5

THE BOOTSTRAP

In the previous chapters we learned about sampling distributions and some ways to compute or estimate them. A common feature of previous examples is that the relevant populations were *known*—for example, a binomial distribution with specified p or exponential distribution with specified λ.

You may protest—What about permutation distributions or goodness-of-fit tests? The populations were not known there. But even in such situations, we were concerned only with sampling distributions when the null hypothesis is true. For example, we assumed that the true means (and spreads and shapes) of two populations are same, or the distribution of birthdays are uniform across quarters, or the home run counts come from a Poisson distribution. These assumptions provided enough additional information that our sampling was from known populations. Thus, in permutation testing, under the null hypothesis, we could then pool the data and proceed to draw samples without replacement, using the pooled data as the known population.

We now move from the realm of probability to statistics, from situations where the population is known to where it is unknown. If all we have are data and a statistic estimated from the data, we need to estimate the sampling distribution of the statistic. In this chapter, we introduce one way to do so, the bootstrap.

5.1 INTRODUCTION TO THE BOOTSTRAP

For the North Carolina data (Case Study in Section 1.2), the mean weight of the 1009 babies in the sample is 3448.26 g. We are interested in μ, the true mean birth weight

Mathematical Statistics with Resampling and R, First Edition. By Laura Chihara and Tim Hesterberg.

```
  times.diff.mean[i] <- mean(Basic.sample) - mean(Ext.sample)
}

hist(times.diff.mean,
      main = "Bootstrap distribution of difference in means")
abline(v = mean(times.Basic) - mean(times.Ext), col = "blue",
       lty = 2)

                                  # Open new graphics device
dev.new()
qqnorm(times.diff.mean)
qqline(times.diff.mean)
```

We find the numeric summaries:

```
> mean(times.Basic) - mean(times.Ext)
[1] 2.34
> mean(times.diff.mean)
[1] 2.344409
> sd(times.diff.mean)
[1] 0.747343
> quantile(times.diff.mean, c(0.025, 0.975))
 2.5% 97.5%
 0.89  3.80
> mean(times.diff.mean) - (mean(times.Basic) -
                          mean(times.Ext)) # bias

[1] 0.004409
```

We will discuss bias in Section 5.6.

The 95% bootstrap percentile confidence interval for the difference in means (basic−extended) is (0.89, 3.80). Thus, we are 95% confident that commercial times on basic channels are, on average, between 0.89 and 3.80 min longer than on extended channels (per half-hour time periods).

We can also conduct a permutation test of the hypothesis that the mean commercial times for the two cable options are the same versus the hypothesis that mean times are not. Figure 5.9 shows the permutation distribution for the difference in mean advertisement time between basic and extended TV channels.

Recall that in permutation testing, we sample *without* replacement from the pooled data. The permutation distribution corresponds to sampling in a way that is consistent with the null hypothesis that the population means are the same. Thus, the permutation distribution is centered at 0. But in bootstrapping, we sample *with* replacement from the individual sample. However, the bootstrap has no restriction in regard to any null hypothesis, so its distribution is centered at the original difference in means.

The permutation distribution is used for a single purpose: calculate a *P*-value to see how extreme an observed statistic is if the null hypothesis is true. The bootstrap is used for estimating standard errors and for answering some other questions we will raise below.

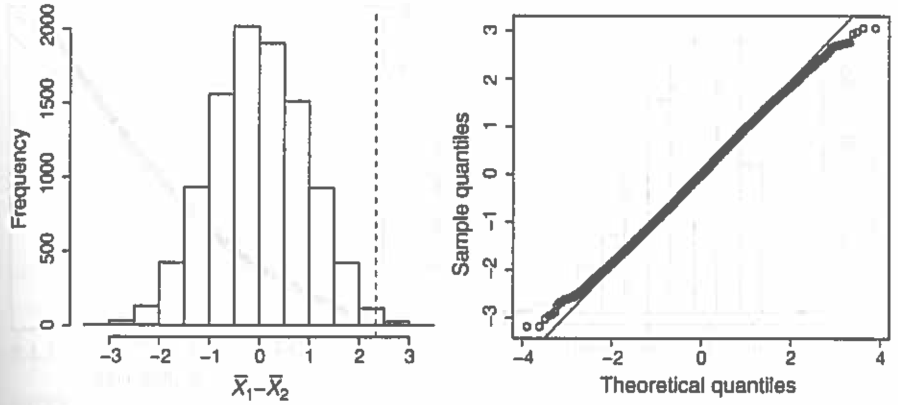

FIGURE 5.9 Histogram and normal quantile plot of the permutation distribution for the difference in mean advertisement time in basic versus extended TV channels. The vertical line in the histogram marks the observed mean difference.

The permutation test for this example results in a *P*-value of 0.0055; thus, we conclude that the mean commercial times are not the same between the two types of cable TV channels. □

Example 5.5 We return again to the Verizon example in Section 3.3. The distribution of the original data is shown in Figure 3.4, the permutation distribution for the difference in means is shown in Figure 3.5, and a permutation test of the difference in medians and trimmed means shown in Figure 3.6.

The bootstrap distribution for the larger ILEC data set ($n = 1664$) is shown in Figure 5.10. The distribution is centered around the sample mean of 8.4, has a relatively narrow spread primarily due to the large sample size, with a bootstrap SE of

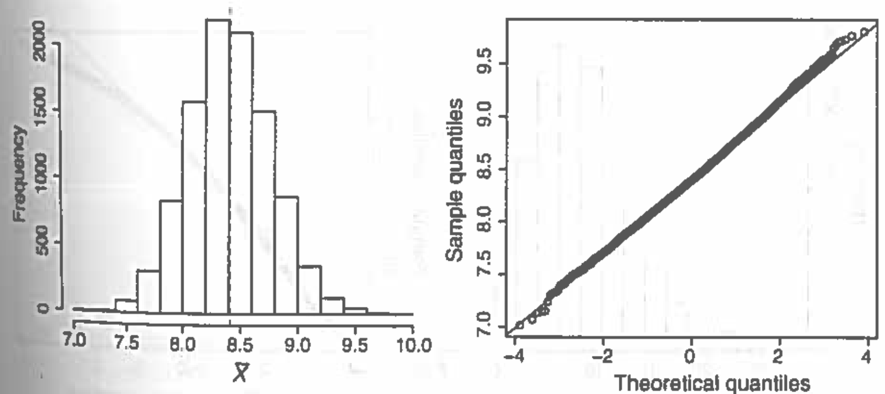

FIGURE 5.10 Bootstrap distribution for the sample mean of the Verizon ILEC data set, $n = 1664$. The vertical line is at the observed mean.

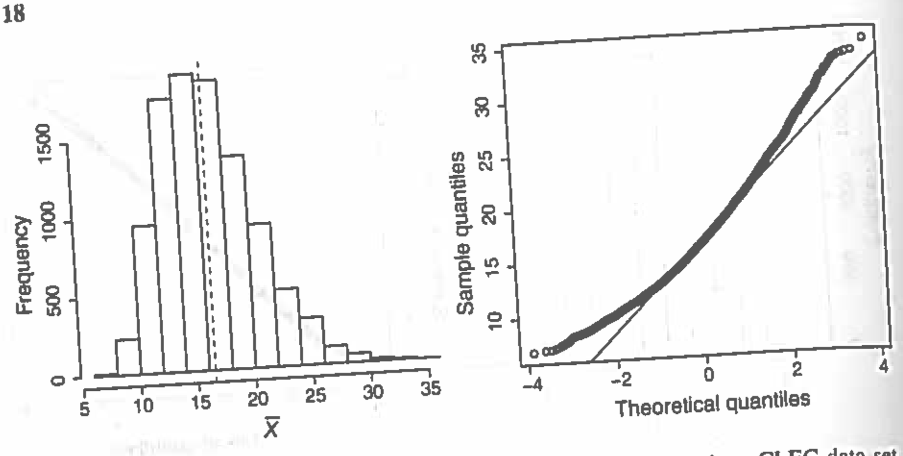

FIGURE 5.11 Bootstrap distribution for the sample mean of the Verizon CLEC data set, $n = 23$. The vertical line is at the observed mean.

0.36 and a 95% bootstrap percentile interval of (7.7, 9.1). The distribution is roughly symmetric, with little skewness.

The bootstrap distribution for the smaller CLEC data set ($n = 23$) is shown in Figure 5.11. The distribution is centered around the sample mean of 16.5, has a much larger spread due to the small sample size, with a bootstrap SE of 3.98 and a 95% bootstrap percentile interval of (10.1, 25.6). The distribution is very skewed.

The bootstrap distribution for the difference in means is shown in Figure 5.12. Note the strong skewness in the distribution. The mean of the bootstrap distribution is −8.1457 with a standard error of 4.0648. A 95% bootstrap percentile confidence interval for the difference in means (ILEC–CLEC) is given by (−17.1838, −1.6114) and so we would say that with 95% confidence, the repair times for ILEC customers are, on average, 1.61–17.18 h shorter than the repair times for CLEC customers. □

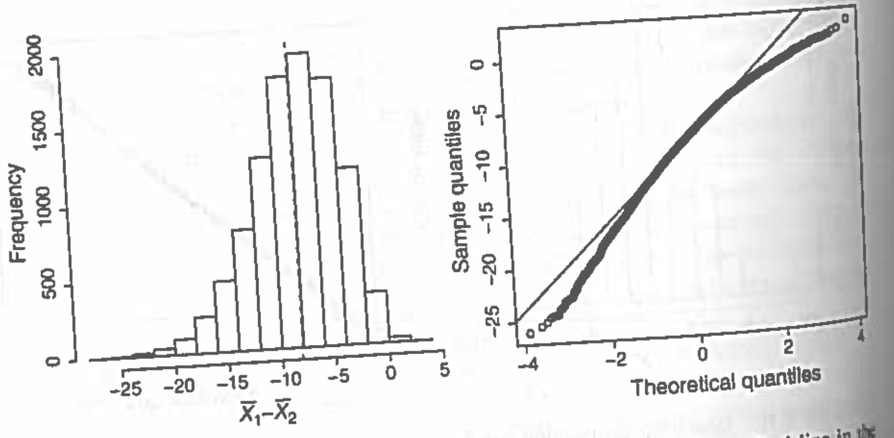

FIGURE 5.12 Bootstrap distribution for the difference in means. The vertical line in the histogram is at the observed mean difference.

