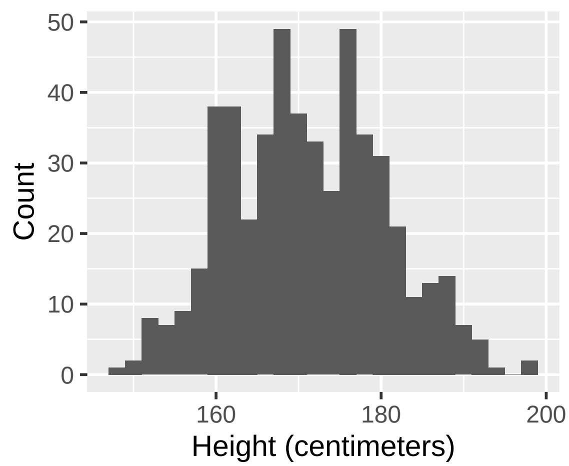
Problem Set 9

Inference for Numerical Data

1. **Heights of adults.** Researchers studying anthropometry collected body measurements, as well as age, weight, height and gender, for 507 physically active individuals. Summary statistics for the distribution of heights (measured in centimeters), along with a histogram, are provided below.

Min	Q1	Median	Mean	Q3	Max	SD	IQR
147.2	163.8	170.3	171.1	177.8	198.1	9.4	14



   a. What is the point estimate for the average height of active individuals? What about the median?

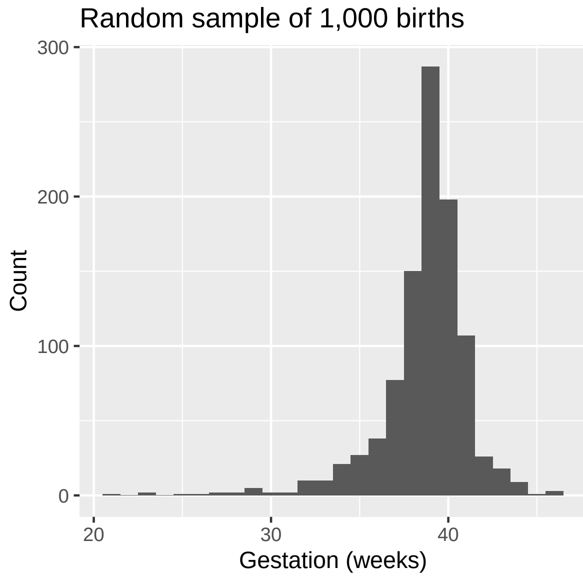
   b. What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?

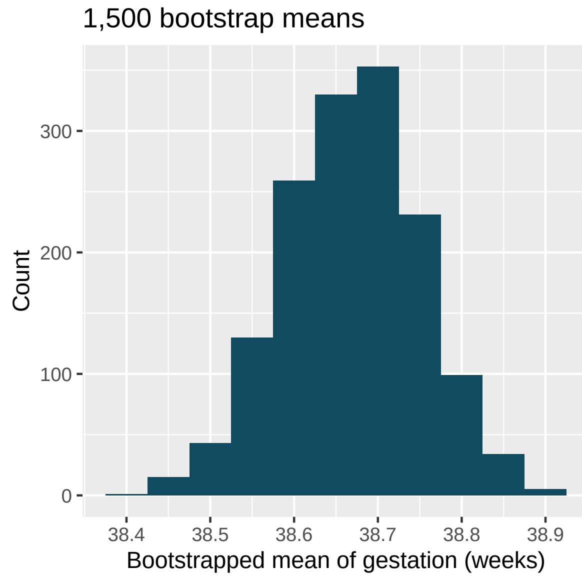
   c. Is a person who is 1m 80cm (180 cm) tall considered unusually tall? And is a person who is 1m 55cm (155cm) considered unusually short? Explain your reasoning.

   d. The researchers take another random sample of physically active individuals. Would you expect the mean and the standard deviation of this new sample to be the ones given above? Explain your reasoning.

   e. The sample means obtained are point estimates for the mean height of all active individuals, if the sample of individuals is equivalent to a simple random sample. What measure do we use to quantify the variability of such an estimate? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

2. **Length of gestation, confidence interval.** Every year, the United States Department of Health and Human Services releases to the public a large dataset containing information on births recorded in the country. This dataset has been of interest to medical researchers who are studying the relation between habits and practices of expectant mothers and the birth of their children. In this exercise we work with a random sample of 1,000 cases from the dataset released in 2014. The length of pregnancy, measured in weeks, is commonly referred to as gestation. The histograms below show the distribution of lengths of gestation from the random sample of 1,000 births (on the left) and the distribution of bootstrapped means of gestation from 1,500 different bootstrap samples (on the right).

