Instructions

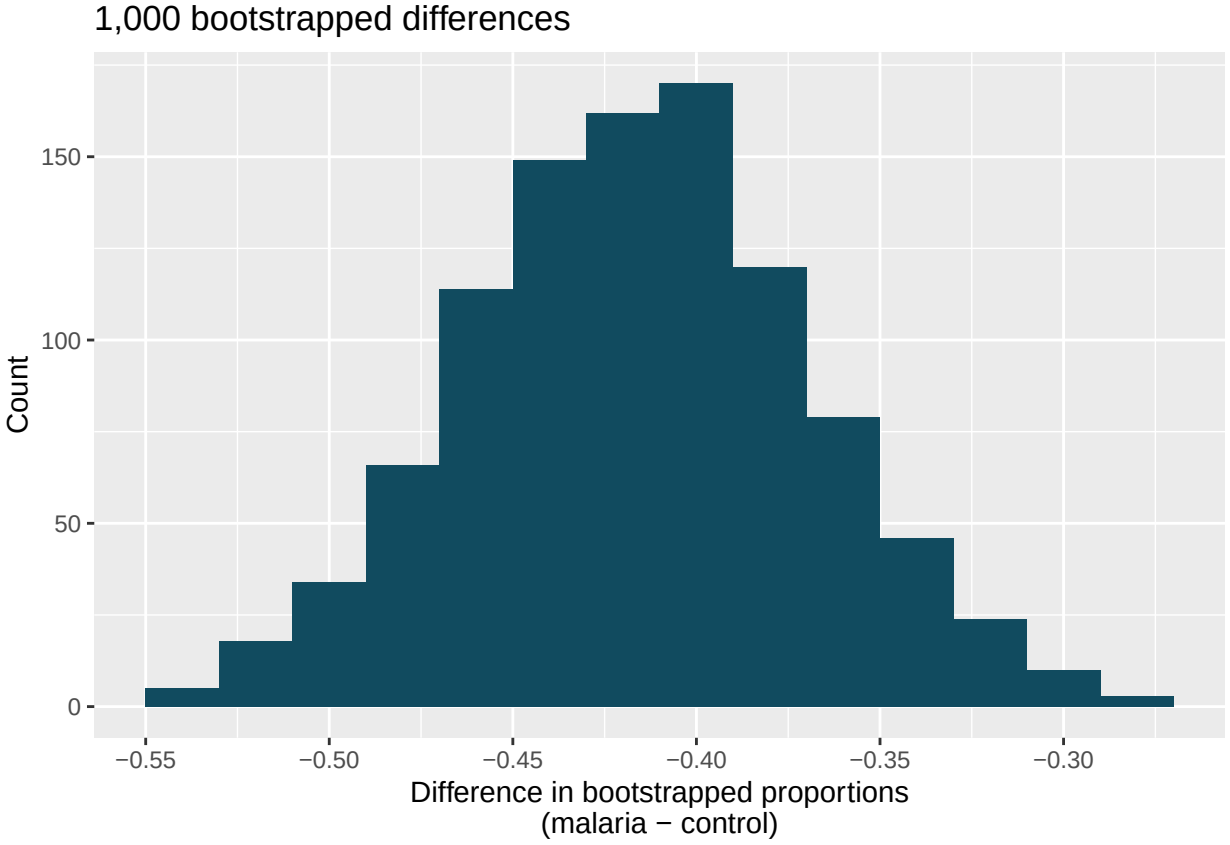
Upload a PDF file, named with your UC Davis email ID and homework number (e.g., sfrei_hw5.pdf), to Gradescope (accessible through Canvas). You will give the commands to answer each question in its own code block, which will also produce output that will be automatically embedded in the output file. All code used to answer the question must be supplied, as well as written statements where appropriate.

All code used to produce your results must be shown in your PDF file (e.g., do not use `echo = FALSE` or `include = FALSE` as options anywhere). Rmd files do not need to be submitted, but may be requested by the TA and must be available when the assignment is submitted.