TABLE 5.5 Partial View of Price Data in File **Cameras**

Item	J & R	B & H
Canon PowerShot A3000	129.99	149.99
Canon PowerShot A495	99.88	96.95
Casio EX-FC150	241.88	241.19
Kodak EasyShare C142	74.94	79.59
⋮		

5.4.1 The Two Independent Populations Assumption

Savvy consumers will often compare prices at different stores before making a purchase. Are some stores really better bargains consistently than others? We compiled the prices of a sample of point-and-shoot digital cameras from two electronic stores with an online presence, J & R and B & H (Table 5.5). The mean price of the cameras was \$155.42 at J & R and \$152.62 at B& H. Does this imply cameras are more expensive at J & R, or could the difference in mean price (\$2.81) be chance variation?

Now, it may be tempting to proceed as in the TV commercials or Verizon repair times examples, looking at the prices as coming from two populations, B & H and J & R. But note that the data are *not independent*! For each camera priced at B & H, we matched it with a price for the *same* camera at JR. Thus, the data are called *matched pairs* or *paired data*.

In this case, for each camera, we compute the difference in price between the two stores (J & R price − B & H price). We then have one variable—the price differences—and we are back to the one-sample setting given on page 101. The price differences are shown in Figure 5.13.

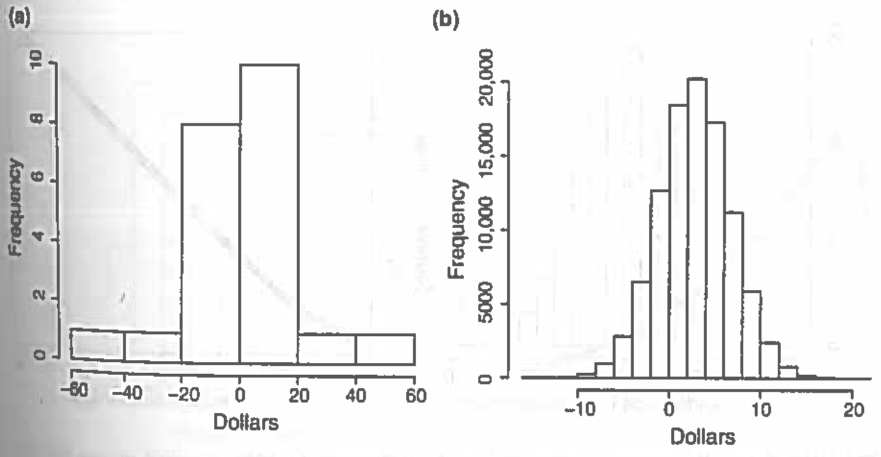

FIGURE 5.13 (a) Distribution of price differences. (b) Bootstrap distribution of price differences. The vertical line is at the observed mean price difference of \$2.81.

Performing a one sample bootstrap with 10^5 resamples, we find a 95% bootstrap percentile interval for the mean price difference to be $(-4.91, 10.62)$. Since 0 is contained in the interval, we cannot conclude that the mean prices for digital point-and-shoot cameras differ between the two stores.

5.5 OTHER STATISTICS

As with permutation testing, when bootstrapping, we are not limited to simple statistics like the simple mean. Once we have drawn a bootstrap sample, we can calculate any statistic for that sample.

For example, instead of a sample mean, we can use more robust statistics that are less sensitive to extreme observations. Figure 5.14 shows the bootstrap distribution for the difference in trimmed means, in this case 25% trimmed means, also known as the *midmean*, the mean of the middle 50% of observations. Compared to the bootstrap difference in ordinary means (Figure 5.12), this distribution has a much smaller spread.

The bootstrap procedure may be used with a wide variety of statistics—means, medians, trimmed means, correlation coefficients, and so on—using the same procedure. This is a major advantage of the bootstrap. It allows statistical inferences such as confidence intervals to be calculated even for statistics for which there are no easy formulas. It offers hope of reforming statistical practice—away from simple but nonrobust estimators like a sample mean or least-squares regression (Chapter 9), in favor of robust alternatives.

Example 5.6 In the Verizon data, rather than looking at the difference in means, suppose we look at the ratio of means. The sample ratio is 0.51, so for ILEC customers, repair times are about half of that for CLEC customers.

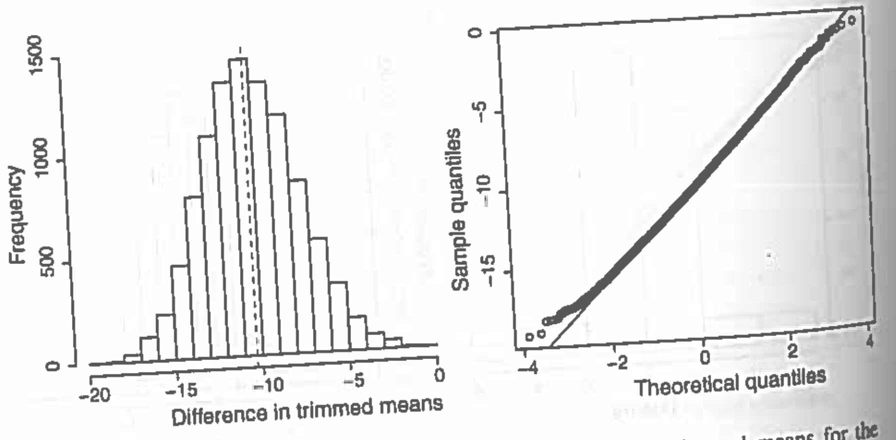

FIGURE 5.14 Bootstrap distribution for the difference in 25% trimmed means for the Verizon data.

R Note:

First create two vectors, one with the time information for the ILEC customers, one for the CLEC customers.

```
Time.ILEC <- subset(Verizon, select = Time, Group == "ILEC",
                    drop = T)
Time.CLEC <- subset(Verizon, select = Time, Group == "CLEC",
                    drop = T)
N <- 10^4
time.ratio.mean <- numeric(N)
for (i in 1:N)
{
  ILEC.sample <- sample(Time.ILEC, 1664, replace = TRUE)
  CLEC.sample <- sample(Time.CLEC, 23, replace = TRUE)
  time.ratio.mean[i] <- mean(ILEC.sample)/mean(CLEC.sample)
}

hist(time.ratio.mean,
     main="Bootstrap distribution of ratio of means")
abline(v=mean(time.ratio.mean, col = "red", lty = 2)
abline(v=mean(Time.ILEC)/mean(Time.CLEC), col = "blue", lty = 4)

dev.new()                  # open new graphics device
qqnorm(time.ratio.mean)
qqline(time.ratio.mean)
```

As in the difference of means example, the bootstrap distribution of the ratio of means exhibits skewness (Figure 5.15).

The 95% bootstrap percentile confidence interval for the ratio of means (ILEC/CLEC) is $(0.3258, 0.8415)$, so with 95% confidence, the true mean repair

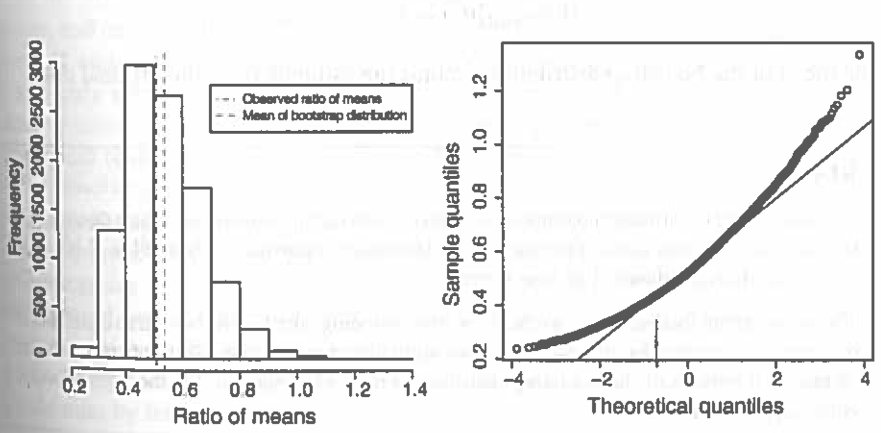

FIGURE 5.15 Bootstrap distribution for the ratio of means.

times for ILEC customers is between 0.33 and 0.84 times less than that for CLEC customers.

R Note (Verizon, Continued):

For the numeric summaries:

```
> mean(time.ratio.mean)
[1] 0.53878
> sd(time.ratio.mean)
[1] 0.1349371
> quantile(time.ratio.mean, c(0.025, 0.975))
   2.5%     97.5%
 0.3258330 0.8415138
> mean(time.ratio.mean) - mean(Time.ILEC)/mean(Time.CLEC) # bias
[1] 0.0292674
```

We will discuss bias in the next section.

□

5.6 BIAS

An estimator $\hat{\theta}$ is biased if, on average, it tends to be too high or too low, relative to the true value of θ. Formally, this is defined using expected values:

Definition 5.2 The *bias* of an estimator $\hat{\theta}$ is

$$\text{Bias}[\hat{\theta}] = \text{E}\left[\hat{\theta}\right] - \theta.$$

The bootstrap estimate of bias is

$$\text{Bias}_{\text{boot}}[\hat{\theta}^*] = \text{E}\left[\hat{\theta}^*\right] - \hat{\theta},$$

the mean of the bootstrap distribution, minus the estimate from the original data. ||

BIAS

A statistic used to estimate a parameter is *biased* when the mean of its sampling distribution is not equal to the true value of the parameter. The bias of a statistic $\hat{\theta}$ is $\text{Bias}[\hat{\theta}] = \text{E}\left[\hat{\theta}\right] - \theta$.

A statistic is **unbiased** if its bias is zero.

The bootstrap method allows us to check for bias by seeing whether the bootstrap distribution of a statistic is centered at the statistic of the original random sample. The bootstrap estimate of bias is the mean of the bootstrap distribution minus the statistic for the original data, $\text{Bias}_{\text{boot}}[\hat{\theta}] = \hat{\text{E}}[\hat{\theta}^*] - \hat{\theta}$.

We have already proven (Theorem A.5) that the sample mean is an unbiased estimator of the population mean μ. In addition, the difference of sample means is also an unbiased estimator of the difference of population means. However, the ratio of sample means is not generally an unbiased estimator of the ratio of population means. The bootstrap distribution for the ratio of means has a long right tail, large observations that occur when the denominator is small. Consequently, the mean of the resample ratio of means is large, causing positive mean bias.

