CHAPTER 6

Normal Distribution

Contents

INTRODUCTION

What is a normal distribution, and why is it important? The term "normal" means different things in everyday conversation and in statistics. In conversation, it implies something that is usual and, if appropriate, healthy.

To determine the "normal" resting blood pressure in healthy 10-year-old girls, take about 1000 healthy girls and measure resting blood pressure accurately. Then set the results out as percentiles, just as in the well-known growth charts for children. One percent of these children will have pressures >99th percentile, 5% of them will have pressures >95th percentile, 10% of them will have pressures >90th percentile, and so on. This percentile distribution is not statistically normal (see below), although it might be fairly close to it. Although these measurements are made in healthy children, it is not clear that those at the extremes of the distribution are necessarily healthy. Children with resting blood pressures in the upper part of the percentile chart may have essential hypertension when they are adults. If this is true, then being above the 95th percentile, for example, may mean illness in the future, even if there is no illness now. This dilemma has been emphasized in relation to the standard growth charts for children (Cole, 2010). Because children are heavier now than they were 20 years ago, growth charts for the weights of "healthy" children at any age have a higher value for a given percentile now than they did

Biostatistics for Medical and Biomedical Practitioners
http://dx.doi.org/10.1016/B978-0-12-802387-7.00006-8

earlier. Thus, a child overweight on a 1990 chart is in the normal range on a 2010 chart, but possibly destined to a variety of diseases in early adult life. The World Health Organization (WHO) distinguishes between a *normal* chart and a *standard* chart, the latter involving only children whose weights indicate future health (not an easy task).

NORMAL OR GAUSSIAN CURVE

In statistical usage, the term "normal" is applied to a distribution specified by the equation

$$f_i = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$$

f_i is the height of the curve at value X_i, μ is the mean of the distribution, and σ is its standard deviation; π and e are constants, μ and σ are parameters, and X_i is a variable. All normal distribution curves have the same bell shape (Figure 6.1), but differ in their means and standard deviations. The whole expression defines the Gaussian curve. Changing the values of the parameters changes the position and width of the curve, but not its bell shape. The value of X_i gives an estimate of the function of the expression at that value.

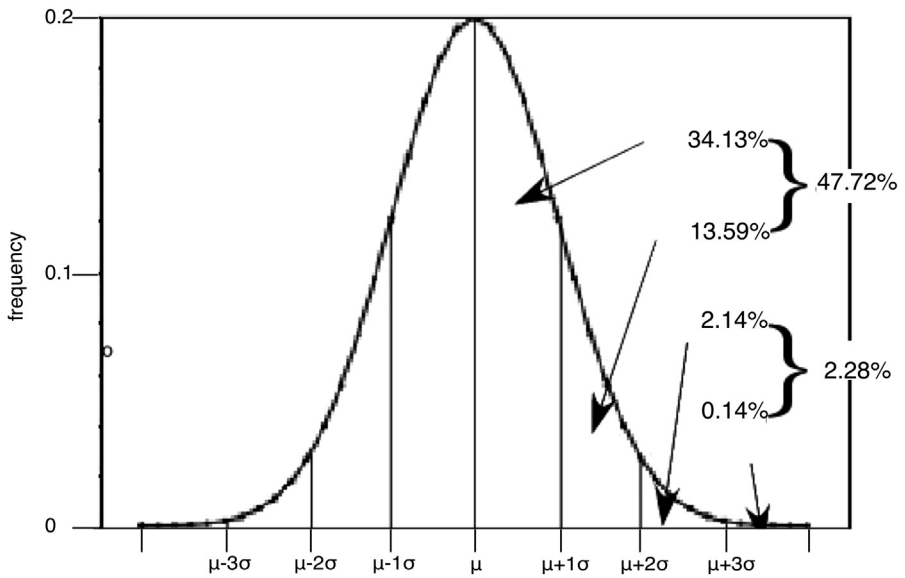

Figure 6.1 Normal probability density plot of X variate against frequency. The areas under the curve as related to deviations from the mean (μ) in units of standard deviation (σ) are shown. 2.28% of the area under the curve is greater than two standard deviations from the mean, and 0.14% of the area is greater than three standard deviations from the mean. 2.5% of the area under the curve is beyond 1.96 standard deviations from the mean. Such a normal curve is often described as $\mathcal{N}(0, 1)$ where the $\mathcal{N}$ refers to the normal distribution, not sample size, 0 is the mean, and 1 is the variance.

The first description of the normal distribution was given in 1733 by Abraham de Moivre (1667–1754), an Anglo-French mathematician who left France for England after the persecution of the Huguenots. He started with probability theory, in particular the binomial theorem (see Chapter 11) and developed the formula so that it would be possible to compute the binomial distribution when the number of binary events (e.g., coin tosses) was very large (Gridgeman, 1966). Later, scientists, particularly astronomers, began to be concerned about how to allow for measurement errors in celestial mechanics, and Laplace, Gauss, and others began to associate the distribution of these errors with the normal curve; the "true" value was the mean, and the errors distributed around the mean produced a normal curve.

This distribution produces the well-known bell-shaped or Gaussian curve (Figure 6.1).

Why is the X-axis labeled in standard deviation units? Different Gaussian curves have different means and standard deviations (Figure 6.2).

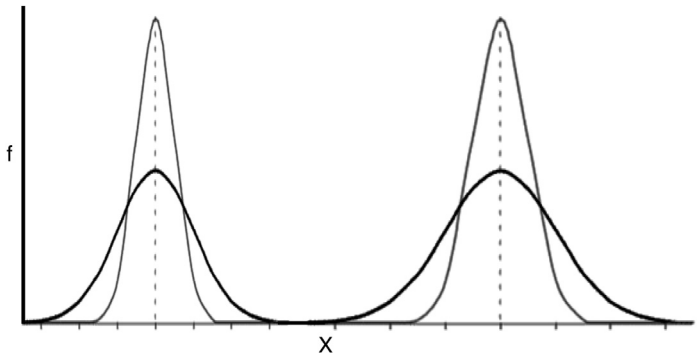

Figure 6.2 Normal Gaussian curves with different means and standard deviations.

