STAT 216 Coursepack

Spring 2025
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: Hancock, S., Carnegie, N., 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: Montana State Introductory Statistics with R, 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/. 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, video notes are provided to aid in taking notes while you complete the required 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!

All 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 the provided videos, taking notes in the coursepack.
  - 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 guided activity, 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.
  - Complete any remaining assigned readings for the week.
  - Complete the week's homework assignment.

MODULE 1

Inference for Two Categorical Variables: Theory-based Hypothesis Testing and Simulation-based and Theory-based Confidence Intervals

1.1 Vocabulary Review and Key Topics

Review the Golden Ticket posted in the resources at the end of the coursepack for a summary of two categorical variables.

1.1.1 Key topics

Module 9 introduces theory-based hypothesis testing methods and both simulation-based and theory-based confidence intervals for two categorical variables.

1.1.2 Vocabulary

Theory-based inference

- **Conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to follow an approximate normal distribution**: The following conditions must be met in order to use theory-based methods for two categorical variables.
  - **Independence**: The sample's observations are independent both within and between the two groups. (*Remember*: This also must be true to use simulation-based methods!)
  - **Success-failure condition**: We *expect* to see at least 10 successes and 10 failures in the *each* sample. We consider this condition to be met if we observe at least 10 successes and 10 failures in our data set in both groups: $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$, $n_2\hat{p}_2 \geq 10$, and $n_2(1-\hat{p}_2) \geq 10$. Equivalently, we check that all four cells in the table have at least 10 observations.
- **Standard error of a difference in sample proportions assuming the null is true**:

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

  where $\hat{p}_{pooled}$ is the **pooled sample proportion**: the total number of successes divided by the total sample size $(n_1 + n_2)$.
- **Standardized difference in sample proportion**:

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{SE_0(\hat{p}_1 - \hat{p}_2)}$$

  - Measures the number of standard errors the sample difference in proportions is above or below the null value of zero
  - If the conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to follow an approximate normal distribution are met, and if the true difference in proportions is equal to zero, the standardized difference in sample proportions, Z, will have an approximate *standard* normal distribution.

  - Use the `pnorm` function in R to find a theory-based p-value for a hypothesis test involving a difference in proportions.

- **Standard error of a difference in sample proportions for a confidence interval** (not assuming the null is true):

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

  - Calculation of the confidence interval for a difference in sample proportions:

$$\hat{p}_1 - \hat{p}_2 \pm z^* \times SE(\hat{p}_1 - \hat{p}_2)$$

  - Use the `qnorm` function in R to find the z^* multiplier.

Simulation-based confidence interval

- R code to find the simulation-based confidence interval using the `two_proportion_bootstrap_CI` function from the `catstats` package.

```
two_proportion_bootstrap_CI(formula = response~explanatory,
                            data=object, # Name of data set
                            first_in_subtraction = "xx", # Order of subtraction: enter the name of G
                            response_value_numerator = "xx", # Define which outcome is a success
                            number_repetitions = 10000, # Always use a minimum of 1000 repetitions
                            confidence_level = xx) # Enter the level of confidence as a decimal
```

1.2 Video Notes: Theoretical Inference for Two Categorical Variables

Read Sections 15.3 and 15.4 in the course textbook. Use the following videos to complete the video notes for Module 9.

1.2.1 Course Videos

- 15.3TheoryTests
- 15.3TheoryIntervals

Hypothesis testing using theory-based methods - Video 15.4TheoryTests

Example: In Modules 3 and 4, we investigated data on higher education institutions in the United States, collected by the Integrated Postsecondary Education Data System (IPEDS) for the National Center for Education Statistics (NCES) (Education Statistics 2018). A random sample of 2900+ higher education institutions in the United States was collected in 2018. Two variables measured on this data set is whether the institution is a land grant university and whether the institution offers tenure. Does the proportion of universities that offer tenure differ between land grant and non-land-grant institutions?

What is the explanatory variable?

What is the response variable?

