Introduction to Inference

Learning Objectives

- Identify the two possible explanations (one assuming the null hypothesis, and one assuming the alternative hypothesis) for a relationship seen in sample data
- Given a research question, construct the null and alternative hypotheses in words and using appropriate statistical symbols
- Describe and perform simulation-based hypothesis tests
- Interpret and evaluate a p-value
- Use a confidence interval to determine the conclusion of a hypothesis test

Steps of Statisical Investigation

We will work through a six step process to complete a hypothesis test for a single proportion.

- **Ask a research question** that can be addressed by collecting data. What are the researchers trying to show.
- **Design a study and collect data**. This step involves selecting the people or objects to be studied and how to gather relevant data on them.
- **Summarize and Visualize the data**. Calculate summary statistics and create graphical plots that best represent the research question.
- **Use statistical analysis methods to draw inferences from the data**. Choose an analysis technique appropriate for the data and identify the p-value. In this study, we will focus on using randomization.
- **Communicate the results and answer the research question**. Using the p-value and confidence interval from the analysis, determine whether the data provide statistical evidence against the null hypothesis.
- **Revisit and look forward** to point out limitations of the study and suggest new studies that could be performed to build on the findings of the study

Handedness of Male Boxers

Left-handedness is a trait that is found in about 10% of the population. The fighting hypothesis states that left-handed men have an advantage in competition. Past studies have shown that left-handed men are over-represented among professional fighters. In this random sample of 500 male boxers we will see if there is an over-prevalence of left-handed fighters.

Summary Statistics Review

1. What are the observational units?

2. What variable are we testing? Is it categorical or quantitative?

3. What type of plot would be appropriate to visually display the data?

4. What statistic will we calculate to summarize the data?

Ask a Research Question.

5. Identify the research question for this study.

Design a Study and Collect Data

6. What is the target population for this study?

7. Did the researchers use a biased or an unbiased method of selection? Explain your answer.

Summarize and Visualize the Data

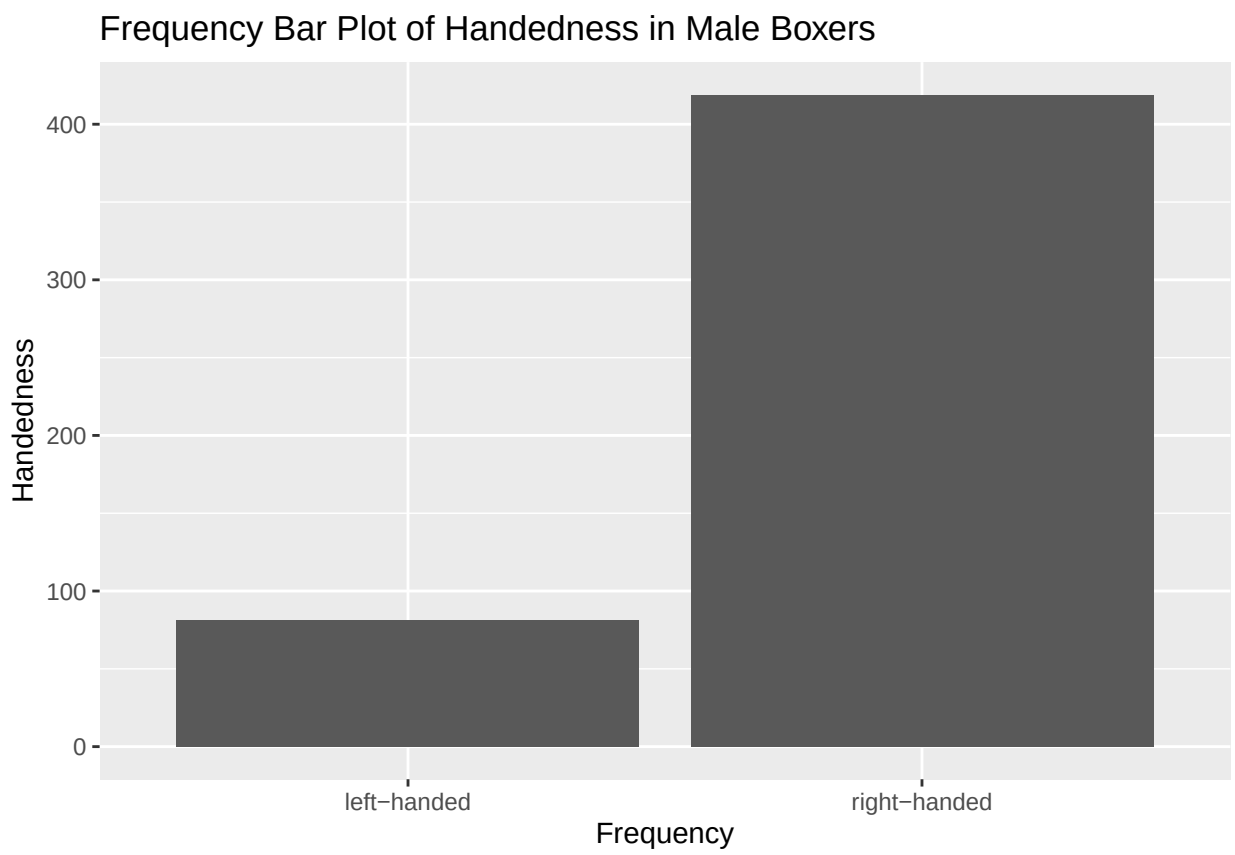
```
# Counts for Handedness
tally(~Stance, data=handedness_sub, margins=T)
```

```
## Stance
##  left-handed right-handed        Total
##           81          419          500
```

```
#Tally creates a table with a count for each level of the variable
```

