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!

Martian Alphabet

Learning Outcomes

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

Activity

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

Steps of Statistical Investigation

The first step of any statistical investigation is to ask a research question. In this study the research question is: can we as a class read Martian? (we will refine this later on!). To answer any research question, we must design a study and collect data. (This will normally be provided for you in class.) For our question, the study consists of each student being presented with two Martian letters and asking which was Bumba. Your responses will become our observed data that we will explore. To answer the research question we will simulate what *could* have happened in our class given random chance, repeat that many times to understand the expected variability between different "randomly guessing" classes, then comparing our class's observed data to the simulation. This gives us an estimate of how often (or the probability of) our class's result would occur if we were all merely guessing, allowing us to determine if we as a class can in fact read Martian.

Let's explore the data. **Observational units** or **cases** are the subjects data is collected on. In a data set the rows will represent a single observational unit.

1. What are the observational units in this study?

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

A **variable** is information collected or measured on each observational unit or case. We will look at two types of variables: **quantitative** and **categorical**. Each column in a data set will represent a different variable.

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

Categorical variables are data that are in groups or categories such as eye color, state of residency, or whether or not a student is considered in-state. Categorical variables with a natural ordering are considered ordinal variables while those without a natural ordering are considered a nominal variable. All variables will be treated as nominal for analysis.

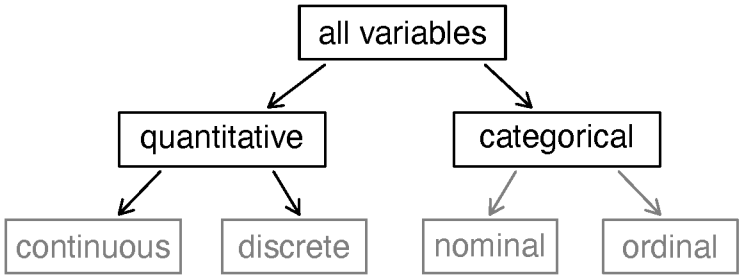


Figure 1: Types of variables.

3. Identify the variable we are collecting on each observational unit in this study, i.e., what are we measuring on each student?

It is important to note that a variable is different than a summary statistic. A variable is measured on a **single observational unit** while a summary statistic is calculated from a group of observational units. For example, the variable **whether or not a student is considered in-state** can be measured on each individual student. In a class of 50 students we can calculate the proportion of students who are considered in-state, the summary statistic. Make sure you wrote the variable in question 3 as a variable **NOT** a summary statistic.

4. Is the variable identified in question 3 categorical or quantitative?

5. Were you correct or incorrect in identifying Bumba?

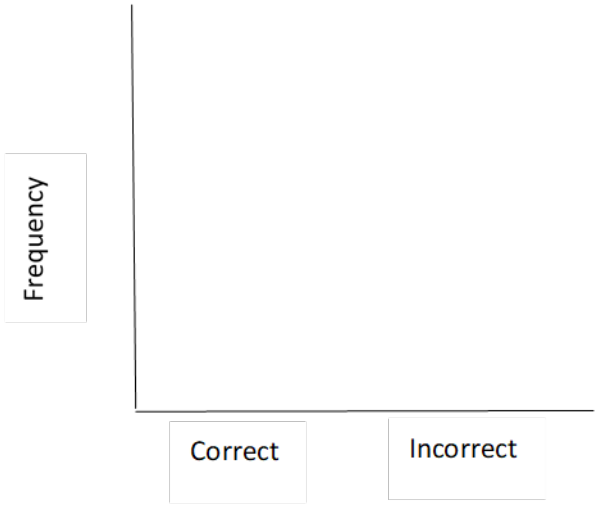
We will now collect the data from the entire class.

6. How many people in your class were correct in identifying Bumba? Using the class size from question 2, calculate the proportion of students who correctly identified Bumba.

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

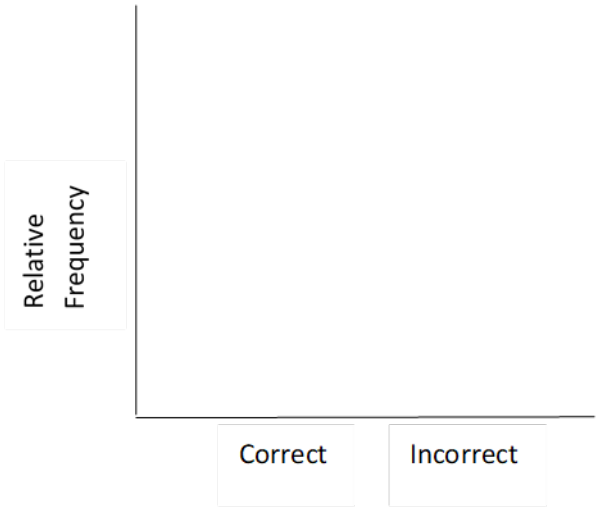
Looking at the data set and the summary statistics is only one way to display the data. We will also want to create a visualization or picture of the data. A **frequency bar plot** is used to display categorical data as a count or frequency. Since our variable has two levels, correct or incorrect, we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

7. Plot the observed class data using a frequency bar plot.



We can also visualize the data as a proportion in a **relative frequency bar plot**. Relative frequency is the proportion calculated for each level of the categorical variable.

