BIOS: 4120 Lab 2

January 23-24, 2018

Objectives

In today's lab we will:

1. Critically think through the study design of the 'diarrhea' dataset using the concepts learned in class
2. Review what we learned last week in lab by reading in the dataset 'diarrhea' into RStudio, and reporting and interpreting summary statistics
3. Briefly discuss hypothesis testing

Study Design

Prior to applying statistical methods to a dataset, it is important to first understand the study design, the question(s) of interest, and determine whether the study could have inherent biases and confounding factors that could reduce the confidence you have in the results. Therefore, it is a good habit to read the description of the dataset to understand the goals of the study and methods of data collection prior to analyses.

Navigate to the URL: http://myweb.uiowa.edu/pbreheny/4120/s18/data.html, find the dataset 'diarrhea', and click on the html link to find the description of the study.

Discussion Questions

1. What is the public health problem of interest? What is the study question or hypothesis?
2. What study design did the research team use? Describe what the name of this study design means.
3. What is the measured outcome of interest? What would you expect to happen if the treatment was effective?
4. How did the study reduce the possibility of biases? What might be some limitations to the results of the study?

Estimate vs Parameter

To reiterate what was explained in class a PARAMETER is a population quantity and is conventionally denoted with a greek letter such as π which represents the population proportion, μ which represents the mean of the population, and σ which represents the standard deviation of the population.

ESTIMATES are quantities derived from the sample (actual data) that we use to ESTIMATE (whoa, creative naming) the population quantities. Symbols that you will see that represent ESTIMATES are $\widehat{\pi}$ which represents the sample proportion, $\widehat{\mu}$ which is the sample mean, and $\widehat{\sigma}$ which is the sample standard deviation. Others you will see throughout the semester include p, $\bar{x}$, and s.

How would you find the $\widehat{\mu}$ for this dataset? How about the μ?

R Exercises

First open RStudio and a new script. Save the file to your H drive (or personal laptop) as Lab2 for future reference.

1. Set a variable named ‘diarrhea’ and read in the data using the read.delim function.

```
diarrhea<- read.delim("http://myweb.uiowa.edu/pbreheny/data/diarrhea.txt")
```

We used several functions last lab to describe our data. Listed below are some helpful functions:

head()= View the column names and first six lines of the data

summary()= Describes summary statistics by column

$= Access a single column within a dataset

mean()= Compute the average or mean

max()= Find the maximum value

min()= Find the minumum value

2. What are the variables in the dataset? How many observations are there?
3. Find the minimum, maximum, and average of the total stool volumes for all infants in the study. Hint: Use the symbol that allows you to access one column at a time. *Is the average you found a parameter or an estimate? Would you call the average a $\widehat{\mu}$ or μ?*

Split a Data Frame by Factors

4. Referring to the goal of the study, we want to know whether Bismuth salicylate was an effective treatment for diarrheal disease in children. Therefore, we want to compare stool volumes between the control and treatment group. How can we find the mean for each of the groups? You can use the ‘by’ function and specify that the mean of stool volumes should be given for the control and treatment group, respectively.

```
#Average stool volume for Control and Treatment Group

by(diarrhea$Stool, diarrhea$Group, mean)
```

Note: The third argument asks what function (FUN) you are interested in. You can replace ‘mean’ with other commands such as summary, min, and max.

Many times in R, you can get the same desired result using different approaches. There are simple and more complicated approaches to achieving the same goal… it just takes time to learn and select the best method. For example, you could also use the function ‘mean’ and subset the data using brackets [] and specify which group you would like the mean for. As you can see below, you will have to write two lines of code instead of one but you will get the same answer as above.

```
#Average stool volume for control group

mean(diarrhea$Stool[diarrhea$Group=="Control"])

#Average stool volume for treatment group

mean(diarrhea$Stool[diarrhea$Group=="Treatment"])
```

Compare the two means for the treatment and control group. Is this what you expected?

Hypothesis Testing - “Null Until Proven Alternative”

In class, you learned that there are a lot of wrong ways to think about p-values. The courtroom is a helpful example that illustrates the correct usage of p-values and hypothesis tests. Look at it in terms of “innocent until proven guilty”: As the person analyzing data, you are the judge. The hypothesis test is the trial, and the null hypothesis is the defendant. The alternative hypothesis is like the prosecution, which needs to make its case beyond a reasonable doubt (say, with 95% certainty).

If the evidence presented doesn’t prove the defendant is guilty beyond a reasonable doubt, you still have not proved that the defendant is innocent. But based on the evidence, you can’t reject that possibility.

So how would that verdict be announced? It enters the court record as “Not guilty.” That phrase is perfect: “Not guilty” doesn’t mean the defendant is innocent, because that has not been proven. It just means the prosecution couldn’t prove its case to the necessary, “beyond a reasonable doubt” standard. It failed to convince the judge to abandon the assumption of innocence.

If you follow that rationale, then you can see that “failure to reject the null” is just the statistical equivalent of “not guilty.” In a trial, the burden of proof falls to the prosecution. When analyzing data, the entire burden of proof falls to the sample data you’ve collected. Just as “not guilty” is not the same thing as “innocent,” neither is “failing to reject” the same as “accepting” the null hypothesis.

This method of thinking about hypothesis tests will come in handy when we start formally testing our own hypotheses.

Source: http://blog.minitab.com/blog/understanding-statistics/things-statisticians-say-failure-to-reject-the-null-hypothesis

What would be the null hypothesis for the ‘diarrheal’ dataset? What about the alternative?