Chapter 6

THE BOOTSTRAP

When testing hypotheses (or computing confidence intervals) with the one-sample Student's T method described in Chapter 5, the central limit theorem tells us that Student's T performs better as the sample size gets large. That is, under random sampling the discrepancy between the nominal and actual Type I error probability will go to zero as the sample size goes to infinity. But unfortunately, for reasons outlined in Chapter 5, there are realistic situations where about 200 observations are needed to get satisfactory control over the probability of a Type I error, or accurate probability coverage when computing confidence intervals. When comparing the population means of two groups of individuals, using Student's T is known to be unsatisfactory when sample sizes are small or even moderately large. In fact, it might be unsatisfactory no matter how large the sample sizes happen to be because under general conditions it does not converge to the correct answer (e.g., Cressie and Whitford, 1986). Switching to the test statistic W, given by Equation (5.3), the central limit theorem now applies under general conditions, so using W means we will converge to the correct answer as the sample sizes get large, but in some cases we again need very large sample sizes to get accurate results. (There are simple methods for improving the performance of W using what are called estimated degrees of freedom, but the improvement remains highly unsatisfactory for a wide range of situations.) Consequently, there is interest in finding methods that beat our reliance on the central limit theorem as it applies to these techniques. That is, we would like to find a method that, in theory at least, converges to the correct answer more quickly as the sample sizes get large, and such a method is described here. (The so-called empirical likelihood method also achieves this goal and is discussed in Chapter12.)

For various reasons, problems with making accurate inferences about the association between two variables are much more difficult than when comparing measures of location. Equation (4.5) of Chapter 4 described Laplace's method for computing a confidence interval for the slope of the least-squares

R.R. Wilcox, *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*, DOI 10.1007/978-1-4419-5525-8_6,

regression. Today a slight variation of this method is used (which was outlined in Chapter 5). But even under normality, we will see that the conventional extension of Laplace's method has serious practical problems in terms of achieving accurate probability coverage. A relatively effective method for dealing with this problem is described in this chapter.

In applied work, it is very common to focus attention not on the slope of a regression line, but instead on what is known as Pearson's correlation coefficient. This chapter introduces this coefficient and notes that problems with making inferences about the slope of a regression line extend to it. Fortunately, there are substantially better methods for making inferences about this correlation coefficient, which will be described. But, unfortunately, there are other more intrinsic problems with this coefficient, described in Chapters 7 and 10, that must also be addressed.

6.1 TWO BOOTSTRAP METHODS FOR MEANS

Both theory and simulation studies tell us that a certain form of a relatively modern method generally offers the improvements we seek when computing a confidence interval or testing hypotheses. It is called a *bootstrap* method, two variations of which are covered here. The bootstrap was first proposed by Julian Simon in 1969, and it was discovered independently a short while later by Brad Efron. It was primarily Efron's work that spurred interest in the method. Based on over 1,000 journal articles, all indications are that the bootstrap has great practical value and should be seriously considered in applied work. It is not a panacea, but when combined with other modern insights (covered in Part II), highly accurate results can be obtained in situations where more traditional methods fail miserably.

The basic idea behind all bootstrap methods is to use the data obtained from a study to approximate the sampling distributions used to compute confidence intervals and test hypotheses. When working with means, for example, one version of the bootstrap uses the data to estimate the probability curve associated with T. This is in contrast to the standard strategy of assuming that, due to normality, this probability curve has a specified form that is completely determined by the sample size only. The other version described here, called the *percentile bootstrap*, estimates the sampling distribution of the sample mean instead. Initially, attention is focused on how the bootstrap is used with means, but it generalizes to all of the applied problems considered in this book.

6.1.1 The Percentile Method

To describe the percentile bootstrap method, we begin with a quick review of a sampling distribution as described in Chapter 5. Consider a single population of individuals from which we randomly sample n observations yielding a sample mean, $\bar{X}$. If we obtain a new sample of subjects, in general we get a different sample mean. The sampling distribution of the sample mean refers to the probability that $\bar{X}$ will be less than 2, less than 6, or less than c for any c we might pick. Put another way, there is uncertainty about the value for the sample mean we will get when collecting data, and the sampling distribution of the sample mean refers to the corresponding probability curve.

Next we consider the notion of a sampling distribution from the point of view that probabilities are relative frequencies. If we could repeat a study billions of times, yielding billions of sample means, a certain proportion of the sample means will be less than 2, less than 6, or less than c. If 10% of the sample means are less than 2, we say that the probability of getting a sample mean less than 2 is 0.1. If the proportion less than 6 is 70%, we take this to mean that the probability of conducting a study and getting a sample mean less than 6 is 0.7. What is important from an applied point of view is that if we know these probabilities, we can compute confidence intervals and test hypotheses about the population mean. But obviously we cannot, in most cases, repeat an experiment even two times, let alone billions of times, so it might seem that this description of the sampling distribution has no practical value. However, this description sets the stage for describing the basic strategy behind the bootstrap.

Although we do not know the probability curve that generates observations, it can be estimated from the data at hand. This suggests a method for repeating our experiment without acquiring new observations. For instance, imagine we conduct a study aimed at rating the overall mental health of college students, so we administer a standard battery of tests and come up with the following 20 ratings:

$$2,\ 4,\ 6,\ 6,\ 7,\ 11,\ 13,\ 13,\ 14,\ 15,\ 19,\ 23,\ 24,\ 27,\ 28,\ 28,\ 28,\ 30,\ 31,\ 43.$$

The sample mean of these 20 ratings is $\bar{X} = 18.6$. Based on these 20 values, we estimate that the probability of observing the value 2 is 1/20 because exactly one of the twenty observations is equal to 2. In a similar manner, two observations have the value 6, so we estimate that the probability of observing a 6 is 2/20. The probability of getting the value 5 is estimated to be zero because the value 5 was not observed. Obviously, these estimates will differ from the actual probabilities, but the issue is whether these estimated probabilities can be used to get more accurate confidence intervals or better control over the probability of a Type I error.

This estimate of the probability curve suggests the following strategy for estimating the probability curve associated with the sample mean. First, randomly sample, with replacement, 20 observations from the 20 values just listed. In our illustration, this means that each time we sample an observation, the value 2 occurs with probability 1/20, the value 4 occurs with probability 1/20, the value 6 occurs with probability 2/20, and so on. That is, we take the observed relative frequencies to be the probabilities. The resulting 20 observations are called a *bootstrap sample.* For example, we might get

14, 31, 28, 19, 43, 27, 2, 30, 7, 27, 11, 13, 7, 14, 4, 28, 6, 4, 28, 19,

and in fact this bootstrap sample was generated on a computer using the original ratings. The mean of this bootstrap sample, called a *bootstrap sample mean,* is $\bar{X}^* = 18.1$, where the notation $\bar{X}^*$ is used to make a clear distinction with the sample mean from our study, $\bar{X} = 18.6$. If we repeat the process of generating a bootstrap sample, we will get a different bootstrap sample mean. And if we repeat this process, say, 600 times, we will have 600 bootstrap sample means. Moreover, if 60 of the 600 bootstrap sample means are less than 3, then this suggests that if we were to actually repeat our study, as opposed to generating bootstrap samples, our estimate is that with probability $60/600 = .1$, we will get a sample mean less than 3. Of course, this estimate will be wrong. The only goal for the moment is to convey the flavor of the percentile bootstrap: Pretend that the observed values give an accurate estimate of the probability curve and then generate bootstrap sample means in an attempt to approximate the sampling distribution of $\bar{X}$.

