Available online at www.sciencedirect.com

SCIENCE DIRECT®

Clinica Chimica Acta 366 (2006) 112 – 129

www.elsevier.com/locate/clinchim

Review

Testing experimental data for univariate normality

A. Ralph Henderson*

Department of Biochemistry, University of Western Ontario, London, Ontario, Canada, N6A 5C1

Received 13 October 2005; received in revised form 7 November 2005; accepted 9 November 2005
Available online 4 January 2006

Abstract

Background: Many experimentally-derived data sets are generated in the practice of clinical chemistry. Graphical presentation is essential to assess the data distribution. The distribution must also be assessed quantitatively. These approaches will determine if the data is Normal or not. Finally the results of these tests of Normality must be shown to be free of sample size effects.
Methods: Four experimentally-derived data sets were used. They represented normal, positive kurtotic, positive- and negatively-skewed distributions. These data sets were examined by graphical techniques, by moment tests, by tests of Normality, and monitored for sample size effects.
Results: The preferred graphical techniques are the histogram and the box-and-whisker plots that may be supplemented, with advantage, by quantile–quantile or probability–probability plots. Classical tests of skewness and kurtosis can produce conflicting and often confusing results and, as a consequence, the alternative use of the newer L-moments is advocated. Normality tests included the Kolmogorov–Smirnov (Lilliefors modification), Cramér-von Mises and Anderson–Darling tests (empirical distribution function statistics) and the Gan–Koehler, Shapiro–Wilk, Shapiro–Francia, and Filliben tests (regression/correlation techniques). Of these only the Anderson–Darling, Shapiro–Wilk, and Shapiro–Francia tests correctly classified all four test samples. The effect of sample size on the resulting p-value was investigated using Royston's V'/v' graphical test.
Conclusions: A systematic approach to Normality testing should follow the route of graphical presentation, the use of L-moments, the use of Anderson–Darling, Shapiro–Wilk, or Shapiro–Francia testing, and Royston's sample size monitoring.

Keywords: Graphical techniques; Skewness; Kurtosis; L-moments; Tests for normality; Royston's v' test

Contents

* Present address: 7 Basildon Crescent, Toronto, Ontario, Canada, M1M 3E1.
E-mail address: ahenders@uwo.ca.

0009-8981/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.cca.2005.11.007

1. Introduction

The usual processes in the statistical assessment of a data set are:

- screen the data for outliers or blunders;
- plot the data to detect asymmetry and tail weight;
- calculate the indices of sample shape (i.e., skewness and kurtosis);
- perform test(s) of Normality;
- if the data is Normal use parametric statistics for further analysis;
- if the data is non-Normal analyze the data by non-parametric statistics including the bootstrap or, if possible, utilize suitable transformations to obtain Normality followed by back-transformation after statistical analysis;
- if the data is transformed and tested for Normality, that test result becomes conservative. This effect was demonstrated by Linnet [1].

There are two circumstances in the practice of clinical chemistry when such testing is necessary during the statistical analysis of experimentally-derived data sets or the generation of population reference ranges [2–4]. In the first situation the experimenter is free to use a variety of data manipulations to statistically analyze the data. In the second situation there are several constraints—sample size, transformations, and tests for Normality (in the case of RefVal [5] the only permitted test of Normality is Anderson–Darling (see Empirical Distribution Function (EDF) statistics, below) as recommended by the IFCC and CSLI [2,3]). However, alternative strategies have been advocated by, for example, Wright and Royston [4].

This review addresses some of the more common procedures that may be used to assess the Normality of an experimental data set. The main sources used in this review, in addition to the primary literature, were the two monographs Goodness-of-fit techniques [6] and Thode's Testing for normality [7]. Thode considered forty tests for Normality in his extremely wide-ranging monograph—although there were some surprising omissions, which are discussed here, such as the treatment of L-moments and estimating departures from Normality—but he remarked that there were very many more such tests in the statistical literature.

Many comprehensive and extensive comparisons of tests of Normality have been reported that are discussed in Power comparisons of tests for Normality, below. Therefore, for illustrative purposes, I have used just four test samples (Fig. 1) derived from unpublished serum amylase quality assessment results and from a study of serum creatine kinase activities following myocardial infarction [8]. These four experimental data sets could be considered as representative of data types commonly experienced in the practice of clinical chemistry. The determination of population reference intervals will not be discussed as it is adequately addressed elsewhere [2–4].

In exploring Normality testing I used S-Plus, version 7.0 (Insightful Corporation) and R, version 2.1.1 [9] with their associated libraries including the car library [10] in S-Plus and the fBasics and nortest packages in R. The majority of the programs used in this review are available in these R packages; all others are listed later. It should be noted that R is freely available [9] and contains a vast resource of statistical libraries.

2. Graphical tools

Tukey's much quoted comment [11]—there is no excuse for failing to plot and look—is a useful starting point for assessing the Normality of data. Pearson and Please [12] provide an extensive diagrammatic review of

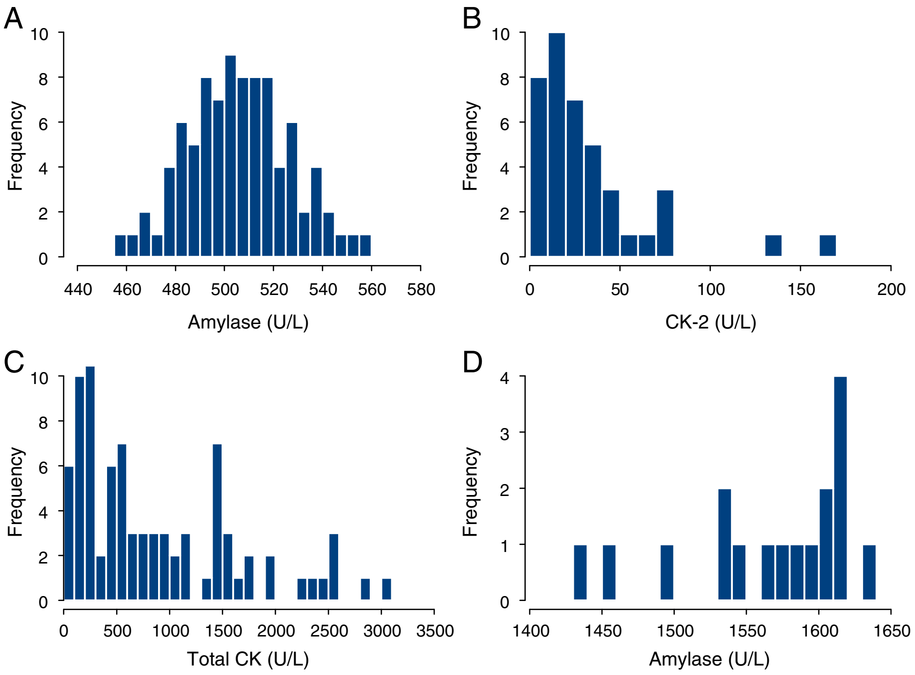

Fig. 1. Sample 1 (A) is essentially Normal (n=89, mean=506, SD=21, skewness=0.12, and kurtosis=2.7), sample 2 (B) has a positive kurtosis (n=40, skewness=2.2, and kurtosis=8.2), sample 3 (C) has a positive skew (n=83, skewness=1.0, and kurtosis=3.1), and sample 4 (D) has a negative skewness (n=17, skewness=−0.9, and kurtosis=2.8). Skewness and kurtosis were calculated using Eqs. (3) and (7), respectively.

population distributions. Essentially, graphical methods provide a qualitative assessment of a sample's Normality.

The most frequently used plot is the *histogram* (Fig. 1); it gives an indication of the symmetry and spread of the sample. A Normal distribution like Fig. 1A (this assumption of Normality will be tested later) can be compared to the other three figures that are clearly non-Normal. Fig. 1B (with positive kurtosis) and Fig. 1C are positively skewed while Fig. 1D has a negative skew. Note also that Fig. 1B contains two obvious outliers. An alternative to the histogram is the simpler stem and leaf plot [11] that provides the same information. The *box-and-whisker plot* (Fig. 2) provides more information than the histogram. Skewness is more easily detected by this type of plot. Outliers (defined as 1.5-fold the IQR) are shown as symbols (Fig. 2B).

The *quantile–quantile (Q–Q) plot* (Fig. 3) compares the ordered distribution of a test sample with the quantiles of a standard Normal distribution indicated by the straight line (other distributions may also be used). If the sample is Normally distributed the points will lie along this line (Fig.

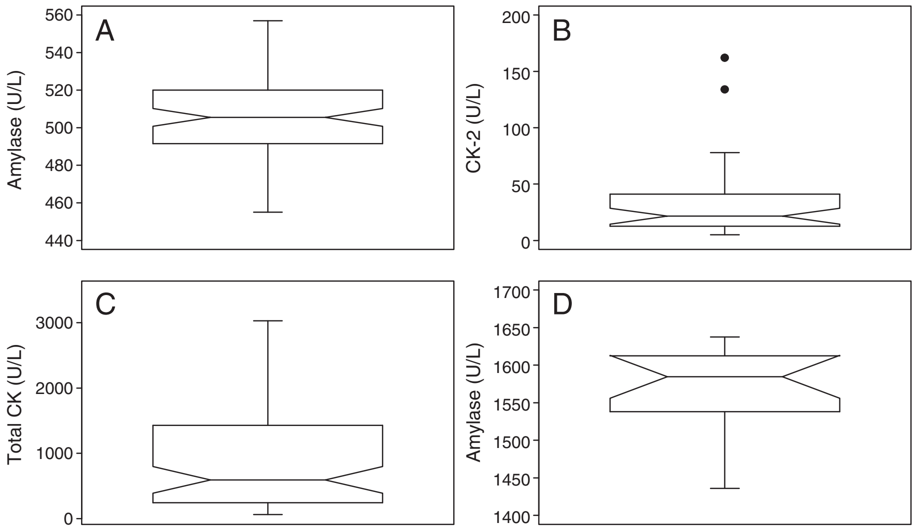

Fig. 2. Boxplots of the data displayed in Fig. 1. The boxes show the lower quartile (25th percentile), the median, and the upper quartile (75th percentile). The difference between the upper and lower quartiles is the inter-quartile range (IQR) and it contains 50% of the sample. The smallest and largest values are indicated by the small horizontal bars at the end of the whiskers (outliers expected).

