A typical statistical estimation problem begins by observing a random sample from a population of interest. Let's call the size of the sample “n”, so n=15 in the right panel of Figure 7.1. From the sample we wish to infer some quantity of interest for the original population, such as the true mean of the 82 population 1st values. In this case we know the answer is 597.5, but usually we would only have the sample data to work with. The whole point of sampling is to avoid the difficulties of a full population census. The sample 1st mean 600.3 is an obvious estimate for the population mean (though later we will discuss some alternatives.) We have just made a statistical inference, a “point estimate” in the usual terminology, that 600.3 is a good guess for the true population 1st mean. But point estimates are not very useful without some idea of their accuracy. We need to make a second-level statistical inference assessing the accuracy of the point estimate 600.3.

7.1 THE BOOTSTRAP

The bootstrap is a modern computer-based algorithm for assessing accuracies. It supplements classical formulas for accuracies, which we will discuss in Chapter ??7. As we shall see the bootstrap can be applied

---

2The article that provided the population data in Figure 7.1 reported that China was employing six million interviewers to conduct a new census. They might save most of this effort by random sampling. A crude way to do this would be to divide China into say two million “sampling units”, and randomly select just 1000 of these for complete census; then add
to almost any point estimate no matter how complicated. We will begin by applying it to a relatively simple case, the accuracy of the 1sat sample mean value.

The algorithm depends on the notion of a "bootstrap sample":

Given an original sample of size n, n=15 in our case, a bootstrap sample is a random sample of size n taken with replacement from the original sample.

In our case we can imagine writing the numbers 1 through 15 on individual slips of paper, putting the slips into a hat, and making 15 draws from the hat, each time replacing the drawn slip before the next draw. The drawn slip numbers indicate the members of the bootstrap sample. In the example shown in Table 6, the first draw was 11, indicating that the first member of the bootstrap sample was the eleventh member of the original sample, 1sat value 653. Successive draws chose the 10th, 3rd, 2nd, 3rd again, 6th, 5th etc members of the original sample.

<table>
<thead>
<tr>
<th>TABLE 7.1</th>
<th>Top: the 15 1sat values for the law schools included in the original random sample, numbered in the order they were drawn from the population. Bottom: A bootstrap sample. The first member of the bootstrap sample was the 11th member of the original sample, the 2nd was the 10th original member, then the third, the second, the third again, the 6th, etc. The mean of the bootstrap sample is 606.7, compared to the original mean of 600.3.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORIGINAL SAMPLE (mean=600.3)</td>
<td>FIRST BOOTSTRAP SAMPLE (mean=606.7)</td>
</tr>
<tr>
<td>#</td>
<td>1</td>
</tr>
<tr>
<td>576</td>
<td>635</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>653</td>
<td>605</td>
</tr>
</tbody>
</table>

Notice that some original members didn’t get into this particular bootstrap sample, #4 for instance, while others entered once or twice. We might easily have seen three or more repetitions, but that didn’t happen with this bootstrap sample.

Bootstrap quantities are usually indicated by a star (asterisk ***) to avoid confusion with the original sample quantities. In this case the bootstrap mean "mean***" (pronounced "mean star") is 606.7, a little bigger than the original 1sat sample mean of 600.3. By drawing more bootstrap samples we can obtain more mean*** values, thousands of them if we wish. The basic idea of the bootstrap is this: the range of deviation of the mean*** values from the original mean indicates the mean’s accuracy. Here is the bootstrap idea applied to the 1sat data.

A computer was used to draw 2000 independent bootstrap samples, and to calculate the mean*** value for each one of them. Figure 7.1 shows the histogram of the 2000 mean***’s. They range from 569 to 636, while the central 955 of them, excluding the lowest and highest 2.5% at each extreme, goes from \(^3\) 580.3 to 620.3.

We can define a meaningful "±" assessment of accuracy for the 1sat sample mean 597.5 in terms of the bootstrap percentiles,

\(^3\)In other words the 2.5th "percentile" of the 2000 mean*** values was 580.3, while the 97.5th percentile equaled 620.3. By definition, 2.5mean***, 50 of them lie below 580.3, and likewise 1950 of them lie below 620.3, leaving 50 above. In general, to find the pth percentile of n numbers, order the numbers and take the \((n\cdot p/100)\)th from the bottom.
FIGURE 7.1 2000 bootstrap replications of the LSAT mean for the sample of 15 law schools. A simple estimate of accuracy for the observed sample mean 600.3 is one-half the distance between the 2.5th and 97.5th percentile, \((620.3-580.3)/2 = 20.0\)
"±" = one-half the distance between the 97.5th and 2.5th bootstrap percentiles.

