ST 516: Foundations of Data Analytics

Introduction to Bootstrap Methods

Big picture

Understanding the sampling distributions for a statistic is crucial to being able to use samples to make inference about populations.

The idea works so well when we are interested in the population mean because we have results that tell us regardless of the population, the shape of the sampling distribution for the sample mean is approximately Normal.

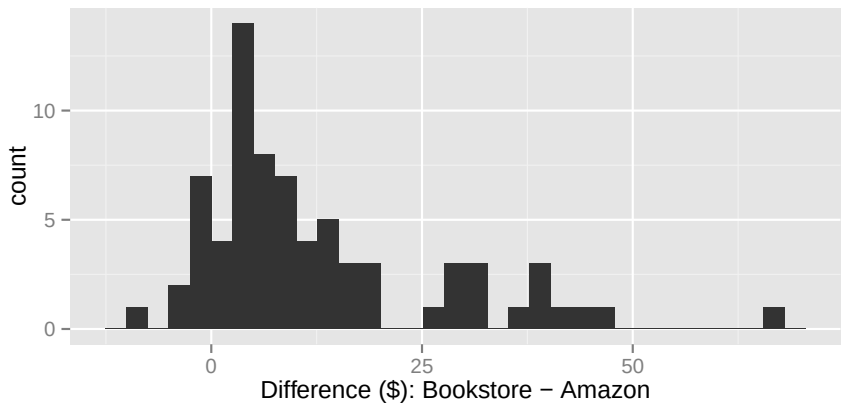
What about for statistics where we don't have general results? Or for populations where the results don't apply?

The **bootstrap** is one method for estimating the sampling distribution from the data at hand.

Example: difference in book prices

The `textbooks` data set has the price of textbooks at the University of California, Los Angeles' (UCLA's) bookstore and prices at Amazon.com for 73 randomly selected courses at UCLA in Spring 2010.

A histogram for the differences in price for each book is shown below:



Example: difference in book prices

Are textbooks cheaper online? To answer questions about the mean difference in price we could use the paired t procedures. With a sample of size 73, we would be relatively confident the t-distribution would well describe the sampling distribution of the sample mean.

What about if we were interested in the standard deviation of the differences? We could imagine estimating this with the sample standard deviation, 14.26, but we haven't presented theory to tell us about the sampling distribution of the sample standard deviation.

Reminder: sampling distributions

How do we find the sampling distribution?

If we want to know the sampling distribution for the sample standard deviations for samples of size 73 from the population of all books required at UCLA, we could:

Repeat many times:

- take a random sample of size 73 from the *population* of all differences in prices between Amazon and the bookstore for all books required at UCLA
- find the sample standard deviation

Then make a histogram of the many sample standard deviations.

The problem is: we have no way to take repeated samples from a population we don't have access to.

The bootstrap

The idea behind the bootstrap, is that although we don't know the population distribution, the distribution of a random sample should be a good guess.

That is, instead of taking many samples from the population distribution, take many samples from our sample distribution. We call these samples from the sample: bootstrap samples.

Repeat many times:

- take a bootstrap sample: take a random sample of size 73 from the *sample of differences* in prices between Amazon and the bookstore for all books required at UCLA
- find the sample standard deviation of this bootstrap sample

Then make a histogram of the many "bootstrapped" sample standard deviations.

The bootstrap

The distribution of the statistic calculated on the many bootstrap samples is called the **bootstrap distribution**.

The bootstrap distribution is an estimate of the sampling distribution in our case of interest: the sample standard deviation of a sample of size 73, from the population of differences in book prices

We can use the bootstrap distribution, much like we would a sampling distribution, to find confidence intervals for our population parameter of interest.

