Chapter 3

THE NORMAL CURVE AND OUTLIER DETECTION

No doubt the reader is well aware that the normal curve plays an integral role in applied research. Properties of this curve, that are routinely described in every introductory statistics course, make it extremely important and useful. Yet, in recent years, it has become clear that this curve can be a potential source for misleading and even erroneous conclusions in our quest to understand data. This chapter summarizes some basic properties of the normal curve that play an integral role in conventional inferential methods. But this chapter also lays the groundwork for understanding how the normal curve can mislead. A specific example covered here is how the normal curve suggests a frequently employed method for detecting outliers that can be highly misleading in a variety of commonly occurring situations. This chapter also describes the central limit theorem, which is frequently invoked in an attempt to deal with nonnormal probability curves. Often the central limit theorem is taken to imply that with about 25 observations, practical problems due to nonnormality become negligible. There are several reasons why this view is erroneous, one of which is given here. The illustrations in this chapter provide a glimpse of additional problems to be covered.

3.1 THE NORMAL CURVE

The equation for the family of normal curves is

$$\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}, \tag{3.1}$$

R.R. Wilcox, *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*, DOI 10.1007/978-1-4419-5525-8_3,

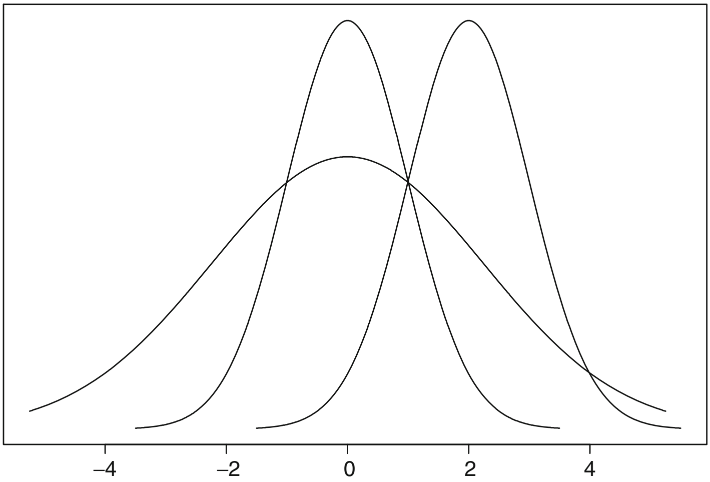

Figure 3.1: The plots provide some sense of how normal probability curves change when the mean and standard deviation are altered. Two of the normal distributions have equal means but unequal standard deviations. Note that increasing the standard deviation from 1 to 1.5 results in a clear and distinct change in a normal curve. Curves with equal standard deviations, but unequal means, are exactly the same, only one is shifted to the right.

where μ is the population mean around which observations are centered, and σ is the population standard deviation, which is a measure of scale introduced in Chapter 2; it determines how tightly the curve is centered around the mean. In Equation (3.1), e is Euler's number (not Euler's constant), which is also called Napier's constant, and is approximately equal to 2.718. (It is the base of natural logarithms.) All normal curves are bell-shaped and symmetric about the population mean. The normal curve with $\mu = 0$ and $\sigma = 1$ is called a *standard normal* distribution.

Figure 3.1 illustrates the effects of varying the mean (μ) and standard deviation (σ). The two normal curves on the left have the same mean (both are centered around zero), but they have different standard deviations. Notice that increasing the standard deviation from 1 to 1.5 results in a clear and noticeable change in the graph of the normal curve. (This property forms the basis of a common misconception discussed in Chapter 7.) The curve with the smaller standard deviation is more tightly centered around the population mean. If two normal curves have the same standard deviation, but unequal means, the shapes of the curves are exactly the same; the only difference is that they are centered around different values.

There is a feature of the normal curve and the standard deviation that plays an integral role in basic inferential methods in statistics. For any normal curve having an arbitrary mean and standard deviation, the probability that an observation is within one standard deviation of the mean is (approximately) 0.68. For example, if a normal probability curve is centered around 100 (the population mean is $\mu = 100$), and its standard deviation is 4 ($\sigma = 4$), then the probability that a randomly sampled observation is between $100 - 4 = 96$ and $100 + 4 = 104$ is 0.68. If instead the standard deviation is 8, then the probability of an observation being between $100 - 8 = 92$ and $100 + 8 = 108$ is again 0.68. This result generalizes to any mean. If the mean is 50 and the standard deviation is 10, then the probability of getting an observation between $50 - 10 = 40$ and $50 + 10 = 60$ is 0.68.

In a similar manner, the probability of being within two standard deviations of the population mean is .954. For example, if again the mean is 100 and the standard deviation is 4, the probability that a randomly sampled observation is between $100 - 2(4) = 92$ and $100 + 2(4) = 108$ is .954. In fact, for any multiple of the standard deviation, the probability remains fixed. For example, the probability of being within 1.96 standard deviations of the mean is .95. Figure 3.2 graphically illustrates this property.

A related result is that probabilities are determined exactly by the mean and standard deviation when observations follow a normal curve. For example, the probability that an observation is less than 20 can be determined (to several decimal places of accuracy) if we are told that the population mean is 22 and the standard deviation is 4. (The computational details are covered in virtually all introductory books on applied statistics.)