Reading the percentiles from Figure 7.1,

\[ \pm = \frac{(620.3-580.3)}{2} = 20.0. \]

The interpretation is that with high probability, approximately 95% of the sample mean 600.3 will lie within 200 points of the true population mean. The shorthand notation for the point estimate and accuracy assessment for the \( \text{lsat} \) mean is just 600.3 ± 20.0. Here we happen to know that indeed the true mean lies well within the range of 600.3 ± 20.0. In fact we have been rather lucky in the particular random sample we chose, giving a difference between estimated value and true value of only 600.3-597.5 = 2.8, much smaller than the typical errors suggested by Figure 7.1.

Let’s state the idea behind the bootstrap more carefully. The top of Figure 7.2 indicates, very schematically, the process of drawing a random sample of size \( n \) from a population of interest. In the present case the population is the 82 law schools, the sample size is \( n=15 \), and we are focusing attention on the \( \text{lsat} \) scores. In particular we would like to learn the true mean of the 82 \( \text{lsat} \) values. To this end we have calculated the sample mean 600.3, but now we want to assess the statistical accuracy of this point estimate.

The bottom of Figure 7.2 indicates the bootstrap sampling process, often called “resampling”. It emphasizes that a bootstrap data set is obtained in the same way as the actual observed data, by drawing a random sample of the same size as the original sample. The difference is that the actual population of interest, the 82 law schools, is replaced by the observed sample. There is a great advantage to resampling from the observed data: because we have all the sample data in hand, we can easily resample as many bootstrap data sets as we wish from it (as represented by the multiple arrows emerging from the Sample Population in Figure 7.2.) The histogram in Figure 7.1 shows the results of 2000 bootstrap samples. For each one we calculated the bootstrap mean value mean*; the variability of the mean*s, measured by half the distance between the 2.5th and 97.5th percentiles of the histogram, gave us an estimate of the original mean’s accuracy, ± 20.0.

It seems like a leap of faith to replace the actual sample, which might be very big, with the sample population, which is often quite small. In fact the size of the target population makes very little difference 4 except as it affects the practical difficulties of obtaining a random sample, while the bootstrap yields useful accuracy assessments for samples as small as \( n=10 \), or even fewer. A naive criticism of statistical methods (that has been heard in the halls of congress) is “you can’t measure a population of millions with a sample of hundreds.” The more sophisticated critic may point out that it is difficult to do good random sampling on large, heterogeneous populations. As with comparisons, the quality of the statistics is more important than the quantity.

4 If the target population is of size “\( N \)”, \( N=82 \) for the law schools, it turns out we can reduce the ± accuracy assessment by a factor equaling the square root of \( 1-n/N \). In our case the square root of 1-15/82 is \( \approx 0.904 \), reducing our previous assessment of 20.0 to 18.1. Literally speaking, the bootstrap ± assessment of 20.0 applies to the usual case where the population size \( N \) is very large compared to the sample size \( n \). (Notice that we never used \( N \) in computing 20.0, so there had to be some hidden assumption involved.) The law school example was used here because the target population of 82 is small enough to picture in a display like Figure 7.1. It is an important fact that the population size \( N \) makes almost no difference to the accuracy assessment as soon as it exceeds 25 times the sample size. For the law school example where \( N \) is only 6 times bigger than \( n \) it does make a difference, and we can adjust the bootstrap accuracy estimate to 18.1 as above, but for simplicity we won’t do so in this discussion. This same point arises in connection with Figure 7.3.
FIGURE 7.2  A schematic sketch of bootstrap accuracy assessments. The multiple arrows in the lower panel represent taking many bootstrap samples, 2000 of them in the context of Figure 7.1.
The bootstrap, and other methods of assessing statistical accuracy, seem to be pulling off the impossible: using only the observed sample data, they say how far away the observed sample estimate is likely to lie from the true (unobserved) population value for the same quantity. The task isn’t really impossible of course. It relies on the fact that a sample population obtained by random sampling usually provides a useful approximation to the target population. This was illustrated in Figure 7.1 where the right panel gave only a rough picture of the full population in the left panel (the “bad” school in the lower left corner didn’t make it into the sample for instance.) But the picture was still good enough to give reasonable estimates of the \textit{l}sat mean and the correlation. Verifying the properties of the bootstrap is the kind of thing mathematical statisticians work on. Since its introduction in 1979, more than 1000 papers have been written clarifying the virtues and limitations of the bootstrap idea. At its heart the bootstrap is based on much older ideas of statistical accuracy that we will touch on in the next chapter.