Let us compare the ratio of bias/SE for different examples. For the arsenic example (page 111), the ratio is only $0.2176/18.2576 = 0.0119$. For the TV example (page 115), the ratio is only $0.0044/0.7473 = 0.0059$, so the bias is less than 0.6% of the standard error. On the other hand, for the Verizon ratio of means (page 121), the ratio is $0.0293/0.1349 = 0.2172$, so the bias is about 22% of the standard error.

If the ratio of bias/SE exceeds ± 0.10, then it is large enough to potentially have a substantial effect on the accuracy of confidence intervals. In applications where accuracy matters, there are other more accurate bootstrap confidence intervals rather than the relatively quick-and-dirty bootstrap percentile intervals. (As it turns out, bootstrap percentile intervals are actually reasonably accurate for the ratio of means.)

The next example also shows noticeable bias.

Example 5.7 A major study of the association between blood pressure and cardiovascular disease found that 55 out of 3338 ($\hat{p}_1 = 0.0165$) men with high blood pressure died of cardiovascular disease during the study period, compared to 21 out of 2676 ($\hat{p}_2 = 0.0078$) with low blood pressure. The estimated *relative risk* is $\hat{\theta} = \hat{p}_1/\hat{p}_2 = 0.0165/0.0078 = 2.12$. Thus, we would say that the risk of cardiovascular disease for men with high blood pressure is 2.12 times greater than the risk for men with low blood pressure.

To bootstrap the relative risk, we draw samples of size $n_1 = 3338$ with replacement from the first group, independently draw samples of size $n_2 = 2676$ from the second group, and calculate the relative risk $\hat{\theta}^*$. In addition, we record the individual proportions $\hat{p}_1^*$ and $\hat{p}_2^*$. The bootstrap distribution for relative risk is shown in Figure 5.16. It is highly skewed, with a long right tail caused by denominator values relatively close to zero. The standard error, from a sample of 10^4 observations, is 0.6188. The theoretical (exhaustive) bootstrap standard error is undefined because some of the $n_1^{n_1} n_2^{n_2}$ bootstrap samples have $\hat{\theta}^*$ undefined: this occurs when the denominator $\hat{p}_2^*$ is zero; this is rare enough to ignore.

The average of the resample relative risks is larger than the sample relative risk, indicating bias. The estimated bias is $2.2107 - 2.10 = 0.1107$, so the ratio of bias to the standard error is 0.1784. While the bias does not appear large in the figure, this amount of bias can have a huge impact on formula-based confidence intervals. While the bootstrap percentile interval is fine, some common symmetric confidence intervals would miss by falling under the true value about twice as often as they should.

Figure 5.17 shows the joint bootstrap distribution of $\hat{p}_1^*$ and $\hat{p}_2^*$. Each point corresponds to one bootstrap resample, and the relative risk is the slope of the line

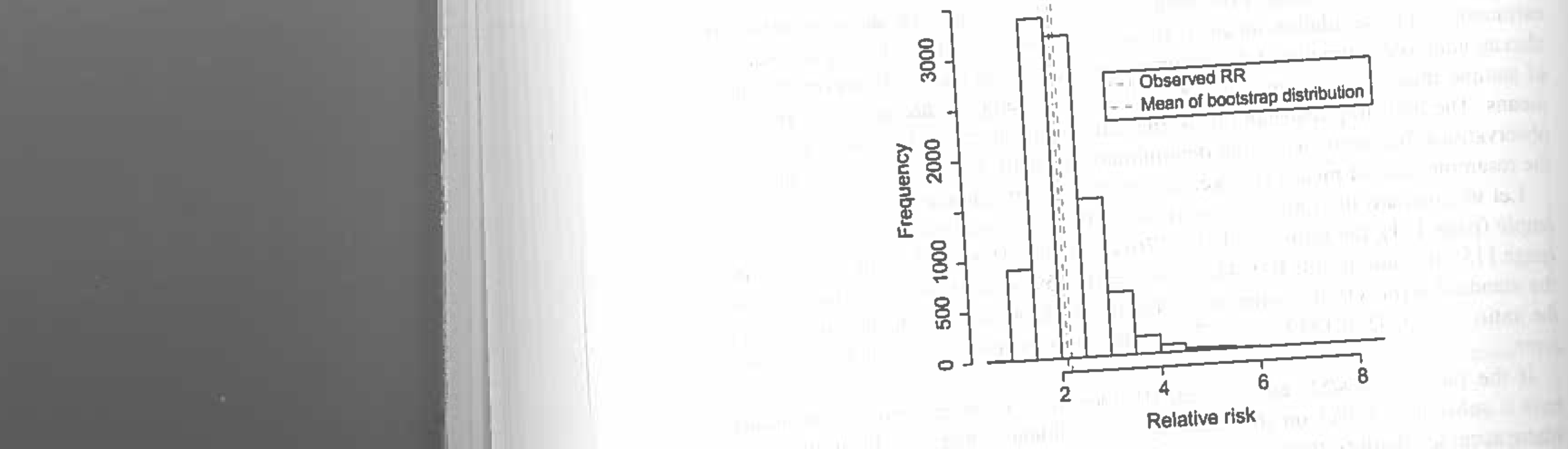

FIGURE 5.16 Bootstrap distribution of relative risk.

between the origin and the point. The original data are at the intersection of horizontal and vertical lines. The 95% bootstrap confidence interval for the true relative risk is (1.3118, 3.6877). Thus, for one bootstrap sample, if the relative risk satisfies $\hat{p}_1^*/\hat{p}_2^* < 1.3118$, then $\hat{p}_1^* < 1.3118\hat{p}_2^*$. Similarly, if $\hat{p}_1^*/\hat{p}_2^* > 3.6877$, then $\hat{p}_1^* > 3.6877\hat{p}_2^*$. This is shown by the points outside the region bounded by the dashed lines of slopes 1.3118 and 3.6877. □

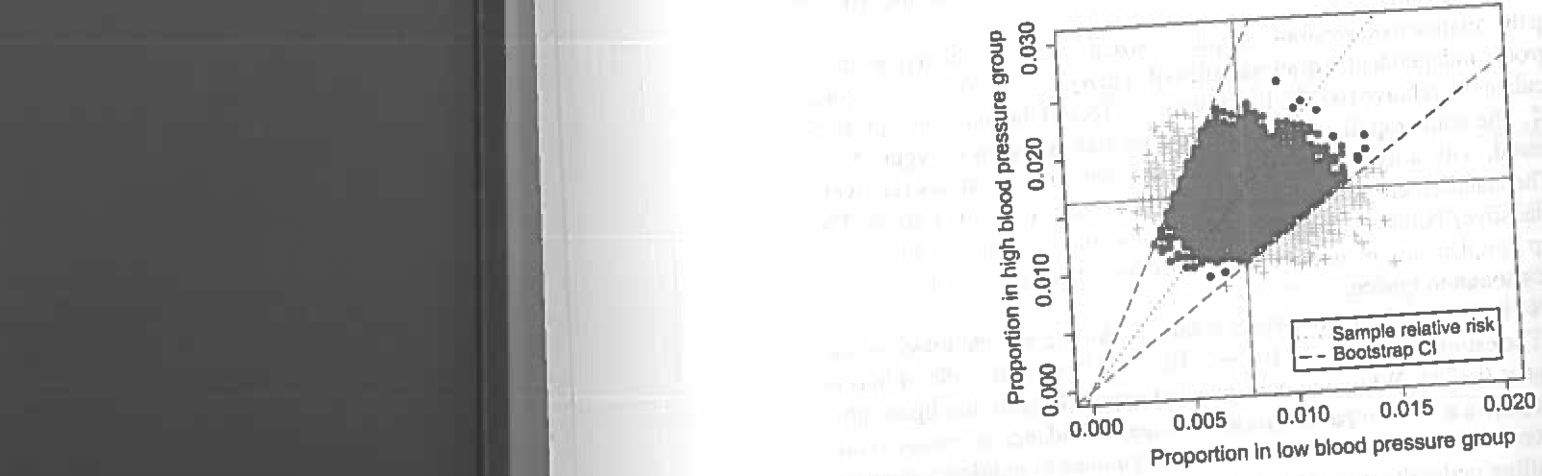

FIGURE 5.17 Bootstrapped proportions of the high blood pressure group against the proportions of the low blood pressure group.

5.7 MONTE CARLO SAMPLING: THE "SECOND BOOTSTRAP PRINCIPLE"

The second bootstrap "principle" is that the bootstrap is implemented by random sampling. This is not actually a principle but an implementation detail.

The name *Monte Carlo* dates from the 1940s, when physicists and applied mathematicians working on the Manhattan Project at Los Alamos Laboratory in New Mexico encountered difficult integrals with no closed form solutions. Stanislaw Ulam and John von Neumann proposed using computer simulations to estimate these integrals. Their conceptual leap was in using a random method to solve a deterministic problem. Because of the use of randomness, they named the method after the casino in Monaco.

Given that we are drawing i.i.d. samples of size n from the observed data, there are at most n^n possible samples ($\binom{2n-1}{n}$, if we disregard the order of observations), and ties in the data can further reduce the number of unique samples. In small samples, we could create all possible bootstrap samples, deterministically. In practice, n is usually too large for that to be feasible, so we use random sampling.

Let N be the number of bootstrap samples used, for example, $N = 10^4$. The resulting N resample statistic values represent a random sample of size N with replacement from the *theoretical bootstrap distribution* consisting of n^n values.

In some cases, we can calculate some aspects of the sampling distribution without simulation. When the statistic is the sample mean, for example, and for the ordinary one-sample bootstrap, the mean and standard deviation of the theoretical bootstrap distribution are $\bar{x}$ and $\hat{\sigma}/\sqrt{n}$, respectively, where $\hat{\sigma}^2 = 1/n \sum(x_i - \bar{x})^2$.

The use of Monte Carlo sampling adds additional unwanted variability that may be reduced by increasing the value of N. We discuss how large N should be in Section 5.9.

5.8 ACCURACY OF BOOTSTRAP DISTRIBUTIONS

How accurate is the bootstrap? This entails two questions:

- How accurate is the theoretical bootstrap?
- How accurately does the Monte Carlo implementation approximate the theoretical bootstrap?

