6-7 Probability Plots

How do we know whether a particular probability distribution is a reasonable model for data? Sometimes this is an important question because many of the statistical techniques presented in subsequent chapters are based on an assumption that the population distribution is of a specific type. Thus, we can think of determining whether data come from a specific probability distribution as **verifying assumptions**. In other cases, the form of the distribution can give insight into the underlying physical mechanism generating the data. For example, in reliability engineering, verifying that time-to-failure data come from an exponential distribution identifies the **failure mechanism** in the sense that the failure rate is constant with respect to time.

Some of the visual displays we used earlier, such as the histogram, can provide insight about the form of the underlying distribution. However, histograms are usually not really reliable indicators of the distribution form unless the sample size is very large. A **probability plot** is a graphical method for determining whether sample data conform to a hypothesized distribution based on a subjective visual examination of the data. The general procedure is very simple and can be performed quickly. It is also more reliable than the histogram for small- to moderate-size samples. Probability plotting typically uses special axes that have been scaled for the hypothesized distribution. Software is widely available for the normal, lognormal, Weibull, and various chi-square and gamma distributions. We focus primarily on normal probability plots because many statistical techniques are appropriate only when the population is (at least approximately) normal.

To construct a probability plot, the observations in the sample are first ranked from smallest to largest. That is, the sample $x_1, x_2, \ldots, x_n$ is arranged as $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, where $x_{(1)}$ is the smallest observation, $x_{(2)}$ is the second-smallest observation, and so forth with $x(n)$ the largest. The ordered observations $x_{(j)}$ are then plotted against their observed cumulative frequency $(j - 0.5)/n$ on the appropriate probability paper. If the hypothesized distribution adequately describes the data, the plotted points will fall approximately along a straight line; if the plotted points deviate significantly from a straight line, the hypothesized model is not appropriate. Usually, the determination of whether or not the data plot is a straight line is subjective. The procedure is illustrated in the following example.

Example 6-7 **Battery Life** Ten observations on the effective service life in minutes of batteries used in a portable personal computer are as follows: 176, 191, 214, 220, 205, 192, 201, 190, 183, 185. We hypothesize that battery life is adequately modeled by a normal distribution. To use probability plotting to investigate this hypothesis, first arrange the observations in ascending order and calculate their cumulative frequencies $(j - 0.5)/10$ as shown in Table 6-6.

TABLE • 6-6 Calculation for Constructing a Normal Probability Plot

j	$x_{(j)}$	$(j - 0.5)/10$	z_j
1	176	0.05	–1.64
2	183	0.15	–1.04
3	185	0.25	–0.67
4	190	0.35	–0.39
5	191	0.45	–0.13
6	192	0.55	0.13
7	201	0.65	0.39
8	205	0.75	0.67
9	214	0.85	1.04
10	220	0.95	1.64

The pairs of values $x_{(j)}$ and $(j-0.5)/10$ are now plotted on normal probability axes. This plot is shown in Fig. 6-22. Most normal probability plots have $100(j-0.5)/n$ on the left vertical scale and (sometimes) $100[1-(j-0.5)/n]$ on the right vertical scale, with the variable value plotted on the horizontal scale. A straight line, chosen subjectively, has been drawn through the plotted points. In drawing the straight line, you should be influenced more by the points near the middle of the plot than by the extreme points. A good rule of thumb is to draw the line approximately between the 25th and 75th percentile points. This is how the line in Fig. 6-22 was determined. In assessing the "closeness" of the points to the straight line, imagine a "fat pencil" lying along the line. If all the points are covered by this imaginary pencil, a normal distribution adequately describes the data. Because the points in Fig. 6-19 would pass the "fat pencil" test, we conclude that the normal distribution is an appropriate model.

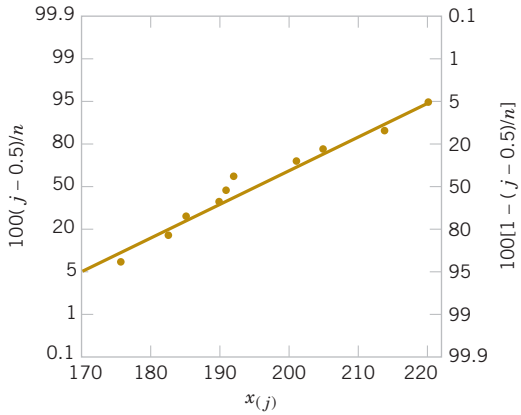

FIGURE 6-22 Normal probability plot for battery life.

A **normal probability plot** can also be constructed on ordinary axes by plotting the standardized normal scores z_j against $x_{(j)}$ where the standardized normal scores satisfy

$$\frac{j-0.5}{n} = P(Z \le z_j) = \Phi(z_j)$$

For example, if $(j-0.5)/n = 0.05, \Phi(z_j) = 0.05$ implies that $z_j = -1.64$. To illustrate, consider the data from Example 6-4. In the last column of Table 6-6 we show the standardized normal scores. Figure 6-23 is the plot of z_j versus $x_{(j)}$. This normal probability plot is equivalent to the one in Fig. 6-22.