Another example might help. Using a computer, let's generate 20 observations from a standard normal curve ($\mu = 0$ and $\sigma = 1$). Theory tells us that the sampling distribution of the sample mean is normal with mean 0 and variance 1/20. But imagine we do not know this and we use the bootstrap to estimate the probability curve using the 20 observations we just generated. This means we repeatedly generate bootstrap samples from these 20 observations and compute a bootstrap sample mean. For illustrative purposes, let's generate 600 bootstrap sample means. Then we plot the bootstrap sample means and compare it to the exact probability curve for the sample mean. That is, we graphically compare the bootstrap estimate of the probability curve to the correct curve. The results are shown in Figure 6.1. As we see, in this particular case the two curves happen to be fairly similar. That is, the bootstrap method gives a reasonable approximation of the true probability curve. But, of course, this one illustration is not convincing evidence that the bootstrap has practical value. Indeed, Figure 6.1 indicates that the plot of the bootstrap sample means does not extend out as far as it should. That is, the probability curve is too light-tailed compared to the correct probability curve being estimated. This foreshadows a problem that must be addressed.

Notice that when we generate bootstrap sample means, they will tend to be centered around the sample mean from our study if each bootstrap sample mean is based on a reasonably large number of observations. That is, a version of the central limit theorem applies to the bootstrap sample means. In the last example, the sample mean is $\bar{X} = 0.01$, so the bootstrap sample means will tend to be centered around 0.01 rather than around the population mean, 0. So, of course, if the sample mean happens to be far from the population mean, the bootstrap sample means will be centered around a value that is far from the population mean as well. Despite this, it will generally be the case that

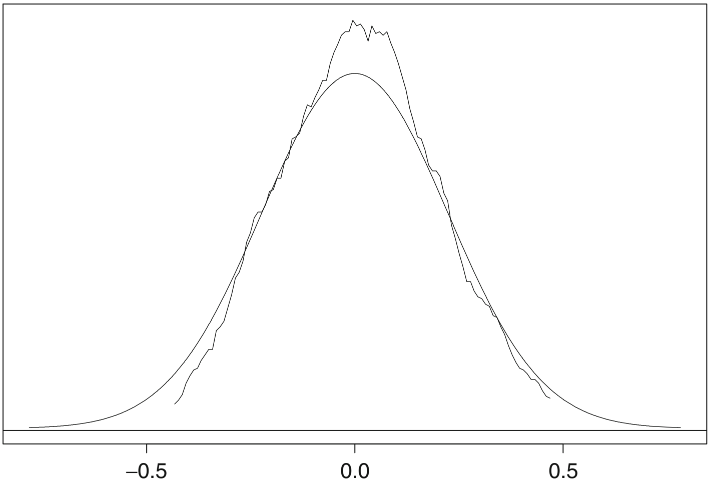

Figure 6.1: Shown are two probability curves. The smooth symmetric curve is what theory tells us we should get for the sampling distribution of the sample mean based on 20 observations. The ragged line is the bootstrap approximation of this curve based on 20 observations randomly sampled from a normal curve.

the middle 95% of the bootstrap sample means will contain the population mean, provided a reasonably large number of observations is available. In our last example, the middle 95% of the bootstrap sample means extend from −0.35 to 0.39, this interval contains 0, and this suggests that we should not rule out the possibility that $\mu = 0$.

Suppose we take the middle 95% of the bootstrap sample means as a 0.95 confidence interval for the population mean. In our last example, we are taking the interval $(-0.35, 0.39)$ to be a .95 confidence interval for μ. This is an example of a *percentile bootstrap confidence interval* for the population mean. Furthermore, consider the rule: Reject the hypothesis H_0: $\mu = 0$ if the bootstrap confidence interval does not contain 0. It can be shown that this rule is reasonable—it can be theoretically justified—provided that the sample size is sufficiently large. That is, if we want the probability of a Type I error to be 0.05, this will be approximately true if a reasonably large sample size is available.

Returning to the mental health ratings of college students, Figure 6.2 shows a plot of 1,000 bootstrap sample means. As indicated, the middle 95% of the bootstrap means lie between 13.8 and 23.35. So the interval (13.8, 23.35) corresponds to a 0.95 percentile bootstrap confidence interval for the unknown population mean.

Unfortunately, when computing confidence intervals for the population mean based on the percentile bootstrap method, large sample sizes are

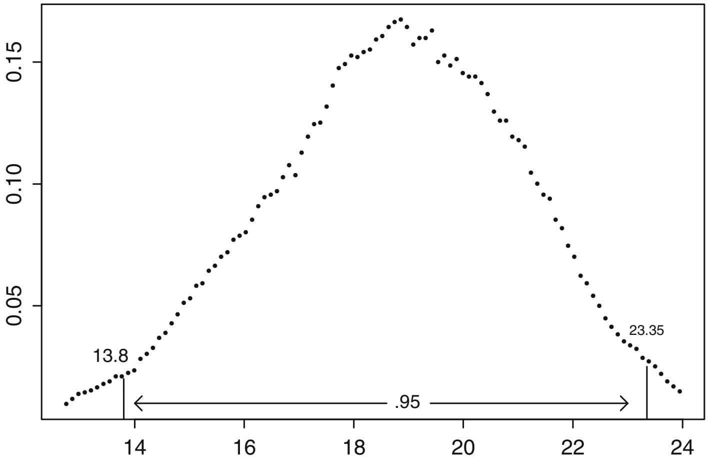

Figure 6.2: Shown is an approximation of the probability curve of the sample mean based on 1,000 bootstrap sample means generated from the ratings data. The middle 95% of the bootstrap sample means lie between 13.8 and 23.35, suggesting that the interval (13.8, 23.35) be used as an approximate 0.95 confidence interval for the population mean.

required to get accurate probability coverage, so we have not yet made any practical progress. But despite this, the percentile bootstrap will be seen to have value for other problems we will consider. The only goal here is to describe the percentile bootstrap for the simplest case.

