

1. Should these observations be considered paired or independent? Explain your answer.

Ask a research question

2. Write out the null hypothesis in words, in the context of this study.

3. Write out the alternative hypothesis in proper notation for this study.

In general, the sampling distribution for a sample mean, $\bar{x}$, 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. For paired data, that means each pairwise difference should be 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 the distribution of the data is approximately normal with 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 in the data, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.
  - $n \geq 100$: If the sample size n is at least 100 (regardless of the presence of skew or outliers), we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

Like we saw in Chapter **5**, we will not know the values of the parameters and must use the sample data to estimate them. Unlike with proportions, in which we only needed to estimate the population proportion, π, quantitative sample data must be used to estimate both a population mean μ and a population standard deviation σ. This additional uncertainty will require us to use a theoretical distribution that is just a bit wider than the Normal distribution. Enter the ***t*-distribution**!

As you can seen from Figure 3.1, the t-distributions (dashed and dotted lines) are centered at 0 just like a standard Normal distribution (solid line), but are slightly wider. The variability of a t-distribution depends on its degrees of freedom, which is calculated from the sample size of a study. (For a single sample of n observations or paired differences, the degrees of freedom is equal to $n - 1$.) Recall from previous classes that larger sample sizes tend to result in narrower sampling distributions. We see that here as well. The larger the sample size, the larger the degrees of freedom, the narrower the t-distribution. (In fact, a t-distribution with infinite degrees of freedom actually IS the standard Normal distribution!)

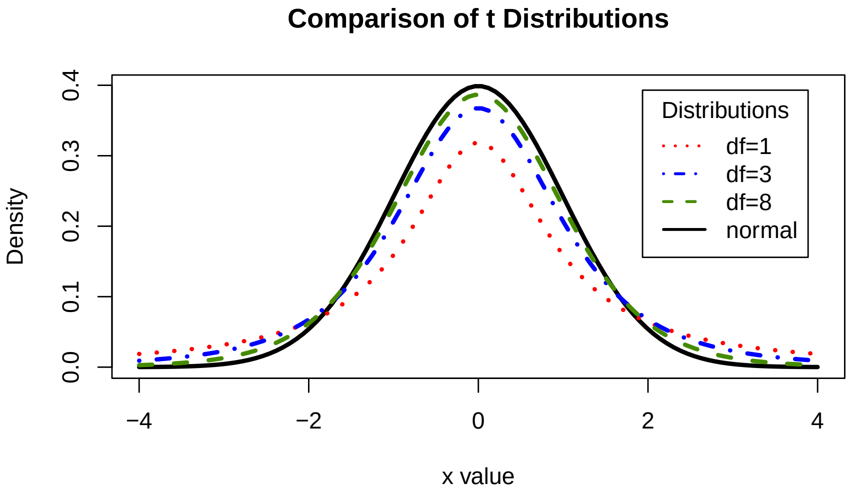


Figure 3.1: Comparison of the standard Normal vs t-distribution with various degrees of freedom

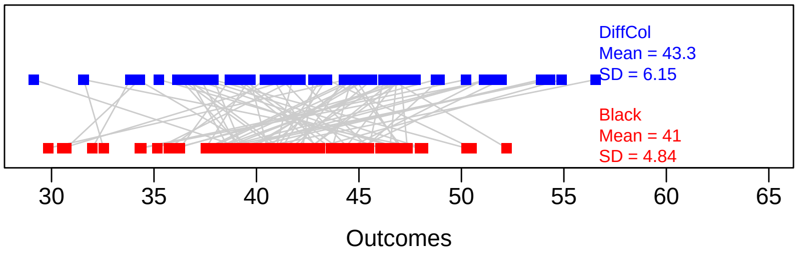
Summarize and visualize the data

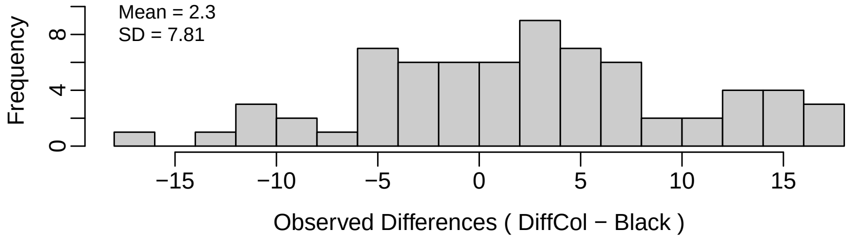
Since the original data from the study are not available, we simulated data to match the means and standard deviations reported in the article. We will use these simulated data in the analysis below.

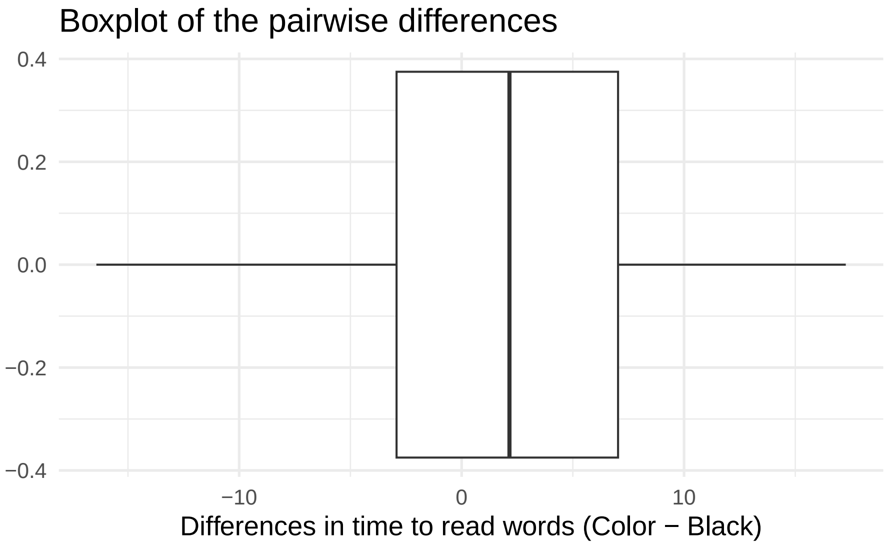
The following code plots each subject's time to read the colored words (above) and time to read the black words (below) connected by a grey line, a histogram of the differences in time to read words between the two conditions, and a boxplot displaying the pairwise differences in time (color − black).

```
color <- read.csv("https://math.montana.edu/courses/s216/data/interference.csv")
paired_observed_plot(color)

color_diff <- color %>%
  mutate(differences = DiffCol-Black)
color_diff %>%
  ggplot(aes(x = differences))+
  geom_boxplot()+
  labs(title="Boxplot of the pairwise differences",
       x = "Differences in time to read words (Color - Black)")
```







The following code gives the summary statistics for the pairwise differences.

```
color_diff %>%
  summarise(favstats(differences))
#>       min      Q1 median      Q3   max mean       sd  n missing
#> 1 -16.42 -2.925   2.15 7.0325 17.27  2.3 7.810196 70       0
```

Check theoretical conditions

4. How do you know the independence condition is met for these data?

5. Is the normality condition met to use the theory-based methods for analysis? Explain your answer.

Use statistical inferential methods to draw inferences from the data

To find the standardized statistic for the paired differences we will use the following formula:

$$T = \frac{\bar{x}_d - \mu_0}{SE(\bar{x}_d)},$$

where the standard error of the sample mean difference is:

$$SE(\bar{x}_d) = \frac{s_d}{\sqrt{n}}.$$

6. Calculate the standard error of the sample mean difference.

7. How many standard errors is the observed mean difference from the null mean difference?

To find the p-value we enter the value for the standardized statistic (T = 2.464) into the pt function in R. If you did not get his answer for question 7, double check your work. For a single sample or paired data, degrees of freedom are found by subtracting 1 from the sample size. You should therefore use `df` $= n_d - 1 = 70 - 1 = 69$ and `lower.tail = FALSE` to find the p-value.

```
pt(2.464, df=69, lower.tail=FALSE)
#> [1] 0.008117801
```

8. Explain why we found the area above the T-score using `lower.tail = FALSE` in the code above.

9. What does this p-value mean, in the context of the study? Hint: it is the probability of what…assuming what?

To calculate a 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()`. The code below will return the 95th percentile of the t distribution with `df` $= n_d - 1 = 70 - 1 = 69$.

```
qt(0.95, df = 69, lower.tail=TRUE)
#> [1] 1.667239
```

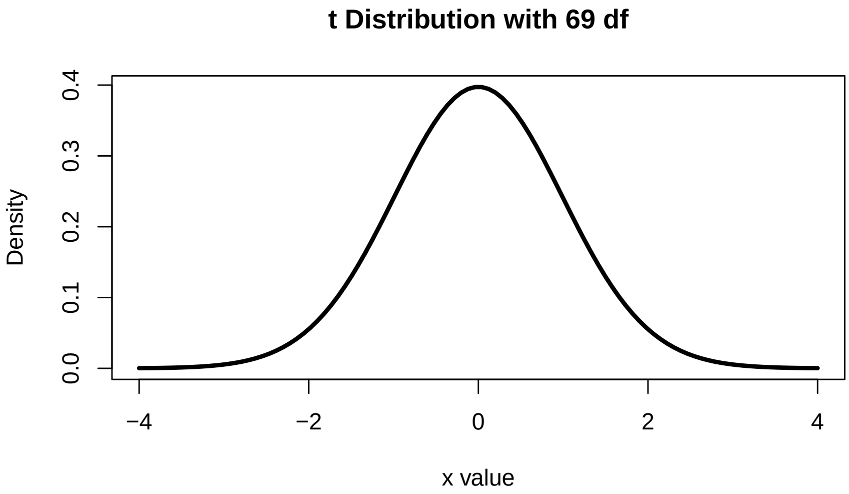


Figure 3.2: t-distribution with 69 degrees of freedom

10. In Figure 3.2, you see a t-distribution with 69 degrees of freedom. Label $t^\star$ and $-t^\star$ on that distribution. Write on the plot the percent of the t_{69}-distribution that is below $-t^\star$, between $-t^\star$ and $t^\star$, and above $t^\star$. Then use your plot to determine the confidence level associated with the $t^\star$ value obtained.

11. Calculate the margin of error for the true paired mean difference using theory-based methods.

12. Calculate the confidence interval for the true paired mean difference using theory-based methods.

13. Interpret the confidence interval in context of the study.

14. Do the results of the CI agree with the p-value? Explain your answer.

15. Write a conclusion to the test in context of the study.

3.3.4 Take-home messages

1. In order to use theory-based methods for dependent groups (paired data), the independent observational units and normality conditions must be met.
2. A T-score is compared to a t-distribution with $n-1$ df in order to calculate a one-sided p-value. To find a two-sided p-value using theory-based methods we need to multiply the one-sided p-value by 2.
3. A t^* multiplier is found by obtaining the bounds of the middle X% (X being the desired confidence level) of a t-distribution with $n-1$ df.

3.3.5 Additional notes

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

3.4 Activity 11: COVID-19 and Air Pollution

3.4.1 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.

3.4.2 Terminology review

In today'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 18 in the textbook.

3.4.3 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 3.3, 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 (US Environmental Protection Agency, n.d.). 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. Use Current - 5-year median as the order of subtraction.

Figure 3.3: The India Gate in New Delhi, India.

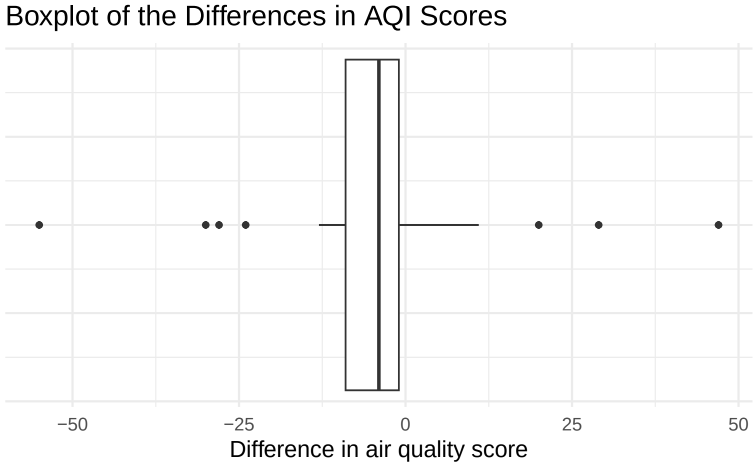


Table 3.1: Summary statistics for current AQI scores, median AQI scores from 2015–2019, and the differences in AQI scores.

	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.

1. Identify the variables in this study. What role (explanatory or response) do each have?

2. Are the differences in AQI scores independent for each case (US city)? Explain.

3. Why is this treated as a paired study design and not two independent samples?

Ask a research question

4. Write the null hypothesis in words.

5. What is the research question?

6. Write the alternative hypothesis in notation.

Summarize and visualize the data

7. Report the summary statistic of interest (mean difference) for the data.

8. What notation is used for the value in question 7?

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.

9. Calculate the difference $\mu_0 - \bar{x}_d$. Will we need to shift the data up or down?

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 9 for `xx` to simulate the null distribution and enter the summary statistic from question 7 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
```

10. Sketch the null distribution created using the R output here.

11. Explain why the null distribution is centered at zero.

12. 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?

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

14. How much evidence does this provide for improved air quality in US cities?

15. 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.

16. Write out the parameter of interest in context of the study.

17. 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.

```r
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

18. Interpret the 99% confidence interval in the context of the problem.

19. Do the results of your confidence interval and hypothesis test agree? What does each tell you about the null hypothesis?

3.4.4 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.

3.4.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today'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 Week 12.

