

# Presentation of Data

**PART 1** IS CONCERNED solely with presenting information about a given sample of data. It contains no discussion of inferences that might be made about the population from which the sample came.

## SUMMARY

Bearing in mind that no rules can be laid down to which no exceptions can be found the committee believes that if the recommendations presented are followed, the presentations will contain the essential information for a majority of the uses made of ASTM data.

## RECOMMENDATIONS FOR PRESENTATION OF DATA

Given a sample of  $n$  observations of a single variable obtained under the same essential conditions:

1. Present as a minimum, the average, the standard deviation, and the number of observations. *Always* state the number of observations.
2. Also, present the values of the maximum and minimum observations. Any collection of observations may contain mistakes. If errors occur in the collection of the data, then correct the data values, but do not discard or change any other observations.
3. The average and standard deviation are sufficient to describe the data, particularly so when they follow a Normal distribution.

To see how the data may depart from a Normal distribution, prepare the grouped frequency distribution and its histogram. Also, calculate skewness,  $g_1$ , and kurtosis,  $g_2$ .

4. If the data seem not to be normally distributed, then one should consider presenting the median and percentiles (discussed in Section 6), or consider a transformation to make the distribution more normally distributed. The advice of a statistician should be sought to help determine which, if any, transformation is appropriate to suit the user's needs.
5. Present as much evidence as possible that the data were obtained under controlled conditions.
6. Present relevant information on precisely (a) the field of application within which the measurements are believed valid and (b) the conditions under which they were made.

## GLOSSARY OF SYMBOLS USED IN PART 1

- $f$  Observed frequency (number of observations) in a single bin of a frequency distribution
- $g_1$  Sample coefficient of skewness, a measure of skewness, or lopsidedness of a distribution
- $g_2$  Sample coefficient of kurtosis
- $n$  Number of observed values (observations)
- $p$  Sample relative frequency or proportion, the ratio of the number of occurrences of a given type to the total possible number of occurrences, the ratio of the number of observations in any stated interval to

- the total number of observations; *sample fraction nonconforming* for measured values the ratio of the number of observations lying outside specified limits (or beyond a specified limit) to the total number of observations
- R* *Sample range*, the difference between the largest observed value and the smallest observed value.
- s* *Sample standard deviation*
- s*<sup>2</sup> *Sample variance*
- cv* *Sample coefficient of variation*, a measure of relative dispersion based on the standard deviation (see Sect. 31)
- X* Observed values of a measurable characteristic; specific observed values are designated  $X_1, X_2, X_3$ , etc. in order of measurement, and  $X_{(1)}, X_{(2)}, X_{(3)}$ , etc. in order of their size, where  $X_{(1)}$  is the smallest or minimum observation and  $X_{(n)}$  is the largest or maximum observation in a sample of observations; also used to designate a measurable characteristic
- $\bar{X}$  *Sample average* or *sample mean*, the sum of the  $n$  observed values in a sample divided by  $n$

**NOTE**

The sample proportion  $p$  is an example of a sample average in which each observation is either a 1, the occurrence of a given type, or a 0, the nonoccurrence of the same type. The sample average is then exactly the ratio,  $p$ , of the total number of occurrences to the total number possible in the sample,  $n$ .

If reference is to be made to the population from which a given sample came, the following symbols should be used.

- $\gamma_1$  *Population skewness* defined as the expected value (see Note) of  $(X - \mu)^3$  divided by  $\sigma^3$ . It is spelled and pronounced "gamma one."
- $\gamma_2$  *Population coefficient of kurtosis* defined as the amount by which the expected value (see Note) of  $(X - \mu)^4$  divided by  $\sigma^4$  exceeds or falls short of 3; it is spelled and pronounced "gamma two."
- $\mu$  *Population average* or *universe mean*

defined as the expected value (see Note) of  $X$ ; thus  $E(X) = \mu$ , spelled "mu" and pronounced "mew."