7.2 CHECKING THE BOOTSTRAP ACCURACY ASSESSMENT

In this case we can check up on the bootstrap directly since we have all the population data as well as our sample of 15. Figure 7.3 shows the \textit{l}sat means for 2000 actual samples, not bootstrap samples, each of size 15, drawn randomly \footnote{These samples were drawn with replacement. As discussed in the previous footnote drawing them without replacement, which would rule out duplicates appearing in the sample, would reduce the spread of the histogram in Figure 7.3 by a factor of .904.} from the full population of 82 law schools. Applying our previous definition of accuracy, half the distance between the 2.5th and 97.5th percentiles, gives the \( \pm \) figure to be 18.97. The bootstrap assessment of 20.0 was only about 55 too large in this case, not bad at all considering how roughly the sample approximates the true population in Figure 7.1. The bootstrap assessment of accuracy for an estimate is an estimate itself, and like all statistical estimators will vary from sample to sample.

7.3 THE “SQUARE ROOT OF N LAW”

A common misunderstanding of the bootstrap goes as follows: since we can take as many bootstrap draws from the observed data as we want, why not select bigger bootstrap samples, say of size 2*\( n\)=30 instead of \( n\)=15, and get better accuracy? Figure 7.4 carries out this (wrong) idea for the \textit{l}sat mean. The open histogram refers to 2000 bootstrap mean*, each the average of 30 bootstrap draws. That is, 30 \textit{l}sat values were drawn randomly and with replacement from the observed sample of 15, and then averaged to give each mean*. Applying the \( \pm \) accuracy definition to the open histogram yields

\[ \pm = (615.1-585.9)/2 = 14.6, \]

compared to 20.0 from Figure 7.1.

Why is this wrong? Because we have spoiled the analogy of Figure 7.2 that underlies the bootstrap computation. The actual size of our sample is 15, and no amount of resampling can increase the amount of observed data. The bootstrap is NOT a device for getting more information out of the sample; it is an algorithm for assessing how much information is in the sample. Our \( \pm \) definition in terms of the percentiles of the bootstrap histogram is a computer-based way of carrying out what could be very difficult information-theoretic calculations.
FIGURE 7.3 Histogram of lsat means, 2000 samples of size 15 from the population of 82 law schools. Measuring accuracy the same way as we did for the bootstrap means, as half the distance between the 2.5th and 97.5th percentiles, we get a true accuracy assessment of \((617.1-579.1)/2 = 18.97\), compared to the bootstrap assessment of 20.0.
**FIGURE 7.4** Doubling the bootstrap sampling size to 30 (open histogram) gives what looks like better accuracy, $\pm = 14.6$ compared to 20.0 for sample size 15 (solid histogram.) However this is spurious since the actual sample size $n=15$ determines the actual accuracy of the mean.
The correct bootstrap histogram from Figure 7.1 has been superimposed on Figure 7.4. Comparing the
n=15 bootstrap histogram (solid) with the more compact n=30 bootstrap histogram (open) shows that
indeed, if we could take samples twice as big we would get improved accuracy for the sample mean. This
reinforces our belief that bigger samples should give more accurate estimates than smaller samples\(^6\).

There is a simple rule for the effect of increased sample size on statistical accuracy:

Accuracy improve proportionatly to the square root of sample size. This is the "square root of n law",
an important fact of scientific life that applies at least roughly to almost any statistic, not just sample
means. Doubling the sample size, going from 15 to 30 in our case, does not double the accuracy of a
statistical estimate, it multiplies it by about the square root of two, 1.41. The ratio of ± values from
Figure 7.4 shows the square root law in action, 20.0/14.6 = 1.37, close enough to 1.41.

The square root law has a gloomy side: it is difficult to get better accuracy by increasing the sample size
n. To get twice as much accuracy requires four times as much data, to get ten times as much accuracy,
"one order of magnitude", requires one hundred times as much data, etc. Good scientific practice tries
to increase the accuracy of each measurement so that one obtains better (smaller) ± values for the same
sample size n.