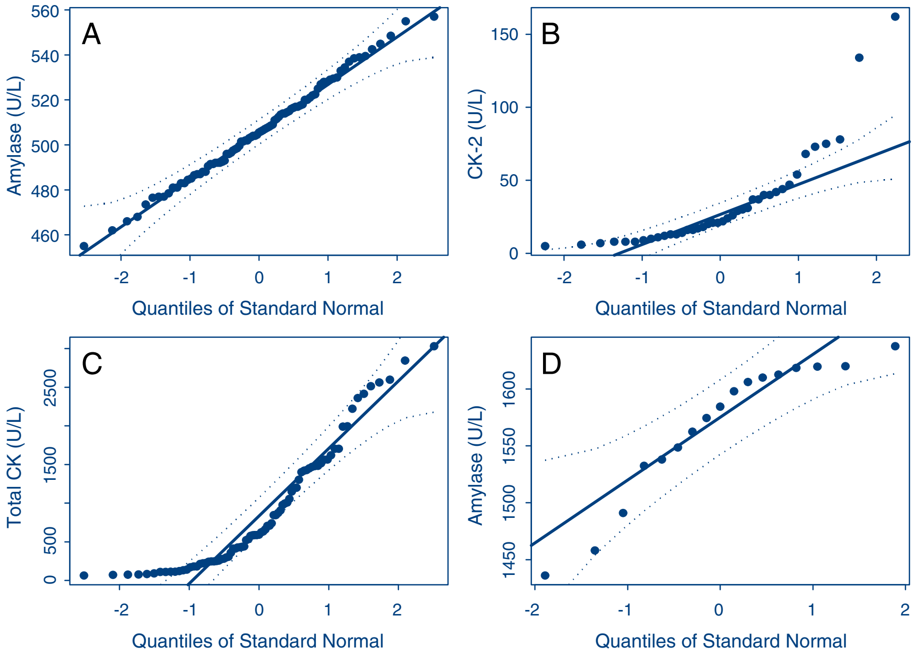

Fig. 3. Quantile–quantile (Q–Q) plots of the data displayed in Fig. 1. The 95% confidence intervals are indicated by the dotted lines.

3A) and would usually lie within the 95% confidence interval. Departures from Normality result in S-curves (Fig. 3D) or banana shapes (Fig. 3B and C) and points lying outside the respective 95% confidence intervals. Note the appearance of the outliers in Fig. 3B. The Q–Q plot is an excellent graphical test of the Normality of a sample and is commonly used for that purpose. The *probability–probability (P–P) plot* (Fig. 4) compares the probability distribution of a test sample to that of a standard Normal probability distribution indicated by the straight line (other distributions may also be used). Its interpretation is similar to that of the Q–Q plot; indeed note the similarity of the plot patterns except for detecting abnormal deviations in Fig. 4D. Again, note the value of providing the 95% confidence intervals.

Finally, the *cumulative distribution plot* (Fig. 5) displays the cumulative probability of a test sample. Typically, a Normal distribution has an S-shaped form as shown in Fig. 5A. This type of plot is helpful as regards probabilities in the tails of the sample distribution such as the deviations seen in Fig. 5B and D. Note that the theoretical curves in Fig. 5B, C, and D deviate markedly from that of Fig. 5A.

In summary therefore, the histogram or box-and-whisker plots are most useful in assessing the distribution of a sample. Both Q–Q and P–P plots with appropriate confidence intervals provide useful, semi-quantitative, evidence of departures from Normality. Of the five graph types the cumulative distribution plot is probably the least useful.

3. Tests using moments

The graphical examination of sample distributions, although essential, does not permit quantitative assessment of deviations from Normality. One such approach utilises the calculation of various moments—(3.1) Central moments (C-moments), (3.2) Linear moments (L-moments), and (3.3) Absolute moments.

3.1. Central moments or C-moments (moments about the mean)

The kth sample moment (m) about the sample mean [7] is (Eq. (1)):

$$m_k = \sum_{i=1}^{n} (x_i - \bar{x})^k / n \tag{1}$$

where x_i are the n observations, $\bar{x}$ the sample mean, and $k \geq 2$. This expression is rendered dimensionless [7] by division by m_2 to give the standardized moment test (Eq. (2)):

$$g_k = \frac{m_k}{m_2^{k/2}}. \tag{2}$$

Note that m_2 is similar to the sample variance except that the denominator for the latter is $(n-1)$: thus m_2 is a *biased* estimator.[1]

3.1.1. Skewness

The third moment test [13], the coefficient of skewness (Eq. (3)):

$$g_1 = \sqrt{\beta_1} = \frac{m_3}{m_2^{3/2}}. \tag{3}$$

Asymptotically g_1 of a Normal distribution has a mean of 0 and variance of $6/n$. For finite samples (note here that

[1] When the mean value of an estimator is equal to the value of the population parameter the estimator is described as unbiased.

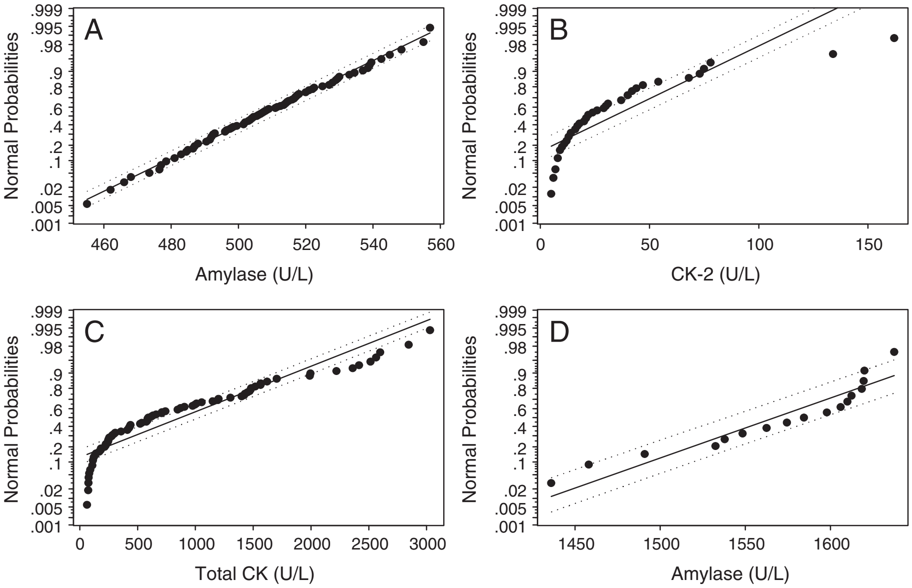

Fig. 4. Probability–probability (P–P) plots of the data displayed in Fig. 1. The 95% confidence intervals are indicated by the dotted lines.

$\sqrt{\beta_1}$ refers to the population while $\sqrt{b_1}$ refers to a sample) the variance [13] (Eq. (4)):

$$\operatorname{var}\left(\sqrt{b_1}\right) = \frac{6(n-2)}{(n+1)(n+3)}. \tag{4}$$

Values of skewness are symmetrically distributed about zero. Negative values indicate a skew to the left (left-tail or negative skewness) while positive values indicate skewing to the right (right-tail or positive skewness).

There are *three* definitions of the third central moment. Equation (3) is the most commonly encountered (i) *biased* definition but there is also an (ii) *unbiased* form [14] used, for example, in the CBStat program (Eq. (5)):

$$\gamma_1 = \frac{m_3}{\mathrm{SD}^3}. \tag{5}$$

Finally there is the (iii) *unbiased* Fisher definition [13] derived from cumulant theory [14,15] (Eq. (6)):

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1. \tag{6}$$

This is available in SAS, SPSS, MINITAB (version 14), S-Plus (the default, although $\sqrt{b_1}$ is available as an option), Excel, Quattro Pro, Analyse-it, and MedCalc.

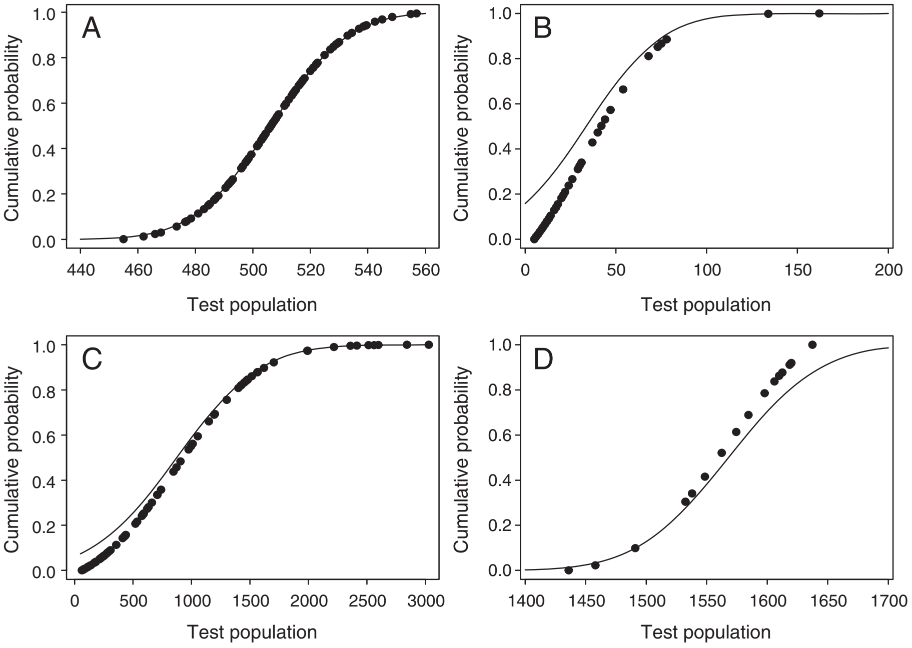

Fig. 5. Cumulative probability plots of the data displayed in Fig. 1. The lines are derived from the samples' mean and SD values using the standard cumulative probability calculation.

Table 1A
C-moments indices of skewness and kurtosis for the test samples

Sample	Indices of skewness			Indices of kurtosis			
	g_1 (Eq. (3))	γ_1 (Eq. (5))	G_1 (Eq. (6))	β_2 (Eq. (7))	g_2 (Eq. (9))	γ_2 (Eq. (10))	G_2 (Eqs. (11) and (12))
Normal	0.12	0.12	0.12	2.7	−0.30	−0.36	−0.25
Positive kurtosis	2.22	2.14	2.31	8.23	5.24	4.83	6.12
Positive skewness	1.01	0.99	1.03	3.13	0.13	0.05	0.21
Negative skewness	−0.93	−0.85	−1.02	2.79	−0.21	−0.53	0.17

3.1.2. Kurtosis

The fourth moment test [13], the coefficient of kurtosis (Eq. (7)):

$$\beta_2 = \frac{m_4}{m_2^2}. \tag{7}$$

Asymptotically β_2 for a Normal distribution has a mean of 3 and a variance of $24/n$. For finite samples (note here that β_2 refers to the population while b_2 refers to a sample) the mean and variance [13] are (Eq. (8)):

$$\mathrm{mean}(b_2) = \frac{3(n-1)}{n+1} \text{ and } \mathrm{var}(b_2) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}. \tag{8}$$

Values of kurtosis range from 0 to ∞. A value <3 indicates a flat-topped distribution (*platykurtic* which has heavy shoulders), values >3 indicate a "pointy" distribution (*leptokurtic* having no shoulders), and a value of 3, as obtained from a Normal distribution, is described as *mesokurtic*.