6.1.2 The Bootstrap *t* Method

Another form of the bootstrap method arises as follows. Recall that when computing a confidence interval for μ, a solution is obtained by assuming that

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has a Student's t distribution. If, for example, $n = 25$, it can be shown that when sampling from a normal curve, there is a 0.95 probability that T will be between -2.064 and 2.064. This result can be used to show that a 0.95 confidence interval for the population mean is

$$\left(\bar{X} - 2.064\frac{s}{\sqrt{n}}, \bar{X} + 2.064\frac{s}{\sqrt{n}}\right)$$

when sampling from a normal distribution. The point is that assuming normality provides an approximation of the probability curve for T that in turn yields an approximate 0.95 confidence interval when sampling from non-normal distributions. But as previously indicated, a practical concern is that this approximation of the probability curve for T performs poorly in some cases, which in turn means we get inaccurate confidence intervals, poor control over the probability of a Type I error, and undesirable power properties. If we could determine the probability curve for T without assuming normality, the problems associated with Type I errors and probability coverage would be resolved. What we need is a better way of approximating the distribution of T.

The *bootstrap t* method, sometimes called a *percentile t bootstrap*, approximates the distribution of T as follows. First, obtain a bootstrap sample as was done when applying the percentile bootstrap method. For this bootstrap sample, compute the sample mean and standard deviation and label the results $\bar{X}^*$ and s^*. As an illustration, consider again the study aimed at assessing the overall mental health of college students based on the 20 ratings

$$2, 4, 6, 6, 7, 11, 13, 13, 14, 15, 19, 23, 24, 27, 28, 28, 28, 30, 31, 43.$$

For the bootstrap sample previously considered, namely

$$14, 31, 28, 19, 43, 27, 2, 30, 7, 27, 11, 13, 7, 14, 4, 28, 6, 4, 28, 19,$$

we get $\bar{X}^* = 18.1$ and a bootstrap standard deviation of $s^* = 11.57$. Next, compute

$$T^* = \frac{\bar{X}^* - \bar{X}}{s^*/\sqrt{n}}. \tag{6.1}$$

In the illustration,

$$T^* = \frac{18.1 - 18.6}{11.57/\sqrt{20}} = -0.19.$$

Repeat this process B times, each time computing T^*. Figure 6.3 shows a plot of $B = 1{,}000$ values obtained in this manner. These B values provide an approximation of the distribution of T without assuming normality.

As indicated by Figure 6.3, 95% of these 1,000 values lie between −2.01 and 2.14. If instead we assume normality, then 95% of the T values would be between −2.09 and 2.09. So in this particular case, there is little difference between the bootstrap t and assuming normality.

Here is a summary of how to compute a 0.95 confidence interval for the mean using the bootstrap t method:

1. Compute the sample mean, $\bar{X}$, and standard deviation, s.

2. Generate a bootstrap sample by randomly sampling with replacement n observations from $X_1, \ldots, X_n$, yielding $X_1^*, \ldots, X_n^*$.

3. Use the bootstrap sample to compute T^* given by Equation (6.1).

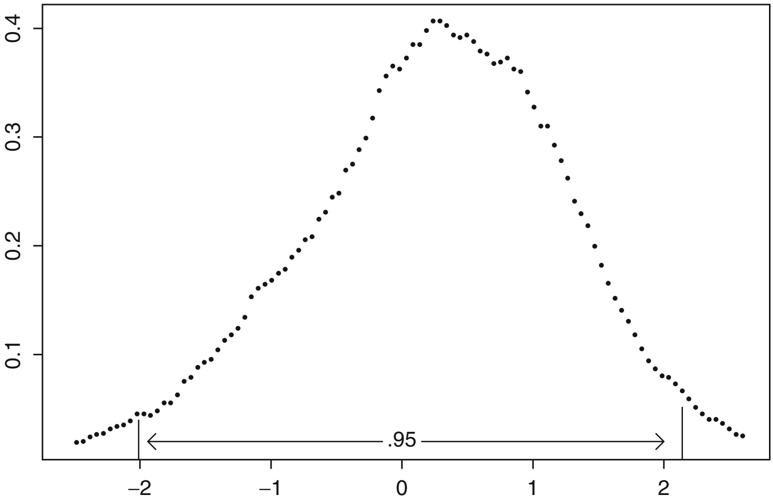

Figure 6.3: A bootstrap estimate of the sampling distribution of T based on the ratings data. The middle 95% of the bootstrap T values lie between -2.01 and 2.14. When sampling from a normal distribution, the T values will lie between -2.09 and 2.09 with probability 0.95. So in this case, the bootstrap t is in close agreement with what we get when assuming normality.

4. Repeat steps 2 and 3 B times yielding $T_1^*, \dots, T_B^*$. For $\alpha = 0.05$, B must be fairly large when working with means. Based on results in Hall (1986), the choice $B = 999$ is recommended rather than the seemingly more natural choice of $B = 1{,}000$. For n small (less than 100), unsatisfactory probability coverage can result when working with means, and increasing B seems to offer little or no advantage in terms of controlling the probability of a Type I error. Smaller values for B can provide good control over the Type I error probability when using some of the methods described in subsequent chapters. However, in terms of power, there are results indicating that choosing B to be relatively large can have practical value (e.g., Jöckel, 1986; Hall and Titterington, 1989; Racine and MacKinnon, 2007).
5. Write the bootstrap T^* values in ascending order, yielding $T_{(1)}^* \leq \cdots \leq T_{(B)}^*$.
6. Set $L = .025B$, $U = .975B$ and round both L and U to the nearest integer.

The bootstrap t confidence interval for μ (also called a bootstrap percentile t interval) is

$$\left(\bar{X} - T_{(U)}^* \frac{s}{\sqrt{n}}, \bar{X} - T_{(L)}^* \frac{s}{\sqrt{n}}\right).$$

[For readers familiar with basic statistics, $T^*_{(L)}$ will be negative, and that is why $T^*_{(L)}s/\sqrt{n}$ is subtracted from the sample mean. Also, it might seem that $T^*_{(L)}$ should be used to define the lower end of the confidence interval, but it can be seen that this is not the case.] In the illustration where $\bar{X} = 18.6$ and $s = 11.14$, a 0.95 confidence interval for the mean based on the bootstrap t method (using software mentioned in the final chapter, which indicates that $T^*_{(U)} = 2.08$ and $T^*_{(L)} = -2.55$) is

$$\left(18.6 - 2.08\frac{11.14}{\sqrt{20}},\, 18.6 + 2.55\frac{11.14}{\sqrt{20}}\right) = (13.42, 24.95).$$

If instead normality is assumed, the confidence interval is (13.3, 23.9).

An important issue is whether the bootstrap t ever gives a substantially different result than assuming normality. If it never makes a difference, of course there is no point in abandoning Student's T for the bootstrap t. The following example, based on data taken from an actual study, illustrates that substantial differences do indeed occur.

M. Earleywine conducted a study on hangover symptoms after consuming a specific amount of alcohol in a laboratory setting. For one of the groups, the results were

0, 32, 9, 0, 2, 0, 41, 0, 0, 0, 6, 18, 3, 3, 0, 11, 11, 2, 0, 11.

