Lab 3: Bootstrapping and Confidence Intervals
KEY

Javier E. Flores

February 15, 2019

Total Possible Points: 41

Bootstrapping

Q1) [2pts] Why do bootstrap samples need to be drawn with replacement? What would happen if the bootstrap samples were drawn without replacement?

A1) The prime motivation for bootstrapping is to simulate the sampling distribution of a statistic. This being said, the case for why we draw with replacement is best made when we consider what happens when we draw without replacement. Knowing that our bootstrap samples are of the same size as our original sample, drawing without replacement implies that we draw our same original sample for every bootstrap sample. When we then compute the statistic of interest for each bootstrap sample, we get the same value every single time. As a result, the distribution is **degenerate**, meaning that it only consists of a single value (this is bad). When we sample with replacement, each bootstrap sample is likely different from the last. This induces variability among statistics computed from bootstrap samples thereby creating a proper distribution. [2 pts]

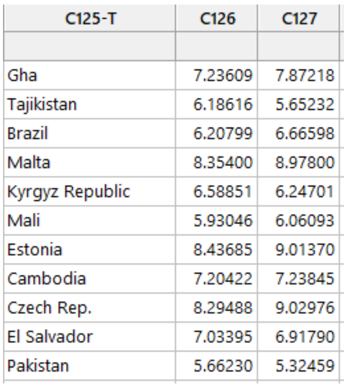
Bootstrap Confidence Intervals

Q2) [2 pts] Find the true population mean human freedom score and the correlation between human freedom score and personal freedom score. In your writeup, express these population parameters using the proper notation (i.e. the greek symbols we referenced in class).

A2) Using Minitab, the computed mean for human freedom score was $\mu = 6.89$. [1 pts] The computed correlation between human freedom score and personal freedom score was $\rho = 0.95$. [1 pts]

Q3) [2 pts] Using Minitab, obtain a random sample of size 40 from the HFI Data. Paste the names of the first 10 countries in your sample into your writeup.

A3) Using Minitab, the following countries were the first 10 contained in my random sample of 40. This will vary from group to group. [2 pts]



C125-T	C126	C127
Gha	7.23609	7.87218
Tajikistan	6.18616	5.65232
Brazil	6.20799	6.66598
Malta	8.35400	8.97800
Kyrgyz Republic	6.58851	6.24701
Mali	5.93046	6.06093
Estonia	8.43685	9.01370
Cambodia	7.20422	7.23845
Czech Rep.	8.29488	9.02976
El Salvador	7.03395	6.91790
Pakistan	5.66230	5.32459

Q4) [2 pts] Find the best estimates of the mean human freedom score and the correlation between human freedom score and personal freedom score.

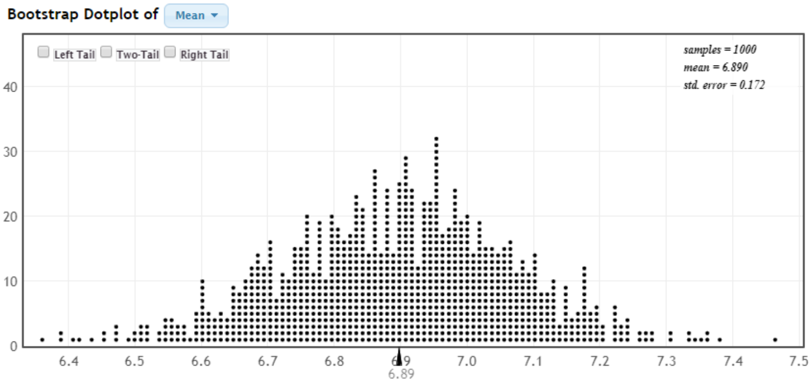
A4) Using Minitab, we use our sample to compute the sample mean human freedom score to obtain $\bar{x} = 6.89$ [1 pts] and sample correlation $r = 0.95$ (yes these are the values my sample actually gave me). [1 pts]

Q5) [2 pts] Use StatKey to construct a 95% bootstrap confidence interval for the mean human freedom score using your sample of size 40. Include a copy of the bootstrap distribution you generated in StatKey along with any other work done.

A5) Using our sample statistic (6.89) and the standard error from below, our interval is:

$$6.89 \pm 2 * 0.172 = (6.546, 7.234)$$

[1 pts]



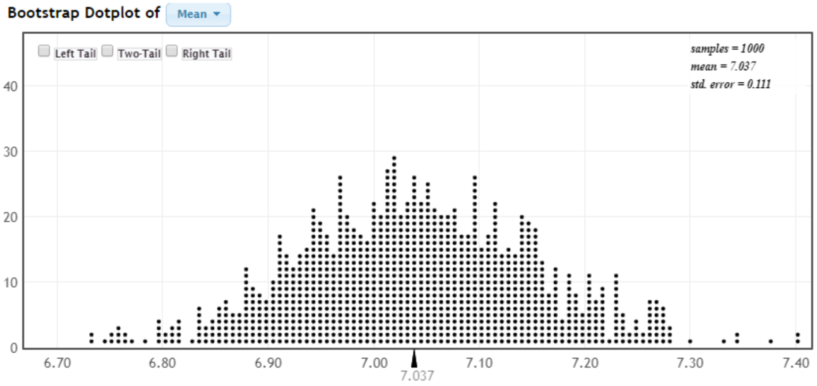
[1 pts]

Q6) [2 pts] Repeat this process in StatKey but with a sample of size 80. Include a copy of the bootstrap distribution you generated in StatKey along with any other work done.

