APPENDIX 4.2: EXPLAINING NORMAL DISTRIBUTIONS

An interesting feature of normal distributions is that each is completely defined by its mean (μ) and its standard deviation (σ). Here is the mathematical definition of a normal distribution:

$$p(x) = \left(\frac{1}{e^{\frac{(x-\mu)^2}{2\sigma^2}}} \right) \left(\frac{1}{\sqrt{2\pi}\sigma} \right). \qquad (4.A2.1)$$

The left side of the equation, $p(x)$, refers to the density of the normal distribution for a given value of x, which is defined on the right side of the equation. [It is unfortunate that statisticians denote density at x with $p(x)$, because density does not mean probability. This is an ambiguity that is usually resolved by context.] Remember, for every value of x, there is only one value of y; in this case, the y value is density, which is denoted by $p(x)$.

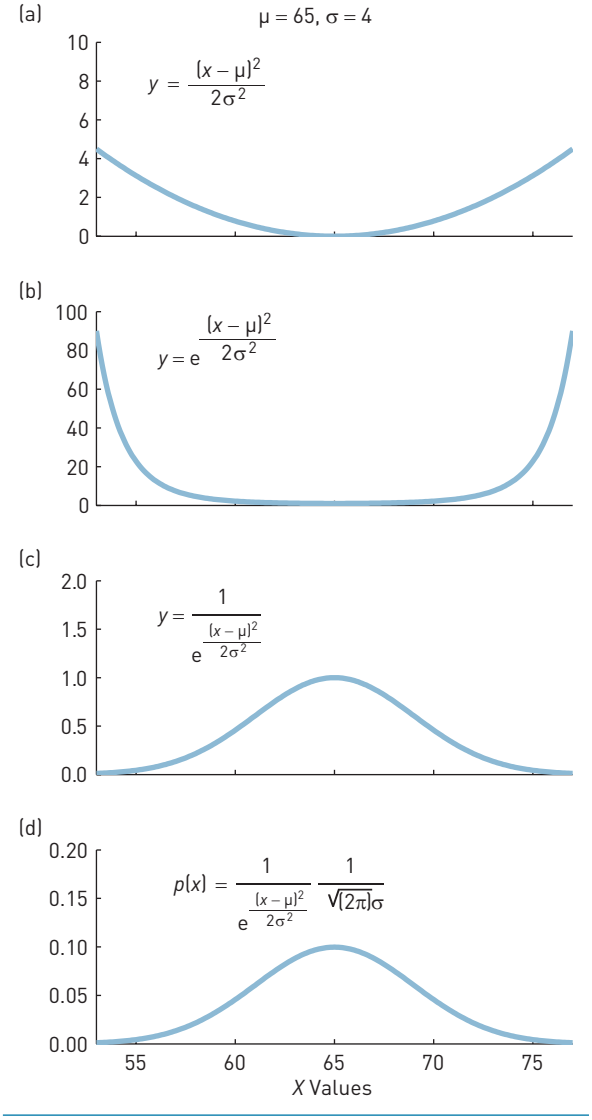
We focus on the first term on the right-hand side of equation 4.A2.1, in which we find the expression $(x - \mu)^2/(2\sigma^2)$. This says that the density of the distribution at x is related to the squared difference between x and μ [i.e., $(x - \mu)^2$] divided by two times the variance (i.e., $2\sigma^2$). So, in this part of the equation, we have x, μ, and σ. The 2 is what we call a *constant*, because it is a number that never changes. A second constant in this part of equation 4.A2.1 is e, which is an irrational number, whose first few digits are 2.71828182. Irrational numbers have an infinite number of non-repeating digits following the decimal place. You might recognize e as the base of the natural logarithm, which is widely used in mathematics and statistics.

In the second term on the right side of equation 4.A2.1, there is a third constant, π. This is the same constant that we use when finding the area of a circle (i.e., area = $2\pi r^2$). It is another irrational number, whose first few digits are 3.14159265.