- $p'$  *Population relative frequency*
- $\sigma$  *Population standard deviation*, spelled and pronounced "sigma."
- $\sigma^2$  *Population variance* defined as the expected value (see Note) of the square of a deviation from the universe mean; thus  $E[(X - \mu)^2] = \sigma^2$
- CV* *Population coefficient of variation* defined as the population standard deviation divided by the population mean, also called the *relative standard deviation*, or *relative error*. (see Sect. 31)

**NOTE**

If a set of data is homogeneous in the sense of Section 3 of **PART 1**, it is usually safe to apply statistical theory and its concepts, like that of an *expected value*, to the data to assist in its analysis and interpretation. Only then is it meaningful to speak of a population average or other characteristic relating to a population (relative) frequency distribution function of  $X$ . This function commonly assumes the form of  $f(x)$ , which is the probability (relative frequency) of an observation having exactly the value  $X$ , or the form of  $f(x)dx$ , which is the probability an observation has a value between  $x$  and  $x + dx$ . Mathematically the *expected value* of a function of  $X$ , say  $h(X)$ , is defined as the sum (for discrete data) or integral (for continuous data) of that function times the probability of  $X$  and written  $E[h(X)]$ . For example, if the probability of  $X$  lying between  $x$  and  $x + dx$  based on continuous data is  $f(x)dx$ , then the expected value is

$$\int h(x) f(x) dx = E[h(x)].$$

If the probability of  $X$  lying between  $x$  and  $x + dx$  based on continuous data is  $f(x)dx$ , then the expected value is

$$\Sigma h(x) f(x) dx = E[h(x)].$$

Sample statistics, like  $\bar{X}$ ,  $s^2$ ,  $g_1$ , and  $g_2$ , also have expected values in most practical cases, but these expected values relate to

the population frequency distribution of *entire samples* of  $n$  observations each, rather than of individual observations. The expected value of  $\bar{X}$  is  $\mu$ , the same as that of an individual observation regardless of the population frequency distribution of  $X$ , and  $E(s^2) = \sigma^2$  likewise, but  $E(s)$  is less than  $\sigma$  in all cases and its value depends on the population distribution of  $X$ .

## INTRODUCTION

### 1. Purpose

**PART 1** of the Manual discusses the application of statistical methods to the problem of: (a) condensing the information contained in a sample of observations, and (b) presenting the essential information in a concise form more readily interpretable than the unorganized mass of original data.

Attention will be directed particularly to quantitative information on measurable characteristics of materials and manufactured products. Such characteristics will be termed *quality characteristics*.

### 2. Type of Data Considered

Consideration will be given to the treatment of a sample of  $n$  observations of a single variable. Figure 1 illustrates two general types: (a) the first type is a series of  $n$  observations representing single measurements of the same quality characteristic of  $n$  similar things, and (b) the second type is a series of  $n$  observations representing  $n$  measurements of the same quality characteristic of one thing.

The observations in Figure 1 are denoted as  $X_i$ , where  $i = 1, 2, 3, \dots, n$ . Generally, the subscript will represent the time sequence in which the observations were taken from a process or measurement. In this sense, we may consider the order of the data in Table 1 as being represented in a time-ordered manner.

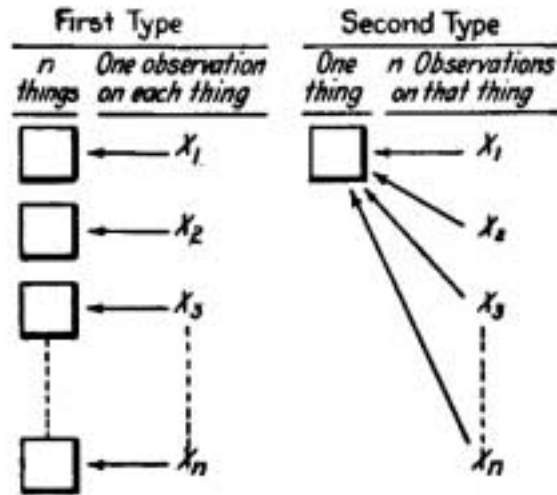


FIG. 1—Two general types of data.

Data of the first type are commonly gathered to furnish information regarding the *distribution* of the quality of the material itself, having in mind possibly some more specific purpose; such as the establishment of a quality standard or the determination of conformance with a specified quality standard, for example, 100 observations of transverse strength on 100 bricks of a given brand.

Data of the second type are commonly gathered to furnish information regarding the errors of measurement for a particular test method, for example, 50-micrometer measurements of the thickness of a test block.

### NOTE

The quality of a material in respect to some particular characteristic, such as tensile strength, is better represented by a frequency distribution function, than by a single-valued constant.