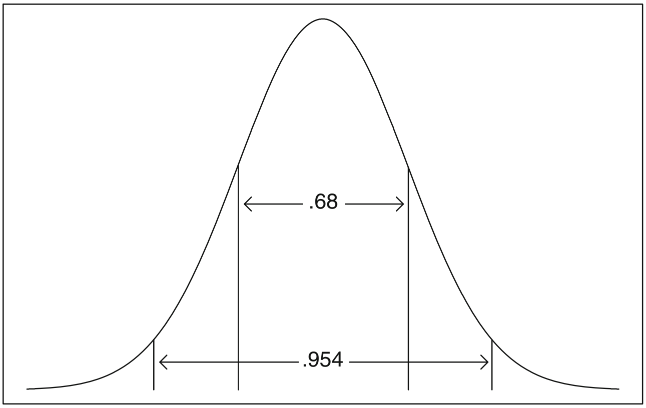

Figure 3.2: For normal probability curves and any positive constant c we might choose, the probability that the distance of an observation from the population mean is less than $c\sigma$ is completely determined by c. That is, regardless of what the values for the population mean and variance might be, the constant c determines this probability. For example, the probability that an observation is within one standard deviation of the mean is exactly 0.68. That is, with probability 0.68, a randomly sampled observation will lie between $\mu - 1\sigma$ and $\mu + 1\sigma$. The probability that an observation is within two standard deviations of the mean is exactly 0.954.

3.2 DETECTING OUTLIERS

The property of the normal curve illustrated in Figure 3.2 suggests a probabilistic approach to detecting outliers that is frequently employed: Declare a value to be an outlier if it is more than two standard deviations from the mean. In symbols, declare the value X to be an outlier if

$$|X - \mu| > 2\sigma, \tag{3.2}$$

the idea being that there is a low probability that the distance of an observation from the mean will be greater than two standard deviations. For a normal probability curve, this probability is 0.046. In practice, the population mean and standard deviation are generally unknown, but they can be estimated as previously indicated, in which case the rule is to declare the value X an outlier if its distance from the sample mean is more than two sample standard deviations. That is, declare X an outlier if

$$|X - \bar{X}| > 2s. \tag{3.3}$$

Unfortunately, this simple method for detecting outliers has a serious problem, which is related to the finite sample breakdown point of $\bar{X}$ and s. To illustrate the problem, consider the values

$$2, 3, 4, 5, 6, 7, 8, 9, 10, 50.$$

The sample mean is 10.4 and the standard deviation is 14.15, so we see that the value 50 is declared an outlier because $|50 - 10.4|$ exceeds 2×14.15. But suppose we add another outlier by changing the value 10 to 50. Then $|\bar{X} - 50| = 1.88s$, so 50 would not be declared an outlier, yet surely it is unusual compared to the other values. If the two largest values in this last example are increased from 50 to 100, then $|\bar{X} - 100| = 1.89s$, and the value 100 still would not be declared an outlier. If the two largest values are increased to 1,000, even 1000 would not be flagged as an outlier! This illustrates the general problem known as *masking*. The problem is that *both* the sample mean and standard deviation are being inflated by the outliers, which in turn masks their presence.

In the illustrations, it might be thought that if we knew the population standard deviation, rather than having to estimate it with s, the problem of masking would no longer be relevant. It turns out that this speculation is incorrect. In fact, even when a probability curve appears to be normal, meaning that it is bell-shaped and symmetric about its mean, practical problems arise. (Details can be found in Chapter 7.)

BETTER METHODS

How can we get a more effective method for detecting outliers that is not subject to masking? A key component is finding measures of location and scale that are not themselves affected by outliers. We know how to get a high breakdown point when estimating location: Replace the mean with the median. But what should be used instead of the sample standard deviation?

One choice that turns out to be relatively effective is the *median absolute deviation* statistic, commonly referred to as MAD. It is computed by first subtracting the median from every observation and then taking absolute values. In symbols, compute

$$|X_1 - M|, |X_2 - M|, \ldots, |X_n - M|.$$

The median of the n values just computed is MAD.

Here is an illustration using the values 2, 4, 6, 7, 9, 12, 16. The median is $M = 7$. Subtracting 7 from each of the seven values, and then taking absolute values, yields

$$|2 - 7| = 5,\ |4 - 7| = 3,\ 1,\ 0,\ 1,\ 4,\ 8.$$

The median of the seven values just computed is MAD. That is, MAD = 4.

Now, for normal probability curves, it turns out that MAD/0.6745 estimates the population standard deviation, σ. Simultaneously, the sample median, M, estimates the population mean, μ. For many purposes, using MAD to estimate the population standard deviation is unsatisfactory. For one, it tends to be a less accurate estimate than the sample standard deviation s when observations do indeed follow a normal curve. However, MAD is much less sensitive to outliers; its finite sample breakdown point is approximately 0.5, the highest possible value, so it is well suited for detecting outliers.