So, we have three constants in equation 4.A2.1 (2, π, and e). We also have the value of x for which we wish to find the density, and we have the parameters of the distribution, μ and σ. Therefore, the density of a normal distribution at x [i.e., $p(x)$] is completely determined by μ and σ, because all other components of the equation are constants.

To understand how equation 4.A2.1 produces a normal distribution, we will unpack its components and illustrate each one with reference to Figure 4.A2.1. The fundamental component of equation 4.A2.1 is the expression $(x - \mu)^2/(2\sigma^2)$, which says, as noted above, that the density of the distribution at x is related to the squared difference between x and μ [i.e., $(x - \mu)^2$] divided by two times the variance (i.e., $2\sigma^2$). The value of $(x - \mu)^2/(2\sigma^2)$ for each value of x is shown in Figure 4.A2.1a. Note that because the difference between x and μ is squared, $(x - \mu)^2/(2\sigma^2)$ will always be positive.

FIGURE 4.A2.1 ■ Unpacking the Formula



The constant e (in equation 4.A2.1) is raised to the power of $(x - \mu)^2/(2\sigma^2)$ for each value of x; i.e., $e^{(x-\mu)^2/(2\sigma^2)}$. The result is shown in Figure 4.A2.1b, in which the value of y increases more quickly as x increases than it did in Figure 4.A2.1a.

Figure 4.A2.1c shows that when the reciprocal of $e^{(x-\mu)^2/(2\sigma^2)}$ is computed $[1/e^{(x-\mu)^2/(2\sigma^2)}]$, small numbers become large numbers and large numbers become small numbers. The result is a function that has the general bell shape of a normal distribution. However, the function shown in Figure 4.A2.1c is not a probability density function because the area under the curve is not 1. To transform the function shown in Figure 4.A2.1c into a probability density function, we multiply the term $1/e^{(x-\mu)^2/(2\sigma^2)}$ by $1/\sqrt{2\pi}\sigma$. The result (shown in Figure 4.A2.1d) is a probability density function. It is a normal distribution with a mean of $\mu = 65$ and standard deviation of $\sigma = 4$.

It is extremely interesting that the parameters used to generate normal distributions using equation 4.A2.1 are also quantities that can be computed from the distribution of scores in a normal population. Remember that μ and σ are computed as follows:

$$\mu = \frac{\Sigma x}{N},$$

and

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}.$$

This is very nice. The parameters of interest are both (i) properties that can be computed from the distribution of scores and (ii) part of the definition the population's density function. Not all probability density functions are this simple.

APPENDIX 4.3: ASSESSING NORMALITY: QQ PLOTS, SKEW, AND KURTOSIS

Many statistical analyses described in this book depend in one way or another on the assumption that scores have been sampled from a normal distribution. Therefore, we need methods to assess whether this assumption is reasonable. One informal approach is to simply create a histogram and try to judge, visually, whether it is obviously non-normal. Figure 4.A3.1 illustrates a normal population and a skewed population. For purposes of illustration, samples of 20 randomly selected scores were drawn from each of these populations. Histograms of these two samples are shown in Figure 4.A3.2.

