Designation: D 4686 – 91 (Reapproved 2003)

Standard Guide for Identification and Transformation of Frequency Distributions[1]

1. Scope

1.1 This guide gives the rudiments of identification of some of the most common and useful frequency distributions. It does not give rigorous identification. To achieve exactitude, the procedures similar to those given by Shapiro[2] should be used.

1.2 This guide provides a key to identify frequency distributions.

1.3 This guide gives ways to select the proper transformation to use to transform a particular set of data to one which can be modeled by the normal distribution, if such a transformation can be found at all.

1.4 This guide includes the following topics:

2. Referenced Documents

2.1 *ASTM Standards:* [3]

D 123 Terminology Relating to Textiles

D 4392 Terminology for Statistically Related Terms

E 456 Terminology Relating to Quality and Statistics

2.2 *ASTM Adjuncts:*

TEX-PAC[4]

NOTE 1—Tex-Pac is a group of PC programs on floppy disks, available through ASTM Headquarters, 100 Barr Harbor Drive, Conshohocken, PA 19428, USA. Many of the transformations described in this Guide can be made using a program in this adjunct.

3. Terminology

3.1 *Definitions:*

3.1.1 *Bernoulli distribution*—see *binomial distribution.*

3.1.2 *binomial distribution, n*—the frequency distribution which has the probability function:

$$P(r) = (n!/[r!\,(n - r)!])p^r q^{n-r} \tag{1}$$

where:

$P(r)$ = the probability of obtaining exactly r "successes" in n independent trials,

p = the probability, constant from trial to trial, of obtaining a "success" in a single trial, and

q = $1 - p$.

(Syn. *Bernoulli distribution*)

3.1.3 *distribution*—see *frequency distribution of a sample* and *frequency distribution of a population.*

3.1.4 *frequency distribution, n—of a population,* a function that, for a specific type of distribution, gives for each value of a random discrete variate, or each group of values of a random continuous variate, the corresponding probability of occurrence.

3.1.5 *frequency distribution, n—of a sample,* a table giving for each value of a discrete variate, or for each group of values of a continuous variate, the corresponding number of observations.

3.1.6 *normal distribution, n*—the distribution that has the probability function:

$$f(x) = (1/\sigma)(2\pi)^{1/2}\exp[-(x - \mu)^2/2\sigma^2] \tag{2}$$

where:

x = a random variate,

μ = the mean of the distribution, and

σ = the standard deviation of the distribution.

(Syn. *Gaussian distribution, law of error*)

3.1.7 *Poisson distribution, n*—the distribution which has as its probability function:

$$P(r) = e^{-\mu}\mu^r/r! \tag{3}$$

[1] This guide is under the jurisdiction of ASTM Committee D13 on Textiles and is the direct responsibility of Subcommittee D13.93 on Statistics.

Current edition approved July 15, 1991. Published September 1991. Originally published as D 4686 – 87. Last previous edition D 4686 – 90.

[2] Shapiro, Samuel S., *How to Test Normality and Other Distribution Assumptions*, American Society for Quality Control, Milwaukee, WI, 1980. Vol. 3 of the series, *The ASQC Basic References in Quality Control: Statistical Techniques*, Edward J. Dudewitz, ed.

[3] 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.

[4] PC programs on floppy disks are available through ASTM. For a 3½ inch disk request PCN:12-429040-18, for a 5¼ inch disk request PCN:12-429041-18.

where:

$P(r)$ = probability of obtaining exactly r occurrences of an event in one unit, such as a unit of time or area,
μ = both mean and variance of distribution, and
e = base of natural logarithms.

3.1.8 *probability function, n—of a continuous variate,* the mathematical expression whose definite integral gives the probability that a variate will take a value within the two limits of integration.

3.1.9 *probability function, n—of a discrete variate,* the mathematical expression which gives the probability that a variate will take a particular value.

3.1.10 *transformation, n*—the change from one set of variables, x, to another set, y, by the use of a function, $y = f(x)$.

3.1.11 For definitions of textile terms used in this guide, refer to Terminology D 123. For definitions of other statistical terms used in this guide, refer to Terminology D 4392, Terminology E 456, or appropriate textbooks on statistics.

4. Significance and Use

4.1 In measuring, testing, and experimenting, statistical tests are made to determine whether the observed effect of the introduction of a factor is real or simply due to chance. The appropriate statistical test to use depends on the kind of distribution used to model the data. Distribution identification is useful in selecting the most powerful statistical test.

NOTE 2—There are statistical tests which can be used for data for which a parametric distribution cannot be selected. But these non-parametric tests do not discriminate as well as the distribution-dependent tests.

4.2 For certain types of data, a transformation can be made which will make it possible to use the hypothesis that the normal distribution is a suitable model for the transformed data. When this hypothesis can be made, the analysis of the data is made much easier.