3.5 Week 11 Lab: Swearing

3.5.1 Learning outcomes

- Identify whether a study is a paired design or independent groups
- Given a research question involving paired data, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Describe and perform a simulation-based hypothesis test for a mean difference.
- Interpret and evaluate a p-value for a hypothesis test for a mean difference.
- Use bootstrapping methods to find a confidence interval for a mean difference.
- Interpret a confidence interval for a mean difference.

3.5.2 Type of samples

For each of the following scenarios, determine whether the samples are paired or independent.

1. Researchers interested in studying the effect of a medical treatment on insulin rate measured insulin rates of 30 patients before and after the medical treatment.

2. **A university is planning to bring emotional support animals to campus during finals week and wants to determine which type of animals are more effective at calming students. Anxiety levels will be measured before and after each student interacts with either a dog or a cat. The university will then compare change in anxiety levels between the 'dog' people and the 'cat' people.**

3. An industry leader is investigating a possible wage gap between male and non-male employees. Twenty companies within the industry are randomly selected and the average salary for all males and non-males in mid-management positions is recorded for each company.

3.5.3 Swearing

Profanity (language considered obscene or taboo) and society's attitude about its acceptableness is a highly debated topic, but does swearing serve a physiological purpose or function? Previous research has shown that swearing produces increased heart rates and higher levels of skin conductivity. It is theorized that since swearing provokes intense emotional responses, it acts as a distracter, allowing a person to withstand higher levels of pain. To explore the relationship between swearing and increased pain tolerance, researchers from Keele University (Staffordshire, UK) recruited 83 native English-speaking participants (Stephens and Robertson 2020). Each volunteer performed two trials holding a hand in an ice-water bath, once while repeating the "f-word" every three seconds, and once while repeating a neutral word ("table"). The order of the word to repeat was randomly assigned. Researchers recorded the length of time, in seconds, from the moment the participant indicated they were in pain until they removed their hand from the ice water for each trial. They hope to find evidence that pain tolerance is greater (longer times) when a person swears compared to when they say a neutral word, on average. Use Swear – Neutral as the order of subtraction.

4. What does μ_d represent in the context of this study?

5. Write out the null hypothesis in proper notation for this study.

6. What sign ($<$, $>$, or $\neq$) would you use in the alternative hypothesis for this study? Explain your choice.

Upload and open the R script file for Week 11 lab. Upload and import the csv file, `pain_tolerance`. Enter the name of the data set (see the environment tab) for datasetname in the R script file in line 6. Highlight and run lines 1–7 to load the data and create a paired plot of the data.

```
swearing <- datasetname
paired_observed_plot(swearing)
```

7. Based on the plots, does there appear to be some evidence in favor of the alternative hypothesis? How do you know?

Enter the outcome for group 1 (`Swear`) for `measurement_1` and the outcome for group 2 (`Neutral`) for `measurement_2` in line 10. Highlight and run lines 9–12 to get the summary statistics for the data.

```
swearing_diff <- swearing %>%
  mutate(differences = measurement_1 - measurement_2)
swearing_diff %>%
    summarise(favstats(differences))
```

8. What is the value of $\bar{x}_d$? What is the sample size?

9. **How far, on average, is each difference in pain tolerance from the mean of the differences in pain tolerance? What is the appropriate notation for this value?**

Use statistical inferential methods to draw inferences from the data

10. Using the provided graphs and summary statistics, determine if both theory-based methods and simulation methods could be used to analyze the data. Explain your reasoning.

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 swearing does not affect pain tolerance, or that the length of time a subject kept their hand in the water would be the same whether the patient was swearing or not.

We will use the `paired_test()` function in R (in the `catstats` package) to simulate the null distribution of sample mean differences and compute a p-value.

11. When using the `paired_test()` function, we need to enter the name of the data set, either the order of subtraction (if the data set has both measurements) or the name of the differences (if the data set contains them). We will also need to provide R with the observed mean difference, the direction of the alternative hypothesis, and the shift required in order to force the null hypothesis to be true. The name of the data set as shown above is `swearing_diff` and the column of differences is called `differences`. What values should be entered for each of the following to create 1000 simulated samples?

- shift:

- As extreme as:

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

- Number of repetitions:

12. Simulate a null distribution and compute the p-value. Using the R script file for this lab, enter your answers for question 11 in place of the `xx`'s to produce the null distribution with 1000 simulations. Highlight and run lines 15–21.

```
paired_test(data = swearing$differences,   # Vector of differences
                                 # or data set with column for each group
        shift = xx,   # Shift needed for bootstrap hypothesis test
        as_extreme_as = xx,  # Observed statistic
        direction = "xx",  # Direction of alternative
        number_repetitions = xx,  # Number of simulated samples for null distribution
        which_first = 1)  # Not needed when using calculated differences
```

Sketch the null distribution created using the `paired_test` code.

Communicate the results and answer the research question

13. **Report the p-value. Based off of this p-value and a 1% significance level, what decision would you make about the null hypothesis? What potential error might you be making based on that decision?**

14. Do you expect the 98% confidence interval to contain the null value of zero? 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.

15. Using bootstrapping and the provided R script file, find a 98% confidence interval. Fill in the missing values/numbers in the `paired_bootstrap_CI()` function to create the 98% confidence interval. Highlight and run lines 24–27. **Upload a copy of the bootstrap distribution created to Gradescope for your group.**

```
paired_bootstrap_CI(data = swearing_diff$differences, # 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
```

Sketch the bootstrap distribution created using the code. Report the 98% confidence interval in interval notation.

16. Interpret the *confidence level* of the interval you calculated in question 15.

17. Write a paragraph summarizing the results of this study as if you were describing the results to your roommate. **Upload a copy of your group's paragraph to Gradescope.** Be sure to describe:

- Summary statistic
- P-value and interpretation
- Conclusion (written to answer the research question)
- Confidence interval and interpretation
- Scope of inference

MODULE 4

Inference for a Quantitative Response with Independent Samples

4.1 Module 12 Reading Guide: Inference for a Difference in Two Means

Chapter 19 (Inference for comparing two independent means)

Videos

- 19.1
- 19.2
- 19.3Tests
- 19.3Intervals

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 independent sample means,

How many cards will you need and how will the cards be labeled?

What do you do with the cards after labeling them?

After shuffling, what value will be plotted on the simulated null distribution?

To create a **bootstrap distribution** of differences in independent sample means,

How many cards will you need and how will the cards be labeled?

What do you do with the cards after labeling them?

After shuffling, what value will be plotted on the bootstrap distribution?

Conditions to use the CLT for a difference in independent sample 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 independent sample 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 from section 19.1: 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 from section 19.2: 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 from section 19.3: North Carolina 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.

4.2 Lecture Notes Week 12: Inference for independent samples

4.2.1 Single categorical, single quantitative variable with independent samples

- In this week, we will study inference for a ________________________________ explanatory variable and a _____________________________________ response variable where the two groups are ____.

- Independent groups: When the measurements in one sample are not ____________________________ to the measurements in the other sample.

- Two random samples taken separately from two populations and the same response variable is recorded. Compare the average number of sick days off from work for people who had a flu shot and people who didn't.

- Participants are randomly assigned to one of two treatment conditions, and the same response variable is recorded.

Rather than analyzing the differences as a single mean we will calculate summary statistics on each sample.

- The summary measure for two independent groups is the ________________________________ in ____.

- Difference in means: the difference in average __________________________________ variable outcome for observational units between _____________________________ variable groups

Parameter of Interest:

- Include:
  - Reference of the population (true, long-run, population, all)
  - Summary measure
  - Context
    - Observational units/cases
    - Response variable (and explanatory variable if present)
      - If the response variable is categorical, define a 'success' in context

$\mu_1 - \mu_2$:

Notation for the Sample Statistics

- Sample mean for group 1:

- Sample mean for group 2:
- Sample difference in means:
- Sample standard deviation for group 1:
- Sample standard deviation for group 2:
- Sample size for group 1:
- Sample size for group 2:

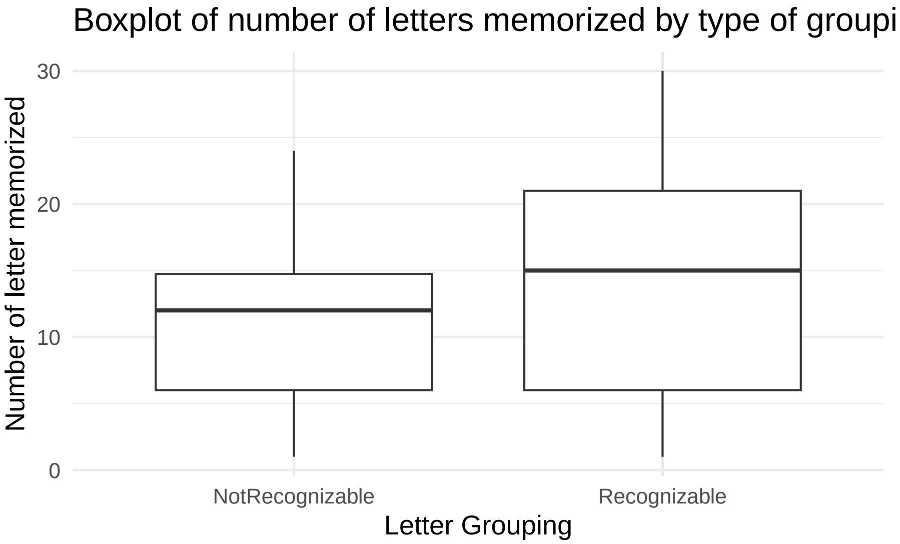
Example: Fifty-one (51) college students volunteered to look at impacts on memorization, specifically if putting letters into recognizable patterns (like FBI, CIA, EDA, CDC, etc.) would increase the number letters memorized. The college students were randomly assigned to either a recognizable or non-recognizable letter group. After a period of study time, the number of letters memorized was collect on each study. Is there evidence that putting letters into recognizable letter groups improve memory?

Why should we treat this as two independent groups rather than paired data?

Parameter of interest:

```
letters<-read.csv("data/letters.csv")
letters %>%
    reframe(favstats(Memorized~Grouped))
#>           Grouped min Q1 median    Q3 max     mean       sd  n missing
#> 1 NotRecognizable   1  6     12 14.75  24 11.15385 6.576883 26       0
#> 2    Recognizable   1  6     15 21.00  30 14.32000 8.518216 25       0
```

```
letters%>%
  ggplot(aes(y = Memorized, x = Grouped))  + #Enter the name of the explanatory and response variable
  geom_boxplot()+
  labs(title = "Boxplot of number of letters memorized by type of grouping", #Title your plot
       y = "Number of letter memorized", #y-axis label
       x = "Letter Grouping") #x-axis label
```



Hypothesis Testing

Conditions:

- Independence:

Null hypothesis assumes "no effect", "no difference", "nothing interesting happening", etc.

Always of form: "parameter" = null value

H_0 :

H_A :

- Research question determines the alternative hypothesis.

Write the null and alternative hypotheses for the letters study:

In words:

H_0 :

H_A :

In notation:

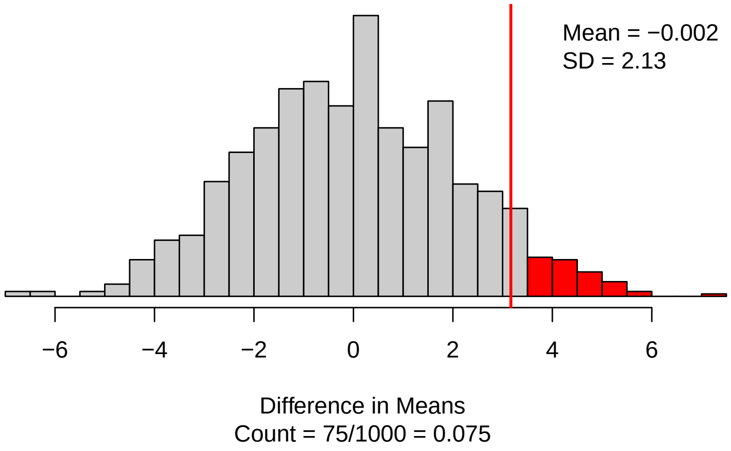
H_0 :

H_A :

Simulation-based method

- Simulate many samples assuming $H_0 : \mu_1 = \mu_2$
  - Write the response variable values on cards
  - Mix the explanatory variable groups together
  - Shuffle cards into two explanatory variable groups to represent the sample size in each group (n_1 and n_2)
  - Calculate and plot the simulated difference in sample means from each simulation
  - Repeat 1000 times (simulations) to create the null distribution
  - Find the proportion of simulations at least as extreme as $\bar{x}_1 - \bar{x}_2$

```
set.seed(216)
two_mean_test(Memorized~Grouped, #Enter the names of the variables
              data = letters,  # Enter the name of the dataset
              first_in_subtraction = "Recognizable", # First outcome in order of subtraction
              number_repetitions = 1000,  # Number of simulations
              as_extreme_as = 3.166,  # Observed statistic
              direction = "greater")  # Direction of alternative: "greater", "less", or "two-sided"
```



Explain why the null distribution is centered at the value of zero:

Interpretation of the p-value:

- Statement about probability or proportion of samples
- Statistic (summary measure and value)
- Direction of the alternative
- Null hypothesis (in context)

Conclusion:

- Amount of evidence
- Parameter of interest
- Direction of the alternative hypothesis

Theory-based method

Conditions:

- Independence
- Large enough sample size

Like with paired data the t-distribution can be used to model the difference in means.

- For a independent samples we use the _________- distribution with ________________________ degrees of freedom to approximate the sampling distribution.

