Designation: E2586 – 16

An American National Standard

Standard Practice for Calculating and Using Basic Statistics[1]

This standard is issued under the fixed designation E2586; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope

1.1 This practice covers methods and equations for computing and presenting basic descriptive statistics using a set of sample data containing a single variable. This practice includes simple descriptive statistics for variable data, tabular and graphical methods for variable data, and methods for summarizing simple attribute data. Some interpretation and guidance for use is also included.

1.2 The system of units for this practice is not specified. Dimensional quantities in the practice are presented only as illustrations of calculation methods. The examples are not binding on products or test methods treated.

1.3 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate safety and health practices and determine the applicability of regulatory limitations prior to use.*

2. Referenced Documents

2.1 *ASTM Standards:*[2]

E178 Practice for Dealing With Outlying Observations

E456 Terminology Relating to Quality and Statistics

E2282 Guide for Defining the Test Result of a Test Method

E3080 Practice for Regression Analysis

2.2 *ISO Standards:*[3]

ISO 3534-1 Statistics—Vocabulary and Symbols, part 1: Probability and General Statistical Terms

ISO 3534-2 Statistics—Vocabulary and Symbols, part 2: Applied Statistics

3. Terminology

3.1 *Definitions*—Unless otherwise noted, terms relating to quality and statistics are as defined in Terminology E456.

3.1.1 *characteristic, n*—a property of items in a sample or population which, when measured, counted, or otherwise observed, helps to distinguish among the items. **E2282**

3.1.2 *coefficient of variation, CV, n*—for a nonnegative characteristic, the ratio of the standard deviation to the mean for a population or sample

3.1.2.1 *Discussion*—The coefficient of variation is often expressed as a percentage.

3.1.2.2 *Discussion*—This statistic is also known as the *relative standard deviation, RSD.*

3.1.3 *confidence bound, n*—see *confidence limit.*

3.1.4 *confidence coefficient, n*—see *confidence level.*

3.1.5 *confidence interval, n*—an interval estimate [L, U] with the statistics L and U as limits for the parameter θ and with confidence level 1 – α, where Pr(L ≤ θ ≤ U) ≥ 1 – α.

3.1.5.1 *Discussion*—The confidence level, 1 – α, reflects the proportion of cases that the confidence interval [L, U] would contain or cover the true parameter value in a series of repeated random samples under identical conditions. Once L and U are given values, the resulting confidence interval either does or does not contain it. In this sense "confidence" applies not to the particular interval but only to the long run proportion of cases when repeating the procedure many times.

3.1.6 *confidence level, n*—the value, 1 – α, of the probability associated with a confidence interval, often expressed as a percentage.

3.1.6.1 *Discussion*—α is generally a small number. Confidence level is often 95 % or 99 %.

3.1.7 *confidence limit, n*—each of the limits, L and U, of a confidence interval, or the limit of a one-sided confidence interval.

3.1.8 *degrees of freedom, n*—the number of independent data points minus the number of parameters that have to be estimated before calculating the variance.

3.1.9 *estimate, n*—sample statistic used to approximate a population parameter.

3.1.10 *histogram, n*—graphical representation of the frequency distribution of a characteristic consisting of a set of rectangles with area proportional to the frequency. **ISO 3534-1**

3.1.10.1 *Discussion*—While not required, equal bar or class widths are recommended for histograms.

[1] This practice is under the jurisdiction of ASTM Committee E11 on Quality and Statistics and is the direct responsibility of Subcommittee E11.10 on Sampling / Statistics.

Current edition approved Nov. 1, 2016. Published November 2016. Originally approved in 2007. Last previous edition approved in 2014 as E2586 – 14. DOI: 10.1520/E2586-16.

[2] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

[3] Available from American National Standards Institute (ANSI), 25 W. 43rd St., 4th Floor, New York, NY 10036, http://www.ansi.org.

3.1.11 *interquartile range, IQR, n*—the 75th percentile (0.75 quantile) minus the 25th percentile (0.25 quantile), for a data set.

3.1.12 *kurtosis, γ_2, g_2, n*—for a population or a sample, a measure of the weight of the tails of a distribution relative to the center, calculated as the ratio of the fourth central moment (empirical if a sample, theoretical if a population applies) to the standard deviation (sample, *s*, or population, σ) raised to the fourth power, minus 3 (also referred to as excess kurtosis).

3.1.13 *mean, n—of a population*, μ, average or expected value of a characteristic in a population – *of a sample*, $\bar{X}$, sum of the observed values in the sample divided by the sample size.

3.1.14 *median, $\tilde{X}$, n*—the 50th percentile in a population or sample.

3.1.14.1 *Discussion*—The sample median is the *[(n + 1)/2]* order statistic if the sample size *n* is odd and is the average of the *[n/2]* and *[n/2 + 1]* order statistics if *n* is even.

3.1.15 *midrange, n*—average of the minimum and maximum values in a sample.

3.1.16 *order statistic, $x_{(k)}$, n*—value of the kth observed value in a sample after sorting by order of magnitude.

3.1.16.1 *Discussion*—For a sample of size *n*, the first order statistic $x_{(1)}$ is the minimum value, $x_{(n)}$ is the maximum value.

3.1.17 *parameter, n*—see *population parameter*.

3.1.18 *percentile, n*—quantile of a sample or a population, for which the fraction less than or equal to the value is expressed as a percentage.

3.1.19 *population, n*—the totality of items or units of material under consideration.

3.1.20 *population parameter, n*—summary measure of the values of some characteristic of a population. **ISO 3534-2**

3.1.21 *prediction interval, n*—an interval for a future value or set of values, constructed from a current set of data, in a way that has a specified probability for the inclusion of the future value.

3.1.22 *quantile, n*—value such that a fraction *f* of the sample or population is less than or equal to that value.

3.1.23 *range, R, n*—maximum value minus the minimum value in a sample.

3.1.24 *residual, n*—observed value minus fitted value, when a model is used. **E3080**

3.1.25 *sample, n*—a group of observations or test results, taken from a larger collection of observations or test results, which serves to provide information that may be used as a basis for making a decision concerning the larger collection.

3.1.26 *sample size, n, n*—number of observed values in the sample.

3.1.27 *sample statistic, n*—summary measure of the observed values of a sample.

3.1.28 *skewness, γ_1, g_1, n*—for population or sample, a measure of symmetry of a distribution, calculated as the ratio of the third central moment (empirical if a sample, and theoretical if a population applies) to the standard deviation (sample, *s*, or population, σ) raised to the third power.

3.1.29 *standard error*—standard deviation of the population of values of a sample statistic in repeated sampling, or an estimate of it.

3.1.29.1 *Discussion*—If the standard error of a statistic is estimated, it will itself be a statistic with some variance that depends on the sample size.

3.1.30 *standard deviation—of a population*, σ, the square root of the average or expected value of the squared deviation of a variable from its mean; *—of a sample, s*, the square root of the sum of the squared deviations of the observed values in the sample from their mean divided by the sample size minus 1.

3.1.31 *statistic, n*—see *sample statistic*.

3.1.32 *variance, σ^2, s^2, n*—square of the standard deviation of the population or sample.

3.1.32.1 *Discussion*—For a finite population, σ^2 is calculated as the sum of squared deviations of values from the mean, divided by *n*. For a continuous population, σ^2 is calculated by integrating $(x - \mu)^2$ with respect to the density function. For a sample, s^2 is calculated as the sum of the squared deviations of observed values from their average divided by one less than the sample size.

3.1.33 *Z-score, n*—observed value minus the sample mean divided by the sample standard deviation.

4. Significance and Use

4.1 This practice provides approaches for characterizing a sample of *n* observations that arrive in the form of a data set. Large data sets from organizations, businesses, and governmental agencies exist in the form of records and other empirical observations. Research institutions and laboratories at universities, government agencies, and the private sector also generate considerable amounts of empirical data.

4.1.1 A data set containing a single variable usually consists of a column of numbers. Each row is a separate observation or instance of measurement of the variable. The numbers themselves are the result of applying the measurement process to the variable being studied or observed. We may refer to each observation of a variable as an item in the data set. In many situations, there may be several variables defined for study.

4.1.2 The sample is selected from a larger set called the population. The population can be a finite set of items, a very large or essentially unlimited set of items, or a process. In a process, the items originate over time and the population is dynamic, continuing to emerge and possibly change over time. Sample data serve as representatives of the population from which the sample originates. It is the population that is of primary interest in any particular study.

4.2 The data (measurements and observations) may be of the variable type or the simple attribute type. In the case of attributes, the data may be either binary trials or a count of a defined event over some interval (time, space, volume, weight, or area). Binary trials consist of a sequence of 0s and 1s in which a "1" indicates that the inspected item exhibited the attribute being studied and a "0" indicates the item did not exhibit the attribute. Each inspection item is assigned either a "0" or a "1." Such data are often governed by the binomial distribution. For a count of events over some interval, the number of times the event is observed on the inspection interval is recorded for each of *n* inspection intervals. The Poisson distribution often governs counting events over an interval.

4.3 For sample data to be used to draw conclusions about the population, the process of sampling and data collection must be considered, at least potentially, repeatable. Descriptive statistics are calculated using real sample data that will vary in repeating the sampling process. As such, a statistic is a random variable subject to variation in its own right. The sample statistic usually has a corresponding parameter in the population that is unknown (see Section 5). The point of using a statistic is to summarize the data set and estimate a corresponding population characteristic or parameter.

4.4 Descriptive statistics consider numerical, tabular, and graphical methods for summarizing a set of data. The methods considered in this practice are used for summarizing the observations from a single variable.

4.5 The descriptive statistics described in this practice are:

4.5.1 Mean, median, min, max, range, mid range, order statistic, quartile, empirical percentile, quantile, interquartile range, variance, standard deviation, *Z*-score, coefficient of variation, skewness and kurtosis, and standard error.

4.6 Tabular methods described in this practice are:

4.6.1 Frequency distribution, relative frequency distribution, cumulative frequency distribution, and cumulative relative frequency distribution.

4.7 Graphical methods described in this practice are:

4.7.1 Histogram, ogive, boxplot, dotplot, normal probability plot, and q-q plot.

4.8 While the methods described in this practice may be used to summarize any set of observations, the results obtained by using them may be of little value from the standpoint of interpretation unless the data quality is acceptable and satisfies certain requirements. To be useful for inductive generalization, any sample of observations that is treated as a single group for presentation purposes must represent a series of measurements, all made under essentially the same test conditions, on a material or product, all of which have been produced under essentially the same conditions. When these criteria are met, we are minimizing the danger of mixing two or more distinctly different sets of data.

4.8.1 If a given collection of data consists of two or more samples collected under different test conditions or representing material produced under different conditions (that is, different populations), it should be considered as two or more separate subgroups of observations, each to be treated independently in a data analysis program. Merging of such subgroups, representing significantly different conditions, may lead to a presentation that will be of little practical value. Briefly, any sample of observations to which these methods are applied should be homogeneous or, in the case of a process, have originated from a process in a state of statistical control.