We can modify our rule for detecting outliers in a simple manner: Declare the value X an outlier if

$$|X - M| > 2\frac{\text{MAD}}{0.6745}. \tag{3.4}$$

As an illustration, consider again the study dealing with the desired number of sexual partners by young males, but to make the illustration more salient, we omit the value 6,000. A portion of the data, written in ascending order, looks like this:

$$0,\ 0,\ 0,\ 0,\ 0,\ 1,\ \ldots,\ 30,\ 30,\ 40,\ 45,\ 150,\ 150.$$

The sample mean is 7.79, and the standard deviation is 21.36. If we use our outlier detection rule based on the mean and standard deviation, we see that the value 150 is declared an outlier, but the other values are not. In contrast, using our rule based on the median and MAD, all values greater than or equal to 4 are declared outliers. That is, 41 values are declared outliers versus only the value 150 when using the mean and standard deviation.

3.3 THE BOXPLOT

One other method for detecting outliers is briefly mentioned because it is frequently employed and recommended. It is based on a graphical method for summarizing data called a boxplot, an example of which is shown in Figure 3.3. The construction of a boxplot begins by computing what are called the lower and upper quartiles, which also are called the .25 and .75 *quantiles*, respectively, but the computational details are not covered here. (For continuous random variables, if the probability of an observation less than or equal to c is q, c is called the qth quantile.) There are, in fact, at least a half-dozen methods for computing quartiles. The important point is that the quartiles are defined so that the middle half of the observed values lie between them. In Figure 3.3, the lower and upper quartiles are approximately 7 and 15, respectively, so about half of the values used to construct the boxplot lie between 7 and 15. (In addition, the lower fourth of the values are less than 7, and the upper fourth are greater than 15.) The boxplot uses the difference between the upper and lower quartiles, called the *interquartile range*, as a measure of dispersion, which in turn plays a role in deciding whether an observation is an outlier. If a value exceeds the upper quartile plus 1.5 times the interquartile range, it is declared an outlier. In symbols, if F_U and F_L are the upper and lower quartiles, $F_U - F_L$ is the interquartile range, and a value is declared an outlier if it is greater than $F_U + 1.5(F_U - F_L)$. Similarly, a value is declared an outlier if it is less than the lower quartile minus 1.5 times the interquartile range. That is, a value is an outlier if it is less than

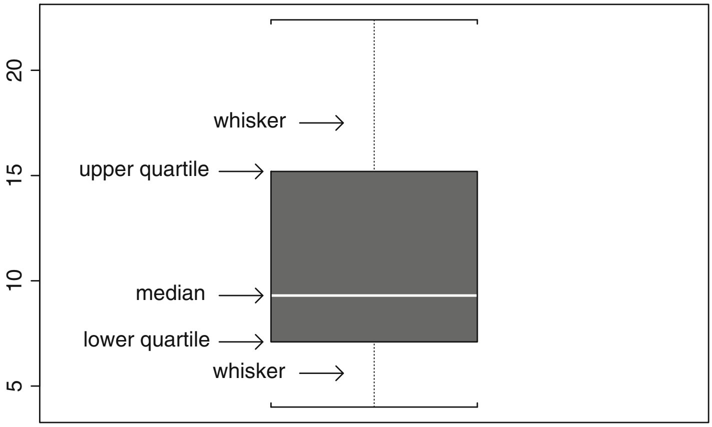

Figure 3.3: An example of a boxplot. The construction of a boxplot begins by computing the lower and upper quartiles, which are defined so that approximately the middle half of the values lie between them. So, about 25% of the values plotted are less than the lower quartile, which in this case is approximately 7. Similarly, about 25% of the values are greater than the upper quartile, which is approximately 15.

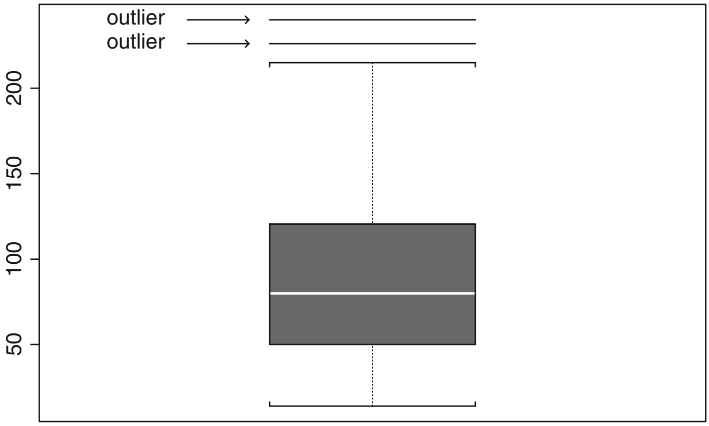

Figure 3.4: Another boxplot, only in contrast to Figure 3.3, two values are flagged as outliers.

$F_L - 1.5(F_U - F_L)$. The lines extending out from the box in Figure 3.3 are called *whiskers*. The ends of the whiskers mark the smallest and largest values not declared outliers. So points lying beyond the end of the whiskers are declared outliers. In Figure 3.3, no outliers are found. Figure 3.4 shows a boxplot where two values are declared outliers.