Theory-based test:

- Calculate the standardized statistic
- Find the area under the t-distribution with the smallest $n - 1$ df [$\min(n_1 - 1, n_2 - 1)$ at least as extreme as the standardized statistic

Standardized difference in sample mean:

Example:

Every year, orange and black monarch butterflies migrate from their summer breeding grounds in the US and Canada to mountain forests in central Mexico, where they hibernate for the winter. Due to abnormal weather patterns and drought affecting monarch habitats and feeding grounds, the population of monarch butterflies is estimated to have decreased by 53% from the 2018-2019 wintering season to the 2019-2020 wintering season (WWF, 2020). While conservationists often resort to captive-rearing with the goal of raising biologically indistinct individuals for release into the wild, tagging studies have shown that captive-reared monarchs have lower migratory success compared to wild monarchs. For this study, the researchers raised 67 monarchs (descended from wild monarchs) from eggs to maturity and then compared them to a group of 40 wild-caught monarchs. The researchers want to explore whether the maximum grip strength (how many Newtons a butterfly exerts at the moment of release when gently tugged from a mesh-covered perch) differs between captive-reared and wild-caught monarchs. Use Captive – Wild for order of subtraction.

Write the null and alternative hypotheses in notation.

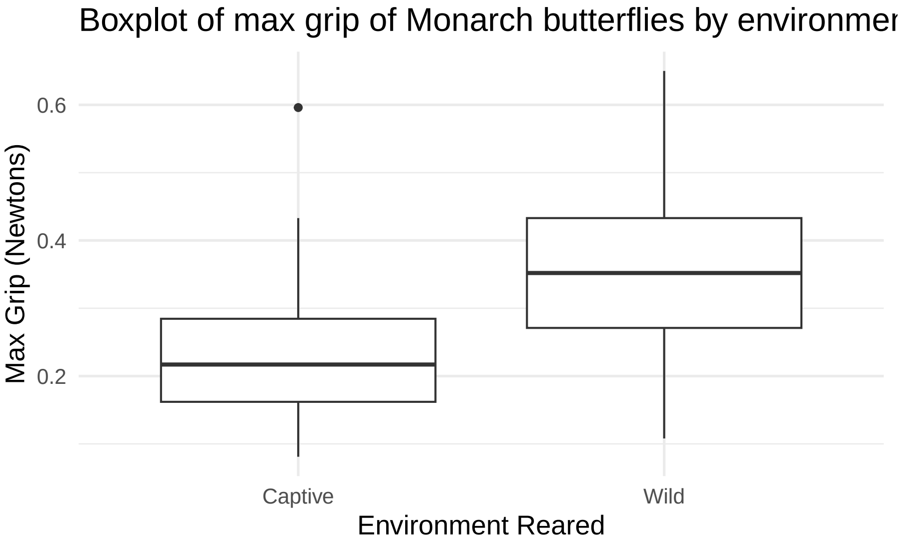
H_0 :

H_A :

```
butterfly <-read.csv("data/butterfly1.csv")
butterflies <- butterfly %>%
  na.omit() %>%
  rename(Monarch_Group = "Monarch.Group",
         MaxGrip = "Max.Grip.Strength..N.") %>%
  mutate(Monarch_Group = factor(Monarch_Group),
         Sex = factor(Sex)) %>%
  mutate(Monarch_Group = fct_collapse(Monarch_Group,
               "Captive" = c("Incubator - Fall conditions",
                             "Rearing room - summer conditions"),
               "Wild" = "Wild migrants"))

butterflies %>%
    reframe(favstats(MaxGrip~Monarch_Group))
#>   Monarch_Group   min    Q1 median     Q3   max      mean         sd  n missing
#> 1       Captive 0.081 0.162  0.217 0.2845 0.596 0.2363731 0.09412948 67       0
#> 2          Wild 0.108 0.271  0.352 0.4330 0.650 0.3607500 0.14066796 40       0
```

```
butterflies%>%
  ggplot(aes(y = MaxGrip, x = Monarch_Group))  + #Enter the name of the explanatory and response variable
  geom_boxplot()+
  labs(title = "Boxplot of max grip of Monarch butterflies by environment reared", #Title your plot
       y = "Max Grip (Newtons)", #y-axis label
       x = "Environment Reared") #x-axis label
```

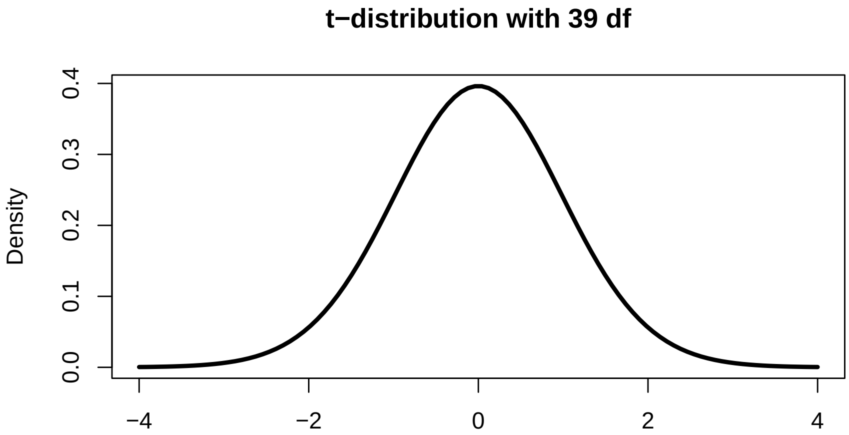


Are the conditions met to analyze the data using theory based-methods?

Calculate the standardized difference in max grip strength.

- 1st calculate the $SE(\bar{x}_1 - \bar{x}_2)$

- Then calculate the T-score



Interpret the standardized statistic:

To find the theory-based p-value:

```
pt(-5, df=39, lower.tail=TRUE)*2
#> [1] 1.252417e-05
```

Confidence interval

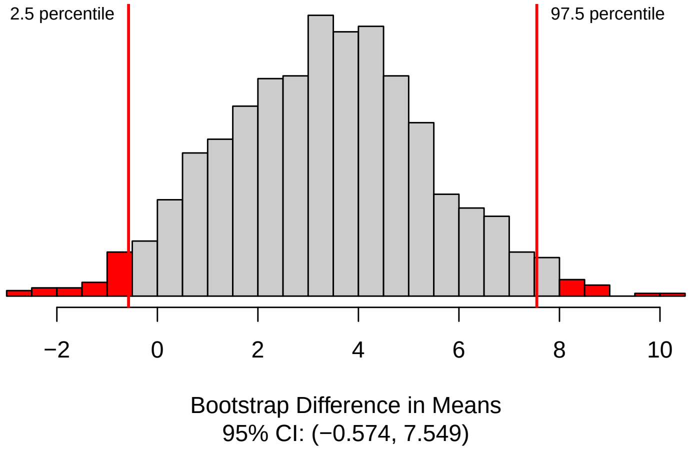
To estimate the difference in true mean we will create a confidence interval.

Simulation-based method

- Write the response variable values on cards
- Keep explanatory variable groups separate
- Sample with replacement n_1 times in explanatory variable group 1 and n_2 times in explanatory variable group 2
- Calculate and plot the simulated difference in sample means from each simulation
- Repeat 1000 times (simulations) to create the bootstrap distribution
- Find the cut-offs for the middle X% (confidence level) in a bootstrap distribution.

Returning to the letters example, we will estimate the difference in true mean number of letters recognized for students given recognizable letter groupings and students given non-recognizable letter groupings.

```
set.seed(216)
two_mean_bootstrap_CI(Memorized ~ Grouped, #Enter the name of the variables
                      data = letters,  # Enter the name of the data set
                      first_in_subtraction = "Recognizable", # First value in order of subtraction
                      number_repetitions = 1000,  # Number of simulations
                      confidence_level = 0.95)
```



Confidence interval interpretation:

- How confident you are (e.g., 90%, 95%, 98%, 99%)
- Parameter of interest
- Calculated interval
- Order of subtraction when comparing two groups

Theory-based method

- Calculate the interval centered at the sample statistic

statistic ± margin of error

Using the Monarch butterfly data, calculate the 99% confidence interval.

```
butterflies %>%
    reframe(favstats(MaxGrip~Monarch_Group))
#>   Monarch_Group   min    Q1 median     Q3   max      mean         sd  n missing
#> 1       Captive 0.081 0.162  0.217 0.2845 0.596 0.2363731 0.09412948 67       0
#> 2          Wild 0.108 0.271  0.352 0.4330 0.650 0.3607500 0.14066796 40       0
```

- Need the t^* multiplier for a 99% confidence interval from a t-distribution with _____________ df.

```
qt(0.995, df=39, lower.tail = TRUE)
#> [1] 2.707913
```

- We will use the same value for the $SE(\bar{x}_1 - \bar{x}_2)$ as calculated for the standardized statistic.

Based on the p-value and confidence interval, write a conclusion to the test.

4.3 Out of Class Activity Week 12: Does behavior impact performance?

4.3.1 Learning outcomes

- 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.

4.3.2 Terminology review

In today's 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

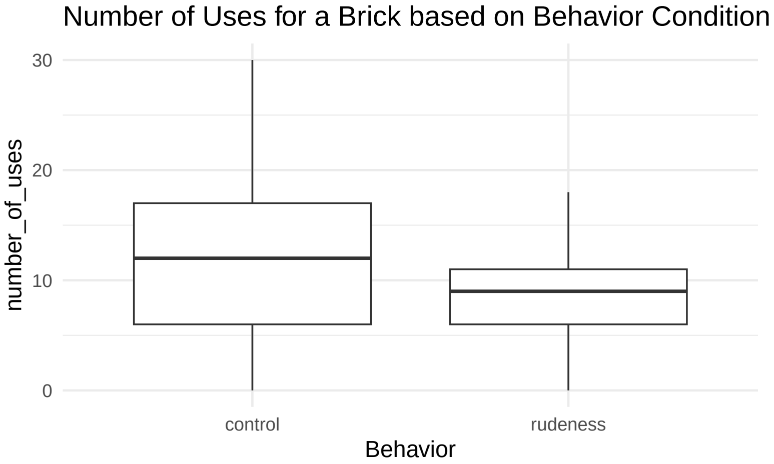
To review these concepts, see Chapter 19 in the textbook.

4.3.3 Behavior and Performance

A study in the Academy of Management Journal (Porath 2017) investigated how rude behaviors influence a victim's task performance. Randomly selected college students enrolled in a management course were randomly assigned to one of two experimental conditions: rudeness condition (45 students) and control group (53 students). Each student was asked to write down as many uses for a brick as possible in five minutes; this value (total number of uses) was used as a performance measure for each student, where higher values indicate better performance. During this time another individual showed up late for class. For those students in the rudeness condition, the facilitator displayed rudeness by berating the students in general for being irresponsible and unprofessional (due to the late-arriving person). No comments were made about the late-arriving person for students in the control group. Is there evidence that the average performance score for students in the rudeness condition is lower than for students in the control group? Use the order of subtraction of rudeness – control.

```r
# Read in data set
rude <- read.csv("https://math.montana.edu/courses/s216/data/rude.csv")
```

```
# Side-by-side box plots
rude %>%
ggplot(aes(x = condition, y = number_of_uses)) +
    geom_boxplot() +
    labs(title = "Number of Uses for a Brick based on Behavior Condition",
         x = "Behavior")
```



```
# Summary statistics
rude %>%
     reframe(favstats(number_of_uses ~ condition))
```

```
#>   condition min Q1 median Q3 max      mean       sd  n missing
#> 1   control   0  6     12 17  30 11.811321 7.382559 53       0
#> 2  rudeness   0  6      9 11  18  8.511111 3.992164 45       0
```

Quantitative variables review

1. The two variables assessed in this study are behavior and number of uses for a brick. Identify the role for each variable (explanatory or response).

2. Which group (control or rudeness) has the highest center in the distributions of number of uses for a brick? Explain which measure of center you are using.

3. Using the side-by-side box plots, which group has the largest spread in number of uses for a brick? 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.

7. Write out the null hypothesis in words.

8. Write the alternative hypothesis in notation.

Summarize and visualize the data

9. Calculate the summary statistic of interest (difference in means). 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 values observed in the data set would have been the same regardless of the behavior condition.

To demonstrate this simulation, we could create cards to simulate a sample.

10. How many cards would we start with?

11. What would we write on each card?

12. Next, we would 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 would be calculated from the cards and plotted on the null distribution? *Hint*: What statistic are we calculating from the data?

14. Would you expect your simulated statistic to be closer to the null value of zero than the difference in means calculated from the sample? Explain why this makes sense.