Write the parameter of interest:

Hypotheses:

In notation:

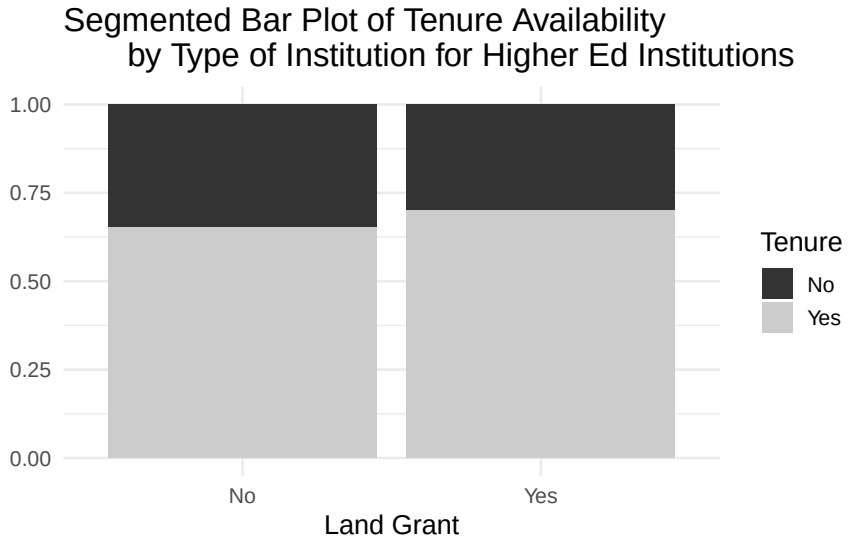
H_0 :

H_A :

```
IPED <-read.csv("https://math.montana.edu/courses/s216/data/IPEDS_2018.csv")

IPEDS <- IPED %>%
    drop_na(Tenure)

IPEDS %>% # Data set piped into...
    ggplot(aes(x = LandGrant, fill = Tenure)) +   # This specifies the variables
  geom_bar(stat = "count", position = "fill") +  # Tell it to make a stacked bar plot
  labs(title = "Segmented Bar Plot of Tenure Availability
       by Type of Institution for Higher Ed Institutions",
       # Make sure to title your plot
       x = "Land Grant",   # Label the x axis
       y = "") + # Remove y axis label
```

```
    scale_fill_grey()

IPEDS %>% group_by(LandGrant) %>% count(Tenure)
```

```
#> # A tibble: 4 x 3
#> # Groups:   LandGrant [2]
#>   LandGrant Tenure     n
#>   <chr>     <chr>  <int>
#> 1 No        No       976
#> 2 No        Yes     1829
#> 3 Yes       No        31
#> 4 Yes       Yes       72
```



Report the summary statistic:

Conditions for inference using theory-based methods for two categorical variables:

- Independence: the response for one observational unit will not influence another observational unit
- Large enough sample size:

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

Steps to use theory-based methods:

- Calculate the standardized statistic
- Find the area under the standard normal distribution at least as extreme as the standardized statistic

Equation for the standard error of the difference in sample proportions assuming the null hypothesis is true:

- This value measures how far each possible sample difference in proportions is from the null value, on average.

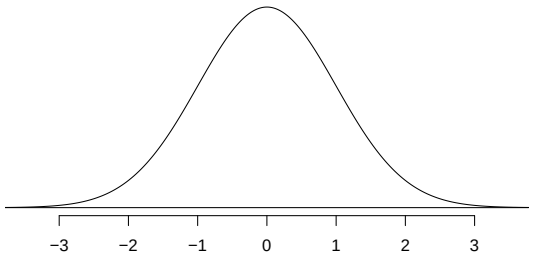
Equation for the standardized difference in sample proportions:

- This value measures how many standard errors the sample difference in proportions is above/below the null value.

Calculate the standardized difference in sample proportion of higher education institutions that offer tenure between land grant universities and non-land grant universities.