The boxplot has a finite sample breakdown point of 0.25, meaning that more than 25% of the values must be outliers before the problem of masking arises. For most situations, it seems that a finite sample breakdown point of 0.25 suffices, but exceptions might occur. For the data on the desired number of sexual partners, using the median and MAD led to declaring 41 values as outliers, and this is about 39% of the 105 values. If a boxplot is used, values greater than or equal to 15—about 10% of the values—are declared outliers.

Research on outlier detection methods continues, a few additional issues will be discussed later, but a more detailed discussion of outlier detection goes far beyond the main thrust of this book.

3.4 THE CENTRAL LIMIT THEOREM

There is a rather remarkable connection between the sample mean and the normal curve based on the central limit theorem derived by Laplace. Let's suppose we are interested in feelings of optimism among all adults living in France. Imagine we have some measure of optimism, and as usual let μ and σ^2 represent the population mean and variance. Further imagine that we randomly sample 20 adults and get a sample mean of 22, so our estimate of the population mean is 22. But suppose a different team of researchers randomly samples 20 adults. Of course, they might get a different sample mean from what we got; they might get 26. If yet another team of researchers sampled 20 adults, they might get yet another value for the sample mean. If this process could be repeated billions of times (and hypothetically infinitely many times), each time yielding a sample mean based on 20 observations, and if we plotted the means, what can be said about the plotted means? Laplace found that provided each mean is based on a reasonably large sample size, the plots of the means will follow, approximately, a normal curve. In fact, the larger the number of observations used to compute each sample mean, the better the approximation. Moreover, this normal curve would be centered around the population mean. That is, the sample means would tend to be centered around the value they are trying to estimate. Furthermore, the variance of the normal curve that approximates the plot of the sample means is determined by the population variance (σ^2) and the number of observations used to compute the mean. If observations follow a probability curve having population variance six, and if again the sample means are based on 20 observations, the variation of the sample means is exactly 6/20. More generally, if n observations are randomly sampled from a curve having population variance σ^2, the variance of the normal curve that approximates the plot of sample means will be σ^2/n.

Now the phrase "reasonably large" is rather vague. How many observations must be used when computing the sample means so that there is fairly good agreement between the plot of the means and a normal curve? There is no theorem giving a precise answer—we must rely on experience, although theory suggests where to look for problems.

We begin with a standard illustration of the central limit theorem where observations follow a so-called uniform distribution. That is, the probability curve is as shown in Figure 3.5. This curve says that all values lie somewhere

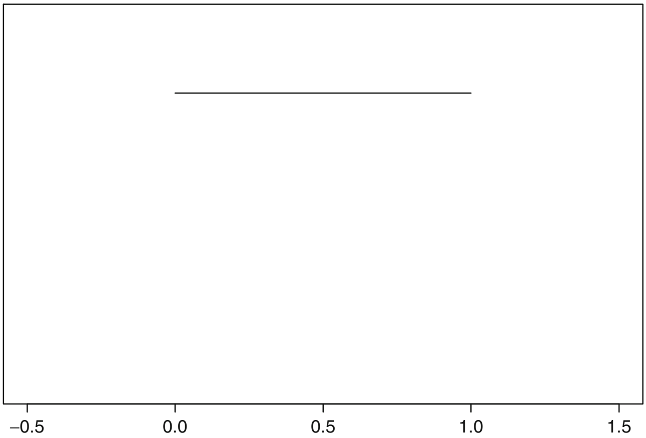

Figure 3.5: A graphical depiction of the so-called uniform distribution. This probability curve is often used to illustrate the central limit theorem. Note that it looks nothing like a normal curve, yet plots of means are approximately normal when a sample size of only 20 is used to compute each mean.

between 0 and 1, and all values are equally likely. As is evident, the curve in Figure 3.5 looks nothing like a normal curve. The population mean for this curve is 0.5, and the variance is 1/12, so the central limit theorem says that if each mean is based on n values, and n is sufficiently large, a plot of the means will be approximately normal and centered around .5 with variance $1/12n$.

Now imagine we randomly sample 20 values and compute the mean. We might get 0.68. If we sample a new set of 20 values, this time we might get 0.42. Of course, we cannot repeat this process infinitely many times, but we can get a fairly accurate sense of what the plot of infinitely many sample means would look like by repeating this process 4,000 times with a computer and plotting the results. Figure 3.6 shows an approximation of the distribution of the sample mean, based 4,000 means, plus the curve we would expect based on the central limit theorem. As can be seen, there is fairly good agreement between the normal curve and the actual distribution of the means, so in this particular case the central limit theorem gives reasonably good results with only 20 observations used to compute each mean.

Let's try another example. We repeat our computer experiment, only this time we sample observations having the probability curve shown in Figure 3.7 (which is called an exponential distribution). Again, this curve looks nothing like a normal curve. Both its mean and variance are 1, so the central limit theorem suggests that plots of means will be centered around 1 with variance $1/n$. Figure 3.8 shows the plot of 4,000 means generated on a computer.