There are *four* definitions of the fourth central moment. The (i) *biased* coefficient of kurtosis (Eq. (7)) is the definition that occurs frequently in the statistical literature. The second (ii) *biased* definition [13] (Eq. (9)):

$$g_2 = \beta_2 - 3 \tag{9}$$

while its (iii) *unbiased* equivalent [14], used by the CBStat program, is centered on zero (Eq. (10)):

$$\gamma_2 = \frac{m_4}{\mathrm{SD}^4} - 3. \tag{10}$$

Thus, Eqs. (9) and (10) define kurtosis as zero for a Normal distribution. This usage is confusing unless the reader is advised that the reported kurtosis index is not centered on the value 3.

The fourth (iv) *unbiased* definition, Fisher's G_2, derived from cumulant theory [14,15] is the form most commonly used in statistical programs (Eq. (11)) as noted above for G_1:

$$G_2 = \frac{n-1}{(n-2)(n-3)}\{(n+1)g_2 + 6\}, \tag{11}$$

or in the alternative format incorporating b_2 (Eq. (12)):

$$G_2 = \frac{(n+1)(n-1)}{(n-2)(n-3)}\left\{b_2 - \frac{3(n-1)}{n+1}\right\}. \tag{12}$$

Joanes and Gill [16] drew attention to a problem seen with small or moderate sample sizes when using these various kurtosis definitions. They used a Normally-distributed sample ($n=20$) with an Anderson–Darling (see Empirical Distribution Function statistics, below) test result of $A=0.376$ (p-value$=0.38$). The following kurtosis results were obtained: $b_2=2.92$ (Eq. (7)), $g_2=-0.07$ (Eq. (9)), $\gamma_2=-0.36$ ((10)), and $G_2=0.27$ (Eq. (11)). The indices b_2 and g_2 are obviously equivalent but the unbiased estimates disagree. Thus three indices indicate that kurtosis is negative but G_2 is positive. This is confusing. These authors do point out, however, that the differences between the various kurtosis definitions are unimportant with respect to large samples.

Joanes and Gill [16] also compared measures of sample skewness and kurtosis by examining their mean-squared errors (MSE) [17] of small to moderate sized Normal and simulated non-Normal samples (Eq. (13)):

$$\mathrm{MSE} = \mathrm{var(sample)} + (\mathrm{bias(sample)})^2. \tag{13}$$

When the estimator is unbiased the MSE is equal to the variance of the estimator and it should be noted that values of a biased variance will always be smaller than the values of unbiased variances thus off-setting to an extent the effect of the bias.

Thus, in the case of the *unbiased* skewness indices γ_1 and G_1 of a Normal sample the following relationship occurs:

$$\mathrm{MSE}(\gamma_1) < \mathrm{MSE}(G_1).$$

Whereas with the kurtosis indices only G_2 is unbiased but it has the largest variance thus:

$$\mathrm{MSE}(g_2) < \mathrm{MSE}(G_2)$$

Non-Normal samples were simulated from chi-squared distributions. For both skewness and kurtosis Joanes and Gill [16] found that the MSE's were smallest for both G_1 and G_2. Thus the choice of the appropriate skewness and kurtosis indices to use (i.e., possessing the smallest MSE values) depend on the type of sample distribution.

The skewness and kurtosis indices of the four test samples are listed in Table 1A. Note the directional agreements between the various indices with the exception of the G_2 value for the negative skewed sample. This is

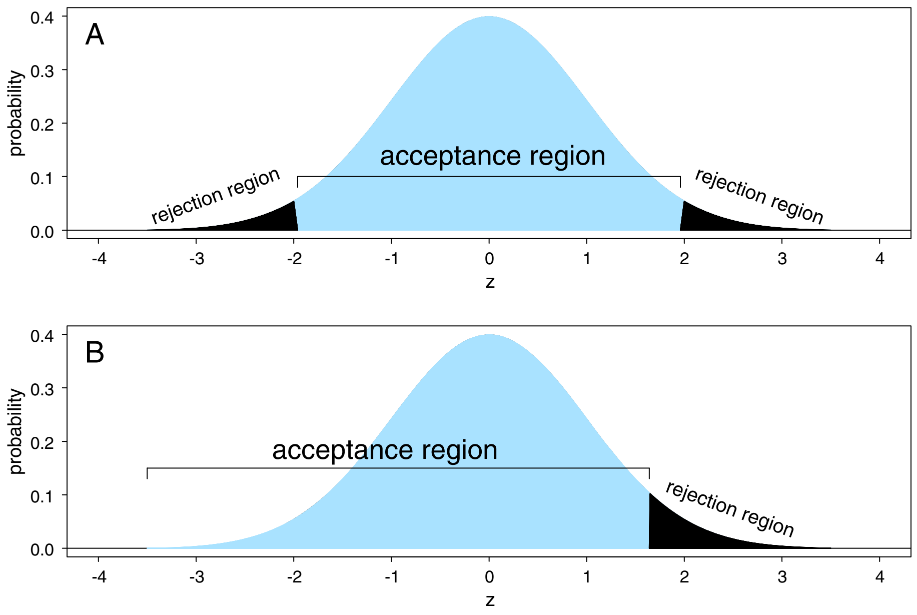

Fig. 6. Two types of regions considered in hypothesis testing. A. Two-tail test: the reject (the Null hypothesis) zone is shaded black. B. One-tail test: in this case the reject zone is shaded black with $z \geq 1.64$ where a directional test is used.

similar to the previously noted effect observed by Joanes and Gill [16].

3.1.3. Probability estimates

In order to determine if a sample skew and kurtosis are significantly different from a Normal distribution it is necessary to determine the standard score or deviate (z-statistic) of the data (Eq. (14)):

$$z = \frac{\text{observed value} - \text{ expected value}}{\text{SD of observed value}}. \qquad (14)$$

This approach, however, depends on the values of both sample skewness and kurtosis being Normally distributed. However, the sample size must be appreciably large (for example, even when the sample size is 5000 a Normal approximation of b_2 may be inappropriate [7]) before both indices are asymptotically Normal.

However, exact expressions of the moments of the sampling distributions of skewness and kurtosis are known in the case of a Normal distribution thus permitting the construction of tables of critical values. For example, D'Agostino [18] lists an extensive set of tables as regards both sample values of n and percentage points.

Table 1B
Probability values for the skewness and kurtosis indices for the test samples

Sample	Skewness		Kurtosis	
	z-value	p-value	z-value	p-value
Normal	0.48	0.36	− 0.38	0.37
Positive kurtosis	4.59	<0.0001	3.56	0.0007
Positive skewness	3.5	0.0009	0.6	0.33
Negative skewness	− 1.85	0.07	0.39	0.37

Alternatively, the development of a transformation to Normality of the null distribution of skewness by D'Agostino [19], for $n \geq 8$, and to a Normal approximation of the null distribution of kurtosis by Anscombe and Glynn [20], for $n \geq 20$, avoids the use of tables. These transformations produce z-statistics (Fig. 6) that are converted to probabilities by reference to a table of the standard Normal distribution (i.e., $N(0,1)$) or by calculation [21] (Eq. (15)):

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}. \qquad (15)$$

The associated probabilities of these z-statistics can be used to test the two-sided composite Null hypothesis[2] illustrated in Fig. 6A (H_0: the tested samples have skewness and kurtosis not significantly different from a Normal population (i.e., p ($|z| < 1.96$) at the 5% level of significance. The alternative hypothesis (H_1) is that the sample belongs to any non-Normal distribution). As will become evident later, the use of a one-sided (in this case upper-tailed) or directional test (such as $\sqrt{\beta_1}$ and β_2) can also be used to test a one-sided composite Null hypothesis (Fig. 6B).

The results of these tests are listed in Table 1B. Where the p-values are >0.05 the Null hypothesis (H_0) is accepted so that the index is compatible with Normality. Thus the sample with negative skewness is compatible with Normality despite the indices of skewness (Table 1A)

[2] A composite hypothesis refers to the situation where one or more of the parameters of the distribution are unspecified whereas a simple hypothesis refers to the situation where all parameters of the distribution are specified.

being ≈ -1. This result is consistent with the appearance of Figs. 3D and 4D.

Are there any advantages to the use of these indices? Apart from their being deeply embedded in classical statistical thought, are very easy to calculate, and are universally available in statistical packages there are probably none.

Are there disadvantages to the uses of $\sqrt{\beta_1}$, β_2, $\sqrt{b_1}$, and b_2? There are several. Neither $\sqrt{b_1}$ nor b$_2$ are Normally distributed until the sample sizes are quite large [7] so that sample transformations are necessary before Normality of a sample can be ascertained using these indices [22]. Secondly, these conventional indices are bounded and cannot attain the full range of values available to the population indices. For example, sample skewness is bounded [23,24] by (Eq. (16)):

$$\left|\sqrt{b_1}\right| \leq \frac{(n-2)}{\sqrt{(n-1)}} \qquad (16)$$

so that the true value cannot be achieved in finite samples. Sample kurtosis is not limited as it is bounded by $\leq n$ [25].

Other disadvantages are listed in Table 2 and have been derived from Balanda and MacGillivray [26], Hosking [27], Royston [29], and Hosking and Wallis [30].

3.2. L-moments

In 1990 Hosking [27] published a review advocating a unified approach to the use of order statistics for the statistical analyses of univariate probability distributions. He suggested that as L-moments are linear functions of the data (hence the use of "L") they are superior to conventional moments (C-moments) because they suffer less from the effects of sampling variability, are more robust to the effects of outliers, and are more reliable for making inferences from small samples.

