PS4

BSS Stat 20

2022-07-18

NOTICE

This problem set is worth zero points; you do not have to turn it in. However, if you want to assess yourself fairly, go ahead and type your answers in an RMarkdown file, knit and save to an html or pdf. Feel free to collaborate or use any kind of help while completing this assignment.

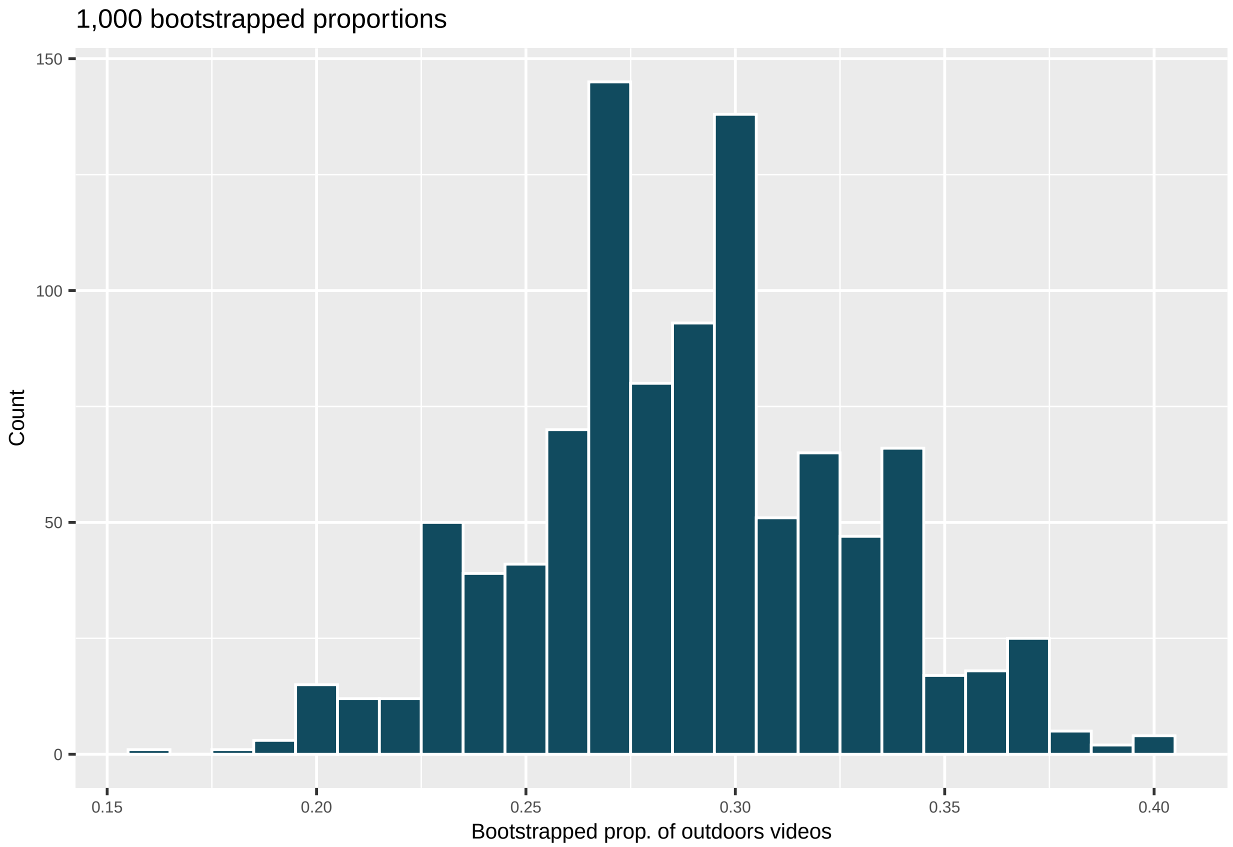
This notice will be pasted at the beginning of every problem set.

Questions

Question 1 - Outside YouTube videos

Let's say that you want to estimate the proportion of YouTube videos which take place outside (define "outside" to be if any part of the video takes place outdoors). You take a random sample of 128 YouTube videos[1] and determine that 37 of them take place outside. You'd like to estimate the proportion of all YouTube videos which take place outside, so you decide to create a bootstrap interval from the original sample of 128 videos.

[1] There are many choices for implementing a random selection of YouTube videos, but it isn't clear how "random" they are.



a. Describe in words the relevant statistic and parameter for this problem. If you know the numerical value for either one, provide it. If you don't know the numerical value, explain why the value is unknown.

b. What notation is used to describe, respectively, the statistic and the parameter?

c. If using software to bootstrap the original dataset, what is the statistic calculated on each bootstrap sample?

d. When creating a bootstrap sampling distribution (histogram) of the bootstrapped sample proportions, where should the center of the histogram lie?

e. The histogram provides a bootstrap sampling distribution for the sample proportion (with 1000 bootstrap repetitions). Using the histogram, estimate a 90% confidence interval for the proportion of YouTube videos which take place outdoors.

f. In words of the problem, interpret the confidence interval which was estimated in the previous part.

Question 2 - Identifying types of Hypotheses

For each of the research statements below, note whether it represents a null hypothesis claim or an alternative hypothesis claim.

a. The number of hours that grade-school children spend doing homework predicts their future success on standardized tests.