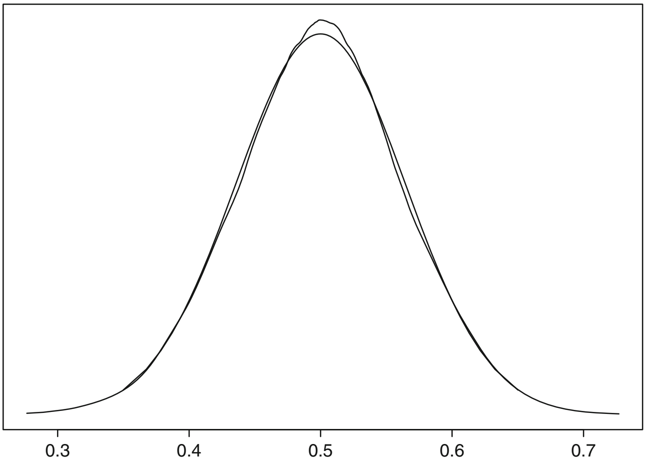

Figure 3.6: The distribution of the sample mean, based on four thousand means, versus the predicted curve based on the central limit theorem when observations are sampled from a uniform distribution. In this case, a normal curve provides a good approximation of the plotted means with only 20 observations used to compute each mean.

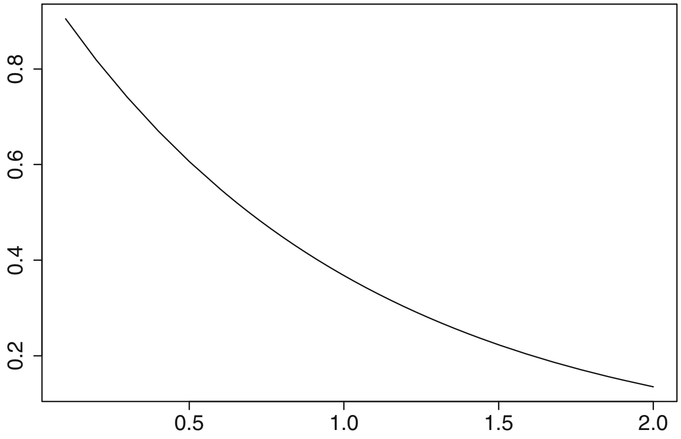

Figure 3.7: A graphical depiction of the so-called exponential distribution. This is another probability curve often used to illustrate the central limit theorem.

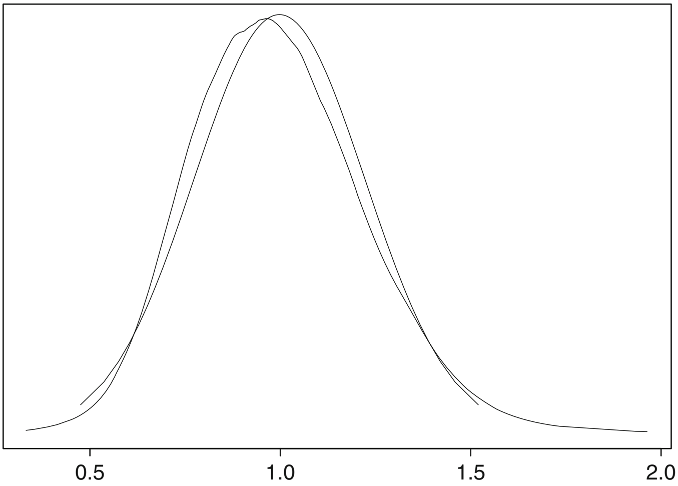

Figure 3.8: The distribution of the sample mean, based on 4,000 means, and the predicted curve based on the central limit theorem when observations are sampled from an exponential distribution. Again, the normal curve suggested by the central limit theorem provides a good approximation with only 20 observations used to compute each mean.

Again, with only 20 values used to compute each mean, the normal curve provides a reasonably good approximation of the plotted sample means.

So we have two examples where we start with a probability curve that looks nothing like a normal curve, yet plots of means are approximately normal when each mean is based on only 20 values. This might suggest that, in general, surely the central limit theorem applies with small sample sizes, but there are at least two problems. A description of one of these problems must be postponed until Chapter 5. To illustrate the other, let's consider yet another example where the probability curve is as shown in Figure 3.9. When 20 observations are used to compute each sample mean, the plot of the means is poorly approximated by a normal curve, particularly in the left tail as indicated in Figure 3.10. If we increase the number of observations

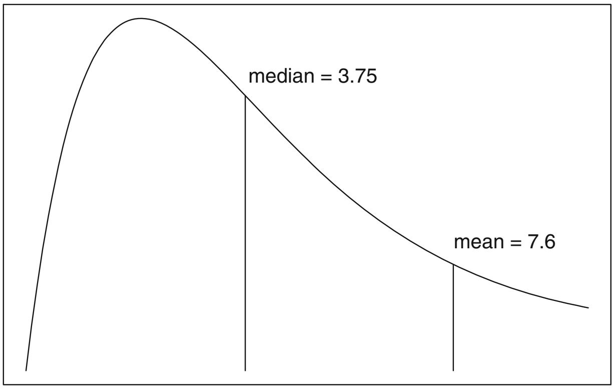

Figure 3.9: An example of an asymmetric probability curve for which outliers are relatively common. Experience indicates that such curves are common in applied work.

