NBS Special Publication 700-2
Industrial Measurement Series

Measurement Evaluation

J. Mandel
National Measurement Laboratory
National Bureau of Standards
Gaithersburg, Maryland 20899

and

L. F. Nanni
Federal University
Porto Alegre, Brazil

March 1986

U.S. Department of Commerce
Malcolm Baldrige, Secretary
National Bureau of Standards
Ernest Ambler, Director

.ibrary of Congress
:atalog Card Number: 86-600510
Jational Bureau of Standards
pecial Publication 700-2
Jatl. Bur. Stand. (U.S.),
pec. Publ. 700-2,
0 pages (March 1986)
:ODEN: XNBSAV

U.S. Government Printing Office
Washington: 1986

For sale by the Superintendent
of Documents
U.S. Government Printing Office,
Washington, DC 20402

FOREWORD

When the National Bureau of Standards was established more than 80 years ago, it was given the specific mission of aiding manufacturing and commerce. Today, NBS remains the only Federal laboratory with this explicit goal of serving U.S. industry and science. Our mission takes on special significance now as the country is responding to serious challenges to its industry and manufacturing--challenges which call for government to pool its scientific and technical resources with industry and universities.

The links between NBS staff members and our industrial colleagues have always been strong. Publication of this new Industrial Measurement Series, aimed at those responsible for measurement in industry, represents a strengthening of these ties.

The concept for the series stems from the joint efforts of the National Conference of Standards Laboratories and NBS. Each volume will be prepared jointly by a practical specialist and a member of the NBS staff. Each volume will be written within a framework of industrial relevance and need.

This publication is an addition to what we anticipate will be a long series of collaborative ventures that will aid both industry and NBS.

Ernest Ambler, Director

INTRODUCTION

This paper was published originally as a chapter in the book entitled "Quality Assurance Practices for Health Laboratories".* It is for that reason that the examples used as illustrations are taken from health-related fields of research. However, the statistical concepts and methods presented here are entirely general and therefore also applicable to measurements originating in physics, chemistry, engineering, and other technical disciplines. The reader should have no difficulty in applying the material of this paper to the systems of measurement in his particular field of activity.

J. Mandel
January, 1986

---

* J. Mandel and L.F. Nanni, Measurement Evaluation Quality Assurance Practices for Health Laboratories. Washington: American Public Health Association; 1978: 209-272.1244 p.

ABOUT THE AUTHORS

John Mandel

John Mandel holds an M.S. in chemistry from the University of Brussels. He studied mathematical statistics at Columbia University and obtained a Ph.D in statistics from the University of Eindhoven.

Dr. Mandel has been a consultant on statistical design and data analysis at the National Bureau of Standards since 1947. He is the author of a book, "The Statistical Analysis of Experimental Data", and has contributed chapters on statistics to several others. He has written numerous papers on mathematical and applied statistics, dealing more particularly with the application of statistical methodology to the physical sciences.

Mandel has served as a Visiting Professor at Rutgers University and at the Israel Institute of Technology in Haifa. He has contributed to the educational program of the Chemical Division of the American Society for Quality Control through lectures and courses.

A fellow of the American Statistical Association, the American Society for Testing and Materials, the American Society for Quality Control, and the Royal Statistical Society of Great Britain, Mandel, is the recipient of a number of awards, including the U.S. Department of Commerce Silver Medal and Gold Medal, the Shewhart Medal, the Dr. W. Edwards Deming Medal, the Frank Wilcoxon Prize and the Brumbaugh Award of the American Society for Quality Control.

He was Chairman of one of the Gordon Research Conferences on Statistics in Chemistry and Chemical Engineering and has served on several ASTM committees and is, in particular, an active member of Committee E-11 on Statistical Methods.

Luis F. Nanni

Luis F. Nanni holds a Civil Engineering degree from the National University of Tucuman, Argentina and the M.A. from Princeton University. He was a member of the faculty of Rutgers University School of Engineering for many years and served there as Professor of Industrial Engineering. Professor Nanni also has extensive experience as an industrial consultant on statistics in the chemical sciences, the physical sciences and the health sciences. He is a member of several professional societies including the American Statistical Association, the Institute of Mathematical Statistics, the Operations Research Society of America, the American Institute of Industrial Engineers the American Society for Engineering Education.

Professor Nanni's fields of specialization are statistical analysis and operations research; his scholarly contributions include statistical methods, random processes and simulation, computer programming and engineering analysis. At the present time he is Professor of Civil Engineering at the Federal University in Porto Alegre, Brazil.

CONTENTS

Measurement Evaluation

J. Mandel (*principal author*), and L. F. Nanni.

Basic Statistical Concepts

Random variables

This chapter is concerned with the evaluation of measurements *by means of statistical methods*. This qualification is important, for the total evaluation of measurements involves many different points of view. What differentiates the statistical viewpoint from all others is that each measurement is considered as only one realization of a hypothetical infinite population of similar measurements. Although, in general, all members of this population refer to the measurements of the same property on the same sample (e.g., the glucose content of a given sample of serum), they are not expected to be identical. The differences among them are attributable to chance effects, due to unavoidable fluctuations in many of the conditions surrounding the measuring process. Alternatively, the members of the population of measurements may refer to different samples, or different individuals. Thus, one may consider the glucose content of serum of all healthy individuals in a certain age range. In such cases, the observed differences among the measured values include what is referred to as *sampling error*, meaning the differences in the measured property among the members of the population of samples or individuals. A variable whose value is associated with a statistical population is called a *random variable* or *variate*.

Frequency distribution and histograms

A mathematical representation can be made of a statistical population, such as the hypothetical infinite population of measurements just mentioned. To obtain this representation, called a *frequency distribution*, one divides all the measurements in the population into group intervals and counts the number of measurements in each interval. Each interval is defined in terms of its lower and upper limit, in the scale in which the measurement is expressed. Since in practice one is always limited to a statistical sample, i.e., a finite number of measurements, one can at best only approximate the frequency distribution. Such an approximation is called a *histogram*. Figure 4.1 contains a histogram of glucose values in serum measurements on a sample of 2,197 individuals. It is worth noting that the frequency tends to be greatest in the vicinity of the mean and diminishes gradually as the distance from the mean

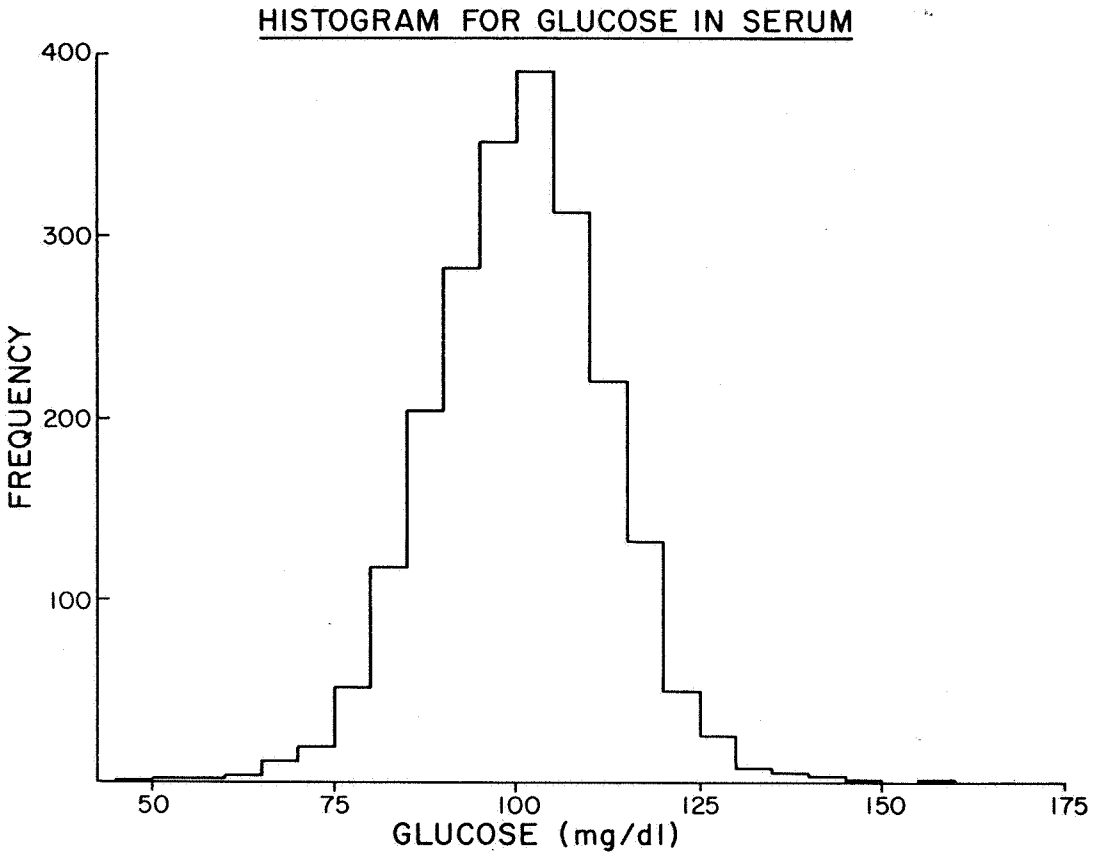

Fig. 4.1. Histogram of glucose serum values on a sample of 2,197 individuals, with a range of 47.5–157.5 mg/dl and a mean of 100.4 mg/dl.

increases. The grouped data on which the histogram is based are given in Table 4.1.

Population parameters and sample estimates

Random samples

The sample of individuals underlying the histogram in Table 4.1 is rather large. A large size, in itself, does not necessarily ensure that the histogram's characteristics will faithfully represent those of the entire population. An additional requirement is that the sample be obtained by a *random selection* from the entire population. A random selection is designed to ensure that each element of the population has an equal chance of being included in the sample. A sample obtained from a random selection is called a *random sample*, although, strictly speaking, it is not the sample but the method of obtaining it that is random. Using the concept of a random sample, it is possible to envisage the population as the limit of a random sample of ever-increasing size. When the sample size N becomes larger and larger, the characteristics of the sample approach those of the entire population. If the random sample is as large as the sample used in this illustration, we may feel confident that its characteristics are quite similar to those of the population.

TABLE 4.1. GROUPED DATA FOR GLUCOSE IN SERUM

Glucose (mg/dl)	Number of individuals	Glucose (mg/dl)	Number of individuals
47.5	1	107.5	313
52.5	2	112.5	220
57.5	2	117.5	132
62.5	3	122.5	50
67.5	12	127.5	26
72.5	20	132.5	8
77.5	52	137.5	6
82.5	118	142.5	4
87.5	204	147.5	1
92.5	281	152.5	0
97.5	351	157.5	1
102.5	390		
		Total number of individuals:	2,197

Thus, upon inspection of Table 4.1, we may feel confident that the mean serum glucose for the entire population is not far from 100.4 mg/dl. We also may feel confident in stating that relatively very few individuals, say about 1 percent of the entire population, will have serum glucose values of less than 70 mg/dl. Our confidence in such conclusions (which, incidentally, can be made more quantitative), however, would have been much less had all of the available data consisted of a small sample, say on the order of five to 50 individuals. Two such sets of data are shown in Table 4.2. Each represents the serum glucose of ten individuals from the population represented in Table 4.1. The mean glucose contents of these samples are 107.57 and 96.37 mg/dl, respectively. If either one of these samples was all the information available

TABLE 4.2. TWO SMALL SAMPLES OF GLUCOSE IN SERUM

Sample I		Sample II	
Individual	Glucose (mg/dl)	Individual	Glucose (mg/dl)
1	134.2	1	88.2
2	119.6	2	82.0
3	91.9	3	96.0
4	96.6	4	94.1
5	118.8	5	96.3
6	105.2	6	108.8
7	103.4	7	106.3
8	112.1	8	101.1
9	97.0	9	89.4
10	96.9	10	101.7
Average	107.57	Average	96.37
Variance	179.44	Variance	70.48
Standard deviation	13.40	Standard deviation	8.40

to us, what could we have concluded about the mean serum glucose of the entire population? And, in that case, what could we have stated concerning the percentage of the population having a serum glucose of less than 70 mg/dl?