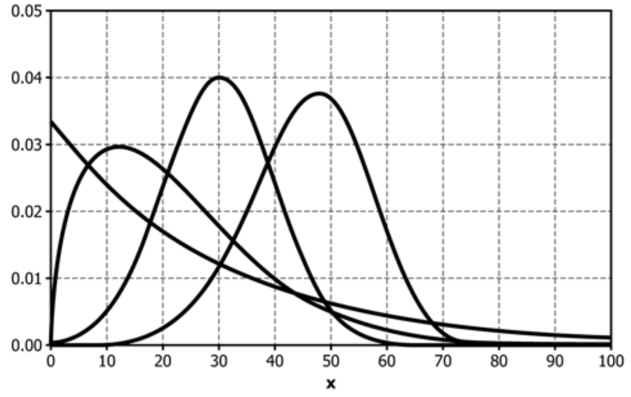

FIG. 1 Probability Density Function—Four Examples of Distribution Shape

4.9 The methods developed in Sections 6, 7, and 8 apply to the sample data. There will be no misunderstanding when, for example, the term "mean" is indicated, that the meaning is sample mean, not population mean, unless indicated otherwise. It is understood that there is a data set containing *n* observations. The data set may be denoted as:

$$x_1, x_2, x_3 \ldots x_n \tag{1}$$

4.9.1 There is no order of magnitude implied by the subscript notation unless subscripts are contained in parenthesis (see 6.7).

5. Characteristics of Populations

5.1 A population is the totality of a set of items under consideration. Populations may be finite or unlimited in size and may be existing or continuing to emerge as, for example, in a process. For continuous variables, *X*, representing an essentially unlimited population or a process, the population is mathematically characterized by a probability density function, *f*(*x*). The density function visually describes the shape of the distribution as for example in Fig. 1. Mathematically, the only requirements of a density function are that its ordinates be all positive and that the total area under the curve be equal to 1.

5.1.1 Area under the density function curve is equivalent to probability for the variable *X*. The probability that *X* shall occur between any two values, say *s* and *t*, is given by the area under the curve bounded by the two given values of *s* and *t*. This is expressed mathematically as a definite integral over the density function between *s* and *t*:

$$P\ (s < X \le t) = \int_s^t f(x)dx \tag{2}$$

5.1.2 A great variety of distribution shapes are theoretically possible. When the curve is symmetric, we say that the distribution is symmetric; otherwise, it is asymmetric. A distribution having a longer tail on the right side is called right skewed; a distribution having a longer tail on the left is called left skewed.

5.1.3 For a given density function, *f*(*x*), the relationship to cumulative area under the curve may be graphically shown in the form of a cumulative distribution function, *F*(*x*). The function *F*(*x*) plots the cumulative area under *f*(*x*) as *x* moves

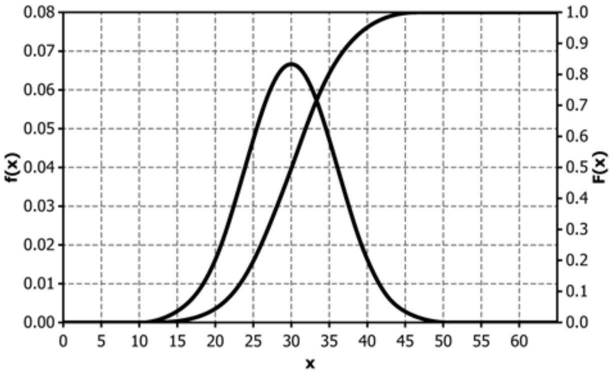

FIG. 2 Cumulative Distribution Function, *F*(*x*), and Density Function, *f*(*x*) Relationship

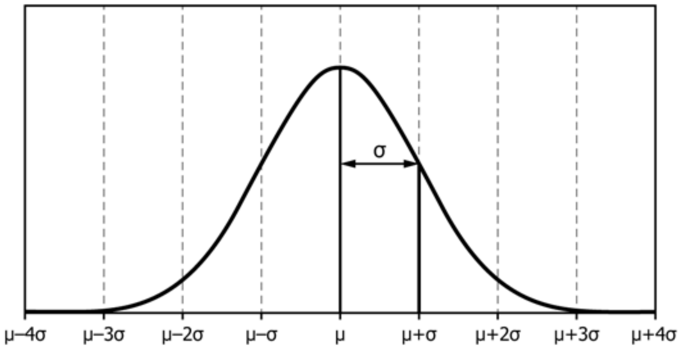

FIG. 3 Normal Distribution and Relationship to Parameters μ and σ

to the right. Fig. 2 shows a symmetric distribution with its density function, $f(x)$, plotted on the left-hand axis and distribution function, $F(x)$, plotted on the right-hand axis.

5.1.4 Referring to the $F(x)$ axis in Fig. 2, observe that $F(30) = 0.5$. The point $x = 30$ divides the distribution into two equal halves with respect to probability (50 % on each side of x). In general, where $F(x) = 0.5$, we call the point x the median or 50th percentile of the distribution. In like manner, we may define any percentile, for example, the 25th or the 90th percentiles. In general, for $0 < p < 1$, a $100p$ % percentile is a location point, Q_p, that divides the distribution into two parts, with $100p$ % lying to the left and $(1 - p)100$ % lying to the right.

5.2 A density function is often given as a equation with one or more parameters, which, when given values, allow the curve to be drawn.[4] For many distributions, two parameters are sufficient (some have one parameter and others have more than two). The parameters may also have meaning with respect to the shape of the curve, the scale used, or some other property of the curve.

5.2.1 The mean or "expected value" of a distribution, denoted by the symbol μ, is a parameter that defines the central location of a distribution. The mean can be thought of as a "center of gravity" for the distribution. When the distribution is symmetric, the mean will coincide with the 50th percentile and occur exactly in the center, splitting the area under the curve into two equal halves of 0.5 each. For right-skewed distributions, the mean will occur to the right of the median; for left-skewed distributions, the mean will occur to the left of the median.

5.2.2 The standard deviation, denoted by the symbol σ, is another important parameter in many distributions. It carries the same units as the variable X, and is also called a scale parameter. Generally, it is a standard measure of variability. The larger the value of σ, the greater will be the variation in the variable X. One of the most important theoretical distributions in statistics is the normal, or Gaussian, distribution. It arises in complex phenomena when many uncontrolled factor effects cause variability and no single effect is of dominating magnitude. The normal distribution is a symmetrical, bell-shaped curve and is completely determined by its mean, μ, and its standard deviation, σ. The parameter μ locates the center, or peak, of the distribution, and the parameter σ determines its spread. The distance from the mean to the inflection point of the curve (maximum slope point) is σ. This is illustrated in Fig. 3.

5.2.3 The probability of obtaining a value in a given interval on the measurement scale is the area under the curve over the interval. This gives some numerical meaning to the parameter σ. Table 1 gives the normal probability for several selected intervals in terms of parameters μ and σ. The first two columns in Table 1 are known as the empirical rule for symmetric and mound-shaped distributions.

5.2.4 The variance of a distribution, σ^2, is the square of the standard deviation. It is the average value of the quantity $(X - \mu)^2$ in the population. It is the variance that is computed first, and then the standard deviation is the positive square root of the variance. For a population specified by a density function, $f(x)$, the theoretical mean and variance are defined mathematically as:

$$\mu = \int_{-\infty}^{\infty} xf(x)dx \tag{3}$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx \tag{4}$$

5.2.5 Here the variable X is assumed to take on all values in the interval (-∞, +∞), but this need not be the case.

5.3 In addition to the mean and standard deviation, measures may be theoretically defined that attempt to describe the general shape of a distribution. Two such quantities are skewness and kurtosis. For a continuous variable, X, skewness is defined as the average value of the quantity $(X - \mu)^3/\sigma^3$, and kurtosis as the average value of the quantity $(X - \mu)^4/\sigma^4$, minus 3. Each of these calculations is taken over the population. The symbols used for the theoretical skewness and kurtosis are γ_1 and γ_2, respectively. For a population specified by a density function, $f(x)$, the theoretical skewness and kurtosis are defined mathematically as:

$$\gamma_1 = \frac{\int_{-\infty}^{\infty} (x - \mu)^3 f(x)dx}{\sigma^3} \tag{5}$$

[4] In the same way a straight line, $y = mx + b$, has "parameters" referred to as the slope, m, and y-intercept, b. Once these parameters are known, the line is completely known and may be drawn precisely.

TABLE 1 Areas Under the Curve for the Normal Distribution

Interval	Area	Interval	Area
μ ± 1σ	0.68270	μ ± 0.674σ	0.50
μ ± 2σ	0.95450	μ ± 1.645σ	0.90
μ ± 3σ	0.99730	μ ± 1.960σ	0.95
μ ± 4σ	0.99994	μ ± 2.576σ	0.99

$$\gamma_2 = \frac{\int_{-\infty}^{\infty} (x-\mu)^4 f(x)dx}{\sigma^4} - 3 \qquad (6)$$

5.3.1 Here again, the variable X is assumed to take on all values in the interval (-∞, +∞).

5.3.2 When a distribution is perfectly symmetric, $\gamma_1 = 0$. This is the case for the normal distribution in Fig. 3. If the distribution has a longer tail on the right, we say that it is right skewed and $\gamma_1 > 0$ as in Fig. 4. If the distribution has a longer tail on the left, we say that it is left skewed and $\gamma_1 < 0$ as in Fig. 5.

5.3.3 For the normal distribution (Fig. 3), $\gamma_2 = 0$. The large base of applications for the normal distribution is the reason for subtracting 3 in the definition of kurtosis. Subtracting of 3 from (6) makes $\gamma_2 = 0$ for the normal distribution. For any distribution the quantity γ_2 cannot be less than –2 (**1**).[5] Several examples of skewness and kurtosis as related to specific distributions are given in Table 2.

5.3.4 Table 2 shows that there is great variation in both skewness and kurtosis for several commonly occurring distributions. Also, for some distributions such as the normal, exponential, and uniform, skewness and kurtosis are constant and not dependent on the value of any other parameter; for others, however, skewness and kurtosis are a function of some other parameter. Here we see that for the Poisson distribution, both γ_1 and γ_2 are functions of the mean, λ. For the Weibull distribution, both γ_1 and γ_2 are functions of the Weibull shape parameter β.

5.4 Statistics is the study of the properties, behavior, and treatment of numerical data. A statistic may be defined as any function of the data values that originate from a sample. In many applications in which one has a specific model in mind, the initial goal is to try to estimate the population (model) parameters using the sample data. These estimates are called descriptive statistics. For example, the sample mean and standard deviation are attempting to estimate the parameters μ and σ, sample skewness and kurtosis are attempting to estimate γ_1 and γ_2, and sample percentiles may be calculated that are attempting to estimate population percentiles. In some cases, there may be more than one statistic that may be used for the same purpose.

