4.5 Bootstrapping to study the standard deviation

We analyzed textbook pricing data in Section 4.2 and found that prices on Amazon were statistically significantly cheaper on average. We might also want to better understand the variability of the price difference from one book to another, which we quantified using the standard deviation: $s = \$14.26$. The sample standard deviation is a point estimate for the population standard deviation. Just as we care about the precision of a sample mean, we may care about the precise of the sample standard deviation.

4.5.1 Bootstrap samples and distributions

The theory required to quantify the uncertainty of the sample standard deviation is complex. In an ideal world, we would sample data from the population again and recompute the standard deviation with this new sample. Then we could do it again. And again. And so on until we get enough standard deviation estimates that we have a good sense of the precision of our original estimate. This is an ideal world where sampling data is free or extremely cheap. That is rarely the case, which poses a challenge to this "resample from the population" approach.

However, we can sample from the sample. In the textbook pricing example, there are 73 price differences. This sample can serve as a proxy for the population: we sample from this data set to get a sense for what it would be like if we took new samples.

A **bootstrap sample** is a sample of the original sample. In the case of the textbook data, we proceed as follows:

1. Randomly sample one observation from the 73 price differences.
2. Randomly sample a second observation from the 73 price differences. There is a 1-in-73 chance that this second observation will be the same one sampled in the first step.

⋮

73. Randomly sample a 73^{rd} observation from the 73 price differences.

This type of sampling is called **sampling with replacement**. Table 4.33 shows a bootstrap sample for the textbook pricing example. Some of the values, such as **16.80**, are duplicated since occasionally we sample the same observation multiple times.

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	

Table 4.33: A bootstrap sample of the textbook price differences, which represents a sample of 73 values from the original 73 observations, where we are sampling with replacement. In sampling with replacement, it is possible for a value to be sampled multiple times. For example, **16.80** was sampled twice in this bootstrap sample.

A bootstrap sample behaves similarly to how an actual sample would behave, and we compute the point estimate of interest. In the textbook price example, we compute the standard deviation of the bootstrap sample: $13.98.

4.5.2 Inference using the bootstrap

One bootstrap sample is not enough to understand the uncertainty of the standard deviation, so we need to collect another bootstrap sample and compute the standard deviation: $16.21. And another: $14.07. And so on. Using a computer, we took 10,000 bootstrap samples and computed the standard deviation for each, and these are summarized in Figure 4.34. This is called the **bootstrap distribution** of the standard deviation for the textbook price differences. To make use of this distribution, we make an important assumption: the bootstrap distribution shown in Figure 4.34 is similar to the sampling distribution of the standard deviation. This assumption is reasonable when doing an informal exploration of the uncertainty of an estimate, and under certain conditions, we can rely on it for more formal inference methods.

Example 4.42 Describe the bootstrap distribution for the standard deviation shown in Figure 4.34.

The distribution is symmetric, bell-shaped, and centered near $14.26, which is the point estimate from the original data. The standard deviation of the bootstrap distribution is $1.60, and most observations in this distribution lie between $11 and $17.

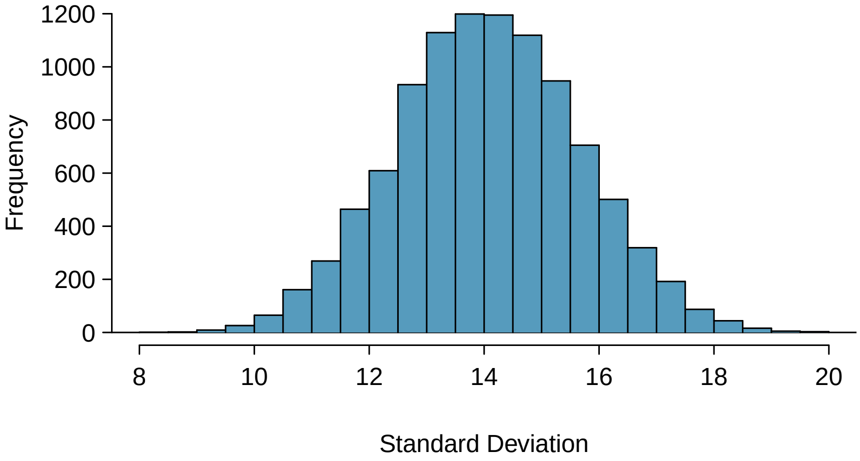

Figure 4.34: Bootstrap distribution for the standard deviation of textbook price differences. The distribution is approximately centered at the original sample's standard deviation, $14.26.