Between any two points on the X-axis, expressed in terms of number of standard deviations from the mean, the area under the curve is the same for all normal curves. Therefore, subtracting the mean μ from each value of X_i and then dividing by the standard deviation σ produces *standard deviates* symbolized by z. Thus

$$z_i = \frac{X_i - \mu}{\sigma}$$

This is one type of linear transformation and it achieves two goals. By subtracting μ from every value of X_i, the numerator becomes 0. The normal curves are shifted so that each curve has a mean of 0 (Figure 6.3). Second, by dividing the numerator by the standard deviation, every normal curve has a unit standard deviation. Therefore, wide curves with big standard deviations and narrow curves with small standard deviations each assume a standard shape with a mean of 0 and a standard deviation of 1 (Figure 6.3).

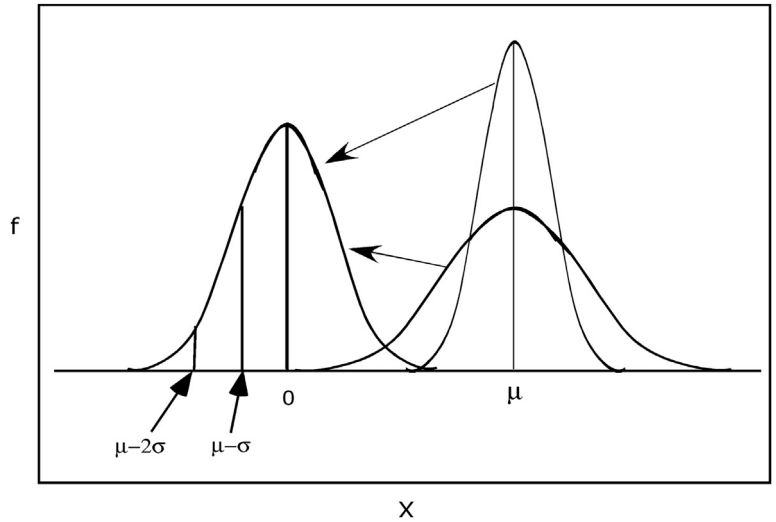

Figure 6.3 *z* Transform. The curve on the left has a mean of 0 and a standard deviation of 1 unit, and is what both the curves on the right would look like after the *z* Transformation.

Once the z transformation has been performed, the new normal curve will be that shown in Figure 6.1. The z transformation is based on population values for the mean and the standard deviation.

The Quincunx

A quincunx is a device resembling a pinball machine that was developed by Sir Francis Galton (1822—1911) to illustrate the theory of random errors (Figure 6.4).

Each pin deflects a falling ball either to the left or to the right. Despite the most careful construction of pins and balls, tiny imperfections or even air currents make each move at random to either side. Intuitively, it is very unlikely for any ball to move always to the left or always to the right, so that the bins at the ends have very few balls. Most balls tend to have roughly equal numbers of leftward or rightward deflections, accounting for the peaks in the central bins. The more pins, bins, and balls there are, the closer the distribution matches the normal Gaussian distribution.

Fascinating pictures of balls moving through the pins and into the bins can be seen at http://www.mathsisfun.com/data/quincunx.html and http://www.jcu.edu/math/isep/quincunx/quincunx.html.

Properties of the Normal Curve

The normal distribution curve has important mathematical properties. It is symmetrical about a mean (μ) of 0, and 68.26% of the measurements are within the limits of one standard deviation (σ) below to one standard deviation above the mean; this one standard

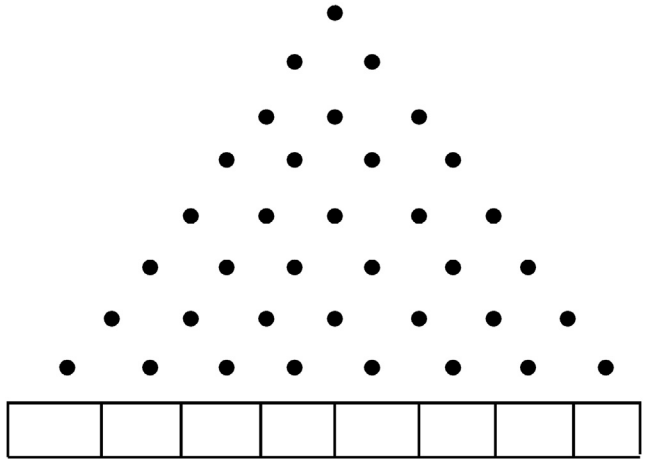

Figure 6.4 Quincunx consisting of regularly placed pins arranged vertically on a board. Steel balls much smaller than the space between the pins are dropped from the top, bounce off the pins, and end up in one of the bins at the bottom. If balls are dropped into bins, the following distributions might occur (Figure 6.5).

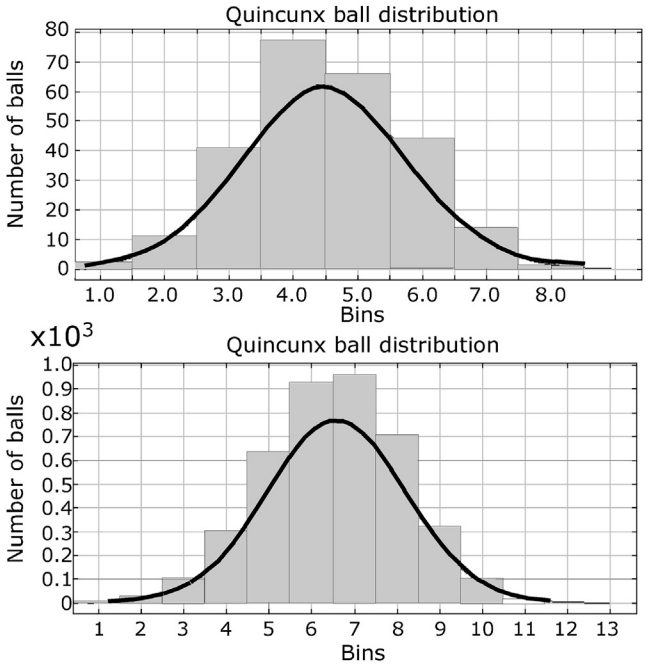

Figure 6.5 Upper: Distribution of 256 balls into 8 bins. The normal Gaussian curve is superimposed on the histogram. The fit is fair. Lower: Results after dropping 4096 balls into 12 bins. The histogram is more symmetrical. *(Constructed with applet from http://www.jcu.edu/math/isep/quincunx/quincunx.html.)*

deviation value is the point of inflection of the curve. The mean value is the most frequent measurement (the mode). 0.025 (or 2.5%) of the area under the curve is above a value of X = mean + 1.96 times the standard deviation; because the curve is symmetrical, there is an equal area below a value of X = mean − 1.96 times the standard deviation. These two areas added together give 0.05 (5%) of the area under the curve that is beyond the limits set by the mean ±1.96 times the standard deviation. Figure 6.6 shows how the areas under the curve are often represented.