Sampling from a sample

An example of a bootstrap sample from the price differences is:

16.80	6.63	5.39	6.39	14.05	6.63	-0.25	12.45	-0.22	9.45	9.45
11.70	39.08	4.80	28.72	9.45	-0.25	-3.88	2.82	45.34	28.72	16.62
38.35	4.74	44.40	3.74	1.75	2.84	30.25	3.35	6.63	30.50	0.00
4.96	6.39	9.48	**16.80**	66.00	44.40	-0.25	-2.55	17.98	2.82	
29.29	9.22	11.70	9.31	4.80	13.63	9.45	38.23	4.96	19.69	
14.26	12.45	5.39	-0.28	8.23	0.42	2.82	4.78	7.01	4.64	
9.12	9.31	9.12	11.70	27.15	28.72	30.71	2.84	-9.53	14.05	

Some of the values, such as **16.80**, are duplicated since occasionally we sample the same observation multiple times.

We call this kind of sampling, **sampling with replacement**.

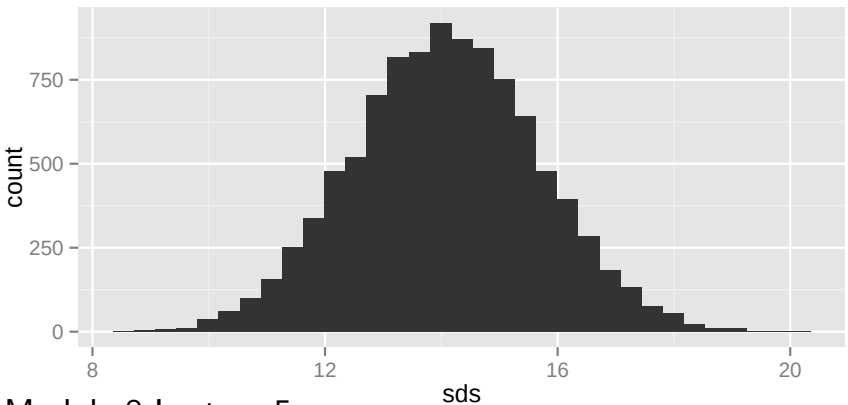
Resampling in R

In R the function `sample` with the argument `replace = TRUE` draws a sample with replacement of the same size as the observations:

```
x <- sample(textbooks$diff, replace = TRUE) # a bootstrap sample
sd(x) # the sample sd of the bootstrap sample
```

```
## [1] 14.32582
```

We can repeat this many times to get the bootstrap distribution of the sample standard deviation.



Return to example

The bootstrap distribution is an estimate of the sampling distribution in our case of interest: the sample standard deviation of a sample of size 73, from the population of differences in book prices

For example, the standard deviation of this distribution is 1.59, and estimates the standard error in the point estimate.

One way to construct a $(1-\alpha)100\%$ confidence interval based on a bootstrap distribution is to use Normal approximation to the bootstrap distribution:

$$\text{point estimate} \pm z_{\alpha/2} SE$$

where the SE is estimated from the standard deviation of the bootstrap distribution.

In our example, a 95% confidence interval for the population standard deviation is

$$14.26 \ \pm \ 1.96 \times 1.59 = (11.14, 17.37)$$

Other confidence interval methods

The confidence interval method presented:

$$\text{point estimate} \pm z_{\alpha/2} SE$$

is appropriate for unbiased point estimates, with nearly Normal sampling distributions.

Another approach is to take the $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrap distribution, a.k.a the *percentile method*. In our example resulting in the 95% CI: (11.01, 17.21)

There are many other ways to construct a confidence interval from a bootstrap distribution. OpenIntro is correct that the percentile method isn't the best, but they are incorrect to say the Normal approximation is better.

We'll use the percentile method exclusively in the next lecture.

Assumptions for bootstrapping

Our sample needs to representative of the population. The easiest way to guarantee this, is to ensure our sample is a simple random sample from the population.

The observations need to be independent.

In general, the bootstrap has better performance for larger samples. Very small samples may fail to capture enough variation to accurately characterize the sampling distribution.