ST 516: Foundations of Data Analytics

More Bootstrap Methods

Bootstrap procedure in general

The general idea in the bootstrap is to learn about a sampling distribution, we use repeated samples from the data at hand rather than the population.

The term *bootstrap* comes from the idea of "pulling oneself up by your own bootstraps" or in other words "to help oneself without the aid of others".

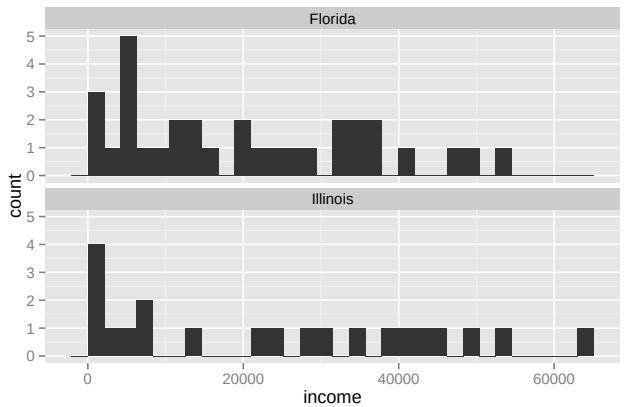
We are using the data at hand to tell us about the sampling distribution without knowing the population distribution.

Example: Median income in Florida

As another example let's get a confidence interval for the median of a population.

In a random sample of 500 people from the US Census 2000, 32 people were from Florida and 21 from Illinois. Below are histograms of their total personal income.

What was the median income for people in Florida in 2000?



Example: Median income in Florida

The sample median income in Florida is \$18,050.

This is our point estimate for the population median, but what about a confidence interval?

We saw one way to calculate one in Module 5:

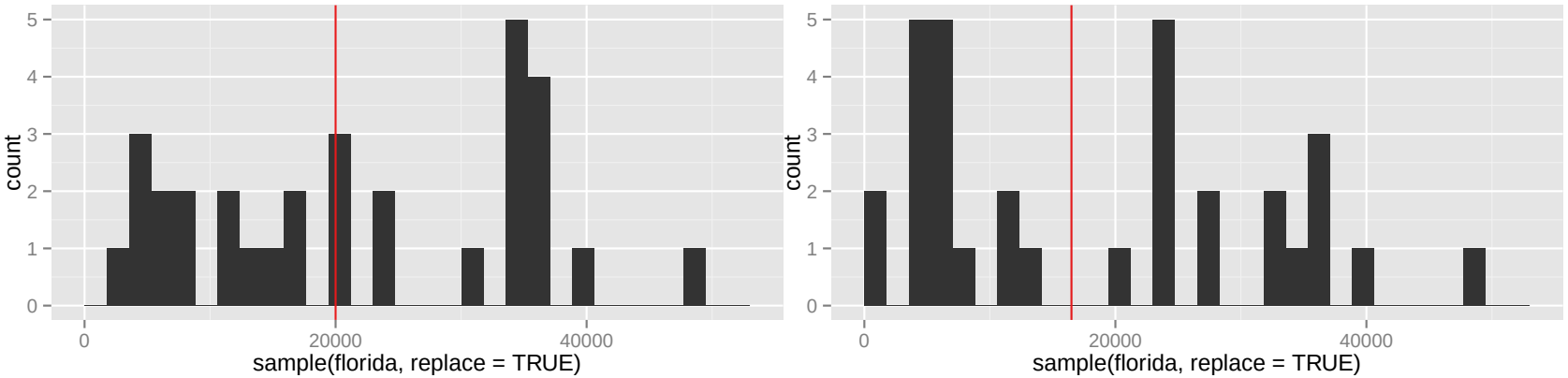
$$\left(\left(\frac{n}{2} - \frac{z_{\alpha/2}\sqrt{n}}{2}\right)\text{th Smallest Observation}, \left(\frac{n}{2} + \frac{z_{\alpha/2}\sqrt{n}}{2}\right)\text{th Smallest Observation}\right)$$

For a 95% CI ($\alpha = 0.05$), with $n = 32$, this results in the 10th and 22th smallest incomes, resulting in the confidence interval (\$7,500, \$28,000)

