Inspiring Greatness

INDEPENDENCE SCHOOL DISTRICT

Math Virtual Learning

AP stats / Bootstrapping

May 22, 2020

Lesson: May 22, 2020

Objective/Learning Target:
Students will be introduced to bootstrap testing and the type of complex thinking typical on the AP test question 6.

Review #1

The SAT math scores for applicants to a particular engineering school are normally distributed with a mean of 680 and a standard deviation of 35. Suppose that only applicants with scores above 700 are considered for admission. What percentage of the applicants considered have scores below 750?

Review #2

Given two event, E and F, such that P(E)= 0.340, P(F)=0.450, and $P(E \cup F)$ = 0.637, then the two events are…

A. Independent and mutually exclusive
B. Independent, but not mutually exclusive
C. Mutually exclusive, but not independent
D. Neither mutually exclusive nor independent
E. There is not enough information to answer

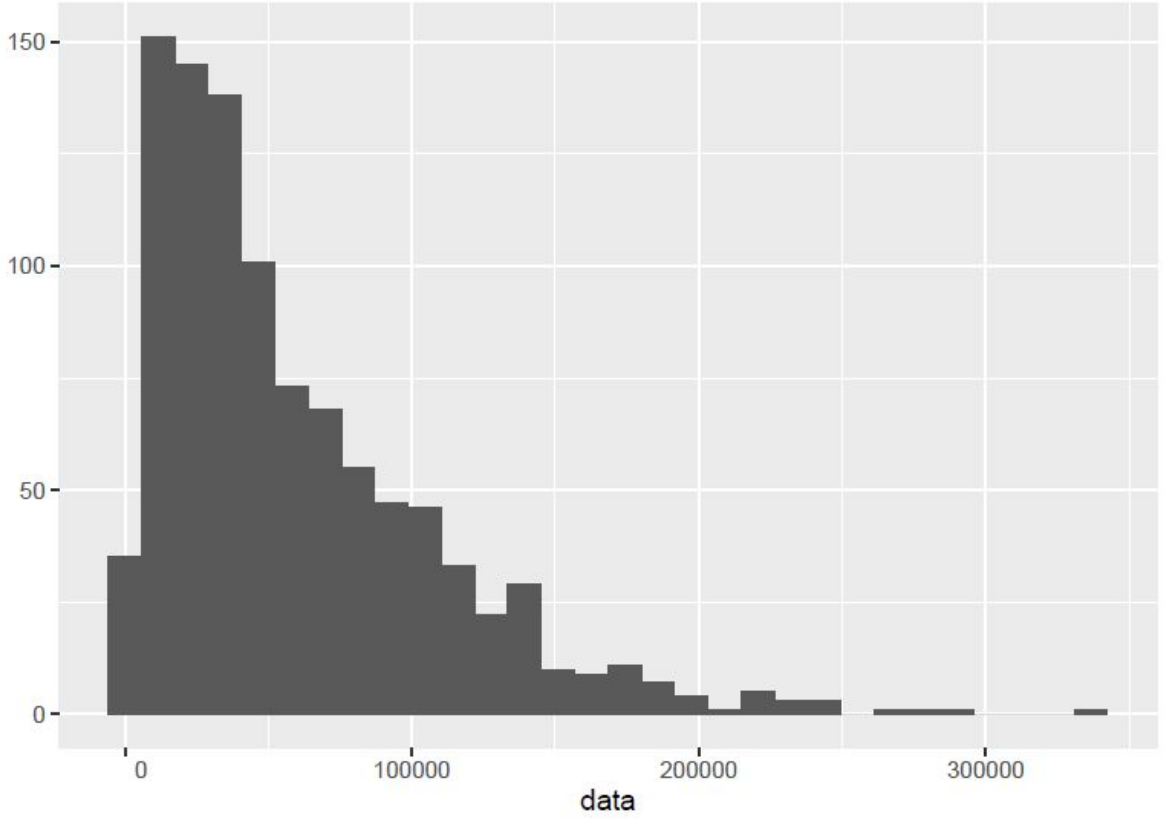
Answers

1. The probability that an applicant is considered is normalcdf(700, 9999, 680, 35) = 0.2839, where 700 is the minimum score, 9999 is just a big number to indicate no upper limit, 680 is the mean, and 35 is the standard deviation. The probability that an applicant has a score above 750 is normalcdf(750, 9999, 680, 35) = 0.0228. So, the probability of a score below 750 given it is considered is P(score<750 | score>700) = (0.2839-0.0228)/0.2839 ≈ 92%
2. Answer is B. $P(E \cup F) = P(E)+P(F)-P(E \cap F)$ so, $0.637 = 0.340+0.450-P(E \cap F)$. Thus $P(E \cap F)=0.153$. Since $P(E \cap F) \neq 0$, E and F are not mutually exclusive. However $P(E \cap F) = 0.153 = 0.340(0.450) = P(E)P(F)$, which implies E and F are independent.

Introduction

For the most part, the AP stats test covers material that we have directly covered in class. That is until question 6 of the regular AP stats test. Question 6 takes longer than any other question to answer and usually asks you to apply your knowledge of statistics in a new and novel way. Often, this application hints towards methods that are applied regularly in higher level statistics. The question is meant to judge the depth at which you understand the material. They have covered complex probability models, inference on multiple means, and most recently inference on medians. Here I plan to walk through what applying your knowledge of statistics to inference on a median might look like.