8. Plot the observed class data using a relative frequency bar plot.



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

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

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

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

10. Each of you will flip a coin one time to simulate your "guess". Let Heads = correct, Tails = incorrect. What was the result of your simulation?

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

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

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

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

    Probability of correct guesses: _____

    Sample size: _____

    Number of repetitions: _____

14. Sketch the distribution displayed by your instructor here, being sure to label each axis appropriately.

15. Is your class particularly good or bad at Martian? How can you use the plot in question 14 to tell?

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

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

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

Take Home Messages

1. In this course we will learn how to evaluate a claim by comparing observed results (classes' "guesses") to a distribution of many simulated results under an assumption like "blind guessing."
2. Blind guessing between two outcomes will be correct only about half the time. We can create data (via computer simulation) to fit the assumption of blind guessing.
3. Unusual observed results will make us doubt the assumptions used to create the simulated distribution. A large number of correct "guesses" is evidence that a person was not just blindly guessing.

Additional Notes

Use this space to take additional notes on today's activity, and to write down the names and contact information of your team mates.

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

	Head Injury	No Head Injury	Total
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.

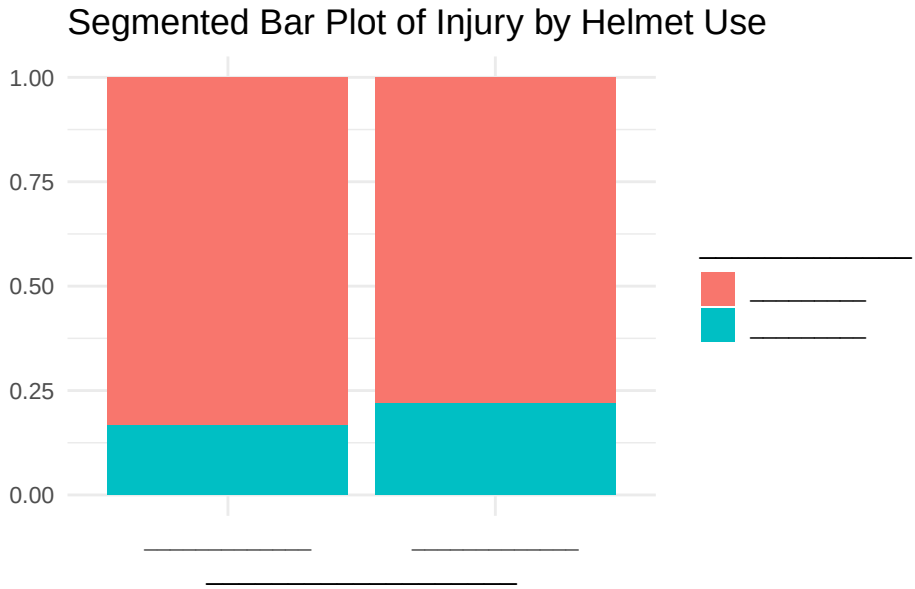
9. Write the null hypothesis out in words using your answers to question 8.

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

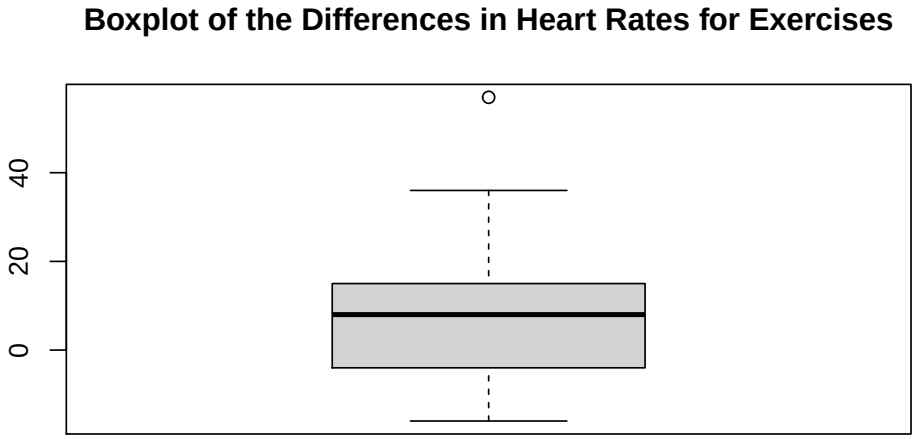
Learning Outcomes

- 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?