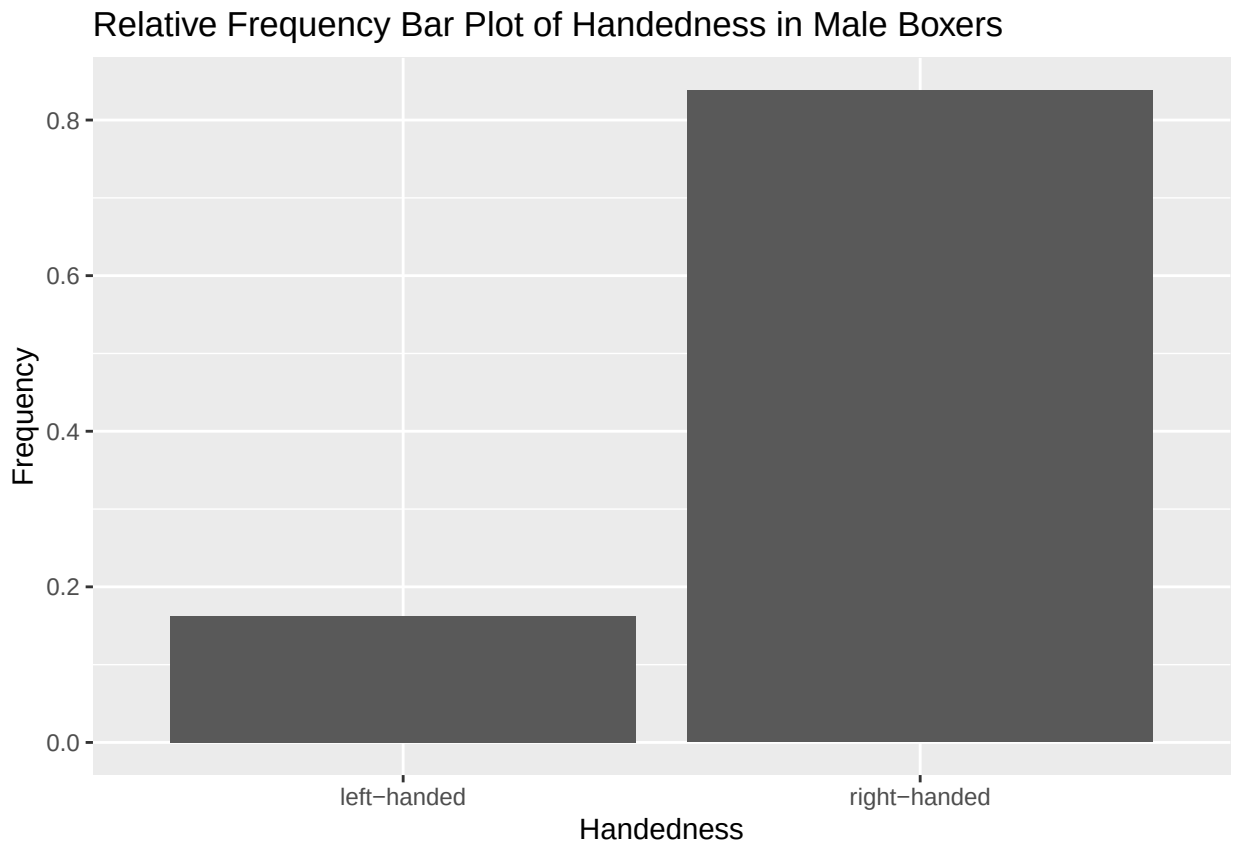
8. Calculate the appropriate summary statistic that represents the research question. Use appropriate notation.

```
ggplot(data = handedness_sub,   #This specifies the dataset
       aes(y = Stance)) +   #This specifies the variable
  geom_bar(stat = "count") +  #Tell it to make a bar plot
  labs(title = "Frequency Bar Plot of Handedness in Male Boxers",  #Give your plot a title
       x = "Handedness",   #Label the x axis
       y = "Frequency")  + #Label the y axis
  coord_flip()  #Turn the bars so they are vertical
```




```
ggplot(data = handedness_sub,   #This specifies the dataset
       aes(x = Stance)) +   #This specifies the variable
```

```
  geom_bar(aes(y = ..prop.., group = 1)) +  #Tell it to make a bar plot with proportions
  labs(title = "Relative Frequency Bar Plot of Handedness in Male Boxers",  #Give your plot a title
       x = "Handedness",   #Label the x axis
       y = "Frequency")  #Label the y axis
```