In this example, the bootstrap distribution's standard deviation, $1.60, quantifies the uncertainty of the point estimate. This is an estimate of the standard error based on the bootstrap. We might be tempted to use it for a 95% confidence interval, but first we must perform some due diligence. As with every statistical method, we must check certain conditions before performing formal inference using the bootstrap.

Bootstrapping for the Standard Deviation

The bootstrap distribution for the standard deviation will be a good approximation of the sampling distribution for the standard deviation when

1. observations in the original sample are independent,
2. the original sample size is at least 30, and
3. the bootstrap distribution is nearly normal.

We're already familiar with checking independence of observations, which we previously checked for this data set, and the second condition is easy to check. The last condition can be checked by examining the bootstrap distribution using a normal probability plot, as shown in Figure 4.35. In this example, we see a very straight line, which indicates the bootstrap distribution is nearly normal, and we can move forward with constructing a confidence interval.

As with many other point estimates, we will use the familiar formula

$$\text{point estimate} \pm t^{\star}_{df} \times SE$$

In the textbook example, using $df = 73 - 1 = 72$ leads to $t^{\star}_{72} = 1.99$ for a 95% confidence level. For bootstrapping, the standard error is computed as the standard deviation of the bootstrap distribution.

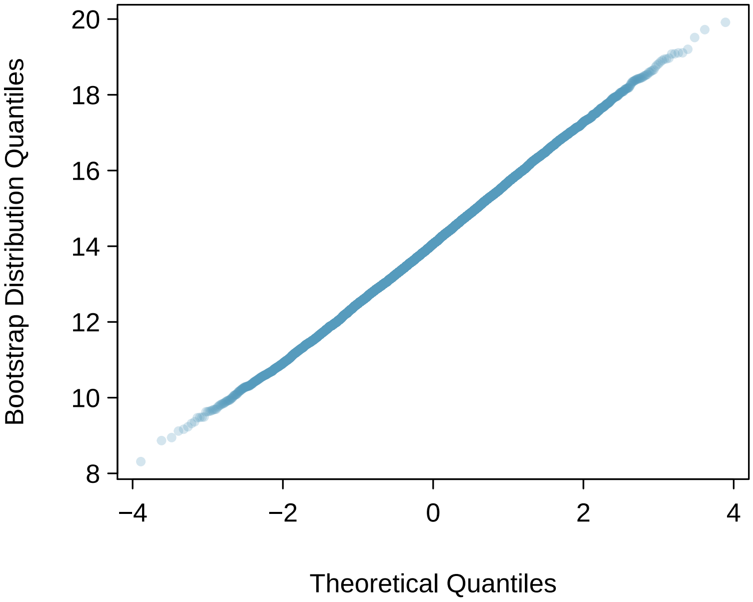

Figure 4.35: Normal probability plot for the bootstrap distribution.

● **Example 4.43** Compute the 95% confidence interval for the standard deviation of the textbook price difference.

We use the general formula for a 95% confidence interval with the t distribution:

$$\text{point estimate} \pm t^{\star}_{df} \times SE$$
$$14.26 \pm 1.99 \times 1.60$$
$$(\$11.08, \$17.44)$$

We are 95% confident that the standard deviation of the textbook price differences is between \$11.08 and \$17.44.

Had we wanted to conduct a hypothesis test, we could have used the point estimate and standard error for a t test as we have in previous sections.

Bootstrap for other parameters

The bootstrap may be used with any parameters using the same conditions as were provided for the standard deviation. However, in other situations, it may be more important to examine the validity of the third condition: that the bootstrap distribution is nearly normal.

4.5.3 Frequently asked questions

There are more types of bootstrap techniques, right? Yes! There are many excellent bootstrap techniques. We have only chosen to present one bootstrap technique that could be explained in a single section and is also reasonably reliable.

Can we use the bootstrap for the mean or difference of means? Technically, yes. However, the methods introduced earlier tend to be more reliable than this particular bootstrapping method and other simple bootstrapping techniques. See the following page for details on an investigation into the accuracy of several bootstrapping methods as well as the t distribution method introduced earlier in this chapter:

www.openintro.org/stat/bootstrap

I've heard a technique called the percentile bootstrap that is very robust. It is a commonly held belief that the percentile bootstrap is a robust bootstrap method. That is false. The percentile method is one of the least reliable bootstrap methods. Instead, use the method described in this section, which is more reliable, or learn about more advanced techniques.