- First calculate the standard error of the difference in proportion assuming the null hypothesis is true

- Then calculate the Z score



Interpret the standardized statistic

To find the p-value, find the area under the standard normal distribution at the standardized statistic and more extreme.

```
pnorm(0.985, lower.tail = FALSE)*2
```

```
#> [1] 0.3246241
```

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 with scope of inference:

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

Confidence interval - Video 15.3TheoryIntervals

- Estimate the _____ in true _____
- $CI =$ statistic $\pm$ margin of error

Theory-based method for a two categorical variables

- $CI = \hat{p}_1 - \hat{p}_2 \pm (z^* \times SE(\hat{p}_1 - \hat{p}_2))$
- When creating a confidence interval, we no longer assume the _____ hypothesis is true. Use the sample _____ to calculate the sample to sample variability, rather than $\hat{p}_{pooled}$.

Equation for the standard error of the difference in sample proportions NOT assuming the null is true:

Example: Estimate the difference in true proportions of higher education institutions that offer tenure between land grant universities and non-land grant universities.

Find a 90% confidence interval:

- 1st find the z^* multiplier

```
qnorm(0.95, lower.tail=TRUE)
```

```
#> [1] 1.644854
```

- Next, calculate the standard error for the difference in proportions **NOT** assuming the null hypothesis is true

- Calculate the margin of error

- Calculate the endpoints of the 90% confidence interval

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

1.2.2 Concept Check

Be prepared for group discussion in the next class. One member from the table should write the answers to the following on the whiteboard.

1. What conditions must be met to use the Normal Distribution to approximate the sampling distribution for the difference in sample proportions?

2. Explain why a theory-based confidence interval for the Good Samaritan study from last module would NOT be similar to the bootstrap interval created.

1.3 Activity 19: Winter Sports Helmet Use and Head Injuries — Theory-based Methods

1.3.1 Learning outcomes

- Assess the conditions to use the normal distribution model for a difference in proportions.
- Create and interpret a theory-based confidence interval for a difference in proportions.
- Calculate and interpret the standardized difference in sample proportion
- Use the standard normal distribution to find the p-value for the test

1.3.2 Terminology review

In today's activity, we will use theory-based methods to estimate the difference in two proportions. Some terms covered in this activity are:

- Standard normal distribution
- Independence and success-failure conditions

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

1.3.3 Winter sports helmet use and head injury

In this activity we will focus on theory-based methods to calculate a confidence interval. The sampling distribution of a difference in proportions can be mathematically modeled using the normal distribution if certain conditions are met.

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

- **Independence**: The data are independent within and between the two groups. (*Remember*: This also must be true to use simulation methods!)
- **Success-failure condition**: This condition is met if we have at least 10 successes and 10 failures in each sample. Equivalently, we check that all cells in the table have at least 10 observations.

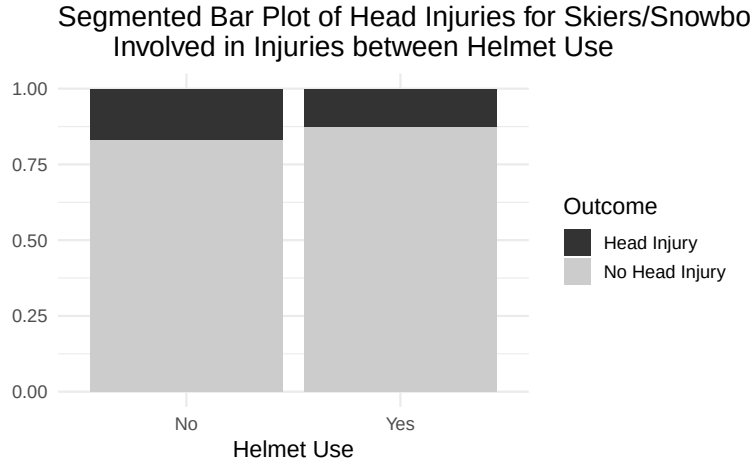
A study was reported in "Helmet Use and Risk of Head Injuries in Alpine Skiers and Snowboarders" by Sullheim et. al., (Sulheim et al. 2017), on the use of helmets and head injuries for skiers and snowboarders involved in accidents. The summary results from a random sample of 3562 skiers and snowboarders involved in accidents is shown in the two-way table below.

