**Calculating Confidence Intervals for Means and Proportions**

The purpose of this activity is to practice producing and interpreting confidence intervals for both means and proportions, both by hand and in R. The types of summaries that you use will depend on the type of data you have and what you want to learn. Often graphical summaries are most easily digestible, though numerical summaries are helpful as well. As you’ve seen and read, we have different numerical and graphical summaries for categorical and quantitative data, and still different summaries for univariate versus bivariate or multivariate data.

**CIs for Proportions**

**Example #1**:

**a.** Earlier in the course, we used the formula to determine the margin of error for a 95% confidence interval. That formula provides what is called the “conservative” estimate of the margin of error.



Suppose that we have a survey based on a randomly selected sample of 400 people. Find the value of the conservative estimate of the margin of error. Don’t convert the answer to a percent.

**b.** The technically precise determination of the margin of error for a 95% confidence interval is

Margin of error = 2 standard errors =. Suppose that we have a sample of *n* = 400 and the sample proportion with a trait of interest is = 0.5. Calculate.



**c.** Now suppose that we have n = 400 and = 0.8. Calculate



**d.** Now suppose that we have n = 400 and= 0.35. Calculate



**e.** Compare the answers to parts b, c, and d to the answer for part a.

(i) When does the margin of error given by the conservative formula (part a) match the more technically precise calculation?

(ii) What does the comparison of parts b, c and d to part a show about why is called a conservative margin of error.



**Example #2**:

A randomly selected sample of 900 U.S. adults is surveyed to determine the proportion that favors stricter penalties for drunk driving.

**a.** Describe the parameter of interest in this situation.

**b.** In the sample of 900 people, 657 individuals are in favor of stricter penalties for drunk driving.

Determine the sample proportion that favors stricter penalties.

**c.** Calculate the standard error of the sample proportion. The formula is .



**d.** Calculate a 95% confidence interval that estimates the population proportion in favor of stricter penalties for drunk driving. [Sample proportion ± 2 standard errors]

**e.** Write a sentence that interprets the 95% confidence interval calculated in the previous part.

**f.** On the basis of the 95% confidence interval computed in this activity, explain whether it is reasonable to conclude that more than 60% of American adults favor stronger penalties for drunk driving.

**Example #3**:

At the course website, access the folder for today’s lab, where I have placed the datafile called CIs\_ex3.R. On a Mac, you can double-click on this .R file directly. If R is not open, it will launch it and produce a line of code, and if it is already open, look in the console and see that it produced a line of code for you with the load command. The load command is the other way to bring a .R data file into R. On a Windows machine, you must use the load command. First, download the data (say to your Desktop), and then change the working directory of R to the location where you downloaded the data (File<Change dir…<). Then simply use

load("CIs\_ex3.R")

The data now exists in R, and to see what it is called, use the ls() command to get a list of all objects currently in R.

The data are from a survey (54 variables) of 1485 students from PSU in 2007.

**a.** The variable LegalMJ\_ asks respondents whether or not they believe marijuana should be legalized. Determine values for the following summary of the sample. (Don’t include missing data.) Use:

with(CIs\_ex3,table(LegalMJ\_))

**Response Number of students Percentage**

No

Yes

Total

**b.** Let R calculate a 95% CI for the proportion of PSU students who believe marijuana should be legalized. Using the number of “Yes” responses and the total number of responses above, we can do this as:

prop.test(# yes, total, correct=F)

The results give a 95% confidence interval that estimates the population proportion. Write the interval given by R, with values rounded to three decimal places.

**c.** Assuming that the sample represents all Penn State students, write a sentence that interprets the confidence interval determined in the previous part.

**d.**  Let R calculate a 99% CI for the proportion of PSU students who believe marijuana should be legalized. Using the number of “Yes” responses and the total number of responses above, we can do this as:

prop.test(# yes, total, correct=F,conf.level=.99), and report the CI below.

**e.** Now we’ll compare the responses of males and females. Use

with(CIs\_ex3,table(Gender,LegalMJ\_))

Give the following values:

Number of females in the sample (not counting missing data) =

Number of females who said “Yes” =

Proportion of females who said “Yes” =

Number of males in the sample (not counting missing data) =

Number of males who said “Yes” =

Proportion of males who said “Yes” =

**f.** Using R, determine a 95% confidence interval that estimates the proportion of all Penn State females who would say “Yes” to the legalization question (see part b above). Report the limits of the interval rounded to three decimal places

**g** Using R, determine a 95% confidence interval that estimates the proportion of all Penn State males who would say “Yes” to the legalization question (see part b above). Report the limits of the interval rounded to three decimal places

**h.** On the basis of the confidence intervals computed in parts f and g, can we conclude that at Penn State, males are more likely than females to favor the legalization of marijuana? Explain.

**i.** Now we’ll find a confidence interval for the difference between the proportions of females and males who said “Yes”. Use:

prop.test( c(# female yes, # male yes), c(total female, total male), correct=F)

The results include a 95% CI for the difference in proportions. Write that interval, rounding values to three decimal places.

**j.** Explain what we learn from the confidence interval in part i.

**CIs for Means**

**Example #4**: Using the same CIs\_ex3 data from above, let’s explore a quantitative variable: HrsStudy, which of course asks students how many hours they study per week.

**a.** We can get a confidence interval that estimates the unknown value of a population mean. We’ll get the interval and interpret it. Use

t.test(CIs\_ex3$HrsStudy)

The output will give a 95% confidence interval. Report the interval.

**b.** On the basis of the confidence interval calculated in part a, can we say that, on average, Penn State students study less than 15 hours per week? Explain.

**c**. Use R to estimate the mean amount that Penn State students spent on textbooks in the Spring 2007 semester. The relevant variable in the dataset is **TextSpd**.

What is the sample mean? What is the correct statistical notation for this sample mean?

Write the 95% confidence interval given on the output.

Write a sentence that interprets the 95% confidence interval in this case. The sentence structure is “We are 95% confident that the mean {describe the characteristic} for {describe the population} is between \_\_\_ and \_\_\_.”

**d.** In this part, we’ll compare the mean GPAs of female and male students at Penn State. The parameter of interest is the difference between means for females and males. Use the following lines of code:

with(CIs\_ex3,t.test(GPA~Gender))

Determine the following:

Sample mean GPA for females =

Sample mean GPA for males =

Sample difference between means for females and males =

95% confidence interval estimate of population difference =

**e.** On the basis of the interval found in part g, do you think that we can conclude the females at Penn state have a higher mean GPA than males? Explain.