(These data differ from the data used to create Figure 5.6, but they are from the same study.) Figure 6.4 shows the bootstrap distribution of T based on $B = 999$ bootstrap samples. The middle 95% of the T^* values are between −4.59 and 1.61. If we assume normality, then by implication, the middle 95% of the T values will be between −2.09 and 2.09 instead. Figure 6.4 also shows the distribution of T assuming normality. As is evident, there is a substantial difference between the two methods. The 0.95 confidence interval based on the bootstrap t method is (−3.13, 11.5), and it is (2.2, 12.7) when assuming normality.

Using yet another set of data from the same study, namely,

0, 0, 0, 0, 0, 0, 0, 0, 1, 8, 0, 3, 0, 0, 32, 12, 2, 0, 0, 0,

the middle 95% of the T values are estimated to lie between −13.6 and 1.42. That is, there is an even bigger discrepancy between the bootstrap and what we get assuming normality.

Because the two methods for computing confidence intervals can differ substantially, there is the issue of which one you should use. If distributions are normal, Student's T offers a very slight advantage. In general, however, including situations where distributions are nonnormal, it seems never to offer a substantial advantage. In contrast, the bootstrap t offers a substantial advantage over Student's T in various realistic situations, so it deserves serious consideration in applied work. A reasonable suggestion is to use Student's T

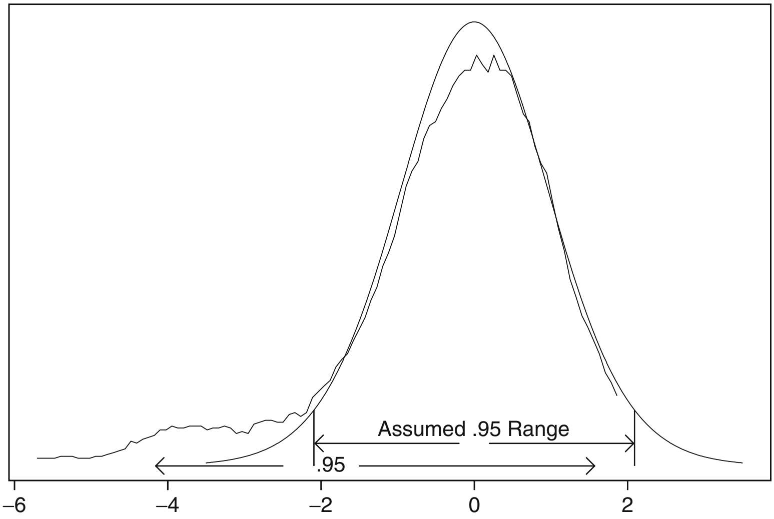

Figure 6.4: An illustration that in applied work, the approximation of the probability curve for T, based on the bootstrap, can differ substantially from the approximation based on the normal curve.

if a distribution seems to be approximately normal, or if the sample size is sufficiently large, but this is too vague. How close to a normal distribution must it be? Currently, there is no satisfactory answer to this question. We can make an educated guess that with 200 observations, Student's T will perform as well as the bootstrap in most situations, but there is no proof that this is the case. So the best advice seems to be always to use the bootstrap t when making inferences about a mean.

6.2 TESTING HYPOTHESES

Bootstrap confidence intervals can be used to test hypotheses. Basically, you proceed as was indicated in Chapter 5 and reject if the hypothesized value is not contained in the confidence interval. If in the last example there is interest in testing H_0: $\mu = 12$, the bootstrap t method would reject with $\alpha = 0.05$ because the 0.95 confidence interval, $(-3.13, 11.5)$, does not contain the hypothesized value, 12. In contrast, Student's T would not reject, because its 0.95 confidence interval, (2.2, 12.8), contains 12.

6.2.1 Why Does the Bootstrap *t* Perform Well Compared to Student's *T*?

The bootstrap t does not always provide more accurate results than Student's T. (Under normality, for example, Student's T is more accurate.) But for a wide range of situations, the bootstrap t is preferable, which is not surprising based on certain theoretical results. To explain, first consider Student's T. Whenever we use it to compute a 0.95 confidence interval, there is generally some discrepancy between 0.95 and the actual probability coverage. Or, when testing some hypothesis with the goal that the probability of a Type I error be 0.05, the actual probability of a Type I error generally differs from 0.05 due to nonnormality. Mathematicians are able to characterize how quickly this discrepancy goes to zero as the sample size gets large. The rate is $1/\sqrt{n}$. That is, $1/\sqrt{n}$ goes to zero as the sample size, n, gets large, and this provides some indication of how quickly errors made with Student's T go to zero as well. This does *not* mean that the difference between the nominal and actual Type I error probability is $1/\sqrt{n}$—we have already seen that in some cases we need two hundred observations when using Student's T. But this ratio is used by mathematicians to measure how well a given method performs.

The point is that when using the bootstrap t, the discrepancy between the actual and nominal Type I error probability goes to zero at the rate $1/n$—it goes to zero more quickly than when using Student's T. So from a theoretical perspective, the bootstrap t beats Student's T. Unfortunately, this by itself is not convincing evidence that in applied work, the bootstrap t beats Student's T when sampling from nonnormal probability curves. The reason is that with small sample sizes, it is not remotely obvious how the performance of the bootstrap t will compare to Student's T based on this one theoretical result. The theory is based on a large-sample comparison of the methods and might give a very poor indication of how they compare when sample sizes are small or even moderately large. Moreover, this result does *not* tell us how large of a sample we need to get accurate results with the bootstrap t. Quantitative experts use simulation studies to answer these questions. The good news is that in simulation studies, typically the bootstrap t performs about as well, and in some cases much better than Student's T. Moreover, there are no indications that Student's T ever offers a substantial improvement over the bootstrap t. The bad news is that when working with the mean, although we get increased accuracy, situations arise where the control over the probability of a Type I error remains unsatisfactory. For example, in Chapter 5 we saw a situation where, when testing a hypothesis about the mean, we need about 200 observations to get accurate control over the probability of a Type I error. If we switch to the bootstrap t, we reduce the required number of observations to 100. So substantial progress has been made, but more needs to be done. We have seen that in some situations, Student's T is biased—its power might actually decline as we move away from the null hypothesis. The bootstrap t reduces this problem as well, but unfortunately it does not eliminate it. Moreover, when using the bootstrap t with means, a fundamental problem described in Chapter 7 remains.

6.3 COMPARING TWO INDEPENDENT GROUPS

The bootstrap methods described in the previous section are easily extended to the problem of comparing two independent groups. Recall from Chapter 5 that Student's T for comparing means assumes groups have equal variances, even when the means differ. One possibility is to use a bootstrap analog of Student's T test, but this approach is not described because it does not correct the technical problems associated with violating the assumption of equal variances. One of the better methods for comparing means is to use a bootstrap t based on the test statistic W given by Equation (5.3). To compute a 0.95 confidence interval for $\mu_1 - \mu_2$, proceed as follows:

1. Compute the sample mean and standard deviation for each group and label the results $\bar{X}_1$ and s_1 for group 1, and $\bar{X}_2$ and s_2 for group 2. Set $d_1 = s_1^2/n_1$ and $d_2 = s_2^2/n_2$, where n_1 and n_2 are the sample sizes.