9. What is the difference between the two plots above?

Use statistical analysis methods to draw inferences from the data

When testing data we must first identify the null hypothesis. The null hypothesis is written about the parameter of interest, the true value of interest.

10. Write out the parameter of interest. (Hint: the true proportion of. . . .)

11. We will assume that the true proportion of male boxers who are left handed is the same as the general population, 0.1. Using the parameter of interest in question 5, write out the null hypothesis in words.

The notation used for a categorical parameter is, π. When writing the null hypothesis in notation we set the parameter equal to the null value, $H_0 : \pi = \pi_0$

12. Write the null hypothesis in notation using the null value of 0.1.

The alternative hypothesis is the claim to be tested and the direction is based on the research question.

13. Based on the research question, are we testing that the parameter is greater than 0.1, less than 0.1 or different than 0.1?

14. Write out the alternative hypothesis in words.

15. Write out the alternative hypothesis in notation.

Remember that when utilizing a hypothesis test, we are evaluating two competing possibilities. For this study the **two possibilities** are either...

- The true proportion of male boxers who are left handed is 0.1 and our results just occurred by random chance or
- The true proportion of male boxers who are left handed is greater than 0.1 and our results reflect this

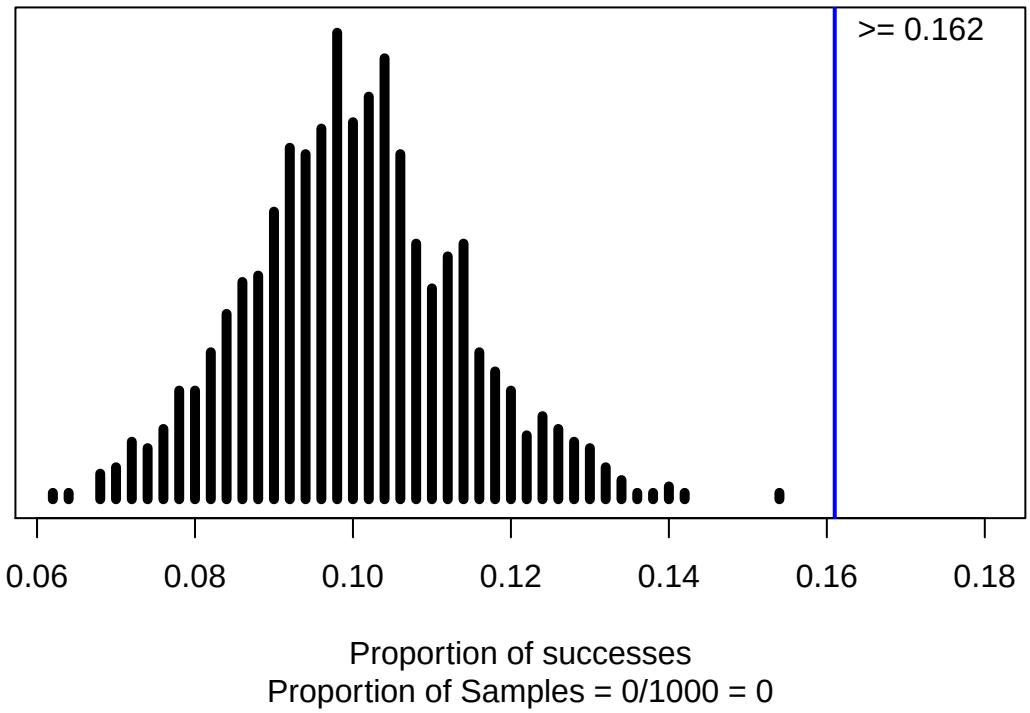
Notice that these two competing possibilities represent the null and alternative hypotheses.

The null distribution is created under the assumption the null hypothesis is true. In this case, we assume the true proportion of male boxers who are left handed is 0.1 so we will create 1000 different simulations of 500 boxers under this assumption.