The data set



The data set

Numerical Analysis

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1023   22755   43810   58857   84240  337593
```

We are using a randomly generated data set of 1000 numbers. The data will be interpreted as a random sample of 1000 household incomes from the midwest. I have created a histogram of the data for you, as well as a numerical summary of the data. Additionally, the standard deviation of the data set is 49083.28. Please complete the following on your own.

Describe the data:

Q1. Write a brief SOCS paragraph.

Q2. What issues might arise in using the mean to describe the distribution of household incomes? Could doing so cause any issues within society at large?

Q3. Since the mean is not a good description of the distribution, what implications does this have on our known inference methods?

Describe the Data

Q1: We should have noted the data is skewed right, no points visually standout as outliers (although maybe by the 1.5 IQR rule), the median is 43810 and IQR is 61845.

Q2: The mean will overestimate what a vast majority of people are earning. We will think that people are a lot better off than they really are. Those in power could be making poor choices for their constituents based on this info.

Q3: Our inference methods have only used mean or proportion. Median does not follow the central limit theorem or a binomial distribution. We have no model to compare it too.

Redefining Our Sampling distribution

We see that our data set is strongly skewed to the right and the mean is going to overestimate where most people's true income is at. We probably do not want the mean income in this case to be used by decision makers. The median seems like a better measure of center. However, we also know that point estimates are likely to be wrong and would like to calculate a confidence interval as well. We have methods for confidence intervals on means and proportions, but not medians. Unfortunately, medians do not follow the central limit theorem, nor do they follow a binomial distribution. We have no way of ensuring a normal distribution. So how do we determine what our sampling distribution looks like?

Critical Thinking Question

Q4. Think back to how we first estimated the sampling distribution from means. How might we estimate the sampling distribution for medians?

Bootstrap samples

We previously created sampling distributions by sampling over and over again and recording the statistic of interest. When given a population of 500 pennies, we took a sample of 10 recorded the results, put them back and resampled. We did this many times until we started getting an idea of what the sampling distribution for the sample means looked like. We could do the same thing here. We could calculate our sample median (43809.78). Then take another sample of 1000 peoples household incomes. Find its median. Then do it again and again and again. This will eventually determine exactly what the sample median should be. However, that will be a very long and expensive process! Think of all the hundreds of thousands of people we would have to contact. Ideally we would end up contacting every household in the midwest. At which point, there is no need to estimate the population median. We could just calculate it.

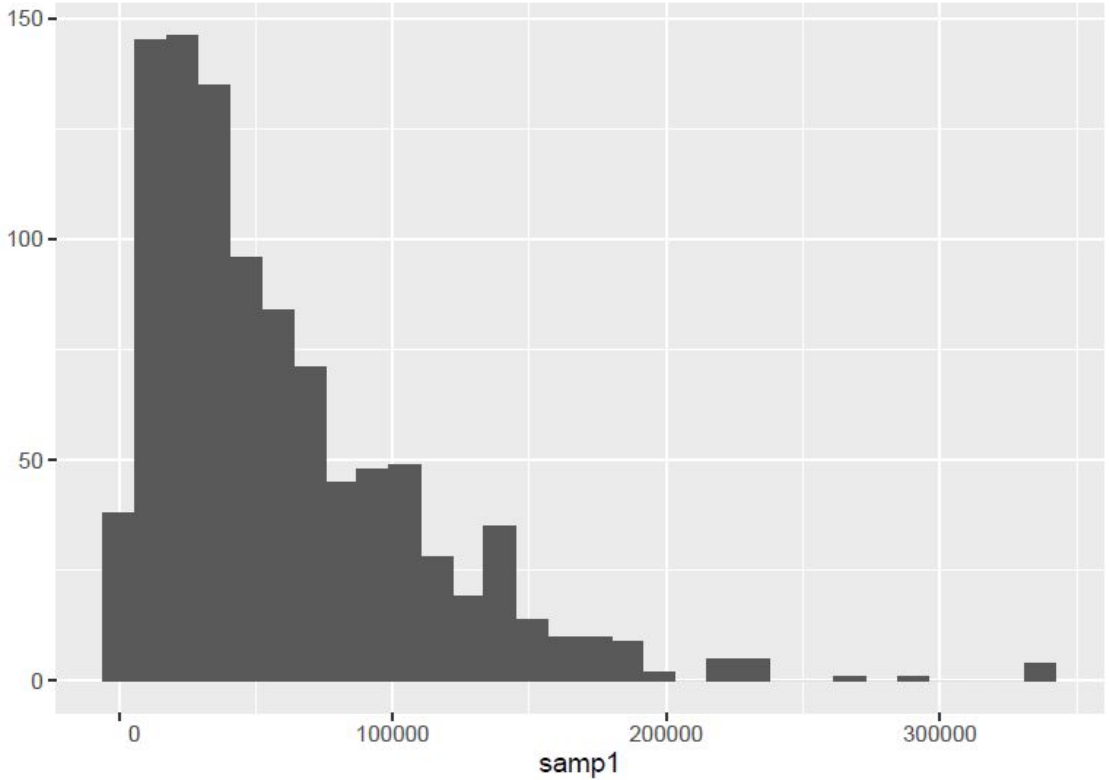
Bootstrap samples

So, we need a different option. This is where the process called bootstrapping steps in. Instead of sampling the population over and over again. We will sample the sample over and over again. To do this we will sample with replacement. So each value can be chosen more than once.

One Bootstrap

This histogram and numerical analysis represents a bootstrap sample taken from sampling our 1000 household incomes with replacement.

Q5. Compare this sample to the distribution from our original data. What has changed and what has stayed the same?



Numerical Analysis

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1023   22530   44520   59591   84470  337593
```