Students may choose to collaborate with each other on the homework, but must clearly indicate with whom they collaborated.

Problem 1: [IMS] 17.2 Malaria vaccine effectiveness (variant)

With no currently licensed vaccines to inhibit malaria, good news was welcomed with a recent study reporting long-awaited vaccine success for children in Burkina Faso. With 450 children randomized to either one of two different doses of the malaria vaccine or a control vaccine, 89 of 292 malaria vaccine and 106 out of 147 control vaccine children contracted malaria within 12 months after the treatment.



Using the entire bootstrap distribution, find a 95% bootstrap percentile confidence interval for the true difference in proportion of children who contract malaria (malaria vaccine minus control vaccine) in the population. Interpret the interval in the context of the problem.

We can use the empirical distribution to estimate the 95% confidence interval by finding the left 2.5th percentile and 97.5th percentile. Since there are 1,000 bootstrapped differences, this corresponds to those differences where about 25 samples lie to the left, and where 25 samples lie to the right. Looking at the plot, this looks to be around -0.50 and -0.33. So we could say a 95% confidence interval for the difference in proportions is (-0.50, -0.33).

The confidence interval doesn't cover zero so it means there is significant evidence of a difference between malaria and control group and the proportions of malaria vaccines are smaller than that of control vaccines with confidence level as 95%. That is, we can be quite confident that the vaccine is effective at reducing malaria.

Problem 2: [IMS] 17.10

According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents. Calculate a 95% confidence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived and interpret it in context of the data. (CDC, 2008)

Solution: Before constructing the confidence interval, let us check the conditions needed are satisfied.

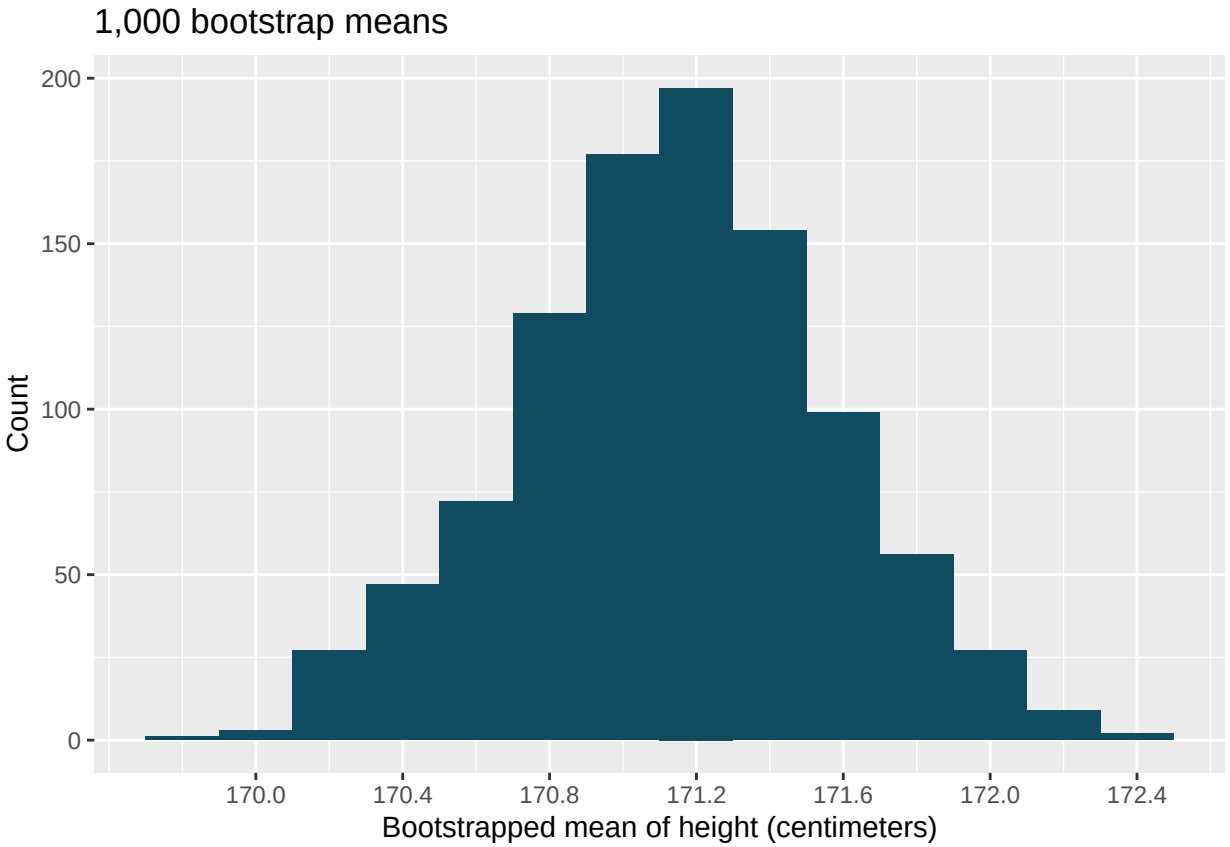
Since the data are randomly sampled, we can assume they are independent. The success-failure condition is clearly met, so we can therefore expect $\hat{p}_{CA} - \hat{p}_{OR}$ is approximately normal. A 95% confidence interval can thus be calculated as,

