STAT 216 Coursepack

Summer 2023
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/. 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. The following will give you an idea of what 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.

WEEK 1

Inference for Two Categorical Variables: Theory-based Methods

1.1 Week 9 Reading Guide: Theory-based Inference for a Difference in Proportions

1.1.1 Section 15.3 (Theory-based inferential methods for $\pi_1 - \pi_2$)

Videos

- 15.3Tests
- 15.3Intervals

Reminders from previous sections

n = sample size

$\hat{p}$ = sample proportion

π = population proportion

General steps of a hypothesis test:

1. Frame the research question in terms of hypotheses.
2. Collect and summarize data using a test statistic.
3. Assume the null hypothesis is true, and simulate or mathematically model a null distribution for the test statistic.
4. Compare the observed test (standardized) statistic to the null distribution to calculate a p-value.
5. Make a conclusion based on the p-value and write the conclusion in context.

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

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

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

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

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

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

Also called a 'significance test'.

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.

$\Longrightarrow$ 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.

Confidence level: how confident we are that the confidence interval will capture the parameter.

Notes

Conditions for the Central Limit Theorem to apply for a difference in proportions

Independence:

Checked by:

Success-failure condition:

Checked by:

Formulas

$SD(\hat{p_1} - \hat{p_2}) =$

Null standard error of the difference in sample proportions: $SE_0(\hat{p_1} - \hat{p_2}) =$

Standardized statistic (or standardized difference in sample proportions): $Z =$

Standard error of the difference in sample proportions when we do not assume the null hypothesis is true: $SE(\hat{p_1} - \hat{p_2}) =$

Theory-based confidence interval for a difference in proportions:

Margin of error of a confidence interval for a difference in proportions:

Notation

Overall (pooled) proportion of successes:

Example: CPR and blood thinners

1. What are the observational units?

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

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

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

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

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

7. Calculate the standard error of the difference in sample proportions without assuming a null hypothesis.

8. Calculate the 90% confidence interval using $z^* = 1.65$ as the multiplier.

Note: A confidence interval interpretation and confidence level interpretation for this example can be found in the Reading Guide solutions for Sections 15.1 and 15.2.

Example: Mammograms

1. What are the observational units?

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

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

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

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

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

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

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

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

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

11. Calculate the standardized statistic.

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

13. Explain how the p-value for this test was calculated.

14. Interpret the p-value in the context of the study.

15. At the 10% significance level, what decision should be made?

16. Write a conclusion for the research question.

1.2 Activity 9A: Winter Sports Helmet Use and Head Injuries — Theory-based Hypothesis Test

1.2.1 Learning outcomes

- Given a research question involving two categorical variables, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Assess the conditions to use the normal distribution model for a difference in proportions.
- Calculate the Z test statistic for a difference in proportions.
- Find, interpret, and evaluate the p-value for a theory-based hypothesis test for a difference in proportions.

1.2.2 Terminology review

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

- Conditional proportion
- Z test
- z^* multiplier
- Null hypothesis
- Alternative hypothesis
- Test statistic
- Standard normal distribution
- Independence and success-failure conditions
- Relative risk

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

1.2.3 Helmet use and head injuries

In "Helmet Use and Risk of Head Injuries in Alpine Skiers and Snowboarders" by Sullheim et. al., (Sulheim et al. 2017), we can see the summary results from a random sample of 3562 skiers and snowboarders involved in accidents in the two-way table below. Is there evidence that safety helmet use is associated with a reduced risk of head injury for skiers and snowboarders?

For this study the observational units are skiers and snowboarders involved in accidents. A success will be considered a head injury in this context and we are comparing the groups helmet use (group 1) and no helmet use (group 2). Use helmet use - no helmet use as the order of subtraction. Highlight and runs lines 1–6 in the provided Rscript file to create the summary data table.

```
injury <- read.csv("https://math.montana.edu/courses/s216/data/HeadInjuries.csv")
injury %>% group_by(Helmet) %>% count(Outcome)
```

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

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

	Helmet Use		
Head Injury	Yes	No	Total
Head Injury			
No Head Injury			
Total			

2. Write the null and alternative hypotheses in notation.

H_0:

H_A:

3. Calculate the summary statistic (difference in proportions) for this study. Use appropriate notation with clear subscripts.

4. Interpret the difference in sample proportions in context of the study.

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 Activity 8A. 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.

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.