Population parameters—general considerations

The answer to these and similar questions requires that we first define some basic characteristics of a statistical sample and relate them to the characteristics of the population. Fortunately, most populations can be characterized in terms of very few quantities, called *parameters*. In many cases, only *two* parameters are required, in the sense that these two parameters contain practically all the pertinent information that is required for answering all useful questions about the population. In cases where more than two parameters are needed, it is often possible to perform a mathematical operation, called a *transformation of scale*, on the measured values, which will reduce the required number of parameters to two. The two parameters in question are the *mean* and the *standard deviation*, measuring, respectively, the location of the center of the population and its spread.

Sample estimates

Let $x_1, x_2, \ldots, x_N$ represent a sample of N measurements belonging to a single population. The *sample mean* is generally denoted by $\bar{x}$ and defined by

$$\bar{x} \equiv \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\Sigma x_i}{N} \tag{4.1}$$

The *sample variance* is denoted by s_x^2 and defined by

$$s_x^2 \equiv \frac{\Sigma(x_i - \bar{x})^2}{N - 1} \tag{4.2}$$

The *sample standard deviation* is denoted by s_x and defined by

$$s_x \equiv \sqrt{s_x^2} \tag{4.3}$$

Table 4.2 contains, for each of the samples, the numerical values of $\bar{x}$, s_x^2, and s_x.

Population parameters as limiting values of sample estimates

The quantities defined by Equations 4.1, 4.2, and 4.3 are not the population parameters themselves but rather are *sample estimates* of these parameters. This distinction becomes apparent by the fact that they differ from sample to sample, as seen in Table 4.2. However, it is plausible to assume that as the sample size N becomes very large, the sample estimates become more and more stable and eventually approach the corresponding population parameters. We thus define three new quantities: the *population mean*, denoted by the symbol μ; the *population variance*, denoted by the symbol σ_x^2 or by the symbol $\mathrm{Var}(x)$; and the *population standard deviation*, denoted by σ_x. Thus:

$$\sigma_x = \sqrt{\sigma_x^2} \quad = \sqrt{\text{Var}(x)} \tag{4.4}$$

It is customary to denote population parameters by Greek letters (e.g., μ, σ) and sample estimates by Latin letters (e.g., $\bar{x}$, s). Another often used convention is to represent sample estimates by Greek letters topped by a *caret* (^); thus s and $\hat{\sigma}$ both denote a sample estimate of σ. It is apparent from the above definitions that the variance and the standard deviation are not two independent parameters, the former being the square of the latter. In practice, the standard deviation is the more useful quantity, since it is expressed in the same units as the measured quantities themselves (mg/dl in our example). The variance, on the other hand, has certain characteristics that make it theoretically desirable as a measure of spread. Thus, the two basic parameters of a population used in laboratory measurement are: (a) its mean, and (b) either its variance or its standard deviation.

Sums of squares, degrees of freedom, and mean squares

Equation 4.2 presents the sample variance as a ratio of the quantities $\Sigma(x_i - \bar{x})^2$ and $(N - 1)$. More generally, we have the relation:

$$MS = \frac{SS}{DF} \tag{4.5}$$

where *MS* stands for *mean square*, *SS* for *sum of squares*, and *DF* for *degrees of freedom*. The term "sum of squares" is short for "sum of squares of deviations from the mean," which is, of course, a literal description of the expression $\Sigma(x_i - \bar{x})^2$, but it is also used to describe a more general concept, which will not be discussed at this point. Thus, Equation 4.2 is a special case of the more general Equation 4.5.

The reason for making the divisor $N - 1$ rather than the more obvious N can be understood by noting that the N quantities

$$x_1 - \bar{x},\ x_2 - \bar{x},\ \ldots,\ x_N - \bar{x}$$

are not completely independent of each other. Indeed, by summing them we obtain:

$$\sum_i (x_i - \bar{x}) = \Sigma x_i - \Sigma \bar{x} = \Sigma x_i - N\bar{x} \tag{4.6}$$

Substituting for $\bar{x}$ the value given by its definition (Equation 4.1), we obtain:

$$\sum_i (x_i - \bar{x}) = \Sigma x_i - N\,\frac{\Sigma x_i}{N} = 0 \tag{4.7}$$

This relation implies that if any $(N - 1)$ of the N quantities $(x_i - \bar{x})$ are given, the remaining one can be calculated without ambiguity. It follows that while there are N independent measurements, there are only $N - 1$ independent deviations from the mean. We express this fact by stating that the sample variance is based on $N - 1$ *degrees of freedom*. This explanation provides at least an intuitive justification for using $N - 1$ as a divisor for the calculation of s^2. When N is very large, the distinction between N and $N - 1$ becomes unimportant, but for reasons of consistency, we always define the

sample variance and the sample standard deviation by Equations 4.2 and 4.3.

Grouped data

When the data in a sample are given in grouped form, such as in Table 4.1, Equations 4.1 and 4.2 cannot be used for the calculation of the mean and the variance. Instead, one must use different formulas that involve the midpoints of the intervals (first column of Table 4.1) and the corresponding frequencies (second column of Table 4.1).

Formulas for grouped data are given below.

To differentiate the regular average (Equation 4.1) of a set of x_i values from their "weighted average" (Equation 4.8), we use the symbol $\tilde{x}$ (x tilde) for the latter.

$$\tilde{x} = \frac{\sum_i f_i x_i}{\sum f_i} \tag{4.8}$$

$$s_x^2 = \frac{\sum_i f_i (x_i - \tilde{x})^2}{\left(\sum f_i\right) - 1} \tag{4.9}$$

$$s_x = \sqrt{s_x^2} \tag{4.10}$$

where f_i (the "frequency") represents the number of individuals in the ith interval, and x_i is the interval midpoint. The calculation of a sum of squares can be simplified by "coding" the data prior to calculations. The coding consists of two operations:

1) Find an approximate central value x_o (e.g., 102.5 for our illustration) and subtract it from each x_i.
2) Divide each difference $x_i - x_o$ by a convenient value c, which is generally the width of the intervals (in our case, $c = 5.0$).

Let the mean

$$u_i = \frac{x_i - x_o}{c} \tag{4.11}$$

The weighted average $\bar{u}$ is equal to $(\tilde{x} - x_o)/c$. Operation (1) alters neither the variance nor the standard deviation. Operation (2) divides the variance by c^2 and the standard deviation by c. Thus, "uncoding" is accomplished by multiplying the variance of u by c^2 and the standard deviation of u by c. The formulas in Equations 4.8, 4.9, and 4.10 are illustrated in Table 4.3 with the data from Table 4.1.

We now can better appreciate the difference between population parameters and sample estimates. Table 4.4 contains a summary of the values of the mean, the variance, and the standard deviation for the population (in this case, the very large sample $N = 2{,}197$ is assumed to be identical with the population) and for the two samples of size 10.

TABLE 4.3. CALCULATIONS FOR GROUPED DATA

x	u	f	x	u	f
47.5	−11	1	107.5	1	313
52.5	−10	2	112.5	2	220
57.5	−9	2	117.5	3	132
62.5	−8	3	122.5	4	50
67.5	−7	12	127.5	5	26
72.5	−6	20	132.5	6	8
77.5	−5	52	137.5	7	6
82.5	−4	118	142.5	8	4
87.5	−3	204	147.5	9	1
92.5	−2	281	152.5	10	0
97.5	−1	351	157.5	11	1
102.5	0	390			
	$\bar{u} = -0.4156$		$\bar{x} = 102.5 + 5\bar{u} = 100.42$		
	$s_u^2 = 5.9078$		$s_x^2 = 25s_u^2 = 147.70$		
	$s_u = 2.4306$		$s_x = 5s_u = 12.15$		

We first deal with the question: "How reliable is a sample mean as an estimate of the population mean?" The answer requires the introduction of two important concepts—the *standard error of the mean* and the method of *confidence intervals*. Before introducing the latter, however, it is necessary to discuss *normal distribution*.

Standard error of the mean

The widely held, intuitive notion that the average of several measurements is "better" than a single measurement can be given a precise meaning by elementary statistical theory.

Let $x_1, x_2, \ldots, x_N$ represent a sample of size N taken from a population of mean μ and standard deviation σ.

Let $\bar{x}_1$ represent the average of the N measurements. We can visualize a repetition of the entire process of obtaining the N results, yielding a new average $\bar{x}_2$. Continued repetition would thus yield a series of averages $\bar{x}_1, \bar{x}_2, \ldots$. (Two such averages are given by the sets shown in Table 4.2). These averages generate, in turn, a new population. It is intuitively clear, and can readily be proved, that the mean of the population of averages is the same as that of the population of single measurements, i.e., μ. On the other hand, the

TABLE 4.4. POPULATION PARAMETER AND SAMPLE ESTIMATES (DATA OF TABLES 4.1 AND 4.2)

Source	Mean (mg/dl)	Variance (mg/dl)^2	Standard Deviation (mg/dl)
Population[a]	100.42	147.70	12.15
Sample I	107.57	179.55	13.40
Sample II	96.37	70.56	8.40

[a]We consider the sample of Table 4.1 as identical to the population.

variance of the population of averages can be shown to be smaller than that of the population of single values, and, in fact, it can be proved mathematically that the following relation holds:

$$\mathrm{Var}(\bar{x}) = \frac{\mathrm{Var}(x)}{N} \tag{4.12}$$

From Equation 4.12 it follows that

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}} \tag{4.13}$$

This relation is known as the law of the *standard error of the mean*, an expression simply denoting the quantity $\sigma_{\bar{x}}$. The term *standard error* refers to the variability of *derived* quantities (in contrast to original measurements). Examples are: the mean of N individual measurements and the intercept or the slope of a fitted line (see section on straight line fitting). In each case, the derived quantity is considered a random variable with a definite distribution function. The standard error is simply the standard deviation of this distribution.

Improving precision through replication

Equation 4.13 justifies the above-mentioned, intuitive concept that averages are "better" than single values. More rigorously, the equation shows that the *precision* of experimental results can be improved, in the sense that the *spread* of values is reduced, by taking the average of a number of replicate measurements. It should be noted that the improvement of precision through averaging is a rather inefficient process; thus, the reduction in the standard deviation obtained by averaging ten measurements is only $\sqrt{10}$, or about 3, and it takes 16 measurements to obtain a reduction in the standard deviation to one-fourth of its value for single measurements.

Systematic errors

A second observation concerns the important assumption of *randomness* required for the validity of the law of the standard error of the mean. The N values must represent a *random* sample from the original population. If, for example, *systematic* errors arise when going from one set of N measurements to the next, these errors are not reduced by the averaging process. An important example of this is found in the evaluation of results from different laboratories. If each laboratory makes N measurements, and if the within-laboratory replication error has a standard deviation of σ, the standard deviation between the *averages* of the various laboratories will generally be *larger* than $\sigma/\sqrt{N}$, because additional variability is generally found between laboratories.

The normal distribution

Symmetry and skewness

The mean and standard deviation of a population provide, in general, a great deal of information about the population, by giving its central location

and its spread. They fail to inform us, however, as to the exact way in which the values are distributed around the mean. In particular, they do not tell us whether the frequency or occurrence of values smaller than the mean is the same as that of values larger than the mean, which would be the case for a *symmetrical* distribution. A nonsymmetrical distribution is said to be *skew*, and it is possible to define a parameter of skewness for any population. As in the case of the mean and the variance, we can calculate a sample estimate of the population parameter of skewness. We will not discuss this matter further at this point, except to state that even the set of three parameters, mean, variance, and skewness, is not always sufficient to completely describe a population of measurements.

The central limit theorem

Among the infinite variety of frequency distributions, there is one class of distributions that is of particular importance, particularly for measurement data. This is the class of *normal*, also known as Gaussian, distributions. All normal distributions are symmetrical, and furthermore they can be reduced by means of a simple algebraic transformation to a single distribution, known as the *reduced normal distribution*. The practical importance of the class of normal distributions is related to two circumstances: (a) many sets of data conform fairly closely to the normal distribution; and (b) there exists a mathematical theorem, known as the *central limit theorem*, which asserts that under certain very general conditions the process of averaging data leads to normal distributions (or very closely so), regardless of the shape of the original distribution, provided that the values that are averaged are independent random drawings from the same population.

The reduced form of a distribution

Any *normal* distribution is completely specified by two parameters, its mean and its variance (or, alternatively, its mean and its standard deviation).