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, `number_of_uses`, and the name of the explanatory variable, `condition`, for the formula. The name of the data set as shown above is `rude`. 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? `"control"` or `"rudeness"`):

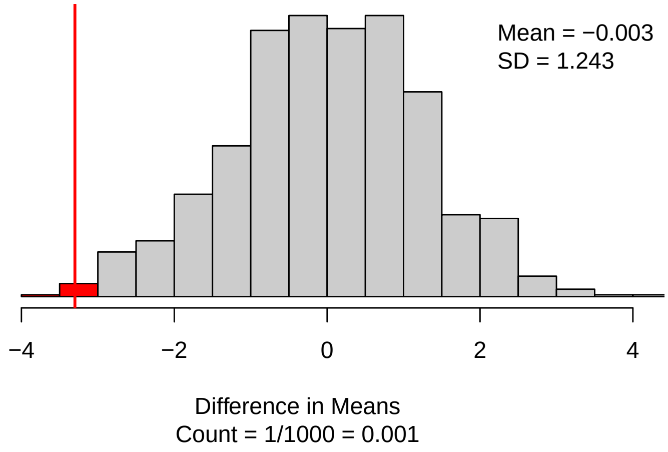
- Number of repetitions:

- As extreme as:

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

The code below will simulate a null distribution and compute the p-value. Check that your answers from question 15 match what is entered below

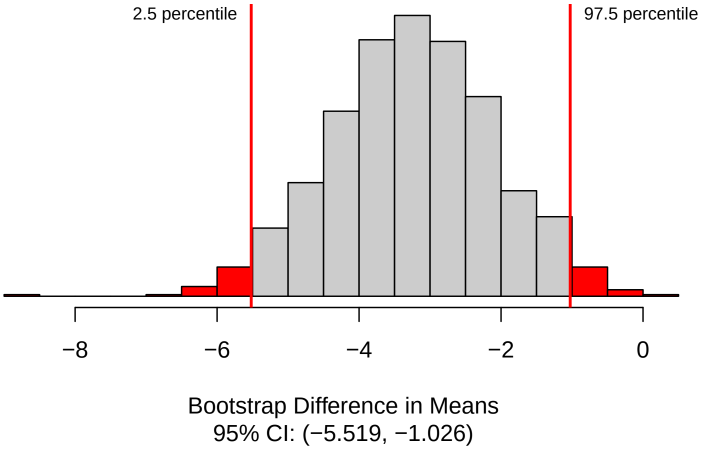
```
set.seed(216)
two_mean_test(number_of_uses~condition, #Enter the names of the variables
              data = rude,  # Enter the name of the dataset
              first_in_subtraction = "rudeness", # First outcome in order of subtraction
              number_repetitions = 1000,  # Number of simulations
              as_extreme_as = -3.3,  # Observed statistic
              direction = "less")  # Direction of alternative: "greater", "less", or "two-sided"
```



16. 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 95% confidence interval.

```
set.seed(216)
two_mean_bootstrap_CI(number_of_uses ~ condition, #Enter the name of the variables
                      data = rude,  # Enter the name of the data set
                      first_in_subtraction = "rudeness", # First value in order of subtraction
                      number_repetitions = 1000,  # Number of simulations
                      confidence_level = 0.95)
```



17. Report the 95% confidence interval.

18. Interpret the interval you calculated in question 17.

4.3.4 Take-home messages

1. This activity differs from the activities in Week 11 because the responses are independent, not paired. These data are analyzed as a difference in means, not a mean difference.

2. 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.

3. 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.

4.3.5 Additional notes

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

4.4 Week 12 Lab: The Triple Crown

4.4.1 Learning outcomes

- 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 theory-based hypothesis test for a difference in means.
- Interpret and evaluate a p-value for a theory-based hypothesis test for a difference in means.
- Use theory-based methods 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.

4.4.2 Terminology review

In today's activity, we will use theory-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:

- Difference in means
- Independence within and between groups
- Normality

To review these concepts, see Chapter 19 in the textbook.

4.4.3 The triple crown

The Triple Crown of "Thru" hiking consists of hiking the Appalachian Trail, the Pacific Crest Trail (PCT), and the Continental Divide Trail (CDT). Each year halfwayanywhere.com conducts a survey to better understand the people who hike these trails. One variable which is queried in the survey is the pre-hike "base weight" of a hiker's pack which is the total weight of gear without food, water, and worn gear. The 131 hikers surveyed who completed the CDT had a mean base weight of 15.266 lbs (sd = 5.128 lbs). The 484 hikers surveyed who completed the PCT had a mean base weight of 17.837 lbs (sd = 7.823 lbs). Is there a difference in average base weight for PCT hikers and CDT hikers? Use order of subtraction CDT - PCT.

1. **Write out the parameter of interest for this study.**

2. Write out the null hypothesis in notation for this study. Be sure to clearly identify the subscripts.

3. Write out the alternative hypothesis in words for this study.

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.
  - $n \geq 30$: 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.
  - $n \geq 100$: If the sample size n is at least 100 (regardless of the presence of skew or outliers), we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

Upload and open the R script file for Week 12 lab. Upload and import the csv file, `Trail_Weight`. Enter the name of the data set (see the environment tab) for datasetname in the R script file in line 7. Write a title for the boxplots in line 11. Highlight and run lines 1–13 to load the data and create plots of the data.

```
hikes <- datasetname
hikes %>%  # Data set piped into...
  ggplot(aes(y = Baseweight, x = Trail))+  # Identify variables
  geom_boxplot()+  # Tell it to make a box plot
  labs(title = "xx",  # Title
       x = "Trail",    # x-axis label
       y = "Baseweight(lbs)")  # y-axis label
```

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

5. Check that the normality condition is met to use theory-based methods to analyze these data.

Enter the name of the explanatory variable for `explanatory` and the name of the response variable for `response` in line 17. Highlight and run lines 16–17 to get the summary statistics for the data.

```
hikes %>%
  summarize(favstats(response~explanatory))
```

6. **Calculate the summary statistic (difference in means) for this study. Use appropriate notation with clearly defined subscripts.**

Use statistical inferential methods to draw inferences from the data

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:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

7. Calculate the standard error for the difference in sample means.

8. **Calculate the standardized statistic for the difference in sample means.**

9. When we are comparing two quantitative variables to find the degrees of freedom to use for the t-distribution, we need to use the group with the smallest sample size and subtract 1. (`df` = minimum of $n_1 - 1$ or $n_2 - 1$). Calculate the `df` for this study.

10. Using the provided R script file, enter the T-score (for `xx`) and the `df` calculated in question 9 for `yy` into the `pt()` function to find the p-value. Highlight and run line 20. Report the p-value calculated.

```
2*pt(xx, df=yy, lower.tail=FALSE)
```

11. **Explain why we multiplied by 2 in the code above.**

12. Do you expect the 95% confidence interval to contain the null value of zero? Explain your answer.

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 (`df` = minimum of $n_1 - 1$ or $n_2 - 1$).

Enter the appropriate percentile value (as a decimal) for `xx` and degrees of freedom for `yy` into the `qt()` function at line 23 to find the appropriate t^* multiplier

```
qt(xx, df = yy, lower.tail=FALSE)
```

13. Report the t^* multiplier for the 95% confidence interval.

14. Calculate the 95% confidence interval using theory-based methods.

15. Do the results of the CI agree with the p-value? Explain your answer.

16. What type of error may be possible?

17. Write a paragraph summarizing the results of the study as if you are reporting the results to your supervisor. **Upload a copy of your paragraph to Gradescope for your group.** Be sure to describe:

- Summary statistic and interpretation
- P-value and interpretation
  - Statement about probability or proportion of samples
  - Statistic (summary measure and value)
  - Direction of the alternative
  - Null hypothesis (in context)
- Confidence interval and interpretation
  - How confident you are (e.g., 90%, 95%, 98%, 99%)
  - Parameter of interest
  - Calculated interval
  - Order of subtraction when comparing two groups
- Conclusion (written to answer the research question)
  - Amount of evidence
  - Parameter of interest
  - Direction of the alternative hypothesis
- Scope of inference

Paragraph:

MODULE 5

Inference for Two Quantitative Variables

5.1 Module 13 Reading Guide: Inference for Slope and Correlation

Chapter 21 (Inference for regression and model conditions)

Videos

- 21.2
- 21.3
- 21.4to21.5Tests
- 21.4to21.5Intervals

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: $\hat{y} = b_0 + b_1 \times x$, 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,

How many cards will you need and how will the cards be labeled?

What do you do with the cards after labeling them?

After shuffling, what value will be plotted on the simulated null distribution?

To create a **bootstrap distribution** of sample slopes or sample correlations,

How many cards will you need and how will the cards be labeled?

What do you do with the cards after labeling them?

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 from sections 21.2 and 21.3: 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 from section 21.4: 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.

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 provided in table 21.2, 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.

5.2 Lecture Notes Week 13: Inference for two quantitative variable

5.2.1 Summary measures and plots for two quantitative variables.

Scatterplot:

- Form: linear or non-linear?
- Direction: positive or negative?
- Strength: how clear is the pattern between the two variables?
- Outliers: points that are far from the pattern or bulk of the data
  - Influential points: outliers that are extreme in the x- variable.

The summary measures for two quantitative variables are:

- ____, interpreted as

- ____, which measures the

- ____, interpreted as

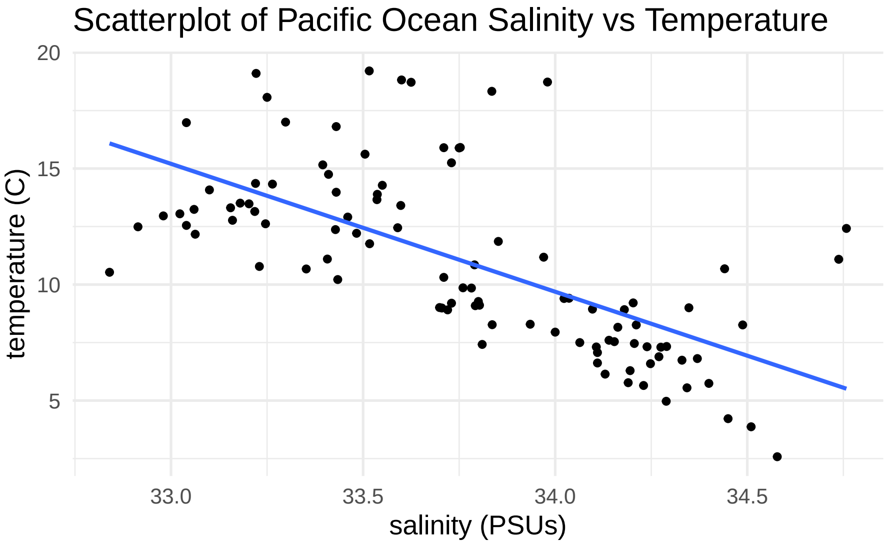
- Least-squares regression line: $\hat{y} = b_0 + b_1 \times x$ (put y and x in the context of the problem)

Notation:

- Population slope:
- Population correlation:
- Sample slope:
- Sample correlation:

Example: Oceanic temperature is important for sea life. The California Cooperative Oceanic Fisheries Investigations has measured several variables on the Pacific Ocean for more than 70 years hoping to better understand weather patterns and impacts on ocean life. For this example, we will look at the most recent 100 measurements of salt water salinity (measured in PSUs or practical salinity units) and the temperature of the ocean measured in degrees Celsius. Is there evidence that water temperature tends to decrease with higher levels of salinity.

```
water %>% # Pipe data set into...
ggplot(aes(x = Salnty, y = T_degC))+  # Specify variables
  geom_point() +  # Add scatterplot of points
  labs(x = "salinity (PSUs)",  # Label x-axis
       y = "temperature (C)",  # Label y-axis
       title = "Scatterplot of Pacific Ocean Salinity vs Temperature") +
               # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```



Describe the four characteristics of the scatterplot:

Linear model output:

```
lm.water <- lm(T_degC~Salnty, data=water) # lm(response~explanatory)
round(summary(lm.water)$coefficients, 3)
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   197.156     21.478    9.18        0
#> Salnty         -5.514      0.636   -8.67        0
```

Correlation:

```
cor(T_degC~Salnty, data=water)
#> [1] -0.6588365
```

Write the least squares equation of the line in context of the problem:

Interpret the value of slope in the context of the problem:

Report and describe the correlation value:

Calculate and interpret the coefficient of determination:

Hypothesis Testing

Conditions:

- Independence:

- Linear relationship:

Null hypothesis assumes “no effect”, “no difference”, “nothing interesting happening”, etc.

Always of form: “parameter” = null value

H_0 :

H_A :

- Research question determines the alternative hypothesis.

Write the null and alternative for the ocean study:

In words:

H_0 :

H_A :

In notation:

H_0 :

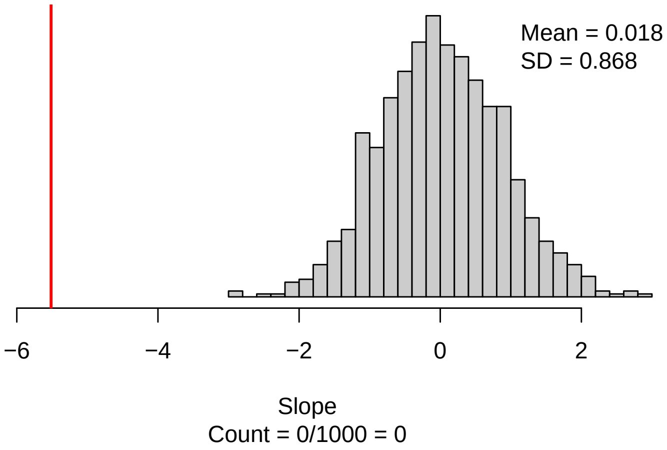
H_A :

Simulation-based method

- Simulate many samples assuming $H_0 : \beta_1 = 0$ or $H_0 : \rho = 0$
  - Write the response variable values on cards
  - Hold the explanatory variable values constant
  - Shuffle a new response variable to an explanatory variable
  - Plot the shuffled data points to find the least squares line of regression
  - Calculate and plot the simulated slope or correlation from each simulation
  - Repeat 1000 times (simulations) to create the null distribution
  - Find the proportion of simulations at least as extreme as b_1 or r

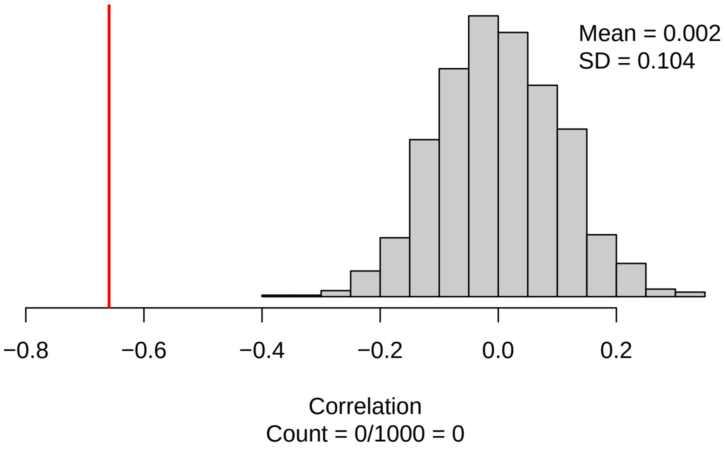
To test slope:

```r
set.seed(216)
regression_test(T_degC ~ Salnty, # response ~ explanatory
               data = water, # Name of data set
               direction = "less", # Sign in alternative ("greater", "less", "two-sided")
               summary_measure = "slope", # "slope" or "correlation"
               as_extreme_as = -5.514, # Observed slope or correlation
               number_repetitions = 1000) # Number of simulated samples for null distribution
```



To test correlation:

```
set.seed(216)
regression_test(T_degC~Salnty, # response ~ explanatory
                data = water, # Name of data set
                direction = "less", # Sign in alternative ("greater", "less", "two-sided")
                summary_measure = "correlation", # "slope" or "correlation"
                as_extreme_as = -0.659, # Observed slope or correlation
                number_repetitions = 1000) # Number of simulated samples for null distribution
```



Explain why the null distribution is centered at the value of zero:

Interpretation of the p-value:

- Statement about probability or proportion of samples
- Statistic (summary measure and value)
- Direction of the alternative
- Null hypothesis (in context)

Conclusion:

- Amount of evidence
- Parameter of interest
- Direction of the alternative hypothesis

Theory-based method

Conditions:

- Linearity (for both simulation-based and theory-based methods): the data should follow a linear trend.
    - Check this assumption by examining the ______________________________ of the two variables, and ____. The pattern in the residual plot should display a horizontal line.
- Independence (for both simulation-based and theory-based methods)
    - One________________________________for an observational unit has no impact on __________________________________.
- Constant variability (for theory-based methods only): the variability of points around the least squares line remains roughly constant
    - Check this assumption by examining the __________________________________. The variability in the residuals around zero should be approximately the same for all fitted values.

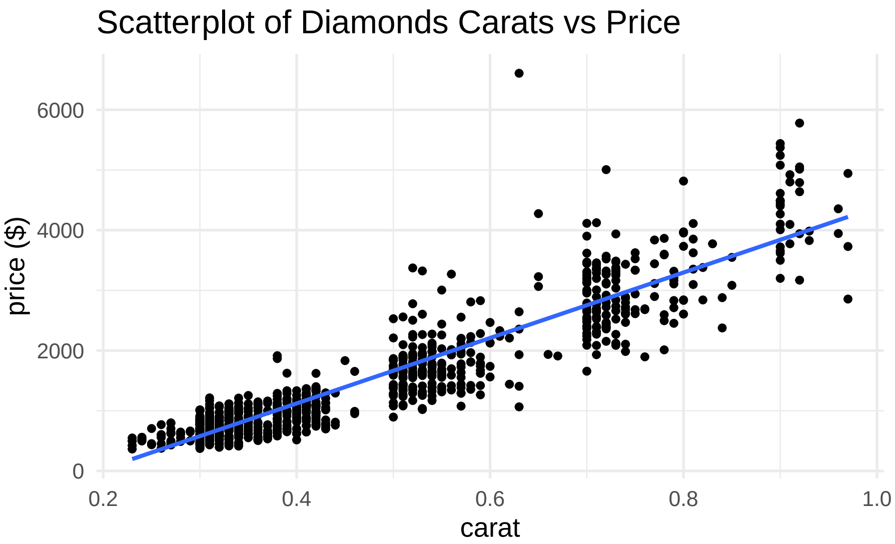
- Nearly normal residuals (for theory-based methods only): residuals must be nearly normal
  - Check this assumption by examining a ____, which should appear approximately normal

Example:

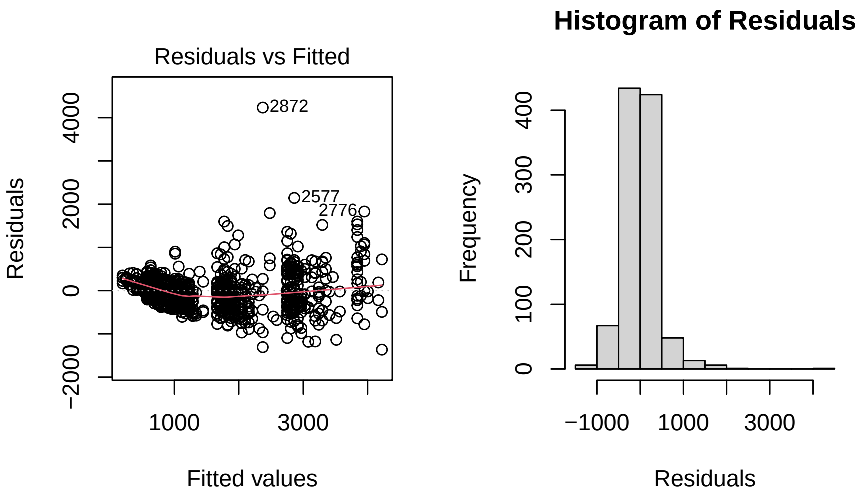
It is a generally accepted fact that the more carats a diamond has, the more expensive that diamond will be. The question is, how much more expensive? Data on thousands of diamonds were collected for this data set. We will only look at one type of cut (“Ideal”) and diamonds less than 1 carat. Does the association between carat size and price have a linear relationship for these types of diamonds? What can we state about the association between carat size and price?

Scatterplot:

```
Diamonds %>% # Pipe data set into...
    ggplot(aes(x = carat, y = price))+  # Specify variables
    geom_point() +  # Add scatterplot of points
    labs(x = "carat",  # Label x-axis
       y = "price ($)",  # Label y-axis
       title = "Scatterplot of Diamonds Carats vs Price") +
               # Be sure to title your plots
    geom_smooth(method = "lm", se = FALSE)  # Add regression line
```



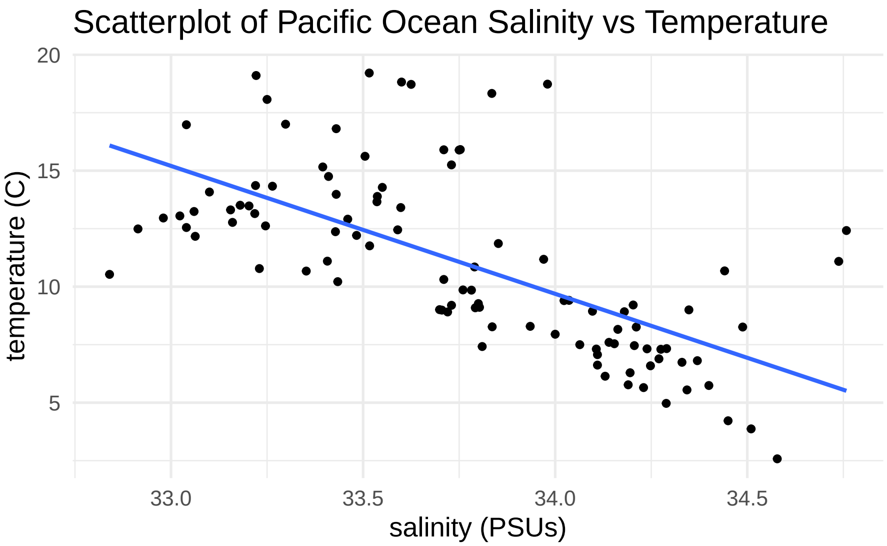
Diagnostic plots:



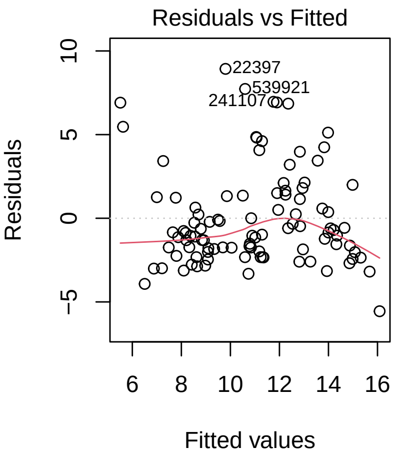
Check the conditions for the ocean data:

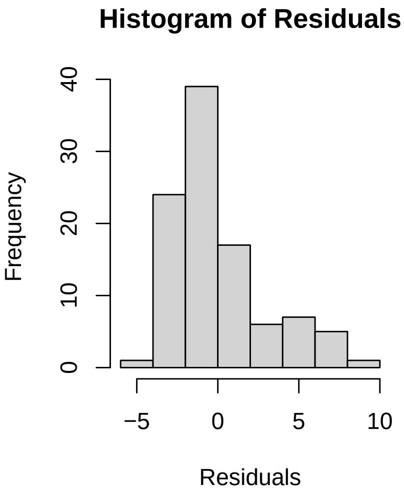
Scatterplot:

```r
water %>% # Pipe data set into...
ggplot(aes(x = Salnty, y = T_degC))+  # Specify variables
  geom_point() +  # Add scatterplot of points
  labs(x = "salinity (PSUs)",  # Label x-axis
       y = "temperature (C)",  # Label y-axis
       title = "Scatterplot of Pacific Ocean Salinity vs Temperature") +
            # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```



Diagnostic plots:





Like with paired data the t-distribution can be used to model slope and correlation.

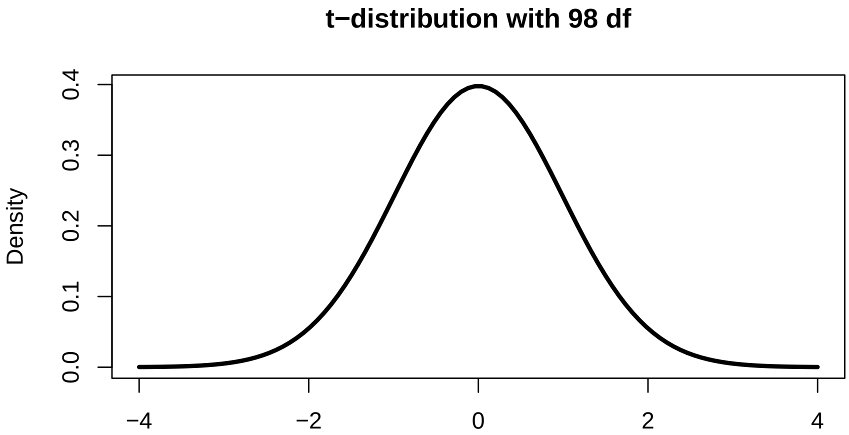
- For two quantitative variables we use the ________-distribution with ______________________________ degrees of freedom to approximate the sampling distribution.

Theory-based test:

- Calculate the standardized statistic
- Find the area under the t-distribution with $n - 2$ df at least as extreme as the standardized statistic

Standardized difference in sample mean:

Calculate the standardized slope for the ocean data



Interpret the standardized statistic:

To find the theory-based p-value:

```
lm.water <- lm(T_degC~Salnty, data=water) # lm(response~explanatory)
summary(lm.water)$coefficients
#>                Estimate Std. Error   t value     Pr(>|t|)
#> (Intercept) 197.156160 21.4778118  9.179527 7.304666e-15
#> Salnty       -5.513691  0.6359673 -8.669770 9.257446e-14
```

or

```
pt(-8.670, df = 98, lower.tail=TRUE)
#> [1] 4.623445e-14
```

Confidence interval

To estimate the true slope (or true correlation) we will create a confidence interval.

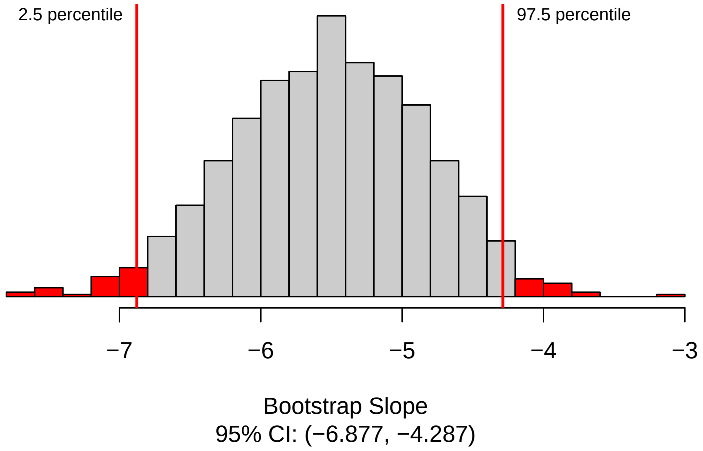
Simulation-based method

- Write the explanatory and response value pairs on cards
- Sample pairs with replacement n times
- Plot the resampled data points to find the least squares line of regression
- Calculate and plot the simulated slope (or correlation) from each simulation

- Repeat 1000 times (simulations) to create the bootstrap distribution
- Find the cut-offs for the middle X% (confidence level) in a bootstrap distribution.