$$\begin{aligned}(\hat{p}_{CA} - \hat{p}_{OR}) \pm z^\star \sqrt{\frac{\hat{p}_{CA}(1-\hat{p}_{CA})}{n_{CA}} + \frac{\hat{p}_{OR}(1-\hat{p}_{OR})}{n_{OR}}} &= (0.08 - 0.088) \pm 1.96 \sqrt{\frac{0.08 \times 0.92}{11,545} + \frac{0.088 \times 0.912}{4,691}} \\ &= -0.008 \pm 0.009 \\ &= (-0.017, 0.001)\end{aligned}$$

We are 95% confident that the difference between proportions of Californians and Oregonians who are sleep deprived is between -1.7% and 0.1%. In other words, we are 95% confident that 1.7% less to 0.1% more Californians than Oregonians are sleep deprived.

Problem 3: [IMS] 19.4

Researchers studying anthropometry collected body measurements, as well as age, weight, height and gender, for 507 physically active individuals. The histogram below shows the sample distribution of bootstrapped means from 1,000 different bootstrap samples.

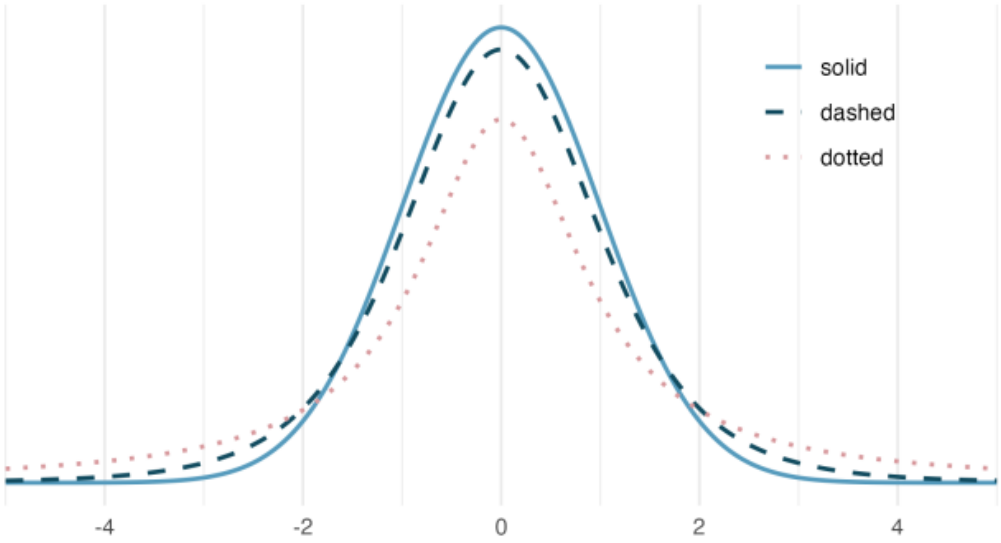


By looking at the bootstrap sampling distribution (1,000 bootstrap samples were taken), find an approximate 90% bootstrap percentile confidence interval for the true average adult height in the population from which the data were randomly sampled. Provide the interval as well as a one-sentence interpretation of the interval.