5.4.1 In addition to estimation, descriptive statistics serve to organize and give meaning to the raw sample data. By itself a set of numbers in columnar format may yield little useful information. The methods of descriptive statistics include numerical, tabular, and graphical methods that will lead to great insight for the underlying phenomena being studied.

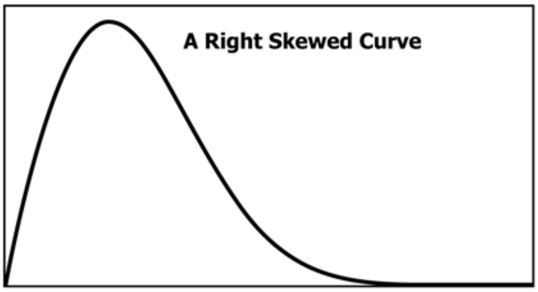

FIG. 4 Curve with Positive Skewness, $\gamma_1 > 0$

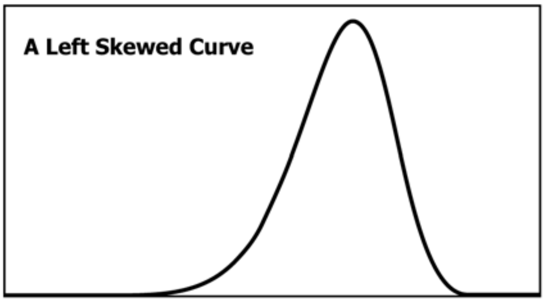

FIG. 5 Curve with Negative Skewness, $\gamma_1 < 0$

TABLE 2 Skewness and Kurtosis for Selected Distribution Forms

Distribution Form	Skewness	Kurtosis
Normal	0	0
Exponential	2	6
Uniform	0	–1.2
Poisson[A]	1/√λ	1/λ
Student's t[B]	0	6/(v – 4)
Weibull[C], β = 3.6	0	–0.28
Weibull, β = 0.5	6.62	84.72
Weibull, β = 50.0	–1	1.9

[A] For the Poisson distribution, λ is the mean.
[B] For the Student's t distribution, v is the degrees of freedom. When $v \le 4$, kurtosis is infinite.
[C] For the Weibull distribution, β is the shape parameter.

6. Descriptive Statistics

6.1 *Mean or Arithmetic Average*—The mean is a measure of centrality or central tendency of a distribution of observations. It is most appropriate for symmetric distributions and is affected by distribution nonsymmetry (shape) and extreme values. The calculation of the mean is the sum of the n sample values divided by the number of values, n. This equation is:

$$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad (7)$$

6.2 *Median or 50th Percentile*—The median is a measure of centrality or central tendency that is generally not affected by the extremes of the distribution. It is a value that divides the distribution into two equal parts. For continuous distributions, 50 % will lie to the left and 50 % to the right of the median. To obtain the 50th percentile of a sample, arrange the n values of a sample in increasing order of magnitude. The median is the $[(n + 1)/2]^{th}$ value when n is odd. When n is even, the median lies between the $(n/2)^{th}$ and the $[(n/2) + 1]^{th}$ values and is not

[5] The boldface numbers in parentheses refer to a list of references at the end of this standard.

TABLE 3 Values of the Constant, d_2, for Converting the Sample Range into an Estimate of Standard Deviation[A]

n	d₂	n	d₂	n	d₂
2	1.128	7	2.704	12	3.258
3	1.693	8	2.847	13	3.336
4	2.059	9	2.970	14	3.407
5	2.326	10	3.078	15	3.472
6	2.534	11	3.173	16	3.532

[A] Source: *ASTM Manual on Presentation of Data and Control Chart Analysis* (**2**).

defined uniquely among the data values. It is then taken to be the arithmetic average of these two values.

6.2.1 As a measure of central tendency, the median is often preferred over the average, particularly for quantities that tend to be skewed in a natural way. Examples include life length of a product, salary, and other monetary quantities or any quantity that has a natural lower or upper bound.

6.3 *Midrange*—Midrange is a measure of central tendency. It is the average of the largest (max) and smallest (min) observed values in a sample of *n* items. It is greatly affected by any outliers in the data set.

6.4 *Max*—The largest observed value in a sample of *n* items.

6.5 *Min*—The smallest observed value in a sample of *n* items.

6.6 *Range*—The difference, *R*, between the largest and smallest observed value in a sample of *n* items is called the sample range and is used as a measure of variation. Its equation is:

$$R = \max(x) - \min(x) \tag{8}$$

6.6.1 The sample range is useful for assessing variation for two basic reasons: (*1*) it is easy to calculate, and (*2*) it is readily understood. But caution is advised when the sample size is modest to large as the min and max then come from the tails of the distribution and can be extremely variable. The sample range is therefore directly affected by extreme values. In general, the standard deviation of a sample is the preferred measure of variation (see 6.12).

6.6.2 The range is particularly useful for small samples, say when $n = 2$ to 12 and there is possibly the burden of calculation, as the standard deviation is more calculation intensive and abstract. An important application occurs when the range is used in quality control applications. For a given sample size, the sample range can be converted into an estimate of the standard deviation. This is done by dividing the range or average range in a group of ranges, by a constant (**2**), d_2, which is the ratio of expected range in a sample of size *n* to standard deviation for a normal distribution. Table 3 contains values of d_2 for sample sizes of 2 through 16.

6.6.3 An important application of this type of estimate for the standard deviation is in quality control charts. When there are available several sample ranges, all with the same sample size, *n*, we take the average range and divide by the appropriate constant, d_2, from Table 3.

6.7 *Order Statistics*—When the observations in a sample are arranged in order of increasing magnitude, the order statistics are:

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \ldots x_{(n-1)} \leq x_{(n)} \tag{9}$$

6.7.1 The bracketed subscript notation indicates that the value is an ordered value. Thus, $x_{(k)}$ is the k^{th} largest value in *n* called the k^{th} order statistic of the sample. This value is said to have a rank of *k* among the sample values. In a sample of size *n*, the smallest observation is $x_{(1)}$ and the largest observation is $x_{(n)}$. The sample range may then be defined in terms of the 1st and n^{th} order statistics:

$$R = x_{(n)} - x_{(1)} \tag{10}$$

6.8 *Empirical Quantiles and Percentiles*—A quantile is a value that divides a distribution to leave a given fraction, *p*, of the observations less than or equal to that value ($0 < p < 1$). A percentile is the same value in which the fraction, *p*, is expressed as a percent, 100*p* %. For example, the 0.5 quantile or 50th percentile (also called the median) is a value such that half of the observations exceed it and half are below it; the 0.75 quantile or 75th percentile is a value such that 25 % of the observations exceed it and 75 % are below it; the 0.9 quantile or 90th percentile is a value such that 10 % of the observations exceed it and 90 % are below it.

6.8.1 The sample estimate of a quantile or percentile is an order statistic or the weighted average of two adjacent order statistics. The i^{th} order statistic in a sample of size *n* is the $i/(n+1)$ quantile or $100i/(n+1)^{th}$ percentile estimate.[6] The quantity $i/(n + 1)$ is referred to as the mean rank for the i^{th} order statistic. In repeated sampling, the expected fraction of the population lying below the i^{th} order statistic in the sample is equal to $i/(n + 1)$ for any continuous population.

6.8.2 To estimate the $100p^{th}$ percentile, compute an approximate rank value using the following equation: $i = (n + 1)p$. If *i* is an integer between 1 and *n* inclusive, then the $100p^{th}$ percentile is estimated as $x_{(i)}$. If *i* is not an integer, then drop the fractional portion and keep the integer portion of *i*. Let *k* be the retained integer portion and *r* be the dropped fractional portion (note that $0 < r < 1$). The estimated $100p^{th}$ percentile is computed from the equation:

$$x_{(k)} + r\left(x_{(k+1)} - x_{(k)}\right) \tag{11}$$

6.8.2.1 *Example*—For a sample of size 20, to estimate the 15th percentile. Calculate $(n + 1)p = 21(0.15) = 3.15$, so $k = 3$ and $r = 0.15$. The 15th percentile is estimated as $x_{(3)} + 0.15(x_{(4)} - x_{(3)})$.

6.9 *Quartile*—The 0.25 quantile or 25th percentile, Q_1, is the 1st quartile. The 0.75 quantile or 75th percentile, Q_3, is the third quartile. The 50th percentile or Q_2, is the 2nd quartile. Note that the 50th percentile is also referred to as the median.

6.10 *Interquartile Range*—The difference between the 3rd and 1st quartiles is denoted as IQR:

$$IQR = Q_3 - Q_1 \tag{12}$$

6.10.1 The IQR is sometimes used as an alternative estimator of the standard deviation by dividing by an appropriate

[6] Several alternatives to the mean rank equation $i/(n + 1)$ are available (**3**), including the median rank and Kaplan-Meier methods. A equation for the exact median rank is available but is computationally intensive. The Behnard approximation equation to the median rank, $(i - 0.3)/(n + 0.4)$, is widely used. The modified Kaplan-Meier equation is $(i - 0.5)/n$.

constant. This is particularly true when several outlying observations are present and may be inflating the ordinary calculation of the standard deviation. The dividing constant will depend on the type of distribution being used. For example, in a normal distribution, the IQR will span 1.35 standard deviations; then dividing the sample IQR by 1.35 will give an estimate of the standard deviation when a normal distribution is used.

6.11 *Variance*—A measure of variation among a sample of *n* items, which is the sum of the squared deviations of the observations from their average value, divided by one less than the number of observations. It is calculated using one of the two following equations:[7]

$$s^2 = \frac{\sum_{i=1}^{n} (x_1 - \bar{x})^2}{n-1} = \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)} \tag{13}$$

6.12 *Standard Deviation*—The standard deviation is the positive square root of the variance.[8] The symbol is *s*. It is used to characterize the probable spread of the data set, but this use is dependent on distribution shape. For mound-shaped distributions that are symmetric, such as the normal form, and modest to large sample size, we may use the standard deviation in conjunction with the empirical rule (see Table 1). This rule states that approximately 68 % of the data will fall within one standard deviation of the mean; 95 % within two standard deviations, and nearly all (99.7 %) within three standard deviations. The approximations improve when the sample size is very large or unlimited and the underlying distribution is of the normal form. The rule is applied to other symmetric mound-shaped distributions based on their resemblance to the normal distribution.