Let x be the result of some measuring process. Unlimited repetition of the process would generate a population of values $x_1, x_2, x_3, \ldots$. If the frequency distribution of this population of values has a mean μ and a standard deviation of σ, then the change of scale effected by the formula

$$z = \frac{x - \mu}{\sigma} \tag{4.14}$$

will result in a new frequency distribution of a mean value of *zero* and a standard deviation of *unity*. The z distribution is called the *reduced* form of the original x distribution.

If, in particular, x is normal, then z will be normal too, and is referred to as the *reduced normal distribution*.

To understand the meaning of Equation 4.14, suppose that a particular measurement x lies at a point situated at k standard deviations above the mean. Thus:

$$x = \mu + k\sigma$$

Then, the corresponding z value will be given by

$$z = \frac{(\mu + k\sigma) - \mu}{\sigma} = k$$

Thus the z value simply expresses the distance from the mean, in units of standard deviations.

Some numerical facts about the normal distribution

The following facts about *normal* distributions are noteworthy and should be memorized for easy appraisal of numerical data:

1) In any normal distribution, the fraction of values whose distance from the mean (in either direction) is more than *one* standard deviation is approximately one-third (one in three).
2) In any normal distribution, the fraction of values whose distance from the mean is more than *two* standard deviations, is approximately 5 percent (one in twenty).
3) In any normal distribution, the fraction of values whose distance from the mean is more than *three* standard deviations is approximately 0.3 percent (three in one thousand).

These facts can be expressed more concisely by using the reduced form of the normal distribution:

1) Probability that $|z| > 1$ is approximately equal to 0.33.
2) Probability that $|z| > 2$ is approximately equal to 0.05.
3) Probability that $|z| > 3$ is equal to 0.003.

The concept of coverage

If we define the *coverage* of an interval from A to B to be the fraction of values of the population falling inside this interval, the three facts (1), (2), and (3) can be expressed as follows (where "sigma" denotes standard deviation):

1) A plus-minus *one*-sigma interval around the mean has a coverage of about 2/3 (67 percent).
2) A plus-minus *two*-sigma interval around the mean has a coverage of about 95 percent.
3) A plus-minus *three*-sigma interval around the mean has a coverage of 99.7 percent.

The coverage corresponding to a $\pm z$-sigma interval around the mean has been tabulated for the normal distribution for values of z extending from 0 to 4 in steps of 0.01, and higher in larger steps. Tabulations of the reduced normal distribution, also known as the "normal curve," or "error curve," can be found in most handbooks of physics and chemistry,[1] and in most textbooks of statistics.[2–5] Since the coverage corresponding to $z = 3.88$ is 99.99 percent, it is hardly ever necessary to consider values of z larger than four.

Confidence intervals

A *confidence interval* aims at bracketing the true value of a population parameter, such as its mean or its standard deviation, by taking into account the uncertainty of the sample estimate of the parameter.

Let $x_1, x_2, \ldots, x_N$ represent a sample of size N from a population of mean μ and standard deviation σ. In general μ and σ are unknown, but can be estimated from the sample in terms of $\bar{x}$ and s, respectively.

Confidence intervals for the mean

A confidence interval for the mean μ is an interval, AB, such that we can state, with a prechosen degree of confidence, that the interval AB brackets the population mean μ.

For example, we see in Table 4.3 that the mean of either of the two samples of size 10 is appreciably different from the (true) population mean (100.42 mg/dl). But suppose that the first of the two small samples is all the information we possess. We then would wish to find two values, A and B, *derived completely from the sample*, such that the interval AB is likely to include the true value (100.42). By making this interval long enough we can always easily fulfill this requirement, depending on what we mean by "likely." Therefore, we first express this qualification in a quantitative way by stipulating the value of a *confidence coefficient*. Thus we may require that the interval shall bracket the population mean "with 95 percent confidence." Such an interval is then called a "95 percent confidence interval."

The case of known σ.—We proceed as follows, assuming for the moment that although μ is unknown, the population standard deviation σ is known. We will subsequently drop this restriction.

We have already seen that the population of averages, $\bar{x}$, has mean μ and standard deviation $\sigma/\sqrt{N}$. The reduced variate corresponding to $\bar{x}$ is therefore:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \tag{4.15}$$

By virtue of the central limit theorem, the variable $\bar{x}$ generally may be considered to be normally distributed. The variable z then obeys the reduced normal distribution. We can therefore assert, for example, that the probability that

$$-1.96 < z < 1.96 \tag{4.16}$$

is 95 percent. Equation 4.16 can be written

$$-1.96 < \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} < 1.96$$

or

$$\bar{x} - 1.96\ \frac{\sigma}{\sqrt{N}} < \mu < \bar{x} + 1.96\ \frac{\sigma}{\sqrt{N}} \tag{4.17}$$

The probability that this double inequality will be fulfilled is 95 percent. Consequently, Equation 4.17 provides a confidence interval for the mean. The *lower limit A* of the confidence interval is $\bar{x} - 1.96\ \sigma/\sqrt{N}$; its *upper limit B* is $\bar{x} + 1.96\ \sigma/\sqrt{N}$. Because of the particular choice of the quantity 1.96, the probability associated with this confidence interval is, in this case, 95

percent. Such a confidence interval is said to be a "95 percent confidence interval," or to have a *confidence coefficient* of 0.95. By changing 1.96 to 3.00 in Equation 4.17, we would obtain a 99.7 percent confidence interval.

General formula for the case of known σ.—More generally, from the table of the reduced normal distribution, we can obtain the proper *critical* value z_c (to replace 1.96 in Equation 4.17) for any desired confidence coefficient. The general formula becomes

$$\bar{x} - z_c \cdot \frac{\sigma}{\sqrt{N}} < \mu < \bar{x} + z_c \cdot \frac{\sigma}{\sqrt{N}} \tag{4.18}$$

Values of z_c for a number of confidence coefficients are listed in tables of the normal distribution.

The *length* L of the confidence interval given by Equation 4.18 is

$$L = \left(\bar{x} + z_c \cdot \frac{\sigma}{\sqrt{N}}\right) - \left(\bar{x} - z_c \cdot \frac{\sigma}{\sqrt{N}}\right) = 2z_c \cdot \frac{\sigma}{\sqrt{N}} \tag{4.19}$$

The larger the confidence coefficient, the larger will be z_c and also L. It is also apparent that L increases with σ, but *decreases* as N becomes larger. This decrease, however, is slow, as it is proportional to only the square root of N. By far the best way to obtain short confidence intervals for an unknown parameter is to choose a measuring process for which the dispersion σ is small—in other words, to choose a measuring process of high precision.

The case of unknown σ. *Student's* t *distribution.*—A basic difficulty associated with the use of Equation 4.18 is that σ is generally unknown. However, the sample of N values provides us with an estimate s of σ. This estimate has $N - 1$ degrees of freedom. Substitution of s for σ in Equation 4.18 is not permissible, since the use of the reduced normal variate z in Equation 4.15 is predicated on a knowledge of σ.

It has been shown, however, that if $\bar{x}$ and s are the sample estimates obtained from a sample of size N, from a normal population of mean μ and standard deviation σ, the quantity, analogous to Equation 4.15, given by

$$t \equiv \frac{\bar{x} - \mu}{s/\sqrt{N}} \tag{4.20}$$

has a well-defined distribution, depending only on the degrees of freedom, $N - 1$, with which s has been estimated. This distribution is known as Student's t distribution with $N - 1$ degrees of freedom.

For σ unknown, it is still possible, therefore, to calculate confidence intervals for the mean μ by substituting in Equation 4.18 s for σ, and t_c for z_c. The confidence interval is now given by

$$\bar{x} - t_c \frac{s}{\sqrt{N}} < \mu < \bar{x} + t_c \frac{s}{\sqrt{N}} \tag{4.21}$$

The critical value t_c, for any desired confidence coefficient, is obtained from a tabulation of Student's t distribution. Tables of Student's t values can

be found in several references.[2–5] The length of the confidence interval based on Student's t distribution is

$$L = 2t_c \frac{s}{\sqrt{N}} \tag{4.22}$$

For any given confidence coefficient, t_c will be larger than z_c, so that the length of the interval given by Equation 4.22 is larger than that given by Equation 4.19. This difference is to be expected, since the interval now must take into account the uncertainty of the estimate s in addition to that of $\bar{x}$.

Applying Equation 4.21 to the two samples shown in Table 4.2, and choosing a 95 percent confidence coefficient (which, for 9 degrees of freedom, gives $t_c = 2.26$), we obtain:

1) For the first sample:

$$107.57 - 2.26 \frac{13.40}{\sqrt{10}} < \mu < 107.57 + 2.26 \frac{13.40}{\sqrt{10}}$$

or

$$98.0 < \mu < 117.2$$

The length of this interval is

$$117.2 - 98.0 = 19.2$$

2) For the second sample:

$$96.37 - 2.26 \frac{8.40}{\sqrt{10}} < \mu < 96.37 + 2.26 \frac{8.40}{\sqrt{10}}$$

or

$$90.4 < \mu < 102.4$$

The length of this interval is

$$102.4 - 90.4 = 12.0$$

Remembering that the population mean is 100.4, we see that the confidence intervals, though very different in length from each other, both bracket the population mean. We also may conclude that the lengths of the intervals, which depend on the sample size, show that a sample of size 10 is quite unsatisfactory when the purpose is to obtain a good estimate of the population mean, unless the measurement process is one of high precision.

Confidence intervals for the standard deviation

The chi-square distribution.—In many statistical investigations, the standard deviation of a population is of as much interest, if not more, than the mean. It is important, therefore, to possess a formula that provides a confidence interval for the unknown population standard deviation σ, given a sample estimate s.

If the number of degrees of freedom with which s is estimated is denoted by v, a confidence interval for σ is given by the formula:

$$s\sqrt{\frac{v}{\chi^2_U}} < \mu < s\sqrt{\frac{v}{\chi^2_L}} \tag{4.23}$$

In this formula, the quantities χ^2_U and χ^2_L are the appropriate upper and lower percentage points of a statistical distribution known as *chi-square*, for the chosen confidence coefficient. These percentage points are found in several references.[2–5]

This formula can be illustrated by means of the two samples in Table 4.2. To calculate 95 percent confidence intervals for σ (the population standard deviation), we locate the limits at points corresponding to the upper and lower 2.5 percentage points (or the 97.5 percentile and the 2.5 percentile) of chi-square. From the chi-square table we see that, for 9 degrees of freedom, the 97.5 percentile is 19.02, and the 2.5 percentile is 2.70. The 95 percent confidence interval in question is therefore:

1) For the first sample:

$$13.40\sqrt{\frac{9}{19.02}} < \sigma < 13.40\sqrt{\frac{9}{2.70}}$$

or

$$9.2 < \sigma < 24.5$$

2) For the second sample:

$$8.40\sqrt{\frac{9}{19.02}} < \sigma < 8.40\sqrt{\frac{9}{2.70}}$$

or

$$5.8 < \sigma < 15.3$$

Here again, both intervals bracket the population standard deviation 12.15, but again the lengths of the intervals reflect the inadequacy of samples of size 10 for a satisfactory estimation of the population standard deviation.

Tolerance intervals

In introducing the data of Table 4.1, we observed that it was possible to infer that about 1 percent of the population has serum glucose values of less than 70 mg/dl. This inference was reliable because of the large size of our sample (N = 2,197). Can similar inferences be made from small samples, such as those shown in Table 4.2? Before answering this question, let us first see how the inference from a very large sample (such as that of Table 4.1) can be made quantitatively precise.

The reduced variate for our data is

$$z = \frac{x - \mu}{\sigma} = \frac{x - 100.42}{12.15}$$

Making $x = 70$ mg/dl, we obtain for the corresponding reduced variate:

$$z = \frac{70 - 100.42}{12.15} = -2.50$$

If we now assume that the serum glucose data are normally distributed (i.e., follow a Gaussian distribution), we read from the table of the normal distribution that the fraction of the population for which z is less than -2.50 is 0.0062, or 0.62 percent. This is a more precise value than the 1 percent estimate we obtained from a superficial examination of the data.

It is clear that if we attempted to use the same technique for the samples of size 10 shown in Table 4.2, by substituting $\bar{x}$ for μ and s for σ, we may obtain highly unreliable values. Thus, the first sample gives a z value equal to $(70 - 107.57)/13.40$ or -2.80, which corresponds to a fraction of the population equal to 0.25 percent, and the second sample gives $z = (70 - 96.37)/8.40 = -3.14$, which corresponds to a fraction of the population equal to 0.08 percent. It is obvious that this approach cannot be used for small samples. It is possible, however, to solve *related* problems, even for small samples. The statistical procedure used for solving these problems is called the method of *tolerance intervals*.