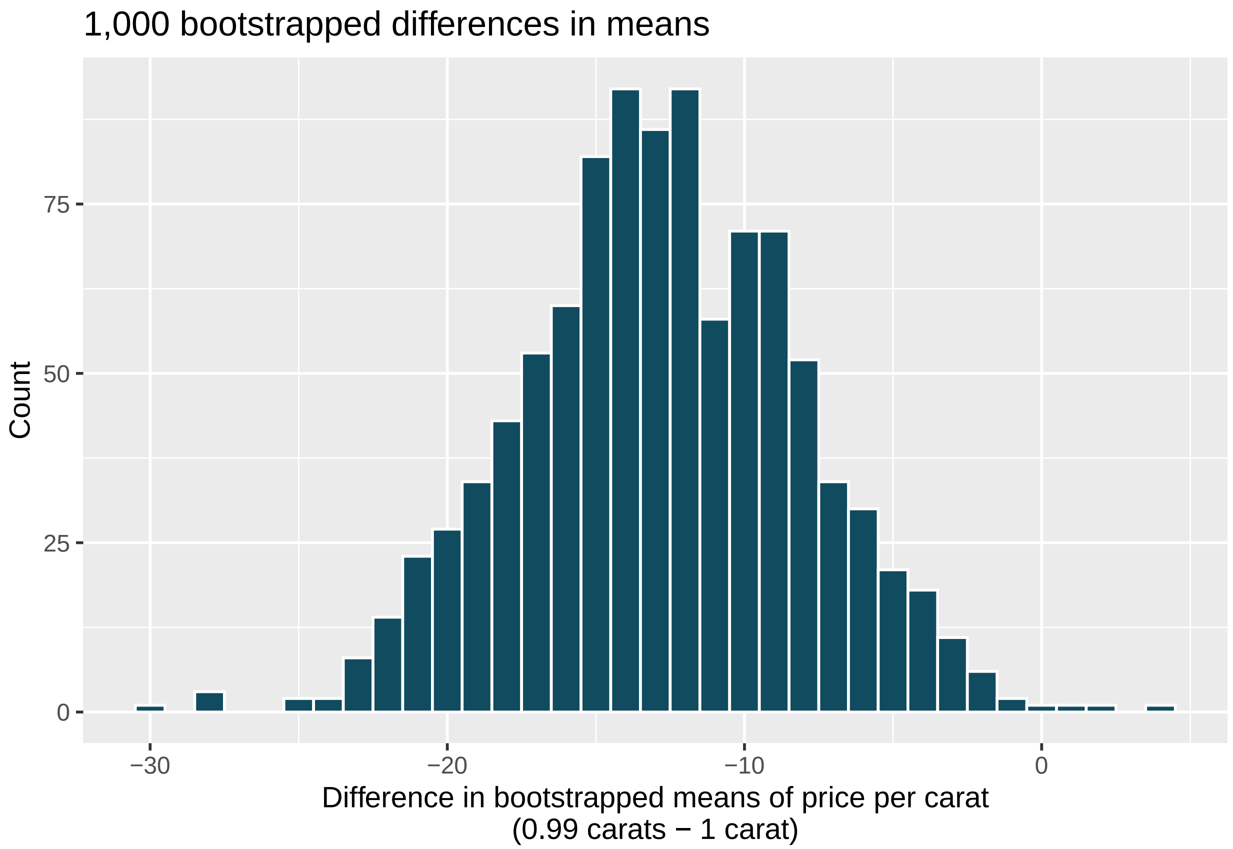
b. King cheetahs on average run the same speed as standard spotted cheetahs.

c. For a particular student, the probability of correctly answering a 5-option multiple choice test is larger than 0.2 (i.e., better than guessing).

d. The mean length of African elephant tusks has changed over the last 100 years.

Question 3 - Diamonds

We have data on two random samples of diamonds: one with diamonds that weigh 0.99 carats and one with diamonds that weigh 1 carat. Each sample has 23 diamonds. Provided below is a histogram of bootstrap differences in means of price per carat of diamonds that weight 0.99 carats and diamonds that weigh 1 carat.



Using the bootstrap distribution, create a (rough) 95% bootstrap percentile confidence interval for the true population difference in prices per carat of diamonds that weigh 0.99 carats and 1 carat. Interpret the interval in the context of the this problem.

Question 4 - Basketball (again!)

In the 2016 NBA season, it was noted that professional basketball player Stephen Curry had a remarkable basket-shooting performance beyond 30 feet. The `curry` data frame (available on the course server) contains his long range shooting performance across 27 attempts. By comparison, the long range shooting percentage of NBA players that season was 7.5%.

In this question, you will assess whether this data is consistent with the notion that Stephen Curry has a long range shooting percentage that is no different from the average NBA player. Said another way, you will assess just how remarkable this Curry's shooting performance was in 2016.

a. Write the null and alternative hypothesis.

b. Compute the observed test statistic.

c. Visualize the observed data using an appropriate plot.

d. Construct a plot featuring 9 subplots, each one featuring a visualization of a data set generated under the null hypothesis. Does your visualization of the observed data from the previous part look like it could be one of these plots?

e. Construct and save the null distribution of statistics.

f. Visualize the null distribution.

g. Compute the p-value.

h. Interpret the p-value. What does it say about the consistency between the null hypothesis and the observed data?