7.4 THE ACCURACY OF AN ESTIMATOR

Not all statistical estimates are created equal. We rather automatically decided to use the sample mean
as our estimation statistic in the \texttt{lsat} example, but in some cases the mean can be quite inefficient,
yielding ± values that are bigger than necessary. More "robust" estimators, like the sample median (the
middle value when the sample is ordered), can, in some circumstances, give more accurate estimates for
the same sample size.

We can use the bootstrap to compare the accuracies of different estimators applied to the \texttt{lsat} data.
A "trimmed mean" is the ordinary mean of a sample after we remove some of the extreme values. For
example we could order the \texttt{lsat} sample, trim off the smallest and largest values, and take the average
of the remaining 13, thereby computing the "one-trimmed mean". (Some olympic sports are judged in
this way.) Likewise we could compute the two-trimmed mean, the three-trimmed mean, etc, up to the
eight-trimmed mean, which for a sample of size 15 is the same as the median. More trimming produces
a more robust estimator, one that better resists large changes due to an occasional wild observation\(^7\),
but whether this increases or decreases accuracy depends upon the situation.

Figure 7.5 shows the bootstrap distribution of 2000 three-trimmed mean\(^8\)'s for the law school \texttt{lsat} sample.
It was constructed by taking 2000 bootstrap samples, just as in Figure 12 (as a matter of fact the same
2000 bootstrap samples as in Figure 7.1, to sharpen the comparison), but then computing the three-
trimmed mean instead of the ordinary mean for each of them. In this case we can see that trimming

\(^6\) As long as they are obtained by random sampling. In practice bigger sample sizes are often associated with poorer
sampling practices, as in Chapter 2, and have to be viewed critically.

\(^7\) The largest value in the \texttt{lsat} sample is 666. Suppose a transcription error caused this to be recorded as 6660. This
would increase the sample mean from 600.3 to 999.9, but the one-trimmed mean would be unaffected, since the largest
value is trimmed off anyway.
FIGURE 7.5 2000 bootstrap replications of the 3-trimmed mean for the sample of 15 lsat values (open histogram). It has worse accuracy than the ordinary mean (solid histogram), with ± value 27.2 compared to 20.0. Trimming does not help accuracy in this case.
harms the accuracy rather than helping it: the $\pm$ value increases from 20.0 to 27.2. The median was even worse, with $\pm = 30.0$.

Ordinary means are not always best. One of the prized accomplishments of statistical theory is "maximum likelihood estimation", which is an automatic algorithm for finding nearly the most efficient estimator even in very complicated situations. Using an efficient estimator helps overcome the square root of n law, and to achieve good accuracy even in small samples.

7.4.1 Correlation Coefficients

Take another look at the scatterplot for the 15 law schools in our random sample, the right panel of Figure 7.1 in Chapter 6. Not too surprisingly, schools with higher sat values also tend to have higher gpa values, and vice-versa. This is even more obvious in the left panel where all 82 schools have been plotted. The relationship isn't perfect - bigger sat values don't always go with bigger gpa values, but the trend seems undeniable.

Statisticians call this kind of trend by a name that has crept into everyday vocabulary, "correlation". The "correlation coefficient" is a number between -1 and +1 that describes the strength of the trend. Figure 7.6 shows three scattergrams that illustrate the range of possibilities. At the middle is correlation 0, where the two coordinates seem to be completely unrelated. This is what you would expect to get if you plotted physically unrelated quantities like the closing Nasdaq average and that day's noontime temperature in New York City (though you might be surprised.)

The right panel of Figure 7.6 shows a perfect positive correlation, with increasing values of one coordinate perfectly predicting the corresponding increase in the other. This situation is assigned correlation +1 on Galton's scale, the largest possible value. The relationship is perfect in the left panel too but goes in the other way, for a correlation of -1.

Taking another look at Figure 7.1, the law school data seems to lie somewhere between correlation 0 and correlation +1. Abreviating correlation as "cor", the correlation coefficient for the 15 sampled points turns out to be $\text{cor} = .776$ according to Pearson's definition. ($\text{cor}$ is pronounced "cor hat"). Statisticians use the hat symbol to distinguish an estimate based on a sample from the corresponding population quantity it is designed to estimate. In this case the population quantity is the correlation coefficient for all 82 schools, which, since we have the whole population in hand, we can compute to be $\text{cor} = .760$.