Tolerance intervals for average coverages

Generally speaking, the method of tolerance intervals is concerned with the estimation of coverages or, conversely, with the determination of intervals that will yield a certain coverage. Let us consider an interval extending from $\bar{x} - ks$ to $\bar{x} + ks$, where k is any given value. The coverage corresponding to this interval will be a random variable, since the end points of the interval are themselves random variables. However, we can find a k value such that, on the average, the coverage for the interval will be equal to any preassigned value, such as, for example, 0.98. These k values, for normal distributions, have been tabulated for various sample sizes and desired average coverages.[5,6] As an illustration, we consider the first sample of size 10 given in Table 4.2, where

$$\bar{x} = 107.57, \; s = 13.40$$

For a coverage of 98 percent and 9 degrees of freedom, the tabulated value is

$$k = 3.053$$

Hence the tolerance interval that, on the average, will include 98 percent of the population is

$$107.57 - (3.053)(13.40) \text{ to } 107.57 + (3.053)(13.40)$$

or

$$66.7 \text{ to } 148.5$$

We can compare this interval to the one derived from the population itself (for all practical purposes, the large sample of 2,197 individuals may be considered as the population). Using the normal table, we obtain for a 98 percent coverage

$$100.42 - (2.326)(12.15) \text{ to } 100.42 + (2.326)(12.15)$$

or

$$72.2 \text{ to } 128.7$$

The fact that the small sample gives an appreciably wider interval is due to the uncertainties associated with the estimates $\bar{x}$ and s.

For a more detailed discussion of tolerance intervals, see Proschan.[6] Tables of coefficients for the calculation of tolerance intervals can be found in Snedecor and Cochran[5] and Proschan.[6]

Non-parametric tolerance intervals—order statistics

The tabulations of the coefficients needed for the computation of tolerance intervals are based on the assumption that the measurements from which the tolerance intervals are calculated follow a normal distribution; the table is inapplicable if this condition is grossly violated. Fortunately, one can solve a number of problems related to tolerance intervals for data from *any* distribution, by using a technique known as *non-parametric* or *distribution-free*. The method always involves an ordering of the data. First one rewrites the observation $x_1, x_2, \ldots, x_N$ in increasing order of magnitude. We will denote the values thus obtained by

$$x_{(1)}, x_{(2)}, \ldots, x_{(N)}$$

For example, Sample I in Table 4.2 is rewritten as:

$x_{(1)} = 91.9$	$x_{(6)} = 105.2$
$x_{(2)} = 96.6$	$x_{(7)} = 112.1$
$x_{(3)} = 96.9$	$x_{(8)} = 118.8$
$x_{(4)} = 97.0$	$x_{(9)} = 119.6$
$x_{(5)} = 103.4$	$x_{(10)} = 134.2$

The values $x_{(1)}, x_{(2)}, \ldots x_{(N)}$ are denoted as the first, second, . . . , Nth *order statistic*. The order statistics can now be used in a number of ways, depending on the problem of interest. Of particular usefulness is the following general theorem.

A general theorem about order statistics.—On the average, the fraction of the population contained between any two *successive* order statistics from a sample of size N is equal to $\frac{1}{N+1}$. The theorem applies to any continuous distribution (not only the Gaussian distribution) and to any sample size N.

Tolerance intervals based on order statistics.—It follows immediately from the above theorem that, *on the average*, the fraction of the population contained between the first and the last order statistics (the smallest and the largest values in the sample) is $\frac{N-1}{N+1}$. For example, on the average, the frac-

tion of the population contained between the smallest and the largest value of a sample of size 10 is $\frac{10-1}{10+1} = \frac{9}{11}$. The meaning of the qualification "on the average" should be properly understood. For any particular sample of size 10, the actual fraction of the population contained in the interval $x_{(N)} - x_{(1)}$ will generally not be equal to $\frac{N-1}{N+1}$. But if the average of those fractions is taken for many samples of size N, it will be close to $\frac{N-1}{N+1}$.

Tolerance intervals involving confidence coefficients

One can formulate more specific questions related to coverages by introducing, in addition to the coverage, the *confidence* of the statement about the coverage. For example, one can propose to find two order statistics such that the confidence is at least 90 percent that the fraction of the population contained between them (the coverage) is 95 percent. For a sample of size 200, these turn out to be the third order statistic from the bottom and the third order statistic from the top (see Table A30 in Natrella[3]). For further discussion of this topic, several references are recommended.[3,5,6]

Non-normal distributions and tests of normality

Reasons for the central role of the normal distribution in statistical theory and practice have been given in the section on the normal distribution. Many situations are encountered in data analysis for which the normal distribution does not apply. Sometimes non-normality is evident from the nature of the problem. Thus, in situations in which it is desired to determine whether a product conforms to a given standard, one often deals with a simple dichotomy: the fraction of the lot that meets the requirements of the standard, and the fraction of the lot that does not meet these requirements. The statistical distribution pertinent to such a problem is the *binomial* (see section on the binomial distribution).

In other situations, there is no a priori reason for non-normality, but the data themselves give indications of a non-normal underlying distribution. Thus, a problem of some importance is to "test for normality."

Tests of normality

Tests of normality should never be performed on small samples, because small samples are inherently incapable of revealing the nature of the underlying distribution. In some situations, a sufficient amount of evidence is gradually built up to detect non-normality and to reveal the general nature of the distribution. In other cases, it is sometimes possible to obtain a truly large sample (such as that shown in Table 4.1) for which normality can be tested by "fitting a normal distribution" to the data and then testing the "goodness of the fit."[5]

Probability plots.—A graphical procedure for testing for normality can be performed using the order statistics of the sample. This test is facilitated through the use of "normal probability paper," a type of graph paper on which the vertical scale is an ordinary arithmetic scale and the horizontal

Let p represent the fraction of individuals having the stated characteristic (serum glucose greater than 110 mg/dl) in the *sample of size* N; and let $q = 1 - p$. It is clear that for a relatively small, or even a moderately large N, p will generally differ from P. In fact, p is a random variable with a well-defined distribution function, namely the *binomial*.

The mean of the binomial (with parameter P) can be shown to be equal to P. Thus

$$E(p) = P \tag{4.24}$$

where the symbol $E(p)$ represents the "expected value" of p, another name for the population mean. Thus the population mean of the distribution of p is equal to the parameter P. If p is taken as an estimate for P, this *estimate* will therefore be *unbiased*.

Furthermore:

$$\text{Var}(p) = \frac{PQ}{N} \tag{4.25}$$

Hence

$$\sigma_p = \sqrt{\frac{PQ}{N}} \tag{4.26}$$

The normal approximation for the binomial distribution

It is a remarkable fact that for a large N, the distribution of p can be approximated by the normal distribution of the same mean and standard deviation. This enables us to easily solve practical problems that arise in connection with the binomial. For example, returning to our sample of 100 individuals from the population given in Table 4.1, we have:

$$E(p) = 0.215$$

$$\sigma_p = \sqrt{\frac{(0.215)(0.785)}{100}} = 0.0411$$

From these values, one may infer that in a sample of $N = 100$ from the population in question, the chance of obtaining p values of less than 0.13 (two standard deviations below the mean) or of more than 0.30 (two standard deviations above the mean) is about 5 percent. In other words, the chances are approximately 95 percent that in a sample of 100 from the population in question the number of individuals found to have serum glucose of more than 110 mg/dl will be more than 13 and less than 30.

Since, in practice, the value of P is generally unknown, all inferences must then be drawn from the sample itself. Thus, if in a sample of 100 one finds a p value of, say, 0.18 (i.e., 18 individuals with glucose serum greater than 110 mg/dl), one will consider this value as an estimate for P, and consequently one will take the value

$$\sqrt{\frac{(0.18)(1 - 0.18)}{100}} = 0.038$$

as an estimate for σ_p. This would lead to the following approximate 95 percent confidence interval for P:

$$0.18 - (1.96)(.038) < P < 0.18 + (1.96)(.038)$$

or

$$0.10 < P < 0.25$$

The above discussion gives a general idea about the uses and usefulness of the binomial distribution. More detailed discussions will be found in two general references.[4,5]

Precision and accuracy

The concept of control

In some ways, a measuring process is analogous to a manufacturing process. The analogue to the raw product entering the manufacturing process is the system or sample to be measured. The outgoing final product of the manufacturing process corresponds to the numerical result produced by the measuring process. The concept of *control* also applies to both types of processes. In the manufacturing process, control must be exercised to reduce to the minimum any random fluctuations in the conditions of the manufacturing equipment. Similarly, in a measuring process, one aims at reducing to a minimum any random fluctuations in the measuring apparatus and in the environmental conditions. In a manufacturing process, control leads to greater uniformity of outgoing product. In a measuring process, control results in higher *precision*, i.e., in less random scatter in repeated measurements of the same quantity.

Mass production of manufactured goods has led to the necessity of interchangeability of manufactured parts, even when they originate from different plants. Similarly, the need to obtain the same numerical result for a particular measurement, regardless of where and when the measurement was made, implies that *local* control of a measuring process is not enough. Users also require *interlaboratory* control, aimed at assuring a high degree of "interchangeability" of results, even when results are obtained at different times or in different laboratories.

Methods of monitoring a measuring process for the purpose of achieving "local" (i.e., within-laboratory) control will be discussed in the section on quality control of this chapter. In the following sections, we will be concerned with a different problem: estimating the precision and accuracy of a *method* of measurement.

Within- and between-laboratory variability

Consider the data in Table 4.6, taken from a study of the hexokinase method for determining serum glucose. For simplicity of exposition, Table

scale is *labeled* in terms of *coverages* (from 0 to 100 percent), but *graduated* in terms of the reduced z-values corresponding to these coverages (see section on the normal distribution). More specifically, suppose we divide the abscissa of a plot of the normal curve into $N + 1$ segments such that the area under the curve between any two successive division points is $\frac{1}{N+1}$. The division points will be $z_1, z_2, \ldots, z_N$, the values of which can be determined from the normal curve. Table 4.5 lists the values $\frac{1}{N+1}, \frac{2}{N+2}, \ldots, \frac{N}{N+1}$, in percent, in column 1, and the corresponding normal z values in column 2, for $N = 10$. According to the general theorem about order statistics, the order statistics of a sample of size $N = 10$ "attempt" to accomplish just such a division of the area into $N + 1$ equal parts. Consequently, the order statistics tend to be linearly related to the z values. The order statistics for the first sample of Table 4.2 are listed in column 3 of Table 4.5. A plot of column 3 versus column 2 will constitute a "test for normality": if the data are normally distributed, the plot will approximate a straight line. Furthermore, the intercept of this line (see the section on straight line fitting) will be an estimate of the mean, and the slope of the line will be an estimate of the standard deviation.[2] For non-normal data, systematic departures from a straight line should be noted. The use of normal probability paper obviates the calculations involved in obtaining column 2 of Table 4.5, since the horizontal axis is graduated according to z but labeled according to the values $\frac{i}{N+1}$, expressed as percent. Thus, in using the probability paper, the ten order statistics are plotted versus the numbers

$$100\,\frac{1}{11}, 100\,\frac{2}{11}, \ldots, 100\,\frac{10}{11}$$

or 9.09, 18.18, . . . , 90.91 percent. It is only for illustrative purposes that we have presented the procedure by means of a sample of size 10. One would generally not attempt to use this method for samples of less than 30. Even then, subjective judgment is required to determine whether the points fall along a straight line.

In a subsequent section, we will discuss transformations of scale as a means of achieving normality.

The binomial distribution

Referring to Table 4.1, we may be interested in the fraction of the population for which the serum glucose is greater than, say, 110 mg/dl. A problem of this type involves partitioning the range of values of a continuous variable (serum glucose in our illustration) into two groups, namely: (a) the group of individuals having serum glucose less than 110 mg/dl; and (b) the group of individuals having serum glucose greater than 110 mg/dl. (Those having serum glucose exactly equal to 110 mg/dl can be attached to one or the other group, or their number divided equally among them.)