Table 2
The disadvantages associated with the use of C-moments

- Neither $\sqrt{\beta_1}$ nor β_2 nor have an intuitively clear meaning as a feature of distributions
- Both $\sqrt{\beta_1}$ and β_2 are very sensitive to small changes in the tails of the distribution
- Both $\sqrt{b_1}$ and b_2 are unreliable for small sample values and may show marked biases
- Identification of the parent distribution by the joint C-moments of the observed random sample is often not useful
- The divisor, m_2, gives relatively more weight to the largest differences due to the squared power
- Both $\sqrt{b_1}$ and b_2 are susceptible to even moderate outliers as they involve the cube and fourth powers, respectively, of extreme deviations
- $\sqrt{b_1}$ can take arbitrarily large positive or negative values as some heavy-tailed distributions can have values approaching infinity
- $\sqrt{b_1}$ is sensitive to extreme tails and is difficult to estimate when the distribution is markedly skewed
- β_2 and b_2 have no unique interpretation and the standard concepts of peakedness of a distribution or as tail weight only apply to closely defined families of symmetric unimodal distributions

Hosking [27], Royston [29], and Hosking and Wallis [30] discuss, in detail, the definitions and basic properties of the *population* L-moments so it is unnecessary to mention them here except to comment on the principle and mention the notation for the population L-moments (location, λ_1; scale, λ_2; CV, τ; skewness, τ_3; and kurtosis, τ_4). L-moments may intuitively be understood [31] as follows—one value in a sample gives a notion of the magnitude of the random variable while the difference between two values gives a sense of how varied the random variable is. When there are three values in a sample they give an indication of how asymmetric the distribution is (i.e., a measure of skewness. When there are four values in the sample they give a notion as to the ratio of the peak to the tails of the distribution (a measure of kurtosis). When many such values are considered the sample's L-moments can be calculated.

Hosking [27] derived L-moments indirectly using probability weighted moments [28] and these have to be introduced when the method of L-moments is described. Thus how L-moments are *estimated* appears to be unrelated to how L-moments are *defined* [31]. Accordingly, Wang [31] derived *direct* estimators of L-moments from their definitions thus eliminating the need for probability weighted moments.

3.2.1. Sample L-moments

The details of the calculation of *sample* L-moments are now described. These are derived from chapter two of the Hosking and Wallis monograph [30]. By analogy with the classical definitions of moments these are location (l_1 or mean), scale or dispersion (l_2), L-CV (t or coefficient of L-variation), L-skewness (t_3, a scale-free ratio measure of skewness), and L-kurtosis (t_4, likewise a scale-free ratio measure of kurtosis). The values l_2, t_3, and t_4 are *nearly* unbiased estimates unlike their moment equivalents. Eq. (17) defines the calculation of l_2:

$$l_2 = 2b_2 - l_1 \text{ where } b_2 = \frac{1}{n(n-1)} \sum_{j=2}^{n} (j-1)x_{j:n}. \qquad (17)$$

Eqs. (18) and (19) define the calculations for l_3 and l_4:

$$l_3 = 6b_3 - 6b_2 + l_1 \text{ where } b_3 = \frac{1}{n(n-1)(n-2)} \sum_{j=3}^{n} (j-1)(j-2)x_{j:n} \qquad (18)$$

$$l_4 = 20b_4 - 30b_3 + 12b_2 - l_1 \text{ where } b_4 = \frac{1}{n(n-1)(n-2)(n-3)} \sum_{j=4}^{n} (j-1)(j-2)(j-3)x_{j:n}. \qquad (19)$$

Eq. (20) displays the calculation for the coefficient of L-variation (L-CV) that ranges in value between 0 and 1:

$$\text{L-CV} = \frac{l_2}{l_1}. \tag{20}$$

The dimensionless measure of L-skewness (Eq. (21)):

$$t_3(\text{L-skewness}) = \frac{l_3}{l_2}. \tag{21}$$

The values of population and sample L-skewness are limited to lie within the interval $(-1, 1)$ for all distributions. As a Normal distribution has a t_3 value close to zero it follows that positive and negative skew distributions will have t_3 values, respectively, above and below zero. Clearly it is easier to interpret a value for t_3 than for $\sqrt{b_1}$ which may take arbitrarily large values. Hosking [32] provides some examples; it is evident that there is a poor relationship between these two indices of skewness.

Royston [29] suggests (Eq. (22)) the use of an alternative skewness index (t_3^*) that is identical to the "shape" index suggested by Efron and Tibshirani [33] as a measure of the asymmetry of a confidence interval.

$$t_3^* = \frac{(1 + t_3)}{(1 - t_3)}. \tag{22}$$

For symmetric, positively skewed, and negatively skewed distributions this alternative index of skewness takes values of 1, >1, and <1, respectively.

Eq. (23) displays the calculation for the dimensionless index of L-kurtosis:

$$t_4(\text{L} = \text{kurtosis}) = \frac{l_4}{l_2}. \tag{23}$$

The value of population and sample L-kurtosis is ≤1 for all distributions. Clearly it is easier to interpret a value for t_4 than for b_2 which may take arbitrarily large values. A Normal distribution will have a t_4 value close to 0.1226 and distributions with negative or positive kurtosis will have t_4 values respectively less than or greater than this value. Hosking [32] compares conventional kurtosis to L-kurtosis values for a range of symmetric and asymmetric distributions; again it is evident that there is a poor relationship between these two indices of kurtosis.

Table 3 lists the L-moments for the four test samples. The normal sample conforms closely to the expected values noted above. Royston's skewness index appears an easier index to review than the L-skewness index itself. When interpreting the L-kurtosis index it has to be recalled that a Normal distribution has a value close to 0.1126. Thus the positive kurtosis sample has a marked kurtosis while the other two samples have a smaller degree of kurtosis than a Normal distribution.

Table 3
L-moments for the test samples

Sample	L1	L2	L-CV	L-skewness	*L-skewness	L-kurtosis
Normal	506.2	12.2	0.0241	0.0287	1.0591	0.1185
Positive kurtosis	33.7	15.9	0.473	0.4249	2.4777	0.2483
Positive skewness	874	417.7	0.4779	0.2734	1.7524	0.0696
Negative skewness	1567.4	33.38	0.0213	−0.2899	0.5505	0.0925

A Monte Carlo simulation (10,000 replications) of the Normal sample (Fig. 1A) was used to compare the distributions of the C- and L-moments (Fig. 7). While the L-moments (Fig. 7B and D) appear Normally distributed (as does the C-moment skewness) it is evident that the C-moment kurtosis distribution has a marked positive skewness (Fig. 7D) even after 10,000 replications.

3.2.2. Probability estimates

Hosking has suggested the use of z-scores as an overall test statistic for the L-moments. Eq. (24) is used when examining a sample that might correspond to a Normal distribution (t_3=0 and t_4=0.1226):

$$\text{statistic} = \frac{(t_3)^2}{\text{var}(t_3)} + \frac{(t_4 - 0.1226)^2}{\text{var}(t_4)} \tag{24}$$

approximated by a χ^2-distribution with two degrees of freedom. This statistic may also be applied to the individual indices using a χ^2-distribution with one degree of freedom. It is necessary to perform a Monte Carlo simulation (10,000 replications) using a distribution appropriate to the test sample.[3] In the present instance the normal test sample can be simulated using a random Normal distribution. The statistic had a p-value of 0.99 as did the individual indices indicating that the test sample was indeed Normal. Hosking and Wallis [30] provide, in an extensive Appendix, the L-moments for many specific distributions thus allowing appropriate values of t_3 and t_4 to be incorporated into Eq. (22).

3.2.3. Advantages of L-moments

Hosking [27] states, for most distributions, that the asymptotic biases are negligible for sample sizes of 20 or more. For example, the Normal distribution (τ_3=0 and τ_4=0.1226) has a t_4 asymptotic bias of 0.03 n^{-1}. However, outliers may have an undue influence on L-moments leading to the development of trimmed L-moments [34] that are much more robust to outliers. When a sample's L-skewness is plotted against L-kurtosis (the L-moment diagram) it is possible to discriminate

[3] Hosking [27] provides an approximation for var(t_3)=0.1866n^{-1}+ 0.8n^{-2}. However there is presently no approximation available for var(t_4) thus it requires calculation from large-sample simulation. Note that the value for var(t_3) obtained by simulation gives exactly the same result as the formal equation.

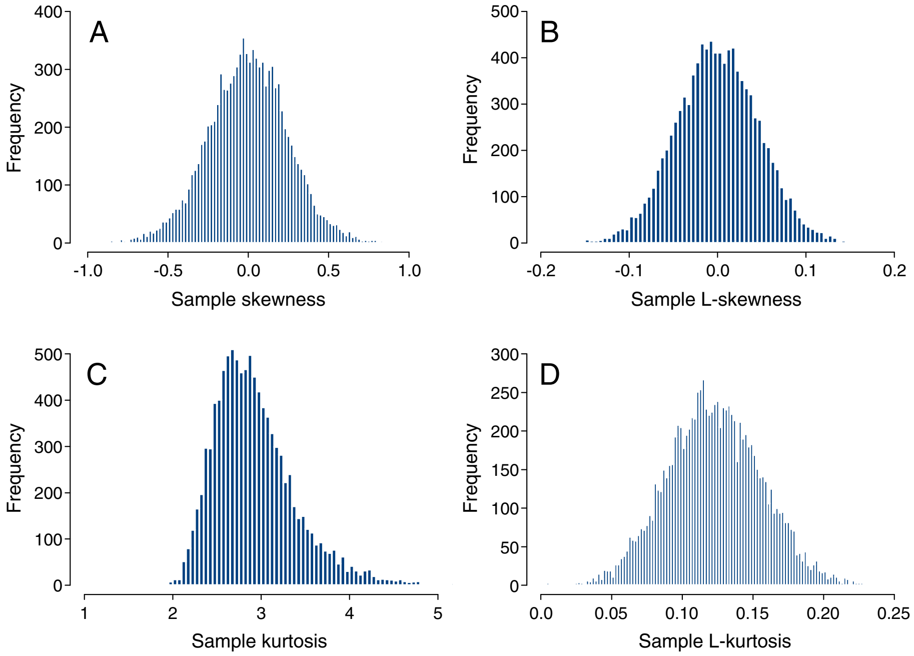

Fig. 7. 10,000 simulations of the data shown in Fig. 1A were used to calculate the replicates of the C-moments (plots A and C) and L-moments (plots B and D) of skewness and kurtosis.

among alternate distributional hypothesis [27,35]. Hosking [27] also suggests the use of L-skewness as a test for Normality against skew alternatives. Advantages of L-moments, listed in Table 4, have been derived from Hosking [27], Royston [29], and Hosking and Wallis [30].

Royston [29] has suggested that the single disadvantage to using L-moments in place of conventional moments is the complexity of the calculations. However, Hosking has made available an R/S-Plus implementation adapted from his LMOMENTS Fortran package [36]. This program can be obtained directly [37] or by searching the R site [9] with the search term “samlmu”. That program only provides l_1, l_2, t_3, and t_4 but can be readily modified to include the indices L-CV and Royston’s t_3*. Wang [31] also provides a Fortran program based on the direct estimation of the L-moments. Wang’s program is available in an R/S-Plus version that provides all the previously-noted L-moment indices.

