Designation: ~~E2586 – 12a~~ E2586 – 12b

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
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:*
3.1.1 Unless otherwise noted, terms relating to quality and statistics are as defined in Terminology E456.
3.1.2 *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**
3.1.3 *coefficient of variation, CV, n*—for a nonnegative characteristic, the ratio of the standard deviation to the mean for a population or sample

[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 ~~Feb. 15, 2012~~Oct. 1, 2012. Published ~~March 2012~~November 2012. Originally approved in 2007. Last previous edition approved in 2012 as ~~E2586 – 12.~~E2586 – 12a. DOI: ~~10.1520/E2586-12A.~~10.1520/E2586-12B.
[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.

3.1.3.1 *Discussion*—

The coefficient of variation is often expressed as a percentage.

3.1.3.2 *Discussion*—

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

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

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

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

3.1.6.1 *Discussion*—

The confidence level, 1 – , 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.

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

3.1.7.1 *Discussion*—

is generally a small number. Confidence level is often 95 % or 99 %.

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

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

~~3.1.9.1 *Discussion*—~~

~~The term 'degrees of freedom' is best defined in the specific context of its use. For a general discussion, the following comments were reprinted from Box, Hunter, and Hunter,~~

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

3.1.11 *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**

3.1.11.1 *Discussion*—

While not required, equal bar or class widths are recommended for histograms.

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

3.1.13 *kurtosis, $_2$, 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,) raised to the fourth power, minus 3 (also referred to as excess kurtosis).

3.1.14 *mean, n*—of a population, μ, average or expected value of a characteristic in a population – *of a sample, x*, sum of the observed values in the sample divided by the sample size.

3.1.15 *median, X , n*—the 50th percentile in a population or sample.

3.1.15.1 *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.

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

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

3.1.17.1 *Discussion*—

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

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

3.1.19 *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.

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

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

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

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

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

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, $_1$, 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,) 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, ,* 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 divided by the sample size minus 1.

3.1.31 *variance, ^{2}, s^2, n*—square of the standard deviation of the population or sample.

3.1.31.1 *Discussion*—

For a finite population, ^{2} is calculated as the sum of squared deviations of values from the mean, divided by *n*. For a continuous population, ^{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.

3.1.32 *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.

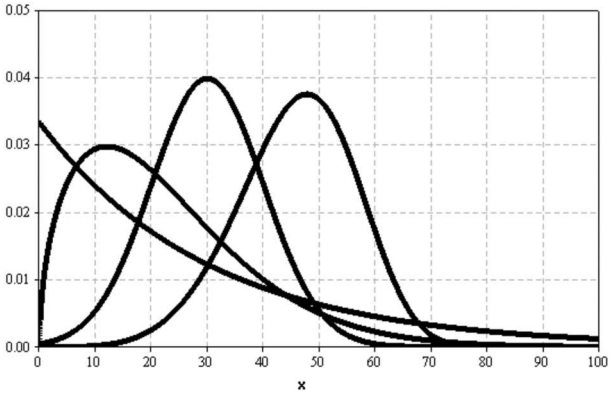

FIG. 1 Probability Density Function—Four Examples of Distribution Shape

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.

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*:

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

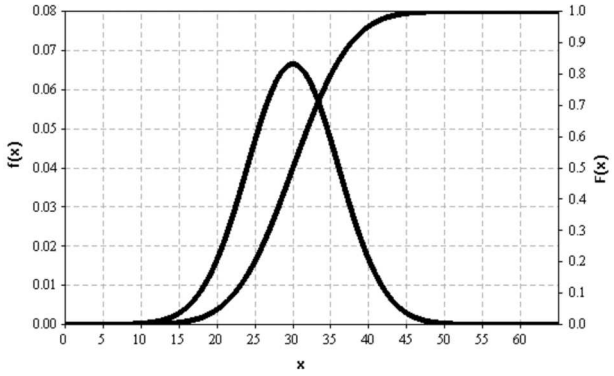

FIG. 2 Cumulative Distribution Function, *F*(*x*), and Density Function, *f*(*x*) Relationship

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 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 , 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 , 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, . The parameter μ locates the center, or peak, of the distribution, and the parameter determines its spread. The distance from the mean to the inflection point of the curve (maximum slope point) is . 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 . Table 1 gives the normal probability for several selected intervals in terms of parameters μ and . 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, ^{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:

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

[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.