SOURCES OF VARIATION IN A BOOTSTRAP DISTRIBUTION

Bootstrap distributions and conclusions based on them include two sources of random variation:

1. The original sample is chosen at random from the population.
2. Bootstrap resamples are chosen at random from the original sample.

We begin this section with a series of pictures intended to illustrate both questions. We conclude this section with a discussion of cases where the theoretical bootstrap is

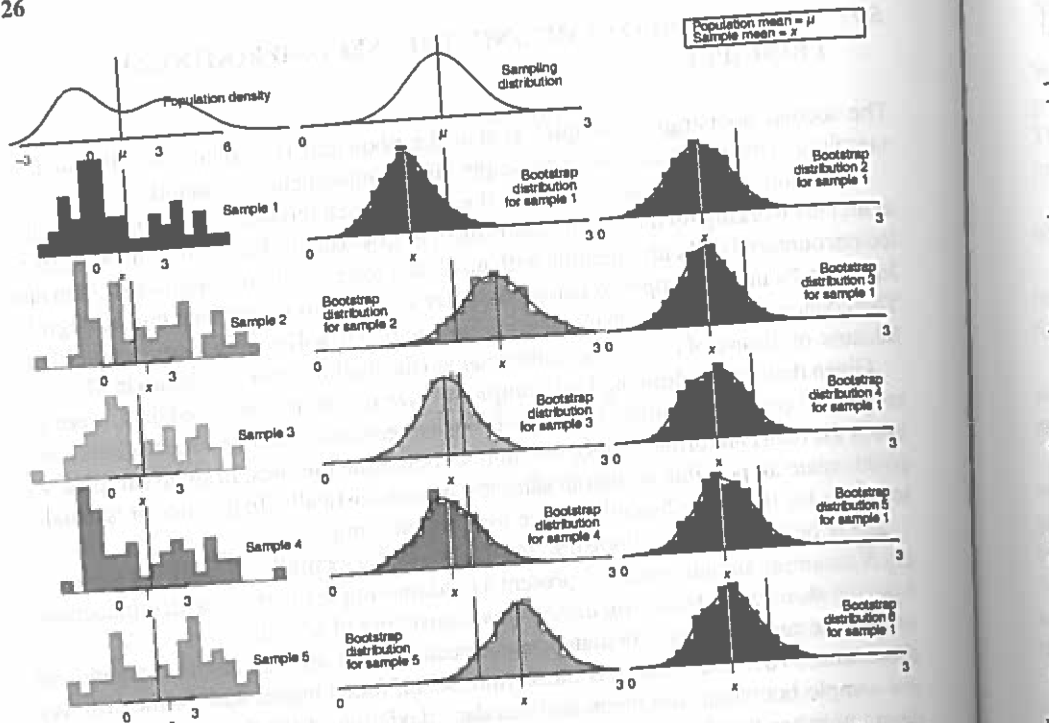

FIGURE 5.18 Bootstrap distribution for the mean, $n = 50$. The left column shows the population and five samples. The middle column shows the sampling distribution and bootstrap distributions from each sample. The right column shows five more bootstrap distributions from the first sample, with $N = 1000$ or $N = 10^4$.

not accurate and remedies. In Section 5.9, we return to the question of Monte Carlo accuracy.

5.8.1 Sample Mean: Large Sample Size

Figure 5.18 shows a population and five samples of size 50 from the population in the left column. The middle column shows the sampling distribution for the mean and bootstrap distributions from each sample, based on $N = 1000$ bootstrap samples. Each bootstrap distribution is centered at the statistic ($\bar{x}$) from the corresponding sample rather than being centered at the population mean μ. The spreads and shapes of the bootstrap distributions vary a bit but not a lot.

This informs what the bootstrap distributions may be used for. The bootstrap does not provide a better estimate of the population parameter μ, because no matter how many bootstrap samples are used, they are centered at $\bar{x}$ (plus random variation), not μ. On the other hand, the bootstrap distributions are useful for estimating the spread and shape of the sampling distribution.

The right column shows five more bootstrap distributions from the first sample, using $N = 1000$ resamples. These illustrate the Monte Carlo variation in the bootstrap. This variation is much smaller than the variation due to different original samples. For many uses, such as quick-and-dirty estimation of standard errors or approximate confidence intervals, $N = 1000$ resamples is adequate. However, there is noticeable variability, particularly in the tails of the bootstrap distributions; so when accuracy matters, $N = 10^4$ or more samples should be used.

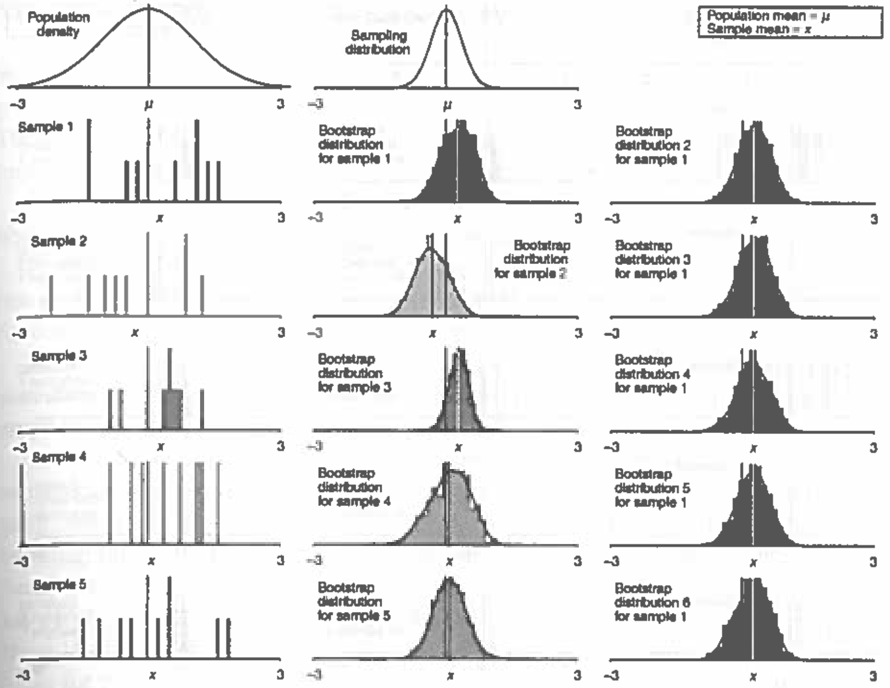

FIGURE 5.19 Bootstrap distributions for the mean, $n = 9$. The left column shows the population and five samples. The middle column shows the sampling distribution and bootstrap distributions from each sample. The right column shows five more bootstrap distributions from the first sample, with $N = 1000$ or $N = 10^4$.

5.8.2 Sample Mean: Small Sample Size

Figure 5.19 is similar to Figure 5.18, but for a smaller sample size, $n = 9$ (and a different population). As before, the bootstrap distributions are centered at the corresponding sample means, but now the spreads and shapes of the bootstrap distributions vary substantially, because the spreads and shapes of the samples vary substantially. As a result, bootstrap confidence interval widths vary substantially (this is also true of nonbootstrap confidence intervals). As before, the Monte Carlo variation is small and may be reduced with more samples.

5.8.3 Sample Median

Now turn to Figure 5.20 where the statistic is the sample median. Here, the bootstrap distributions are poor approximations of the sampling distribution. In contrast, the sampling distribution is continuous, but the bootstrap distributions are discrete with the only possible values being values in the original sample (here n is odd). The

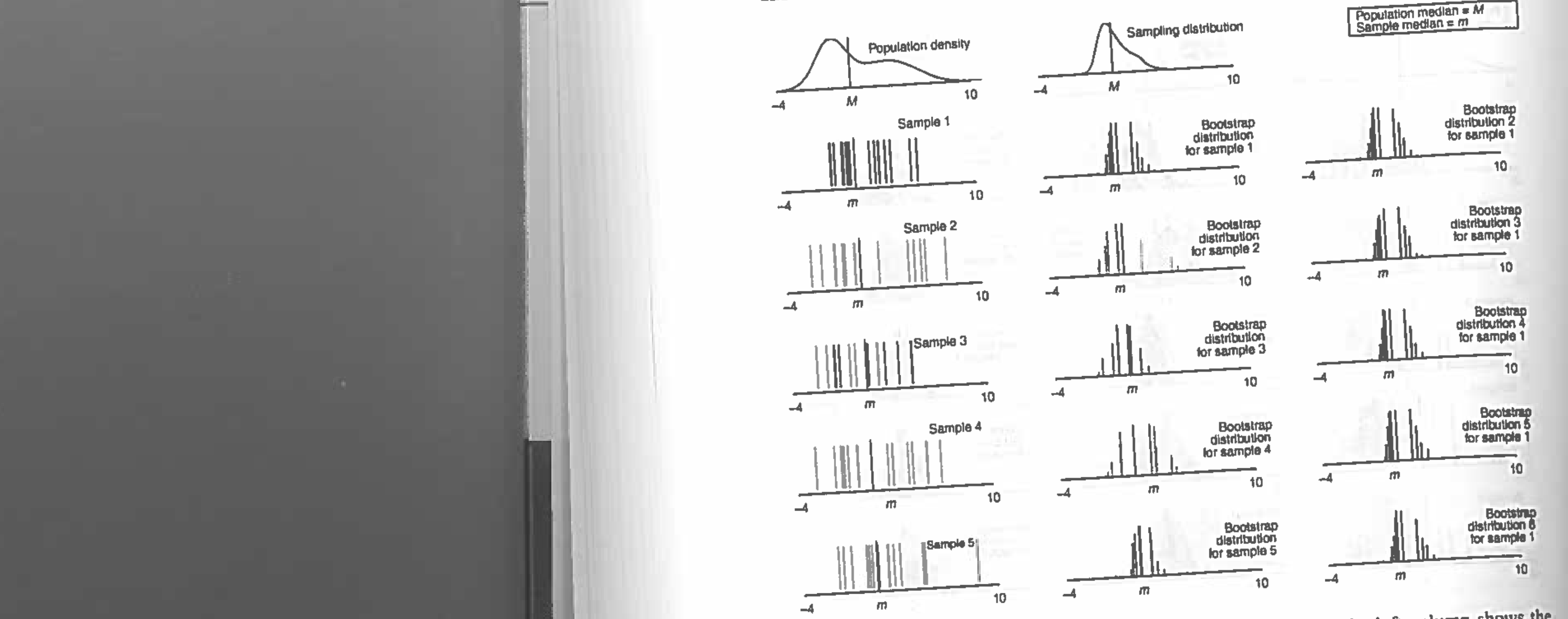

FIGURE 5.20 Bootstrap distributions for the median, $n = 15$. The left column shows the population and five samples. The middle column shows the sampling distribution and bootstrap distributions from each sample. The right column shows five more bootstrap distributions from the first sample.

bootstrap distributions are very sensitive to the sizes of gaps among the observations near the center of the sample (see Exercise 9).