Table 4
The advantages of using L-moments

- L-moment indices are easy to interpret
- Sample L-moments are not bounded — they can take any values found in the corresponding population
- L-moments of an observed random sample can identify the parent distribution (see text)
- Asymptotic approximations to sampling distributions are superior to C-moments
- Robust to outliers (but see text)
- Approximately normally distributed so that transformations are unnecessary unlike C-moments
- Unbiased even for small samples
- Can indicate the type of departure from normality

3.3. Population absolute moments

The quantity (Eq. (25)):

$$v_c = \int_{-\infty}^{\infty} |x-\mu|^c dF \tag{25}$$

is called the absolute moment of order c about μ [14]. The cth sample absolute moment is (Eq. (26)):

$$v_c = \sum_{i=1}^{n} |x_i - \bar{x}|^c / n. \tag{26}$$

It is thus possible [7] to define absolute test statistics (Eq. (27)):

$$a(c) = v_c / v_2^{c/2} = v_c / m_2^{c/2}, \quad c \neq 0 \text{ or } 2. \tag{27}$$

A test based on absolute moments, where $c=1$, was proposed by Geary [38] in 1935.

4. Geary’s test

Geary’s test [38,39] is the ratio of the mean deviation to the unbiased standard deviation that can be used as a test for Normality (Eq. (28)):

$$a = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} |x_i - \bar{x}| \Bigg/ \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{28}$$

The asymptotic mean is $\sqrt{\frac{2}{\pi}} = 0.797885$ and SD 0.2123/$\sqrt{n}$. Geary [38] provided a table of upper and lower 1%

and 5% probability points for sample sizes from 6 in intervals to 1000. In a later publication [39] he provided two graphs covering values of n from 11 to 1000 and a table in which upper and lower 10%, 5%, and 1% critical values were listed.

D'Agostino and Rosman [40] pointed out that Geary's test had not been subjected to comparative studies like many other tests of normality (see Power comparisons of tests for Normality, below) and they performed a comparison between the Shapiro–Wilk, Shapiro–Francia, D'Agostino's D-test, and Geary's one- and two-sided tests.

The use of the Geary graphs or tables described above does not provide a p-value. Accordingly, D'Agostino [41] described a transformation (Eq. (29)) of "a" to standard Normality thus providing a p-value for the test:

$$z = \frac{\sqrt{n}(a - 0.7979)}{0.2123}. \tag{29}$$

The probability is then calculated from Eq. (15). Eq. (29) is less reliable for the range $11 \leq n \leq 31$ but is better for $n \geq 41$ as it then approximates to the critical values provided by Geary's probability table [39].

The four test samples were assessed using the Geary test and the results tabulated (Table 5). It is evident that this test detects a Normal sample distribution and sample kurtosis but fails to identify obvious positive or negative skewness.

5. Chi-square (χ^2) goodness-of-fit test

The chi-square test can be applied to discrete or continuous, univariate or multivariate data. It is the oldest goodness-of-fit test and was described by Karl Pearson [71]. The test compares observed and expected (i.e., the hypothesised distribution) frequencies for individual categories, where m is the number of cells or bins, thus (Eq. (30)):

$$X^2 = \sum_{i=1}^{m} \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}. \tag{30}$$

Note that the observed data is collected in a series of bins or cells. Because of the relationship [14,42] between the χ^2 distribution and the sum of squares of the individual z-scores (as defined by Eq. (14)) where "expected value" is the sample mean) of the individual cells, the following alternative relationship can be used (Eq. (31)):

$$X^2 = \sum_{i=1}^{m} (z\text{-score}_i)^2. \tag{31}$$

This formulation was used by Hosking (see Eq. (24)) for assessing the probability value of the L-moments. As the chi-square test partitions the data into cells it looses available information. Indeed, Moore [43] recommends that the test *not be used* on univariate distributions in favour of special purpose tests of fit or tests based on empirical distribution function (EDF) statistics.

6. Empirical distribution function statistics

Empirical distribution function statistics (EDF) refer to "a step function, calculated from the sample, which

Table 5
Results of Geary, EDF, and Regression tests for univariate Normality

Test	Sample type→	Normal	Positive kurtosis	Positive skewness	Negative skewness
Absolute moment test					
Geary	a	0.8037	0.6976	0.8258	0.8237
	p-value	0.386	0.005	0.194	0.352
Empirical distribution function tests					
Kolmogorov–Smirnov	D	0.0452	0.1954	0.1533	0.1657
(Lilliefors modification)	p-value	0.9252	0.0005	0.0001	0.2463
Cramér–von Mises	W^2	0.0179	0.5104	0.5363	0.1144
	p-value	0.9829	<0.01	<0.01	0.0644
Anderson–Darling	A^2	0.123	3.0069	3.2977	0.728
	p-value	0.9868	<0.01	<0.01	0.0465
Regression/correlation tests					
Gan–Koehler	k_o^2	0.9978	0.8536	0.9245	0.9236
	p-value	>0.10	<0.01	<0.01	>0.05
Shapiro–Wilk	W	0.9948	0.7422	0.8758	0.8859
	p-value	0.9811	<0.01	<0.01	0.0397
Shapiro–Francia	W'	0.9967	0.7337	0.8807	0.8927
	p-value	0.9943	<0.01	<0.01	0.0502
Filliben	r	0.9984	0.8556	0.9388	0.9451
	p-value	<0.01	<0.01	<0.01	<0.01

estimates the population distribution function. EDF statistics are measures of the discrepancy between the EDF and a given distribution function." (Stephens [44]). There are two classes of EDF statistics—the Kolmogorov–Smirnov type and the Quadratic type.

6.1. Kolmogorov–Smirnov type

6.1.1. Kolmogorov–Smirnov test

This test is based on the maximum vertical difference between the EDF and the Normal cumulative distribution curve (when the null hypothesis is that the EDF demonstrates normality). The sample values are ordered, then (Eq. (32)):

$$\left.\begin{array}{l} D^{+} = \max_{i=1,\ldots,n}\left[i/n - p_{(i)}\right] \\ D^{-} = \max_{i=1,\ldots,n}\left[p_{(i)} - (i-1)/n\right] \\ D = \max\left[D^{+}, D^{-}\right] \end{array}\right\} \quad (32)$$

D^{+} and D^{-} are, respectively, the largest vertical distances above and below the Normal cumulative distribution curve. The Kolmogorov–Smirnov statistic is D. The null hypothesis (Normality) is rejected when the value of D does not exceed the particular chosen critical value, i.e., an upper-tailed test. These critical values are provided in tables but Stephens [46] has provided a modification for all sample sizes for a range of critical values for a range of significance points that obviates the need to consult tables (Eq. (33)):

$$D^{*} = D\left(\sqrt{n} - 0.01 + 0.85/\sqrt{n}\right) \quad (33)$$

The critical values of D for the significance points 0.1, 0.05, and 0.01 are 0.819, 0.895, and 1.035, respectively. The Kolmogorov–Smirnov test is "intended for use only when the hypothesized distribution function is completely specified, that is, when there are no unknown parameters that must be estimated from the sample. Otherwise the test becomes conservative" [45]. In accord with Stephens' remark, Massey [47] suggests that the critical value of D is reduced if the parameters of the distribution must be estimated from the sample as the probability of a type 1 error (α) will be smaller than that given in the tables of D. Massey further notes that the conventionally used Kolmogorov–Smirnov test has superior power to the chi-square test.

6.1.2. Kolmogorov–Smirnov (Lilliefors modification) test

As a consequence of the foregoing, Lilliefors [48] revised the Kolmogorov–Smirnov tables for samples with mean and variance unknown using a Monte Carlo procedure (1000 or more samples) for each value of n. The table covers the levels of significance (0.2, 0.15, 0.1, 0.05, and 0.01) for sample sizes from 4 to 30. For $n=31$ onwards the provided values of D have to be divided by $\sqrt{n}$. The Lilliefors table has subsequently been revised by Mason and Bell [49] who used a more extensive Monte Carlo simulation. For comparison, the calculation of the critical value of D is (Eq. (34)):

$$D = \frac{\text{critical value for } n \geq 31}{\left(\sqrt{n} - 0.01 + 0.83/\sqrt{n}\right)}. \quad (34)$$

Dallal and Wilkinson [50] have also provided an extensive corrected table of critical values and Iman [51] has prepared graphs for tests at the 0.1, 0.05, and 0.01 levels of significance.

The Lilliefors testing procedure is identical to that for the Kolmogorov–Smirnov test (Eq. (30)), but must have >4 test samples. However, the z-statistic is calculated as part of the procedure (Eq. (35)):

$$z_{(i)} = \left(x_{(i)} - \hat{\mu}\right)/\hat{\sigma}. \quad (35)$$

Lilliefors demonstrated the superior power of his modification compared to the Kolmogorov–Smirnov test [48].

6.2. Quadratic type

6.2.1. Cramér–von Mises test

This test was developed due to the contributions of Cramér [72], von Mises [73], and Smirnov [74]. The test samples (>7), mean and variance unknown, are ordered and the statistic, W^2, calculated as follows (Eqs. (36) and (37)):

$$p_i = \Phi((x_i - \bar{x})/\sigma) \quad (36)$$

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[p_{(i)} - \frac{(2i-1)}{2n}\right]^2. \quad (37)$$

Here Eq. (36) is the cumulative distribution function of the standard Normal distribution. Stephens [46] has provided a modification for all sample sizes for a range of critical values for a range of significance points (Eq. (38)):

$$W^{2^*} = W^2\left(1.0 + \frac{0.5}{n}\right) \quad (38)$$

The calculation of the respective p-values is as follows:

when $W^{2^*} < 0.0275$

$$p\text{-value} = 1 - \exp\left(-13.953 + 775.5 \times W^{2^*} - 12542.61 \times \left(W^{2^*}\right)^2\right)$$

when $W^{2^*} < 0.0051$

$$p\text{-value} = 1 - \exp\left(-5.903 + 179.546 \times W^{2^*} - 1515.29 \times \left(W^{2^*}\right)^2\right)$$

when $W^{2^*} < 0.092$

$$p\text{-value} = \exp\left(0.886 - 31.62 \times W^{2^*} + 10.897 \times \left(W^{2^*}\right)^2\right)$$

otherwise

$$p\text{-value} = \exp\left(1.111 - 34.242 \times W^{2^*} + 12.832 \times \left(W^{2^*}\right)^2\right).$$

6.2.2. Anderson–Darling test

The Anderson–Darling goodness-of-fit test was introduced in 1952 [52,53]. The test samples (>7), mean and variance unknown, are ordered and the statistic A^2 calculated (Eqs. (36) and (39)):

$$A^2 = -n - n^{-1} \sum_{i=1}^{n} [2i-1][\ln(p_i) + \ln(1 - p_{(n-i+1)})]. \tag{39}$$

This test differs from the Cramér–von Mises test in the type of weighting function incorporated in its formulation. This type of weighting provides more influence to the tails of the distribution than does the Cramér–von Mises test.