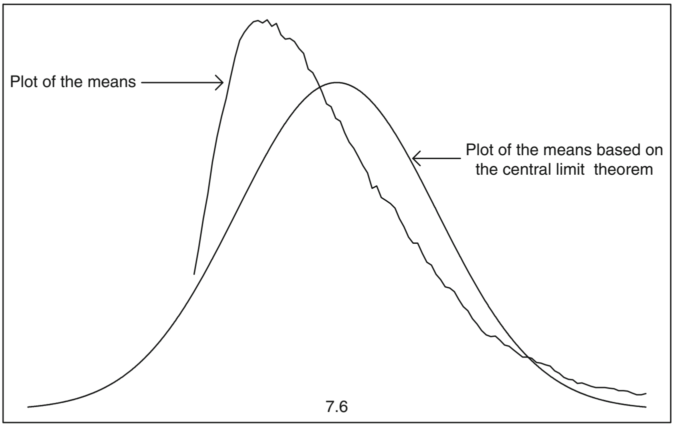

Figure 3.10: A plot of 4,000 means versus the predicted curve based on the central limit theorem when observations are sampled from the distribution shown in Figure 3.9. Now the normal curve for approximating the plot of the means, suggested by the central limit theorem, performs poorly when 20 observations are used to compute each mean.

used to compute each mean, then according to the central limit theorem, the approximation will improve. But if we use 50 observations to compute each sample mean, the approximation remains poor. If instead each mean is based on 100 observations, the plot of means is now reasonably well approximated by a normal curve. So we see that in some cases, with 20 observations we get a good approximation, but there are situations where we need about 100 observations instead.

In Figure 3.9, which is just a reproduction of Figure 2.4, the population mean is in a fairly extreme portion of the right tail, as noted in Chapter 2. Observe that despite this, the sample means are centered around this extreme value. That is, the sample means satisfy their intended goal of estimating the population mean, even though the population mean might be far removed from the bulk of the observations.

What distinguishes the three illustrations of the central limit theorem? The answer is that the first two are based on probability curves characterized by what are called light tails. This roughly means that outliers tend to be rare when sampling observations. In the last example, sampling is from a heavy-tailed probability curve where outliers are common—a situation that occurs frequently in applied work. So a tempting generalization is that we can assume sample means follow a normal curve if sampling is from a light-tailed probability curve, even when each mean is based on only 20 observations, but the approximation might be poor if sampling is from an asymmetric probability curve that is likely to produce outliers. However, there is one more problem that must be taken into consideration, but we must cover other details before it is described. For now, it is merely remarked that even when sampling from a light-tailed probability curve, practical problems arise in situations to be covered. (The details are explained in Chapter 5; see in particular the text regarding Figures 5.5 and 5.6.)

3.4.1 Normality and the Median

Mathematical statisticians have derived many measures of location in addition to the mean and median. It is noted that there is a version of the central limit theorem that applies to most of them, including all the measures of location considered in this book. For example, if we sample 20 values and we compute one of these measures of location, and if we repeat this process billions of times, the plot of these measures of location will be approximately normal provided each is based on a reasonably large number of observations. If, for example, we randomly sample observations from a normal distribution, the resulting medians will be centered around the population mean, which is equal to the population median. For asymmetric probability curves, such as those in Figure 3.7 and 3.9, the sample medians will be centered around the population median, which in general is not equal to the population mean, as explained in Chapter 2.

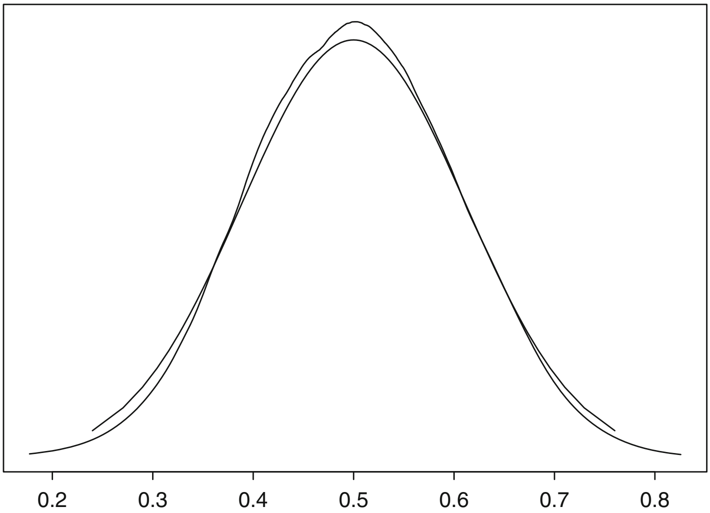

Figure 3.11: There is a version of the central limit theorem that applies to medians and weighted means as well. This figure shows that when sampling observations from the uniform distribution in Figure 3.5, the distribution of the medians is well approximated by a normal curve when each median is based on only 20 observations.

Let's repeat our computer experiment where we sample values having the uniform probability curve shown in Figure 3.5, only now we compute medians rather than means. A plot of the medians appears in Figure 3.11, and again we get a good approximation with a normal curve.