TABLE 4.5. TEST OF NORMALITY USING ORDER STATISTICS[a]

Expected cumulative areas[b] in percent	Reduced normal variate	Order statistics of sample
9.09	–1.335	91.9
18.18	–0.908	96.6
27.27	–0.604	96.9
36.36	–0.348	97.0
45.45	–0.114	103.4
54.54	0.114	105.2
63.64	0.348	112.1
72.73	0.604	118.8
81.82	0.908	119.6
90.91	1.335	134.2

Straight Line Fit of column 3 versus column 2:

Intercept = 107.6 = $\hat{\mu}$

Slope = 15.5 = $\hat{\sigma}$

[a]The example is merely illustrative of the method. In practice one would never test normality on a sample of size 10.

[b]values of 100 $\frac{i}{N+1}$, where $N = 10$.

Suppose now that we have a random sample of only 100 individuals from the entire population. What fraction of the 100 individuals will be found in either group? It is seen that the binomial distribution has shifted the emphasis from the continuous variable (serum glucose) to the *number of individuals* (or the corresponding *fraction*, or *percentage*) in each of the two groups. There are cases in which no continuous variable was ever involved: for example, in determining the number of times a six appears in throwing a die. However, the theory of the binomial applies equally to both types of situations.

The binomial parameter and its estimation

Let P represent the *fraction* (i.e., a number between zero and one) of individuals in one of the two groups (e.g., serum glucose *greater* than 110 mg/dl) *in the population*. It is customary to represent the fraction for the other group by Q. Then it is obvious that $Q = 1 - P$. (If the fractions are expressed as percentages, we have percent Q = 100 − percent P.) For the data in Table 4.1 and the dividing value 110 mg/dl, we can calculate P by using the normal distribution:

The reduced value corresponding to 110 mg/dl is

$$\frac{110 - 100.42}{12.15} = 0.79$$

From the table of the normal distribution, we then obtain for P:

$$P = 0.215$$

Hence $Q = 1 - 0.215 = 0.785$

TABLE 4.6. DETERMINATION OF SERUM GLUCOSE

Laboratory	Serum sample			
	A	B	C	D
1	40.9[a]	76.0	137.8	206.3
	42.3	78.6	137.4	208.5
	42.3	77.5	138.5	204.9
	40.5	77.8	138.5	210.3
2	43.4	78.6	135.2	211.6
	43.8	76.0	131.3	201.2
	43.1	76.8	146.7	201.2
	42.3	75.7	133.4	208.7
3	41.3	75.0	134.5	205.1
	40.2	76.1	134.8	200.3
	40.6	76.4	131.5	206.9
	42.0	76.4	133.4	199.9

[a]All results are expressed in mg glucose/dl.

4.6 contains only a portion of the entire set of data obtained in this study. Each of three laboratories made four replicate determinations on each of four serum samples. It can be observed that, for each sample, the results obtained by *different* laboratories tend to show greater differences than results obtained through replication in the same laboratory. This observation can be made quantitative by calculating, for each sample, two standard deviations: the standard deviation "within" laboratories and the standard deviation "between" laboratories. Within-laboratory precision is often referred to as *repeatability*, and between-laboratory precision as *reproducibility*.[7] We will illustrate the method for serum sample A.

The data for serum A can first be summarized as follows:

Laboratory	Average	Standard Deviation
1	41.50	0.938
2	43.15	0.635
3	41.02	0.793

The three standard deviations could be averaged to obtain an "average within-laboratory"standard deviation. However, if one can assume that these three standard deviations are estimates of one and the same population standard deviation, a better way is to "pool" the variances,[2] and take the square root of the pooled variance. Using this procedure, we obtain for the best estimate of the within-laboratory standard deviation s_w:

$$s_w = \sqrt{\frac{(0.938)^2 + (0.635)^2 + (0.793)^2}{3}} = 0.798$$

Let us now calculate the standard deviation among the three average values 41.50, 43.15, and 41.02. Denoting this standard deviation by $s_{\bar{x}}$, we obtain:

$$s_{\bar{x}} = 1.117$$

If the laboratories displayed no systematic differences, this standard deviation, being calculated from averages of four individual results, should be equal to $s_w/\sqrt{4} = 0.798/\sqrt{4} = 0.399$. The fact that the calculated value, 1.117, is appreciably larger than 0.399 can be explained only through the presence of an additional, *between-laboratory* component of variability. This component, expressed as a standard deviation and denoted by s_L (where L stands for "laboratories"), is calculated by subtracting the "anticipated" variance, $(0.399)^2$, from the "observed" variance, $(1.117)^2$, and taking the square root:

$$s_L = \sqrt{(1.117)^2 - (0.399)^2} = 1.04$$

The calculations for all four serum samples are summarized in Table 4.7, in which standard deviations are rounded to two decimal places.

It may be inferred from Table 4.7 that s_w tends to increase with the glucose content of the sample. The between-laboratory component, s_L, shows no such trend. However, the data are insufficient to establish these facts with reasonable confidence. Since our purpose is to discuss general principles, and the use of these data is only illustrative, we will ignore these shortcomings in the discussion that follows.

Accuracy—comparison with reference values

The two components, s_w and s_L, define the *precision* of the method. To estimate its *accuracy*, one requires *reference values* for all samples. Let us assume that such values have been established and are as follows:

Serum Sample	Reference Value
A	40.8
B	76.0
C	133.4
D	204.1

The values given here as "reference values" are actually only tentative. We will assume, however, in our present discussion, that they can be considered to be free of systematic errors. Our task is to decide whether the values obtained in our study are, *within random experimental error*, equal to these reference values. The grand average value for sample A, 41.89 mg/dl, which

TABLE 4.7. SUMMARY OF ANALYSIS FOR SERUM GLUCOSE DATA

		Standard deviation	
Serum sample	Average (mg/dl)	s_w (mg/dl)	s_L (mg/dl)
A	41.89	0.80	1.04
B	76.74	1.05	0.54
C	136.08	4.08	1.07
D	205.41	3.91	1.08

we denote by the symbol $\bar{\bar{x}}$, involves 12 individual determinations and four laboratories. Its variance, therefore, can be estimated by the formula:

$$s_{\bar{\bar{x}}} = \sqrt{\frac{(0.80)^2}{12} + \frac{(1.04)^2}{4}} = 0.57$$

Now, $\bar{\bar{x}}$ differs from the reference value by the amount:

$$41.89 - 40.8 = 1.09$$

Corresponding values for all four samples are shown in Table 4.8.

It can be seen that, on the one hand, all four grand averages are larger than the corresponding reference values but, on the other hand, the differences D are of the order of only one or two standard errors $s_{\bar{\bar{x}}}$. One would tentatively conclude that the method shows a positive systematic error (bias) but, as has been pointed out above, the data are insufficient to arrive at definite conclusions.

Straight line fitting

The fitting of straight lines to experimental data is a subject of great importance, particularly in analytical laboratories. Many analytical and clinical methods make extensive use of linear calibration curves for the purpose of converting a measured quantity, such as an optical absorbance or a ratio of peaks–heights on a mass-spectrometer scan, into a concentration value for an unknown constituent. Calibration curves are established by subjecting samples of known concentrations to the measuring process and fitting lines to the resulting data. Let x be the known concentration, and y the measurement (e.g., optical absorbance). The data will consist of a set of paired values, as shown for an illustrative example in the columns labeled x and y in Table 4.9.

Inspection of the table shows that there is a "blank": for zero concentration, one finds a nonzero absorbance value. If one "corrected" the subsequent two values for the blank, one would obtain $0.189 - 0.050 = 0.139$, and $0.326 - 0.050 = 0.276$. If the "corrected" absorbance were proportional to concentration (as required by Beer's law), these two corrected absorbances should be proportional to 50 and 100, i.e., in a ratio of 1 to 2. Actually, 0.139 is slightly larger than (0.276/2). We will assume that this is due

TABLE 4.8. STUDY OF ACCURACY OF GLUCOSE DETERMINATION

Serum sample	Reference value (R)	Grand average ($\bar{\bar{x}}$)	D ($\bar{\bar{x}} - R$)	$s_{\bar{\bar{x}}}$
A	40.8	41.89	1.09	0.57
B	76.0	76.74	0.74	0.41
C	133.4	136.08	2.68	1.29
D	204.1	205.41	1.31	1.25

TABLE 4.9. CALIBRATION CURVE FOR GLUCOSE IN SERUM

x	y	$\hat{y}$	d
0	0.050	0.0516	−0.0016
50	0.189	0.1895	−0.0005
100	0.326	0.3273	−0.0013
150	0.467	0.4652	0.0015
200	0.605	0.6030	0.0020
400	1.156	1.1545	0.0015
600	1.704	1.7059	−0.0019
214.29	0.6425	0.6425	0

$\hat{y} = 0.0516 + 0.0027571\,x$
$s_\epsilon = 0.0019$

x = concentration of glucose, in mg/dl
y = absorbance
$\hat{y}$ = "fitted value"
d = residual

solely to experimental error in the measured absorbance values, thus assuming that any errors in the concentration values are negligibly small.

A general model

If α represents the true value of the "blank" and β the absorbance per unit concentration, we have, according to Beer's law:

$$E(y) - \alpha = \beta x \tag{4.27}$$

where $E(y)$ is the expected value for absorbance, i.e., the absorbance value freed of experimental error. Now the actual absorbance, y, is affected by an experimental error, which we will denote by e. Hence:

$$y = E(y) + e \tag{4.28}$$

Combining Equations 4.27 and 4.28 we obtain the "model" equation

$$y = \alpha + \beta x + e \tag{4.29}$$

This equation should hold for all x-values, i.e., $x_1, x_2, \ldots, x_N$, with the same values of α and β. Hence

$$y_i = \alpha + \beta x_i + e_i \tag{4.30}$$

where $i = 1$ to N.
The errors e_i should, on the average, be zero, but each one departs from zero by a random amount. We will assume that these random departures from zero do not increase with the absorbance (in some cases, this assumption is not valid) and that their distribution is Gaussian with standard deviation σ_e.

The object of the analysis is to estimate: (a) α and β, as well as the uncertainties (standard errors) of these estimates; and (b) the standard deviation of e; i.e., σ_e.

Formulas for linear regression

The fitting process is known in the statistical literature as the "linear regression of y on x." We will denote the estimates of α, β, and σ_e by $\hat{\alpha}$, $\hat{\beta}$, and s_e, respectively. The formulas involve the following three quantities:

$$U = \Sigma(x_i - \bar{x})^2 \tag{4.31}$$

$$W = \Sigma(y_i - \bar{y})^2 \tag{4.32}$$

$$P = \Sigma(x_i - \bar{x})(y_i - \bar{y}) \tag{4.33}$$

In terms of these three quantities, we have the formulas:

$$\hat{\beta} = \frac{P}{U} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{4.34}$$

$$s_e = \sqrt{\frac{W - (P^2/U)}{N - 2}} \tag{4.35}$$

$$s_{\hat{\beta}} = \frac{s_e}{\sqrt{U}} \qquad s_{\hat{\alpha}} = s_e \sqrt{\frac{1}{N} + \frac{\bar{x}^2}{U}} \tag{4.36}$$

For the data of Table 4.9, the calculations result in the following values: $\hat{\alpha} = 0.0516$, $s_{\hat{\alpha}} = 0.0010$, $\hat{\beta} = 0.0027571$, $s_{\hat{\beta}} = 0.0000036$, $s_e = 0.0019$. Since $\hat{\alpha}$ and $\hat{\beta}$ are now available, we can calculate, for each x, a "calculated" (or "fitted") value, $\hat{y}$, given by the equation $\hat{y} = \hat{\alpha} + \hat{\beta}x$. This is, of course, simply the ordinate of the point on the fitted line for the chosen value of x.

The differences between the observed value y and the calculated value $\hat{y}$ is called a "residual." Table 4.9 also contains the values of $\hat{y}$ and the residuals, denoted by the symbol "d."

It is important to observe that the quantity $(W - P^2/U)$, occurring in Equation 4.35, is simply equal to Σd_i^2. Thus:

$$s_e = \sqrt{\frac{\Sigma d^2}{N - 2}} \tag{4.37}$$

This formula, though mathematically equivalent to Equation 4.35, should be used in preference to Equation 4.35, unless all calculations are carried out with many significant figures. The reason for this is that the quantities d_i are less affected by rounding errors than the quantity $(W - P^2/U)$.