Stephens [44] described a modification to obtain critical values for all sample values (Eq. (40)):

$$A^{2^*} = A^2\left(1.0 + 0.75/n + 2.25/n^2\right) \tag{40}$$

The calculation of the respective p-values is as follows:

when $A^{2^*} < 0.2$

$$p\text{-value} = 1 - \exp\left(-13.436 + 101.14 \times A^{2^*} - 223.73 \times \left(A^{2^*}\right)^2\right)$$

when $A^{2^*} < 0.34$

$$p\text{-value} = 1 - \exp\left(-8.318 + 42.796 \times A^{2^*} - 59.938 \times \left(A^{2^*}\right)^2\right)$$

when $A^{2^*} < 0.6$

$$p\text{-value} = \exp\left(0.9177 - 4.279 \times A^{2^*} - 1.38 \times \left(A^{2^*}\right)^2\right)$$

otherwise

$$p\text{-value} = \exp\left(1.2937 - 5.709 \times A^{2^*} + 0.0186 \times \left(A^{2^*}\right)^2\right).$$

6.3. Summary of test results

Results of the three EDF tests (Lilliefors, Cramér–von Mises, and Anderson–Darling) on the four test samples are listed in Table 5. All three tests correctly identify the normal, positive kurtosis, and positive skewed samples. However, only the Anderson–Darling test correctly identifies the negatively skewed sample.

7. Regression/Correlation tests [54]

7.1. Gan–Koehler tests

The Gan–Koehler tests [55] are two goodness-of-fit statistics based on measures of linearity for standardized P–P plots. Their second test statistic is (Eqs. (36), (41), and (42)):

$$p_i = i/(n+1) \tag{41}$$

$$k_0^2 = \frac{\left[\sum_{i=1}^{n} (z_i - 0.5)(p_i - 0.5)\right]^2}{\left[\sum_{i=1}^{n} (z_i - 0.5)^2 \sum_{i=1}^{n} (p_i - 0.5)^2\right]}. \tag{42}$$

Note that z_i is defined by Eq. (35) and k_0^2 is a modified squared correlation coefficient. Critical values for k_0^2 are calculated for the lower pth percentiles from Eq. (43):

$$\text{critical } k_p^2 = 1 - \left(\alpha_p + n\beta_p\right)^{-1}. \tag{43}$$

The terms, α and β, are provided, in a table for several percentiles, by Gan and Koehler.

7.2. Shapiro–Wilk test

The Shapiro–Wilk test was introduced in 1965 [56]. Essentially the test statistic W is the square of the Pearson correlation coefficient computed between the order statistics of the sample and scores that represent what the order statistics should look like if the population were Gaussian. Thus, if the value of W is close to 1.0 the sample behaves like a Normal sample whereas if W is below 1.0 the sample is non-Gaussian.

The original formulation of the W-test [56] was limited to sample sizes of n=3(1)50 tabulated for percentage points of the null distribution for p-values of 0.01, 0.02, 0.05, 0.1, 0.5, 0.9, 0.95, 0.98, and 0.99. Calculation, and interpretation, of the W-test required the use of tables. Subsequently, Shapiro and Wilk [57] proposed a normalizing transformation for W in the region n=7(1)50 although tables were still required for n=4(1)6. In 1982 Royston [58] produced an extension to the W-test allowing sample sizes up to 2000; subsequently, he raised this limit to 5000 [59,60].

The Shapiro–Wilk test statistic is defined as (Eq. (44)):

$$W = \frac{1}{D}\left[\sum_{i=1}^{k} a_i\left(x_{(n-i+1)}-x_{(i)}\right)\right]^2 \tag{44}$$

where D is (Eq. (45)):

$$D = \sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{45}$$

and k (Eq. (44)) is approximately $n/2$ and the a_i coefficients represent what the statistics would look like if the population was normal. These are obtained from volume two of the *Biometrika Tables* [61] by entering the value of n.

The p-value of W is calculated (Eqs. (46) and (15)) where the values for b, c, and d are obtained from *Biometrika Tables* [61] by entering the value of n:

$$z = b_n + c_n\left[\ln\frac{W-d_n}{1-W}\right]. \tag{46}$$

As noted, the original configuration of the W-test required the use of tables. Royston [58] developed an approximate Normalizing transformation suitable for computer implementation that calculated the W value and its significance level for any sample size between 3 and 2000 [62]. Later an improved algorithm was published that covered the range $3 \leq n \leq 5000$ [60].

7.3. Shapiro–Francia test

The Shapiro–Francia W'-test test was described in 1972 [63]. Like the W-test it is the square of the Pearson correlation coefficient computed between the order statistics of the sample values and the expected Normal order statistics (as illustrated by the straight line in Fig. 4A). The W' test is readily calculated unlike the more complex calculation of the a_i coefficients in the W-test. Because of this difference between the W and W'-tests, the values of W and W' differ slightly (see Table 5). The major advantage of the W'-test, over the *original* formulation of the W-test, was that it was not limited to a sample size <51.

Royston [64,65] proposed an easy-to-calculate approximation to the Shapiro–Francia test and its p-value for sample sizes $5 \leq n \leq 5000$. The value of W' was first calculated as noted above and the p-value was obtained as follows (Eq. (47)):

$$z = [\ln(1-W') - \hat{\mu}]/\hat{\sigma} \tag{47}$$

where (Eq. (48)):

$$\left.\begin{array}{l}\hat{\mu} = -1.2725 + 1.0521(v-u)\\ \hat{\sigma} = 1.0308 + 0.26758(v+2/u)\\ \text{where } u = \ln(n) \text{ and } v = \ln(u)\end{array}\right\} \tag{48}$$

and the p-value is calculated using Eq. (15). Note that z refers to the upper tail of the distribution.

7.4. Filliben's r test

In 1975 Filliben [66] described the probability plot correlation coefficient as a test for the composite hypothesis for Normality. He used the correlation between the sample order statistics and the estimated median values of the theoretical order statistics.

The data is sorted in ascending order and indexed (i). The uniform $[0,1]$ order statistic medians (m_i) are calculated where n is the number of data samples (Eq. (49)):

$$m_i = \begin{cases} 1-0.5^{(1/n)} & i=1\\ (i-0.3175)/(n+0.365) & i=2,3,\ldots,n-1\\ 0.5^{(1/n)} & i=n \end{cases} \tag{49}$$

The Normal $N\ (0,1)$ order statistic medians M_i are next computed (Eq. (50)):

$$M_i = \Phi^{-1}(m_i) \tag{50}$$

The Pearson correlation coefficient was then calculated using the values of the original data and M_i. The critical values for the percentage points 0.005, 0.01, 0.025, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 0.975, 0.99, and 0.995 are provided by Filliben for values of n from 3 to 100.

Filliben suggested that the advantage of the r-test lay in its conceptual simplicity—the use of the probability plot and the correlation coefficient. Both the W- and W'-tests rely on the equivalence of two measures of variation—the squared slope of Normal probability plot regression line and the residual mean square of the regression line. Further he showed that the power of the r-test compared well with the W- and W'-tests.

7.5. Summary of test results

The Gan–Koehler test correctly classified the normal, positive kurtosis, and positive skewed samples. However, it did not detect the abnormality in the negatively skewed sample. By contrast, the Shapiro–Wilk and Shapiro–Francia tests correctly classified all four test samples. The Filliben test correctly classified the three non-Normal samples but unaccountably classified the normal test sample as non-Normal.

8. The effect of sample size on the resulting *p*-value

Using a random Normal population, $N(0,1)$, of varying sample sizes it is possible to detect changes in the p-values as the sample sizes increase calculated from the D'Agostino/Anscombe and Glynn transformations [19,20]. In contrast, using the same Normal population, and the Hosking's statistic (Eq. (24)), the p-values remain constant over the range $5 \leq n \leq 1000$.

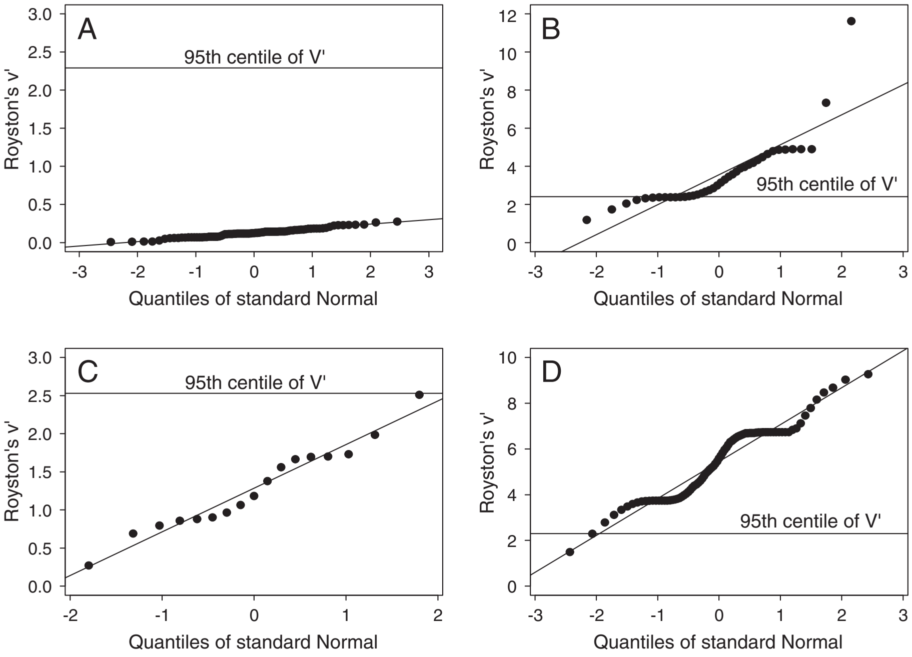

Fig. 8. Royston's v plots for the data displayed in Fig. 1. The quantiles of the standard normal distribution are plotted against Royston's v'. The horizontal line indicates the respective 95th centiles for V'.