2. Generate a bootstrap sample from the first group, compute the bootstrap sample mean and standard deviation, and label the results $\bar{X}_1^*$ and s_1^*. Do the same for the second group, yielding $\bar{X}_2^*$ and s_2^*. Set $d_1^* = (s_1^*)^2/n_1$ and $d_2^* = (s_2^*)^2/n_2$.

3. Compute
$$W^* = \frac{(\bar{X}_1^* - \bar{X}_2^*) - (\bar{X}_1 - \bar{X}_2)}{\sqrt{d_1^* + d_2^*}}.$$

4. Repeat steps 2 and 3 B times yielding $W_1^*, \ldots, W_B^*$. For a Type I error of 0.05, which corresponds to computing a .95 confidence interval, $B = 999$ is recommended. (Smaller values for B can be used in situations to be covered.)

5. Put the $W_1^*, \ldots, W_B^*$ values in ascending order yielding $W_{(1)}^* \leq \cdots \leq W_{(B)}^*$.

6. Set $L = 0.025B$, $U = 0.975B$ and round both L and U to the nearest integer.

The bootstrap t confidence interval for $\mu_1 - \mu_2$ is

$$\left(\bar{X}_1 - \bar{X}_2 + W_{(L)}^*\sqrt{d_1 + d_2},\ \bar{X}_1 - \bar{X}_2 + W_{(U)}^*\sqrt{d_1 + d_2}\right).$$

6.3.1 Hypothesis Testing

Reject H_0: $\mu_1 = \mu_2$, the hypothesis that two groups have equal means, if the confidence interval just computed does not contain 0. If, for example, the confidence interval is (1.2, 2.3), the estimate is that the difference between the means ($\mu_1 - \mu_2$) is at least 1.2, so in particular the situation $\mu_1 - \mu_2 = 0$ seems unlikely in light of the data.

It is stressed that if groups do not differ, and in particular they have identical probability curves, bootstrap methods offer little or no advantage over nonbootstrap methods in terms of Type I errors. However, this does not salvage nonbootstrap methods because, of course, you do not know whether the groups differ. If the groups do differ, the bootstrap tends to provide more accurate confidence intervals. In some situations the improvement is substantial. And as just indicated, it seems that standard methods offer a minor advantage in some cases but never a major one. Consequently, the bootstrap t is recommended for comparing means.

6.4 COMPARING MEDIANS

Chapter 5 noted that when tied values occur, known methods for estimating the standard error of the sample median can be highly unsatisfactory even with large sample sizes. And Chapter 3 noted that the sampling distribution of the median can be poorly approximated by a normal curve. One practical consequence is that methods for comparing the medians of two groups, based on estimates of the standard errors, can be highly unsatisfactory as well. There is, however, a method for comparing medians that has been found to perform well: the percentile bootstrap method, which does not use or require an estimate of the standard errors.

Briefly, generate a bootstrap sample from each group and compute the sample medians, which we label M_1^* and M_2^*. Let $D^* = M_1^* - M_2^*$ be the difference between these bootstrap estimates. If the goal is to compute a 0.95 confidence interval for the difference between the population medians, then repeat this process many times and use the middle 95% of the value D^* after they are put in ascending order. To compute a 0.9 confidence interval, use the middle 90%.

6.5 REGRESSION

Equation (4.5) of Chapter 4 describes Laplace's method for computing a confidence interval for the slope of a regression line based on the least-squares estimator. Today a slight variation of this method is routinely used and is described in most introductory texts. The method is again based on the least-squares estimate of the slope, but the value 1.96 in Equation (4.5) is replaced by a larger value, the magnitude of which depends on the sample size and is read from tables of Student's t distribution (with $n - 2$ degrees of freedom). Often this method is used to test H_0: $\beta_1 = 0$, the hypothesis that the slope of the regression line is zero, and this hypothesis is rejected if the confidence interval for the slope does not contain zero. Unfortunately, this relatively simple method can be disastrous in terms of Type I errors and probability coverage, even under normality. If, for example, there is heteroscedasticity (meaning that the variance of the outcome measure, Y, changes with the value of the predictor, X, as described in Chapter 4), the actual probability of a Type I error can exceed .5 when testing at the 0.05 level. Some authorities would counter that in applied work, it is impossible to simultaneously have heteroscedasticity and a slope of zero. That is, Type I errors are never made when there is heteroscedasticity because the null hypothesis of a zero slope is virtually impossible. However, highly inaccurate confidence intervals for the slope can result. So even if we accept the argument about Type I errors under heteroscedasticity, another concern is that heteroscedasticity can mask an association of practical importance. Serious problems arise even under normality. The reason is that Laplace's method and its modern extension assume homoscedasticity which leads to an expression for the variance of the least-squares estimate of the slope. The concern is that under heteroscedasticity, this expression is no longer valid, and this leads to practical problems that were impossible to address in an effective manner until fairly recently. For example, a confidence interval for the slope might contain zero even when the population value of the slope is not equal to zero, resulting in a Type II error.

The bootstrap can be extended to the problem of computing a confidence interval for the slope of a regression line in a manner that takes heteroscedasticity into account. There are, in fact, several strategies that seem to perform relatively well, two of which are outlined here.

6.5.1 A Modified Percentile Bootstrap Method

One of the simpler methods begins by randomly sampling, with replacement, n pairs of observations from the data at hand. To illustrate the process, again consider Boscovich's data on meridian arcs, which were described in Chapter 2. For convenience, we list the five observed points here: (0.0000, 56,751), (0.2987, 57,037), (0.4648, 56,979), (0.5762, 57,074) and (0.8386, 57,422). A bootstrap sample consists of randomly selecting, with replacement, five pairs of observations from the five pairs available to us. Using a computer, the first pair we select might be (.4648, 56,979). When we draw the second pair of values, with probability 1/5 we will again get the pair (0.4648, 56,979). More generally, when we have n pairs of observations, a bootstrap sample consists of randomly selecting a pair of points, meaning each point has probability $1/n$ of being chosen, and repeating this process n times.

For completeness, there is another approach to generating bootstrap samples based on residuals. Theoretical results tell us, however, that it should not be used when there is heteroscedasticity, and studies that assess how the method performs with small sample sizes also indicate that the method can be highly unsatisfactory. We could test the assumption of homoscedasticity, but it is unknown how to determine whether such tests have enough power to detect situations where this assumption should be discarded. Consequently, details about this other bootstrap method are not given here.