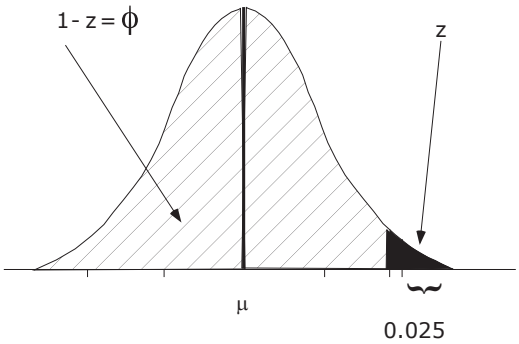

Figure 6.6 Components of normal curve. *z* is the area beyond some value on the horizontal X-axis.

Most texts and tables give the values of the total area under the curve beyond any given value of z (black area). The remaining area (cross-hatched area) to the left of the z demarcation is $1 - z$, often termed φ; this is the cumulative area from the left-hand end of the curve to the value of z. Occasionally some tables have different shaded areas, and the reader should check to see what the listed values refer to. Cumulative areas can be calculated at http://stattrek.com/online-calculator/normal.aspx and http://www.danielsoper.com/statcalc3/calc.aspx?id=2.

For a continuous distribution of this type, it makes little sense to ask about the probability of obtaining a given X value. The probability of an adult male human weighing 68.83997542001 kg is virtually 0. What does matter is the area under the curve between different values of X. As examples:

1. What proportion of the area under the curve lies between the mean μ and one standard deviation σ below the mean? From Figure 6.1, the area between these limits is 0.3413 or 34.13% of the total area.
2. What proportion of the area under the curve lies between the μ and 0.5 standard deviations below the mean? From tabulated areas under the curve, the area under the curve for $\mu - 0.5\sigma$ is 0.1914 (Figure 6.7(a)).
3. What proportion of the area under the curve lies between 0.5 and 1 standard deviations below the mean? From tabulated areas under the curve, the area under the curve for $\mu - 0.5\sigma$ is 0.1914, and for $\mu - \sigma$ is 0.3413. Therefore, the required area is $0.3413 - 0.1914 = 0.1499$ (Figure 6.7(b)).

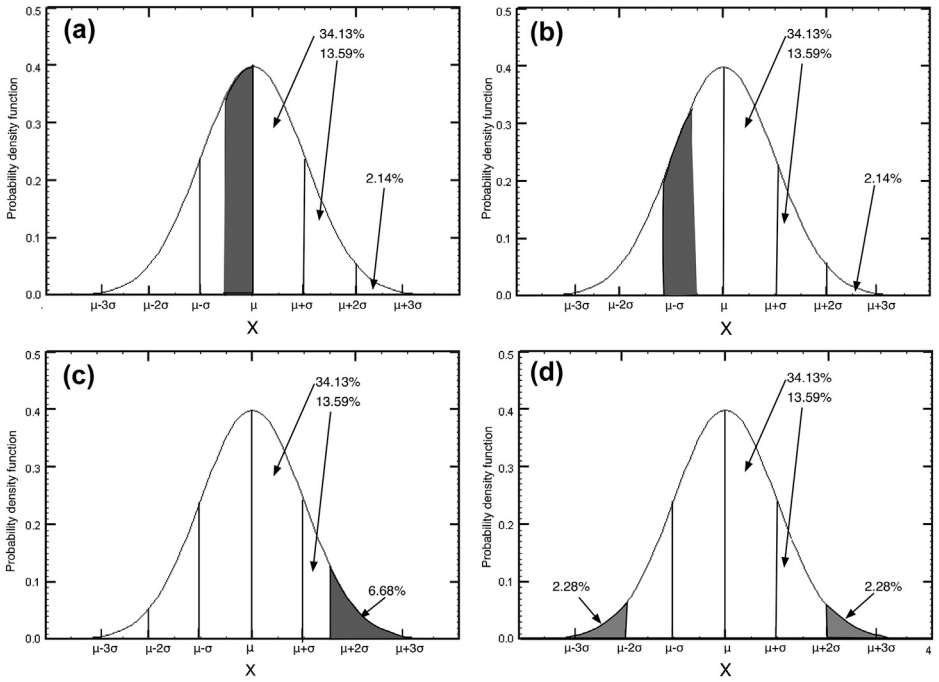

Figure 6.7 Areas under the curve.

4. What proportion of the area under the curve is >1.5 standard deviations above the mean? From tabulated areas under the curve, the area under the curve for $\mu + 1.5\sigma$ is 0.9332. The whole area under the curve is 1. Therefore the required area is $1 - 0.9332 = 0.0668$ (Figure 6.7(c)).
5. What proportion of the area under the curve is >2 standard deviations above and below the mean? From tabulated areas under the curve, the area under the curve for $\mu \pm 2\sigma$ is 0.0456 (Figure 6.7(d)).
6. 99.73% of the area under the curve lies between $\mu \pm 3\sigma$, 99.9937% between $\mu \pm 4\sigma$, and 99.999942% between $\mu \pm 5\sigma$.

Problem 6.1

Use the online calculator to determine the area under the normal curve between the limits of 0.75σ below the mean to 1.5σ above the mean.

All these areas can be calculated easily using http://davidmlane.com/hyperstat/z_table.html, http://easycalculation.com/statistics/normal-distribution.php, and http://psych.colorado.edu/~mcclella/java/normal/accurateNormal.html.

The normal curve can also be converted into a cumulative probability density curve with its characteristic S or sigmoid shape (Figure 6.8).

