STAT 216 Activity Coursepack

Fall 2020

Contents

Preface

This coursepack accompanies the textbook for STAT 216: Introduction to Statistics at Montana State University. 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. Bring this workbook with you to class each week, and take notes in the workbook as you would your own notes. A well-written complete workbook will provide an optimal study guide for exams!

Statistical Investigations for Two Categorical Variables

Learning Objectives.

- Write out the null and alternative hypothesis for two categorical variables
- Assess the conditions to use the standard normal distributions
- Calculate the Z test statistic for a difference in proportions
- Find the p-value and assess the strength of evidence
- Create and interpret a confidence interval for the difference in proportions

Terminology

Here are a few terms we will use in today's activity.

- Conditional proportion
- Z test
- z* multiplier
- Null Hypothesis
- Alternative Hypothesis
- Test statistic

Review Chapter 5 in your textbook for more information on these topics.

Background

In "Helmet Use and Risk of Head Injuries in Alpine Skiers and Snowboarders" by Sullheim et. al., in the Journal of the American Medical Association, Vol. 295, No. 8, we can see the results from a random sample 3562 skiers and snowboarders involved in accidents.

	Head Injury	No Head Injury	Total
Wore Helmet	96	656	752
Did Not Wear Helmet	480	2330	2810
Total	576	2986	3562

Is there evidence that safety helmet use reduces the risk of head injury for skiers and snowboarders?

Vocabulary Review

1. What is the explanatory variable?

2. What is the response variable?

3. Is this an experiment or observational study?

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

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

5. What is the conditional proportion of skiers/snowboarders with a head injury that wore a helmet?

6. What is the conditional proportion of skiers/snowboarders with a head injury that did not wear a helmet?

Ask a Research Question

In this study we are looking at the relationship between two groups or two parameters (π_1 and π_2). Remember we define the parameter as the true proportion of observational units that represent the variable of interest.

7. What is the variable of interest in this study?

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

   π_1 -

   π_2 -

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

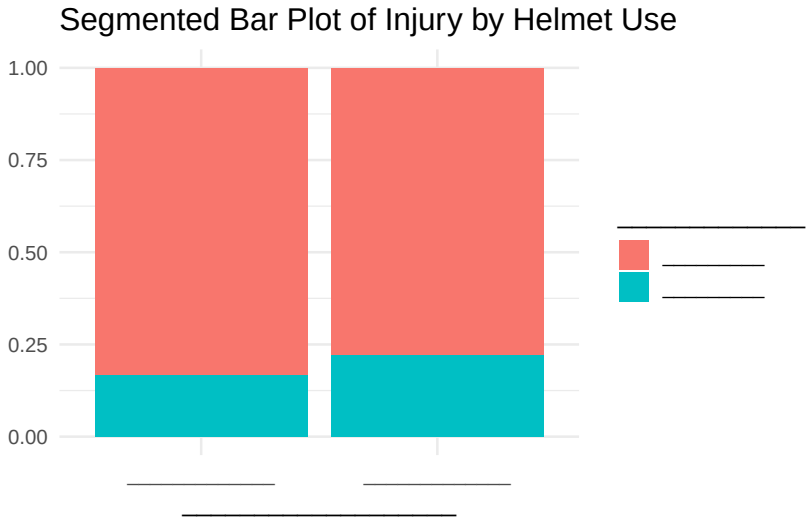
9. Write the null hypothesis out in words using your answers to question 8.

10. What is the research question?

11. Based on the research question fill in the appropriate sign for the alternative hypothesis:

$H_A : \pi_1 - \pi_2$ _____ 0

Summarize and Visualize the data



12. Fill in the blanks on the graph with the appropriate variables and values to plot a segmented bar plot of injury by helmet use.

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

14. Calculate the point estimate for this study. We will use helmet use minus no helmet use as the order of subtraction.

15. What is the notation used for the value calculated in question 14?

Use statistical analysis methods to draw inferences from the data

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

Conditions for the sample distribution of $\hat{p}_1 - \hat{p}_2$

- Independence: The data are independent within and between the two groups.
- Success-Failure Condition: The success-failure condition holds for each group.

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

17. Is the success-failure condition met for each group? Explain your answer.

To calculate the test statistic we use:

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

where the standard error is calculated using the pooled proportion of successes.

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}_{pool}(1-\hat{p}_{pool})(\frac{1}{n_1} + \frac{1}{n_2})}, \text{where}$$

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

18. Calculate the $SE(\hat{p}_1 - \hat{p}_2)$.

19. Calculate the test statistic.

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

```
#> [1] 0.002118205
```

20. Report the p-value.

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

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

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

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

Note that the formula changes when calculating the variability around the statistic in order to calculate a confidence interval! Here use the sample proportions for each group to calculate the standard error for the difference in proportions. The z^* multiplier is found under the normal distribution. We find the values that encompass the middle 95% of the data.