Once we have a bootstrap sample of n pairs of points, we can compute a bootstrap estimate of the slope. For example, if the bootstrap sample for Boscovich's data happens to be (0.4648, 56,979), (0.0000, 56,751), (0.8386, 57,422), (0.4648, 56,979), and (0.0000, 56,751), the least-squares estimate of the slope based on these five pairs of observations is 737.4. That is, 737.4 is a bootstrap estimate of the slope. If we obtain a new bootstrap sample, typically it will differ from the first bootstrap sample and yield a new bootstrap estimate of the slope.

Next, we proceed as was done with the percentile bootstrap method for the mean. That is, we generate many bootstrap estimates of the slope and take the middle 95% to be a 0.95 confidence interval for the true slope. This method improves upon the conventional approach based on Student's T, but unfortunately it requires about 250 pairs of observations to get reasonably accurate results over a wide range of situations. There is, however, a simple modification of the method that has been found to perform well when sample sizes are small. It is based on the observation that for a given sample size, the actual probability coverage obtained with the percentile bootstrap method is fairly stable. If, for example, the actual probability coverage is 0.9 under normality, it will be approximately 0.9 when sampling from a nonnormal curve instead. This suggests that if we expand our confidence interval so that under normality the actual probability coverage will be 0.95, then it will be about 0.95 under nonnormality, and this has been found to be true for a wide range of situations. In terms of testing hypotheses, the actual probability of a Type I error will be reasonably close to 0.05.

The method is implemented as follows. First generate 599 bootstrap estimates of the slope and label them $\hat{\beta}_1^*, \hat{\beta}_2^* \ldots \hat{\beta}_{599}^*$. Next, put these values in order and label them $\hat{\beta}_{(1)}^* \leq \hat{\beta}_{(2)}^* \leq \cdots \leq \hat{\beta}_{(599)}^*$. The 0.95 confidence interval for slope, based on the least-squares estimator, is $(\hat{\beta}_{(a)}^*, \ \hat{\beta}_{(c)}^*)$ where for $n < 40$, $a = 7$ and $c = 593$; for $40 \leq n < 80$, $a = 8$ and $c = 592$; for $80 \leq n < 180$, $a = 11$ and $c = 588$; for $180 \leq n < 250$, $a = 14$ and $c = 585$; while for $n \geq 250$, $a = 15$ and $c = 584$. If, for example, $n = 20$, the lower end of the 0.95 confidence interval is given by $\hat{\beta}_{(7)}^*$, the seventh of the 599 bootstrap estimates after they are put in ascending order. This method becomes the standard percentile bootstrap procedure when $n \geq 250$. It is stressed that although this method performs fairly well in terms of Type I errors, any method based on the least squares estimator might be unsatisfactory for reasons outlined in Chapter 7.

The success of the method just described, in terms of Type I errors, is somewhat surprising. As noted in Chapter 2, the least squares estimate of the slope is just a weighted mean of the outcome (Y) values. This suggests that the modified percentile bootstrap method for the slope might also work well when trying to test hypotheses about the population mean using $\bar{X}$. But it has been found that this is not the case. Using the percentile bootstrap to compute a confidence interval for μ is very unstable, so any simple modification along the lines considered here is doomed to failure.

To illustrate the practical difference between the conventional method for computing a confidence interval for the slope, and the percentile bootstrap, again consider Boscovich's data. The conventional method yields a 0.95 confidence interval of $(226.58, 1,220.30)$. In contrast, the modified percentile bootstrap method gives $(-349.19, 1,237.93)$. The upper ends of the two confidence intervals are similar, but the lower ends differ substantially, so we see that the choice of method can make a practical difference.

6.5.2 The Wild Bootstrap

An alternative method for computing a confidence for the slope, which allows heteroscedasticity and has received considerable attention in recent years, is based on what is called a wild bootstrap method.

For convenience, let

$$T_{hc4} = \frac{b_1}{s_{hc4}},$$

where b_1 is the least-squares estimate of the slope, and s_{hc4} is the HC4 estimate of the standard error of b_1. (Easy-to-use software is described in the final chapter of this book.) Roughly, the strategy is to determine the distribution of T_{hc4} when the null hypothesis of a zero slope is true. This is done by generating bootstrap pairs of observations that mimic the situation where the null hypothesis is true. Once this is done, a confidence interval is computed in a manner similar to how we computed a bootstrap t confidence interval for the mean. But rather than resample with replacement pairs of points, as done by the percentile bootstrap method just described, bootstrap values for Y are generated by multiplying the residuals by values randomly generated by a computer, which yields bootstrap Y values.

To elaborate a bit, recall that the residuals for the least squares regression line are given by

$$r_i = Y_i - b_0 - b_1 X_i,$$

$i = 1, \ldots, n$. The wild bootstrap multiplies r_i by a value randomly generated from a distribution that must have certain properties, the details of which are not important here. (See, for example, Godfrey, 2006, for details.) Let's call this value e_i^*. (A common approach is to take $e_i^* = 1$ with probability 0.5; otherwise, $e_i^* = -1$.) Then a bootstrap value for Y_i is computed, which we label

$$Y_i^* = e_i r_i.$$

Now we have n bootstrap pairs of observations: $(Y_1^*, X_1), \ldots, (Y_n^*, X_n)$. Notice that in the bootstrap world, the Y values are generated in a manner for which both the intercept and slope are 0. Also note that bootstrap Y values are generated, but bootstrap X values are not. Next, compute T_{hc4} based on this bootstrap sample, which yields a bootstrap test statistic for testing H_0: $\beta_1 = 0$, the hypothesis that the slope is zero. Call this test statistic T^*. Repeat this many times. For illustrative purposes, imagine that we repeat this 1,000 times, yielding $T_1^*, \ldots, T_{1000}^*$, and let T be the test statistic based on the original data. Then a confidence for both the slope and intercept can be computed, which is similar in form to the bootstrap-t confidence interval for the mean.

6.6 CORRELATION AND TESTS OF INDEPENDENCE

When studying the association between two variables, it is common for researchers to focus on what is called Pearson's correlation coefficient rather than the least-squares estimate of the slope. The two methods are intimately connected, but the information conveyed by the correlation coefficient differs from the least squares estimate of the slope (except in a certain special case that is discussed in Chapter 10). The immediate goal is to introduce this measure of association and note that again heteroscedasticity plays a role when applying a conventional method aimed at establishing whether two variables are dependent.

Given n pairs of observations, $(X_1, Y_1), \ldots, (X_n, Y_n)$, the sample covariance between X and Y is

$$\mathrm{COV}(X, Y) = \frac{1}{n-1}[(X_1 - \bar{X})(Y_1 - \bar{Y}) + \cdots + (X_n - \bar{X})(Y_n - \bar{Y})].$$

For Boscovich's data, $\bar{X} = 0.436$, $\bar{Y} = 57{,}052.6$, there are $n = 5$ pairs of points, so the sample covariance is

$$\frac{1}{5-1}[(0.000 - 0.436)(56{,}751 - 57{,}052.6) + \cdots$$
$$+(0.8386 - 0.436)(57{,}422 - 57{,}052.6)] = 70.8.$$

Covariance is a generalization of the sample variance in the sense that the covariance of the variable X with itself, COV(X,X), is just s_x^2, the sample variance of the X values.