6.13 *Z-Score*—In a sample of *n* distinct observations, every sample value has an associated *Z*-score. For sample value, x_i, the associated *Z*-score is computed as the number of standard deviations that the value x_i lies from the sample mean. Positive *Z*-scores mean that the observation is to the right of the average; negative values mean that the observation is to the left of the average. *Z*-scores are calculated as:

$$Z_i = \frac{(x_i - \bar{x})}{s} \tag{14}$$

6.13.1 Sample *Z*-scores are often useful for comparing the relative rank or merit of individual items in the sample. *Z*-scores are also used to help identify possible outliers in a set of data. There is a much-used rule of thumb that a *Z*-score outside the bounds of ±3 is a possible outlier to be examined for a special cause. Care should be exercised when using this rule, particularly for very small as well as very large sample sizes. For small sample sizes, it is not possible to obtain a *Z*-score outside the bounds of ±3 unless *n* is at least 11. Eq 15 and Table 4 illustrates this theory:

$$|Z_i| \leq (n-1)/\sqrt{n} \tag{15}$$

TABLE 4 Maximum *Z*-Scores Attainable for a Selected Sample Size, *n*

n	3	5	10	11	15	18
Z(n)	1.155	1.789	2.846	3.015	3.615	4.007

6.13.2 Table 4 was constructed using the equation for the maximum (contained in Ref (**4**)).

6.13.3 On the other hand, for very large sample sizes, such as *n* = 250 or more, it is a common occurrence in practice to find at least one *Z*-score outside the range of ±3. Where we can claim a normal distribution is the underlying model, the approximate probability of at least one *Z*-score beyond ±3 is approximately 50 % when the sample size is around 250. At *n* = 300, it is approximately 55 %. A thorough treatment of the use of the sample *Z*-score for detecting possible outlying observations may be found in Practice E178.

6.14 *Coefficient of Variation*—For a non-negative characteristic, the coefficient of variation is the ratio of the standard deviation to the average.

6.15 *Skewness, g_1*—Skewness is a measure of the shape of a distribution. It characterizes asymmetry or skew in a distribution. It may be positive or negative. If the distribution has a longer tail on the right side, the skewness will be positive; if the distribution has a longer tail on the left side, the skewness will be negative. For a distribution that is perfectly symmetrical, the skewness will be equal to 0; however, if the skewness is equal to 0, this does not imply that the distribution is symmetric.[9]

6.16 *Kurtosis, g_2*—Kurtosis is a measure of the combined weight of the tails of a distribution relative to the rest of the distribution.

6.16.1 Sample skewness and kurtosis are given by the equations:

$$g_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{n\, s^3},\ g_2 = \frac{\sum (x_i - \bar{x})^4}{n\, s^4} - 3 \tag{16}$$

6.16.2 Alternative estimates of skewness and kurtosis are defined in terms of *k*-statistics. The *k*-statistic equations have the advantage of being less biased than the corresponding moment estimators. These statistics are defined by:

$$k_1 = \bar{x},\ k_2 = s^2,\ k_3 = \frac{n\sum_{i=1}^{n} (x_i - \bar{x})^3}{(n-1)(n-2)} \tag{17}$$

$$k_4 = \frac{n(n+1)\sum_{i=1}^{n} (x_i - \bar{x})^4}{(n-1)(n-2)(n-3)} - \frac{3\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)^2}{(n-2)(n-3)} \tag{18}$$

6.16.3 From the *k*-statistics, sample skewness and kurtosis are calculated from Eq 19. Notice than when *n* is large, g_1 and g_2 reduce to approximately:

$$g_1 \approx k_3/k_2^{1.5},\ g_2 \approx k_4/k_2^2 \tag{19}$$

[7] These equations are algebraic equivalents, but the second form may be subject to round off error.

[8] When the denominator of the sample variance is taken as *n* instead of *n* – 1, the square root of this quantity is called the root mean squared deviation (RMS).

[9] For example, an *F* distribution having four degrees of freedom in the denominator always has a theoretical skewness of 0, yet this distribution is not symmetric. Also, see Ref (**5**), Chapter 27, for further discussion.

6.16.4 One cannot definitely infer anything about the shape of a distribution from knowledge of g_2 unless we are willing to assume some theoretical distribution such as the Pearson or other distribution family provides.

6.17 *Degrees of Freedom:*

6.17.1 The term 'degrees of freedom' is used in several ways in statistics. First, it is used to denote the number of items in a sample that are free to vary and not constrained in any way when estimating a parameter. For example, the deviations of *n* observations from their sample average must of necessity sum to zero. This property, that $\Sigma(y - \bar{y})=0$, constitutes a *linear constraint* on the sum of the *n* deviations or *residuals* $y_1 - \bar{y}, y_2 - \bar{y}, ..., y_n - \bar{y}$ used in calculating the sample variance, $s^2 = \Sigma(y - \bar{y})^2/(n - 1)$. When any *n*–1 of the deviations are known, the *n*th is determined by this constraint – thus only *n*–1 of the *n* sample values are free to vary. This implies that knowledge of any *n*–1 of the residuals completely determines the last one. The *n* residuals, $y_1 - \bar{y}$, and hence their sum of squares $\Sigma(y_i - \bar{y})^2$ and the sample variance $\Sigma(y - \bar{y})^2/(n - 1)$ are said to have *n*–1 *degrees of freedom*. The loss of one degree of freedom is associated with the need to replace the unknown population mean μ by the sample average $\bar{y}$. Note that there is no requirement that $\Sigma(y_i - \mu)=0$. In estimating a parameter, such as a variance as described above, we have to estimate the mean μ using the sample average ȳ. In doing so, we lose 1 degree of freedom.

6.17.1.1 More generally, when we have to estimate *k* parameters, we lose *k* degrees of freedom. In simple linear regression where there are *n* pairs of data (x_i, y_i) and the problem is to fit a linear model of the form $y = mx + b$ through the data, there are two parameters (*m* and *b*) that must be estimated, and we effectively lose 2 degrees of freedom when calculating the residual variance. The concept is further extended to multiple regression where there are *k* parameters that must be estimated and to other types of statistical methods where parameters must be estimated.

6.17.2 Degrees of freedom are also used as an indexing variable for certain types of probability distributions associated with the normal form. There are three important distributions that use this concept: the Student's *t* and chi-square distributions both use one parameter in their definition. The parameter in each case is referred to as its "degrees of freedom." The F distribution requires two parameters, both of which are referred to as "degrees of freedom." In what follows we assume that there is a process in statistical control that follows a normal distribution with mean μ and standard deviation σ.

6.17.2.1 *Student's t Distribution*—For a random sample of size *n* where ȳ and *s* are the sample mean and standard deviation respectively, the following has a Student's *t* distribution with *n*–1 degrees of freedom:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{20}$$

The *t* distribution is used to construct confidence intervals for means when Σ is unknown and to test a statistical hypothesis concerning means, among other uses.

6.17.2.2 *The Chi-Square Distribution*—For a random sample of size *n* where *s* is the sample standard deviation, the following has a chi-square distribution with *n*–1 degrees of freedom:

$$q = \frac{(n - 1)s^2}{\sigma^2} \tag{21}$$

The chi-square distribution is used to construct a confidence interval for an unknown variance; in testing a hypothesis concerning a variance; in determining the goodness of fit between a set of sample data and a hypothetical distribution; and in categorical data analysis, among other uses.

6.17.2.3 *The F Distribution*—There are two independent samples of sizes n_1 and n_2. In the most common variant the samples are selected from normal distributions having the same standard deviation. In that case the following has an F distribution with n_1–1 and n_2–1 degrees of freedom:

$$F(n_1 - 1, \ n_2 - 1) = \frac{s_1^2}{s_2^2} \tag{22}$$

Both degrees of freedom are required to use the F distribution. It is common to specify one as associated with the numerator and one as associated with the denominator. If the two populations being sampled have differing standard deviations, say σ_1 for population 1 and σ_2 for population 2, then the F ratio above is multiplied by σ_2^2/σ_1^2. The F distribution is used to construct confidence intervals for a ratio of two variances, and in hypothesis testing associated with designed experiments, among other uses.

6.18 *Statistics for Use with Attribute Data:*

6.18.1 *Case 1*—Binomial simple count data occurs in an inspection process in which each inspection unit is classified into one of two dichotomous categories. The population being sampled is either very large relative to the sample or a process (essentially unlimited). Often we use "0" or "1" to stand for the categories. Other designations are: conforming and nonconforming unit or nondefective and defective unit. In all cases, there is a sample size, *n*, and the interest lies in the fraction of nonconforming units in the sample. This fraction is an estimate of the probability, *p*, that a future randomly selected unit will be a nonconforming unit. Often, the population being sampled is conceptual—that is, a process with some unknown nonconforming fraction, *p*.

6.18.1.1 If an indicator variable, *X*, is defined as *X* = 1 when the unit is nonconforming and 0 if not, then the statistic of interest may be defined as:

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{23}$$

6.18.1.2 In some applications, such as in quality control, there are *k* samples each of size *n*. Each sample gives rise to a separate estimate of *p*. Then the statistic of interest may be defined as:

$$\bar{p} = \frac{\sum_{i=1}^{k} P_i}{k} \tag{24}$$

6.18.1.3 The bar over the "p" indicates that this is an average of the sample fractions which estimates the unknown probability p. The binomial distribution is the basis of the p and np charts found in classical quality control applications.

6.18.2 *Case 2—Poisson Simple Count Data*—If an inspection process counts the number of nonconformities or "events" over some fixed inspection area (either a fixed volume, area, time, or spatial interval), the estimate of the mean is identical to the equation in 6.1. We refer to this as the estimate of the mean number of events expected to occur within the interval, volume, area, weight, or time period sampled. The Poisson distribution is the basis of the c and u charts found in classical quality control applications.

6.19 *Standard Error Concept*—When a statistic is calculated from a set of sample data there is usually some population parameter that is of interest and for which the statistic or some simple function therefore serves as the estimate of the parameter. We know that when a second sample is taken, we will not get the same result as the first sample provided. This is because the sample values are different every time a sample is taken. Different sample values will necessarily give us different values for the statistic. A statistic is a random variable subject to variation in repeated sampling. The standard error of the statistic is the standard deviation of the statistic in repeated sampling.

6.19.1 In using or reporting any statistic, it is good practice to also report a standard error for that statistic. This gives the user some idea of the uncertainty in the results being stated. For example, suppose that a sample mean and standard deviation of 29.7 and 2.8 is obtained from a sample of $n = 20$. Suppose further that the sample data originate from a process so that the population is conceptually unlimited. It may be shown that the standard error of the mean (sample average) is specified as:

$$se(\bar{x}) = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} = \frac{2.8}{\sqrt{20}} = 0.63 \qquad (25)$$

6.19.1.1 Here the quantity σ represents the unknown population standard deviation, s is the sample standard deviation and estimates σ, and n is the sample size. In this example, the estimated standard error of the mean is approximately 0.63.