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

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

23. Using the multiplier of $z^* = 1.96$, calculate the 95% confidence interval for the difference in true proportion of head injuries for those that used helmets minus those who did not.

24. Interpret the confidence interval found in question 23 in context of the problem.

25. Write a paragraph summarizing the results of the study. Be sure to include:

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

Types of Errors

Hypothesis tests are not flawless. In a hypothesis test, there are two competing hypotheses: the null and alternative. We make a decision about which might be true, but we may choose incorrectly.

		Test Conclusion	
Truth	H_0 true	good decision	Type 1 Error
	H_A true	Type 2 Error	good decision

A Type 1 Error is rejecting the null hypothesis when $H_0$is actually true. A Type 2 Error is failing to reject the null hypothesis when the alternative is actually true.

26. Using a significance level of 0.05, what decision do you make in regards to the null hypothesis?

27. What type of error could we have made?

28. Write this error in context of the problem.

Statistical Investigations for Paired Data

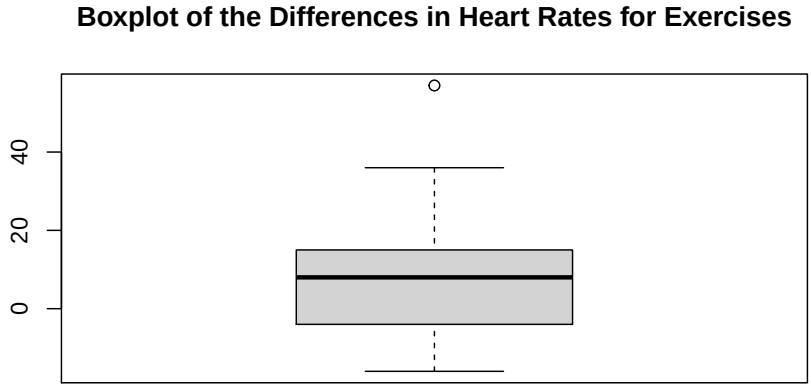
Learning Outcomes

- Given a research question, construct the null and alternative hypotheses in words and using appropriate statistical symbols
- Describe and perform simulation-based hypothesis
- Interpret and evaluate a p-value
- Construct and interpret a theory-based confidence interval
- Use a confidence interval to determine the conclusion of a hypothesis test

Mean Difference in Heart Rates for Jumping Jacks and Bicycle Kicks

Which exercise, jumping jacks or bicycle kicks will raise your heart rate more? Students in an introductory statistics class were asked to participate in an experiment to answer this question. Each student flipped a coin to determine which exercise to complete first. If the coin landed on heads the student would do jumping jacks for 30 seconds and then measure their heart rate. After a 2 minute break the student would do bicycle kicks for 30 seconds and then record their heart rate. If the coin landed on tails the student would complete bicycle kicks first followed by jumping jacks using the same times as above.

Review




```
#>  min Q1 median Q3 max     mean       sd  n missing
#>  -16 -4      8 15  57 7.604651 15.91666 43       0
```

1. What is the sample size?

2. Identify the variables in this study. What role do each have?

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

4. What is the purpose of randomizing the order of jumping jacks and bicycle kicks before measuring heart rates?

Ask a Research Question

5. What are the two competing possibilities to run a hypothesis test?

6. Write the null hypothesis in words.

7. What is the research question?

8. Write the alternative hypothesis in notation.

Summarize and Visualize the Data

9. Report the summary statistic for the data.

10. What notation is used for the value in question 9?

Use statistical inferential methods to draw inferences from the data

To simulate the null distribution we will use a bootstrapping method - sampling with replacement from the data set. Before bootstrapping we will need to shift the each data point by the difference $\mu_0 - \bar{x}$. This will ensure that the simulated null distribution will be centered at the null value.

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

Add simulation here

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

13. What proportion of samples are beyond the sample mean difference in heart beats for jumping jacks minus bicycle kicks?

Communicate the results and answer the research question.

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

To calculate a confidence interval to estimate the mean difference in heart rates, we will use this equation:

$\bar{x}_d \pm t^* SE(\bar{x}_d)$, where $SE(\bar{x}_d) = \sqrt{(\frac{\bar{x}_d}{n_d})}$

To find the t^* multiplier for a 95% confidence interval we will find the value in the t-distribution that represents the endpoints for the middle 95% of the data.

```
#> [1] 2.018082
```

15. Calculate the $SE(\bar{x}_d)$.

16. Calculate a 95% confidence interval for the parameter of interest.

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

18. Based off your p-value and confidence interval, write a conclusion.

Revisit and Look Forward

Suppose we had a sample of 90 students instead of 43 resulting in the same summary statistic.

19. Would this new data provide more or less evidence against the null hypothesis? Explain your answer.

20. Would this result in a wider or narrower confidence interval?