The estimate of Pearson's correlation coefficient is

$$r = \frac{\mathrm{COV}(X, Y)}{s_x s_y}, \tag{6.2}$$

where s_x and s_y are the sample standard deviations corresponding to X and Y, respectively. For Boscovich's data, $r = 0.94$. The population analog of r (the value of r if all subjects or objects could be measured) is typically labeled ρ. It can be shown that the value of ρ always lies between -1 and 1, and that when X and Y are independent, $\rho = 0$. So if one can reject the hypothesis that $\rho = 0$, dependence between X and Y is implied. In addition, it can be shown that if $\rho > 0$ the least squares regression line will have a positive slope (meaning that according to the least squares line, Y tends to increase as X increases), and if $\rho < 0$, the reverse is true.

The conventional test of H_0: $\rho = 0$ is derived under the assumption that X and Y are independent. An implication of this assumption is when predicting Y from X, there is homoscedasticity. If X or Y has a normal distribution, homoscedasticity makes it possible to derive a commonly used test statistic:

$$T = r\sqrt{\frac{n-2}{1-r^2}}. \tag{6.3}$$

Under normality, and when $\rho = 0$, T has a Student's t distribution (with $n - 2$ degrees of freedom). If, for example, 26 pairs of observations are used to compute r, then with probability .95, T will have a value between -2.064 and 2.064. So if we reject when $|T| > 2.064$, the probability of a Type I error will be 0.05, still assuming normality. (Computational details can be found in virtually any introductory text.)

There are many pitfalls associated with r and ρ, most of which are described in Part II of this book. For the moment we focus on an issue related to testing the hypothesis that $\rho = 0$. If, indeed, the population correlation coefficient is zero, does this imply independence? The answer is no, not necessarily. For example, suppose X and Y are independent, standard normal random variables when $X \leq 1$, but that for $X > 1$, the standard deviation of Y is X. So given that $X = 0.5$, say, the standard deviation of Y is 1, but if $X = 2$, the standard deviation of Y is 2. Then $\rho = 0$, yet X and Y are dependent because knowing the value of X can alter the probabilities associated with Y.

Figure 6.5 shows a plot of 400 values generated on a computer in the manner just described. That is, four hundred pairs of values for both X and

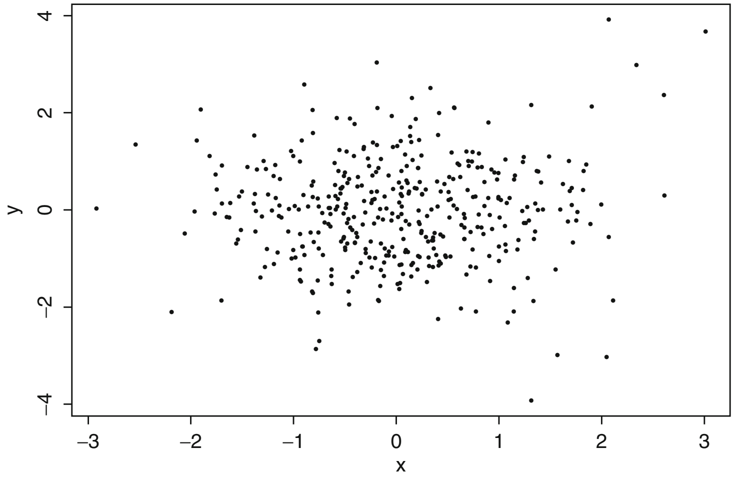

Figure 6.5: This cloud of points was generated from a situation where X and Y are dependent, yet $\rho = 0$. The cloud of points might suggest that there is independence, but for $X > 0$, the variance of Y increases with X.

Y were generated from a standard normal probability curve, and if the value for X is greater than one, Y was multiplied by X. Notice that there is little or no indication that there is an association between X and Y. Not only is the population correlation zero, but the population slope of the least squares regression line (β_1) is zero as well. Yet, when using T to test the hypothesis of a zero correlation at the 0.05 level, the actual probability of rejecting is 0.15. So in this particular case, when we reject, the correct conclusion is that X and Y are dependent, but it would be incorrect to conclude that $\rho \neq 0$, and it would be incorrect to conclude that the mean of Y increases or decreases with X. The problem is that the test of H_0: $\rho = 0$ is based on the assumption that there is homoscedasticity, but in reality there is heteroscedasticity, and this causes the probability of rejecting to exceed .05 even though the hypothesis being tested is true.

To take a more extreme example, suppose e and X are independent, standard normal variables, and that $Y = |X|e$. (So to generate a value for Y, we generate a value from the standard normal probability curve, call it e, generate a value for X, independent of e, and then set $Y = |X|e$.) Then again, X and Y have zero correlation even though X and Y are dependent. (They are dependent because the variance of Y changes with X.) Now when testing H_0: $\rho = 0$ at the 0.05 level, the actual probability of a Type I error, based on a sample of $n = 20$ points, is 0.24. Increasing the sample size to 400, the probability of rejecting is 0.39. That is, the probability of rejecting *increased* even though the hypothesis of a zero correlation is true. That is, we are more likely to reject with a larger sample size even though the hypothesis about Pearson's correlation is true and should not be rejected. The probability of incorrectly rejecting increases due to heteroscedasticity. So when we reject, it is correct to conclude that X and Y are dependent, but we must be careful about any inferences we draw about how are X and Y related. If we reject and $r > 0$, for example, it is certainly true that the estimated least-squares regression will have a positive slope, but this does not necessarily mean that in reality, it is generally the case that the expected (or average) value of Y increases with X. In our example, it does not, and we will see a variety of other ways the value of r might mislead us.

Put another way, when we reject the hypothesis that $\rho = 0$, this might be because $\rho \neq 0$, but an additional factor causing us to reject might be heteroscedasticity. Today there are at least three ways that appear to be reasonably effective at separating the influence of these two factors. One approach is the (modified) percentile bootstrap method used to compute a confidence interval for the slope of the least-squares regression line. The method is exactly the same as before, only for each bootstrap sample we simply compute the correlation coefficient rather than the least squares estimate of the slope. In particular, generate a bootstrap sample by sampling with replacement n pairs of observations. Then compute Pearson's correlation coefficient and label it r^*. Repeat this 599 times and label the resulting correlation coefficients $r_1^*, \ldots r_{599}^*$. Next, put these values in order and label them $r^*_{(1)} \leq \cdots \leq r^*_{(599)}$. The .95 confidence interval for ρ is $(r^*_{(a)},\ r^*_{(c)})$ where for $n < 40$, $a = 7$ and $c = 593$; for $40 \leq n < 80$, $a = 8$ and $c = 592$; for $80 \leq n < 180$, $a = 11$ and $c = 588$; for $180 \leq n < 250$, $a = 14$ and $c = 585$; while for $n \geq 250$, $a = 15$ and $c = 584$.