6.19.2 Any standard error calculation or equation will typically be a function of the sample size (as it is for the mean) as well other items such as the kind of distribution being sampled. Tables 5 and 6 contain a short list of commonly required statistics along with associated standard errors

6.19.3 Many other equations for finding or approximating the standard error for a given statistic are available in the literature. When a statistic is complicated to the point at which a closed-form solution or even an approximate equation may be very difficult to find, computer-intensive methodology can be used. Monte Carlo simulation methods are very useful for such purposes. In particular, the technique known as a parametric bootstrap (**6**) uses the original data to generate many new samples (the so-called bootstrap samples) each of the same size n as the original sample. For each bootstrap sample, the statistic of interest is again calculated and saved to a file. Following this process, the standard deviation is calculated for the set of bootstrap estimates, and this number is taken as the standard error.

TABLE 5 Commonly Required Statistics and Their Standard Errors—Data Is of the Variable Type and Population Is Normal

NOTE 1—For skewness and kurtosis,[A] the range for the sample size is $n = 5$ through 1000. The constant c_4 is a function of the sample size n and is widely available in tables. Alternatively, this approximate equation may be used. See Table 7 and Ref (**5**).

Skewness, $g_1 = k_3 / k_2^{1.5}$, let $v = \ln(n)$

$\ln(se) = 0.54 - 0.3718v - 0.01144\, v^2$

Kurtosis, $g_2 = k_4 / s^4$, let $v = \ln(n)$

$\ln(se) = 1.641 - 0.6752v - 0.05498\, v^2 - 0.004492v^3$

Statistic	Estimated Standard Error
Mean	
$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$se(\bar{x}) = \frac{s}{\sqrt{n}}$
Variance	
$s^2 = \frac{\sum_{i=1}^{n} (x_i - x)^2}{n-1}$	$se(s^2) = \sqrt{\frac{2s^4}{n-1}}$
Standard Deviation	$se(s) = s\sqrt{1 - c_4^2}$
$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - x)^2}{n-1}}$	$\approx \frac{s\sqrt{8n-7}}{4n-3}$

[A] The standard error equations for these statistics were determined using a Monte Carlo simulation.

TABLE 6 Commonly Required Statistics and Their Standard Errors—Data Is of the Attribute Type

Statistic	Estimated Standard Error
Binomial Distribution, Mean	
$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$	$se(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}$
Poisson Distribution, Mean	
$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$	$se(\hat{\lambda}) = \sqrt{\hat{\lambda}}$

6.20 *Confidence Intervals*—A confidence interval for an unknown population parameter is constructed using sample data and provides information about the uncertainty of an estimate of that parameter in the form of a probability statement. The confidence interval consists of a set of plausible values for the parameter, bounded by a lower limit (L) and an upper limit (U). The limit values that make up the confidence interval are referred to as confidence limits.

6.20.1 Since the limits of a confidence interval are sample statistics, they will vary in repeated sampling. A confidence interval is said to include, cover or capture the parameter of interest if the upper and lower confidence limits fall on opposite sides of the true parameter value. The probability of

TABLE 7 Values for the Constant, c_4, Used in Calculating the Standard Error of a Sample Standard Deviation When Sampling from a Normal Distribution

n	c_4	n	c_4	n	c_4
		11	0.975350	25	0.989640
2	0.797885	12	0.977559	30	0.991418
3	0.886227	13	0.979406	35	0.992675
4	0.921318	14	0.980971	40	0.993611
5	0.939986	15	0.982316	45	0.994335
6	0.951533	16	0.983484	50	0.994911
7	0.959369	17	0.984506	75	0.996627
8	0.965030	18	0.985410	100	0.997478
9	0.969311	19	0.986214	150	0.998324
10	0.972659	20	0.986934	200	0.998745

this coverage is called the confidence coefficient or confidence level. The term "confidence" refers to the long run fraction of such intervals that would actually cover the parameter in repeating the experiment a large number of times for a fixed value of the parameter. The confidence level is calculated theoretically or by means of computer simulations. Confidence levels are most often expressed as percentages, up to but not including 100 %. Commonly used confidence coefficients are 90 %, 95 %, and 99 %. Generally, the greater the confidence level, the wider (more conservative) will be the confidence interval.

6.20.2 An approximate confidence interval for an unknown parameter, θ, can be expressed in terms of the standard error:

$$\hat{\theta} \pm z_{1-\alpha/2} \times se\left(\hat{\theta}\right) \tag{26}$$

The quantity $\hat{\theta}$ is a statistic, the estimator of the unknown parameter θ; $se(\hat{\theta})$ is an estimate of the standard error of $\hat{\theta}$; and the multiplier $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile selected from the standard normal distribution (5.3) for a $(1 - \alpha)$ two sided confidence interval. For example, when 95 % confidence level is used ($\alpha = 0.05$), $z_{0.975} = 1.960$; when 99 % confidence level is used, $z_{0.995} = 2.576$.

6.20.3 To construct a confidence interval for an unknown proportion, p, using the observed sample proportion $\hat{p}$ from a sample of size n, the general approximate Eq 26 may be used with the standard error as specified in Table 6. For the approximation to be adequate, $n\hat{p}$ and $n(1 - \hat{p})$ should be 5 or more. The equation for this interval is:

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\hat{p}(1-\hat{p})/(n-1)} \tag{27}$$

6.20.4 When the parameter is the mean of a normal distribution, use the standard error estimate in Eq 25 or Table 5 and a multiplier based on Student's t distribution. This gives a theoretically exact confidence interval when the population distribution is a normal curve (5.2.2):

$$\bar{x} \pm t_{1-\alpha/2,\, df}\, s/\sqrt{n} \tag{28}$$

$t_{1\text{-}\alpha/2,\ df}$ is the 1-α/2 quantile of Student's t distribution with df degrees of freedom when the standard deviation s has df degrees of freedom.

6.20.4.1 *Example*—For a sample of size 20, having sample mean 29.7 and sample standard deviation 2.8 (6.19.1), a 95 % confidence interval for the mean is:

$$29.7 \pm 2.093 \times 2.8/\sqrt{20}$$

or 28.4 to 31.0. The multiplier 2.093 comes from a table of Student's t distribution. The confidence interval may be expressed as (28.4, 31.0) or as 29.7 ± 1.3.

6.20.5 One-sided confidence intervals are used when only an upper or a lower bound on the plausible range of values of the parameter is of interest. For example, when the characteristic of interest is the strength of a material, a lower confidence limit can be provided. If the characteristic is a proportion of defective units, and interest is on how large this might be, an upper confidence limit can be provided.

6.20.5.1 *Example*—The lower one-sided 95 % confidence limit for the example of (6.19.1) and (6.20.4.1) is:

$$\bar{x} - t_{1-\alpha,\, df}\, s/\sqrt{n} = 29.7 - 1.729 \times 2.8/\sqrt{20}$$

or 28.6.

6.20.6 Procedures for calculating confidence intervals from sample data are available in textbooks and in the literature for parameters of a variety of distribution functions and for a variety of scenarios (for example, single parameter, difference between two parameters, ratio of two parameters, etc.). Widely available published tables are used to construct confidence intervals for cases involving the binomial, Poisson, exponential and normal distributions. For the common cases as well as others, tables of Student's t, the chi-square and F distributions are required for construction of the interval. Generally, the coverage probability depends on the correctness of the assumed distribution from which the data have arisen.

6.21 *Prediction-Type Intervals for a Normal Distribution*—It may sometimes be the case that we have a sample of n observations from a normal distribution and we want to construct an interval that would contain one or more future observations with some stated confidence C. Such intervals are called prediction intervals.

6.21.1 *Two-Sided Prediction Intervals for a Single Future Value From a Normal Population*—A prediction interval for a single future observation, y, from a normal population is constructed using a sample of n observations from a normal distribution and provides the limits within which the future value is expected to fall with some confidence $C = 1 - \alpha$. We can have both single sided and double sided limits. Let y be the future value. The prediction limits for the two sided interval for the future value are $P_L \leq y \leq P_U$. Equations for these limits are:

$$P_L = \bar{x} - t_{1-\alpha/2}\, s\sqrt{1+1/n} \tag{29}$$

$$P_U = \bar{x} + t_{1-\alpha/2}\, s\sqrt{1+1/n} \tag{30}$$

$t_{1\text{-}\alpha/2}$ is the $1 - \alpha/2$ quantile from Student's t distribution with $n - 1$ degrees of freedom; $\bar{x}$ and s are the sample mean and standard deviation from the original sample of the x values; and the sample size is n. The interval $[P_L, P_U]$ is the region wherein the next observation is expected to fall with confidence $C = 100\,(1 - \alpha/2)$ %.

6.21.2 *Single-Sided Prediction Intervals For a Single Future Value From a Normal Population*—A prediction interval for a single future for the one sided case uses the on of the following forms:

6.21.2.1 For the lower limit use:

$$P_L = \bar{x} - t_{1-\alpha}\, s\sqrt{1+1/n} \tag{31}$$

The future value satisfies $y \geq P_L$ with confidence 100 (1 – α) %. $t_{1\text{-}\alpha}$ is the 1-α quantile from Student's *t* distribution with *n* – 1 degrees of freedom.

6.21.2.2 For the upper limit use:

$$P_U = \bar{x} + t_{1-\alpha} s \sqrt{1+1/n} \tag{32}$$

The future value satisfies $y \leq P_U$ with confidence 100 (1 – α) %. $t_{1\text{-}\alpha}$ is the 1 – α quantile from Student's *t* distribution with *n* – 1 degrees of freedom.

6.21.3 *Prediction Intervals For More Than One Future Value(s) From a Normal Population*—The prediction intervals discussed in 6.21.1 and 6.21.2 can be modified to apply to more than 1 future value. There is only a slight modification to the quantile level for the *t* value used in equations Eq 29-32. When a prediction interval is to apply to *m* future observations using a confidence level of *C* = 1–α, the Student's *t* value is modified as follows.

Use $t_{1-\alpha/(2m)}$ for a two-sided interval.

Use $t_{1-\alpha/(m)}$, for a one-sided interval.

6.21.3.1 The degrees of freedom remain *n* – 1. The modification of the quantile level is an application of the Bonferroni inequality (see Ref (**7**)). Many variations on the theme of prediction intervals are possible. Note that the interval methodology in this section should not be used unless the underlying distribution is normal and stable. For further information on this topic, see Refs (**7, 8,** or **9**).

