Data 310: Stat. Methods for PSUniversity of Pennsylvania
Spring 2021 Professor Marc Meredith

Homework Week #4

Question 1

(a) Consider a sample of 20 people for a random variable X which has a mean $\bar{x} = 4$ and a sample standard deviation $s_x = 6$. Estimate the symmetrical 95% confidence interval for this estimate using the t-distribution with the correct degrees of freedom.

(b) Calculate the 95% confidence interval instead using a normal distribution centered on $\bar{x}$ and a variance of s_x^2. Is this different then what you got above?

(c) What if we have two independent samples of 10 people each, where $\bar{x} = 5$ and $s_x = 2$; $\bar{y} = 6$ and $s_y = 2$. What is the 95% confidence interval around the difference in these two means? (You can calculate the degrees of freedom for the t distribution using the provided formula, but note that it will come out to 18.)

Question 2

(a) This question is going to have you further investigate some polling data. Upload "AZPoll-Fake.Rdata" into R. Note: this file began as real polling data from the NYT, but I've created a fake variable "clinton.thermometer" so that we have a continous measure to work with. So don't take too seriously the conclusions that we draw from this question!

(b) "clinton.thermometer" is a (again, fake) measure of how each respondent feels about Secretary Clinton, with 0 indicating that they feel very "cool" towards her, and 100 indicating they feel very "warm" towards her. Pretending for a moment that this is a simple random sample, calculate using our known equations: (i) the 95% confidence interval for "clinton.thermometer"; (ii) the 95% confidence interval for "clinton.thermometer" among those voting for clinton ("clinton"==1); (iii) the 95% confidence interval for "clinton.thermometer" among those not voting for clinton ("clinton"==0).

(c) Next, using the survey package, calculate the same three confidence intervals, but this time applying the provided weights. How does each estimate change? What does this tell us about the probability of those voting for each candidate ending up in our sample?

(d) Now perform a ttest using the survey package to find the confidence interval for the difference in means of "clinton.thermometer" for those who voted for Clinton and those who voted for Trump.

(e) Confirm for yourself that using the lm() command gives you the same answer for a difference as means as what you got in (d). What does the estimate for "(Intercept)" mean in this output? What does the estimate for "clinton" mean in this output? (Do not forget to include weights in the regression.)

(f) Create a scatterplot where clinton.thermometer is on the x-axis and final_weight is on the y-axis. Just looking at the data, are you able to form any conclusions about the relationship between these two variables?

(g) Add a regression line to your plot (don't worry about weighting), does this change your interpretation of the relationship?

(h) Consider the output of the regression where clinton.thermometer is the independent variable and final weight is the dependent variable.(Again, don't worry about weighting.) What is the interpretation of the Estimate for "clinton.thermometer" from this output? What is the the interpretation of "(Intercept)"? What sort of conclusion can you make about the sampling process for this survey based on this output?