Returning to the ocean example, we will estimate the true slope between salinity and temperature of the Pacific Ocean.

```
set.seed(216)
regression_bootstrap_CI(T_degC~Salnty, # response ~ explanatory
   data = water, # Name of data set
   confidence_level = 0.95, # Confidence level as decimal
   summary_measure = "slope", # Slope or correlation
   number_repetitions = 1000) # Number of simulated samples for bootstrap distribution
```

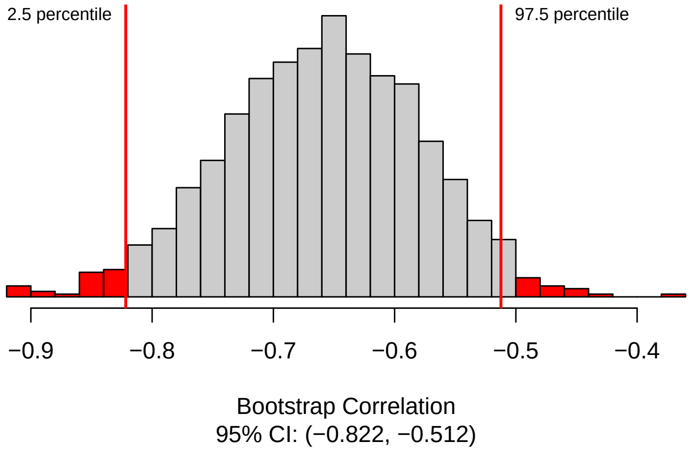


Confidence interval interpretation:

- How confident you are (e.g., 90%, 95%, 98%, 99%)
- Parameter of interest
- Calculated interval
- Order of subtraction when comparing two groups

Now we will estimate the true correlation between salinity and temperature of the Pacific Ocean.

```
set.seed(216)
regression_bootstrap_CI(T_degC~Salnty, # response ~ explanatory
   data = water, # Name of data set
   confidence_level = 0.95, # Confidence level as decimal
   summary_measure = "correlation", # Slope or correlation
   number_repetitions = 1000) # Number of simulated samples for bootstrap distribution
```



Confidence interval interpretation:

- How confident you are (e.g., 90%, 95%, 98%, 99%)
- Parameter of interest
- Calculated interval
- Order of subtraction when comparing two groups

Theory-based method

- Calculate the interval centered at the sample statistic

statistic ± margin of error

```
lm.water <- lm(T_degC~Salnty, data=water) # lm(response~explanatory)
round(summary(lm.water)$coefficients, 3)
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  197.156     21.478    9.18        0
#> Salnty        -5.514      0.636   -8.67        0
```

Using the ocean data, calculate the confidence interval for the true slope.

5.3 Out of Class Activity Week 13: Prediction of Crocodylian Body Size

5.3.1 Learning outcomes

- Given a research question involving two quantitative variables, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Describe and perform a simulation-based hypothesis test for slope or correlation.
- Interpret and evaluate a p-value for a simulation-based hypothesis test for a slope or correlation.
- Use bootstrapping to find a confidence interval for the slope or correlation.
- Interpret a confidence interval for a slope or correlation.

5.3.2 Terminology review

In today's activity, we will use simulation-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 Chapter 21 in the textbook.

5.3.3 Crocodylian Body Size

Much research surrounds using measurements of animals to estimate body-size of extinct animals. Many challenges exist in making accurate estimates for extinct crocodylians. The term crocodylians refers to all members of the family Crocodylidae ("true" crocodiles), family Alligatoridae (alligators and caimans) and family Gavialidae (gharial, Tomistoma). The researchers in this study (**obrien2019?**) state, "Among extinct crocodylians and their precursors (e.g., suchians), several methods have been developed to predict body size from suites of hard-tissue proxies. Nevertheless, many have limited applications due to the disparity of some major suchian groups and biases in the fossil record. Here, we test the utility of head width (HW) as a broadly applicable body-size estimator in living and fossil suchians." Is there evidence that head width is a good predictor of body size for crocodylians?

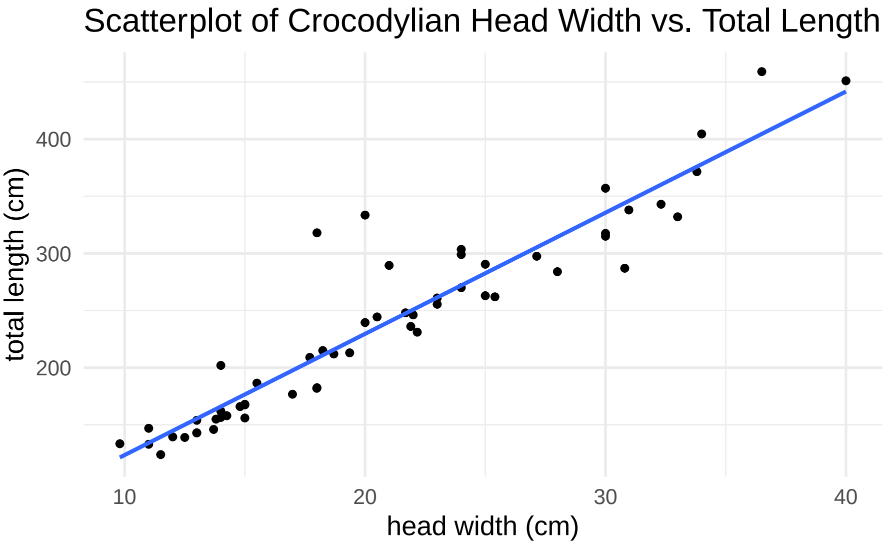
```
# Read in data set
croc <- read.csv("https://math.montana.edu/courses/s216/data/Crocodylian_headwidth.csv")
croc <- croc %>%
    na.omit()
```

Vocabulary review

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.

To create a scatterplot to examine the relationship between head width and total body length we will use `HW_cm` as the explanatory variable and `TL_cm` as the response variable.

```
croc %>% # Pipe data set into...
ggplot(aes(x = HW_cm, y = TL_cm))+  # Specify variables
  geom_point() +  # Add scatterplot of points
  labs(x = "head width (cm)",  # Label x-axis
       y = "total length (cm)",  # Label y-axis
       title = "Scatterplot of Crocodylian Head Width vs. Total Length") +
               # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```



2. 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?

Ask a research question

3. Write out the null hypothesis in words to test **slope**.

4. Using the research question, write the alternative hypothesis in notation using **slope** as the summary measure.

Summarize and visualize the data

The linear model output for the data is given below.

```
lm.croc <- lm(TL_cm~HW_cm, data=croc) # lm(response~explanatory)
round(summary(lm.croc)$coefficients, 5)
#>             Estimate Std. Error  t value Pr(>|t|)
#> (Intercept) 17.61250   11.36269  1.55003  0.12687
#> HW_cm       10.59983    0.51294 20.66494  0.00000
```

The value of correlation is given below.

```
cor(croc$HW_cm, croc$TL_cm)
#> [1] 0.9412234
```

5. Using the output from the evaluated R code above, write the equation of the regression line in the context of the problem using appropriate statistical notation.

6. Interpret the estimated slope in context of the problem.

7. Report the value of correlation between head width and total body length.

Use statistical inferential methods to draw inferences from the data

In this activity, we will focus on using simulation-based methods for inference in regression.

Simulation-based hypothesis test

Let's start by thinking about how one simulation would be created on the null distribution using cards. First, we would write the values for the response variable, total length, 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 (x, y) pairs and calculate either the slope or correlation of the shuffled sample.

We will use the `regression_test()` function in R (in the `catstats` package) to simulate the null distribution of shuffled slopes (or shuffled 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 `croc`), the summary measure 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 `TL_cm` and the explanatory variable name is `HW_cm` for these data.

8. What inputs should be entered for each of the following to create the simulation to test regression slope?

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

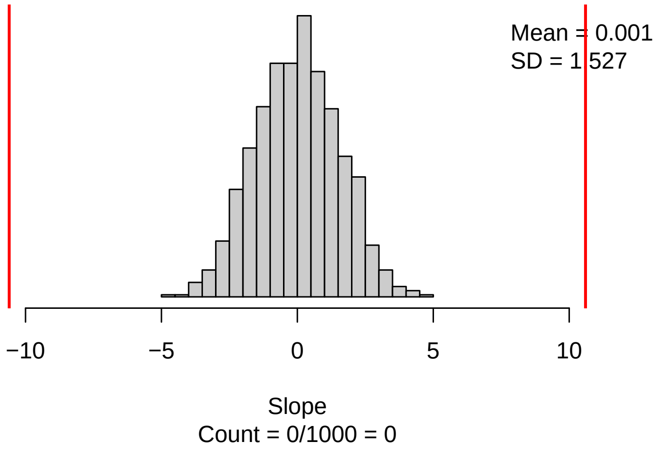
- Summary measure (choose `"slope"` or `"correlation"`):

- As extreme as (enter the value for the sample slope):

- Number of repetitions:

Check that your answers to question 8 reflect what is shown below in the R code to produce the null distribution for slope.

```
set.seed(216)
regression_test(TL_cm~HW_cm, # response ~ explanatory
                data = croc, # Name of data set
                direction = "two-sided", # Sign in alternative ("greater", "less", "two-sided")
                summary_measure = "slope", # "slope" or "correlation"
                as_extreme_as = 10.600, # Observed slope or correlation
                number_repetitions = 1000) # Number of simulated samples for null distribution
```

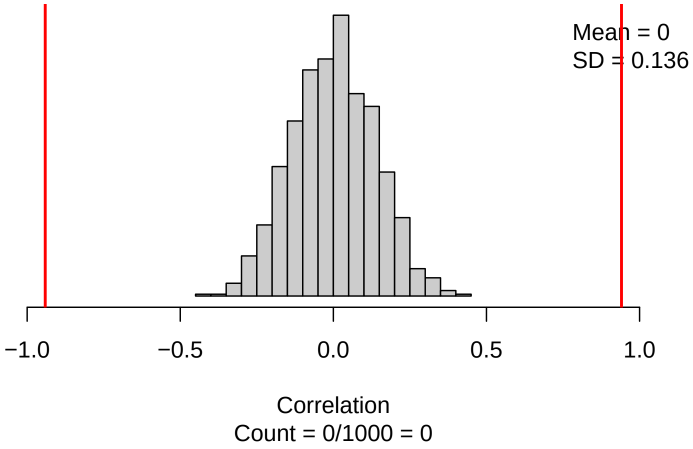


9. Report the p-value from the R output.

10. Suppose we wanted to complete the simulation test using correlation as the summary measure, instead of slope. Which two inputs in #8 would need to be changed to test for correlation? What inputs should you use instead?

Check that your answers to question 10 reflect what is shown below in the R code to produce the null distribution for correlation.

```
set.seed(216)
regression_test(TL_cm~HW_cm, # response ~ explanatory
               data = croc, # Name of data set
               direction = "two-sided", # Sign in alternative ("greater", "less", "two-sided")
               summary_measure = "correlation", # "slope" or "correlation"
               as_extreme_as = 0.941, # Observed slope or correlation
               number_repetitions = 1000) # Number of simulated samples for null distribution
```

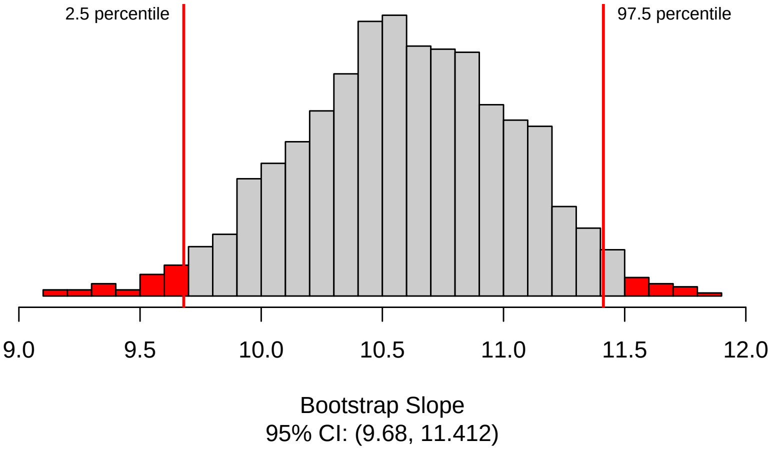


11. The p-values from the test of slope and the test of correlation should be similar. Explain why the two p-values should match. *Hint: think about the relationship between slope and correlation!*

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. The following code gives the 95% confidence interval for the true slope.

```
set.seed(216)
regression_bootstrap_CI(TL_cm~HW_cm, # response ~ explanatory
   data = croc, # Name of data set
   confidence_level = 0.95, # Confidence level as decimal
   summary_measure = "slope", # Slope or correlation
   number_repetitions = 1000) # Number of simulated samples for bootstrap distribution
```



12. Report the bootstrap 95% confidence interval in interval notation.

13. Interpret the interval in question 12 in context of the problem. *Hint: use the interpretation of slope in your confidence interval interpretation.*

Communicate the results and answer the research question

14. Based on the p-value, write a conclusion in context of the problem.

15. Does the conclusion based on the p-value agree with the results of the 95% confidence interval? What does each tell you about the null hypothesis?

5.3.4 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 interpret a confidence interval for the slope, think about how to interpret the sample slope and use that information in the confidence interval interpretation for slope.

3. To create one simulated sample on the null distribution when testing for a relationship between two quantitative variables, 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.

4. To create one simulated sample on the bootstrap distribution when assessing two quantitative variables, 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.

5.3.5 Additional notes

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

5.4 Activity 13: Golf Driving Distance

5.4.1 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.
- Find the T test statistic (T-score) for a slope based off of `lm()` output in R.
- Find, interpret, and evaluate the p-value for a theory-based hypothesis test for a slope.
- Create and interpret a theory-based confidence interval for a slope.
- Use a confidence interval to determine the conclusion of a hypothesis test.

5.4.2 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. Some terms covered in this activity are:

- Slope
- Regression line

To review these concepts, see Chapter 21 in the textbook.

5.4.3 Golf driving distance

In golf the goal is to complete a hole with as few strokes as possible. A long driving distance to start a hole can help minimize the strokes necessary to complete the hole, as long as that drive stays on the fairway. Data was collecting on 354 PGA and LPGA players in 2008 ("Average Driving Distance and Fairway Accuracy" 2008). For each player, the average driving distance (yards), fairway accuracy (percentage), and sex was measured. Use these data to assess, "Does a professional golfer give up accuracy when they hit the ball farther?"

```r
# Read in data set
golf <- read.csv("https://math.montana.edu/courses/s216/data/golf.csv")
```

Plot review.

Use the provided R script file to create a scatterplot to examine the relationship between the driving distance and percent accuracy by filling in the variable names (`Driving_Distance` and `Percent_Accuracy`) for `xx` and `yy` in line 9. Highlight and run lines 1–15.