Royston [67] pointed out that the p-values obtained from the Shapiro–Francia W' or Shapiro–Wilk W tests are not quantitative measures of departures from normality. Indeed, "samples of differing size drawn from the same non-Normal distribution could result in p-values respectively above and below the magic 0.05 level". He proceeded to derive two indices based on the W' and W tests with their associated plots [67]. Only the index V' and the v'-plot used with the Shapiro–Francia test will be briefly described and utilized here. In essence, V' is the sample residual variance of the differences between the ordered data and their expected values assuming normality. It is thus an index of departure from normality that the Shapiro–Francia test is not. Royston showed that the 95th centile of the V' index from a Normal population (identical to that used earlier) varied little (2.8 to 2.04) over the range of sample sizes 5 to 1000 whereas the 95th centile of $1-W'$ (a surrogate index of W') varied by a factor of >70-fold. p-values varied between 0.554 (i.e., Normal) and 0.026 (non-Normal). Royston used the v'-plot (Fig. 8) to determine if the sample departs significantly from Normality by the plotted points exceeding the 95th centile (i.e., the value of W' is significant at the 0.05 level). Clearly, plot A is Normal while both plots B and D are definitely non-Normal. Observe that plot C is just within the 95th centile although all EDF and Regression/Correlation tests (Table 5) showed the distribution to be non-Normal.

9. Power comparisons of tests for Normality

Several extremely comprehensive studies of the comparative effectiveness of various tests of normality have been published [56,68,69,46,40,66,70,55] (those mentioned, but not discussed here, are described in Thode's monograph). Many of these studies have been comprehensively summarized by Thode [7]. Generally the studied distributions fell into three main categories—long-tailed symmetric, short-tailed symmetric, and asymmetric. Sample sizes generally ranged from 10 to 100. Shapiro and Wilk [56] studied the empirical power of nine test procedures ($\sqrt{\beta_1}, \beta_2$ u,[4] Kolmogorov–Smirnov, Cramér–von Mises, Anderson–Darling, chi-squared, and Shapiro–Wilk tests) on fifteen non-Normal populations. The more extensive 1968 study (Shapiro et al. [68]) compared the power of these nine test procedures using 45 distributions in twelve families for several sample sizes ($n=10$, 15, 20, 35, and 50). They concluded: that the Shapiro–Wilk test provided a generally superior measure of non-Normality, that the u statistic was excellent against symmetric short-tailed distributions but insensitive to asymmetry, and while $\sqrt{\beta_1}$ and β_2 were sensitive their performance was usually dominated by the Shapiro–Wilk test. Chen [69] demonstrated that the Shapiro–Wilk test showed adequate sensitivity when applied to scale- or location-contaminated Normal distributions.[5]

Stephens [46] studied power comparisons ($n=10$, 20, 50, 100, and 8) of seven tests (D'Agostino D, Cramér–von Mises, Anderson–Darling, Shapiro–Wilk, Shapiro–Fran-

[4] Otherwise known as the range test defined as: $u = (x_{\max} - x_{\min})/SD$.

[5] A scale-contaminated distribution is composed of two superimposed normal distributions with the same means but differing SD's. A location-contaminated distribution is composed of two normal distributions with the same SD's but differing means.

cia, Kuiper, and Watson tests under three situations—where the hypothesised distribution (of which there were nine) is completely specified or one or more parameters require to be estimated. When both the mean and SD have to be estimated Stephens reported that the Anderson–Darling and Cramér–von Mises tests appeared to possess the highest power. However, both the Shapiro–Wilk and Shapiro–Francia tests were reported to have a superior performance to these EDF tests.

D'Agostino and Rosman [40] observed that Geary's *a*-test was not included in the earlier noted set of comparisons. They performed comparisons with sample sizes of 20, 50, and 100 on ten distribution types and concluded that Geary's test matched the Shapiro–Wilk and Shapiro–Francia tests as a two-sided test on symmetric distributions but performed better as a one-sided test. With the present availability of small-sample points of β_2 there is no requirement for Geary's test as a test for kurtosis. Also Geary's test does not perform as well as the Shapiro–Wilk or Shapiro–Francia on skewed distributions (a finding confirmed here). Finally, $\sqrt{\beta_1}$ usually exceeds Geary's test for all sample sizes as do the other tests on contaminated Normal distributions.

Filliben [66] compared the power of his r-test to several tests including Shapiro–Wilk, Shapiro–Francia, and D'Agostino's *D* statistic using samples sizes of 20, 50, and 100. He used varieties of three distribution types—(i) short-tailed symmetric, (ii) long-tailed symmetric, and (iii) skewed. He concluded that, in comparison with the Shapiro–Wilk test, the *r*-test was poorer for type (i), marginally better for type (ii), and marginally poorer for type (iii).

Pearson et al.[70] compared the powers of some omnibus (K^2, *R* [70],[6] Shapiro–Wilk's *W*, and D'Agostino's *D*) and directional ($\sqrt{\beta_1}$(upper tail), β_2 (upper and lower tail), and D'Agostino's *D* (upper and lower tail)) tests. They used 58 non-Normal populations derived from 12 distributions (almost equally divided between symmetrical (of which 9 were scale-contaminated) and skewed (9 location-contaminated populations) for samples of $n=20$, 50, and 100. The sampled populations were classified, in the symmetrical group, by their degree of kurtosis (ranging from 1.6 to infinity) and by both $\sqrt{\beta_1}$ and β_2 with the former ranging from 0.7 to 6). Interestingly, given the complexity of the results, the authors emphasized the difficulty in disentangling the relationship between particular tests and their power. For symmetric platykurtic ($\beta_2<3$) populations the Shapiro–Wilk (*W*) test was the best omnibus test while the best directional test was the lower tailed β_2 test but overall this test had superior power to the *W*-test. For symmetric leptokurtic ($\beta_2>3$) populations the omnibus K^2 test was best but for long-tailed the directional tests β_2 and D'Agostino's *D* test were equally good. For the skewed populations the best test is the upper-tailed $\sqrt{\beta_1}$ when the population is positively skewed ($\sqrt{\beta_1}>0$) while the lower-tailed $\sqrt{\beta_1}$ test is better. The above noted conclusions applied to sample sizes of 20 and 50. The comparative results on samples of 100 were quite different as many tests showed powers of 100% or so and because values of $\sqrt{\beta_1}$ and β_2 were not recorded the authors concluded that the findings for sample sizes of 20 and 50 applied to symmetric leptokurtic populations and to skewed populations.

Gan and Koehler [55] described two goodness-of-fit tests based on measures of linearity for standardized *P–P* plots (see Regression/Correlation Tests, above) and compared their performance with Filliben's *r*, Shapiro–Wilk, Anderson–Darling, Cramér–von Mises, Watson, *R*, and Kolmogorov–Smirnov tests using 91 distributions with sample sizes of 20, 50, and 100. They showed that Shapiro–Wilk was the best overall test for Normality and that the Kolmogorov–Smirnov test was the weakest. The *r* test is useful for heavy-tailed distributions that are either symmetrical or not too skewed but is poor for distributions with $\beta_2<3$. The Anderson–Darling, Cramér–von Mises, Watson, and *R* tests performed well and the Gan–Koehler tests were slightly less powerful than the Cramér–von Mises test.

Unlike the considerable literature reviewing the comparative performance of C-moments and tests of Normality there are only two articles reviewing the comparative performance of L-moments [32,29]. These authors took diametrically different approaches—Hosking [32] using a series of simulated distributions and Royston [29] using three medical data sets.

Hosking used 21 distributions ($n=20$) derived from the Gan and Koelher collection [55] that were symmetric and long-tailed and compared β_2 (with values ≥ 3) and τ_4 values against the Shapiro–Wilk *W* test results. Likewise he compared the power of the Shapiro–Wilk test with the $\sqrt{\beta_1}$, β_2, τ_3, and τ_4 values of 31 skewed distributions ($n=20$) also derived from the Gan and Koehler collection.

An analysis of Hosking's data tabulations follows. The L-kurtosis values of the symmetrical distributions correlated well ($r=0.985$) with the power of the Shapiro–Wilk test while the correlation with the kurtosis values was poor ($r=0.3$). Plots of the skewed distributions (L-skewness/power and skewness/power) displayed a U-shaped relationship although each arm showed a superior correlation for the L-skew values ($r=0.9$ and 0.99) than for the skew values ($r=0.66$ and 0.95). Hosking concluded that the L-moments were more concordant with the power of the Shapiro–Wilk results than were the C-moments.

Royston [29] examined three data sets comparing C-moments with L-moments. He demonstrated, using 15 sets of highly-skewed and leptokurtic maternal serum–fetoprotein results, that C-moment skewness and kurtosis varied 30- and 300-fold, respectively, whereas L-skewness and L-kurtosis

[6] The K^2 test is a graphical test showing the bivariate relationship between the indices of skewness and kurtosis using a series of contours for a range of sample sizes. The *R* test is a sequential procedure that estimates the likely maximum number of outliers before calculating a trimmed mean and variance. The *R* value is compared to a table of critical values to determine whether any outliers exist in the original sample.

vales were tightly controlled. A second data set, possessing a single outlier, was randomly sampled obtaining subsamples of a range of percentages of the original sample ($n = 251$). While C-moment skewness and kurtosis showed considerable variability both the L-skewness and L-kurtosis showed less variability especially when the sample size increased. The third data set—a skewed distribution of 216 bilirubin values—was used to illustrate the bias in C-moment skewness and kurtosis in small to moderate size samples as indicated by Eq. (16). Such a bias is almost negligible in the L-skewness and L-kurtosis indices. The joint conclusions of these authors are listed in Tables 2 and 4 regarding the advantages of L-moments.

It is not easy to draw firm conclusions from the foregoing regarding the "best" test for Normality. In general, however, the Anderson–Darling, Shapiro–Wilk, and Shapiro–Francia tests appear to be the most frequently favoured tests. Certainly these three tests perform well when used on the four test samples of the type commonly encountered in clinical chemistry when studying experimentally-derived results.

10. Concluding remarks

The following steps are suggested when examining experimental data for Normality:

- Identify all programs used in the calculations thus avoiding ambiguity regarding indices of skewness and kurtosis.
- Plot the data using histograms or box-and-whisker diagrams and supplwement these with Q–Q or P–P plots.
- Consider the advantages of using L-moments in place of C-moments.
- Test for Normality with Anderson–Darling, Shapiro–Wilk, or Shapiro–Francia tests.
- Monitor the effect of sample size on the resulting p-value using Royston's V/v or V'/v' tests thus suggesting the use of the Shapiro–Wilk or Shapiro–Francia tests for assessing Normality.