Building a sampling distribution

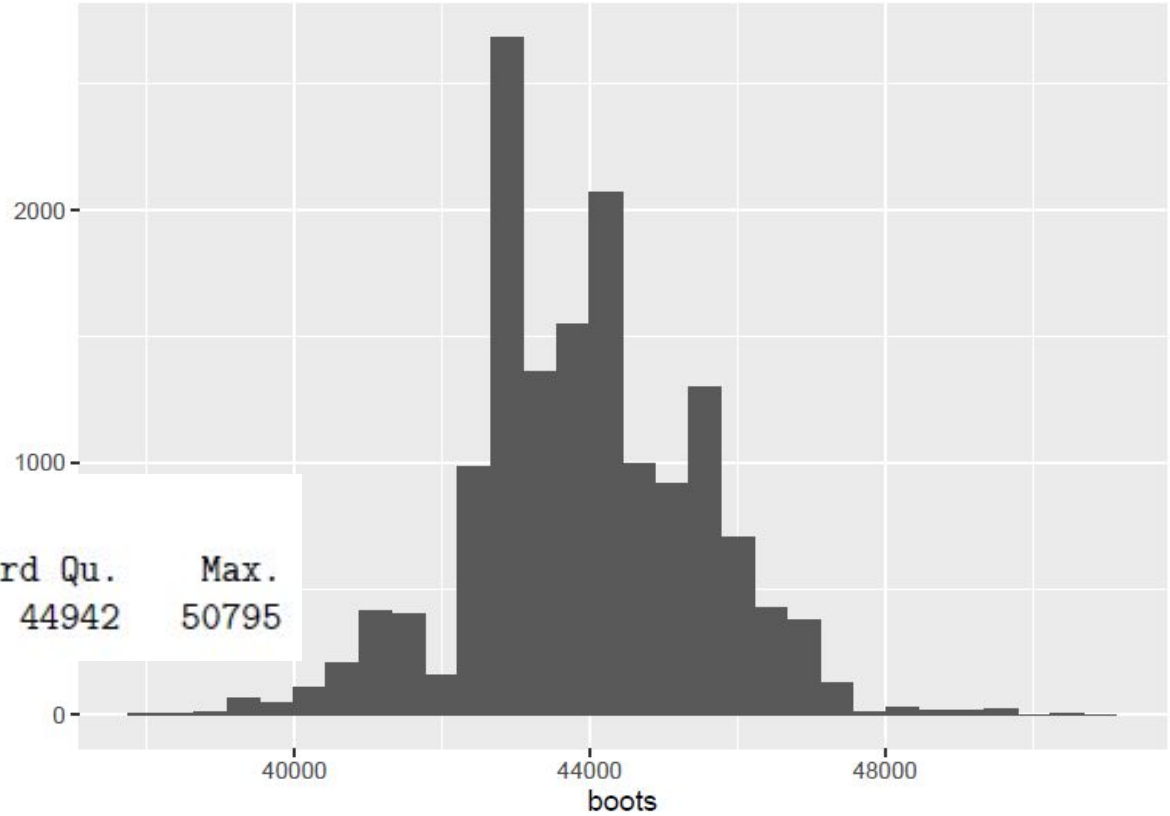
Now we have a method of obtaining a “new” sample of the household incomes. The statistic of interest is the median, so the next step is to obtain a large number of new samples and record the median for each. This is something that computers are great at doing! In fact, this method was not really possible until we had powerful computers.

The new sampling distribution

This distribution is created from 15000 bootstrap samples. It is definitely not something you would want to do by hand, but it took my computer about a second to complete.