```r
golf %>% # Pipe data set into...
ggplot(aes(x = xx, y = yy))+  # Specify variables
  geom_point() +  # Add scatterplot of points
  labs(x = "Driving Distance",  # Label x-axis
       y = "Percent Accuracy",  # Label y-axis
       title = "Scatterplot of Driving Distance by Percent Accuracy") +
               # Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```

1. Sketch the plot created below. Based on your plot, does it appear that there is a relationship between driving distance and percent accuracy? Note: `Driving Distance` should be on the x-axis.

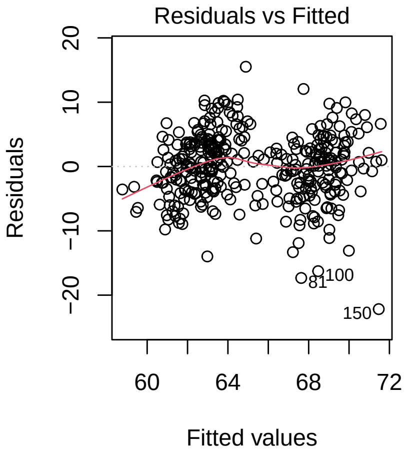
Conditions for the least squares line

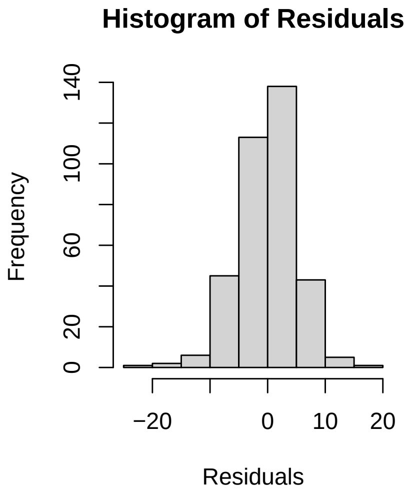
When performing inference on a least squares line, the follow conditions are generally required:

- *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].

[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

The scatterplot generated in question 1 and the residual plots shown below will be used to assess these conditions for approximating the data with the t-distribution.





2. Are the conditions met to use the t-distribution to approximate the sampling distribution of the standardized statistic? Justify your answer.

Ask a research question

3. Write out the null hypothesis in words to test the slope.

4. 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, `Percent_Accuracy`, into the `lm()` (linear model) function for `response` and the explanatory variable name, `Driving_Distance`, for `explanatory` in line 25 to get the linear model output. Highlight and run lines 25–26.

```r
lm.golf <- lm(response~explanatory, data=golf) # lm(response~explanatory)
round(summary(lm.golf)$coefficients, 5)
```

5. Using the output from the evaluated R code above, write the equation of the regression line in the context of the problem using appropriate statistical notation.

6. Interpret the estimated slope in context of the problem.

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.

7. What are the values of b_1 and $SE(b_1)$? Where in the linear model R output can you find these values?

8. Calculate the standardized statistic for slope. Identify where this calculated value is in the linear model R output.

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

10. The p-value in the linear model R output is the two-sided p-value for the test of significance for slope. Report the p-value to answer the research question.

11. 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 354 so we will use the degrees of freedom 352 ($n-2$).

```
qt(0.975, 352) # 95% t* multiplier
```

```
#> [1] 1.966726
```

12. Calculate the 95% confidence interval for the true slope.

13. Interpret the 95% confidence interval in context of the problem.

Communicate the results and answer the research question

14. Write a conclusion to answer the research question in context of the problem.

Multivariate plots

Another variable that may affect the percent accuracy is the sex of the golfer. We will look at how this variable may change the relationship between driving distance and percent accuracy. Highlight and run lines 32–39 to produce the multivariate plot.

```
golf %>%
  ggplot(aes(x = Driving_Distance, y = Percent_Accuracy, color=Sex))+  # Specify variables
  geom_point(aes(shape = Sex), size = 3) +  # Add scatterplot of points
  labs(x = "Driving Distance (m)",  # Label x-axis
       y = "Percent Accuracy",  # Label y-axis
       color = "Sex", shape = "Sex",
       # Be sure to title your plots
       title = "Scatterplot of Golf Driving Distance and Percent Accuracy by Sex") +
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```

15. Does the association between driving distance and percent accuracy change dependent on sex of the golfer? Explain your answer.

16. Explain the association between sex and each of the other two variables.

5.4.4 Take-home messages

1. 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.

2. To interpret a confidence interval for the slope, think about how to interpret the sample slope and use that information in the confidence interval interpretation for slope.

3. Use the explanatory variable row in the linear model R output to obtain the slope estimate (`estimate` column) and standard error of the slope (`Std. Error` column) to calculate the standardized slope, or T-score. The calculated T-score should match the `t value` column in the explanatory variable row. The standardized slope tells the number of standard errors the observed slope is above or below 0.

4. The explanatory variable row in the linear model R output provides a **two-sided** p-value under the `Pr(>|t|)` column.

5. The standardized slope is compared to a t-distribution with $n-2$ degrees of freedom in order to obtain a p-value. The t-distribution with $n-2$ degrees of freedom is also used to find the appropriate multiplier for a given confidence level.

5.4.5 Additional notes

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

5.5 Week 13 Lab: Big Mac Index

5.5.1 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 determine in theory or simulation-based methods should be used.
- Find, interpret, and evaluate the p-value for a hypothesis test for a slope or correlation.
- Create and interpret a confidence interval for a slope or correlation.

5.5.2 Big Mac Index

Can the relative cost of a Big Mac across different countries be used to predict the Gross Domestic Product (GDP) per person for that country? The GDP per person and the adjusted dollar equivalent to purchase a Big Mac was found on a random sample of 55 countries in January of 2022. The cost of a Big Mac in each country was adjusted to US dollars based on current exchange rates. Is there evidence of a positive relationship between Big Mac cost and the GDP per person?

Upload and open the R script file for Week 13 lab. Upload and import the csv file, `big_mac_adjusted_index_S22.csv`. Enter the name of the data set (see the environment tab) for datasetname in the R script file in line 7. Highlight and run lines 1–7 to load the data.

```
# Read in data set
mac <- datasetname
```

Summarize and visualize the data

To find the correlation between the variables, `GDP_dollar` and `dollar_price` highlight and run lines 10–13 in the R script file.

```
mac %>%
  select(c("GDP_dollar", "dollar_price")) %>%
  cor(use="pairwise.complete.obs") %>%
  round(3)
```

1. Report the value of correlation between the variables.

2. **Calculate the value of the coefficient of determination between `GDP_dollar` and `dollar_price`.**

3. Interpret the value of the coefficient of determination in context of the problem.

In the next part of the activity we will assess the linear model between Big Mac cost and GDP. Enter the variable `GDP_dollar` for `response` and the variable `dollar_price` for `explanatory` in line 17. Highlight and run lines 17–18 to get the linear model output.

```
# Fit linear model: y ~ x
bigmacLM <- lm(response~explanatory, data=mac)
summary(bigmacLM)$coefficients # Display coefficient summary
```

4. Give the value of the slope of the regression line. Interpret this value in context of the problem.

Conditions for the least squares line

Highlight and run lines 22–35 to produce the diagnostic plots needed to assess conditions to use theory-based methods. Use the scatterplot and the residual plots to assess the validity conditions for approximating the data with the *t*-distribution.

```
#Scatterplot
mac %>% # Pipe data set into...
  ggplot(aes(x = dollar_price, y = GDP_dollar))+  # Specify variables
  geom_point() +  # Add scatterplot of points
  labs(x = "Big Mac Cost",  # Label x-axis
       y = "GDP",  # Label y-axis
       title = "Scatterplot of Big Mac Cost vs. GDP per person") +  # Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line

#Diagnostic plots
bigmacLM <- lm(GDP_dollar~dollar_price, data = mac) # Fit linear regression model
par(mfrow=c(1,2)) # Set graphics parameters to plot 2 plots in 1 row
plot(bigmacLM, which=1) # Residual vs fitted values
hist(bigmacLM$resid, xlab="Residuals", ylab="Frequency",
     main = "Histogram of Residuals") # Histogram of residuals
```

5. **Are the conditions met to use the *t*-distribution to approximate the sampling distribution of the standardized statistic? Justify your answer.**

Ask a research question

6. Write out the null and alternative hypotheses in notation to test *correlation* between Big Mac cost and country GDP.

H_0 :

H_a :

Use statistical inferential methods to draw inferences from the data

Hypothesis test

Use the `regression_test()` function in R (in the `catstats` package) to simulate the null distribution of 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 `mac`), the summary measure used for the test, number of repetitions, the sample statistic (value of correlation), and the direction of the alternative hypothesis.

The response variable name is `GDP_dollar` and the explanatory variable name is `dollar_price`.

7. What inputs should be entered for each of the following to create the simulation to test correlation?

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

- Summary measure (choose `"slope"` or `"correlation"`):

- As extreme as (enter the value for the sample correlation):

- Number of repetitions:

Using the R script file for this activity, enter your answers for question 7 in place of the `xx`'s to produce the null distribution with 1000 simulations. Highlight and run lines 38–43. **Upload a copy of your plot showing the p-value to Gradescope for your group.**

```r
regression_test(GDP_dollar~dollar_price, # response ~ explanatory
                data = mac, # Name of data set
                direction = "xx", # Sign in alternative ("greater", "less", "two-sided")
                summary_measure  = "xx", # "slope" or "correlation"
                as_extreme_as = xx, # Observed slope or correlation
                number_repetitions = 1000) # Number of simulated samples for null distribution
```

8. Report the p-value from the R output.

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

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 **correlations** and calculate a confidence interval. Fill in the `xx`'s in the the provided R script file to find a 90% confidence interval. Highlight and run lines 46–50.

```r
regression_bootstrap_CI(GDP_dollar~dollar_price, # response ~ explanatory
   data = mac, # Name of data set
   confidence_level = xx, # Confidence level as decimal
   summary_measure = "xx", # Slope or correlation
   number_repetitions = 1000) # Number of simulated samples for bootstrap distribution
```

10. Report the bootstrap 90% confidence interval in interval notation.

11. Interpret the 90% confidence interval in context of the problem.

Communicate the results and answer the research question

12. Based on the p-value, write a conclusion in context of the problem.

13. Using a significance level of 0.1, what decision would you make?

14. What type of error is possible?

15. Interpret this error in context of the problem.

16. Write a paragraph summarizing the results of the study as if you are reporting these results in your local newspaper. **Upload a copy of your paragraph to Gradescope for your group.** Be sure to describe:

- Summary statistic
- P-value and interpretation
- Confidence interval and interpretation
- Conclusion (written to answer the research question)
- Scope of inference

MODULE 6

Probability and Relative Risk

6.1 Module 14 Reading Guide: Special Topics

Chapter 23 (Probability with tables)

Videos

- Chapter23

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 from section 23.4: 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. $1/10000 = 0.0001 = P(\qquad\qquad\qquad)$
   b. $100\% = 1.0 = P(\qquad\qquad\qquad)$
   c. $99.98\% = 0.9998 = P(\qquad\qquad\qquad)$

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.

Section 15.1.4 revisited (Simulation-based inference for a relative risk)

Vocabulary

Relative risk:

Notes

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.

Formulas

Relative risk =

Example: Malaria Vaccine

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 relative risk? Interpret the value in the context of the study.
6. What is the parameter (using relative risk) 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.

6.2 Lecture Notes Week 14: Probability and Relative Risk

6.2.1 Probability

- Event: something that could occur, something we want to find the probability of
  - Getting a four when rolling a fair die
- Complement: opposite of the event
  - Getting any value but a four when rolling a fair die
- The probability of an event is the ____________________ proportion of times the event would occur if the ______________________ process were repeated indefinitely.
  - For example, the probability of getting a four when rolling a fair die is _______________.
- Unconditional probabilities
  - An ____________________________probability is calculated from the entire population not____ on the occurrence of another event.
  - Examples:
    - ∗ The probability of a single event
      - · The probability a selected Stat 216 student is a computer science major.
    - ∗ An "And" probability
      - · The probability a selected Stat 216 student is a computer science major and a freshman.
- Conditional probabilities
  - A ______________________________ probability is calculated _________________________________ on the occurrence of another event.
  - Examples:
    - ∗ The probability of event A given B
      - · The probability a selected freshman Stat 216 student is a computer science major.
    - ∗ The probability of event B given A
      - · The probability a selected computer science Stat 216 student is a freshman

Finding probabilities from a table

	A	A^c	Total
B	A and B	A^c and B	Total B
B^c	A and B^c	A^c and B^c	Total B^c
Total	Total A	Total A^c	TOTAL

Calculating unconditional probabilities:

$P(A) =$

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

Calculating conditional probabilities:

$P(A|B) =$

$P(B|A) =$

Example: A random sample of people who had ever been married, demonstrating the proportions who smoked and who had ever been divorced. The numbers are shown in the following table. Because this survey was based on a random sample in the United States in the early 1990s, the data should be representative of the adult population who had ever been married at that time.

- Let event D be a person has gone through a divorce
- Let event S be a person smokes

	Has divorced	Has never divorced	Total
Smokes	238	247	Total 485
Does not smoke	374	810	Total 1184
Total	612	Total 1057	1669

- What is the approximate probability that the person smoked?

- What is the approximate probability that the person had ever been divorced?

- Given that the person had been divorced, what is the probability that he or she smoked?

- Given that the person smoked, what is the probability that he or she had been divorced?

Calculate and interpret each of the following:

- $P(S^c) =$

- $P(D^c|S^c) =$

Creating a hypothetical two-way table

Steps:

- Start with a large number like 100000.
- Then use the unconditional probabilities to fill in the row or column totals.
- Now use the conditional probabilities to begin filling in the interior cells.
- Use subtraction to find the remaining interior cells.
- Add the column values together for each row to find the row totals.
- Add the row values together for each column to find the column totals.

Example: An airline has noticed that 30% of passengers pre-pay for checked bags at the time the ticket is purchased. The no-show rate among customers that pre-pay for checked bags is 5%, compared to 15% among customers that do not pre-pay for checked bags.

- Let event B = customer pre-pays for checked bag
- Let event N = customer no shows

Start by identifying the probability notation for each value given.

- 0.30 =
- 0.05 =
- 0.15 =

	B	B^c	Total
N			
N^c			
Total			100,000

- What is the probability that a randomly selected customer who shows for the flight, pre-purchased checked bags?

Diagnostic tests

- Sensitivity:
- Specificity:
- Prevalence:

6.2.2 Relative Risk

- Relative risk is the ratio of the risks in two different categories of an explanatory variable.

Relative Risk:

- Interpretation:
  - The proportion of ____________________________ in group 1 is the RR ________________________ the proportion of ______________________________ in group 2.

Increase in risk:

- Interpretation:
  - The proportion of ___________________________ in group 1 is the (RR-1) ____________________ higher/lower than the proportion of ___________________________ in group 2.

Percent increase in risk:

- Interpretation:
  - The proportion of ________________________ in group 1 is the (RR-1)*100% higher/lower than the proportion of _________________________ in group 2.

Example: One-hundred fifty (150) children who had shown sensitivity to peanuts were randomized to receive a flour containing a peanut protein or a placebo flour for 2.5 years. At age 5 years, children were tested with a standard skin prick to see if they had an allergic reaction to peanut protein (yes or no). 71% of those in the peanut flour group no longer demonstrated a peanut allergy compared to 2% of those in the placebo group.

- Calculate the relative risk of desensitization comparing the peanut flour group to the placebo group.

- Interpret the value of relative risk in context of the problem.

- Find the increase (or decrease) in risk of desensitization and interpret this value in context of the problem.

- Find the percent increase (or decrease) in risk of desensitization and interpret this value in context of the problem.

Within the peanut flour group, the percent desensitized within each age group (at start of study) is as follows:

1-year-olds: 71%; 2-year-olds: 35%; 3-year-olds: 19%

- Calculate the relative risk of desensitization comparing the 3 year olds to the 2 year olds within the peanut flour group.

- Interpret the percent increase (or decrease) in risk of desensitization comparing the 3 year olds to the 2 year olds within the peanut flour group.

Relative risk in the news

People 50 and older who have had a mild case of covid-19 are 15% more likely to develop shingles (herpes zoster) within six months than are those who have not been infected by the coronavirus, according to research published in the journal Open Forum Infectious Diseases.

- What was the calculated relative risk of developing shingles when comparing those who has mild COVID-19 to those who had not had COVID-19, among the 50 and older population?

Testing Relative Risk

In Unit 2, we tested for a difference in proportion. We could also test for relative risk.

Null Hypothesis:

H_0 :

Alternative Hypothesis:

H_A :

6.3 Out of Class Activity Week 14: Titanic Survivors — Relative Risk

6.3.1 Learning outcomes

- Interpret the value of relative risk in terms of a percent increase or decrease.
- Evaluate the association between two categorical variables using relative risk.

6.3.2 Terminology review

In today's activity, we will look another summary. Some terms covered in this activity are:

- Conditional proportion
- Relative risk

To review these concepts, see Chapter 15 in your textbook.

6.3.3 Titanic Survivors

A complete data set exists listing all those aboard HMS Titanic and includes related facts about each person including age, how much they paid for their ticket, which boat they survived in (if they survived), and their job if they were crew members. Stories, biographies and pictures can be found on the site: www.encyclopedia-titanica.org/. Did all passengers aboard the Titanic have the same chance of survival? Was the risk of death higher among 3rd class passengers compared to 1st class passengers?

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