Now look at Figure 3.12, which shows the distributions of the mean (the solid line) and median (the dashed line) when sampling from what is called a lognormal distribution, with each mean and median based on a sample of 20 observations. (A lognormal distribution is similar in shape to the probability curve in Figure 3.9, but has lighter tails, roughly meaning that outliers are less common.) We see that the distribution of the medians resembles a normal distribution much more than the distribution of the means. Put another way, in this particular case, convergence to normality is quicker when using medians. Roughly, practical problems associated with assuming normality, via the central limit theorem, are more likely when the finite sample breakdown point is low and observations are sampled from a skewed distribution where even a few outliers are likely to occur. As previously indicated, the sample mean has the lowest possible finite sample breakdown point (only one outlier can completely dominate its value) compared to the median, which has the highest possible breakdown point, 0.5. So in situations where inferences are based on the central limit theorem (using methods to be covered), larger sample sizes might be needed to avoid practical problems when using means rather than medians.

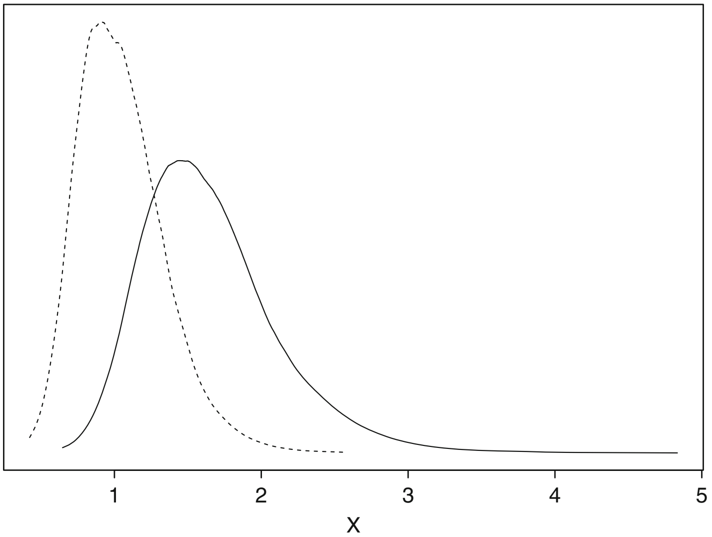

Figure 3.12: When sampling observations from an asymmetric distribution, the distribution of the median (indicated by the dashed line) can better resemble a normal curve than the distribution of the means (indicated by the solid line).

At this point, it might seem that dealing with nonnormal distributions is trivial: Replace the mean with the median. But as will become evident, there are several practical concerns associated with the median that are eliminated when using some of the alternative measures of location introduced in Chapter 8. One of these fundamental concerns arises when dealing with tied (duplicated) values.

Imagine we observe the values 5, 34, 2, 32, 15, and 23. Each value occurs only once, and so we say there are no tied values. But if we observe the values 5, 34, 5, 32, 15, 23, and 16, the value 5 occurs twice. That is, now we have tied values. Unlike other measures of location to be described, tied values can create special problems when using the median. To illustrate one source of concern, we repeat our computer experiment one more time, but now we focus on a situation where the possible observed values are limited to the integers 0, 1, 2, 3, 4, 5, 6; and the probabilities associated with these seven values are as shown in Figure 3.13.

Now we compute the median based on a sample size of 30 drawn from the distribution in Figure 3.13 and we repeat this 4,000 times. Of course tied values will occur simply because there are only 7 possible values each time an observation is made. The upper left panel of Figure 3.14 shows a plot of the medians. As is evident, the plot does not resemble a normal distribution. Indeed, the sample medians have only four distinct values: 4, 4.5, 5, and 6. The upper right panel shows a plot of 4,000 medians, with each median based

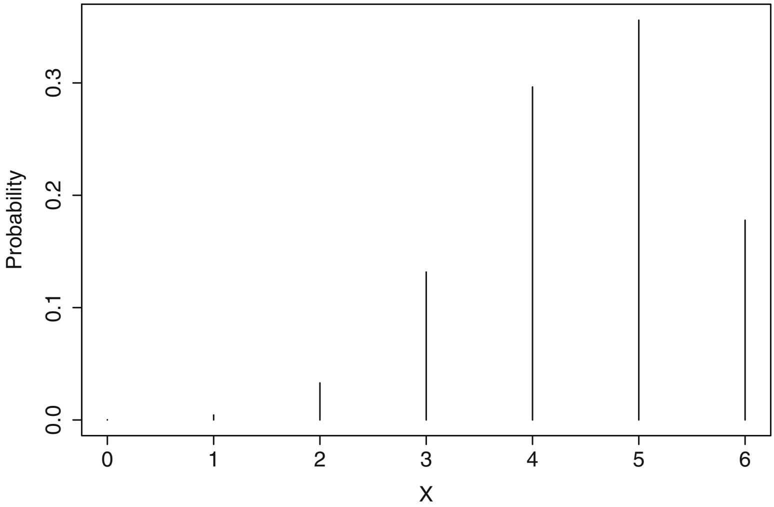

Figure 3.13: A discrete distribution used to illustrate the effects of tied values when using the median.