Numerical Analysis

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   37843   42955   43813   43912   44942   50795
```



Critical Thinking Questions

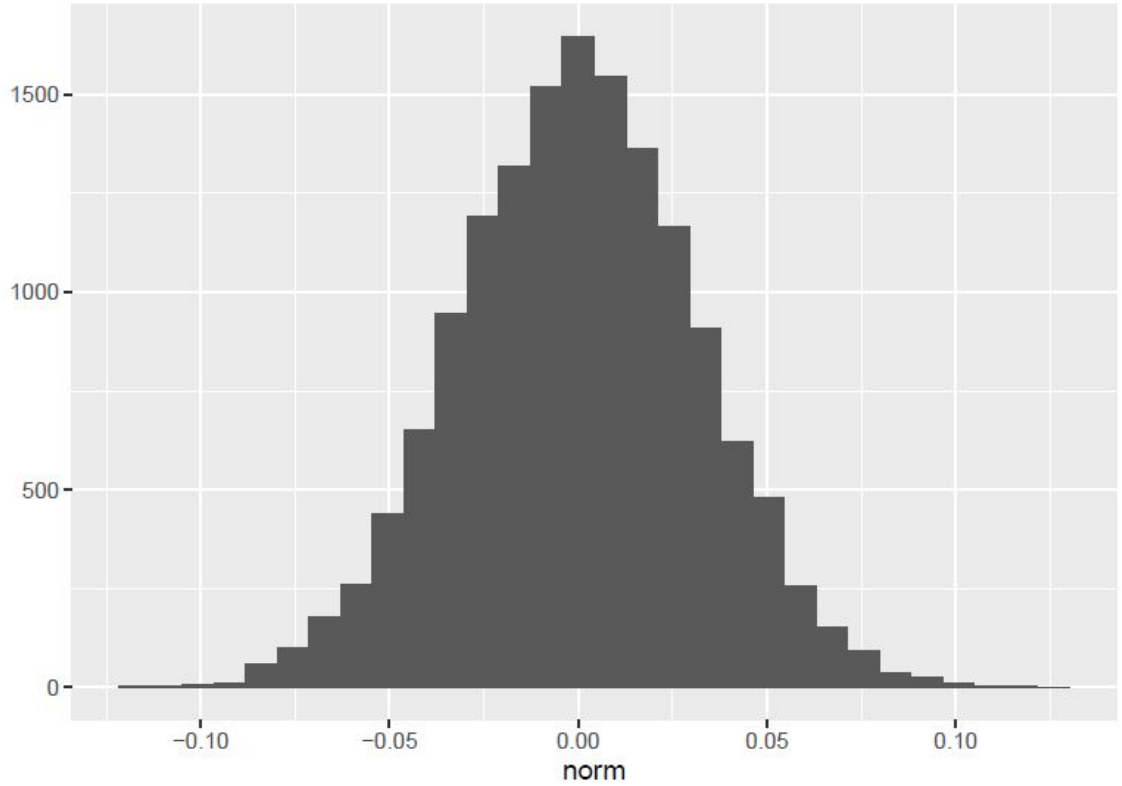
Q6. Describe the distributions of bootstrap samples

Q7. Think about what a sampling distribution of n=1000 would look like for means? How close to normal would we expect it to be? Does this appear to be that close to normal?

Describing our distribution

We might be tempted to describe this distribution as normal, but we should be wary. On the next slide is a distribution of 15000 sample means with n=1000. We look at it and have no doubt in our mind that it follows a normal distribution. This is not quite true for our bootstraps.