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

- Download the R script file from D2L and upload to the RStudio server
- Highlight and run 1–13 to import the data set and create the segmented bar plot

```
skiers <- read.csv("https://www.math.montana.edu/courses/s216/data/HeadInjuries.csv") # Read data set in
skiers %>% # Data set piped into...
  ggplot(aes(x = Helmet, fill = Outcome)) +   # This specifies the variables
  geom_bar(stat = "count", position = "fill") +  # Tell it to make a stacked bar plot
  labs(title = "Segmented Bar Plot of Head Injuries for Skiers/Snowboarders
       Involved in Injuries between Helmet Use",  # Make sure to title your plot
```

```
      x = "Helmet Use",    # Label the x axis
      y = "") +  # Remove y axis label
scale_fill_grey()  # Make figure black and white
```



1. Verify the independence condition is met.

2. Verify the success failure condition is met to use theory-based methods.

3. Calculate the difference in sample proportion of skiers and snowboarders involved in accidents with a head injury for those who wear helmets and those who do not. Use appropriate notation with informative subscripts.

Hypothesis test

4. Write the null and alternative hypotheses in notation.

H_0:

H_A:

Use statistical analysis methods to draw inferences from the data

To test the null hypothesis, we could use simulation-based methods as we did in the activities in Module 8. In this activity, we will focus on theory-based methods. Like with a single proportion, the sampling distribution of a difference in sample proportions can be mathematically modeled using the normal distribution if certain conditions are met.

To calculate the standardized statistic we use:

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - \text{null value}}{SE_0(\hat{p_1} - \hat{p}_2)},$$

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

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

For this study we would first calculate the pooled proportion of successes.

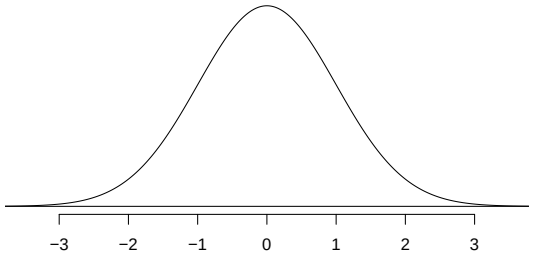
$$\hat{p}_{pool} = \frac{\text{number of "successes"}}{\text{number of cases}}$$

5. Calculate the pooled proportion of head injuries.

6. Use the value for the pooled proportion of successes to calculate the $SE_0(\hat{p}_1 - \hat{p}_2)$ assuming the null hypothesis is true.

7. Use the value of the null standard error to calculate the standardized statistic (standardized difference in proportion).

8. Mark the value of the standardized difference in proportion on the standard normal distribution shown below. Interpret the standardized statistic in context of the problem.



We will use the `pnorm()` function in R to find the p-value.

- Enter the value of z from question 7 for xx
- Highlight and run lines 18–20

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

9. Write a conclusion to the test.

How would an increase in sample size impact the p-value of the test?

	Helmet Use	No Helmet Use	Total
Head Injury	135	674	809
No Head Injury	921	3270	4191
Total	1056	3944	5000

Note that the sample proportions for each group are the same as the smaller sample size.

$$\hat{p}_h = \frac{135}{1056} = 0.128, \quad \hat{p}_n = \frac{674}{3944} = 0.171$$

First calculate the pooled proportion of successes.

$$\hat{p}_{pool} = \frac{\text{number of "successes"}}{\text{number of cases}} = \frac{809}{5000} = 0.162$$

We use the value for the pooled proportion of successes to calculate the $SE_0(\hat{p}_1 - \hat{p}_2)$.