6.21.4 *Example 1*—A certain type of material tensile strength exhibits a sample mean and standard deviation from a sample of *n* = 7 observations of 17,580 and 795 lbs, respectively. This characteristic has historically been shown to be normally distributed. A two-sided 95 % prediction interval for the tensile strength of the next observation is calculated from Eq 26 and Eq 27. For *n* = 7, use 6 degrees of freedom and a quantile level of 1 – 0.05⁄2 = 0.975. A standard table of Student's *t* values shows that $t_{0.975} = 2.447$. The corresponding prediction interval is:

$$17{,}580 \pm 2.447\,(795)\,\sqrt{1+1/7}$$
$$17{,}580 \pm 2079.7$$
$$\text{The interval is } 15{,}550 \text{ to } 19{,}659.7$$

6.21.5 *Example 2*—For the data in Example 1, calculate a 90 % lower prediction interval for the next 10 individual observations. Use Eq 28 with 6 degrees of freedom. The quantile level for Student's *t* is 1 – 0.10⁄10 = 0.99. The Student's *t* value is therefore $t_{0.99}$ = 3.143. The lower bound for the next 10 individual observations from this normal distribution is:

$$17{,}580 - 3.14\,(795)\,\sqrt{1+1/7}$$
$$17{,}580 - 2671.2$$
$$14{,}908.8$$

The lower bound is therefore 14,909, rounding to the nearest unit. The next 10 individual observations are therefore expected to be at least as large as this with 90 % confidence.

7. Tabular Methods

7.1 Given a set of data, a tabular display called a frequency distribution may be constructed that summarizes the data in terms of what values occur and how often. The frequency distribution consists of several non-overlapping classes or categories. (The terms "cell" or "bin" are also used.) Each class has an upper and lower class boundary, and the class width is defined as the difference between the boundaries for any class (typically, equal class widths are used). Associated with each class is a frequency value that gives the count or frequency of data values in the data set lying within the boundaries of that class. The frequency for a class divided by the total number of observations in the data set defines the relative frequency for that class. Adjacent classes share a common boundary where the upper boundary of one class is the lower boundary of the following class. When possible, class boundaries should be selected so that no data value falls on a boundary. When this is not possible, values falling on a boundary are placed in the class with the larger values.

7.2 To construct the frequency distribution one needs to decide on two quantities: (*1*) the fixed class width and (*2*) the number of classes. Typically, the number of classes in a frequency distribution should be between 13 and 20, but there is no limit to the number of classes that may be defined if the data set is large enough. For data sets of 25 or fewer observations, a frequency distribution will provide little information and is not recommended. There are several rules of thumb available for determining the number of classes in a frequency distribution in preparation for constructing a histogram. For example, there is Sturge's rule, Scott's rule, and the rule of Freedman and Diaconis (**10**). Selection of the number and width of classes is a matter of judgment. Too many classes will create a fragmented view with some classes perhaps empty; too few classes will be too coarse to be of any use. Conventional guidance would suggest between 13 and 20 cells for a number of observations of 250 or more; for less than 250 as few as 10 cells may be used.

7.3 Once the number of classes, *k*, is determined, the class width may be calculated by dividing the range of the data values by *k*. This gives an approximate class width which should be adjusted to a convenient number.

7.4 It is recommended that cell boundaries be chosen using one more significant digit than the data have. In this manner, the problem of deciding which of two adjacent cells to assign a value when that value is equal to the boundary between the two cells will be avoided. For example, suppose that the data values are presented to the nearest tenth of an inch and that a boundary for two cells exists as 74.8. To which class should an actual value of 74.8 be assigned? We can prevent such a question from ever arising by using cell boundaries that have one more significant digit than the data do (in this case, two will do). One should set the boundary between such cells as 74.85. Boundaries between sets of other adjacent cells are similarly adjusted.

7.5 From the core frequency distribution table, a column corresponding to the relative frequency for a class may be easily added by dividing the frequency column by the sample size, *n*. It is often important to report the cumulative behavior of the data, and for such requirements, we can construct a cumulative frequency (CF) column and a cumulative relative frequency (CRF) column. The CF column is constructed from the frequency column by adding the frequencies cumulatively through the several classes. In this process, the cumulative frequency for the last class should be equal to the sample size. The CRF column is equal to the CF column divided by the sample size *n*. The CRF for the last class should be 1.

7.6 These ideas are further illustrated in Section 9.

8. Graphical Methods

8.1 *Histogram*—From the frequency distribution and descriptive statistics for a set of variable data, a number of useful plots may be constructed that greatly aid in the interpretation of the data set. The first and most fundamental graph that may be constructed from the frequency distribution is the frequency or relative frequency histogram. This chart is a bar graph whose bars are typically centered on the midpoints of the class intervals and whose heights are equal to the frequency (or relative frequency) of the class. The bars should be contiguous and of equal width.

8.1.1 The principal information to be derived from such a plot is the estimation of the probability of occurrence between two values. If *a* and *b* are two values of the variable, where $a < b$, then the area contained within the bars between *a* and *b* is proportional to the approximate probability that the value of the variable, *X*, will be observed between *a* and *b*. In theory, this estimate of probability gets better as the sample size increases and as the bar width (class width) shrinks in size; however, any probability estimate will also be a function of the data quality (resolution) and quantity.

8.1.2 The second purpose for constructing a histogram is to assess the general shape of the distribution from which the sample originated. Here the analysis is mostly visual. The histogram may suggest both questions and answers. For example, has the data originated from a symmetrical distribution? Might there be any outliers among our data?

8.2 *Ogive or Cumulative Frequency Distribution*—Often, the interest is in approximating the cumulative probability of occurrence. Using the frequency distribution, a graph constructed with the class upper bounds as the abscissa and the cumulative relative frequency as the ordinate is referred to as an Ogive plot. Start this plot using the lower class bound for the leftmost class plotted against 0. The distribution function, $F(x)$, for the random variable *X* gives the probability that the random variable will be less than or equal to *x*. The Ogive is the integral of the histogram and graphically approximates the true distribution function.

8.2.1 An alternative to the Ogive plot is the empirical distribution function. When the data values are arranged in increasing numerical order, we have constructed the order statistics of the sample. Let $X_{(i)}$ be the i^{th} order statistic. The empirical distribution function is a step function which takes the value *i/n* for values from $X_{(i)}$ up to (but not including) $X_{(i+1)}$. The plot is necessarily less "smooth" than the Ogive. It is more useful for larger data sets, say *n* at least 100.

8.3 *Boxplot*—Another useful plot for depicting distribution shape is the boxplot or "box and whisker" plot. To construct a boxplot, we need four numbers from the sample: the minimum, maximum, 25th, and 75th percentiles. These percentiles will be denoted as the first and third quartiles (Q1 and Q3). The median, Q2, may also be calculated and depicted on the graph. It may also be useful to plot the mean of the data using the symbol "·" or "+" to visualize whether or not the distribution is truly symmetrical.

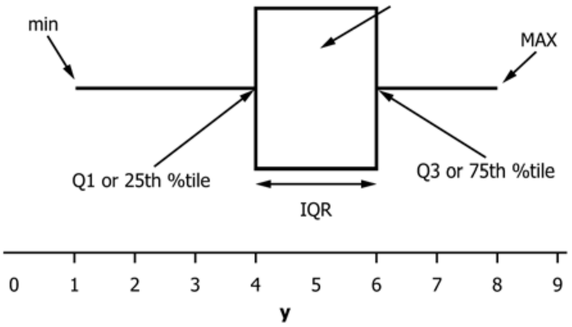

FIG. 6 Boxplot Construction with Horizontal Axis Equal to the Scaled Axis

8.3.1 The boxplot is plotted along one axis (either vertical or horizontal may be used). This axis will be referred to as the scaled axis. The second axis is typically used to identify groups when more than one boxplot is to be presented. This axis will be referred to as the unscaled axis. The boxplot consists of a central box whose dimension along the unscaled axis may be any convenient size. The box dimension along the scaled axis has length equal to the interquartile range, IQR = Q3 – Q1. The leftmost box edge is anchored at Q1 and the rightmost box edge is anchored at Q3. With this construction, the box is said to "contain" the middle 50 % of the data. A line splitting the box is drawn at the value of the median, Q2. From each side of the box along the scaled axis (see Fig. 6) construct a line parallel to the scaled axis. These lines or "whiskers" are continued to the point of the largest and smallest sample values that lie within 1.5 times the IQR from the box edges. Thus, each "whisker" can never exceed 1.5 times the interquartile range. If all sample values are within 1.5 times the IQR of the box edges, whiskers will end at the sample max (on the right) and sample min (on the left).

8.3.2 Any data point exceeding a distance of 1.5 times the IQR from either side (from Q1 or Q3) is plotted using a point plotting symbol, and this indicates that the point is a potential outlier. This rule should be considered as a graphic method to identify potential outliers and not as an outlier test (consult Practice E178 for rigorous outlier tests). If the sample originates from an underlying normal distribution model, the probability of individual points exceeding the 1.5 IQR rule may be derived. For modest to large sample size, these probabilities are large enough that a value outside the 1.5 IQR range is not necessarily an outlier.

8.3.3 Boxplots are particularly useful when several samples are to be compared. The several boxplots can be plotted on a single page using the same scaled axis making for easy graphical comparison. Fig. 7 is a comparison of eight samples of bearing failure time for a certain bearing type using eight different grease formulations. Vertical lines within boxes mark

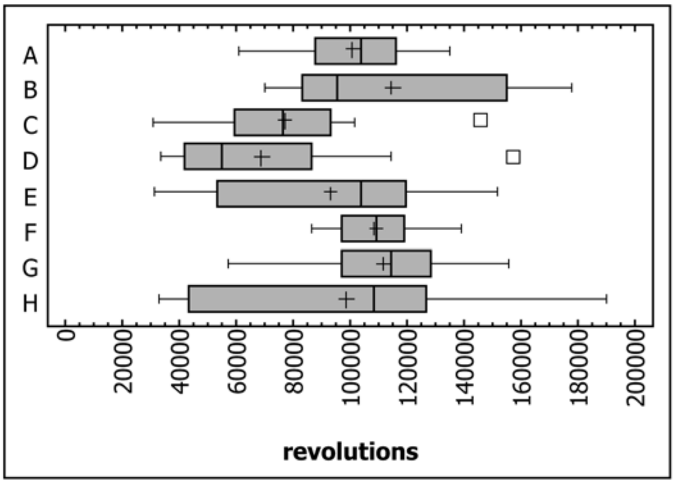

FIG. 7 Bearing Life Data—Illustration of Sample Comparison Using Boxplots, Sample Size, n = 30 Each Group

the median, "+" signs mark the average, and small squares mark points outside of the 1.5 IQR whisker regions for each plot.

8.4 *Dot Plot*—An alternative to the histogram is the dot plot. In a dot plot, the frequency of a class is plotted as a series of stacked dots, as opposed to the bars in a histogram. For large data sets, a single dot may stand for more than one value. The dot plot is also useful for comparing several sample distributions and assessing the density of the data relative to the several classes.

8.5 *Quantile-Quantile (Q-Q) and Probability Plots*—A sample quantile is numerically the same as a sample percentile, but the later is expressed as a percent while the former is expressed as a fraction between 0 and 1. For example, to say that the sample 10th percentile is 106 is to say that the sample quantile of order 0.1 is 106. While the term "percentile" is typically associated with simple descriptive statistics, the term "quantile" plays an important role in graphical methods.