FIGURE 4.A3.1 ■ Normal and Skewed Populations

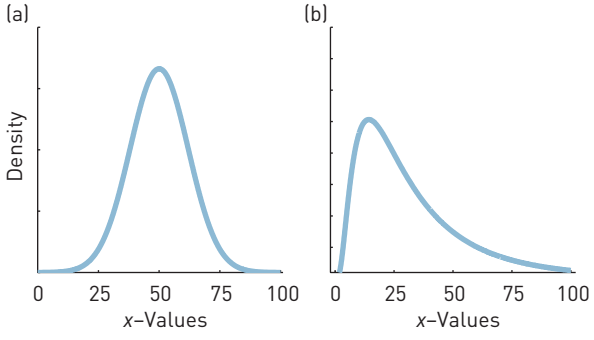
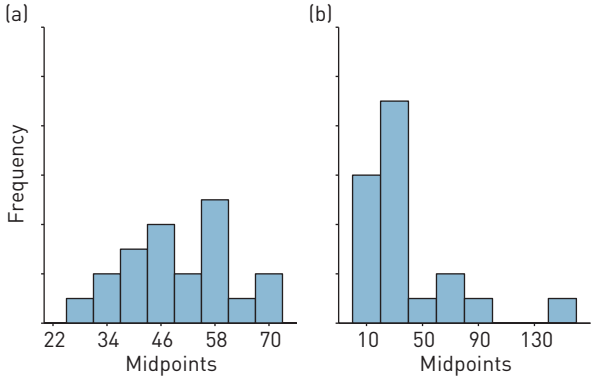


FIGURE 4.A3.2 ■ Normal and Skewed Samples



It is clear from Figure 4.A3.2 that a distribution of sample scores shares characteristics of the population from which the scores were drawn (Figure 4.A3.1). Therefore, based on only the samples, we might suspect that the scores in Figure 4.A3.2b were drawn from a skewed distribution. A problem with this approach, and indeed almost all approaches, is that when samples are small, their distribution can look very non-normal even if the scores came from a normal distribution.

QQ Plots

A common method for judging the normality of a sample is to create a *QQ plot*. The letter Q stands for *quantile*. The quantiles of normal distributions are *z*-scores. Therefore, QQ plots compare two kinds of quantiles, which in this case means two sets of *z*-scores. These two sets of *z*-scores are described in Table 4.A3.1.

The left and right sides of Table 4.A3.1 present a QQ analysis of the normal distribution and a QQ analysis of the skewed distribution, respectively. We will work through the left side first. The first column shows the scores in the sample. At the bottom of this column, we find the mean of the sample ($m = 50$) and its standard deviation ($s = 12$). The scores in the sample are converted to *z*-scores by dividing the difference between each score and the mean by the sample standard deviation [i.e., $z = (x - m)/s$]. These *z*-scores are shown in the column labeled *z*. Notice that the scores in the *x* column are sorted from largest to smallest.

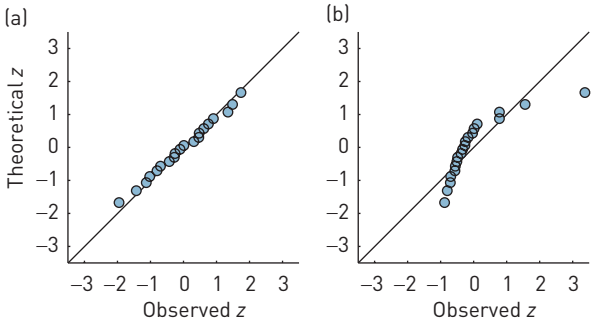
TABLE 4.A3.1 ■ Illustration of QQ Data

Normal Distribution				Skewed Distribution			
x	*z*	*P*(*x*)	*q*	*x*	*z*	*P*(*x*)	*q*
70.87	1.74	0.952	1.67	142.83	3.38	0.952	1.66
67.84	1.49	0.905	1.31	85.78	1.56	0.905	1.31
66.20	1.35	0.857	1.07	61.33	0.78	0.857	1.07
60.85	0.90	0.810	0.88	61.31	0.78	0.810	0.88
58.99	0.75	0.762	0.71	40.22	0.11	0.762	0.71
57.34	0.61	0.714	0.57	37.16	0.01	0.714	0.57
55.78	0.48	0.667	0.43	35.67	−0.04	0.667	0.43
55.56	0.46	0.619	0.30	31.43	−0.17	0.619	0.30
53.77	0.31	0.571	0.18	28.67	−0.26	0.571	0.18
50.10	0.01	0.524	0.06	28.36	−0.27	0.524	0.06
48.78	−0.10	0.476	−0.06	26.40	−0.33	0.476	−0.06
46.88	−0.26	0.429	−0.18	24.92	−0.38	0.429	−0.18
46.61	−0.28	0.381	−0.30	21.78	−0.48	0.381	−0.30
44.86	−0.43	0.333	−0.43	21.00	−0.50	0.333	−0.43
41.58	−0.70	0.286	−0.57	19.39	−0.55	0.286	−0.57
40.43	−0.80	0.238	−0.71	19.03	−0.57	0.238	−0.71
37.67	−1.03	0.191	−0.88	14.95	−0.70	0.191	−0.87
36.43	−1.13	0.143	−1.07	14.69	−0.70	0.143	−1.07
32.89	−1.43	0.095	−1.31	11.76	−0.80	0.095	−1.31
26.59	−1.95	0.048	−1.67	9.38	−0.87	0.048	−1.66
50.00				**36.80**			
12.00				**31.38**			

