

Winter Sports Helmet Use and Head Injuries

7.1 Learning objectives

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

7.2 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
- Type 1 and Type 2 errors
- Decisions

To review these concepts, see Chapter 5 in your textbook.

7.3 Helmet use and head injuries

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 (2006), we can see the summary results from a random sample 3562 skiers and snowboarders involved in accidents in the two-way table below. Is there evidence that safety helmet use reduces the risk of head injury for skiers and snowboarders?

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

These counts can be found in R by using the `count()` function:

```
injury <- read.csv("data/head_injury.csv") # Read data set in
injury <- # Write over original data with the following
  injury %>% # Pipe data set into
  mutate(Helmet <- factor(Helmet),
         Injury <- factor(Injury)) # Convert to factors

injury %>% group_by(Helmet) %>% count(Injury)
```

```
#> # A tibble: 4 x 3
#> # Groups:   Helmet [2]
#>   Helmet      Injury      n
#>   <chr>      <chr>    <int>
#> 1 No_Helmet Head_Injury    480
#> 2 No_Helmet No_Head_Injury 2330
#> 3 Wore_Helmet Head_Injury     96
#> 4 Wore_Helmet No_Head_Injury  656
```

Vocabulary review

1. What is the explanatory variable?
2. What is the response variable?
3. Is this an experiment or observational study? Justify your answer.
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 that wore a helmet with a head injury?
6. What is the conditional proportion of skiers/snowboarders that did not wear a helmet with a head injury?

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 for a categorical variable as the true proportion of observational units that are labeled as a “success” in the response variable.

7. 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.

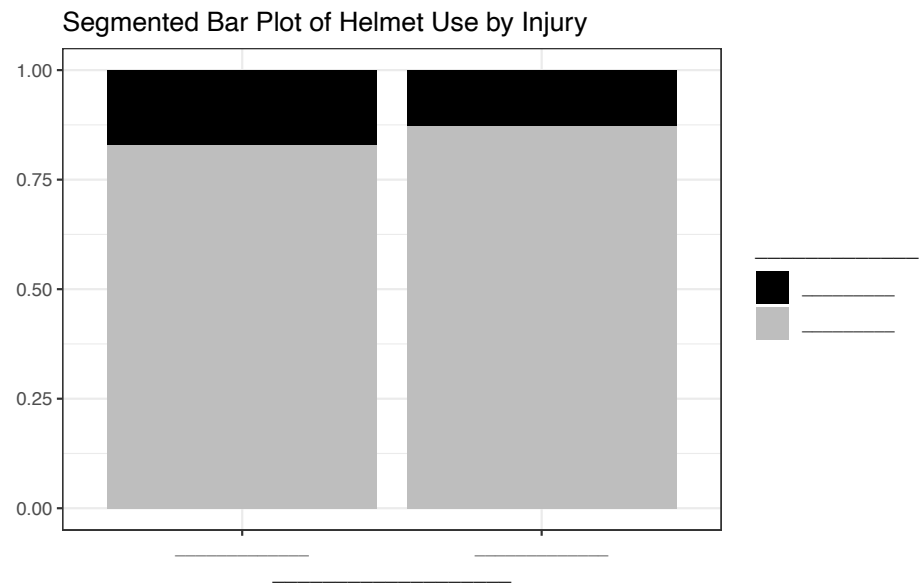
8. Write the null hypothesis out in words using your answers to question 7.

9. What is the research question?

10. Based on the research question fill in the appropriate sign for the alternative hypothesis ($<$, $>$, or \neq):

$$H_A : \pi_1 - \pi_2 \text{ _____ } 0$$

Summarize and visualize the data



- 11. Fill in the blanks on the graph with the appropriate variables and values to complete the segmented bar plot showing the proportion of head injuries between those who use helmets and those who do not use helmets. *Hint:* Use the conditional proportions from questions 5 and 6.
- 12. Based on the segmented bar plot, Does there appear to be an association between helmet use and head injury? Explain.
- 13. Calculate the point estimate for this study. Use helmet use minus no helmet use as the order of subtraction.
- 14. 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.

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

16. 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.

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}_{pool}(1 - \hat{p}_{pool})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \text{ where}$$
$$\hat{p}_{pool} = \frac{\text{number of "successes"}}{\text{number of cases}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

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

18. Calculate the test statistic.

We will use the `pnorm` function in R to find the p-value. Use the provided R markdown file and enter the value of the test statistic at `xx`.

```
pnorm(xx, # 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
```

19. Report the p-value.
20. 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.

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

The z^* multiplier is found under the standard normal distribution. We find the values that encompass the middle 95% of the distribution. If 95% of the standard normal distribution should be in the middle, that leaves 5% in the tails, or 2.5% in each tail. The `qnorm` function in R will tell us the z^* value for the desired percentile (in this case, 95% + 2.5% = 97.5% percentile).

```
qnorm(0.975) # Multiplier for 95% confidence interval
```

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

22. 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.

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

Table 7.3: Four different possible scenarios for hypothesis tests.

		Test conclusion	
		Fail to reject H_0	Reject H_0
	H_0 true	Good decision	Type 1 Error
Truth	H_A true	Type 2 Error	Good decision

24. Write a paragraph summarizing the results of the study. Be sure to describe:

- Summary statistic
- Test statistic and interpretation
- P-value and interpretation
- Conclusion (written to answer the research question)
- Confidence interval and interpretation
- 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.

Shown in the table above, 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.

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

267. What type of error could we have made?

27. Write this error in context of the problem.

7.4 Additional notes

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