We have constructed our probability plots with the probability scale (or the z-scale) on the vertical axis. Some computer packages "flip" the axis and put the probability scale on the horizontal axis.

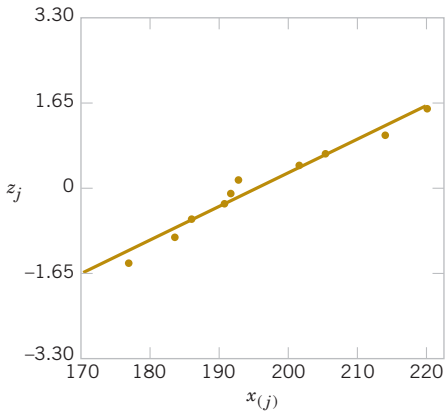

FIGURE 6-23 Normal probability plot obtained from standardized normal scores.

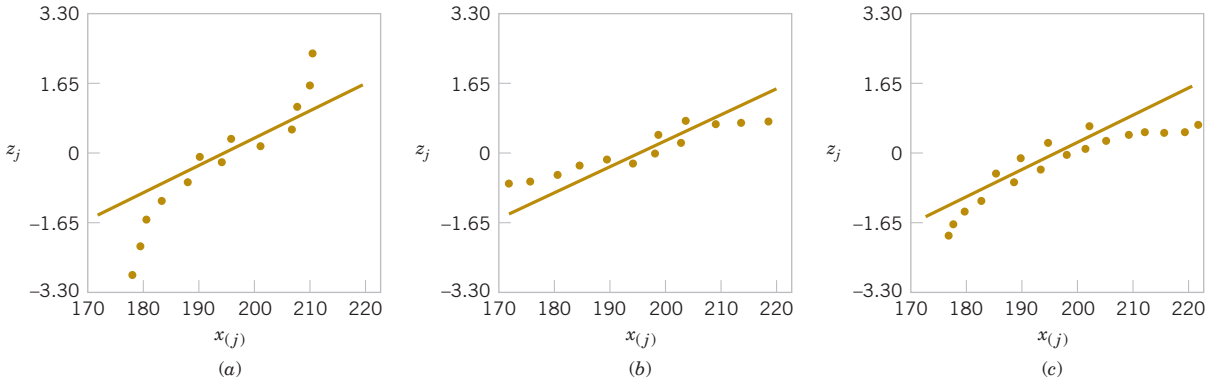

FIGURE 6-24 Normal probability plots indicating a nonnormal distribution. (a) Light-tailed distribution. (b) Heavy-tailed distribution. (c) A distribution with positive (or right) skew.

Normal Probability Plots of Small Samples Can Be Unreliable

The normal probability plot can be useful in identifying distributions that are symmetric but that have tails that are "heavier" or "lighter" than the normal. They can also be useful in identifying skewed distributions. When a sample is selected from a light-tailed distribution (such as the uniform distribution), the smallest and largest observations will not be as extreme as would be expected in a sample from a normal distribution. Thus, if we consider the straight line drawn through the observations at the center of the normal probability plot, observations on the left side will tend to fall below the line, and observations on the right side will tend to fall above the line. This will produce an S-shaped normal probability plot such as shown in Fig. 6-24(a). A heavy-tailed distribution will result in data that also produce an S-shaped normal probability plot, but now the observations on the left will be above the straight line and the observations on the right will lie below the line. See Fig. 6-24(b). A positively skewed distribution will tend to produce a pattern such as shown in Fig. 6-24(c), where points on both ends of the plot tend to fall below the line, giving a curved shape to the plot. This occurs because both the smallest and the largest observations from this type of distribution are larger than expected in a sample from a normal distribution.

Even when the underlying population is exactly normal, the sample data will not plot exactly on a straight line. Some judgment and experience are required to evaluate the plot. Generally, if the sample size is $n < 30$, there can be significant deviation from linearity in normal plots, so in these cases only a very severe departure from linearity should be interpreted as a strong indication of nonnormality. As n increases, the linear pattern will tend to become stronger, and the normal probability plot will be easier to interpret and more reliable as an indicator of the form of the distribution.

Exercises FOR SECTION 6-7

Problem available in *WileyPLUS* at instructor's discretion.
Go Tutorial Tutoring problem available in *WileyPLUS* at instructor's discretion.

6-93. Construct a normal probability plot of the piston ring diameter data in Exercise 6-7. Does it seem reasonable to assume that piston ring diameter is normally distributed?

6-94. Construct a normal probability plot of the insulating fluid breakdown time data in Exercise 6-8. Does it seem reasonable to assume that breakdown time is normally distributed?

6-95. Construct a normal probability plot of the visual accommodation data in Exercise 6-11. Does it seem reasonable to assume that visual accommodation is normally distributed?

6-96. Construct a normal probability plot of the solar intensity data in Exercise 6-12. Does it seem reasonable to assume that solar intensity is normally distributed?

6-97. Construct a normal probability plot of the O-ring joint temperature data in Exercise 6-19. Does it seem reasonable to assume that O-ring joint temperature is normally distributed? Discuss any interesting features that you see on the plot.

6-98. Construct a normal probability plot of the octane rating data in Exercise 6-30. Does it seem reasonable to assume that octane rating is normally distributed?