5. Key to Distributions

5.1 Table 1 is a key to the identification of frequency distributions. The table consists of a series of pairs of statements. The statement in the pair that is true either (*1*) directs the user to another pair of statements to solicit additional information or (*2*) identifies the distribution.

TABLE 1 Key to Frequency Distributions[A,B]

	Section
1a. Variates are discrete. (2)	
1b. Variates are continuous. (8)	
2a. Variates are counts of events. (3)	
2b. Variates are on an arbitrary scale . . . . .UNIDENTIFIED	10, 11
3a. Counts are of one of a pair of mutually exclusive events per unit. (4)	
3b. Counts are of other events per unit. (6)	
4a. Experiment consists of n identical trials, each conducted independently. (5)	
4b. Experiment otherwise . . . . .UNIDENTIFIED	10, 11
5a. Probability of occurrence of counted event constant from trial to trial . . . . BINOMIAL	6
5b. Probability otherwise . . . . .UNIDENTIFIED	10, 11
6a. Probability of event occurrence constant from unit to unit. (7)	
6b. Probability otherwise . . . . .UNIDENTIFIED	10, 11
7a. Number of events independent from unit to unit . . . . POISSON	7
7b. Number of events otherwise . . . . .UNIDENTIFIED	10, 11
8a. Variates individual observations. (9)	
8b. Variates sample averages . . . . NORMAL	9
9a. Distribution symmetrical. (10)	
9b. Distribution otherwise . . . . .UNIDENTIFIED	10, 11
10a. Distribution unimodal. (11)	
10b. Distribution multimodal . . . . .UNIDENTIFIED	10, 11
11a. Distribution bell shaped . . . . NORMAL	8
11b. Distribution shape otherwise . . . . .UNIDENTIFIED	10, 11

[A] See Section 5 for instructions for using this table.
[B] See Note 1 concerning unidentified distributions.

6. Binomial Frequency Distribution

6.1 A binomial distribution is the probability distribution of a binomial experiment. This fact provides a means for its identification. Such an experiment has the following four characteristics:

6.1.1 The experiment consists of n independent, identical trials. Each trial is conducted independently.

6.1.2 There are only two possible outcomes on each trial. These may be success and failure, good or bad, pass or fail.

6.1.3 The probability of occurrence of one of the two possible outcomes remains the same from trial to trial. This probability is usually denoted by p and the probability of nonoccurrence by $q = (1 - p)$.

6.1.4 The experimenter is interested in r, the number of successes observed during the n trials.

6.2 Counting the number of successes when trying to ignite eight specimens of carpet is one trial from a binomial experiment. Data from a series of such trials will produce a binomial distribution, provided all of the trials are made on essentially the same kind of material and in the same manner.

6.3 *Approximation of the Binomial Distribution by a Normal Distribution*—Binomial data can be transformed to a near normal distribution. (See Section 11.) Under a certain condition, the binomial distribution can be approximated by the normal distribution without transformation. This condition is met when the interval, $p \pm 3\sigma$, or $p \pm 3\sqrt{(p(1-p))}$ does not contain zero or one.

7. Poisson Frequency Distribution

7.1 A Poisson distribution is the probability distribution of a Poisson experiment. This fact provides a means for its identification. Such an experiment has the following four characteristics:

7.1.1 The experiment consists of counting the number of times a particular event occurs in a given unit. The unit may be time, area, volume, a single item, or any other unit of measure.

7.1.2 The probability that an event occurs in a given unit is the same for all units.

7.1.3 The number of events occurring in one unit is independent of the numbers occurring in other units.

7.1.4 The number of potential events per unit is essentially unlimited.

7.2 A count of the number of knots faced on the top of a cone made from a single supply package from a source of essentially uniform supply packages is one trial in a Poisson experiment (see 7.1). A series of such trials will produce data with a Poisson distribution, provided that the trials are all made independently on units from the same material and are made in the same manner.

7.3 *Approximation of the Poisson by a Normal Distribution*—Poisson data can be transformed to a near normal distribution. (See Section 11.) Under a certain condition, the Poisson distribution can be approximated by the normal distribution without transformation. This condition is met when the average, μ, of a Poisson distribution is equal to or greater than nine.

8. Normal Frequency Distribution

8.1 A normal frequency distribution is the symmetrical, bell-shaped probability distribution described by a specific mathematical equation (Eq 2). The equation provides a means for its identification. Measurement data occasionally behave as if they could be modeled by a normal distribution. The characteristics of a normal distribution are:

8.1.1 The distribution is symmetrical about its mode.

8.1.2 The distribution has only one mode.

8.1.3 The distribution's mode, median, and mean all coincide.

8.1.4 The distribution is bell-shaped, and is not peaked or flat.

9. Sample Average Distribution

9.1 The central limit theorem states that the sampling distribution of the average of n observations will be approximately normally distributed regardless of the distribution from which the observations come. The approximation will become more accurate as the value of n becomes larger. Unfortunately, there is no clear-cut solution to the problem of the approximate value of n required. Generally speaking, the central limit theorem serves very well, even for small sample sizes, but not always. In relying on the central limit theorem, it is essential that the average values of observations be used in any analysis, and not individual observations.