The variability in a group of observed values of such a quality characteristic is made up of two parts: variability of the material itself, and the errors of measurement. In some practical problems, the error of measurement may be large compared with the variability of the material; in others, the converse may be true. In any case, if one is interested in discovering the objective frequency distribution of the quality of the material, consideration must be given to correcting

the errors of measurement (This is discussed in Ref. 1, pp. 379-384, in the seminal book on control chart methodology by Walter A. Shewhart.)

### 3. Homogeneous Data

While the methods here given may be used to condense any set of observations, the results obtained by using them may be of little value from the standpoint of interpretation unless the data are good in the first place and satisfy certain requirements.

To be useful for inductive generalization, any sample of observations that is treated as a single group for presentation purposes should represent a series of measurements, all made under essentially the same test conditions, on a material or product, all of which has been produced under essentially the same conditions.

If a given sample of data consists of two or more subportions collected under different test conditions or representing material produced under different conditions, it should be considered as two or more separate subgroups of observations, each to be treated independently in the analysis. Merging of such subgroups, representing significantly different conditions, may lead to a condensed presentation that will be of little practical value. Briefly, any sample of observations to which these methods are applied should be *homogeneous*.

In the illustrative examples of **PART 1**, each sample of observations will be assumed to be homogeneous, that is, observations from a common universe of causes. The analysis and presentation by control chart methods of data obtained from several samples or capable of subdivision into subgroups on the basis of relevant engineering information is discussed in **PART 3** of this Manual. Such methods enable one to determine whether for practical

**TABLE 1.** Three groups of original data.

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

purposes a given sample of observations may be considered to be homogeneous.

#### 4. Typical Examples of Physical Data

Table 1 gives three typical sets of observations, each one of these datasets represents measurements on a sample of units or

specimens selected in a random manner to provide information about the quality of a larger quantity of material—the general output of one brand of brick, a production lot of galvanized iron sheets, and a shipment of hard drawn copper wire. Consideration will be given to ways of arranging and condensing these data into a form better adapted for practical use.

TABLE 1. Continued.

(b) Weight of Coating of 100 Sheets of Galvanized Iron Sheets, oz/ft <sup>2</sup> <sup>b</sup>					(c) Breaking Strength of Ten Specimens of 0.104-in. Hard- Drawn Copper Wire, lb <sup>c</sup>
1.467	1.603	1.577	1.563	1.437	578
1.623	1.603	1.577	1.393	1.350	572
1.520	1.383	1.323	1.647	1.530	570
1.767	1.730	1.620	1.620	1.383	568
1.550	1.700	1.473	1.530	1.457	572
1.533	1.600	1.420	1.470	1.443	570
1.377	1.603	1.450	1.337	1.473	570
1.373	1.477	1.337	1.580	1.433	572
1.637	1.513	1.440	1.493	1.637	576
1.460	1.533	1.557	1.563	1.500	584
1.627	1.593	1.480	1.543	1.607	
1.537	1.503	1.477	1.567	1.423	
1.533	1.600	1.550	1.670	1.573	
1.337	1.543	1.637	1.473	1.753	
1.603	1.567	1.570	1.633	1.467	
1.373	1.490	1.617	1.763	1.563	
1.457	1.550	1.477	1.573	1.503	
1.660	1.577	1.750	1.537	1.550	
1.323	1.483	1.497	1.420	1.647	
1.647	1.600	1.717	1.513	1.690	

<sup>a</sup> Measured to the nearest 10 psi. Test method used was ASTM Method of Testing Brick and Structural Clay (C 67). Data from *ASTM Manual for Interpretation of Refractory Test Data*, 1935, p. 83.

<sup>b</sup> Measured to the nearest 0.01 oz/ft<sup>2</sup> of sheet, averaged for three spots. Test method used was ASTM Triple Spot Test of Standard Specifications for Zinc-Coated (Galvanized) Iron or Steel Sheets (A 93). This has been discontinued and was replaced by ASTM Specification for General Requirements for Steel Sheet, Zinc-Coated (Galvanized) by the Hot-Dip Process (A 525). Data from laboratory tests.

<sup>c</sup> Measured to the nearest 2 lb. Test method used was ASTM Specification for Hard-Drawn Copper Wire (B 1). Data from inspection report.