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

6. Is the success-failure condition met for each group? Explain in context of the study.

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:

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

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

$$\hat{p}_{pool} = \frac{\text{number of "successes"}}{\text{number of cases}} = \frac{576}{3562} = 0.162$$

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

$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{0.162(1 - 0.162)\left(\frac{1}{752} + \frac{1}{2810}\right)} = 0.015$$

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

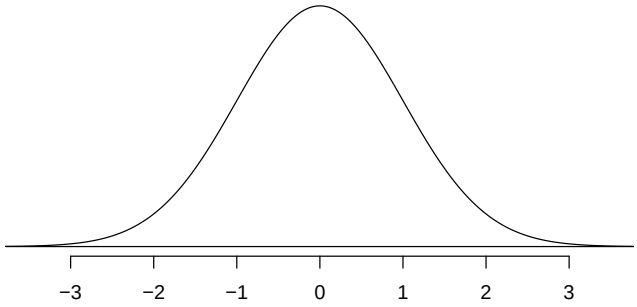

Figure 1.1: A standard normal curve.

8. Mark the value of the standardized statistic on the standard normal distribution above and shade the area to find the p-value.

We will use the `pnorm()` function in R to find the p-value. Use the provided R script file and enter the value of the standardized statistic found in question 7 at `xx` in line 11; highlight and run lines 11–13.

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

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

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

12. Would a 90% confidence interval contain the null value of zero? Explain your answer.

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

Impacts on the p-value

Suppose that we want to show that there is a **difference** in true proportion of head injuries for those that wear helmets and those that do not.

14. Write out the alternative hypothesis in notation for this new research question.

15. How would this impact the p-value?

Suppose in a larger sample of skiers and snowboarders involved in accidents we saw the following results.

	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.

$$\hat{p}_1 = \frac{135}{1056} = 0.127, \hat{p}_2 = \frac{674}{3944} = 0.171$$

16. The standard error for the difference in proportions for this new sample is 0.013 ($SE(\hat{p}_h - \hat{p}_n) = 0.013$). Calculate the standardized statistic for this new sample.

Use Rstudio to find the p-value for this new sample. Enter the value of the standardized statistic found in question 16 for xx in line 18. 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 greater than the standardized statistic
```

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

18. Suppose another sample of 3562 skiers and snowboarders was taken. In this new sample a difference in proportions of head injuries was found to be -0.009, ($\hat{p}_h - \hat{p}_n = -0.009$) with a standard error for the difference in proportions of 0.015, ($SE(\hat{p}_h - \hat{p}_n) = 0.015$). Calculate the standardized statistic for this new sample.

Use Rstudio to find the p-value for this new sample. Enter the value of the standardized statistic found in question 18 for xx in line 25. Highlight and run lines 25–27.

```
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 greater than the standardized statistic
```

19. How does a statistic closer to the null value affect the p-value?

20. Summarize how each of the following affected the p-value:

a) Switching to a two-sided test.

b) Using a larger sample size.

c) Using a sample statistic closer to the null value.

1.2.4 Take-home messages

1. When comparing two groups, we are looking at the difference between two parameters. In the null hypothesis, we assume the two parameters are equal, or that there is no difference between the two proportions.
2. The standardized statistic when the response variable is categorical is a Z-score and is compared to the standard normal distribution to find the p-value. To find the standardized statistic, we take the value of the statistic minus the null value, divided by the null standard error of the statistic. The standardized statistic measures the number of standard errors the statistic is from the null value.
3. The p-value for a two-sided test is approximately two times the value for a one-sided test. A two-sided test provides less evidence against the null hypothesis.
4. The larger the sample size, the smaller the sample to sample variability. This will result in a larger standardized statistic and more evidence against the null hypothesis.
5. The farther the statistic is from the null value, the larger the standardized statistic. This will result in a smaller p-value and more evidence against the null hypothesis.

1.2.5 Additional notes

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

1.3 Activity 9B: Winter Sports Helmet Use and Head Injuries — Theory-based Confidence Interval

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.

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. Recall from Activity 9A, 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.

1. Explain why a theory-based confidence interval for the data set in Activities 8A and 8B would NOT be similar to the bootstrap interval created.

For this activity we will again use the Helmet Use and Head Injury data set. In Activity 9A we saw that there was evidence that helmet use is associated with a reduced risk of head injury. Today we will estimate the difference in proportion of head injuries for those who wore helmets and those who did not.

In "Helmet Use and Risk of Head Injuries in Alpine Skiers and Snowboarders" by Sullheim et. al., (Sulheim et al. 2017), we can see the summary results from a random sample of 3562 skiers and snowboarders involved in accidents 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

2. Write the parameter of interest for this study in context of the problem.

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.

$$\hat{p}_1 - \hat{p}_2 \pm z^* \times SE(\hat{p}_1 - \hat{p}_2), \text{ where}$$
$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Note that the formula changes when calculating the variability around the statistic in order to calculate a confidence interval from the formula used in Activity 9A! Here, we use the sample proportions for each group to calculate the standard error for the difference in proportions since we are not assuming that the true difference is zero.

To calculate the standard error for a difference in proportions to create a 90% confidence interval we substitute in the two sample proportions and the sample size for each group into the equation above.

$$n_1 = 752, n_2 = 2810, \hat{p}_1 = \frac{96}{752} = 0.128, \hat{p}_2 = \frac{480}{2810} = 0.171$$

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{0.128(1-0.128)}{752} + \frac{0.171(1-0.171)}{2810}} = 0.014$$

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) # Multiplier for 90% confidence interval
```