To create one simulation we could have 50 blue cards and 450 red cards, where a blue card represents left-handed. Mix cards, draw 1 card, write down if it's red or blue, replace the card, repeat 499 times. The proportion of blue cards out of the 500 draws represents the simulated proportion of male boxers who are left handed and will be plotted on the simulated null distribution.

We will use the computer to simulate 1000 simulated proportions of male boxers who are left handed for a sample size of 500 based on the assumption that the true proportion of male boxers who are left handed is 0.1. This is called the null distribution because it is created based on the assumption that the null hypothesis is true.

```
one_proportion_test(probability_success = 0.1, #Null hypothesis value
                    sample_size = 500, #Enter sample size
                    number_repetitions = 1000, #Enter number of simulations
                    as_extreme_as = 81/500, #observed statistic
                    direction = "greater", #specify direction of alternative hypothesis
                    report_value = "proportion") #Reporting proportion or number of successes?
```



16. At what value is the null distribution centered? Why does that make sense?

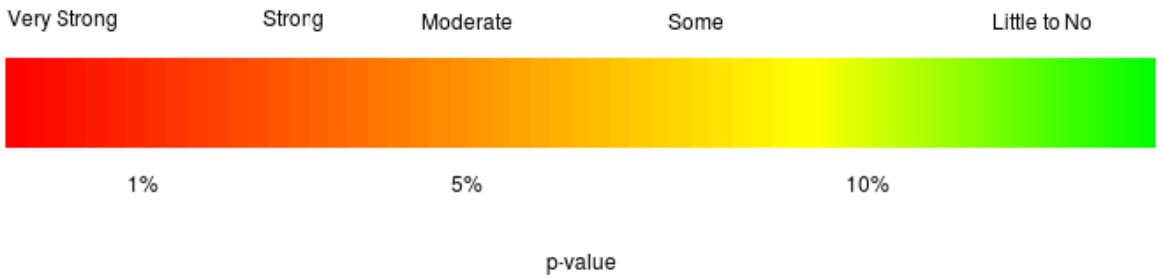
17. Where does the statistic (value from question 8) fall in the null distribution? Is it towards the center or in one of the tails?

18. Is the statistic likely to happen or unlikely to happen if the true proportion of male boxers is 0.1? Explain your answer.

19. Using the simulation, what is the probability that we find this summary statistic or greater, if the true proportion of male boxers is 0.1?

This is the p-value. The smaller the p-value the more evidence we have against the null hypothesis.

20. Using the following guidelines for the strength of evidence, how much evidence does the statistic provide against the null hypothesis?



21. Is there evidence that there is a higher proportion of male boxers that are left handed than the general population? Explain your answer.