Examination of residuals—weighting

The residuals should behave like a set of random observations with a mean of *zero* and a standard deviation σ_e. It follows that the algebraic signs should exhibit a random pattern similar to the occurrence of heads and tails in the flipping of a coin. In our example, the succession of signs raises some suspicion of nonrandomness, but the series is too short to decide on this matter one way or the other. In any case, the errors are quite small, and the calibration curve is quite satisfactory for the intended purpose.

The assumptions underlying this procedure of fitting a straight line are not always fulfilled. The assumption of homoscedasticity (i.e., all e_i have the same standard deviation), in particular, is often violated. If the standard deviation of the error e_i is nonconstant and depends on x_i, the fitting of the straight line requires the application of "weighted regression analysis." Briefly, assuming a different value of σ_{e_i} for each i, one defines a "weight" w_i equal to the reciprocal of the square of σ_{e_i}. Thus:

$$w_i = 1/\sigma_{e_i}^2 \tag{4.38}$$

The weights w_i are then used in the regression calculations, leading to formulas that are somewhat different from those given in this section. For further details, two references can be consulted.[2,5]

Propagation of errors

It is often necessary to evaluate the uncertainty of a quantity that is not directly measured but is derived, by means of a mathematical formula, from other quantities that *are* directly measured.

An example

As an example, consider the determination of glucose in serum, using an enzymatic reaction sequence. The sequence generates a product, the optical absorbance of which is measured on a spectrophotometer. The procedure consists of three steps: (a) apply the enzyme reaction sequence to a set of glucose solutions of known concentrations, and establish in this way a calibration curve of "absorbance" versus "glucose concentration," (b) by use of the same reaction sequences, measure the absorbance for the "unknown," and (c) using the calibration curve, convert the absorbance for the unknown into a glucose concentration.

It turns out that the calibration curve, for this sequence of reactions, is *linear*. Thus, if y represents absorbance, and x concentration, the calibration curve is represented by the equation:

$$y = \alpha + \beta x \tag{4.39}$$

The calibration curve is established by measuring y for a set of known x values. We will again use the data of Table 4.9 for illustration. Fitting a straight line to these data, we obtain:

$$y = 0.0516 + 0.0027571x \tag{4.40}$$

Let us now suppose that an unknown sample of serum is analyzed m times (for example, $m = 4$), and that the average absorbance found is $y_u = 0.3672$ (where y_u stands for absorbance for the unknown). Using the calibration line, we convert the value y_u into a concentration value, $\hat{x}_u$, by solving the calibration equation for x:

$$x_{\hat{u}} = \frac{y_u - \alpha}{\beta} = \frac{0.3672 - 0.0516}{0.0027571} = 114.47 \text{ mg/dl} \tag{4.41}$$

How reliable is this estimate?

Let us assume, at this point, that the uncertainty of the calibration line is negligible. Then the only quantity affected by error is y_u, and it is readily seen from Equation 4.41 that the error of $\hat{x}_u$ is equal to that of y_u, divided by β. If we assume that the standard deviation of a single measured y-value is 0.0019 absorbance units, then the standard error of y_u, the average of four determinations, is

$$0.0019/\sqrt{4} = 0.00095$$

Hence the standard deviation of $\hat{x}_u$ is

$$0.00095/\beta = 0.00095/0.0027571 = 0.34 \text{ mg/dl}$$

A more rigorous treatment would also take account of the uncertainty of the calibration line.

The general case

More generally, a calculated quantity z can be a function of several measured quantities $x_1, x_2, x_3, \ldots$, each of which is affected by experimental error. The problem to be solved is the calculation of the standard deviation of the error of z as a function of the standard deviations of the errors of $x_1, x_2, x_3, \ldots$.

We will only deal with the case of *independent* errors in the quantities $x_1, x_2, x_3, \ldots$; i.e., we assume that the error of any one of the x's is totally unaffected by the errors in the other x's. For independent errors in the measured values $x_1, x_2, x_3, \ldots$, some simple rules can be applied. They are all derived from the application of a general formula known as "the law of propagation of errors," which is valid under very general conditions. The reader is referred to Mandel[2] for a general discussion of this formula.

Linear relations.—For

$$y = a_1x_1 + a_2x_2 + a_3x_3 + \ldots \tag{4.42}$$

the law states:

$$\text{Var}(y) = a_1^2 \text{Var}(x_1) + a_2^2 \text{Var}(x_2) + a_3^2 \text{Var}(x_3) + \ldots \tag{4.43}$$

As an example, suppose that the weight of a sample for chemical analysis has been obtained as the difference between two weights: the weight of an empty crucible, W_1, and the weight of the crucible containing the sample, W_2. Thus the sample weight S is equal to

$$S = W_2 - W_1 \tag{4.44}$$

This is in accordance with Equation 4.42 by writing:

$$S = (1)W_1 + (-1)W_2$$

Hence, according to Equation 4.43,

$$\text{Var}(S) = (1)^2\text{Var}(W_1) + (-1)^2\text{Var}(W_2)$$

or

$$\text{Var}(S) = \text{Var}(W_1) + \text{Var}(W_2)$$

Hence

$$\sigma_S = \sqrt{\sigma_{W_1}^2 + \sigma_{W_2}^2} \tag{4.45}$$

Note that in spite of the negative sign occurring in Equation 4.44, the variances of W_1 and W_2 in Equation 4.45 are *added* (not subtracted from each other).

It is also of great importance to emphasize that Equation 4.43 is valid *only* if the errors in the measurements $x_1, x_2, x_3, \ldots$, are *independent* of each other. Thus, if a particular element in chemical analysis was determined as the difference between 100 percent and the sum of the concentrations found for all other elements, the error in the concentrations for that element would *not* be independent of the errors of the other elements, and Equation 4.43 could *not* be used for any linear combination of the type of Equation 4.42 involving the element in question and the other elements. But in that case, Equations 4.42 and 4.43 could be used to evaluate the error variance for the element in question by considering it as the *dependent* variable y. Thus, in the case of three other elements x_1, x_2, and x_3, we would have:

$$y = 100 - (x_1 + x_2 + x_3)$$

where the errors of x_1, x_2, and x_3 are independent. Hence:

$$\text{Var}(y) = \text{Var}(x_1) + \text{Var}(x_2) + \text{Var}(x_3)$$

since the constant, 100, has zero-variance.

Products and ratios.—For products and ratios, the law of propagation of errors states that the squares of the coefficients of variation are additive. Here again, independence of the errors is a necessary requirement for the validity of this statement. Thus, for

$$y = x_1 \cdot x_2 \tag{4.46}$$

with independent errors for x_1 and x_2, we have:

$$\left(100\,\frac{\sigma_y}{y}\right)^2 = \left(100\,\frac{\sigma_{x_1}}{x_1}\right)^2 + \left(100\,\frac{\sigma_{x_2}}{x_2}\right)^2 \tag{4.47}$$

We can, of course, divide both sides of Equation 4.47 by 100^2, obtaining:

$$\left(\frac{\sigma_y}{y}\right)^2 = \left(\frac{\sigma_{x_1}}{x_1}\right)^2 + \left(\frac{\sigma_{x_2}}{x_2}\right)^2 \tag{4.48}$$

Equation 4.48 states that for products of independent errors, the squares of the *relative* errors are additive.

The same law applies for ratios of quantities with independent errors. Thus, when x_1 and x_2 have independent errors, and

$$y = \frac{x_1}{x_2} \tag{4.49}$$

we have

$$\left(\frac{\sigma_y}{y}\right)^2 = \left(\frac{\sigma_{x_1}}{x_1}\right)^2 + \left(\frac{\sigma_{x_2}}{x_2}\right)^2 \tag{4.50}$$

As an illustration, suppose that in a gravimetric analysis, the sample weight is S, the weight of the precipitate is W, and the "conversion factor" is F. Then:

$$y = 100F\frac{W}{S}$$

The constants 100 and F are known without error. Hence, for this example,

$$\left(\frac{\sigma_y}{y}\right)^2 = \left(\frac{\sigma_W}{W}\right)^2 + \left(\frac{\sigma_S}{S}\right)^2$$

If, for example, the coefficient of variation for S is 0.1 percent, and that for W is 0.5 percent, we have:

$$\frac{\sigma_y}{y} = \sqrt{(0.005)^2 + (0.001)^2} = 0.0051$$

It is seen that in this case, the error of the sample weight S has a negligible effect on the error of the "unknown" y.

Logarithmic functions.—When the calculated quantity y is the natural logarithm of the measured quantity x (we assumed that $x > 0$):

$$y = \ln x \tag{4.51}$$

the law of propagation of error states

$$\sigma_y = \frac{\sigma_x}{x} \tag{4.52}$$

For logarithms to the base 10, a multiplier must be used: for

$$y = \log_{10} x \tag{4.53}$$

the law of propagation of error states:

$$\sigma_y = \frac{1}{2.30} \cdot \frac{\sigma_x}{x} \tag{4.54}$$

Sample sizes and compliance with standards

Once the repeatability and reproducibility of a method of measurement are known, it is a relatively simple matter to estimate the size of a statistical sample that will be required to detect a desired effect, or to determine whether a given specification has been met.

An example

As an illustration, suppose that a standard requires that the mercury content of natural water should not exceed 2μg/l. Suppose, furthermore, that the standard deviation of reproducibility of the test method (see section on precision and accuracy, and Mandel[7]), at the level of 2μg/l, is 0.88μg/l. If subsamples of the water sample are sent to a number of laboratories and

each laboratory performs a single determination, we may wish to determine the number of laboratories that should perform this test to ensure that we can detect noncompliance with the standard. Formulated in this way, the problem has no definite solution. In the first place, it is impossible to guarantee *unqualifiedly* the detection of any noncompliance. After all, the decision will be made on the basis of measurements, and measurements are subject to experimental error. Even assuming, as we do, that the method is *unbiased*, we still have to contend with random errors. Second, we have, so far, failed to give precise meanings to the terms "compliance" and "noncompliance"; while the measurement in one laboratory might give a value less than 2μg/l of mercury, a second laboratory might report a value greater than 2μg/l.

General procedure—acceptance, rejection, risks

To remove all ambiguities regarding sample size, we might proceed in the following manner. We consider two situations, one definitely acceptable and the other definitely unacceptable. For example, the "acceptable" situation might correspond to a *true* mercury content of 1.5μg/l, and the "unacceptable" situation to a mercury content of 2.5μg/l (see Fig. 4.2).

Because of experimental errors, we must consider two *risks*: that of *rejecting* (as noncomplying) a "good" sample (1.5μg/l); and that of *accepting* (as complying) a "bad" sample (2.5μg/l). Suppose that both risks are set at 5 percent.

Let us now denote by N the number of laboratories required for the test. The average of the N measurements, which we denote by $\bar{x}$, will follow a normal distribution whose mean will be the true value of the mercury content of the sample and whose standard deviation will be $\sigma/\sqrt{N}$, or $0.88/\sqrt{N}$. For the "acceptable" situation the mean is 1.5μg/l, and for the "unacceptable" situation it is 2.5μg/l. We now stipulate that we will *accept*

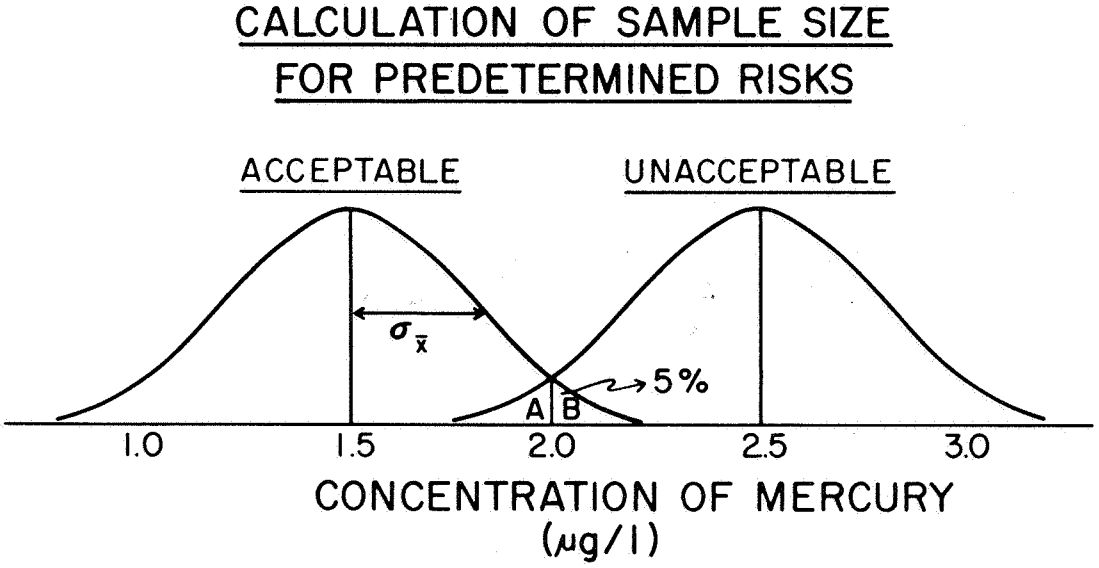

Fig. 4.2. Distribution of measurements of mercury in subsamples of a water sample sent to N laboratories.