Answer

There are 1000 bootstrap samples, so the center 90% of the bootstrap means will leave 50 bootstrap means in the left tail and 50 bootstrap means in the right tail. A rough approximation (using count on the y-axis) leads to a 90% confidence interval for the true average gestation in the population of (170.3 cm, 171.8 cm). We can be 90% confident that, in the population from which the data were randomly sampled, the true mean height is between 170.3 cm to 171.8 cm

Problem 4 The figure below shows three unimodal and symmetric curves: the standard normal (z) distribution, the t-distribution with 5 degrees of freedom, and the t-distribution with 1 degree of freedom. Determine which is which, and explain your reasoning.



We can compare the tails, when the degree of freedom increases in t-distribution, the tail will be lighter and the normal distribution has the lightest tail so it can be determined by the tails. By a "light" tail we mean less large values on the y-axis when the x value is large. Thus the solid line is standard normal, the dashed line is t distribution with 5 degrees of freedom, and the dotted line is t distribution with 1 degree of freedom.

Problem 5 In this exercise we work with a random sample of 1,000 cases from the dataset released by the United States Department of Health and Human Services in 2014. Provided below are sample statistics for gestation (length of pregnancy, measured in weeks) of births in this sample.

Min	Q1	Median	Mean	Q3	Max	SD	IQR
21	38	39	38.7	40	46	2.6	2

a. What is the point estimate for the average length of pregnancy for all women?

It's the mean, 38.7 weeks.

b. You might have heard that human gestation is typically 40 weeks. Using the data, describe why the assumptions required to perform a hypothesis test are satisfied, perform a complete hypothesis test, using mathematical models, to assess the 40 week claim. State the null and alternative hypotheses, find the T score, find the p-value, and provide a conclusion in context of the data.

Answer Null hypothesis: $H_0 : \mu = 40$, alternative: $H_A, \mu \neq 40$, where μ is the average length of pregnancy in weeks.

The null hypothesis is that the average length of pregnancy in the population is 40 weeks; the alternative hypothesis is that the average length of pregnancy in the population is not 40 weeks.

T score: (38.7 - 40) / (2.6 / sqrt(1000)) = -15.8.

P-value is extremely small - software will report $<$ 2e-16 or equivalent.

Conclusion: Reject 40 weeks as the average length of pregnancy for all births from which these data were randomly selected.

Problem 6 A 90% confidence interval for a population mean is (65, 77). The population distribution is approximately normal and the population standard deviation is unknown. This confidence interval is based on a simple random sample of 25 observations. Calculate the sample mean, the margin of error, and the sample standard deviation. Assume that all conditions necessary for inference are satisfied. Use the t-distribution in any calculations.

Answer We know that the sample mean is (65+77)/2=71, since the sample mean is the point estimate and thus the center for the confidence interval. The margin of error is thus 6, since the confidence interval takes the form point estimate +/- standard error.

We can derive the sample standard deviation from the margin of error. The margin of error is equal to $t^*_{24} \times SE$. The t^*_{24} is the value corresponding to the 95th percentile for the t distribution with 24 degrees of freedom, since we have a 90% confidence interval and 25 samples. We can calculate this using R,

```
qt(0.95, df=24)
```

```
[1] 1.710882
```

The standard error is equal to the sample standard deviation divided by $\sqrt{n}$, so we have

$$MarginError = 6 = t^* \times SE = 1.71 \times s/\sqrt{25}$$

Solving for s we get $s = 5 \times 6/1.71 = 17.5$.