Communicate the results and answer the research question

A point estimate provides a single plausible value for a parameter. However, a point estimate is rarely perfect; usually there is some error in the estimate. In addition to supplying a point estimate of a parameter, a next logical step would be to provide a plausible range of values for the parameter.

This plausible range of values for the population parameter is called a confidence interval.

To calculate a 95% confidence interval, we will build the interval around the point estimate.

$\hat{p} \pm SE(\hat{p})$ where $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

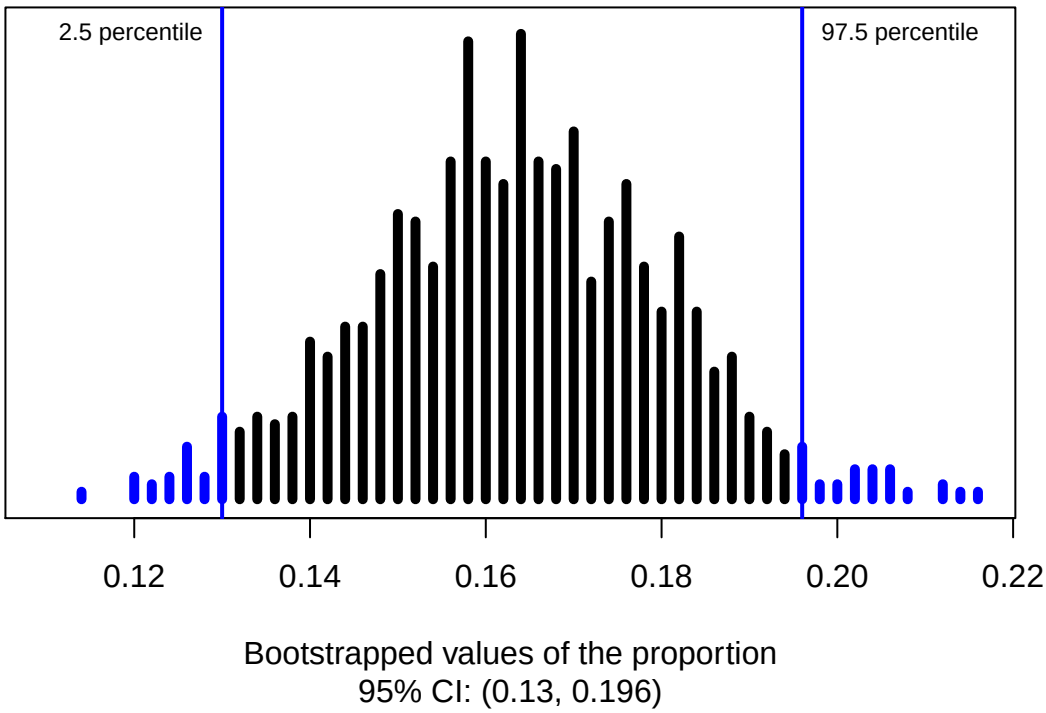
The standard error of $\hat{p}$ measures the uncertainty or variability associated with the point estimate.

When using a normal distribution to model the variability of a statistic we need to be sure certain conditions are not violated.

The sampling distribution for $\hat{p}$ based on a sample of size n from a population with a true proportion, π is nearly normal when:

1. The sample's observations are independent, e.g. are from a simple random sample.
2. We expected to see at least 10 successes and 10 failures in the sample, i.e. $n\pi \geq 10$ and $n(1-\pi) \geq 10$. This is called the success/failure condition.

```
one_proportion_bootstrap_CI(sample_size = 500, #Sample size
                    number_successes = 81, #Observed number of successes
                    number_repetitions = 1000, #Number of bootstrap samples to use
                    confidence_level = 0.95) #Confidence level as a decimal
```



22. Are the conditions met to use the normal approximation? Explain your answer.

23. Calculate the $SE(\hat{p})$.

24. Using the multiplier of 1.96 and the standard error found in question 23, calculate a 95% confidence interval.

25. What are we 95% confident is contained within this interval?

When we write a conclusion we answer the research question by stating how much evidence there is for the alternative hypothesis.

26. Write a conclusion to this test based on the p-value and confidence interval found.

Revisit and look forward

27. Suggest a new research question that you might investigate, building on what you learned in this study.