the sample, as complying, whenever $\bar{x}$ is less than 2.0, and *reject* it, as noncomplying, whenever $\bar{x}$ is greater than 2.0. As a result of setting our risks at 5 percent, this implies that the areas A and B are each equal to 5 percent (see Fig. 4.2). From the table of the normal distribution, we read that for a 5 percent one-tailed area, the value of the reduced variate is 1.64. Hence:

$$z = \frac{2.0 - 1.5}{0.88/\sqrt{N}} = 1.64$$

(We could also state the requirement that $(2.0 - 2.5)/(0.88/\sqrt{N}) = -1.64$, which is algebraically equivalent to the one above.) Solving for N, we find:

$$N = \left(\frac{1.64 \cdot 0.88}{0.5}\right)^2 = 8.3 \tag{4.55}$$

We conclude that nine laboratories are required to satisfy our requirements. The general formula, for equal risks of accepting a noncomplying sample and rejecting a complying one, is:

$$N = \left(\frac{z_c \cdot \sigma}{d}\right)^2 \tag{4.56}$$

where σ is the appropriate standard deviation, z_c is the value of the reduced normal variate corresponding to the risk probability (5 percent in the above example), and d is the departure (from the specified value) to which the chosen risk probability applies.

Inclusion of between-laboratory variability

If the decision as to whether the sample size meets the requirements of a standard must be made in a single laboratory, we must make our calculations in terms of a different standard deviation. The proper standard deviation, for an average of N determinations in a *single* laboratory, would then be given by:

$$\sigma = \sqrt{\frac{\sigma_w^2}{N} + \sigma_L^2} \tag{4.57}$$

The term σ_L^2 must be included, since the laboratory mean may differ from the true value by a quantity whose standard deviation is σ_L. Since the between-laboratory component σ_L^2 is not divided by N, σ cannot be less than σ_L no matter how many determinations are made in the single laboratory. Therefore, the risks of false acceptance or false rejection of the sample cannot be chosen at will. If in our case, for example, we had $\sigma_w = 0.75\mu$g/l and $\sigma_L = 0.46\mu$g/l, the total σ cannot be less than 0.46. Considering the favorable case, $\mu = 1.5\mu$g/l, the reduced variate (see Fig. 4.2) is:

$$\frac{2.0 - 1.5}{0.46} = 1.09$$

This corresponds to a risk of 13.8 percent of rejecting (as noncomplying) a sample that is actually complying. This is also the risk probability of accept-

ing (as complying) a sample that is actually noncomplying. The conclusion to be drawn from the above argument is that, in some cases, testing error will make it impossible to keep the double risk of accepting a noncomplying product and rejecting a complying product below a certain probability value. If, as in our illustration, the purpose of the standard is to protect health, the proper course of action is to set the specified value at such a level that, even allowing for the between-laboratory component of test error, the risk of declaring a product as complying, when it is actually noncomplying, is low. If, in our illustration, a level of 2.5μg/l is such that the risk of false acceptance of it (as complying) should be kept to 5 percent (and $\sigma_L = 0.46\mu$g/l), then the specification limit should be set at a value x such that:

$$\frac{2.5 - x}{0.46} = 1.64$$

which, solved for x, yields 1.75 μg/l.

Transformation of scale

Some common transformations

Non-normal populations are often *skew* (nonsymmetrical), in the sense that one tail of the distribution is longer than the other. Skewness can often be eliminated by a *transformation of scale*. Consider, for example, the three numbers 1, 10, and 100. The distance between the second and the third is appreciably larger than that between the first and the second, causing a severe asymmetry. If, however, we convert these numbers to logarithms (base 10), we obtain 0, 1, and 2, which constitute a symmetrical set. Thus, if a distribution is *positively skewed* (long-tail on the right), a logarithmic transformation will reduce the skewness. (The simple logarithmic transformation is possible only when all measured values are positive). A transformation of the logarithmic type is not confined to the function $y = \log x$. More generally, one can consider a transformation of the type:

$$y = K \log (A + Bx) \tag{4.58}$$

or even

$$y = C + K \log (A + Bx) \tag{4.59}$$

where C, K, A, and B are properly chosen constants. It is necessary to choose A and B such that $A + Bx$ is positive for all x values. Other common types of transformations are:

$$y = \sqrt{x} \tag{4.60}$$

and

$$y = \arcsin \sqrt{x} \tag{4.61}$$

Robustness

The reason given above for making a transformation of scale is the presence of skewness. Another reason is that certain statistical procedures are

valid only when the data are at least approximately normal. The procedures may become grossly invalid when the data have a severely non-normal distribution.

A statistical procedure that is relatively insensitive to non-normality in the original data (or, more generally, to any set of specific assumptions) is called "robust." Confidence intervals for the mean, for example, are quite robust because, as a result of the central limit theorem, the distribution of the sample mean $\bar{x}$ will generally be close to normality. On the other hand, tolerance intervals are likely to be seriously affected by non-normality. We have seen that nonparametric techniques are available to circumvent this difficulty.

Suppose that, for a particular type of measurement, tests of normality on many sets of data always show evidence of non-normality. Since many statistical techniques are based on the assumption of normality, it would be advantageous to transform these data into new sets that are more nearly normal.

Fortunately, the transformations that reduce skewness also tend, in general, to achieve closer compliance with the requirement of normality. Therefore, transformations of the logarithmic type, as well as the square root and arcsine transformations, are especially useful whenever a nonrobust analysis is to be performed on a set of data that is known to be seriously non-normal. The reader is referred to Mandel[2] for further details regarding transformations of scale.

Transformations and error structure

It is important to realize that any nonlinear transformation changes the *error structure* of the data, and transformations are, in fact, often used for the purpose of making the experimental error more uniform over the entire range of the measurements. Transformations used for this purpose are called "variance-stabilizing" transformations. To understand the principle involved, consider the data in Table 4.10, consisting of five replicate absorbance values at two different concentrations, obtained in the calibration of

TABLE 4.10. ERROR STRUCTURE IN A LOGARITHMIC TRANSFORMATION OF SCALE

	Original data (Absorbance)		Transformed data ($\log_{10}$ Absorbance)	
	Set A[a]	Set B[b]	Set A	Set B
	0.2071	1.6162	−0.6838	0.2085
	0.2079	1.5973	−0.6821	0.2034
	0.1978	1.6091	−0.7038	0.2066
	0.1771	1.7818	−0.7518	0.2509
	0.2036	1.6131	−0.6912	0.2077
Average	0.1987	1.6435	−0.7025	0.2154
Standard deviation	0.0127	0.0776	0.0288	0.0199

[a]Absorbance values for a solution of concentration of 50 mg/dl of glucose.

[b]Absorbance values for a solution of concentration of 600 mg/dl of glucose.

spectrophotometers for the determination of serum glucose. At the higher concentration level, the absorbance values are of course higher, but so is the standard deviation of the replicate absorbance values. The ratio of the average absorbance values is 1.6435/0.1987 = 8.27. The ratio of the standard deviations is 0.0776/0.0127 = 6.11. Thus the standard deviation between replicates tends to increase roughly in proportion to the level of the measurement. We have here an example of "heterogeneity of variance." Let us now examine the two sets of values listed in Table 4.10 under the heading "transformed data." These are simply the logarithms to the base 10 of the original absorbance values. This time, the standard deviations for the two levels are in the proportion 0.0199/0.0288 = 0.69. Thus, the logarithmic transformation essentially has eliminated the heterogeneity of variance. It has, in fact, "stabilized" the variance. The usefulness of variance stabilizing transformations is twofold: (a) a single number will express the standard deviation of error, regardless of the "level" of the measurement; and (b) statistical manipulations whose validity is contingent upon a uniform error variance (homoscedasticity) and which are therefore inapplicable to the original data, can be applied validly to the transformed data.

Presentation of data and significant figures

The law of propagation of errors (see that section) enables one to calculate the number of significant figures in a calculated value. A useful rule of thumb is to report any standard deviation or standard error with two significant figures, and to report a calculated value with as many significant figures as are required to reach the decimal position of the second significant digit of its standard error.

An example

Consider the volumetric determination of manganese in manganous cyclohexanebutyrate by means of a standard solution of sodium arsenite. The formula leading to the desired value of percent Mn is

$$\text{Percent Mn} = 100\frac{v(\text{ml}) \cdot t\left(\frac{\text{mg}}{\text{ml}}\right) \cdot \frac{200\ (\text{ml})}{15(\text{ml})}}{w(\text{mg})}$$

where w is the weight of the sample, v the volume of reagent, and t the titer of the reagent, and the factor 200/15 is derived from taking an aliquot of 15 ml from a total volume of 200 ml.

For a particular titration, the values and their standard errors are found to be:

$v = 23.67$	$\sigma_v = 0.0040$
$t = 0.41122$	$\sigma_t = 0.000015$
200	$\sigma = 0.0040$
15	$\sigma = 0.0040$
$w = 939.77$	$\sigma_w = 0.0060$

The values are reported as they are read on the balance or on the burettes and pipettes; their standard errors are estimated on the basis of previous experience. The calculation gives:

$$\text{Percent Mn} = 13.809872$$

The law of propagation of errors gives:

$$\sigma_{\%\text{Mn}} =$$

$$13.8099\sqrt{\left(\frac{0.0040}{23.67}\right)^2+\left(\frac{0.000015}{0.41122}\right)^2+\left(\frac{0.0040}{200}\right)^2+\left(\frac{0.0040}{15}\right)^2+\left(\frac{0.0060}{939.77}\right)^2}$$

$$= 0.0044$$

On the basis of this standard deviation, we would report this result as:

$$\text{Percent Mn} = 13.8099;\ \sigma_{\%\text{Mn}} = 0.0044$$

It should be well understood that this calculation is based merely on weighing errors, volume reading errors, and the error of the titer of the reagent. In repeating the determination in different laboratories or even in the same laboratory, uncertainties may arise from sources other than just these errors. They would be reflected in the standard deviation calculated from such repeated measurements. In general, this standard deviation will be larger, and often considerably larger, than that calculated from the propagation of weighing and volume reading errors. If such a standard deviation from repeated measurements has been calculated, it may serve as a basis to redetermine the precision with which the reported value should be recorded.

In the example of the manganese determination above, the value given is just the first of a series of repeated determinations. The complete set of data is given in Table 4.11. The average of 20 determinations is 13.8380. The

TABLE 4.11. MANGANESE CONTENT OF MANGANOUS CYCLOHEXANEBUTYRATE

Determination number	Result (Percent Mn)	Determination number	Result (Percent Mn)
1	13.81	11	13.92
2	13.76	12	13.83
3	13.80	13	13.73
4	13.79	14	13.99
5	13.94	15	13.89
6	13.76	16	13.76
7	13.88	17	13.88
8	13.81	18	13.82
9	13.84	19	13.87
10	13.79	20	13.89

Average = $\bar{x}$ = 13.838

$s_x = 0.068$

$s_{\bar{x}} = 0.068/\sqrt{20} = 0.015$

standard deviation of the replicate values is 0.068; therefore, the standard error of the mean is $0.068/\sqrt{20} = 0.015$. The final value reported for this analysis would therefore be:

$$\text{Percent Mn} = \bar{x} = 13.838; s_{\bar{x}} = 0.015$$

This example provides a good illustration of the danger of basing an estimate of the precision of a value solely on the *reading* errors of the quantities from which it is calculated. These errors generally represent only a small portion of the total error. In this example, *the average of 20 values* has a true standard error that is still more than three times larger than the reading error of a *single determination*.