Beside each *z* is the approximate proportion of scores *in the sample* at or below *x* (or *z*). In the simplest case, when there are no ties, these proportions would range from $1/n$ to $n/n = 1$, in steps of $1/n$. In our example, this would mean that the $P(x)$ values would range from 1/20 to 1, in steps of 1/20. This creates a problem because the *z*-score corresponding to $P(x) = 1$ would be ∞, and that's not very useful. There are many tweaks that can be made to address this issue. The simplest solution is to adjust the $P(x)$ values slightly and have them range from $1/(n+1)$ to $n/(n+1)$. Using this adjustment, the $P(x)$ values now range from 1/21 (.048) to 20/21 (.952) in steps of 1/21 = .0476.

The next step is the interesting one. The numbers in the column labeled *q* represent the *z*-scores from a normal distribution that correspond to the $P(x)$ values in the third column. [The *q*-values were obtained using the Excel **NORM.S.INV** function on the $P(z)$ values in the third column; see Figure 4.A3.4.] When scores in a sample are drawn from a normal distribution, we would expect the *z*-scores and *q*-scores to be very similar. When we plot one against the other, as done in Figure 4.A3.3a, we would expect the pairs of scores to fall on a line going from bottom left to top right, as shown. Although we wouldn't expect the plot of *z* and *q* to make a perfectly straight line, there should be no *systematic deviations* from linearity. Therefore, Figure 4.A3.3a shows what one would expect to find when the sample has been drawn from a normal distribution.

FIGURE 4.A3.3 ■ Two QQ Plots



(a and b) Plots of theoretical quantiles (*z*-scores) on the horizontal axis against observed quantiles (*z*-scores) on the vertical axis. When the sample is drawn from a normal distribution, we expect the scatter of points to vary randomly about the diagonal line, as in (a). When the sample is drawn from a non-normal distribution, we expect the scatter of points to vary systematically about the diagonal line, as in (b).

When a sample is drawn from a non-normal distribution, we would expect the plot of z and q to deviate systematically from a straight line. Figure 4.A3.3b plots the z- and q-scores calculated in the right part of Table 4.A3.1 for the sample drawn from the skewed distribution. This plot clearly deviates systematically from a straight line. This is very strong evidence that the sample was not drawn from a normal distribution.

Most statistical packages, such as SPSS and SAS, provide routines to create QQ plots. Making QQ plots in Excel is also very straightforward (see Figure 4.A3.4).