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

3. Draw and label a standard normal distribution. Mark the value of the z^* multiplier adn the percentages used to find this muliplier.

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.

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

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

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

6. Interpret the confidence interval found in question 5 in context of the problem.

7. Interpret the level of confidence in context of the problem. What does it mean to be 90% confident in the confidence interval?

8. What decision (reject or fail to reject the null hypothesis) would you make based on your confidence interval? Explain your answer.

1.3.4 Effect of sample size

Suppose in another sample of skiers and snowboards involved in accidents we saw these results:

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

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

$$\hat{p}_1 = \frac{135}{1056} = 0.127, \hat{p}_2 = \frac{674}{3944} = 0.171$$

9. Calculate the standard error for the difference in sample proportions for this new sample.

10. Calculate the margin of error for a 90% confidence interval using a multiplier of $z^* = 1.645$ for this new sample. Is the margin of error larger or smaller than the margin of error for the original study?

11. Calculate the 90% confidence interval for this new study using the margin of error from question 10.

12. Is the confidence interval calculated in question 11 with the larger sample size wider or narrower than the confidence interval in question 5? Why?

1.3.5 Take-home messages

1. Simulation-based methods and theory-based methods should give the same results for a study *if the validity conditions are met*. For both methods, observational units need to be independent. To use theory-based methods, additionally, the success-failure condition must be met. Check the validity conditions for each type of test to determine if theory-based methods can be used.

2. When calculating the standard error for the difference in sample proportions when doing a hypothesis test, we use the pooled proportion of successes, the best estimate for calculating the variability *under the assumption the null hypothesis is true*. For a confidence interval, we are not assuming a null hypothesis, so we use the values of the two conditional proportions to calculate the standard error. Make note of the difference in these two formulas.

3. Increasing sample size will result in less sample-to-sample variability in statistics, which will result in a smaller standard error, and thus a narrower confidence interval.

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 Week 9 Lab: Diabetes

1.4.1 Learning outcomes

- Given a research question involving two categorical variables, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Assess the conditions to use the normal distribution model for a difference in proportions.
- Describe and perform a simulation-based hypothesis test for a difference in proportions.
- Calculate the Z test statistic for a difference in proportions.
- Find, interpret, and evaluate the p-value for a hypothesis test 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 for Week 9 lab. Upload and import the csv file, `diabetes`. 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 1–8 to get the counts for each combination of categories.

```
glycemic <- datasetname
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			
Insulin Required			
Total			

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

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

5. Using the research question, write the alternative hypothesis in notation.

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

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

```
two_proportion_test(formula = outcome~treatment, # response ~ explanatory
         data= glycemic, # 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")
```

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

8. **Will the theory-based p-value be similar to the simulation p-value? Explain your answer.**

9. **Calculate the number of standard errors the sample difference in proportion is from the null value of zero.**

10. **Will a 95% simulation confidence interval contain the null value of zero? Explain your answer.**

11. Calculate the standard error for a difference in proportions to create a 95% confidence interval.

12. Use the multiplier of $z^* = 1.96$ and the standard error found in question 11 to calculate a 95% confidence interval for the parameter of interest.

13. Write a paragraph summarizing the results of the study. 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
  - 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 a copy of your group's p-value interpretation and scope of inference to Gradescope.

Paragraph (continued):

WEEK 2

Group Exam 2 Review

Use the provided data set from the Islands (ExamReviewData.csv) and the appropriate Exam 1 Review R script file to answer the following questions. Each adult (>21) islander was selected at random from all adult islanders. Note that some islanders choose not to participate in the study. These islanders that did not consent to be in the study are removed from the dataset before analysis. Variables and their descriptions are listed below.

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: "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			

e. Calculate the value of the 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. Using a significance level of $\alpha = 0.05$, what statistical decision will you make about the null hypothesis?

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?

"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.

"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.

O'Brien, Lynch, H. D. 2019. "Crocodylian Head Width Allometry and Phylogenetic Prediction of Body Size in Extinct Crocodyliforms." *Integrative Organismal Biology* 1.

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.

"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.

"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).