Math 152 - Statistical Theory - Homework 8

write your name here

Due: Friday, October 16, 2020, midnight PDT

Important Note:

You should work to turn in assignments that are clear, communicative, and concise. Part of what you need to do is not print pages and pages of output. Additionally, you should remove these exact sentences and the information about HW scoring below.

Click on the *Knit to PDF* icon at the top of R Studio to run the R code and create a PDF document simultaneously. [PDF will only work if either (1) you are using R on the network, or (2) you have LaTeX installed on your computer. Lightweight LaTeX installation here: https://yihui.name/tinytex/]

> Either use the college's RStudio server (https://rstudio.pomona.edu/) or install R and R Studio on to your personal computer. See: https://research.pomona.edu/johardin/math152f20/ for resources.

Assignment

1: PodQ

Describe one thing you learned from someone in your pod this week (it could be: content, logistical help, background material, R information, etc.) 1-3 sentences.

2: Understanding confidence intervals.

(Thanks to Allan Rossman's askgoodquestions.blog !)

Assume that an alien has landed on Earth and wants to understand the gender diversity of humans. Fortunately, the alien took a good statistics course on its home planet, so it knows to take a sample of human beings and produce a confidence interval for the proportion of people who self-identify as female. Unfortunately, the alien happens upon the 2020 US Senate as its sample of human beings. The US Senate has 25 senators who self-identify as having a female gender (its most ever!) among its 100 members in 2020. (Ok, ok, I didn't personally ask any of them, but I do believe that in this particularly bonkers example, if I had been able to ask them, which is certainly the only way to know someone's gender, they would have reported female.)

(a) Calculate the alien's 95% confidence interval using the following formula (not everyone will have seen this formula, but it is the equivalent formula for the true proportion as the one we saw for the mean):

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

(b) Interpret the interval (include a clear statement for the value the interval is trying to capture).

(c) Is the interval consistent with you experience living on this planet?

(d) What went wrong?

(e) As we saw in class in the applet (http://www.rossmanchance.com/applets/ConfSim.html), about 5% of all 95% confidence intervals (i.e., 5% of samples of data) fail to capture the actual value of the population parameter. Is that the explanation for what went wrong here?

(f) Would it be reasonable for the alien to conclude, with 95% confidence, that between 16.5% and 33.5% of US senators in the year 2020 self-identify as female?

3: Wrong ideas

Explain what is wrong with each of the following statements.

a. The standard deviation of the bootstrap sampling distribution (of the statistic) will be a good approximation for the standard deviation of the original sample (the data).
b. The bootstrap distribution is created by resampling without replacement from the original sample.
c. When generating the resamples, it is best to use a sample size larger than the size of the original sample.
d. The bootstrap distribution is created by resampling with replacement from the population.
e. The main reason we bootstrap is in settings where the sample is small enough that the limit theorems don't hold.

4: R - blood pressure

Consider a large study of the association between blood pressure and cardiovascular disease that found:

- 55 out of 3338 men with high bp died of cardiovascular disease ($\hat{p}_1 = 0.0165$)
- 21 out of 2676 men with low bp died of cardiovascular disease ($\hat{p}_2 = 0.0078$)

The parameter of interest in this study is called the relative risk:

$$RR = \theta = \frac{P(\text{death from CD} \,|\, \text{high bp})}{P(\text{death from CD} \,|\, \text{low bp})}$$

$$\hat{\theta} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{0.0165}{0.0078} = 2.12$$

a. Why is bootstrapping a single time from the high bp group the same as taking 3338 random values from a Bernoulli distribution with probability 0.0165? (Or said differently, 1 binomial sample with `n=3338` and `p=0.0165`.)
b. Why is bootstrapping repeatedly (e.g., B = 500) from the high bp group the same as taking 500 binomial random values with `n=3338` and `p=0.0165`?
c. Use the code below to fill in the necessary values (also, change to `eval = TRUE` so that the code will run).
d. Create three histograms (use `hist()`) describing (1) bootstrapped statistic, (2) bootstrapped SEs, (3) bootstrapped T values.
e. Find three 95% intervals (note the `qnorm()` and `quantile()` functions in R):
   i. normal CI with BS SE
   ii. BS-t CI (don't forget to subtract the 0.975 quantile on the lower end)
   iii. BS percentile interval
f. Based on the distributions (histograms) seen above, interpret the intervals (using words like relative risk and blood pressure). Also, comment on whether or not each of the three intervals seems appropriate.
g. Repeat c, d, e, f for a new statistic: $\ln(RR)$. n.b. the correct R function is `log()`.

$$\hat{\theta} = \ln\left(\frac{\hat{p}_1}{\hat{p}_2}\right) = \ln\left(\frac{0.0165}{0.0078}\right) = \ln(0.0165) - \ln(0.0078) = 0.7418$$

```r
B = ____
M = ____

highbp.bs <- rbinom(___, ___, ___)  # print this to see what it looks like!
lowbp.bs <- rbinom(___, ___, ___)
```

```
RR.bs <- _______  # this should be a vector B long

RR.SE.bs <- c()  # will eventually be a vector that is B long

for(b in 1:B){
  highbp.bsbs <- rbinom(____, ____, prob from b^(th) element in highpb.bs)
  lowbp.bsbs <- ____

  RR.bsbs <- ______

  RR.SE.bs <- c(RR.SE.bs, ____) # keep the SE of the statistic from the double BS

}

T.bs <- ____
```