FIGURE 4.A3.4 ■ Computing QQ Scores in Excel

	A	B	C	D	E	F	G	H
1	**Cum. F**	**x**	**z**	**Formula**	**P(x)**	**Formula**	**q**	**Formula**
2	20	70.87	1.74	=(B2-$B$22)/$B$23	0.952	=A2/21	1.67	=NORM.S.INV(E2)
3	19	67.84	1.49	=(B3-$B$22)/$B$23	0.905	=A3/21	1.31	=NORM.S.INV(E3)
4	18	66.20	1.35	=(B4-$B$22)/$B$23	0.857	=A4/21	1.07	=NORM.S.INV(E4)
5	17	60.85	0.90	=(B5-$B$22)/$B$23	0.810	=A5/21	0.88	=NORM.S.INV(E5)
6	16	58.99	0.75	=(B6-$B$22)/$B$23	0.762	=A6/21	0.71	=NORM.S.INV(E6)
7	15	57.34	0.61	=(B7-$B$22)/$B$23	0.714	=A7/21	0.57	=NORM.S.INV(E7)
8	14	55.78	0.48	=(B8-$B$22)/$B$23	0.667	=A8/21	0.43	=NORM.S.INV(E8)
9	13	55.56	0.46	=(B9-$B$22)/$B$23	0.619	=A9/21	0.30	=NORM.S.INV(E9)
10	12	53.77	0.31	=(B10-$B$22)/$B$23	0.571	=A10/21	0.18	=NORM.S.INV(E10)
11	11	50.10	0.01	=(B11-$B$22)/$B$23	0.524	=A11/21	0.06	=NORM.S.INV(E11)
12	10	48.78	-0.10	=(B12-$B$22)/$B$23	0.476	=A12/21	-0.06	=NORM.S.INV(E12)
13	9	46.88	-0.26	=(B13-$B$22)/$B$23	0.429	=A13/21	-0.18	=NORM.S.INV(E13)
14	8	46.61	-0.28	=(B14-$B$22)/$B$23	0.381	=A14/21	-0.30	=NORM.S.INV(E14)
15	7	44.86	-0.43	=(B15-$B$22)/$B$23	0.333	=A15/21	-0.43	=NORM.S.INV(E15)
16	6	41.58	-0.70	=(B16-$B$22)/$B$23	0.286	=A16/21	-0.57	=NORM.S.INV(E16)
17	5	40.43	-0.80	=(B17-$B$22)/$B$23	0.238	=A17/21	-0.71	=NORM.S.INV(E17)
18	4	37.67	-1.03	=(B18-$B$22)/$B$23	0.190	=A18/21	-0.88	=NORM.S.INV(E18)
19	3	36.43	-1.13	=(B19-$B$22)/$B$23	0.143	=A19/21	-1.07	=NORM.S.INV(E19)
20	2	32.89	-1.43	=(B20-$B$22)/$B$23	0.095	=A20/21	-1.31	=NORM.S.INV(E20)
21	1	26.59	-1.95	=(B21-$B$22)/$B$23	0.048	=A21/21	-1.67	=NORM.S.INV(E21)
22	**m**	**50.00**						
23	**s**	**12.00**						

The original scores are shown under the heading **x**. The cumulative frequencies (i.e., number of scores at or below x) are to the left (**Cum. F**). The z-scores corresponding to the x values are shown under the heading **z**, and the **Formulas** for computing the z-scores are in the next column. The approximate cumulative proportions computed as (**Cum. F**)/(n+1) are shown under **P(x)**, and **Formulas** used to compute these are provided on the right. The theoretical quantiles, which are the z-scores corresponding to the approximate cumulative proportions, are shown under **q**. As shown in the next column, these were computed with **NORM.S.INV** (see Appendix 4.1).

An alternative to the QQ plot is the *PP plot*. Rather than plot theoretical and observed *quantiles* against each other (as in Figure 4.A3.3), a PP plot plots theoretical and observed proportions [$P(x)$] against each other. The observed proportion is the proportion of the sample falling at or below each score in the sample. (This proportion would have to be corrected using a method like the one described earlier.) The theoretical proportion is the proportion of a normal distribution, having the same mean and standard deviation as the sample, falling below each score in the sample. These theoretical and observed proportions can be plotted against each other just as we did for theoretical and observed quantiles in the QQ plot.

