Homework 4: Sections 3.1 - 3.4

STA209-04: Applied Statistics

February 15, 2019

From the Book:

Due Date: February 22, 2019
Questions: 3.31, 3.39, 3.63, 3.67, 3.74, 3.91, 3.133

3.31 One of the many wonderful things about studying statistics is that graduate programs in statistics often pay their graduate students, which means that many graduate students in statistics are able to attend graduate school tuition free with an assistantship or fellowship. In 2009, there were 82 US statistics or biostatistics doctoral programs for which enrollment data were available. The dataset StatisticsPhD lists all these schools together with the total enrollment of full-time graduate students in each program in 2009.

(a) Use *StatKey* or other technology to select a random sample of 10 of the 82 enrollment values. Indicate which values you've selected and compute the sample mean.

(b) Repeat part (a) by taking a second sample and calculating the mean.

(c) Find the mean enrollment for the entire population of these 82 graduate programs. Use correct notation for your answer. Comment on the accuracy of using the sample means found in parts (a) and (b) to estimate the population mean.

(d) Give a rough sketch of the sampling distribution if we calculate many sample means taking samples of size $n = 10$ from this population of enrollment values. What shape will it have and where will it be centered?

3.39 Exercise 3.37 tells us that 47 of the 303 inductees to the Rock and Roll Hall of Fame have been female or have included female members. The data are given in RockandRoll. Using all inductees as your population:

(a) Use *StatKey* or other technology to take many random samples of size $n = 10$ and compute the sample proportion that are female or with female members. What is the standard error for these sample proportions? What is the value of the sample proportion farthest from the population proportion of $p = 0.155$? How far away is it?

(b) Repeat part (a) using samples of size $n = 20$.

(c) Repeat part (a) using samples of size $n = 50$.

(d) Use your answers to parts (a), (b), and (c) to comment on the effect of increasing the sample size on the accuracy of using a sample proportion to estimate the population proportion.

3.63 A random sample of $n = 1483$ adults in the US were asked whether they consider a car a necessity or a luxury, and we find that a 95% confidence interval for the proportion saying that it is a necessity is 0.83 to 0.89. Explain the meaning of this confidence interval in the appropriate context.

3.67 A new study provides some evidence that playing action video games strengthens a person's ability to translate sensory information quickly into accurate decisions. Researchers had 23 male volunteers with an average age of 20 to look at moving arrays on a computer screen and indicate the direction in which the dots were moving. Half of the volunteers (11 men) reported playing action video games at least five times a week for the previous year, while the other 12 reported no video game playing in the previous year. The *response time* and the *accuracy score* were both measured. A

95% confidence interval for the mean response time for game players minus the mean response time for non-players is -1.8 to -1.2 seconds, while a 95% confidence interval for mean accuracy score for game players minus mean accuracy score for non-players is -4.2 to +5.8.

(a) Interpret the meaning of the 95% confidence interval for difference in mean response time.

(b) Is it plausible that game players and non-game players are basically the same in response time? Why or why not? If not, which group is faster (with smaller response time)?

(c) Interpret the meaning of the 95% confidence interval for difference in mean accuracy score.

(d) Is it plausible that game players and non-game players are basically the same in accuracy? why or why not? If not, which group is more accurate?

3.74 Suppose that a student is working on a statistics project using data on pulse rates collected from a random sample of 100 students from her college. She finds a 95% confidence interval for mean pulse rate to be (65.5, 71.8). Discuss how each of the statements below would indicate an *improper* interpretation of this interval.

(a) I am 95% sure that all students will have pulse rates between 65.5 and 71.8 beats per minute.

(b) I am 95% sure that the mean pulse rate for this sample of students will fall between 65.5 and 71.8 beats per minute.

(c) I am 95% sure that the confidence interval for the average pulse rate of all students at this college goes from 65.5 to 71.8 beats per minute.

(d) I am sure that 95% of all students at this college will have pulse rates between 65.5 and 71.8 beats per minute.

(e) I am 95% sure that the mean pulse rate for all US college students is between 65.5 and 71.8 beats per minute.

(f) Of the mean pulse rates for students at this college, 95% will fall between 65.5 and 71.8 beats per minute.

(g) Of random samples of this size taken from students at this college, 95% will have mean pulse rates between 65.5 and 71.8 beats per minute.

3.91 Researchers suspect that drinking tea might enhance the production of interferon gamma, a molecule that helps the immune system fight bacteria, viruses, and tumors. A recent study involved 21 healthy people who did not normally drink tea or coffee. Eleven of the participants were randomly assigned to drink five or six cups of tea a day, while 10 were asked to drink the same amount of coffee. After two weeks, blood samples were exposed to an antigen and production of interferon gamma was measured. The results are shown in Table 3.11 and are available in ImmuneTea. We are interested in estimating the effect size, the increase in average interferon gamma production for drinking tea when compared to coffee. Use *StatKey* or other technology to estimate the difference in mean production for tea drinkers minus cofee drinkers. Give the standard error for the difference and a 95% (bootstrap) confidence interval. Interpret the result in context.

3.133 In Exercise 3.70 on page 227, we describe a study in which participatns ate significantly more and exercise significantly less for a month. Two and a half years later, participants weighed an average of 6.8 pounds more than at the start of the experiment (while the weights of a control group had not changed). Is the amount of weight gained over the following 2.5 years directly related to how much weight was gained furing the one-month period? For the 18 participants, the correlation between increase of body weight during the one-month intervention and increase of body weight after 30 months is $r = 0.21$. We want to estimate, for the population of all adults, the correlation between weight gain over one month of bingeing and the effect of that month on a person's weight 2.5 years later.

(a) What is the population parameter of interest? What is the best estimate for that parameter?

(b) To find the sample correlation $r = 0.21$, we used a dataset containing 18 ordered pairs (weight gain over the one month and weight gain 2.5 years later for each individual in the study). Describe how to use this data to obtain one bootstrap sample.

(c) What statistic is recorded for the bootstrap sample?

(d) Suppose that we use technology to calculate the relevant statistic for 1000 bootstrap samples. Describe how to find the standard error using those bootstrap statistics.

(e) The standard error for one set of bootstrap statistics is 0.14. Calculate a 95% confidence interval for the correlation.

(f) Use the confidence interval from part (e) to indicate whether you are confident that there is a positive correlation between the amount of weight gain during the one-month intervention and amount of weight gained over the next 2.5 years, or whether it is plausible that there is no correlation at all. Explain your reasoning.

(g) Will a 90% confidence interval most likely be *wider* or *narrower* than the 95% confidence interval found in part (e)?

Miscellaneous:

Due Date: February 20, 2019

S1 Generate a list of *at least* ten questions about any concepts discussed during class or in lab over the past four weeks. In the textbook, this should be equivalent to any concept discussed in the first three chapters. We will spend some (or all) of class on Wednesday, February 20, answering these questions in preparation for the first exam, which is scheduled for Friday, February 22.