The ordinary bootstrap tends not to work well for statistics such as the median or other quantiles that depend heavily on a small number of observations out of a larger sample.

VARIATION IN BOOTSTRAP DISTRIBUTIONS

For most statistics, almost all the variation in bootstrap distributions comes from the selection of the original sample from the population. Reducing this variation requires collecting a larger original sample.

Bootstrapping does not overcome the weakness of small samples as a basis for inference. Some bootstrap procedures are more accurate than others (we will discuss this later) and more accurate than common nonbootstrap procedures, but still they may not be accurate for very small samples. Use caution in any inference—including bootstrap inference—from a small sample.

The bootstrap resampling process using 1000 or more resamples introduces little additional variation, but for good accuracy use 10,000 or more.

5.9 HOW MANY BOOTSTRAP SAMPLES ARE NEEDED?

We suggested in Section 5.8 that 1000 bootstrap samples are enough for rough approximations, but more are needed for greater accuracy. We elaborate on this here. The focus here is on Monte Carlo accuracy—how well the usual random sampling implementation of the bootstrap approximates the theoretical bootstrap distribution.

A bootstrap distribution based on N random samples corresponds to drawing N observations with replacement from the theoretical bootstrap distribution.

Brad Efron, inventor of the bootstrap, suggested in 1993 that $N = 200$, or even as few as $N = 25$, suffices for estimating standard errors and that $N = 1000$ is enough for confidence intervals (Efron and Tibshirani (1993)).

We argue that more resamples are appropriate, on two grounds. First, those criteria were developed when computers were much slower; with faster computers it is much easier to take more resamples.

Second, those criteria were developed using arguments that combine the random variation due to the original sample with the random variation due to bootstrap sampling. We prefer to treat the data as given and look just at the variability due to bootstrap sampling. For typical 95% bootstrap percentile confidence intervals, to reduce Monte Carlo variability to the point that a supposed 95% confidence interval has a high probability of missing between 2.5% and ± 0.25% on each side requires about 15,000 bootstrap samples.

So for routine practice we recommend at least 10,000 bootstrap resamples and more when accuracy matters.

5.10 EXERCISES

For all exercises that ask you to perform exploratory data analysis (EDA), you should plot the data (histogram, normal quantile plots), describe the shape of the distribution (bell-shaped, symmetric, skewed, etc.), and provide summary statistics (mean, standard deviation). For bootstrapping questions, always provide plots and describe the shape, spread, and bias of the distribution.

1. Consider the sample 1–6. Use a six-sided die to obtain three different bootstrap samples and their corresponding means.
2. Consider the sample 1, 3, 4, 6 from some distribution.
   (a) For one random bootstrap sample, find the probability the mean is 1.
   (b) For one random bootstrap sample, find the probability the maximum is 6.
   (c) For one random bootstrap sample, find the probability that exactly two elements in the sample are less than 2.

   Assume order matters.
3. Consider the sample 1–3.
   (a) List all the bootstrap samples from this sample. How many are there?

for all North Carolina babies born in 2004; this is probably not the same as the sample mean. We have already mentioned that different samples of the same size will yield different sample means, so how can we gauge the accuracy of 3448.26 as an estimate to μ?

If we knew the sampling distribution of sample means for samples of size 1009 from the population of all 2004 North Carolina births, then this would give us an idea of how means vary from sample to sample, for the standard error of the sampling distribution tells us how far means deviate from the population mean μ. But of course, since we do not have *all* the birth weights, we cannot generate the sampling distribution (and if we did have all the weights, we would know the true μ!).

The bootstrap is a procedure that uses the given sample to create a new distribution, called the bootstrap distribution, that approximates the sampling distribution for the sample mean (or for other statistics).

We will begin by considering only a small subset of the birth weights—three observations, 3969, 3204, and 2892.

To find the bootstrap distribution of the mean, we draw samples (called *resamples* or *bootstrap samples*) of size n, with replacement, from the original sample and then compute the mean of each resample. In other words, we now treat the original sample as the population. For instance, Table 5.1 shows all $3^3 = 27$ samples of size 3, taking order into account (the notation x^* indicates a resampled observation, and $\bar{x}^*$ or $\hat{\theta}^*$ the statistic for a bootstrap sample).

The idea behind the bootstrap is that if the original sample is representative of the population, then the bootstrap distribution of the mean will look approximately like the sampling distribution of the mean; that is, have roughly the same spread and shape. However, the mean of the bootstrap distribution will be same as the mean of the *original sample* (Theorem A.5), not necessarily that of the original population.

THE BOOTSTRAP IDEA

The original sample approximates the population from which it was drawn. So resamples from this sample approximate what we would get if we took many samples from the population. The bootstrap distribution of a statistic, based on many resamples, approximates the sampling distribution of the statistic, based on many samples.

Thus, the standard deviation of all the resample means listed in Table 5.1 is 266 and we use this value as an estimate of the actual standard error (standard deviation of the true sampling distribution).

Of course, it is hard for three observations to accurately approximate the population. Let us work with the full data set; we draw resamples of size $n = 1009$ from the 1009 birth weights and calculate the mean for each.

There are now 1009^{1009} samples, too many for exhaustive calculation. Instead, we draw samples randomly, of size 1009 with replacement from the data, and calculate the mean for each. We repeat this many times, say 10,000, to create the bootstrap distribution. You can imagine that there is a table like Table 5.1 with 1009^{1009} rows, and we are randomly picking 10, 000 rows from that table.

TABLE 5.1 **All Possible Samples of Size 3 from 3969, 3204, and 2892.**

x_1^*	x_2^*	x_3^*	$\bar{x}^*$
3969	3969	3969	3969
3969	3969	3204	3714
3969	3969	2892	3610
3969	3204	3969	3714
3969	3204	3204	3459
3969	3204	2892	3355
3969	2892	3969	3610
3969	2892	3204	3355
3969	2892	2892	3251
3204	3969	3969	3714
3204	3969	3204	3459
3204	3969	2892	3355
3204	3204	3969	3459
3204	3204	3204	3204
3204	3204	2892	3100
3204	2892	3969	3355
3204	2892	3204	3100
3204	2892	2892	2996
2892	3969	3969	3610
2892	3969	3204	3355
2892	3969	2892	3251
2892	3204	3969	3355
2892	3204	3204	3100
2892	3204	2892	2996
2892	2892	3969	3251
2892	2892	3204	2996
2892	2892	2892	2892

BOOTSTRAP FOR A SINGLE POPULATION

Given a sample of size n from a population,

1. Draw a resample of size n with replacement from the sample. Compute a statistic that describes the sample, such as the sample mean.
2. Repeat this resampling process many times, say 10,000.
3. Construct the bootstrap distribution of the statistic. Inspect its spread, bias, and shape.

Remark One technical point—there are n^n samples where order matters, but only $\binom{2n-1}{n}$ unordered samples (see Exercise 5), a much smaller number. For exhaustive calculations we could use unordered samples and be careful to keep track of the

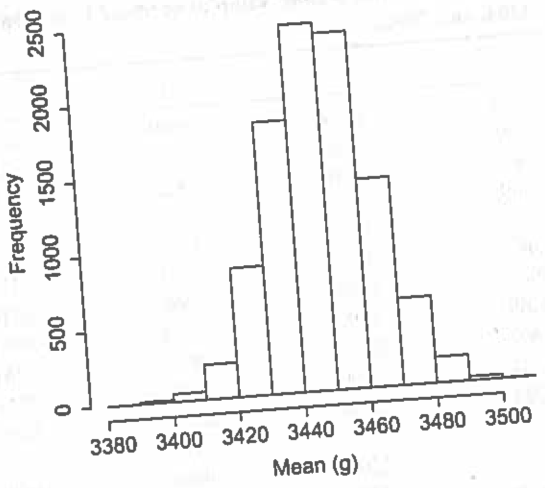

FIGURE 5.1 Bootstrap distribution of means for the North Carolina birth weights.

different probabilities for each sample. But when programming the random sampling procedure, it is easier to sample as if order matters.

In Figure 5.1, we note that the bootstrap distribution is approximately normal. Second, with mean 3448.206, it is centered at approximately the same location as the original mean, 3448.26. Third, we get a rough idea of the amount of variability. We can quantify the variability by computing the standard deviation of the bootstrap distribution, in this case 15.379. This is the bootstrap standard error.

For comparison, the standard deviation of the data is 487.736. The bootstrap standard error is smaller—this reflects the fact that an average of 1009 observations is more accurate (less variable) than is a single observation.

BOOTSTRAP STANDARD ERROR

The *bootstrap standard error* of a statistic is the standard deviation of the bootstrap distribution of that statistic.

To highlight some key features of the bootstrap distribution, we begin with two examples in which the theoretical sampling distributions of the mean are known.

Example 5.1 Consider a random sample of size 50 drawn from $N(23, 7^2)$. From Corollary A.2, we know the sampling distribution of the sample means is normal with mean 23 and standard error $\sigma/\sqrt{n} = 7/\sqrt{50} = 0.99$. Figure 5.2 shows the distribution of one such random sample with sample mean and standard deviation $\bar{x} = 24.13$, $s = 6.69$, respectively.