A6) Using our sample statistic (7.03) and the standard error from below, our interval is:

$$7.03 \pm 2 * 0.111 = (6.808, 7.252)$$

[1 pts]



samples = 1000
mean = 7.037
std. error = 0.111

[1 pts]

Q7) [2 pts] Compare the widths of the intervals generated in the previous two questions. Which is wider? Why? Explain in no more than three sentences.

A7) The width of the first interval is $2 * (2 * 0.172) = 0.688$ and the width of the second interval is $2 * (2 * 0.111) = 0.444$. The first interval is wider. [1 pts] This is explained by the fact that the size of each bootstrap sample was smaller in the first case relative to the second. [1 pts]

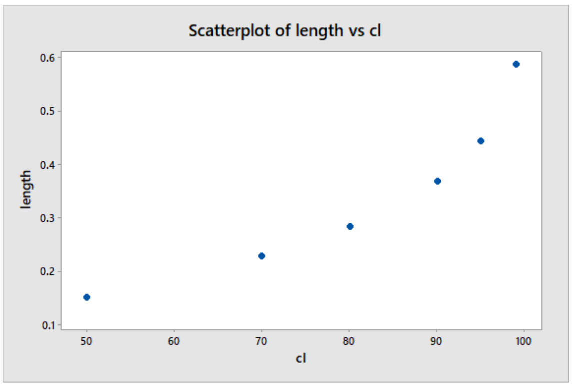
Q8) [6 pts] Fill out the following table relating confidence level and interval width using the percentile bootstrap approach on your most recent sample ($n = 80$) to construct interval estimates for the mean human freedom score.

A8) Give 1 point for each row filled out. [6 pts]

Confidence Level:	**Interval: (A,B)**	**Length: B-A**
50%	(6.962, 7.112)	0.150
70%	(6.924, 7.151)	0.227
80%	(6.900, 7.182)	0.282
90%	(6.857, 7.225)	0.368
95%	(6.818, 7.260)	0.442
99%	(6.750, 7.336)	0.586

Q9) [2 pts] Create a scatterplot relating the confidence level and length from the completed table in the previous question. Is the relationship linear or non-linear? Describe the relationship in no more than three sentences. Include the scatterplot in your response.

A9) The relationship is not linear. It appears exponential. To see this, focus on the rightmost datapoints. [1 pts]



[1 pts]

Q10) [5 pts] Based upon the various bootstrap intervals we've investigated in this lab, fill out the following table summarizing the impact of changing various factors on confidence interval width:

A10) Give 1 point for each row filled out. [5 pts]

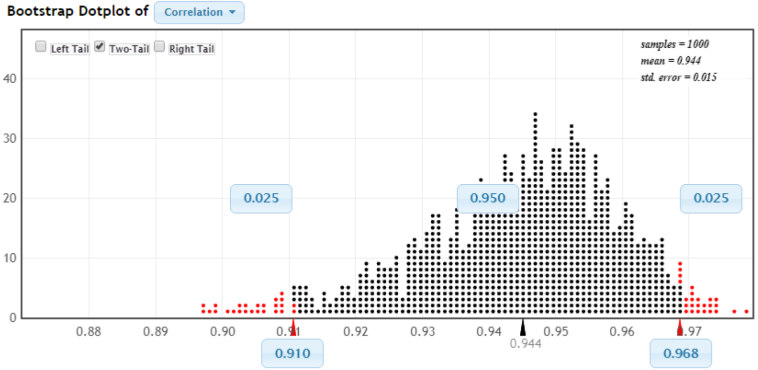
Change:	**Impact: wider/narrower/negligible?**
Increase n	narrower
Increase number of bootstrap samples	negligible
Increase confidence level	wider
Increase standard error	wider
Decrease confidence level	narrower

Q11) [2 pts] Given what you've seen, do you think that all values in a 95% confidence interval are equally plausible? Explain your answer in 1-2 sentences.

A11) All values in the interval are not equally plausible. Remember that our most likely estimate is towards the center of our sampling distribution, where data are most concentrated. Larger width intervals created as a result of increased confidence levels incorporate more of the tails of our bootstrap distribution. Datapoints in the extreme (tails) are not as frequently occurring as those towards the center. [2 pts]

Q12) [4 pts] Our investigation thus far has largely centered around the mean. Create either a percentile or standard error 95% bootstrap confidence interval of the correlation between personal freedom score and human freedom score. Choose one or other other depending on what you observe about the bootstrap distribution. Include a plot of your bootstrap distribution in your writeup.

A12) The bootstrap distribution I obtained was left-skewed. As a result, I obtained the 95% percentile bootstrap confidence interval (0.910, 0.968). [2 pts]



[2 pts]

Q13) [8 pts] When introducing the HFI dataset, I mentioned the data contained a vast number of indicators of personal and economic freedom. Visit this site for the codebook containing all variables included in the HFI dataset (scroll through the "Columns" section). Using any two of these variables, formulate a research question requiring you to use either correlation/regression or a difference in means. In your writeup, include 3-5 sentences addressing:

a) Your choices of variables and the research question

b) Your sample size

c) Your bootstrap confidence interval

d) Your interpretation of the interval in the context of your research question

A13) Several answers possible. Award 2 points for every completed component. [8 pts]