In our illustrations where T, given by Equation (6.3), does not control the probability of a Type I error, the probability of rejecting H_0: $\rho = 0$ is close to 0.05, as intended, even though there is heteroscedasticity. That is, the bootstrap separates inferences about the population correlation coefficient from a factor that contributes to our probability of rejecting. In so far as we want a test of H_0: $\rho = 0$ to be sensitive to ρ only, the bootstrap is preferable.

Some might argue that it is impossible to have heteroscedasticity with ρ exactly equal to zero. That is, we are worrying about a theoretical situation that will never arise in practice. Regardless of whether one agrees with this view, the sensitivity of the T test of H_0: $\rho = 0$ is relatively minor compared to other problems to be described. What might be more important is whether heteroscedasticity masks an association. That is, from an applied point of view, perhaps there are situations where we fail to reject with T, not because there is no association, but because there is heteroscedasticity. But even if this concern has no relevance, several other concerns have been found to be extremely important in applied work.

We mention one of these concerns here, but we must postpone a discussion of the others until more basic principles are described. (These additional concerns are described in Chapters 7 and 10.) The concern illustrated here is that the sample breakdown point of the correlation coefficient, r, is only $1/n$. That is, one point, properly placed, can cause the correlation coefficient to take on virtually any value between -1 and 1. So care must be taken when interpreting that value of r.

Figure 6.6 shows a scatterplot of the logarithm of the surface temperature of 47 stars versus the logarithm of its light intensity. The scatterplot suggests that, generally, there is a positive association between temperature and light intensity, yet $r = -0.21$. The value of r is negative because of the four outliers in the upper left corner of Figure 6.6. That is, from a strictly numerical point of view, r suggests there is a negative association, but for the bulk of the points the scatterplot indicates that the reverse is true. One could argue that these outliers are interesting because they hint at the possibility that the association changes for relatively low X values. It cannot be emphasized too strongly that of course outliers can be interesting. Theoreticians who work on robust statistical methods assume that this is obvious. In this particular case, the outliers happen to be red giant stars, and perhaps this needs to be taken into account when studying the association between light intensity and surface temperature. Note that for the six X values less than 4.2 (shown in the left portion of Figure 6.6), the scatterplot suggests that there might be a negative association, otherwise, the association appears to be positive. So perhaps there is a nonlinear association between temperature and light intensity. That is, fitting a straight line to these data might be ill advised when considering the entire range of X values available to us.

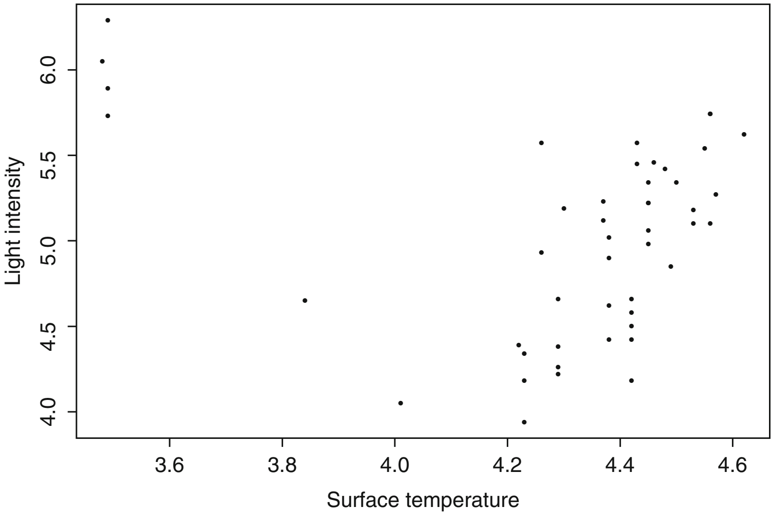

Figure 6.6: A scatterplot of the star data illustrating that outliers can greatly influence r. Here, r is negative even though for the majority of the points there is a positive association.

We have just seen that simply looking at a scatterplot can give us an indication that the value of r is misleading. But in Chapter 7 we will see examples where even a scatterplot is potentially deceptive. (See, in particular, the discussion of Figures 7.10 and 7.11.) Understanding the association between two or more variables requires a library of tools. Subsequent chapters will elaborate on this important point and illustrate some of the tools one might use.

Finally, recent results suggest that there are two alternative methods for making inferences about Pearson's correlation that appear to have an advantage over the percentile bootstrap method just described. Both are based on the HC4 estimate of the standard error. No details are given here, but appropriate software is described in Chapter 12.

6.7 A SUMMARY OF KEY POINTS

- Both theory and simulations indicate that the bootstrap t (also called the percentile t bootstrap) beats our reliance on the central limit theorem when computing confidence intervals for the population mean. Practical problems with Student's T are reduced but not eliminated. It was illustrated that Student's T and the bootstrap t can yield substantially different results.

- The percentile bootstrap method is not recommended when working with the sample mean, but it has practical value when making inferences about the slope of a regression line. For example, when using the least-squares estimate of the slope, the modified percentile bootstrap method provides relatively accurate probability coverage even under heteroscedasticity. Even under normality, heteroscedasticity can invalidate the conventional method based on Student's T. When dealing with least-squares, an effective alternative to the percentile bootstrap is the wild bootstrap used in conjunction with the HC4 estimate of the standard error. The wild bootstrap seems a bit better for general use.

- Heteroscedasticity also affects the conventional T test of the hypothesis that $\rho = 0$. Again this problem is corrected substantially by using the modified percentile bootstrap method or a method based on the HC4 estimator.

- For a wide range of situations, some type of bootstrap method can have practical value. But care must be used because not all variations perform well for a given goal. For example, when computing a confidence interval for the mean, use a bootstrap t method, not the percentile bootstrap. But when dealing with the median, the reverse is true, particularly when there are tied values. (Summaries of the relative merits of various bootstrap methods, when dealing with commonly occurring goals, can be found in Wilcox, 2005.)

6.8 BIBLIOGRAPHIC NOTES

There are many variations of the bootstrap beyond those described here. For books completely dedicated to the bootstrap, see Efron and Tibshirani (1993), Davison and Hinkley (1997), and Shao and Tu (1995). In this chapter methods were described for obtaining bootstrap samples in regression when there is heteroscedasticity. For relevant theoretical results, see Wu (1986). The modification of the percentile method for computing a confidence for the slope of a least-squares regression line comes from Wilcox (1996a). Godfrey (2006) reports results supporting the use of the wild bootstrap in conjunction with the HC4 estimator when dealing with least squares. But a more extensive investigation by Ng (2009) indicates that no one method is always best. A nonbootstrap method using the HC4 estimator was found to perform best in some situations, but in other situations, the reverse is true.