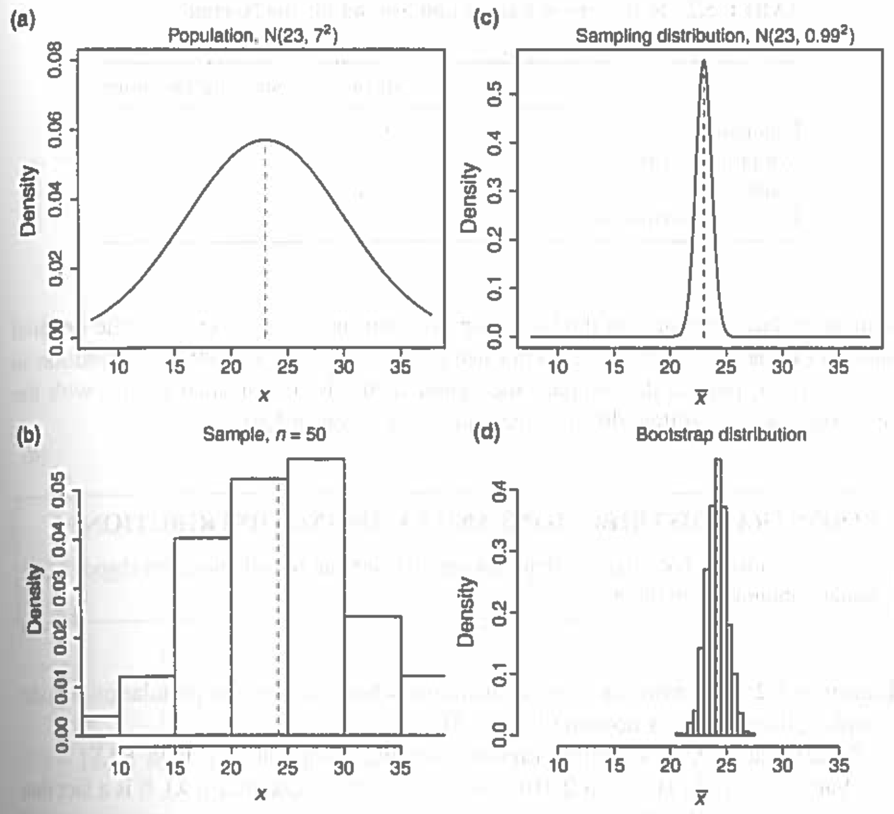

FIGURE 5.2 Sampling and bootstrap distributions of the mean for $N(23, 7^2)$. (a) The population distribution, $N(23, 7^2)$. (b) The distribution of one sample of size 50 from $N(23, 7^2)$. (c) The theoretical sampling distribution of $\bar{X}$, $N(23, 7^2/50)$. (d) The bootstrap distribution. Vertical lines mark the means.

In this example, we use software to run the algorithm on page 101 drawing 1000 resamples of size 50 from the original sample, and computing the mean of each resample. The bootstrap distribution is then the distribution of these 1000 resample means. The bootstrap standard error is the standard deviation of these 1000 resample means. We will also look at the center and shape of the bootstrap distribution.

From Figure 5.2, we can see that the bootstrap distribution has roughly the same spread and shape as the theoretical sampling distribution, but the centers are different. □

This example illustrates some important features of the bootstrap that hold for other statistics besides the mean: the bootstrap distribution of a particular statistic $\hat{\theta}$ has approximately the same spread and shape as the sampling distribution of the

TABLE 5.2 Summary of Center and Spread for the Normal Distribution Example

	Mean	Standard Deviation
Population	23	7
Sampling distribution of $\bar{X}$	23	0.99
Sample	24.13	6.69
Bootstrap distribution	24.15	0.92

statistic $\hat{\theta}$, but the center of the bootstrap distribution is at the center of the original sample (Table 5.2). Hence we do not use the center of the bootstrap distribution in its own right, but we do compare the center of the bootstrap distribution with the observed statistic; if they differ, it indicates *bias* (Section 5.6).

BOOTSTRAP DISTRIBUTIONS AND SAMPLING DISTRIBUTIONS

For most statistics, bootstrap distributions approximate the spread, bias, and shape of the actual sampling distribution.

Example 5.2 We now consider an example where neither the population nor the sampling distribution is normal (Table 5.3).

Recall that if X is a gamma random variable, Gamma(r, λ), then $E[X] = r/\lambda$ and $\text{Var}[X] = r/\lambda^2$ (Theorem B.10). Let $X_1, \ldots, X_n \sim \text{Gamma}(r, \lambda)$. It is a fact that that the sampling distribution of the mean $\bar{X}$ is Gamma($nr, n\lambda$) (a consequence of Theorem B.11 and Proposition B.3).

We draw a random sample of size $n = 16$ from the gamma distribution Gamma(1,1/2) (population mean 2, standard deviation 2.) Figure 5.3a and c shows a graph of the population and the distribution of one random sample ($\bar{x} = 2.73$ and $s = 2.61$), respectively. Figure 5.3b displays the theoretical sampling distribution of the means, Gamma(16, 8). Figure 5.3d displays the bootstrap distribution, based on 10,000 resamples.

Even though the distribution of the sample does not exactly match the population distribution, the bootstrap distribution is similar to the sampling distribution: it has roughly the same shape, a slightly larger spread (because the data have a slightly larger standard deviation than does the population), and the mean of the bootstrap distribution matches the empirical distribution rather than the population.

TABLE 5.3 Summary of Center and Spread for the Gamma Distribution Example

	Mean	Standard Deviation
Population	2	2
Sampling distribution of $\bar{X}$	2	0.5
Sample	2.74	2.61
Bootstrap distribution	2.73	0.642

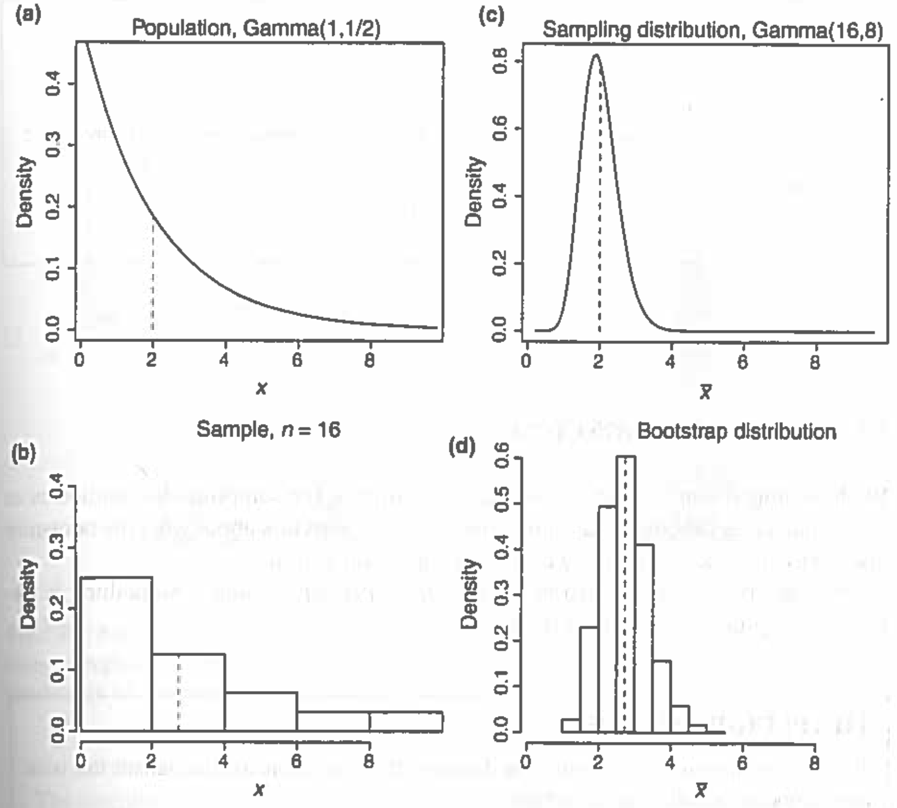

FIGURE 5.3 Sampling and bootstrap distribution from a gamma distribution. (a) The population distribution Gamma(1, 1/2). (b) A single sample of size 16 from Gamma(1, 1/2). (c) The theoretical sampling distribution of $\bar{X}$. (d) The bootstrap distribution for means of size 16, drawn from the sample.

R Note:

Draw a random sample of size 16 from Gamma(1, 1/2):

```
my.sample <- rgamma(16, 1, 1/2)
```

The following simulates a bootstrap distribution based on 10^5 resamples.

```
N <- 10^5
```

```
my.boot <- numeric(N)
for (i in 1:N)
{
  x <- sample(my.sample, 16, replace = TRUE) # draw resample
  my.boot[i] <- mean(x)               # compute mean, store in my.boot
}
hist(my.boot)
mean(my.boot)
sd(my.boot)
```

□

5.2 THE PLUG-IN PRINCIPLE

We have hinted that we use the bootstrap to estimate the sampling distribution or at least some things about the sampling distribution. Let us talk about what the bootstrap does, why it works, and what we can and cannot do with it.

The idea behind the bootstrap is the *plug-in principle*—that if something is unknown, we plug in an estimate for it.

THE PLUG-IN PRINCIPLE

To estimate a parameter, a quantity that describes the population, use the statistic that is the corresponding quantity for the sample.

This principle is used all the time in statistics. For example, the standard error for $\bar{X}$ calculated from i.i.d. observations from a population with standard deviation σ is $\sigma/\sqrt{n}$; when σ is unknown, we plug in an estimate s to obtain the usual standard error $s/\sqrt{n}$.

What is different in the bootstrap is that we plug in an estimate for the whole population, not just for a numerical summary of the population. We use the observed data as an estimate of the whole population; we will come to this in Section 5.2.1, and alternatives in Chapter 11, but for now we will continue with the main idea, that we plug in an estimate and what follows from that.

Our goal is to estimate a sampling distribution of some statistic. The sampling distribution depends on

1. the underlying population(s),
2. the sampling procedure (e.g., sampling with or without replacement), and
3. the statistic, such as $\bar{X}$.

Figure 5.4 contains a diagram of this process.