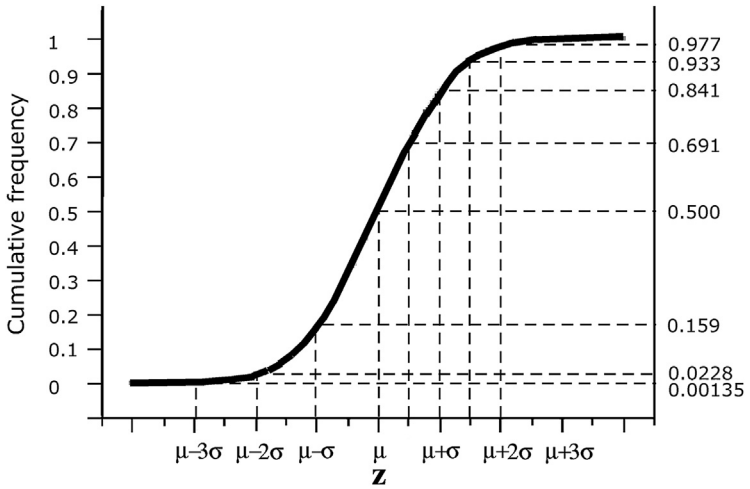

Figure 6.8 Cumulative probability density curve (frequency curve). Because of the changing slope of the curve, a change in the X-axis from μ to $\mu + 1\sigma$ changes the cumulative frequency from 0.500 to 0.691 for a difference of 0.191, whereas a one standard deviation change from $\mu + 2\sigma$ to $\mu + 3\sigma$ changes the cumulative frequency from 0.977 to 0.99865 (not shown) for a difference of 0.02165. These cumulative frequencies can be obtained easily online. The lower scale shows standard deviates (see below).

POPULATIONS AND SAMPLES

Initially, attention was paid to the sampling features of the mean of large samples; the sample standard deviation and the population standard deviation were assumed to be virtually identical. Gosset realized that some correction was required if inferences about small samples were to be made from the normal Gaussian curve. He introduced a value termed t, similar to z but with one important difference. Whereas for means

$$z = \frac{\overline{X_i} - \mu}{\sigma_{\overline{X}}},$$

where $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}}$ and $\sigma = \frac{\sum (X_i - \mu)^2}{N}$,

Gosset used an analogous expression

$$t = \frac{\overline{X_i} - \mu}{s_{\overline{X}}},$$

where $s_{\overline{X}} = \frac{s}{\sqrt{N}}$ and $s = \frac{\sum (X_i - \overline{X})^2}{N-1}$.

There were two added features to his "Student's" t distribution. One was that $N - 1$ is a specific example of $N - k$, where k is the number of degrees of freedom, and second that the areas under the normal curve varied with the degrees of freedom. For large

sample size, >200, the t and z distributions were identical, but for a sample of 11 with 10 degrees of freedom 95% of the area under the normal curve lies within the limits of $\mu \pm 2.228 s_{\overline{X}}$, not $\pm 1.96 \sigma_{\overline{X}}$ as for the z table.

DESCRIPTION OF THE DISTRIBUTION SHAPE

The mean μ gives the average size of the measurements, a measure of central tendency, and the square root of the variance gives the standard deviation σ, which is a measure of variability. Unfortunately, most real distributions are not exactly normal, and may even be very far removed from it. One common form of nonnormality is to have a few very large measurements, with the effect shown diagrammatically in Figure 6.9.

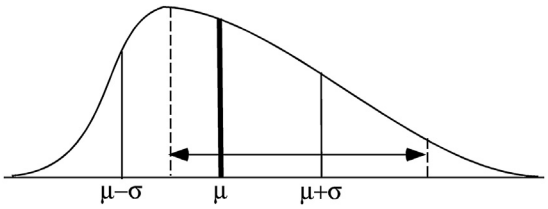

Figure 6.9 The dotted lines and double-headed arrow indicate the range within which two-thirds of the measurements lie.

The curve is no longer symmetrical, but is pulled to the right toward the larger measurements; it is *skewed* to the right. The mean is no longer near the most frequent measurement, but to the right of it. More importantly, the range from one standard deviation below to one standard deviation above the mean is not symmetrical and does not include a known proportion of the measurements. For some distributions with extreme skewing, the mean gives no useful information.

Many other types of nonnormality exist. One might guess that the distribution of serum electrolyte concentrations in healthy people approximates a normal distribution, but Elveback and her colleagues (Elveback et al., 1970; Elveback, 1972) showed that the distributions of several commonly obtained laboratory biochemical values are not normal in the Gaussian sense. Furthermore, they pointed out that to calculate mean and standard deviations from the data in the hope that the 2.5% with the highest values and the 2.5% with the lowest values could be declared abnormal (and therefore unhealthy) would lead to serious underdiagnosis of illness. In their studies, the distributions of commonly determined biochemical values were often leptokurtotic (excessively peaked), with more than 68% of the area between the limits of $\mu - \sigma$ and $\mu + \sigma$ and with excessively long tails (Figure 6.10). Therefore, the standard deviation of the leptokurtotic distribution is wider than that for a normal distribution, and this can have serious clinical consequences. For example, if the (leptokurtotic) distribution of serum calcium is

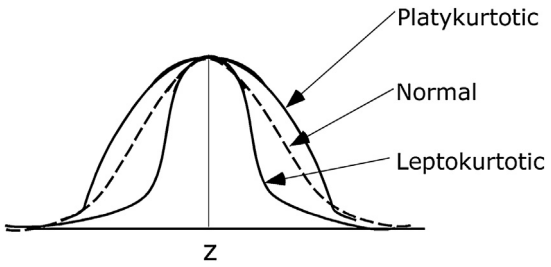

Figure 6.10 Normal, **L**eptokurtotic (for **L**ong-tailed), and Pl**at**ykurtotic (for fl**at** top) curves.

regarded as normal, the "normal" limits are too wide, and the upper critical normal value based on the calculated standard deviation would lead investigators to miss about 20% of patients with hyperparathyroidism. The problems of using the "normal range" to make clinical diagnoses has been discussed many times (Mainland, 1971; Murphy and Abbey, 1967). The International Federation of Clinical Chemistry recommends the term "reference range" rather than the poorly defined term "normal range" (Strike, 1981).

Other symmetrical curves are short-tailed and flat-topped (platykurtotic) so that fewer than 68% of the measurements are between the limits of $\mu - \sigma$ and $\mu + \sigma$ (Figure 6.10).

One approach to defining shape is by computing *moments*. The first moment about the mean is $\frac{\sum (X_i - \mu)}{N}$, that is, the average value of the deviations from the mean, and this is zero. The second moment about the mean is $\frac{\sum (X_i - \mu)^2}{N}$, the average of the squared deviations of X_i from the population mean, that is, the population variance. The third moment about the mean is $\frac{\sum (X_i - \mu)^3}{N}$, the average value of the sum of the cubed deviations from the mean; it is designated as m_3 or k_3. The fourth moment about the mean is $\frac{\sum (X_i - \mu)^4}{N}$, the average value of the sum of the deviations about the mean to the fourth power; it is designated as m_4 or k_4.