```r
# Read data set in
survive <- read.csv("https://math.montana.edu/courses/s216/data/Titanic.csv")
survive <- survive %>%
  filter(Class_Dept == "1st Class Passenger" | Class_Dept == "3rd Class Passenger")
survive %>% group_by(Class_Dept) %>% count(Survived)
```

```
#> # A tibble: 4 x 3
#> # Groups:   Class_Dept [2]
#>   Class_Dept          Survived     n
#>   <chr>               <chr>    <int>
#> 1 1st Class Passenger Alive      166
#> 2 1st Class Passenger Dead       108
#> 3 3rd Class Passenger Alive      147
#> 4 3rd Class Passenger Dead       509
```

Data Exploration

1. Fill in the data from the R output to complete the two-way table.

	Class		
Outcome	1st Class Passenger	3rd Class Passenger	Total
Dead			
Alive			
Total			

2. Calculate the conditional proportion of 1st class passengers that died.

3. Calculate the conditional proportion of 3rd class passengers that died.

4. Calculate the difference in conditional proportions of death for 3rd and 1st class passengers. Use 3rd − 1st as the order of subtraction.

5. Interpret the difference in proportions in context of the problem.

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:

$$\text{relative risk} = \frac{\hat{p}_1}{\hat{p}_2}.$$

6. Calculate the relative risk of death for 3rd class passengers compared to 1st class passengers.

7. Interpret the value of relative risk in context of the problem.

8. Calculate the percent increase or percent decrease in death.

9. Interpret the value of relative risk as a percent increase or percent decrease in death.

10. Based on the summary statistic, was the risk of death higher among 3rd class passengers compared to 1st class passengers? By what percent?

6.3.4 Risk in the News

11. Find a recent news article discussing ‘risk’. Summarize the article below by answering the following questions.

- What is the article discussing the risk of? (This is the a *success* for the study.)

- What two groups are being compared? (These are the two levels of the *explanatory* variable.)

- What is the percent increase/decrease in risk reported? What is the relative risk comparing the two groups?

- Does the news report appear to indicate that the reported difference in the groups is statistically significant? Do you agree with the report? If so, explain why. If not, what further information would you need to assess statistical significance?

- Does the news report appear to indicate a causal relationship exists based on the reported relative risk? Do you agree with the report? Justify your answer.

6.3.5 Take-home messages

1. Relative risk calculates the ratio of the proportion of successes in group 1 compared to the proportion of successes in group 2.
2. Relative risk evaluates the percent increase or percent decrease in the response variable attributed to the explanatory variable. To find the percent increase or percent decrease we calculate the following percent change $= (RR - 1) \times 100\%$. If relative risk is less than 1 there is a percent decrease. If relative risk is greater than 1 there is a percent increase.

6.3.6 Additional notes

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

6.4 Activity 14: What's the probability?

6.4.1 Learning outcomes

- Recognize and simulate probabilities as long-run frequencies.
- Construct two-way tables to evaluate conditional probabilities.

6.4.2 Terminology review

In today's activity, we will cover two-way tables and probability. Some terms covered in this activity are:

- Proportions
- Probability
- Conditional probability
- Two-way tables

To review these concepts, see Chapter 23 in the textbook.

6.4.3 Probability

1. In a large general education class, 60% of students are science majors and 40% are liberal arts majors. Twenty percent of the science majors are seniors, while 30% of the liberal arts majors are seniors. Given the following two-way table answer the following questions.

	Senior	Not a Senior	Total
Science	12,000	48,000	60,000
Liberal Arts	12,000	28,000	40,000
Total	24,000	76,000	100,000

a. What is the probability that a randomly selected senior is a science major? Use appropriate probability notation.

b. What is the probability that a randomly selected student is both a senior and a science major. Use appropriate probability notation.

c. What is the probability that a randomly selected student is not a senior given they are a liberal arts major. Use appropriate probability notation.

2. 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. (Stewart et al. 2014)

   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.

	A	A^c	Total
B			
B^c			
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?

3. 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.

	L	L^c	Total
S			
S^c			
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.

6.4.4 Take home messages

1. 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.
2. When creating a two-way table we typically want to put the explanatory variable on the columns of the table and the response variable on the rows.
3. To fill in the two-way table, always start with the unconditional variable in the total row or column and then use the conditional probabilities to fill in the interior cells.

6.4.5 Additional notes

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

6.5 Week 14 Lab: Efficacy of the COVID Vaccination

6.5.1 Learning outcomes

- Recognize and simulate probabilities as long-run frequencies.
- Use two-way tables to calculate conditional probabilities.
- Interpret the value of relative risk in terms of a percent increase or decrease.
- Evaluate the association between two categorical variables using relative risk.

6.5.2 Efficacy of the COVID vaccination

In November 2021, it was estimated that 59.1% of all US adults ($\geq$ 18 years old) were fully vaccinated against COVID-19 ("US COVID-19 Vaccine Tracker: See Your State's Progress" 2021). While vaccination is not 100% effective at protection against COVID-19, there are also other benefits to the vaccine. What impact does vaccination have on hospitalization rates for COVID? The following hypothetical two-way table was created based on CDC data on adult hospitalizations for COVID in the US ("Rates of Laboratory-Confimed COVID-19 Hospitalizations by Vaccination Status" 2021) in the same time period.

Let A = the event the US adult is vaccinated, and B = the event the US adult is hospitalized with COVID.

	Vaccinated	Not Vaccinated	Total
Hospitalized with COVID	2.3049	27.7302	30.0351
Not hospitalized with COVID	59,097.6951	40,872.2698	99,969.9649
Total	59,100	40,900	100,000

1. What is the probability that a US adult is both hospitalized with COVID-19 and vaccinated? Use proper probability notation.

2. **What is the probability that a US adult hospitalized with a COVID infection is vaccinated? Use proper probability notation.**

3. What is the probability that a US adult is hospitalized with a COVID infection in November 2021? Use proper probability notation.

4. **Give the probability notation for the calculation** $\frac{27.7302}{30.0392} = 0.923$**. Write out what this probability measures in words.**

5. What is the probability that a vaccinated US adult is hospitalized with COVID?

6. What is the probability that a un-vaccinated US adult is hospitalized with COVID?

7. **Calculate the relative risk for hospitalization with COVID in November 2021 for US adults fully vaccinated compared to US adults not vaccinated.**

8. Calculate the percent increase (or decrease) in hospitalization rate for US adults fully vaccinated compared to US adults not vaccinated.

9. **Interpret the relative risk as a percent increase/decrease in context of the problem.**

10. **Does it appear that there is an association with the risk of hospitalization due to COVID-19 and vaccination status? Explain.**

11. **Explain why a hypothesis test would not be appropriate in this case.**

“Average Driving Distance and Fairway Accuracy.” 2008. https://www.pga.com/ and https://www.lpga.com/.

Bulmer, M. n.d. “Islands in Schools Project.” https://sites.google.com/site/islandsinschoolsprojectwebsite/home.

Darley, J. M., and C. D. Batson. 1973. “”From Jerusalem to Jericho”: A Study of Situational and Dispositional Variables in Helping Behavior.” *Journal of Personality and Social Psychology* 27: 100–108.

Education Statistics, National Center for. 2018. “IPEDS.” https://nces.ed.gov/ipeds/.

Group, TODAY Study. 2012. “A Clinical Trial to Maintain Glycemic Control in Youth with Type 2 Diabetes.” *New England Journal of Medicine* 366: 2247–56.

Hamblin, J. K., K. Wynn, and P. Bloom. 2007. “Social Evaluation by Preverbal Infants.” *Nature* 450 (6288): 557–59.

Hirschfelder, A., and P. F. Molin. 2018. “I Is for Ignoble: Stereotyping Native Americans.” Retrieved from https://www.ferris.edu/HTMLS/news/jimcrow/native/homepage.htm.

Hutchison, R. L., and M. A. Hirthler. 2013. “Upper Extremity Injuies in Homer’s Iliad.” *Journal of Hand Surgery (American Volume)* 38: 1790–93.

“IMDb Movies Extensive Dataset.” 2016. https://kaggle.com/stefanoleone992/imdb-extensive-dataset.

Keating, D., N. Ahmed, F. Nirappil, Stanley-Becker I., and L. Bernstein. 2021. “Coronavirus Infections Dropping Where People Are Vaccinated, Rising Where They Are Not, Post Analysis Finds.” *Washington Post*. https://www.washingtonpost.com/health/2021/06/14/covid-cases-vaccination-rates/.

Moquin, W., and C. Van Doren. 1973. “Great Documents in American Indian History.” Praeger.

National Weather Service Corporate Image Web Team. n.d. “National Weather Service – NWS Billings.” https://w2.weather.gov/climate/xmacis.php?wfo=byz.

Porath, Erez, C. 2017. “Does Rudeness Really Matter? The Effects of Rudeness on Task Performance and Helpfulness.” *Academy of Management Journal* 50.

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.

“Rates of Laboratory-Confimed COVID-19 Hospitalizations by Vaccination Status.” 2021. CDC. https://covid.cdc.gov/covid-data-tracker/#covidnet-hospitalizations-vaccination.

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.

Stephens, R., and O. Robertson. 2020. “Swearing as a Response to Pain: Assessing Hypoalgesic Effects of Novel ”Swear” Words.” *Frontiers in Psychology* 11: 643–62.

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.

Stroop, J. R. 1935. “Studies of Interference in Serial Verbal Reactions.” *Journal of Experimental Psychology* 18: 643–62.

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.

“Titanic.” n.d. http://www.encyclopedia-titanica.org.

“US COVID-19 Vaccine Tracker: See Your State’s Progress.” 2021. Mayo Clinic. https://www.mayoclinic.org/coronavirus-covid-19/vaccine-tracker.

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.

“Welcome to the Navajo Nation Government: Official Site of the Navajo Nation.” 2011.Retrieved from https://www.navajo-nsn.gov/.