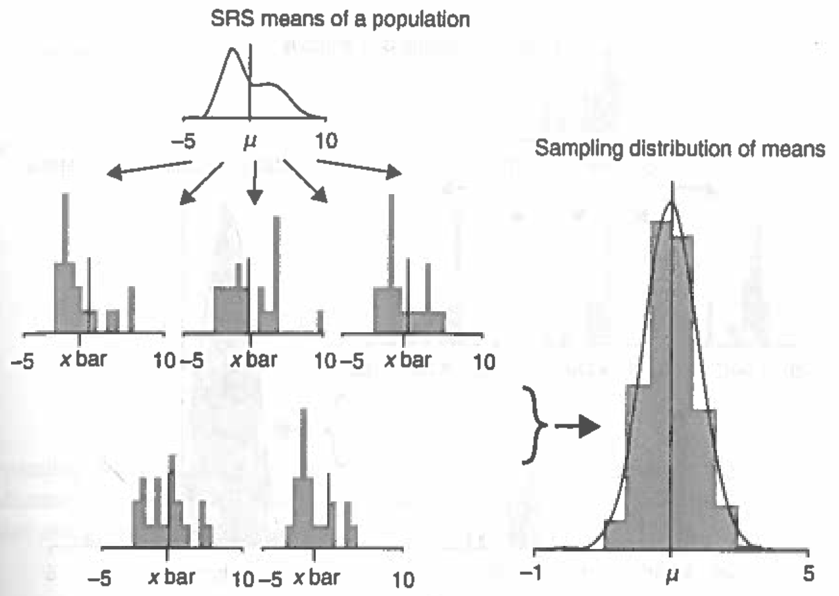

FIGURE 5.4 Diagram of the process of creating a sampling distribution. Many (infinitely many) samples are drawn from the population, a statistic like $\bar{X}$ is calculated for each. The distribution of the statistics is the sampling distribution.

The sampling distribution of a statistic is the result of drawing many samples from the population and calculating the statistic for each. The problem in most statistical applications is that the population is unknown.

The bootstrap principle is to plug in an estimate for the population and then mimic the real-life sampling procedure and statistic calculation (Figure 5.5). The bootstrap distribution depends on

1. an estimate for the population(s),
2. the sampling procedure, and
3. the statistic, such as $\bar{X}$.

In this chapter, we use the empirical distribution as an estimate for the population; that is, we use the empirical cumulative distribution function (Section 2.5) as an estimate for the unknown cumulative distribution function. This corresponds to sampling from the data. Let us look at this more closely.

5.2.1 Estimating the Population Distribution

We are using the data as an estimate for the population. Let us look more carefully at this; we will also see other alternatives in Chapter 11.

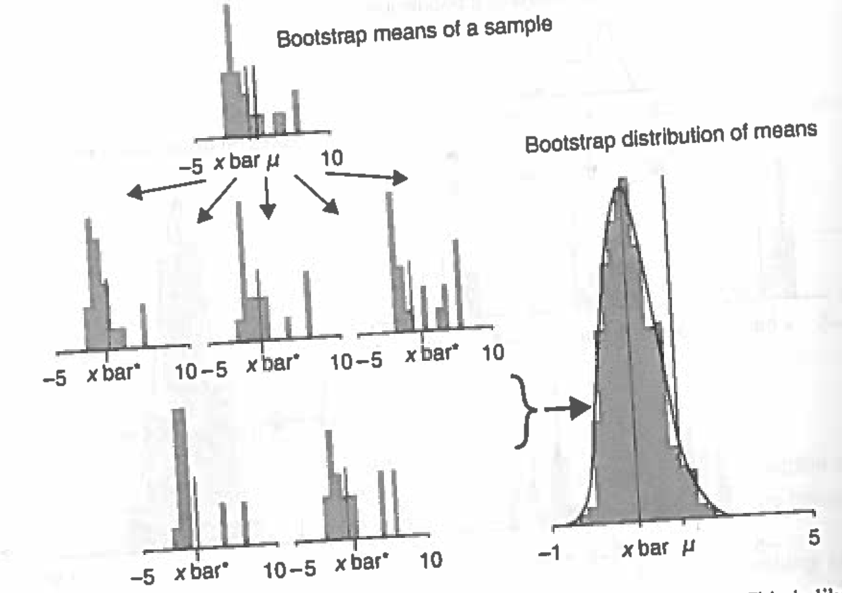

FIGURE 5.5 Diagram of the process of creating a bootstrap distribution. This is like Figure 5.4, except that the original data take the place of the population. We draw many samples from the original data, calculate $\bar{X}^*$ or another statistic for each, and collect the statistics to form the bootstrap distribution.

Let F and f denote the cdf and pdf for some unknown distribution and let $x_1, x_2, \ldots, x_n$ denote a random sample.

If we were willing to make assumptions about the population, say that it followed an exponential distribution, we could estimate the parameter λ from the data and then draw bootstrap samples from an exponential distribution with the estimated λ. This would be a parametric bootstrap, discussed in Section 11.2.

But most often when bootstrapping, we want to make as few assumptions as possible about the population. We want the data to tell us what it can, not introduce bias by making assumptions that may be wrong. So we resort to the empirical distribution, introduced in Section 2.5:

$$\hat{F}(s) = \frac{1}{n}\{\text{number of points } \leq s\}.$$

This is a discrete distribution, with probability $1/n$ at each of the observed data points, and empirical mass function

$$\hat{f}(s) = \frac{1}{n}\{\text{number of points } = s\}.$$

For instance, if the sample is 5, 5, 6, 10, 11, 11, 11, 12, then $\hat{f}(5) = 2/8$, $\hat{f}(6) = 1/8$, $\hat{f}(10) = 1/8$, $\hat{f}(11) = 3/8$, $\hat{f} = 1/8$, and $\hat{f}(s) = 0$ for all other values s.

For most bootstrap applications, we never bother to define or write down $\hat{F}$ and $\hat{f}$—instead we just need to know how to draw samples, which we do by sampling from the original observations, with equal probabilities $1/n$ for each.

In some cases we need the mean and variance of $\hat{F}$. For comparison, the mean for F is

$$\mathrm{E}_F[X] = \mu_F = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

or

$$\mathrm{E}_F[X] = \mu_F = \sum_x x\, f(x),$$

depending whether the population is continuous or discrete, where the subscript F indicates expected value based on F. Since $\hat{F}$ is discrete, we calculate the expected value using summation,

$$\begin{aligned}\mathrm{E}_{\hat{F}}[X] &= \mu_{\hat{F}}\\ &= \sum_x x\, \hat{f}(x)\\ &= \sum_{i=1}^{n} x_i(1/n) = \bar{x}.\end{aligned}$$

Similarly, the population variance under $\hat{F}$ is

$$\begin{aligned}\mathrm{Var}_{\hat{F}}[X] &= \sigma^2_{\hat{F}}\\ &= \mathrm{E}_{\hat{F}}[(X - \mu_{\hat{F}})^2]\\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2(1/n).\end{aligned}$$

This is like the sample variance s^2 (Equation 2.2), but with a denominator of n instead of $n - 1$.

5.2.2 How Useful Is the Bootstrap Distribution?

A fundamental question is how well the bootstrap distribution approximates the sampling distribution. We discuss this question in greater detail in Section 5.8, but note a few key points here.

First, the statistics that we bootstrap are generally *estimators*, statistics that estimate a parameter. For example, $\bar{X}$ is an estimator for μ, whereas a chi-squared test statistic (Equation 3.3) is not an estimator.

Definition 5.1 If $X_1, X_2, \dots, X_n$ are random variables from a distribution with parameter θ and $g(X_1, X_2, \dots, X_n)$ an expression used to estimate θ, then we call this function an *estimator*. ||

For most common estimators and under fairly general distribution assumptions, the following need to be noted:

Center The center of the bootstrap distribution is *not* an accurate approximation for the center of the sampling distribution. For example, the center of the bootstrap distribution for $\bar{X}$ is centered at approximately $\bar{x} = \mu_{\hat{F}}$, the mean of the sample, whereas the sampling distribution is centered at μ.

Spread The spread of the bootstrap distribution does reflect the spread of the sampling distribution.

Bias The bootstrap bias estimate (see Section 5.6) does reflect the bias of the sampling distribution. Bias occurs if a sampling distribution is not centered at the parameter.

Skewness The skewness of the bootstrap distribution does reflect the skewness of the sampling distribution.

The first point bears emphasis. It means that *the bootstrap is not used to get better parameter estimates* because the bootstrap distributions are centered around statistics $\hat{\theta}$ calculated from the data (e.g., $\bar{x}$) rather than the unknown population values (e.g., μ). Drawing thousands of bootstrap observations from the original data is not like drawing observations from the underlying population, it does not create new data.

Instead, the bootstrap sampling is useful for *quantifying the behavior of a parameter estimate*, such as its standard error, skewness, bias, or for calculating confidence intervals.

Example 5.3 Arsenic is a naturally occurring element in the groundwater of Bangladesh. However, much of this groundwater is used for drinking water by rural populations, so arsenic poisoning is a serious health issue. Figure 5.6 displays the distribution of arsenic concentrations from 271 wells in Bangladesh.[1]

The sample mean and standard deviation are $\bar{x} = 125.31$ and $s = 297.98$, respectively (measured in micrograms per liter).

We draw resamples of size 271 with replacement from the data and compute the mean for each resample. Figure 5.7 shows a histogram and normal quantile plot of the bootstrap distribution. The bootstrap distribution looks quite normal, with some skewness. This is the Central Limit Theorem (CLT) at work—when the sample size is large enough, the sampling distribution for the mean is approximately normal, even if the population is not normal.

[1] Data provided solely for illustrative purposes and to enable statistical analysis. Full data are available from the British Geological Survey web site http://www.bgs.ac.uk/arsenic/bphase2/datadownload.htm.

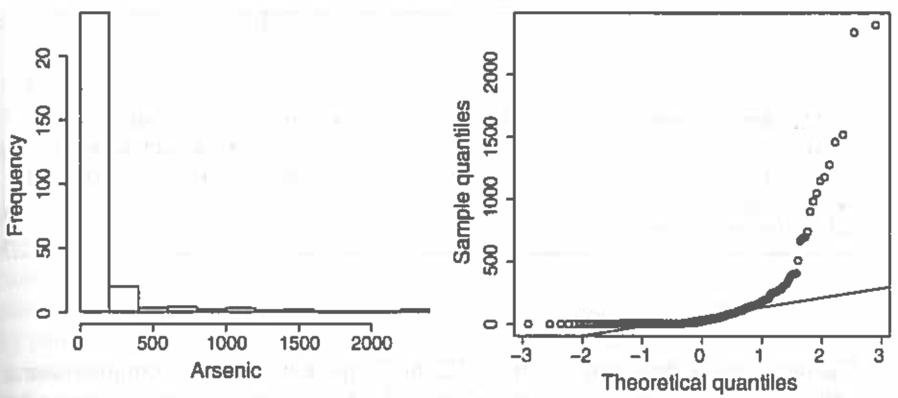

FIGURE 5.6 Arsenic levels in 271 wells in Bangladesh.