Skewness

With perfect symmetry, the third moment is zero, and if the distribution is nearly symmetrical the cubed numerator yields a small number because pluses and minuses almost cancel out. If the curve is skewed to the left, there will be more negative than positive numbers and k_3 will be negative. Conversely, skewing to the right yields a positive result. Because this expression has cubed units, it is customary to divide by the cube of the standard deviation to obtain a dimensionless measurement termed γ_1:

$$\gamma_1 = \frac{k_3}{\sigma^3}.$$

With sample data, the expression becomes:

$$g_1 = \frac{k_3}{s^3}, \text{ which is the above equation corrected for sample size.}$$

Standard computer programs perform the calculations. A free online calculator can be found at http://www.wessa.net/skewkurt.wasp, and skewness can be calculated by the Skew function in Excel. The program gives the probability that $g_1 = 0$, or tables of g_1 can be consulted (see http://mvpprograms.com/help/mvpstats/distributions/SkewnessCriticalValues or http://www.engl.unt.edu/~leubank/researchmethods/appendicesa&b.html) to indicate if a positive or negative value of g_1 is large enough that the hypothesis that $g_1 = 0$ can be rejected.

Kurtosis

If the distribution is symmetrical, then kurtosis can be assessed with the fourth moment about the mean to calculate k_4 or m_4, and this can be made dimensionless and standardized to the standard deviation by

$$\gamma_2 = \frac{k_4}{\sigma^4}: \quad \text{the equivalent sample values are } g_2 = \frac{k_4}{s^4}.$$

γ_2 for a normal curve should be 3. It is customary to subtract 3 from the value of g_2; if $g_2 - 3$ is significantly below 0 then the curve is platykurtotic and if it is significantly above 0 then it is leptokurtotic. If either of these distorted normal curves is encountered, it is worth considering why they have occurred. A leptokurtotic curve might indicate the superimposition of two normal curves with the same mean but different standard deviations, and a platykurtotic curve might indicate the superimposition of two normal curves with similar standard deviations but different means (Zar, 2010). Platykurtosis often occurs when batches of data are collected at different times, with a slight shift in mean from one batch to the other. A free online calculator can be found at http://www.wessa.net/rwasp_skewness_kurtosis.wasp#output and http://www.calculatorsoup.com/calculators/statistics/descriptivestatistics.php. Kurtosis can also be calculated by the KURT function in Excel.

A rough assessment of kurtosis can be made with the ratio interquartile distance/1.35 (pseudostandard deviation or PSD). If the standard deviation $\ll$PSD, the distribution is platykurtotic, and if the standard deviation $\gg$PSD, the distribution is leptokurtotic (Hamilton, 1990).

Small sample sizes produce wide confidence limits.

Problem 6.2

Take the data from Table 4.15 and test them for skewness and kurtosis. Also calculate the PSD and decide if the curve is leptokurtotic or platykurtotic.

DETERMINING NORMALITY

Initially inspect the histogram, stem-and-leaf diagram, or box plots for asymmetry and outliers. Not only is the distribution seen clearly but any outliers are identified. Now

that all of these graphics are incorporated into standard software programs, there is no excuse for not determining if a set of observations appears to be normal.

Skewing is easy to detect, but symmetrical distributions with straggling of the highest and lowest measurements are more difficult to judge by eye. The PSD described above is an easy way of assessing this. This is important to know because these straggling tail values can grossly distort the standard deviation and make comparisons between groups inefficient. In addition, the mean $\pm 0.25\ s$ should include about 40% of the measurements. If many more are included, this suggests that the upper part of the curve is narrower than it should be from a Gaussian distribution.

Some calculations yield a number that can be used to determine the likelihood that a given data set could represent a normal distribution. Shapiro and Wilk's test (Shapiro and Wilk, 1965), Lilliefors test (Lilliefors, 1967), and D'Agostino's test (D'Agostino and Pearson, 1973) are available in most statistical programs. Also see http://www.wessa.net/Ian.Holliday/rwasp_Shapiro-Wilks%20Test%20for%20Normality.wasp for the Shapiro–Wilk's test, http://www.wessa.net/rwasp_skewness_kurtosis.wasp#output for Agostino's test, and http://in-silico.net/statistics/lillieforstest, http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Normality.htm for the Lilliefors test. These tests are easy to do and interpret, but may not indicate where the distribution has departed from normality and thus may not indicate what to do about the problem.

A test for both skewness and kurtosis combined is the Jarque–Bera test (Jarque and Bera, 1980). The test value *JB* is calculated from:

$$JB = \frac{N}{6}\left(s^2 + \frac{(k-3)^2}{4}\right),$$

where s is the sample skewness and k is the sample kurtosis. The statistic *JB* has an asymptotic chi-square distribution with two degrees of freedom (Chapter 7), and tests the assumption that the data come from a normal distribution. It can be performed online at http://www.wessa.net/rwasp_skewness_kurtosis.wasp#output.

Graphic tests. Some tests are graphic: for example, the use of probability paper or normal quantile plots. In Figure 6.8, the typical sigmoid cumulative frequency curve is shown. If data points could be plotted on such a curve with an excellent fit, it is reasonable to conclude that the distribution from which those points came was normal. It is, however, difficult to assess S-shaped curves, and easier to assess straight lines. Fortunately the S-shaped curve can be made straight by plotting its points on probability paper (Figure 6.11).