The majority of the described programs are available within the freely-available R site except for Wang's L-moments and Royston's V'/v'. R/S-Plus versions of these two programs are available on request to the author.

Acknowledgements

I am most grateful to Dr Jon Hosking (IBM Research Division, TJ Watson Research Center, Yorktown Heights, NY) for his goodwill and patience in constructively responding to my questions, and to the technical support staff at Insightful Corporation for their advice and assistance.

References

[1] Linnet K. Testing normality of transformed data. Appl Stat 1988; 37:180–6.

[2] Horn P, Pesce AJ. Reference intervals. A user's guide. Washington, DC: AACC Press; 2005. p. 1–123.

[3] Solberg HE. Chapter 16: establishment and use of reference values. In: Burtis CA, Ashwood ER, Bruns DE, editors. Tietz textbook of clinical chemistry and molecular diagnostics. 4th ed. St. Louis: Elsevier Saunders; 2005. p. 425–48.

[4] Wright EM, Royston P. Calculating reference intervals for laboratory measurements. Stat Methods Med Res 1999;8:93–112.

[5] Solberg HE. RefVal: a program implementing the recommendations of the International Federation of Clinical Chemistry on the statistical treatment of reference values. Comput Methods Programs Biomed 1995;48:247–56.

[6] D'Agostino RB, Stephens MA, editors. Goodness-of-fit techniques. New York: Marcel Dekker, Inc.; 1986. p. 1–560.

[7] Thode HCJ. Testing for normality. New York: Marcel Dekker, Inc.; 2002. p. 1–479.

[8] Leung FY, Galbraith LV, Jablonsky G, Henderson AR. Re-evaluation of the diagnostic utility of serum total creatine kinase and creatine kinase-2 in myocardial infarction. Clin Chem 1989;35:1435–40.

[9] R Development Core Team. R: A language and environment for statistical computing. http://www.R-project.org. (Accessed 19-10-2005).

[10] Fox J. An R and S-PLUS companion to applied regression. Thousand Oaks, CA: Sage Publications, Inc.; 2002. p. 1–312.

[11] Tukey JW. Exploratory data analysis. Reading, MA: Addison-Wesley Publishing Company; 1977. p. 1–688.

[12] Pearson ES, Please NW. Relation between the shape of population distribution and the robustness of four simple test statistics. Biometrika 1975;62:223–41.

[13] Cramér H. Mathematical methods of statistics. Princeton: Princeton University Press; 1946. p. 1–575.

[14] Stuart A, Ord JK. Kendall's advanced theory of statistics. Distribution theory, volume 1, 6th ed. London: Arnold; 1994. p. 1–676.

[15] Cornish EA, Fisher RA. Moments and cumulants in the specification of distributions. Rev Int Stat Inst 1937;5:307–22.

[16] Joanes DN, Gill CA. Comparing measures of sample skewness and kurtosis. Statistician 1998;47:183–9.

[17] Stuart A, Ord JK, Arnold S. Kendall's advanced theory of statistics. Classical inference and the linear model, volume 2A.6th ed. London: Arnold; 1999. p. 1–885.

[18] D'Agostino RB. Chapter 9: tests for the normal distribution. Goodness-of-fit techniques. New York: Marcel Dekker, Inc.; 1986. p. 367–419.

[19] D'Agostino RB. Transformation to normality of the null distribution of g_2. Biometrika 1970;57:679–81.

[20] Anscombe FJ, Glynn WJ. Distribution of the kurtosis statistic b2 for normal samples. Biometrika 1983;70:227–34.

[21] Hald A. Statistical theory with engineering applications. New York: John Wiley and Sons, Inc.; 1952. p. 1–783.

[22] D'Agostino RB, Belanger A, D'Agostino Jr RB. A suggestion for using powerful and informative tests of normality. Am Stat 1990;44:316–21.

[23] Kirby W. Algebraic boundedness of sample statistics. Water Resour Res 1974;10:220–2.

[24] Dalén J. Algebraic bounds on standardized sample moments. Stat Probab Lett 1987;5:329–31.

[25] Johnson ME, Lowe VW. Bounds on the sample skewness and kurtosis. Technometrics 1979;21:377–8.

[26] Balanda KP, MacGillivray HL. Kurtosis: a critical review. Am Stat 1988;42:111–9.

[27] Hosking JRM. L-moments: analysis and estimation of distributions using linear combinations of order statistics. JR Stat Soc B 1990;52:105–24.

[28] Greenwood JA, Landwehr JM, Matalas NC, Wallis JR. Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form. Water Resour Res 1979;15:1049–54.

[29] Royston P. Which measures of skewness and kurtosis are best? Stat Med 1992;11:333–43.

[30] Hosking JRM, Wallis JR. Regional frequency analysis. An approach based on L-moments. Cambridge: Cambridge University Press; 1997. p. 1–224.

[31] Wang QJ. Direct sample estimators of L-moments. Water Resour Res 1996;32:3617–9.

[32] Hosking JRM. Moments or L-moments? An example comparing two measures of distributional shape. Am Stat 1992;46:186–9.

[33] Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman and Hall; 1993. p. 1–436.

[34] Elamir EAH, Seheult AH. Trimmed L-moments. Comp Stat Data Anal 2003;43:299–314.

[35] Vogel RM, Fennessey NM. L-moment diagrams should replace product moment diagrams. Water Resour Res 1993;29:1745–52.

[36] Hosking JRM. Fortran routines for use with the method of L-moments. http://lib.stat.cmu.edu/general/lmoments. (Accessed 19-4-2005).

[37] Hosking JRM. SAMLMU: estimating sample L-moments. http://www.r-project.org/nocvs/mail/r-help/2001/6042.html. (Accessed 20-11-2005).

[38] Geary RC. The ratio of the mean deviation to the standard deviation as a test for normality. Biometrika 1935;27:310–32.

[39] Geary RC. Moments of the ratio of the mean deviation to the standard deviation for normal samples. Biometrika 1936;28:295–305.

[40] D'Agostino RB, Rosman B. The power of Geary's test of normality. Biometrika 1974;61:181–4.

[41] D'Agostino RB. Simple compact portable test of normality: Geary's test revisited. Psychol Bull 1970;74:138–40.

[42] van Belle G, Fisher LD, Heagerty PJ, Lumley T. Biostatistics: a methodology for the health sciences.2nd ed. New York: Wiley Interscience; 2004. p. 1–871.

[43] Moore DS. Chapter 3: tests of chi-squared type. In: D'Agostino RB, Stephens MA, editors. Goodness-of-fit techniques. New York: Marcel Dekker, Inc.; 1986. p. 63–95.

[44] Stephens MA. Chapter 4: tests based on EDF statistics. In: D'Agostino MA, Stephens MA, editors. Goodness-of-fit techniques. New York: Marcel Dekker, Inc.; 1986. p. 97–193.

[45] Conover WJ. Chapter 6: statistics of the Kolmogorov–Smirnov type. Practical nonparametric statistics.2nd ed. New York: John Wiley and Sons; 1980. p. 344–93.

[46] Stephens MA. EDF statistics for goodness of fit and some comparisons. J Am Stat Assoc 1974;69:730–7.

[47] Massey FJ. The Kolmogorov–Smirnov test for goodness-of-fit. J Am Stat Assoc 1951;46:68–78.

[48] Lilliefors HW. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. J Am Stat Assoc 1967;62:399–402.

[49] Mason AL, Bell CB. New Lilliefors and Srinivasan tables with applications. Comm Stat—Simul 1986;15:451–67.

[50] Dallal GE, Wilkinson L. An analytic approximation to the distribution of Lilliefor's test statistic for normality. Am Stat 1986; 40:294–6.

[51] Iman RL. Graphs for use with the Lilliefors test for normal and exponential distributions. Am Stat 1982;36:109–12.

[52] Anderson TW, Darling DA. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. Ann Math Stat 1952;23:193–212.

[53] Anderson TW, Darling DA. A test of goodness of fit. J Am Stat Assoc 1954;49:765–9.

[54] Stephens MA. Chapter 5: tests based on regression and correlation. In: D'Agostino RB, Stephens MA, editors. Goodness-of-fit techniques. New York: Marcel Dekker, Inc.; 1986. p. 195–233.

[55] Gan FF, Koehler KJ. Goodness-of-fit tests based on *P*–*P* probability plots. Technometrics 1990;32:289–303.

[56] Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 1965;52:591–611.

[57] Shapiro SS, Wilk MB. Approximations for the null distribution of the *W* statistic. Technometrics 1968;10:861–6.

[58] Royston JP. An extension of Shapiro and Wilk's *W* test for normality to large samples. Appl Stat 1982;31:115–24.

[59] Royston JP. Approximating the Shapiro–Wilk *W*-test for non-normality. Stat Comput 1982;2:117–9.

[60] Royston P. A remark on algorithm AS 181: the *W*-test for normality. Appl Stat 1995;44:547–51.

[61] Pearson ES, Hartley HO. Biometrika tables for statisticians, volume two, reprinted with corrections. London: Biometrika Trust; 1976. p. 1–385.

[62] Royston JP. Algorithm AS 181: the *W* test for normality. Appl Stat 1982;31:176–80.

[63] Shapiro SS, Francia RS. An approximate analysis of variance test for normality. J Am Stat Assoc 1972;67:215–6.

[64] Royston P. A toolkit for testing for non-normality in complete and censored samples. Statistician 1993;42:37–43.

[65] Royston P. A pocket-calculator algorithm for the Shapiro–Francia test for non-normality: an application to medicine. Stat Med 1993;12:181–4.

[66] Filliben JJ. The probability plot correlation coefficient test for normality. Technometrics 1975;17:111–7.

[67] Royston P. Estimating departure from normality. Stat Med 1991;10:1283–93.

[68] Shapiro SS, Wilk MB, Chen HJ. A comparative study of various tests for normality. J Am Stat Assoc 1968;63:1343–72.

[69] Chen EH. The power of the Shapiro–Wilk *W* test for normality in samples from contaminated normal distributions. J Am Stat Assoc 1971;66:760–2.

[70] Pearson ES, D'Agostino RB, Bowman KO. Tests for departure from normality: comparison of powers. Biometrika 1977;64:231–46.

[71] Pearson K. On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen in random sampling. Phil Mag 1900;50:157–75.

[72] Cramér H. On the composition of elementary errors. II: statistical applications. Skand Aktuartidskr 1928;11:141–80.

[73] von Mises, R. Wahrscheinlichkeitsrechnung. Leipzig-Wein 1931.

[74] Smirnov, NV. Sur la distribution de w^2. C R Acad Sci Paris 1936;202:449–52.