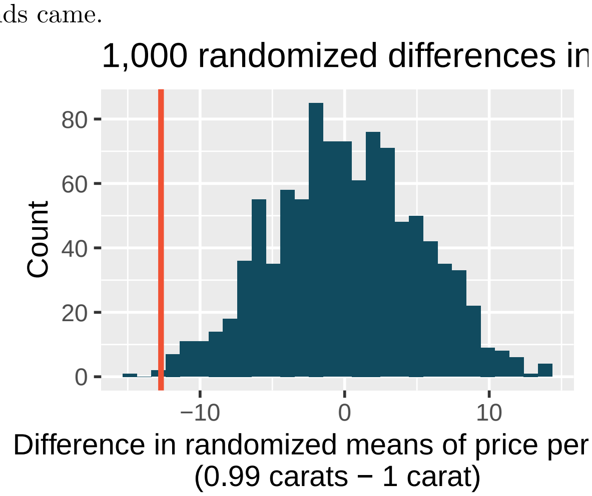




a. Given the bootstrap sampling distribution for the sample mean, find an approximate value for the standard error of the mean.

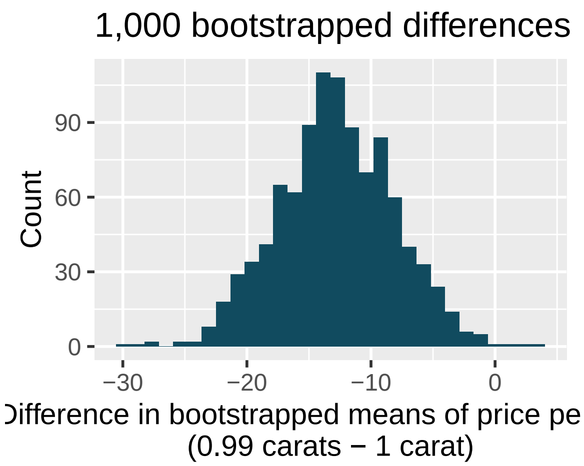
b. By looking at the bootstrap sampling distribution (1,500 bootstrap samples were taken), find an approximate 99% bootstrap percentile confidence interval for the true average gestation length in the population from which the data were randomly sampled. Provide the interval as well as a one-sentence interpretation of the interval.

3. **Diamonds, randomization test.** The prices of diamonds go up as the carat weight increases, but the increase is not smooth. For example, the difference between the size of a 0.99 carat diamond and a 1 carat diamond is undetectable to the naked human eye, but the price of a 1 carat diamond tends to be much higher than the price of a 0.99 diamond. In this question we use two random samples of diamonds, 0.99 carats and 1 carat, each sample of size 23, and randomize the carat weight to the price values in order compare the average prices of the diamonds to a null distribution. In order to be able to compare equivalent units, we first divide the price for each diamond by 100 times its weight in carats. That is, for a 0.99 carat diamond, we divide the price by 99. or a 1 carat diamond, we divide the price by 100. The randomization distribution (with 1,000 repetitions) below describes the null distribution of the difference in sample means (of price per carat) if there really was no difference in the population from which these diamonds came.



Using the randomization distribution of the difference in average price per carat (1,000 randomizations were run), conduct a hypothesis test to evaluate if there is a difference between the prices per carat of diamonds that weigh 0.99 carats and diamonds that weigh 1 carat. Make sure to state your hypotheses clearly and interpret your results in context of the data. [@ggplot2]

4. **Diamonds, bootstrap interval.** 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.