There are other ways of plotting and assessing normal distributions. Normal quantile plots are standard on most computer programs. Figure 6.8 shows how for equal standard deviation increments the incremental change in the cumulative percentages became progressively less as the distance from the mean increased. This was rectified in the

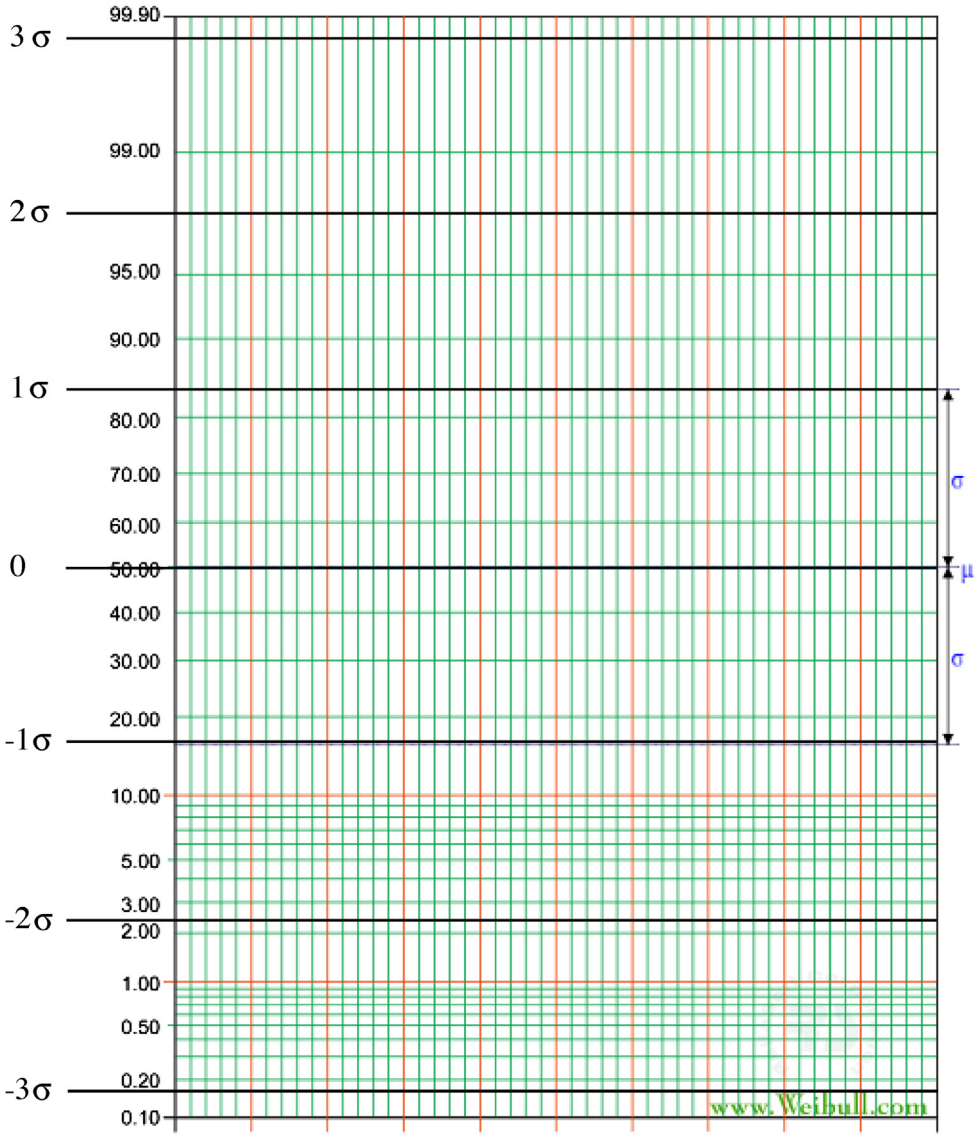

Figure 6.11 Normal probability paper. The X-axis is linear, but the Y-axis becomes magnified as one goes further away from the mean. The thick horizontal lines are placed 1, 2, and 3 standard deviations above and below the mean, as defined by the areas under the normal curve. Graphs for this and other distributions may be obtained free online from the ReliaSoft Corporation at http://www.weibull.com. (see Download Probability Plotting Papers).

probability paper that features in Figure 6.11 where the heavy horizontal lines demarcate standard deviation units rather than cumulative percentages. Thus, the vertical Y-axis is linear in standard deviation units, sometimes called standard normal deviates or normal equivalent deviates (NEDs). (The transformation of the S-shaped curve to the linear

NED scale is attributed to the pharmacologist J.H. Gaddum (1900–1965) (Eggert and Stick, 1984).) The cumulative percentages corresponding to various z values are shown in Table 6.1.

Table 6.1 Normal equivalent deviates (NEDs). The NED is the same as the area defined as $1 - z = \varphi$ in Figure 6.4

Cumulative percentage	Normal equivalent deviate
0.00135	−3
0.02775	−2
0.1589	−1
0.5	0
0.6915	0.5
0.8413	1
0.933	1.5
0.977	2
0.9986	3

With this information, plot the data on probability graph paper. The X-axis is the measurement of the variable, and on the Y-axis plot the observed cumulative percentages of the X variable. As an example, Table 6.2 gives data on the distribution of heights of eighteenth-century English soldiers in America (Komlos and Cinnirella, 2005).

Table 6.2 Distribution of heights. Discussion of columns in text below

Height (in)	Frequency	Cumulative frequency	% Cumulative frequency	NED
59	10	10	0.99	−2.330
60	14	24	2.15	−2.024
61	36	60	5.38	−1.609
62	50	110	9.87	−1.289
63	98	208	18.65	−0.891
64	172	380	34.08	−0.410
65	174	554	49.69	−0.008
66	184	738	66.19	0.418
67	119	857	76.86	0.734
68	127	984	88.25	1.188
69	60	1044	93.63	1.524
70	31	1075	96.41	1.800
71	22	1097	98.39	2.142
72	14	1111	99.64	2.687
73	4	1115	100.00	
Mean = 65.6 Sd = 2.54				

The mean of grouped data can be calculated using http://www.easycalculation.com/statistics/group-arithmetic-mean.php, and the mean and standard deviation from http://www.knowpapa.com/sd-freq.

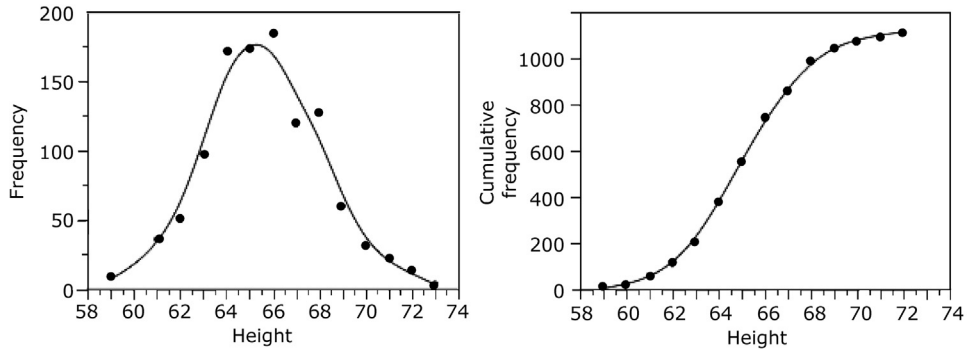

Figure 6.12 Frequency and cumulative frequency distributions of heights. The cumulative curve evens out the irregularities shown in the original distribution.

