Introduction to Inference

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

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

- Frame the research question in terms of hypotheses. Hypothesis tests are appropriate for research questions that can be summarized in two competing hypotheses. The null hypothesis (H_0) usually represents a skeptical perspective or a perspective of no difference. The alternative hypothesis (H_A)usually represents a new view or a difference.
- Collect data with an observational study or experiment. If a research question can be formed into two hypotheses, we can collect data to run a hypothesis test. If the research question focuses on associations between variables but does not concern causation, we would run an observational study. If the research question seeks a causal connection between two or more variables, then an experiment should be used.
- Model the randomness as if the null hypothesis was true. We will use a computer to generate the null distribution from many different randomizations in order to quantify the null variability.
- Analyze the data. Choose an analysis technique appropriate for the data and identify the p-value. In this study, we will focus on using randomization.
- Form a conclusion. Using the p-value from the analysis, determine whether the data provide statistically significant evidence against the null hypothesis. Also, be sure to write the conclusion in plain language so casual readers can understand the results.

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 overrepresented among professional fighters. In this sample of 500 male boxers we will see if there is an overprevalence of left-handed fighters.

Summary Statistics Review

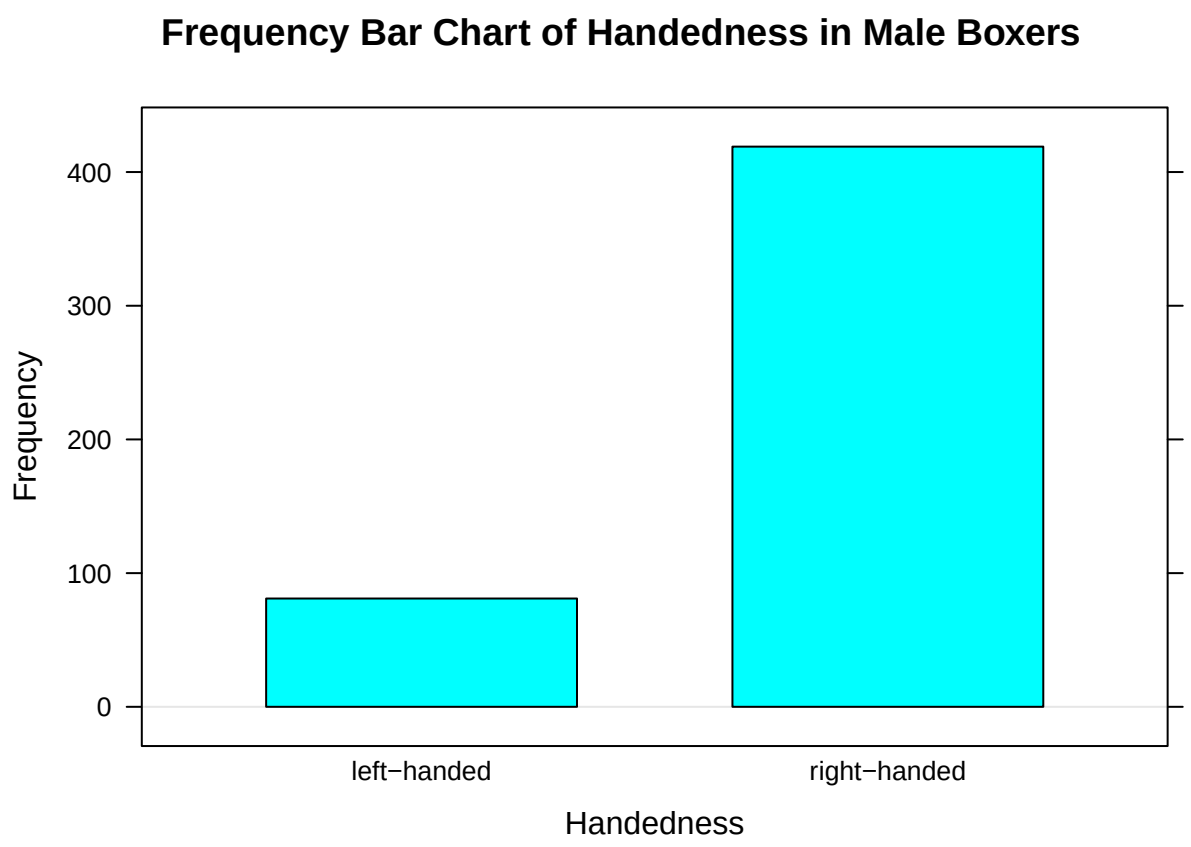
```
# Counts for Handedness
tally(~Stance, data=handedness_sub, margins=T) #Tally creates a table with a count in each level of the
```

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

```r
# Percentages of Handedness
tally(~Stance, data=handedness_sub, format="proportion", margins=T) #Using the format proportion create
```

```
## Stance
##  left-handed right-handed        Total
##        0.162        0.838        1.000
```

```r
barchart(handedness_sub$Stance, #This specifies the dataset and the variable
         horizontal = FALSE, #Turn the bars so they are vertical
         main = "Frequency Bar Chart of Handedness in Male Boxers", #Give your graph a title
         xlab = "Handedness", #Label the x axis
         ylab = "Frequency", #Label the y axis
         )
```




```r
barplot(table(handedness_sub$Stance)/nrow(handedness_sub), #divide the frequency counts by the total
        main = "Relative Frequency Bar Chart of Handedness in Male Boxers", #Give your chart a title
        xlab = "Handedness", #Label the x axis
        ylab = "Relative Frequency", #Label the y axis
        )
```

1. What are the observational units?

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

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

4. What proportion of Male Boxers in the sample are left-handed? Remember this is the statistic calculated from the data. Give the proper notation for this value.

Frame the research question.

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.

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

6. 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, p. When writing the null hypothesis in notation we set the parameter equal to the null value, $H_0 : p = p_0$

7. 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.

8. Identify the research question for this study.

9. 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?

10. Write out the alternative hypothesis in words.

11. 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 give evidence against the null hypothesis

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

Collect Data

12. What is the summary statistic for this data? Write using proper notation. **repeat question??**

Model the Randomness

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.

Add simulation here

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

14. Where does the statistic (value from question 3) fall in the null distribution? Is it towards the center or in one of the tails?

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

Analyze the data

16. Using the simulation, what is the probability that we find this summary statistic or greater, if the true proprotion 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.

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

Add image of guidelines

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

Form Conclusions

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.

Add conditions?

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

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

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

20. Using the multiplier of 1.96 and the standard error found in question 17, calculate a 95% confidence interval.

21. 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.

Conclusion as evidence against the null and statement amount alternative??? similar to 217