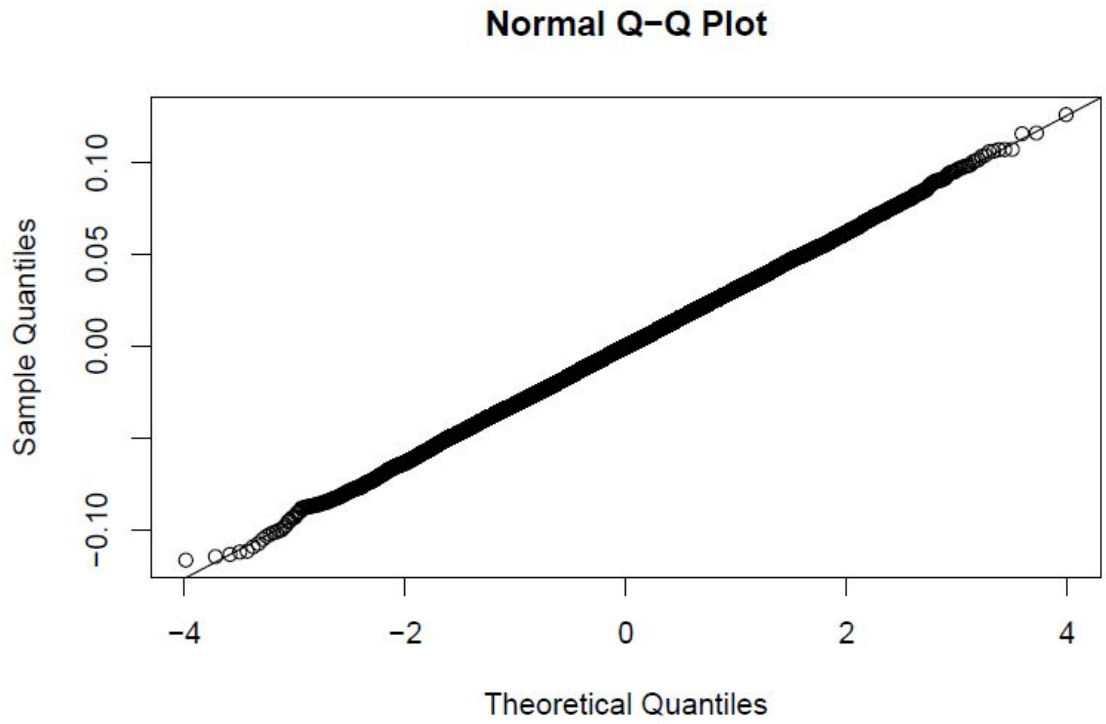
Normal Distribution



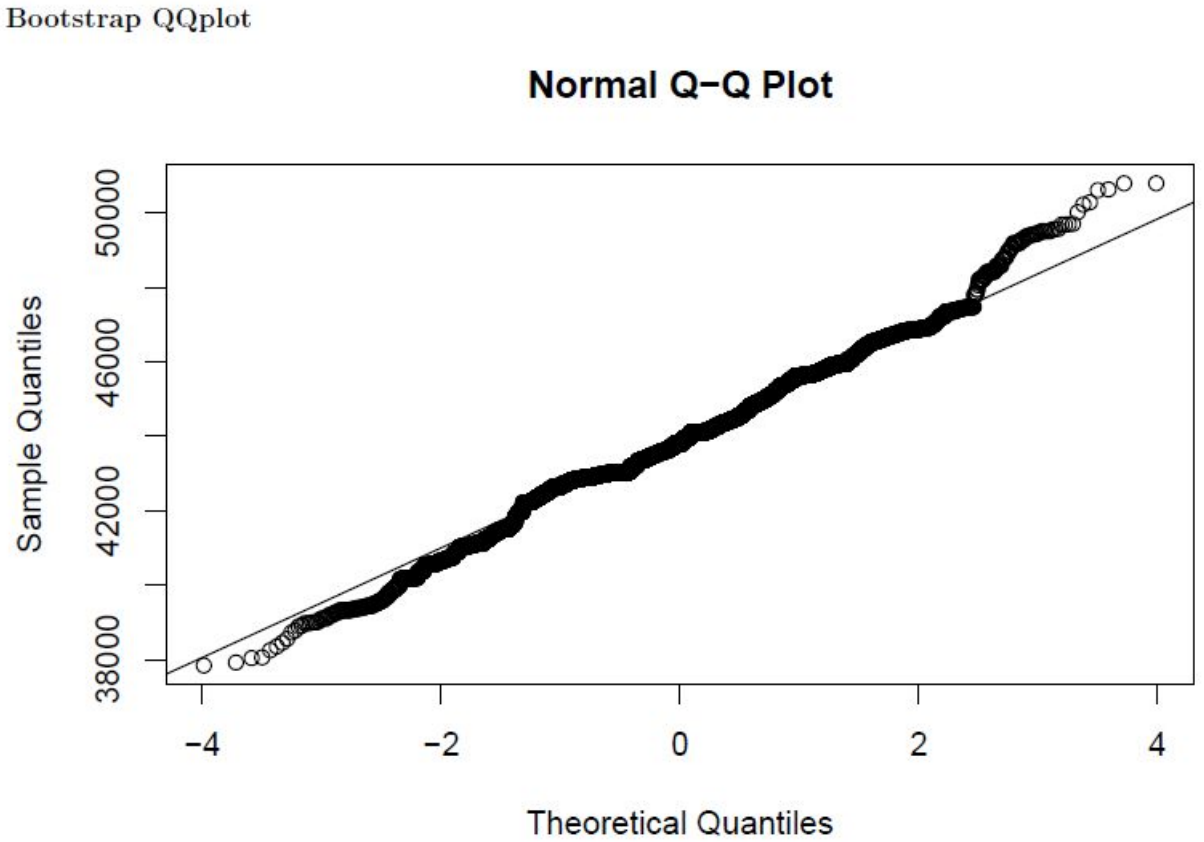
Normality?

We see that the distributions do not show the same level of normality. We can also examine normal quantile plots for the two sampling distributions.To the right is the QQ plot for the means. It is almost perfectly normal

Sample means QQplot



Bootstrap QQ-plot



Normality

Examining the two QQ plots we see that the sample means almost exactly follow the normal distribution. The Bootstrap samples deviate from the linear pattern we are expecting to see near the tails. This is because the bootstrap samples of the medians is not normal. Unfortunately, this requires a more hands on approach to creating a confidence interval.

Critical Thinking Question

Q8. If a large sample size does make this distribution normal, what do you think a smaller sample size would do to the shape of this distribution?

Q9. Since we cannot use z-scores. What more direct method might we have at calculating a confidence interval?

Creating a confidence Interval

We know that we have 15000 samples. We also know that if we want to be 95% confident in our interval. We only want 5% of our samples outside our interval. Splitting this, we should have 2.5% above and 2.5% below our interval. In other words, we want to only have $15000 \cdot 0.025 = 375$ median measures below our interval, and another 375 above our interval.

Frequency

This is a frequency table of our bootstrap medians