The resultant frequency and cumulative frequency distributions are shown in Figure 6.12.

One way of testing the normality of the distribution is to plot the cumulative percentage against the height on probability paper (Figure 6.13).

Another method that does not involve probability paper is to plot the NEDs on the Y-axis and the height on the X-axis. This involves calculating the cumulative percentage, transforming these values into NEDs (as shown in Table 6.1), from detailed tables, from Figure 6.6, or from online calculators http://stattrek.com/online-calculator/normal.aspx, http://sampson.byu.edu/courses/zscores.html or http://davidmlane.com/hyperstat/z_table.html (Figure 6.14). The calculators are more accurate but the results are similar.

UNGROUPED DATA

Table 6.3 shows the weight (in pounds) of 13 dogs. For ungrouped data, the individual frequencies are 1,1,1…1; the cumulative frequencies are 1,2,3…N; and the relative cumulative frequencies are $1/N$, $2/N$…N/N. Thus, because in Table 6.3 $N = 13$, the first relative cumulative frequency is $1/13 = 0.077$, the second is $2/13 = 0.154$, and so on.

Plotting the NED in column 4 against the actual values in column 2 gives Figure 6.15.

This is not as good a fit as shown in Figure 6.14, and some points are quite far from the theoretical line of normality. To determine if these data are consistent with a normal distribution, enter them into the programs http://stattrek.com/online-calculator/normal.aspx or http://www.wessa.net/rwasp_skewness_kurtosis.wasp#output. These show that skewness and kurtosis are not abnormal enough to allow rejection of the null hypothesis.

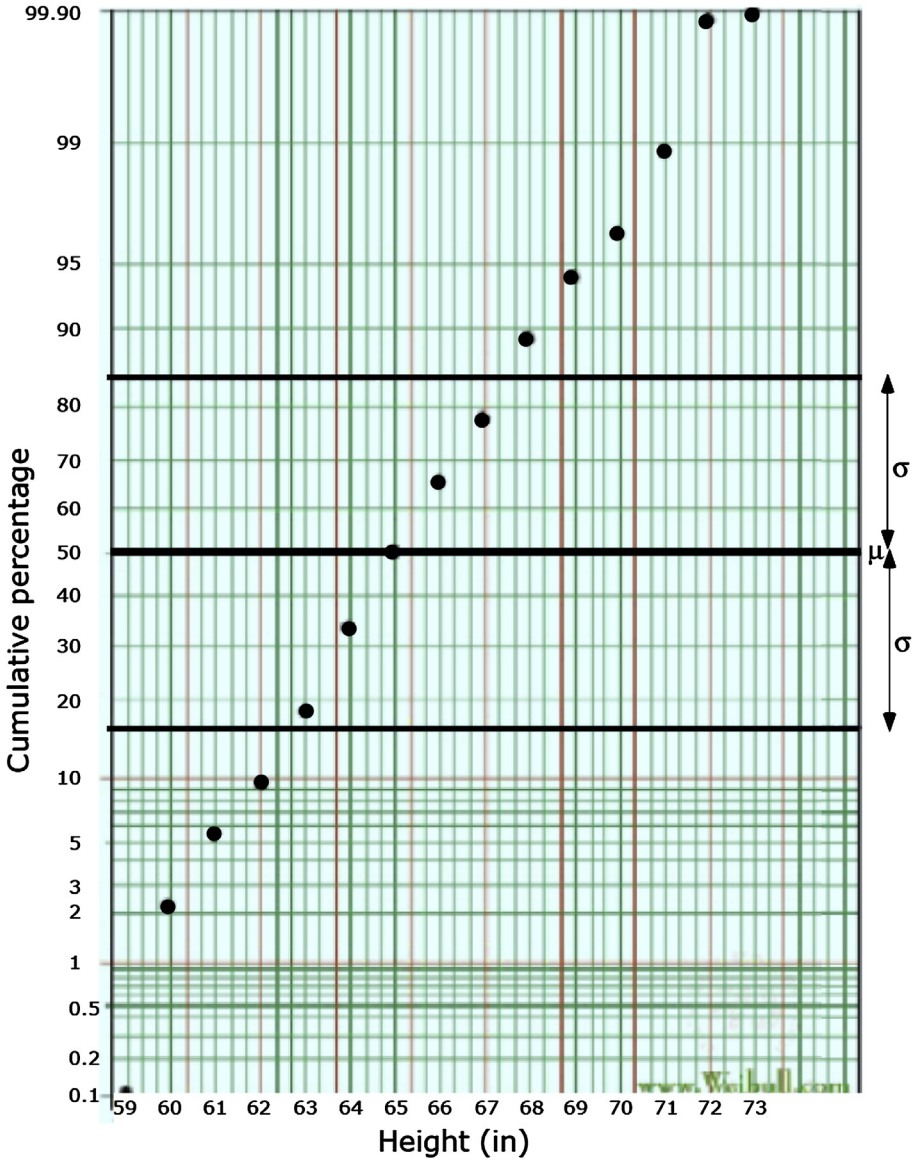

Figure 6.13 Probability paper showing cumulative heights. Apart from the ends, which are based on very small numbers, the distribution is reasonably linear. This suggests that the distribution is approximately normal, although it cannot be truly normal because it is truncated at each end.

The graphs in Figures 6.14 and 6.15 are *quantile* or Q–Q plots. To determine if these data are still compatible with a normal distribution, some programs, such as JMP, insert 95% confidence limits (Figure 6.16). Quantile plots can be implemented online by http://www.wessa.net/rwasp_harrell_davis.wasp#output, but without confidence

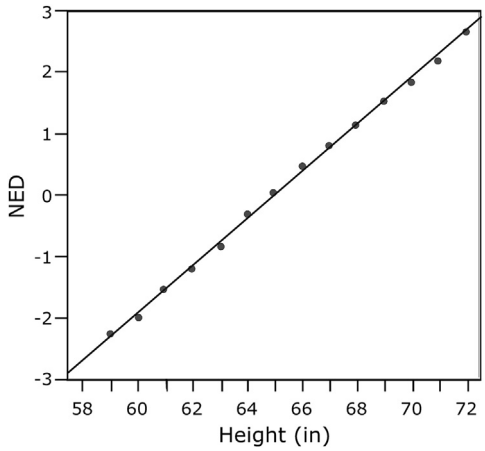

Figure 6.14 Using normal equivalent deviate (NED) to determine normality on ordinary graph paper. This is the equivalent of using probability paper.