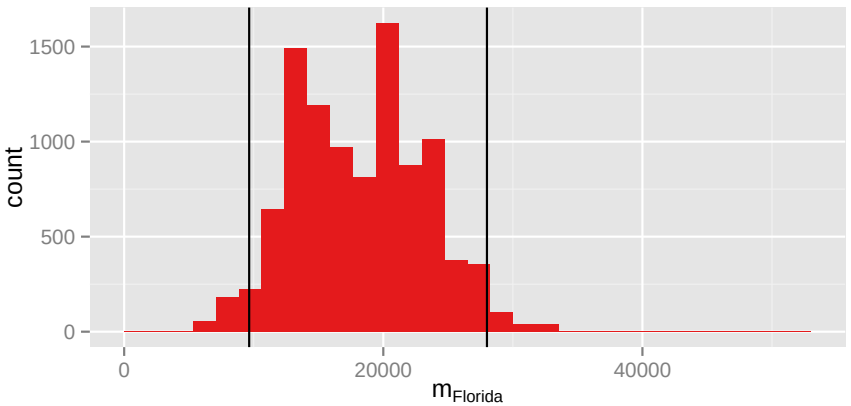
The bootstrap distribution

An alternative would be to construct a bootstrap interval.

Two bootstrap samples



The bootstrap distribution



A bootstrap confidence interval

To construct a 95% confidence interval using the percentile method we take the 0.025 and 0.975 qunatiles of the bootstrap distribution.

This results in the interval: ($9,650, $28,000)

This is quite close to the previous interval.

What about a situation where we don't have an exact method?

A two population example: Florida and Illinois

Let's get a 95% CI for the difference in median income between Florida and Illinois.

Again, a sensible point estimate, is the difference in sample medians: $\$24,000 - \$18,050 = \$5,950$.

The form for a confidence interval for the difference of two population medians is complicated (which is why you didn't see it in Module 6), but let's try to get one with bootstrapping.

Bootstrapping two populations

The bootstrap resampling mimics the population sampling we are trying to emulate.

In this example, we have two populations, and two random samples: one random sample of size 32 from Florida, and one random sample of size 21 from Illinois.

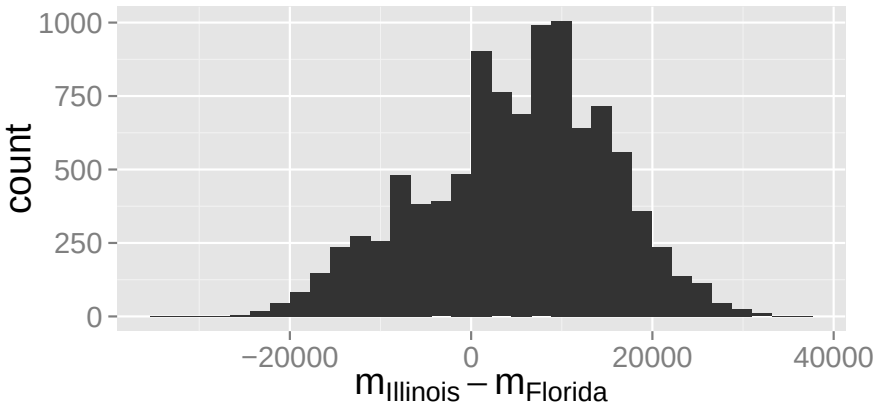
Our bootstrap procedure is,

Repeat many times:

- Sample 32 Florida incomes with replacement from those observed, and calculate the sample median, $m_{Florida}$
- Sample 21 Illinois incomes with replacement from those observed, and calculate the sample median, $m_{Illinois}$
- Calculate the difference in sample medians, $m_{Illinois} - m_{FLorida}$.

Examine the distribution of differences in sample medians

Bootstrapping two populations



The 95% confidence interval for the difference in medians using the percentile method is : $ ($-16,250, $23,250)$

With 95% confidence the median income in Florida is between $16,250 higher and $23,250 lower than the median income in Illinois.

Bootstrap Testing

We can use the relationship between confidence intervals and hypothesis tests to perform a bootstrap test.

If we are interested in:

H_0 : the median income is the same in Florida and Illinois, $M_{Florida} = M_{Illinois}$

H_A : the median income is not the same in Florida and Illinois, $M_{Florida} \neq M_{Illinois}$

We would fail to reject the null hypothesis at the 0.05 level, since \$0 is inside our 95% confidence interval.

When doesn't bootstrapping work?

Remember to keep an eye out for reasons the bootstrap might fail.

Dependence

Independence is still a key assumption for the bootstrap.

If dependence is present, the resampling of the observed data doesn't accurately capture the dependence, and the variability in the bootstrap distribution won't capture the true uncertainty in the point estimate.

Non-representative sample

If the sample data wasn't sampled at random from the population, then it won't be a good proxy for the population. A common way this can occur is that the sample was randomly obtained, but data is missing in a non-random way, perhaps due to non-response.