```
##                    .
## [36000,37000)    0
## [37000,38000)    2
## [38000,39000)   13
## [39000,40000)  127
## [40000,41000)  339
## [41000,42000)  922
## [42000,43000) 2634
## [43000,44000) 3954
## [44000,45000) 3452
## [45000,46000) 2352
## [46000,47000)  947
## [47000,48000)  163
## [48000,49000)   52
## [49000,50000)   36
## [50000,51000)    7
## [51000,52000)    0
```

Confidence Intervals

If we examine the frequency table above, we see that somewhere between 46000 and 47000, we will hit a point where there will only be 375 bootstrap medians above it. That will be the upper bound for our interval. We can also see that somewhere between 40000 and 41000, there will also be a value for which only 375 medians will fall below this point. That will be the lower bound. So if we err on the side of caution we have a 95% confidence interval no larger than 40000 to 48000. A computer can sort our list of bootstrap medians and give a slightly more precise answer of 40699 to 46851.

Therefore, We are 95% confident that the true median income is captured by the interval 40699 to 46851.

Mean vs. Median

We see that our interval, from about 40,000 to 47,000 is much lower than the original samples mean of almost 59,000. Given the differences between these two estimates for the center of the data, we can see why we have to be careful when interpreting statistics. The two different “average” incomes present very different views.

Critical Thinking Question

Q10. All of our methods have had lists of assumptions being made, one usually far more important than the rest. The bootstrap method is what is called non-parametric. It does not require any assumptions about a sampling distribution. But there is one very important assumption we still have to make about our sample. What do you think it is?

Critical Thinking Question

The one assumption, and oftentimes most important assumption we have to make is that our sample is representative of what is actually happening in the population. If we somehow manage to get a “strange” sample, our interval will not be representative no matter what we do.

Extra Practice

Free Response Question - complete #6

Answers