Table 6.3 Dog data

Number	Weight (lbs)	Cumulative relative frequencies	NED
1	17.2	0.077	−1.426
2	20.8	0.154	−1.019
3	21.0	0.231	−0.736
4	21.2	0.308	−0.502
5	21.5	0.385	−0.292
6	24.2	0.462	−0.095
7	24.3	0.539	0.098
8	25.6	0.616	0.295
9	27.7	0.693	0.504
10	31.0	0.77	0.739
11	34.5	0.857	1.067
12	38.7	0.924	1.433
13	39.2	1.00	—
	$\sum X_i = 346.9$ $\overline{X} = 26.68$ $s = 7.12$		

limits. They can also be produced in Excel http://facweb.cs.depaul.edu/cmiller/it223/normQuant.html.

If the distribution had been perfectly normal, then all the points would have been on the line. These plots show exactly where the deviations from normality occur, and give a probability that can be used to decide if the distribution is sufficiently far from a normal distribution.

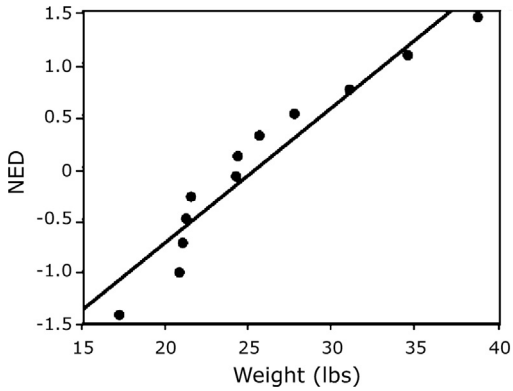

Figure 6.15 Dog data in normal equivalent deviate (NED) plot.

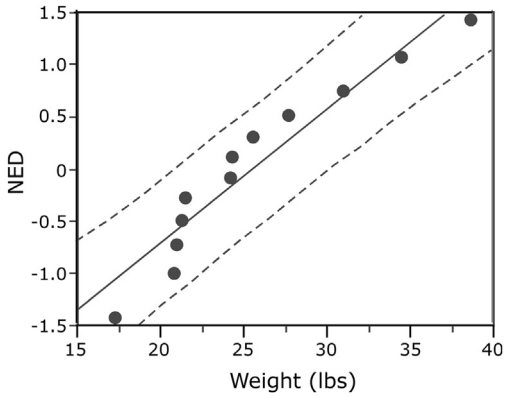

Figure 6.16 Normal quantile plot. The diagonal line indicates perfect normality. The vertical scale shows normal equivalent deviates (NEDs), and the horizontal scale shows the weight. The dashed curved lines show the 95% confidence limits.

Most computer programs produce similar figures. They can also be performed with Excel: see http://facweb.cs.depaul.edu/cmiller/it223/normQuant.html.

Problem 6.3
Draw a quantile plot of the data from Table 4.15.

HOW IMPORTANT IS NORMALITY?

The normal distribution curve plays a central part in statistical thinking and modeling. If the distribution is markedly abnormal, then even though a test statistic can be calculated, inferences drawn from it may be wrong. Furthermore, because statistics based on the

normal distribution are very efficient, most people try to use normalizing transformations so that conventional statistics can be used. Alternatively, there are tests to use when distributions are markedly abnormal; these are called nonparametric tests.

How necessary is it to test normality and how much abnormality of the distribution can be tolerated? If a stem-and-leaf diagram, a box plot, or a histogram appear roughly normal and symmetrical by eye, then, as long as there are no extreme outliers, the distribution is normal enough, and more elaborate tests may not be needed. This attitude is supported by an article by Sall (with the appropriate title "Leptokurtophobia: irrational fear of non-normality") that appeared in the technical publication for JMP users, JMPer Cable (Sall, 2004). He pointed out that: "In large samples it is easy to detect non-normality, but it doesn't matter. In small samples, non-normality may matter, but you can't detect it." Sall concluded, however, that graphical testing, even if of limited use for detecting nonnormality, was of value looking for anomalies or a pattern that might be a clue to some hidden structure of the distribution.

REFERENCES

Cole, T.J., 2010. Babies, bottles, breasts: is the WHO growth standard relevant? Significance, Virtual Medical Issue 6–10.

D'Agostino Jr., R.B., Pearson, E.S., 1973. Tests of departure from normality. Biometrika 60, 613–622.

Eggert, P., Stick, C., 1984. The pattern of bilirubin response to phototherapy for neonatal hyperbilirubinemia. Pediatr. Res. 18, 682.

Elveback, L., 1972. A discussion of some estimation problems encountered in establishing "normal" values. In: Gabrieli, E.R. (Ed.), Clinically Oriented Documentation of Laboratory Data. Academic Press, New York.

Elveback, L.R., Guillier, C.L., Keating Jr., F.R., 1970. Health, normality, and the ghost of Gauss. J. Am. Med. Assoc. 211, 69–75.

Gridgeman, N.T., July 28, 1966. The normal curve. New Sci. 211–213.

Hamilton, L.C., 1990. Modern Data Analysis. A First Course in Applied Statistics. Brooks/Cole Publishing Co, Pacific Grove, CA.

Jarque, C.M., Bera, A.K., 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Econ. Lett. 6, 3.

Komlos, J., Cinnirella, F., 2005. European Heights in the Early 18th Century (Online). Available: http://epub.ub.uni-muenchen.de/.

Lilliefors, H.W., 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. J. Am. Stat. Assoc. 64, 399–402.

Mainland, D., 1971. Remarks on clinical "norms". Clin. Chem. 17, 267–274.

Murphy, E.A., Abbey, H., 1967. The normal range—a common misuse. J. Chronic Dis. 20, 79–88.

Sall, J., 2004. Leptokurtophobia: Irrational Fear of Non-normality. JMPer Cable.

Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika 52, 591–611.

Strike, P.W., 1981. Medical Laboratory Statistics. John Wright and Sons, Ltd, Bristol.

Zar, J.H., 2010. Biostatistical Analysis. Prentice Hall, Upper Saddle River, NJ.