$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{0.162 \times (1 - 0.162) \times \left(\frac{1}{1056} + \frac{1}{3944}\right)} = 0.013$$

Standardized Statistic Calculation:

$$Z = \frac{0.128 - 0.171 - 0}{0.013} = -3.308$$

Use Rstudio to find the p-value for this new sample.

```
pnorm(-3.308, # Enter value of standardized statistic
      m=0, s=1, # Using the standard normal mean = 0, sd = 1
      lower.tail=TRUE) # Gives a p-value greater than the standardized statistic
```

```
#> [1] 0.000469824
```

10. How does the increase in sample size affect the p-value?

1.3.4 Take-home messages

1. Simulation-based methods and theory-based methods should give similar results for a study *if the validity conditions are met*. For both methods, observational units need to be independent. To use theory-based methods, additionally, the success-failure condition must be met. Check the validity conditions for each type of test to determine if theory-based methods can be used.
2. When calculating the standard error for the difference in sample proportions when doing a hypothesis test, we use the pooled proportion of successes, the best estimate for calculating the variability *under the assumption the null hypothesis is true*.
3. Increasing sample size will result in less sample-to-sample variability in statistics, which will result in a smaller standard error, and a larger standardized statistic.

1.3.5 Additional notes

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

1.4 Activity 20: Diabetes

1.4.1 Learning outcomes

- Assess the conditions to use the normal distribution model for a difference in proportions.
- Describe and perform a simulation-based confidence interval for a difference in proportions.
- Create and interpret a theory-based confidence interval for a difference in proportions.

1.4.2 Glycemic control in diabetic adolescents

Researchers compared the efficacy of two treatment regimens to achieve durable glycemic control in children and adolescents with recent-onset type 2 diabetes (Group 2012). A convenience sample of patients 10 to 17 years of age with recent-onset type 2 diabetes were randomly assigned to either a medication (rosiglitazone) or a lifestyle-intervention program focusing on weight loss through eating and activity. Researchers measured whether the patient still needs insulin (failure) or had glycemic control (success). Of the 233 children who received the Rosiglitazone treatment, 143 had glycemic control, while of the 234 who went through the lifestyle-intervention program, 125 had glycemic control. Is there evidence that there is difference in proportion of patients that achieve durable glycemic control between the two treatments? Use Rosiglitazone – Lifestyle as the order of subtraction.

- Upload and open the R script file. Upload the csv file, `diabetes`.
- Enter the name of the data set for `datasetname.csv` in the R script file in line 7.
- Highlight and run lines 1–8 to get the counts for each combination of categories.

```
glycemic <- read.csv("datasetname.csv")
glycemic %>% group_by(treatment) %>% count(outcome)
```

1. Is this an experiment or an observational study?

2. Complete the following two-way table using the R output.

	Treatment		
Outcome	rosiglitazone	lifestyle	Total
glycemic control (success)			
insulin required (failure)			
Total			

3. Is the independence condition met for this study? Explain your answer.

4. Is the success failure condition met for this study? Explain your answer.

5. Write the parameter of interest for the research question.

6. **Calculate the summary statistic (difference in proportions). Use appropriate notation.**

Simulation methods

First we will use simulation methods to find the confidence interval. This will give an interval estimate for the parameter of inference.

We will use the `two_proportion_bootstrap_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of differences in sample proportions and calculate a 90% confidence interval. We will need to enter the response variable name and the explanatory variable name for the formula, the data set name (identified above as `glycemic`), the outcome for the explanatory variable that is first in subtraction, number of repetitions, the outcome for the response variable that is a success (what the numerator counts when calculating a sample proportion), and the confidence level as a decimal.