8.5.1 A Q-Q plot may be used to show the relationship between the same quantiles of two samples or to demonstrate that a sample comes from some assumed distribution. In the latter case, the plot is called a probability plot. In probability plotting, we assume some distribution, such as the normal distribution, and plot the sample quantiles against the theoretical quantiles from the assumed distribution. The theoretical quantiles will be a function of the assumed distribution and the sample size.

8.5.2 *Q-Q Plots*—For a given sample of size n, each r^{th} order statistic is a sample quantile of order $r/(n + 1)$, on the average. Note that the r^{th} order statistic is also the $100r/(n + 1)^{th}$ sample percentile. In a quantile-quantile plot, the quantiles from one sample are plotted against the corresponding quantiles of another sample. With two samples of equal size, the order statistics from one sample are plotted against the order statistics of the second sample. If both samples are exactly the same, then the resulting plot will be straight line with slope 1 and y-intercept 0. If the mean of one sample (plotted on the horizontal axis) is shifted to the right, say k units, but otherwise the samples are exactly the same, the resulting plot would be a line of slope 1 and y-intercept $-k$. A slope less than 1 would indicate that the sample plotted as the horizontal coordinate has more variability than the sample plotted as the vertical coordinate. In this manner, fundamental differences between the two samples may be discerned graphically.

8.5.3 When the sample sizes are not equal, we use the smaller sample size to determine the quantiles that are to be plotted. Let two samples be denoted through the variables X and Y; further, let the smaller sample size, n, belong to X, and the larger sample size, m, belong to Y. The n order statistics of the variable X determine the quantiles to be used. These are quantiles of orders $r/(n + 1)$ for $r = 1, 2, \ldots n$. To find the associated quantiles of the same orders from sample of Y values use the method outlined in 6.8. Using this method, two sets of n sample quantiles are determined and may be plotted in manner described previously.

8.5.4 *Probability Plots*—To prepare and use a probability plot, a distribution must be assumed for the variable being studied. Important cases of distributions that are used for this purpose include the normal, log-normal, exponential, Weibull, and extreme value distributions. In most cases, the special probability paper needed for each distribution is readily available or construction is available in a wide variety of software packages. The utility of a probability plot lies in the property that the sample data will generally plot as a straight line given that the assumed distribution is true. From this property, use as an informal and graphic hypothesis test that the sample arose from the assumed distribution is in frequent use.[10] The underlying theory will be illustrated using the normal distribution. Illustrations appear in the section on examples.

8.5.5 *Normal Distribution Case*—Given a sample of n observations assumed to come from a normal distribution with unknown mean and standard deviation (μ and σ), let the variable be Y and the order statistics be $y_{(1)}, y_{(2)}, \ldots y_{(n)}$. Plot the order statistics $y_{(i)}$ against the inverse standard normal distribution function, $\Phi^{-1}(p)$, evaluated at $p = i/(n + 1)$, where $i = 1, 2, 3, \ldots n$. This is because $i/(n + 1)$ is the expected fraction of a population lying below the order statistic $y_{(i)}$ in any sample of size n. The resulting relationship is:

$$y_{(i)} = \Phi^{-1}(i/(n+1))\sigma + \mu \tag{33}$$

8.5.5.1 There is a linear relationship between $y_{(i)}$ and $z = \Phi^{-1}[i/(n + 1)]$ and this establishes a pairing between the ordered y and z values. For example, when a sample of $n = 5$ is used, the z values to use are: –0.967, –0.432, 0, 0.432, and 0.967. Notice that the z values will always be symmetric because of the symmetry of the normal distribution about the median. With the five sample values, form the ordered pairs (y_i, z_i) and plot these on ordinary coordinate paper. If the normal distribution assumption is true, the points will plot as an approximate straight line. The method of least squares may also be used to fit a line through the paired points. When this is done, the slope of the line will approximate the standard deviation. Such a plot is called a normal probability plot.

[10] Formal methods for testing the hypothesis that the data arise from the assumed distribution are available. Such tests include the Anderson-Darling, the Shapiro-Wilks, and a chi-square test among others.

TABLE 8 Breaking Strength in Pounds of Ten Items of 0.104-in. (0.264-cm) Hard-Drawn Copper Wire

578	570	572	570	576
572	568	570	572	584

8.5.5.2 In practice, it is more common to find the cumulative normal probability on the vertical axis instead of the *z* values. With this plot, the normal distribution assumption may be visually verified and estimates of the cumulative probability readily obtained. For this practice, special normal probability paper or widely available software is in use.

8.5.6 *Other Distributions*—The probability plotting technique can be extended to several other types of distributions, most notably the Weibull distribution. In a Weibull probability plot we use the theory that the cumulative distribution function *F(x)* is related to *x* through $F(x) = 1 - \exp(x/\eta)^{\beta}$. Here the quantities η and β are parameters of the Weibull distribution. For a given order statistic $x_{(i)}$ associate the mean rank f_i (or use some other rank method). Algebraic manipulation of the equation for the Weibull distribution function *F(x)*, shows that $\ln\{-\ln(1 - F(x))\} = \beta\ln(x) - \beta\ln(\eta)$. In practice the median rank equation $f_i = (i - 0.3)/(n + 0.4)$ is often used to estimate $F(x_{(i)})$. When the distribution is Weibull, the variables $\ln\{-\ln(1 - F(x))\}$ and $\ln(x_{(i)})$ will plot as an approximate straight line. Other distributions may also be used with this technique.

9. Examples

9.1 *Example 1*—Calculation of descriptive statistics (Table 8).

9.1.1 Mean, variance, and standard deviation calculation. Refer to Table 9.

9.1.1.1 Table 9 contains columns for *X* and X^2 for the data in Table 8. These are used to compute the sample mean, variance, and standard deviation statistics.

9.1.1.2 The mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{5732}{10} = 573.2 \tag{34}$$

9.1.1.3 The variance:

$$s^2 = \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)} \tag{35}$$

$$= \frac{10\,(3{,}285{,}792) - 5732^2}{10\,(9)} = 23.29$$

9.1.1.4 The standard deviation:

$$s = \sqrt{23.29} = 4.83 \tag{36}$$

9.1.2 *Calculation of Order Statistics, Min, Max, and Range*—The order statistics are the items arranged in increasing magnitude. For example, in Table 9 these are: 568, 570, 570, 570, 572, 572, 572, 576, 578, and 584. The smallest of the order statistics is the min, in this case, 568; the largest of the order statistics is the max, in this case, 584. The sample range is max-min = 584 – 568 = 16.

9.1.3 *Calculation of Median and Sample Quartiles:*

TABLE 9 Calculations of the Sample Mean, Variance, and Standard Deviation

item	*X*	*X*²
1	578	334,084
2	572	327,184
3	570	324,900
4	568	322,624
5	572	327,184
6	570	324,900
7	570	324,900
8	572	327,184
9	576	331,776
10	584	341,056
sum	**5732**	**3,285,792**

9.1.3.1 The first quartile is the 25th empirical percentile. When *p* = 0.25 and *n* = 10, *r* = 2.75. The integer portion of *r* is 2 and the fractional portion is 0.75. The 25th empirical percentile is estimated using the second-order statistic and 75 % of the distance between the second and third order statistic. This is:

$$Q_1 = 570 + 0.75\,(570 - 570) = 570 \tag{37}$$

9.1.3.2 When the sample size is even, as here, the 50th percentile or median is the mean of the two middle order statistics. Here this is: (572 + 572)/2 = 572. The third quartile is the 75th empirical percentile. When *p* = 0.75 and *n* = 10, *r* = 8.25. The integer portion of *r* is 8 and the fractional portion is 0.25. This gives for the 75th percentile the eighth order statistic plus 25 % of the distance between the eighth and ninth order statistic. This is:

$$Q_3 = 576 + 0.25\,(578 - 576) = 577.5 \tag{38}$$

9.1.3.3 Suppose we want the 90th percentile of the sample. Then with *p* = 0.9, we find that *r* = 9.9. The 90th empirical percentile is thus equal to the ninth-order statistic and 90 % of the distance between the ninth and tenth order statistics. This is 578 + 0.9 (584 – 578) = 583.4.

9.1.4 The interquartile range is Q3–Q1.

9.1.5 *Five-Number Summary*—It is often useful to present five numbers as a short summary of a set of data. These numbers are called the five-number summary and include the min, Q1, Q2, Q3, and max. Note that the five numbers are also useful in the construction of a box plot.

9.1.6 The sample *Z*-scores or standardized values are computed using the equation in 6.13. The *Z*-scores for the data in Table 8 are shown in Table 10.

9.2 *Example 2: Tabular and Graphical Methods:*

9.2.1 Table 11 contains 270 observations of transverse strength in psi of specimens of brick. Note that the data were recorded to the nearest 10 psi so that any data point has an uncertainty error[11] of at least ±5 psi. (Observe that every number in the table has a 0 as the units' digit place.) In constructing a frequency distribution, we should therefore be advised to round cell boundaries to the nearest 5 psi.

[11] The uncertainty considered here is only related to the significant digits of the reported data and does not include other sources of uncertainty such as measurement error.

TABLE 10 *Z*-Scores Calculated Using the Data from Table 8

item	X	*z*-score
1	578	0.99464
2	572	–0.24866
3	570	–0.66309
4	568	–1.07753
5	572	–0.24866
6	570	–0.66309
7	570	–0.66309
8	572	–0.24866
9	576	0.58021
10	584	2.23794

TABLE 11 Strength of 270 Bricks of a Typical Brand, psi[A]

860	1320	820	1040	1000
920	1100	1250	1480	1150
1200	830	1100	890	270
850	920	940	1310	1330
920	1070	1630	670	1150
1090	700	910	1170	800
830	880	870	1340	840
1040	1080	1040	980	1240
1510	1060	840	940	1110
740	1230	1020	1060	990
1150	860	1100	840	1060
1000	720	800	1170	970
1140	1080	990	570	790
1030	960	870	800	1040
700	860	660	1180	780
920	1100	1080	980	760
860	990	890	940	910
950	880	970	1000	990
1020	750	1070	920	870
1300	970	800	650	1180
890	1030	1060	1610	1190
1080	970	960	1180	1050
910	1100	870	980	730
870	970	910	830	1030
810	1070	1100	460	860
1010	1190	1180	1080	1100
740	1080	860	1000	810
1070	830	1380	960	1360
1020	1390	830	820	980
1170	920	1120	1170	1160
960	1020	1090	2010	890
1180	740	880	790	1100
800	860	1010	1130	970
1240	1290	870	1260	1050
1020	820	1030	860	850
1030	990	1100	1080	1070
690	1020	890	700	880
1070	820	580	820	1060
820	1180	1350	1180	950
1230	950	900	760	1380
830	1220	1100	1090	1380
1100	1020	1380	1010	1030
830	850	630	710	900
1010	1230	780	1000	1150
860	1150	1400	880	730
1400	850	1010	1010	1240
920	1110	780	780	1190
800	800	1140	940	980
1050	710	890	1010	1120
1070	880	1240	940	860
1130	1330	1260	890	980
1000	1090	1140	970	1110
730	930	900	1150	900
1360	910	890	950	1270

[A] Source: *ASTM Manual on Presentation of Data and Control Chart Analysis* (**2**).