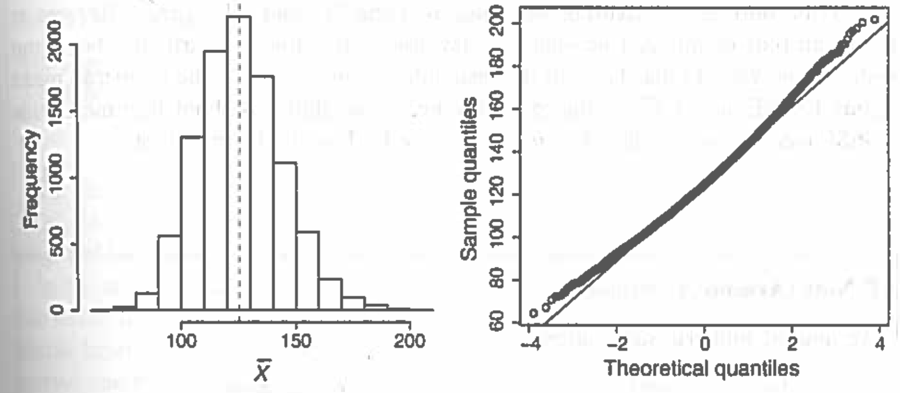

FIGURE 5.7 Histogram and QQ plot of the bootstrap distribution for the arsenic concentrations.

R Note:

Import the data set `Bangladesh` into R, then:

```
Arsenic <- Bangladesh$Arsenic
hist(Arsenic)
dev.new()                 # New graphics device
qqnorm(Arsenic)
qqline(Arsenic)

n <- length(Arsenic)
N <- 10^4
arsenic.mean <- numeric(N)
for (i in 1:N)
{
  x <- sample(Arsenic, n, replace = TRUE)
  arsenic.mean[i] <- mean(x)
}
hist(arsenic.mean, main = "Bootstrap distribution of means")
abline(v = mean(Arsenic), col = "blue", lty = 2) # observed mean
dev.new()                                # open new graphics device
qqnorm(arsenic.mean)
qqline(arsenic.mean)
```

The mean of the bootstrap means is 125.5375, quite close to the sample mean $\bar{x}$ (the difference is 0.2176, to four decimal places). The *bootstrap standard error* is the standard deviation of the bootstrap distribution; in this case, the bootstrap standard error is 18.2576.

For the normal distribution, we know that the 2.5 and 97.5 percentiles are at the mean plus or minus 1.96 standard deviations. But for this particular bootstrap distribution, we find that 1.5% of the resample means are below the bootstrap mean minus 1.96SE, and 3.4% of the resample means are above the bootstrap mean plus 1.96SE (see below). In this case, relying on the CLT would be inaccurate.

R Note (Arsenic, Continued):

We find the numeric summaries:

```
> mean(arsenic.mean)                      # bootstrap mean
[1] 125.5375
> mean(arsenic.mean)-mean(Arsenic)  # bias
[1] 0.2175773
> sd(arsenic.mean)                        # bootstrap SE
[1] 18.25759
```

Compute the points that are 1.96 standard errors from the mean of the bootstrap distribution:

```
> 125.5375-1.96*18.25759                # mark of 1.96SE from mean
[1] 89.75262
> 125.5375+1.96*18.25759
[1] 161.3224
> sum(arsenic.mean > 161.3224)/N
[1] 0.0337
> sum(arsenic.mean < 89.75262)/N
[1] 0.0153
```

5.3 BOOTSTRAP PERCENTILE INTERVALS

The sample mean $\bar{x}$ gives an estimate of the true mean μ, but it probably does not hit it exactly. It would be nice to have a *range* of values for the true μ that we are 95% sure includes the true μ.

In the North Carolina birth weights case study, the bootstrap distribution (Figure 5.1) shows roughly how sample means vary for samples of size 1009. If most of the sample means are concentrated within a certain interval of the bootstrap distribution, it seems reasonable to assume that the true mean is most likely somewhere in that same interval. Thus, we can construct what is called a 95% confidence interval by using the 2.5 and 97.5 percentiles of the bootstrap distribution as endpoints. We would then say that we are 95% confident that the true mean lies within this interval. These are *bootstrap percentile confidence intervals*.[2]

BOOTSTRAP PERCENTILE CONFIDENCE INTERVALS

The interval between 2.5 and 97.5 percentiles of the bootstrap distribution of a statistic is a 95% *bootstrap percentile confidence interval* for the corresponding parameter.

For the NC birth weights, the interval marked by the 2.5 and 97.5 percentiles is (3419, 3478). Thus, we would state that we are 95% confident that the true mean weight of NC babies born in 2004 is between 3419 and 3478 g.

In the arsenic example, the 2.5% and 97.5% points of the bootstrap distribution give us the interval (92.9515, 164.4418), so we are 95% confident that the true mean arsenic level is between 92.95 and 164.44 μg/L. Note that with $\bar{x} = 125.5375$, this interval can be written $(\bar{x} - 32.586, \bar{x} + 38.9043)$; in particular, this interval is *not* symmetric about the mean, reflecting the asymmetry of the bootstrap distribution.

R Note:

```
> quantile(arsenic.mean, c(0.025, 0.975))
   2.5%    97.5%
92.9515 164.4418
```

The arsenic data illustrate an interesting point. A good confidence interval for the mean need not necessarily be symmetric: an endpoint will be farther from the sample mean in the direction of any outliers. A confidence interval is an insurance policy: rather than relying on a single statistic, the sample mean, as an estimate of μ, we give a range of plausible values for μ. We can see that there are some extremely

[2] We will discuss the logic of confidence intervals more formally in Chapter 7.

large arsenic measurements: of the 271 observations, 8 are above 1000 μg/L and 2 are above 2200 μg/L (remember, the sample mean is only 125.31!). What we do not know is just how huge arsenic levels in the population can be, or how many huge ones there are. It could be that huge observations are *underrepresented* in our data set. In order to protect against this—that is, to have only a 2.5% chance of missing a true big mean, the interval of plausible values for μ must stretch far to the right. Conversely, there is less risk of missing the true mean on the low side, so the left endpoint need not be as far away from the mean.

5.4 TWO SAMPLE BOOTSTRAP

We now turn to the problem of comparing two samples. In general, bootstrapping should mimic how the data were obtained. So if the data correspond to independent samples from two populations, we should draw two samples that way. Then we proceed to compute the same statistic comparing the samples as for the original data, for example, difference in means or ratio of proportions.

BOOTSTRAP FOR COMPARING TWO POPULATIONS

Given independent samples of sizes m and n from two populations,

1. Draw a resample of size m with replacement from the first sample and a separate resample of size n from the second sample. Compute a statistic that compares the two groups, such as the difference between the two sample means.
2. Repeat this resampling process many times, say 10,000.
3. Construct the bootstrap distribution of the statistic. Inspect its spread, bias, and shape.

Example 5.4 A high school student was curious about the total number of minutes devoted to commercials during any given half-hour time period on basic and extended cable TV channels (Rodgers and Robinson (2004)). Table 5.4 contains some data that he collected to study this.

The means of the basic and extended channel commercial times are 9.21 and 6.87 min, respectively, so on average, commercials on basic channels are 2.34 min longer than on extended channels. Is this difference of 2.34 min statistically significant? The poor student could only stand to watch 10 h of random TV, so his observations may not accurately represent the TV universe.

TABLE 5.4 Length of Commercials (Minutes) During Random Half-Hour Periods from 7a.m. to 11p.m.

Basic	7.0	10.0	10.6	10.2	8.6	7.6	8.2	10.4	11.0	8.5
Extended	3.4	7.8	9.4	4.7	5.4	7.6	5.0	8.0	7.8	9.6

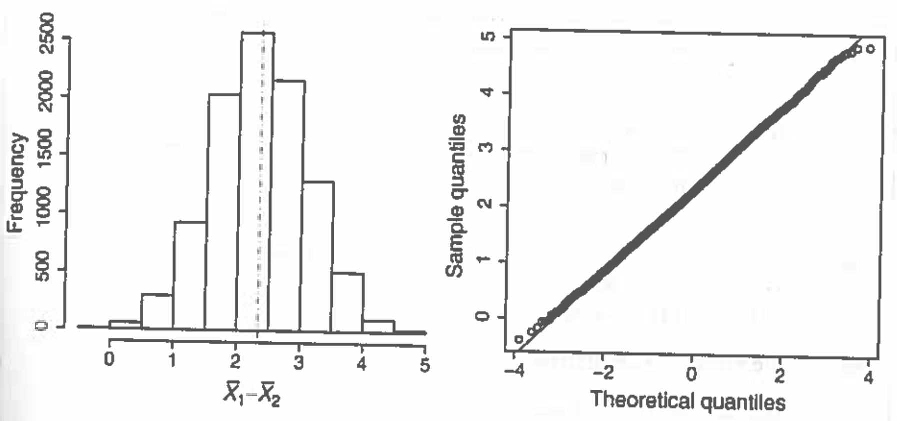

FIGURE 5.8 Histogram and normal quantile plot of the bootstrap distribution for the difference in mean advertisement time in basic versus extended TV channels. The vertical line in the histogram marks the observed mean difference.

The original data are simple random samples of size 10 from two populations. We draw a bootstrap sample from the basic channel data and independently draw a bootstrap sample from the extended channel data, compute the means for each sample, and take the difference.

Figure 5.8 shows the bootstrap distribution of the difference of sample means. As in the single-sample case, we see that the bootstrap distribution is approximately normal and centered at the original statistic (the difference in sample means). We also get a quick idea of how much the difference in sample means varies due to random sampling. We may quantify this variation by computing the bootstrap standard error, which is 0.76. Again, the bootstrap standard error is the standard error of the sampling distribution.

The right panel of Figure 5.8 shows a normal quantile plot for the bootstrap distribution: the distribution is very close to normal.

R Note:

Import the TV data into R, then

```
times.Basic <- subset(TV, select = Times,
                      subset = Cable == "Basic", drop = T)
times.Ext <- subset(TV, select = Times,
                    subset = Cable == "Extended", drop = T)
N <- 10^4
times.diff.mean <- numeric(N)
for (i in 1:N)
{
  Basic.sample <- sample(times.Basic, 10, replace = TRUE)
  Ext.sample <- sample(times.Ext, 10, replace = TRUE)
```