QQ Plots in SPSS

To illustrate QQ plots in SPSS, we will make use of the two sets of 20 raw scores in Table 4.A3.1. These two sets of scores were named xNormal and xSkewed in an SPSS

data file. To compute a QQ plot, we do the following: Analyze→Descriptive Statistics→Q-Q Plots . . . and the dialog in Figure 4.A3.5 appears. The two variables have been moved into the Variable(s): panel in the usual way. The drop-down list under Test Distribution allows us to compare our data against many different distributions, but the Normal distribution is the one of interest. Below the Variable(s): panel, there is a check beside the option to Standardize values. This means that *z*-scores will be used as quantiles as in Figure 4.A3.2. Clicking OK begins the analysis, and the output is shown in Figure 4.A3.6.

FIGURE 4.A3.5 ■ The QQ Plots Dialog in SPSS

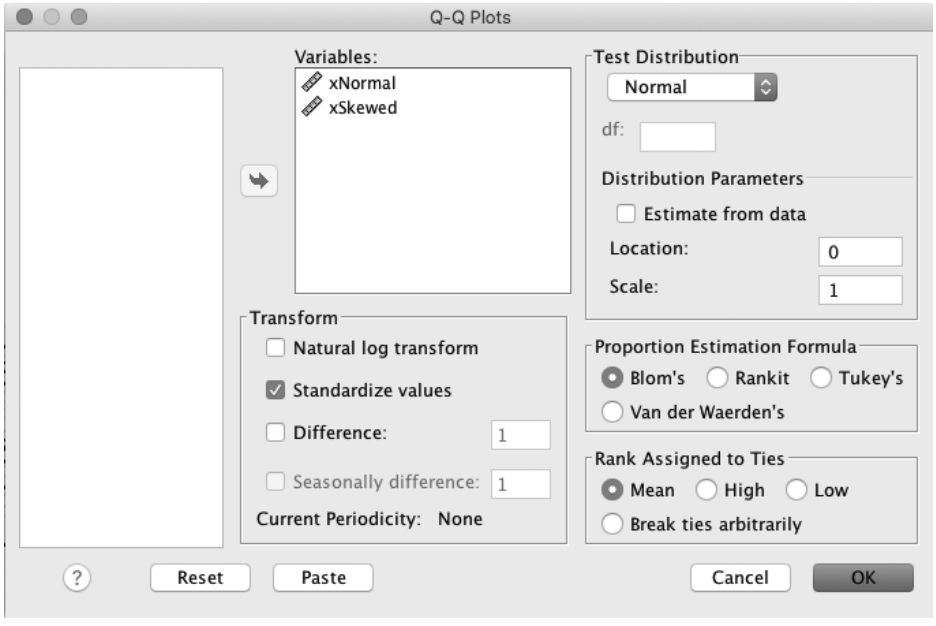
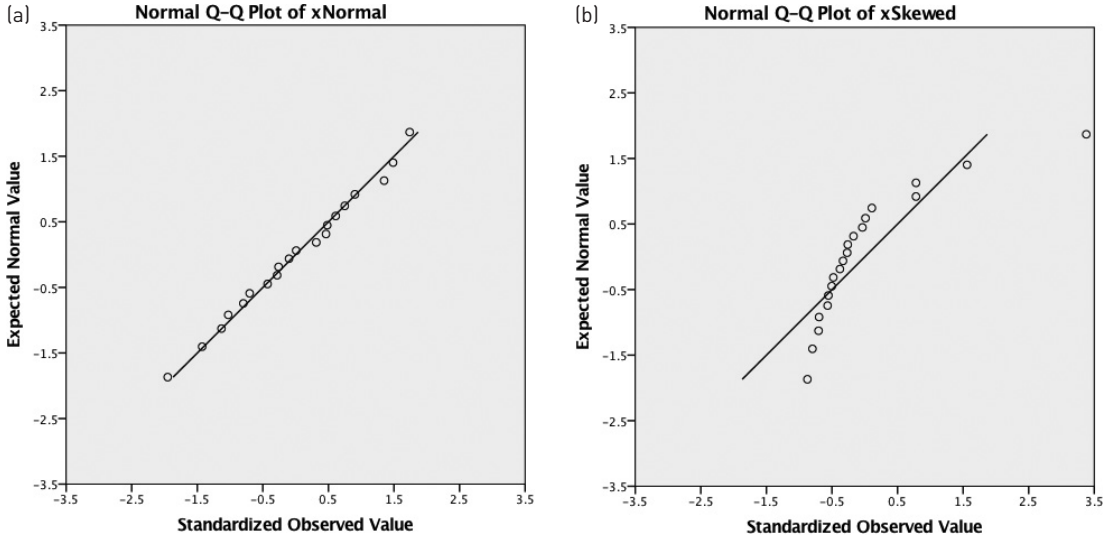