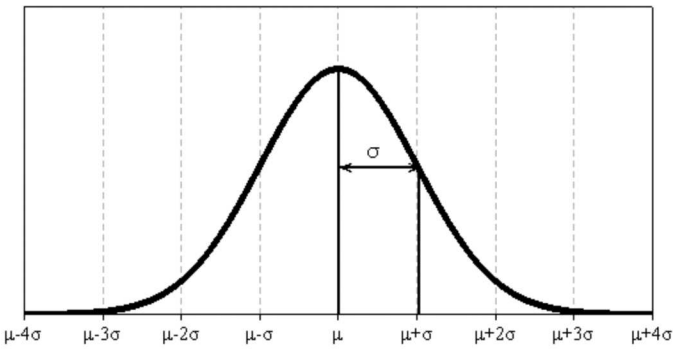

FIG. 3 Normal Distribution and Relationship to Parameters μ and σ

TABLE 1 Areas Under the Curve for the Normal Distribution

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

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

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:

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

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

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, . For the Weibull distribution, both γ_1 and γ_2 are functions of the Weibull shape parameter .

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

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

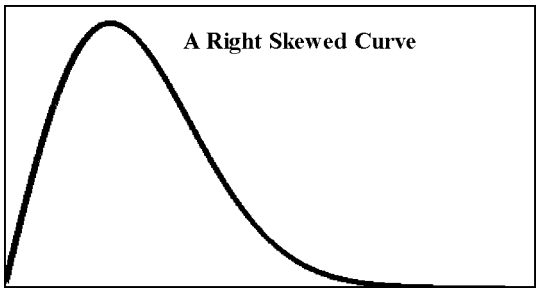

FIG. 4 Curve with Positive Skewness, $\gamma_1 > 0$

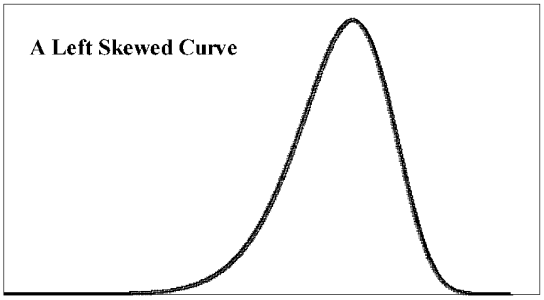

FIG. 5 Curve with Negative Skewness, $\gamma_1 < 0$

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/=	1/
Student's t[B]	0	6/(v – 4)
Weibull[C] , = 3.6	0	–0.28
Weibull, = 0.5	6.62	84.72
Weibull, = 50.0	–1	1.9

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

example, the sample mean and standard deviation are attempting to estimate the parameters μ and , sample skewness and kurtosis are attempting to estimate $_1$ and $_2$, 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.

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:

$$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{7}$$

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

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

n	d_2	n	d_2	n	d_2
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)**.

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:

$$x_{\{1\}} \le x_{\{2\}} \le x_{\{3\}} \le \ldots x_{\{n-1\}} \le x_{\{n\}} \tag{9}$$

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 1^{st} 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, $100p$ %. For example, the 0.5 quantile or 50^{th} 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 75^{th} percentile is a value such that 25 % of the observations exceed it and 75 % are below it; the 0.9 quantile or 90^{th} 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] Several alternatives to the mean rank equation $i/(n + 1)$ are available **(7)**, 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 $(i - 0.5)/n$.

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

6.8.2 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:

$$x_{~k!}1r\sim x_{~k11!}2x_{~k!}! \tag{11}$$

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

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

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

$$IQR\,5\,Q_3\,2\,Q_1 \tag{12}$$

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

$$s^2\,5\,\frac{\left(\sum_{i51}^{n}\sim x_1\,2\,\bar{x}!^2\right)}{n\,2\,1}\,5\,\frac{n\left(\sum_{i51}^{n}x_i^2\,2\left(\sum_{i51}^{n}x_i\right)^2\right)}{n\sim n\,2\,1!} \tag{13}$$

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:

$$Z_i\,5\,\frac{\sim x_i\,2\,\bar{x}!}{s} \tag{14}$$

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 63 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 63 unless n is at least 11. Eq 15 and Table 4 illustrates this theory:

$$?Z_i?\,\leq\sim n\,2\,1!/\bar{n} \tag{15}$$

6.13.2 Table 4 was constructed using the equation for the maximum (contained in Ref. **(3)**).

[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).

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 63. Where we can claim a normal distribution is the underlying model, the approximate probability of at least one *Z*-score beyond 63 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:

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

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:

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

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 $\sum(y - \bar{y}) = 0$, constitutes a *linear constraint* on the sum of the *n* deviations or *residuals* $y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_n - \bar{y}$ used in calculating the sample variance, $s^2 = \sum(y - \bar{y})^2/(n-1)$. 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, $y_i - \bar{y}$, and hence their sum of squares $\sum(y_i - \bar{y})^2$ and the sample variance $\sum(y - \bar{y})^2/(n-1)$ 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 $\sum(y_i - \mu) = 0$. In estimating a parameter, such as a variance as described above, we have to estimate the mean μ using the sample average $\bar{y}$. 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

[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. **(4)**, Chapter 27, for further discussion.