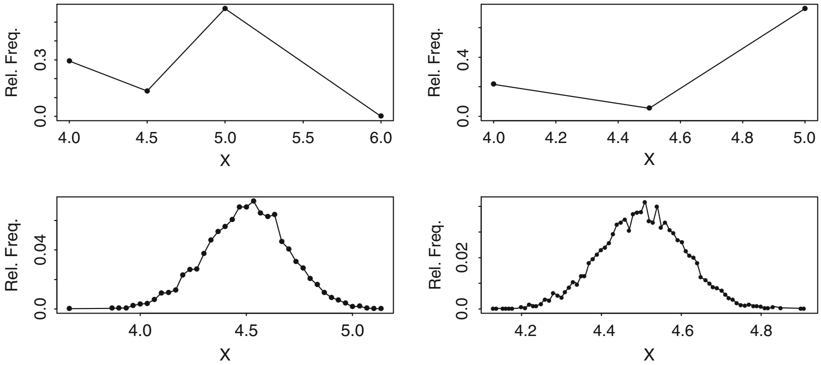

Figure 3.14: The upper panels show a plot of the medians when sampling from the distribution in Figure 3.13. In the upper left panel, each sample median is based on a sample size of 30. In the upper right panel, $n = 100$. As is evident, the plots of the medians do not resemble a normal curve, even when increasing the sample size from 30 to 100. The lower panels are the same as the upper panels, only means are used instead. In contrast to the median, the plots of the means have, approximately, a normal curve.

on a sample size of 100. Despite increasing the sample size, again the plot does not resemble a normal distribution. Indeed, now only three values for the sample median occur.

In contrast, imagine that instead we compute means when sampling from the distribution in Figure 3.13. The lower left panel of Figure 3.14 shows a plot of 4,000 sample means generated in this manner. And the lower right panel shows a plot of the sample means when each mean is based on a sample size of 100 ($n = 100$). As is evident, both plots have the shape of a normal distribution.

For most measures of location discussed in this book, if we increase the sample size and plot the resulting values of the measure of location over many studies, the plot will look more like a normal distribution, in accordance with the central limit theorem. This includes the situation considered in Figure 3.13. But exceptions occur when using the median, as just illustrated.

A practical concern is that, when analyzing data using medians, some methods assume that the medians have a normal distribution, which is unreasonable for the situation at hand. But this is not to suggest that the median should be abandoned or has no practical value. The only point is that when tied values can occur, special techniques might be required, some of which are described in Chapter 6. But even when tied values never occur, there are practical reasons for considering other measures of location that are introduced in Part II.

It is remarked that versions of the central limit theorem also apply when dealing with regression. Imagine repeating an experiment infinitely many times, each time randomly sampling n points and computing the slope. Chapter 2 pointed out that the least-squares estimate of the slope of a regression line can be viewed as a weighted mean of the outcome (Y) values. This suggests that we would get a good approximation of the plotted slopes if in each experiment a reasonably large number of pairs of observations are used, and this speculation turns out to be correct under fairly general conditions. Again, there is the issue of how many points need to be sampled to get good results with the central limit theorem in applied work. For now, suffice it to say that this issue turns out to be nontrivial.

3.5 THREE POINTS WORTH STRESSING

Three points should be stressed before concluding this chapter. First, although problems associated with tied values can be addressed, it is not being suggested that medians be routinely used instead of means when dealing with nonnormality or outliers. Simultaneously, it is not recommended that medians be completely ruled out. They do have value, but there are practical concerns described in Chapter 5. What is needed is some measure of location that performs about as well as the mean when the probability curve is normal, but continues to perform well when outliers are common.

Second, a tempting strategy is to check for outliers and use means if none are found, but this can lead to serious practical problems for reasons described in Chapter 5.

Third, if we were to repeat our computer experiment by sampling observations from a symmetric, heavy-tailed probability curve, it would appear that the central limit theorem performs well with means using only 20 observations to compute each mean. There are, however, serious concerns that are discussed in Chapter 7.

3.6 A SUMMARY OF KEY POINTS

- For all normal probability curves and any constant c, the probability that an observation does not differ from the population mean by more than $c\sigma$ is completely determined by c. For example, if $c = 1$, then with probability 0.68, $|X - \mu| < \sigma$. This property of normal curves suggests a commonly used rule for detecting outliers: Declare the value X an outlier if it is more than two standard deviations away from the sample mean, as described by Equation (3.3). This rule can be highly unsatisfactory, however, due to masking.

- An outlier detection rule that avoids the problem of masking can be obtained by using location and scale that have a high finite sample breakdown point. One example was the MAD-median rule, given by Equation (3.4), and another is the boxplot rule.

- The central limit theorem says that with a sufficiently large sample size, it can be assumed that the sample mean has a normal distribution. In some cases a good approximation of the sample mean is obtained with $n = 20$. But it was illustrated that in other situations, $n = 100$ is required. (In subsequent chapters we will see that even $n = 160$ may not be sufficiently large.)

- A version of the central limit theorem applies to the sample median. It was illustrated that situations arise where the distribution of the median approaches a normal curve more quickly, as the sample size increases, than does the distribution of the mean. However, when tied values can occur, assuming that the median has, approximately, a normal distribution can be highly unsatisfactory. (A method for dealing with tied values, when comparing groups via the median, is described in Chapter 6.)