FIGURE 4.A3.6 ■ QQ Plots SPSS Output



QQ plots produced by SPSS for variables xNormal (a) and xSkewed (b).

Assessing Skew and Excess Kurtosis in SPSS

In Appendix 3.3 of Chapter 3, we described how skew and excess kurtosis can be computed from samples. Normal distributions have 0 skew and 0 excess kurtosis. In Appendix 3.1, we saw that skew and excess kurtosis are calculated through the Analyze→Descriptive Statistics→Descriptives . . . dialog in SPSS. Figure 4.A3.7 shows this analysis for xNormal and xSkewed. For xNormal, skew and excess kurtosis (−0.070 and −0.602, respectively) seem quite close to the values of 0 expected from normal distributions. For xSkewed, skew and excess kurtosis (2.382 and 6.488, respectively) seem very far from the values of 0 expected from normal distributions.

FIGURE 4.A3.7 ■ SPSS Output

Descriptive Statistics

	N	Skewness		Kurtosis	
	Statistic	Statistic	Std. Error	Statistic	Std. Error
xNormal	20	-.070	.512	-.602	.992
xSkewed	20	2.382	.512	6.488	.992
Valid N (listwise)	20				

Analysis of skew and excess kurtosis in SPSS for variables xNormal and xSkewed.

When we say that −0.070 and −0.602 are close to 0 and that 2.382 and 6.488 are not, one immediately wonders what metric allows us to judge close and far. A rule of thumb is that when the ratio of a statistic (skew or excess kurtosis) to its standard error (shown under the headings Std. Error) is outside the interval ±2, there is strong evidence that the sample was not drawn from a normal distribution. We will call this ratio z. For the sample drawn from the normal distribution, the z-values corresponding to skew and excess kurtosis are

$$z_{\text{Skew}} = \frac{-0.07}{0.512} = -0.14,$$

and

$$z_{\text{ExcessKurtosis}} = \frac{-0.602}{0.992} = -0.61.$$

Because neither ratio is outside the interval ±2, there is little concern that the sample was drawn from a nonnormal population.

For the sample drawn from the skewed distribution, the z-values corresponding to skew and excess kurtosis are

$$z_{\text{Skew}} = \frac{2.382}{0.512} = 4.65$$

and

$$z_{\text{ExcessKurtosis}} = \frac{6.488}{0.992} = 6.54.$$

Because both ratios are outside the interval ±2, there is serious concern that the sample was drawn from a nonnormal population.

The logic of this rule of thumb won't become clear until we reach Chapter 7, where we cover significance tests. However, the fact that the quantities we've computed are denoted with z_{Skew} and $z_{\text{ExcessKurtosis}}$ may provide a clue to the logic. Remember, most z-scores fall within ±2 standard deviations of the mean of the distribution. Therefore, it is unusual for a z-score to be outside the interval ±2. So, if it were true that a sample was drawn from a normal distribution with zero skew and zero excess kurtosis, it would be very unusual for z_{Skew} and $z_{\text{ExcessKurtosis}}$ to be outside the interval ±2. If one of these *is* outside the interval ±2, there is a strong possibility that the assumption that the sample was drawn from a normal distribution is wrong.