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

- First in subtraction (What is the outcome for the explanatory variable that is used as first in the order of subtraction? `"rosi"` or `"lifestyle"`):
- Number of repetitions:
- Response value numerator (What is the outcome for the response variable that is considered a success? `"success"` or `"failure"`):
- confidence_level:
- Fill in the missing values/names in the R script file in the two_proportion_bootstrap_CI function to create a simulation 90% confidence interval.
- Highlight and run lines 12–17

```
two_proportion_bootstrap_CI(formula = response~explanatory,
         data=glycemic, # Name of data set
         first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1
         response_value_numerator = "xx", # Define which outcome is a success
         number_repetitions = 10000, # Always use a minimum of 1000 repetitions
         confidence_level = xx) # Enter the level of confidence as a decimal
```

8. Report the 90% confidence interval.

9. Interpret the confidence interval in context of the problem.

Theory-based Methods

Next we will use theory-based methods to find the 90% confidence interval. Review the conditions for using theory-based methods from Activity 19.

10. Is the sample size large enough to use theory-based methods to find the confidence interval? Explain in context of the study,

To find a confidence interval for the difference in proportions we will add and subtract the margin of error from the point estimate to find the two endpoints.

$$\hat{p}_1 - \hat{p}_2 \pm z^* \times SE(\hat{p}_1 - \hat{p}_2), \quad \text{where}$$

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

In this formula, we use the sample proportions for each group to calculate the standard error for the difference in proportions since we are not assuming that the true difference is zero.

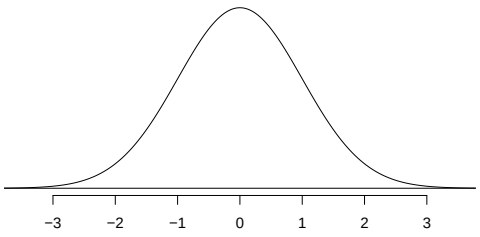
11. Calculate the standard error of the sample proportion not assuming the null hypothesis is true.

Recall that the z^* multiplier is the percentile of a standard normal distribution that corresponds to our confidence level. If our confidence level is 90%, we find the Z values that encompass the middle 90% of the standard normal distribution. If 90% of the standard normal distribution should be in the middle, that leaves 10% in the tails, or 5% in each tail. The `qnorm()` function in R will tell us the z^* value for the desired percentile (in this case, 90% + 5% = 95% percentile).

```
qnorm(0.95, lower.tail = TRUE) # Multiplier for 90% confidence interval
```

```
#> [1] 1.644854
```

12. Mark the value of the z^* multiplier and the percentages used to find this multiplier on the standard normal distribution shown below.



Remember that the margin of error is the value added and subtracted to the sample difference in proportions to find the endpoints for the confidence interval.

$$ME = z^* \times SE(\hat{p}_1 - \hat{p}_2)$$

13. Using the multiplier of $z^* = 1.645$ and the calculated standard error, calculate the margin of error for a 90% confidence interval.

14. Calculate the 90% confidence interval for the parameter of interest.

1.4.3 Take-home messages

1. When calculating the standard error for the difference in sample proportions when doing a hypothesis test, we use the pooled proportion of successes, the best estimate for calculating the variability *under the assumption the null hypothesis is true*. For a confidence interval, we are not assuming a null hypothesis, so we use the values of the two conditional proportions to calculate the standard error. Make note of the difference in these two formulas.
2. Increasing sample size will result in less sample-to-sample variability in statistics, which will result in a smaller standard error, and a larger standardized statistic.
3. Since we add and subtract the margin of error to the point estimate, the margin of error is half the width of the confidence interval.

1.4.4 Additional notes

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

1.5 Module 9 Lab: Poisonous Mushrooms

1.5.1 Learning outcomes

- Given a research question involving two categorical variables, 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 proportions.
- Interpret and evaluate a p-value for a simulation-based hypothesis test for a difference in proportions.
- Interpret and evaluate a confidence interval for a simulation-based confidence interval for a difference in proportions.