9.2.1.1 To determine the number of classes and the class width to use, we first determine the sample range. The range may be determined from the sample min and max as 2010 – 270, yielding a span of 1740 psi. It may be desirable in this case to create a distribution in increments of 100 units. This would give us 18 classes. Keeping in mind that we shall like any cell boundary to have a 5 in its units place, start at some convenient location and add 100 consecutively to create the cell boundaries through the data. For example if we start with 255, the boundaries of the first class would be 255 to 355, the second class 355 to 455, and to forth. In this way, the last class would be 1955 to 2055.

TABLE 12 Frequency Distribution of Brick Strength Data (Table 11)

lower	upper	Freq.	Rel. Freq.	Cume Freq.	CumeRel. Freq.
255	355	1	0.0037	1	0.0037
355	455	0	0.0000	1	0.0037
455	555	1	0.0037	2	0.0074
555	655	4	0.0148	6	0.0222
655	755	16	0.0593	22	0.0815
755	855	37	0.1370	59	0.2185
855	955	56	0.2074	115	0.4259
955	1055	55	0.2037	170	0.6296
1055	1155	50	0.1852	220	0.8148
1155	1255	25	0.0926	245	0.9074
1255	1355	11	0.0407	256	0.9482
1355	1455	9	0.0333	265	0.9815
1455	1555	2	0.0074	267	0.9889
1555	1655	2	0.0074	269	0.9963
1655	1755	0	0.0000	269	0.9963
1755	1855	0	0.0000	269	0.9963
1855	1955	0	0.0000	269	0.9963
1955	2055	1	0.0037	270	1.0000

9.2.1.2 Do not start with the number 250, since this would give boundaries for the first class of 250 to 350, for the second class of 350 to 450, and so forth. In this case, we would not be able to decide on the basis of the boundaries alone to which class a sample value of 350 belongs. This problem is rectified when boundaries are constructed having a "5" in the units place.

9.2.1.3 When this plan is followed, a set of classes and associated frequencies can easily be determined. Once the frequency column is determined, other columns that define the relative frequency, the cumulative frequency, and cumulative relative frequency are also easily determined. Table 12 contains a frequency distribution for the brick strength data set in Table 11.

9.2.1.4 Table 12 meets all the requirements for a frequency distribution: frequencies add up to the sample size, n; relative frequencies add up to 1; the last cumulative frequency is equal to the sample size n = 270; and the last cumulative relative frequency equals 1.

9.2.2 Using the information in the frequency distribution, a histogram and Ogive curve are easily constructed. To construct a frequency histogram for this data, use a bin width of 100 and set the bin left and right boundary according to "lower" and "upper" columns in Fig. 8. The bars should be made to look contiguous as shown in Fig. 9.

9.2.2.1 The Ogive is constructed from the cumulative relative frequency column (CRF) of Table 12. In constructing the Ogive, plot CRF against the "upper" column values that define right boundaries for each class. This plot is illustrated in Fig. 10.

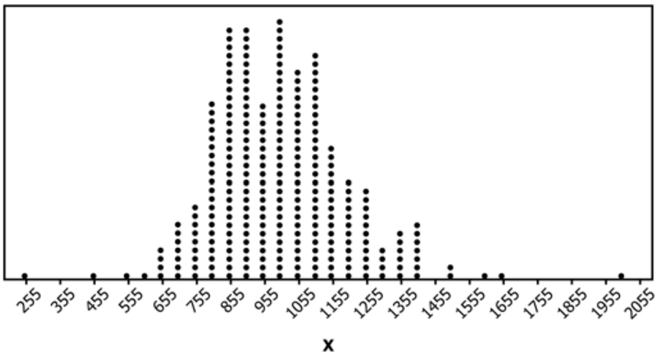

FIG. 8 Dot Plot for Table 11 Data

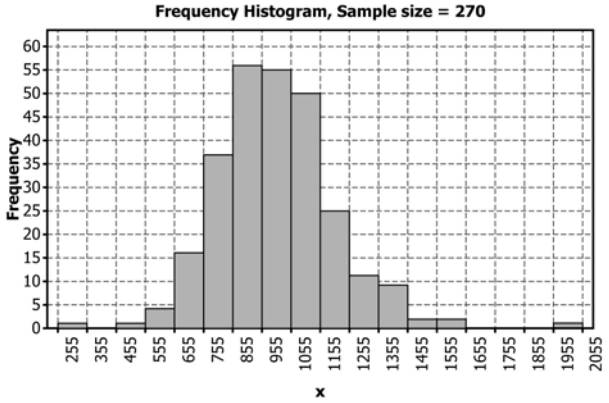

FIG. 9 Frequency Histogram Constructed from Table 12 Information

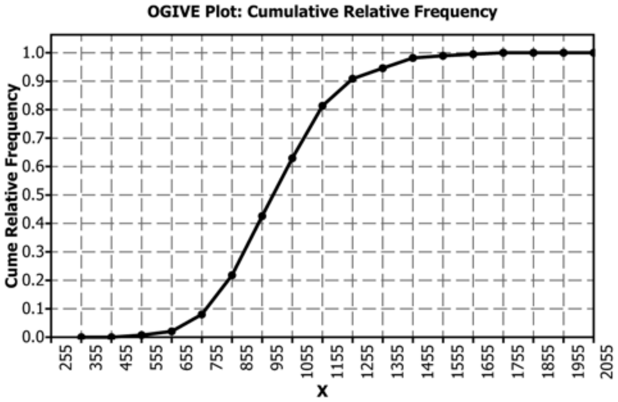

FIG. 10 Relative Frequency Ogive Constructed from Table 12 Information

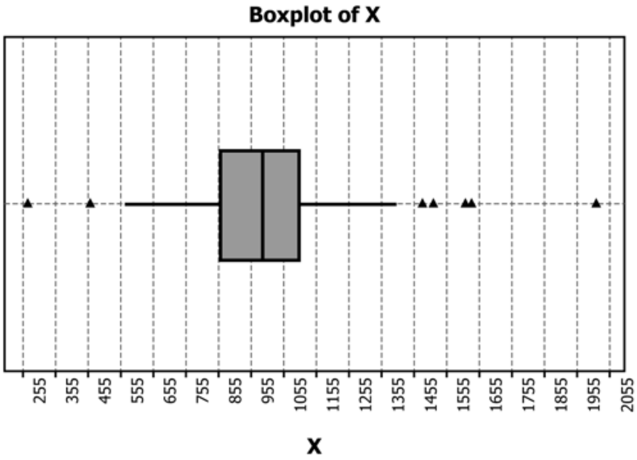

FIG. 11 Boxplot for Table 11 Data

9.2.3 A box plot for these data would look like Fig. 11. Notice that there are several points indicated by the triangle symbol, and this may indicate that these points are potential outliers since they lay more than 1.5 times the IQR from the 25th (to the left) and the 75th (to the right) percentiles.

9.2.4 Fig. 12 is the normal probability plot for the data in Table 11. The plot was constructed using the theory outlined in 8.5.3, except here the probability scale is used. It is also apparent that several outliers may be present.

9.2.5 Fig. 8 is a dot plot of the data in Table 11.

10. Keywords

10.1 boxplot; dot plot; empirical percentile; frequency distribution; histogram; kurtosis; mean; median; midrange; Ogive; order statistic; population parameter; prediction; probability plot; q-q plot; range; sample statistic; skewness; standard deviation; standard error; variance

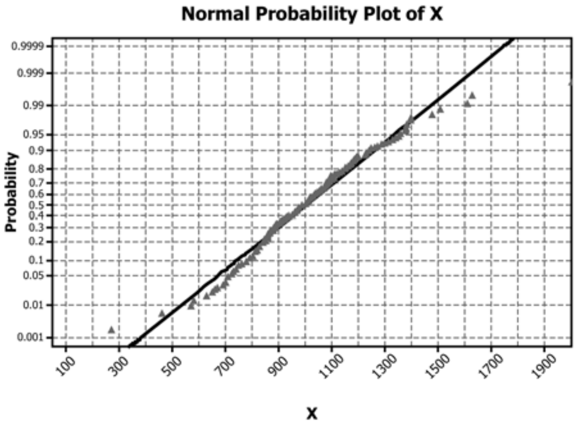

FIG. 12 Normal Probability Plot for Table 11 Data

REFERENCES

(1) Johnson, N.L., and Kotz, S., eds., *Encyclopedia of Statistical Sciences*, Vol 4, s.v. "Kurtosis," Wiley-Interscience, 1983.

(2) *Manual on Presentation of Data and Control Chart Analysis*, 8th ed., ASTM International, West Conshohocken, PA, 2010.

(3) Hyndman, R. J., and Fan, Y., "Sample Quantiles in Statistical Packages," *The American Statistician*, Vol 50, 1996, pp. 361–365.

(4) Shiffler, R. E., "Maximum Z Scores and Outliers," *The American Statistician*, Vol 42, No. 1, February 1988, pp. 79–80.

(5) Duncan, A. J., *Quality Control and Industrial Statistics*, 5th ed., Irwin, Homewood, IL, 1986.

(6) Efron, B. Y., *The Jackknife, Bootstrap and Other Resampling Plans*, Regional Conference Series in Applied Mathematics, No. 38, SIAM, 1982.

(7) Hahn, G., and Meeker, W., *Statistical Intervals, A Guide for Practitioners*, John Wiley & Sons, 1991.

(8) Whitmore, G. A., "Prediction Limits for a Univariate Normal Observation," *The American Statistician*, Vol 40, No. 2, 1986, pp. 141–143.

(9) Hahn, G. J., "Finding an Interval for next Observation from a Normal Distribution," *Journal of Quality Technology*, Vol 1, No. 3, 1969, pp. 168–171.

(10) Wand, M. P., "Data-Based Choice of Histogram Bin Width," *The American Statistician*, Vol 51, No. 1, February 1997, pp. 59–64.

ASTM International takes no position respecting the validity of any patent rights asserted in connection with any item mentioned in this standard. Users of this standard are expressly advised that determination of the validity of any such patent rights, and the risk of infringement of such rights, are entirely their own responsibility.

This standard is subject to revision at any time by the responsible technical committee and must be reviewed every five years and if not revised, either reapproved or withdrawn. Your comments are invited either for revision of this standard or for additional standards and should be addressed to ASTM International Headquarters. Your comments will receive careful consideration at a meeting of the responsible technical committee, which you may attend. If you feel that your comments have not received a fair hearing you should make your views known to the ASTM Committee on Standards, at the address shown below.