---

8 Originally "co-relation". Francis Galton, an eccentric but brilliant British scientist working in the 1870's, provided an unforgettable first example of correlation. His scatterplot had more than 1000 points, each of which reported the height at adulthood of a father and his first-born son. A clear trend emerged from the scatterplot: fathers two inches taller than the British mean had sons averaging one inch taller than the mean, those four inches taller had sons two inches taller on average, etc. Galton, with substantial later help from Karl Pearson, went on to develop the theory of correlation still in use. The father-son relationship coined another familiar term, "regression to the mean".

9 We are avoiding the mathematical definition of the correlation coefficient, but here it is anyway: (1) call the two coordinates "x" and "y". (2) Subtract the mean of the x's from each x value and the mean of the y's from each y value, and call the new coordinates "X" and "Y". (3) Then the correlation coefficient between x and y is the mean value of the quantity "X times Y", divided by the square root of the quantity "mean(X squared) times mean(Y squared)". This specific definition is called the Pearson correlation coefficient, after Galton's great successor Karl Pearson.
FIGURE 7.6  The range of possible correlation coefficients.
How accurate is the estimate $c_{\theta r} = .776$? Here we can see that it is very accurate, nearly equalling the true value $cor = .760$, but in the usual situation we would only have the sample data to go on. Once again we can use the bootstrap to get a sample-based assessment of accuracy, this time for $c_{\theta r}$.

![Histogram of bootstrap correlation coefficients](image)

**Figure 7.7** 2000 bootstrap correlation coefficients for the sample of 15 law schools. The ± accuracy assessment is .248. Notice that in this case the histogram is quite asymmetric, and that the 2.5th percentile point is much further than the 97.5th percentile from the point estimate .776.

Figure 7.7 is the histogram of 2000 bootstrap correlation coefficients obtained from the sample of 15 law schools. The bootstrap samples were drawn exactly as before, as random samples of size 15 drawn with replacement from the list of 15 law schools. For example we might have drawn the same bootstrap sample as in Table 6, schools 11,10,3,2,3,6,...8,10,6. Now though we compute the correlation coefficient for each set of bootstrap data, not the raw mean. Applying our definition of accuracy to the bootstrap percentiles in Figure 7.7 gives

$$
\pm = (0.965 - 0.469)/2 = .248
$$

for $c_{\theta r}$. The actual random sample that gave us the 15 law schools was indeed a fortunate one: it produced a point estimate $c_{\theta r} = .776$ much closer to the true value $cor = .760$, only .016 units away, than the ± Figure lead us to expect on average. Another sample of 15 law schools might easily given us $c_{\theta r} = .65$ or .85 or worse. Using three digits to report the results is misleading, and it would be fairer to
report

c\hat{r} = .78 + -.25.

You may have noticed that .78 + .25 = 1.03, greater than the largest possible value 1.00 for a correlation coefficient. Figure 7.7 shows that the bootstrap histogram for the correlation is quite asymmetric, unlike Figure 7.1 for the mean. We can see that in this case \( c\hat{r} \) can err further in the too-small direction than too-big. We could use the bootstrap percentiles to capture this asymmetry and give a more accurate interval for the true correlation than \( c\hat{r} \pm .248 \), but this would take us into the realm of "confidence intervals", beyond our story here.

### 7.4.2 Complicated Statistical Estimates

The correlation coefficient is more complicated to define and compute than a sample mean. Notice however that the mathematical complications of a summary statistic are fundamentally unimportant to the bootstrap. Suppose you have used some terribly complicated computer subroutine, let's call it "comp", to compute the value of a statistical estimate of interest for your observed data. In our previous example, "comp" was the subroutine that computed the Pearson correlation coefficient, giving the estimate .776 for the observed sample of 15 law schools. Then in the bootstrap stage, at the bottom of Figure 13, you simply call "comp" again for each of the bootstrap samples. This is what was done in Figure 7.7 where the correlation subroutine was invoked 2000 times. It took the computer more time to bootstrap the correlation than the \texttt{lsat} mean because the correlation coefficient is more complicated, actually about 10 times as long, but no additional human thought was required. This is a good bargain in an age of busy people and fast and cheap computation.

Statisticians and the scientists who use statistics now do computations involving massive data sets and enormously complicated statistical estimators. Methods like permutation tests and the bootstrap have been developed in response to the increased demands of scientific investigation in the computer era. However the ideas behind these methods go back at least 200 years, as described in the next chapter.