1.5.2 Poisonous Mushrooms

Wild mushrooms, such as chanterelles or morels, are delicious, but eating wild mushrooms carries the risk of accidental poisoning. Even a single bite of the wrong mushroom can be enough to cause fatal poisoning. An amateur mushroom hunter is interested in finding an easy rule to differentiate poisonous and edible mushrooms. They think that the mushroom's gills (the part which holds and releases spores) might be related to a mushroom's edibility. They used a data set of 8124 mushrooms and their descriptions. For each mushroom, the data set includes whether it is edible (e) or poisonous (p) and the size of the gills (broad (b) or narrow (n)). Is there evidence gill size is associated with whether a mushroom is poisonous? PLEASE NOTE: According to The Audubon Society Field Guide to North American Mushrooms, there is no simple rule for determining the edibility of a mushroom; no rule like "leaflets three, let it be'' for Poisonous Oak and Ivy.

- Upload and open the R script file for the Module 9 lab. Upload and import the csv file, `mushrooms_edibility`.
- Enter the name of the data set (see the environment tab) for datasetname.csv in the R script file in line 8.
- Highlight and run lines 1–9 to get the counts for each combination of categories.

```
mushrooms <- read.csv("datasetname.csv") # Read data set in
mushrooms %>% group_by(gill_size) %>% count(edibility) #finds the counts in each group
```

1. What is the explanatory variable? How are the two levels of the explanatory variable written in the data set?

2. What is the response variable? How are the two levels of the response variable written in the data set?

3. Write the parameter of interest in words, in context of the study.

4. Write the null hypothesis for this study in notation.

5. **Using the research question, write the alternative hypothesis in words.**

6. Fill in the following two-way table using the R output.

	Gill Size		
Edibility	Broad (b)	Narrow (n)	Total
Poisonous (p)			
Edible (e)			
Total			

7. **Calculate the difference in proportion of mushrooms that are poisonous for broad gill mushrooms and narrow gill mushrooms. Use broad - narrow for the order of subtraction. Use appropriate notation.**

- Fill in the missing values/names in the R script file for the `two-proportion_test` function to create the null distribution and find the p-value for the test.

```
two_proportion_test(formula = response~explanatory, # response ~ explanatory
    data= mushrooms, # Name of data set
    first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1
    number_repetitions = 1000, # Always use a minimum of 1000 repetitions
    response_value_numerator = "xx", # Define which outcome is a success
    as_extreme_as = xx, # Calculated observed statistic (difference in sample proportions)
    direction="xx") # Alternative hypothesis direction ("greater","less","two-sided")
```

8. Report the p-value for the study.

9. **Do you expect that a 90% confidence interval would contain the null value of zero? Explain your answer.**

- Fill in the missing values/names in the R script file in the two_proportion_bootstrap_CI function to create a simulation 90% confidence interval.
- **Upload a copy of the bootstrap distribution to Gradescope.**

```
two_proportion_bootstrap_CI(formula = response~explanatory,
         data=mushrooms, # Name of data set
         first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1
         response_value_numerator = "xx", # Define which outcome is a success
         number_repetitions = 1000, # Always use a minimum of 1000 repetitions
         confidence_level = xx) # Enter the level of confidence as a decimal
```

10. Report the 90% confidence interval.

11. Write a paragraph summarizing the results of the study as if writing a press release. Be sure to describe:

- Summary statistic and interpretation
  - Summary measure (in context)
  - Value of the statistic
  - Order of subtraction when comparing two groups
- 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
  - To what group of observational units do the results apply (target population or observational units similar to the sample)?
  - What type of inference is appropriate (causal or non-causal)?

Upload your group's confidence interval interpretation and conclusion to Gradescope.

