Inference for a Difference in Proportion

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.

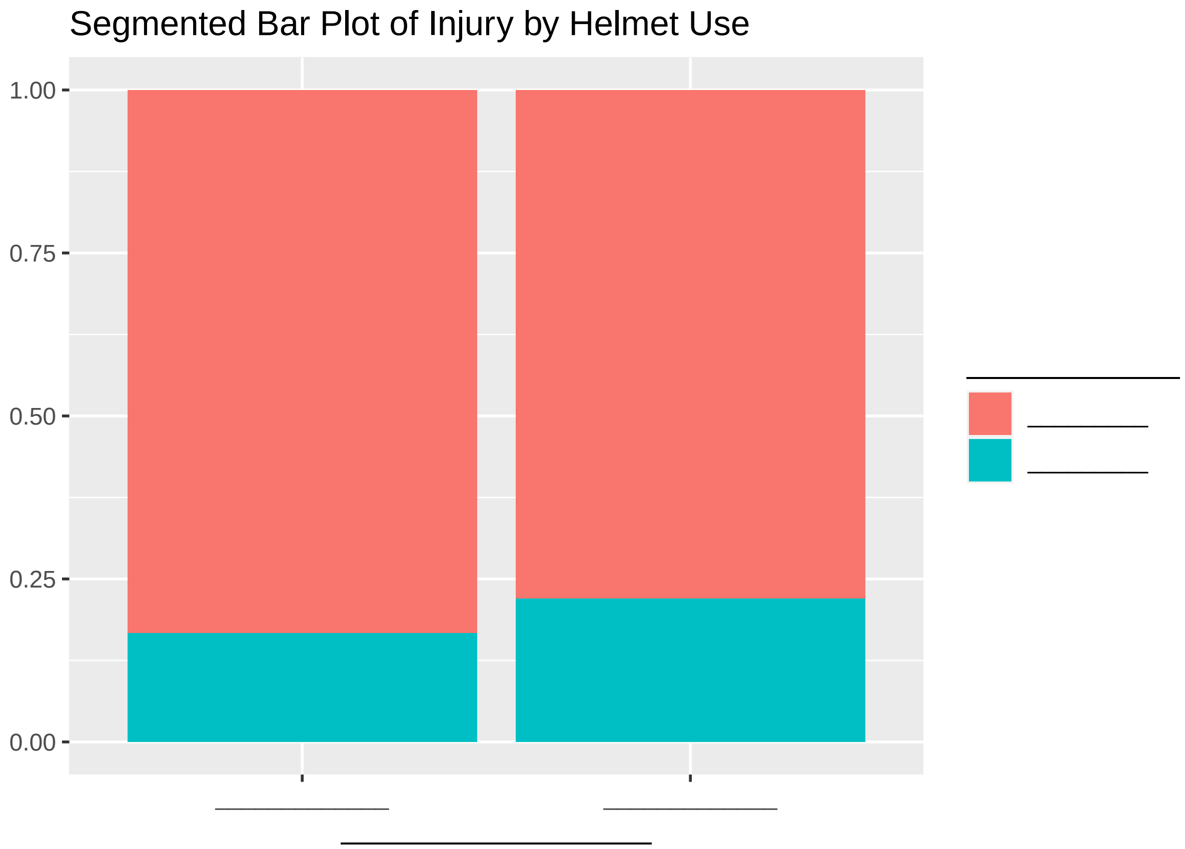
10. What is the research question?

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

```
ggplot(data = injury,   #This specifies the dataset
       aes(x = Injury, fill = Helmet)) +   #This specifies the variables
  geom_bar(stat = "count", position = "fill") +  #Tell it to make a stacked bar plot
  labs(title = "Segmented Bar Plot of Injury by Helmet Use",  #Make sure to title your plot
       x = "___________________",   #Label the x axis
       y = "",  #Remove y axis label
       fill = "_____________") + #Change legend label
  scale_fill_discrete(labels = c("_________", "_________")) +
  scale_x_discrete(labels = c("_____________", "_____________"))
```



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?

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.

```r
pnorm(-2.86 #enter value of test statistic
      , m=0, s=1 #using the standard normal mean = 0, sd = 1
      , lower.tail=TRUE) # gives a p-value less than the test statistic
```

```
## [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.

```r
qnorm(0.95 + 0.025) #multiplier for 95% confidence interval
```

```
## [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

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.