10. Other Distributions

10.1 The occurrence of data which can be modeled using the distributions discussed in the previous sections are fairly common. These and other distributions are described along with graphical illustrations in Juran,[5] Figure 22-3.

11. Transformations of Data

11.1 Sometimes raw data come from an underlying distribution which seems to produce non-normal individual observations and averages. In some of these cases a transformation can be made so that the transformed data may be modeled by the normal distribution (See Shapiro[2]). This will allow statistical tests to be made on the transformed data, using the assumption of normality.

11.2 *Transformation of Binomial to Normal Distribution*—To transform a binomial distribution to a near normal distribution, use the following:

$$y = \arcsin\,[(r + 3/8)/(n + 3/4)]^{1/2} \tag{4}$$

11.2.1 *Numerical Example*—A sample of 15 rolls of fabric is taken from every lot. The number of rolls having at least one nonconformance is counted. At the end of a month a frequency distribution is tabulated showing the number of lot samples having 0, 1, 2, ...15 rolls with nonconformances. This frequency distribution would be expected to be binomial with $n = 15$ and r = the number of nonconformances. Table 2 shows r and its transformed value y, using Eq 4.

11.3 *Relationship of Standard Deviation to Average*—Calculate the standard deviation at various averages of adjacent values in an ordered set of observations, or calculate the standard deviations and averages of logically grouped sets of data. Plot the standard deviations against the averages. Estimate the relationship of the standard deviations to the averages. This relationship gives indications of specific transformations to use in those particular cases.

11.3.1 *Standard Deviation Independent of the Mean*—No transformation required.

11.3.2 *Standard Deviation Proportional to the Square Root of the Mean*—When the standard deviation of the observations is proportional to the square root of the mean, then use the transformation:

$$y = \sqrt{x} \tag{5}$$

11.3.2.1 *Transformation of Poisson to Normal Distribution*—In the case of the Poisson distribution, the standard deviation is proportional to the square root of the mean. The approximation to the normal distribution can be improved by using the following transformations:

When the mean value $\mu \leq 3$ use:

$$y = \sqrt{r} + \sqrt{(r = 1)} \tag{6}$$

When the mean value is $3 < \mu < 4$ use:

TABLE 2 Number of Rolls, *r*, in Sample, and the Arc Sine Transformation, *y*, Having at Least One Nonconformance

r	y
0	0.1549
1	0.2999
2	0.3988
3	0.4813
4	0.5551
5	0.6239
6	0.6896
7	0.7536
8	0.8172
9	0.8812
10	0.9469
11	1.0157
12	1.0895
13	1.1720
14	1.2709
15	1.4159

[5] Juran, J. M. ed., *Quality Control Handbook*, Third Edition, McGraw-Hill Book Company, New York, NY, 1974.

$$y = \sqrt{(r + 3/8)} \quad (7)$$

When the mean value μ ≥ 4 use:

$$y = \sqrt{r} \quad (8)$$

11.3.3 *Standard Deviation Proportional to the Mean*—When the standard deviation is proportional to the mean, then use the transformation:

$$y = \log x \quad (9)$$

11.3.3.1 *Measurements of Particle Sizes*—Measurements of particle sizes almost invariably produce data which can be modeled by the normal distribution when the transformation of Eq 5 is used.

11.3.4 *Standard Deviation Proportional to the Square of the Mean*—When the standard deviation of x is proportional to the square of the mean, then use the transformation:

$$y = 1/x \quad (10)$$

11.3.4.1 *Measurements of Rates*—Measurements of rates almost invariably produce data which can be modeled by the normal distribution when the transformation of Eq 10 is used.

11.3.5 *Standard Deviation Increases to Some Value of the Mean then Decreases (Data a Proportion)*—When the standard deviation of x increases to some value of the mean then decreases, then use the transformation:

$$y = \log\left[x/(1 = x)\right] \quad (11)$$

11.4 Almost invariably whenever data are transformed, the resulting data behave as if they require a different model than that for the original data.

11.4.1 It may happen that ends down were counted on one frame during 4-h periods, and reported as ends down per 4 h. These data would be expected to behave as if they could be modeled by a Poisson distribution. Subsequently these data might be transformed to ends down per 1000 h. The transformed data would not behave as a Poisson distribution. Statistical tests made on these transformed data, based on the assumption that they may be modeled by a Poisson distribution, would not be valid. On the other hand, data observed on ends down per 1000 h and reported on that basis could be modeled by a Poisson distribution.