Paragraph:

"Average Driving Distance and Fairway Accuracy." 2008. https://www.pga.com/ and https://www.lpga.com/.
Banton, et al, S. 2022. "Jog with Your Dog: Dog Owner Exercise Routines Predict Dog Exercise Routines and Perception of Ideal Body Weight." *PLoS ONE* 17(8).
Bhavsar, et al, A. 2022. "Increased Risk of Herpes Zoster in Adults ≥50 Years Old Diagnosed with COVID-19 in the United States." *Open Forum Infectious Diseases* 9(5).
Bulmer, M. n.d. "Islands in Schools Project." https://sites.google.com/site/islandsinschoolsprojectwebsite/home.
"Bureau of Transportation Statistics." 2019. https://www.bts.gov/.
"Child Health and Development Studies." n.d. https://www.chdstudies.org/.
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.
Davis, Smith, A. K. 2020. "A Poor Substitute for the Real Thing: Captive-Reared Monarch Butterflies Are Weaker, Paler and Have Less Elongated Wings Than Wild Migrants." *Biology Letters* 16.
Du Toit, et al, G. 2015. "Randomized Trial of Peanut Consumption in Infants at Risk for Peanut Allergy." *New England Journal of Medicine* 372.
Edmunds, et al, D. 2016. "Chronic Wasting Disease Drives Population Decline of White-Tailed Deer." *PLoS ONE* 11(8).
Education Statistics, National Center for. 2018. "IPEDS." https://nces.ed.gov/ipeds/.
"Great Britain Married Couples: Great Britain Office of Population Census and Surveys." n.d. https://discovery.nationalarchives.gov.uk/details/r/C13351.
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.
Kalra, et al., Dl. 2022. "Trustworthiness of Indian Youtubers." Kaggle. https://doi.org/10.34740/KAGGLE/DSV/4426566.
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/.
Laeng, Mathisen, B. 2007. "Why Do Blue-Eyed Men Prefer Women with the Same Eye Color?" *Behavioral Ecology and Sociobiology* 61(3).
Levin, D. T. 2000. "Race as a Visual Feature: Using Visual Search and Perceptual Discrimination Tasks to Understand Face Categories and the Cross-Race Recognition Deficit." *Journal of Experimental Psychology* 129(4).
LUETKEMEIER, et al., M. 2017. "Skin Tattoos Alter Sweat Rate and Na+ Concentration." *Medicine and Science in Sports and Exercise* 49(7).
Madden, et al, J. 2020. "Ready Student One: Exploring the Predictors of Student Learning in Virtual Reality." *PLoS ONE* 15(3).
Miller, G. A. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review* 63(2).
Moquin, W., and C. Van Doren. 1973. "Great Documents in American Indian History." Praeger.
"More Americans Are Joining the 'Cashless' Economy." 2022. https://www.pewresearch.org/short-reads/2022/10/05/more-americans-are-joining-the-cashless-economy/.
National Weather Service Corporate Image Web Team. n.d. "National Weather Service – NWS Billings." https://w2.weather.gov/climate/xmacis.php?wfo=byz.
O'Brien, Lynch, H. D. 2019. "Crocodylian Head Width Allometry and Phylogenetic Prediction of Body Size in Extinct Crocodyliforms." *Integrative Organismal Biology* 1.
"Ocean Temperature and Salinity Study." n.d. https://calcofi.org/.

"Older People Who Get Covid Are at Increased Risk of Getting Shingles." 2022. https://www.washingtonpost.com/health/2022/04/19/shingles-and-covid-over-50/.

"Physician's Health Study." n.d. https://phs.bwh.harvard.edu/.

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.

Subach, et al, A. 2022. "Foraging Behaviour, Habitat Use and Population Size of the Desert Horned Viper in the Negev Desert." *Soc.Open Sci* 9.

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.

Wahlstrom, et al, K. 2014. "Examining the Impact of Later School Start Times on the Health and Academic Performance of High School Students: A Multi-Site Study." *Center for Applied Research and Educational Improvement*.

Watson, et al., N. 2015. "Recommended Amount of Sleep for a Heathy Adult: A Joint Consensus Statement of the American Academy of Sleep Medicine and Sleep Research Society." *Sleep* 38(6).

Weiss, R. D. 1988. "Relapse to Cocaine Abuse After Initiating Desipramine Treatment." *JAMA* 260(17).

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

Wilson, Woodruff, J. P. 2016. "Vertebral Adaptations to Large Body Size in Theropod Dinosaurs." *PLoS ONE* 11(7).