General recommendations

It is good practice to retain, for *individual* measurements, *more* significant figures than would result from calculations based on error propagation, and to use this law only for reporting the final value. This practice enables any interested person to perform whatever statistical calculations he desires on the individually reported measurements. Indeed, the results of statistical manipulations of data, when properly interpreted, are never affected by unnecessary significant figures in the data, but they may be seriously impaired by too much rounding.

The practice of reporting a measured value with a ± symbol followed by its standard error should be avoided at all costs, unless the meaning of the ± symbol is specifically and precisely stated. Some use the ± symbol to indicate a standard error of the value preceding the symbol, others to indicate a 95 percent confidence interval for the mean, others for the standard deviation of a single measurement, and still others use it for an uncertainty interval including an estimate of bias added to the 95 percent confidence interval. These alternatives are by no means exhaustive, and so far no standard practice has been adopted. It is of the utmost importance, therefore, to define the symbol whenever and wherever it is used.

It should also be borne in mind that the same measurement can have, and generally does have, more than one precision index, depending on the framework (statistical population) to which it is referred. For certain purposes, this population is the totality of (hypothetical) measurements that would be generated by repeating the measuring process over and over again on the same sample in the same laboratory. For other purposes, it would be the totality of results obtained by having the sample analyzed in a large number of laboratories. The reader is referred to the discussion in the section on precision and accuracy.

Tests of significance

General considerations

A considerable part of the published statistical literature deals with significance testing. Actually, the usefulness of the body of techniques classified under this title is far smaller than would be inferred from its prominence

in the literature. Moreover, there are numerous instances, both published and unpublished, of serious misinterpretations of these techniques. In many applications of significance testing, a "null-hypothesis" is formulated that consists of a statement that the observed experimental result—for example, the improvement resulting from the use of a drug compared to a placebo—is not "real," but simply the effect of chance. This null-hypothesis is then subjected to a statistical test and, if rejected, leads to the conclusion that the beneficial effect of the drug is "real," i.e., *not* due to chance. A closer examination of the nature of the null-hypothesis, however, raises some serious questions about the validity of the logical argument. In the drug-placebo comparison, the null-hypothesis is a statement of *equality of the means of two populations*, one referring to results obtained with the drug and the other with the placebo. All one infers from the significance test is a probability statement regarding the observed (sample) difference, on the hypothesis that the *true* difference between the population means is *zero*. The real question, of course, is related *not* to the *means* of hypothetical populations but rather to the benefit that any particular subject, selected at random from the relevant population of patients, may be expected to derive from the drug. Viewed from this angle, the usefulness of the significance test is heavily dependent on the *size* of the sample, i.e., on the number of subjects included in the experiment. This size will determine how large the difference between the two populations must be, *as compared to the spread of both populations*, before the statistical procedure will pick it up with a reasonable probability. Such calculations are known as the determination of the "power" of the statistical test of significance. Without indication of power, a test of significance may be very misleading.

Alternative hypotheses and sample size—the concept of power

An example of the use of "power" in statistical thinking is provided by our discussion in the section on sample sizes. Upon rereading this section, the reader will note that *two* situations were considered and that a probability value was associated with each of the two situations, namely, the probability of accepting or rejecting the lot. In order to satisfy these probability requirements, it was necessary to stipulate a value of N, the sample size. Smaller values of N would not have achieved the objectives expressed by the stipulated probabilities.

In testing a drug versus a placebo, one can similarly define two situations: (a) a situation in which the drug is hardly superior to the placebo; and (b) a situation in which the drug is definitely superior to the placebo. More specifically, consider a very large, *hypothetical* experiment in which subjects are paired at random, one subject of each pair receiving the placebo and the other the drug. Situation (a) might then be defined as that in which only 55 percent of all pairs shows better results with the drug than with the placebo; situation (b) might be defined as that in which 90 percent of the pairs shows greater effectiveness of the drug.

If we now perform an *actual* experiment, similar in nature but of moderate size, we must allow for random fluctuations in the percentage of pairs

that show better results with the drug as compared to the placebo. Therefore, our acceptance of the greater effectiveness of the drug on the basis of the data will involve risks of error. If the true situation is (a), we may wish to have only a small probability of declaring the drug superior, say, a probability of 10 percent. On the other hand, if the true situation is (b), we would want this probability to be perhaps as high as 90 percent. These two probabilities then allow us to calculate the required sample size for our experiment. Using this sample size, we will have assurance that the power of our experiment is sufficient to realize the stipulated probability requirements.

An example

An illustration of this class of problems is shown in Table 4.12. The data result from the comparison of two drugs, S (standard) and E (experimental), for the treatment of a severe pulmonary disease. The data represent the reduction in blood pressure in the heart after administration of the drug. The test most commonly used for such a comparison is Student's t test.[2–5] In the present case, the value found for t is 3.78, for DF = 142 (DF = number of degrees of freedom). The probability of obtaining a value of 3.78 or larger by pure chance (i.e., for equal efficacy of the two drugs) is less than 0.0002. The smallness of this probability is of course a strong indication that the hypothesis of equal efficacy of the two drugs is unacceptable. It is then generally concluded that the experiment has demonstrated the superior efficacy of E as compared to S. For example, the conclusion might take the form that "the odds favoring the effectiveness of E over S are better than M to 1" where M is a large number (greater than 100 in the present case). However, both the test and the conclusion are of little value for the solution of the real problem underlying this situation, as the following treatment shows. If we assume, as a first approximation, that the standard deviation 3.85 is the "population parameter" σ, and that the means, 0.10 for S and 2.53 for E, are also population parameters, then the probability of a single patient being better off

TABLE 4.12. TREATMENT OF PULMONARY EMBOLISM—COMPARISON OF TWO DRUGS

	Decrease in Right Ventricular Diastolic Blood Pressure (mm Hg)	
	Standard treatment (S)	Experimental treatment (E)
Number of patients	68	76
Average	0.10	2.53
Standard deviation	3.15	4.28
Standard error of average	0.38	0.50
True mean	μ_1	μ_2

t test for H_0: $\mu_1 = \mu_2$, (H_0 = null hypothesis)

$s_{pooled} = 3.85$ DF = 67 + 75 = 142, (DF = degrees of freedom)

$$t = \frac{2.53 - 0.10}{3.85\sqrt{\frac{1}{68} + \frac{1}{76}}} = 3.78 \; (P < 0.0002)$$

with E than with S is a function of the quantity Q defined by $Q \equiv (\mu_2 - \mu_1)/\sigma$. In the present case:

$$Q = \frac{2.53 - 0.10}{3.85} = 0.63$$

This can be readily understood by looking at Figure 4.3, in which the means of two populations, S and E, are less than one standard deviation apart, so that the curves show a great deal of overlap. There is no question that the two populations *are* distinct, and this is really all the *t* test shows. But due to the overlap, the probability is far from overwhelming that treatment E will be superior to treatment S for a randomly selected pair of individuals. It can be shown that this probability is that of a random normal deviate exceeding the value $\left(-\frac{Q}{\sqrt{2}}\right)$, or, in our case $\left(-\frac{0.63}{\sqrt{2}}\right) = -0.45$. This probability is 0.67, or about 2/3. Thus, in a large population of patients, two-thirds would derive more benefit from S than from E. Viewed from this perspective, the significance test, with its low "P value" (of 0.0002 in our case) is seen to be thoroughly misleading.

The proper treatment of a problem of this type is to raise the question of interest within a logical framework, derived from the nature of the problem, rather than perform standard tests of significance, which often merely provide correct answers to trivial questions.

Evaluation of diagnostic tests

The concepts of precision and accuracy are appropriate in the evaluation of tests that result in a quantitative measure, such as the glucose level of serum or the fluoride content of water. For medical purposes, different types of tests denoted as "diagnostic tests" are also of great importance. They dif-

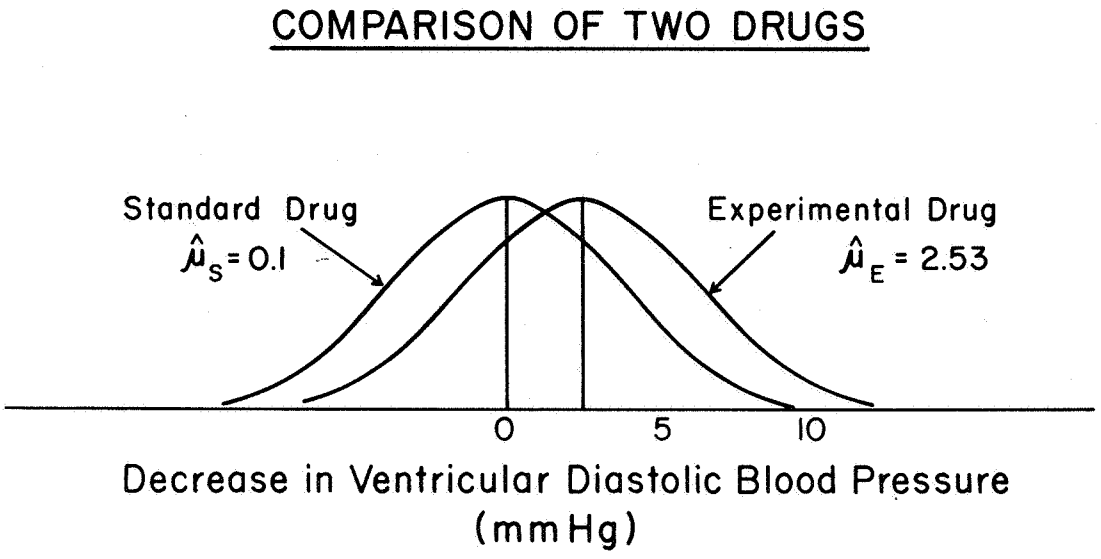

Fig. 4.3. Comparison of two drugs for the treatment of pulmonary disease, as measured by the reduction in right ventricular diastolic blood pressure (mm Hg).

fer from quantitative types of tests in that their outcome is characterized by a simple dichotomy into *positive* or *negative* cases.

As an example, consider Table 4.13, representing data on the alpha-fetoprotein (AFP) test for the diagnosis of hepatocellular carcinoma.[8] What do these data tell us about the value of the AFP test for the diagnosis of this disease?

Sensitivity and specificity

The statistical aspects of this type of problem are best understood by introducing a number of concepts that have been specifically developed for these problems.[8]

Sensitivity is the proportion of positive results among the subjects affected by the disease. Table 4.13 provides as an estimate of sensitivity:

$$\text{Sensitivity} = \frac{90}{107} = 0.8411 = 84.11\%$$

Specificity is the proportion of negative results among the subjects who are free of the disease. From Table 4.13:

$$\text{Specificity} = \frac{2079}{2118} = 0.9816 = 98.16\%$$

The concepts of sensitivity and specificity are useful descriptions of the nature of a diagnostic test, but they are not, in themselves, sufficient for providing the physician with the information required for a rational medical decision.

For example, suppose that a particular subject has a positive AFP test. What is the probability that this subject has hepatocarcinoma? From Table 4.13 we infer that among all subjects for whom the test is positive a proportion of 90/129, or 69.77 percent, are affected by the disease. This proportion is called the *predictive value of a positive test*, or *PV+*.

Predictive values—the concept of prevalence

Predictive value of a positive test.—(PV+) is defined as the proportion of subjects affected by the disease among those showing a positive test. The (PV+) value cannot be derived merely from the sensitivity and the specificity of the test. To demonstrate this, consider Table 4.14, which is fictitious and was derived from Table 4.13 by multiplying the values in the "Present"

TABLE 4.13. RESULTS OF ALPHA-FETOPROTEIN TESTS FOR DIAGNOSIS OF HEPATOCELLULAR CARCINOMA

	Hepatocarcinoma		
Test result	Present	Absent	Total
+	90	39	129
−	17	2,079	2,096
Total	107	2,118	2,225

TABLE 4.14. VALUES FOR ALPHA-FETOPROTEIN TESTS DERIVED FROM TABLE 4.13

Test Result	Hepatocarcinoma		Total
	Present	Absent	
+	900	39	939
−	170	2,079	2,249
Total	1,070	2,118	3,118

column by 10, and by leaving the values in the "Absent" column unchanged. Table 4.14 leads to the same sensitivity and specificity values as Table 4.13. However, the (PV+) value is now 900/939 = 95.85 percent.

It is seen that the (PV+) value depends not only on the sensitivity and the specificity but also on the *prevalence of the disease* in the total population. In Table 4.13, this prevalence is 107/2225 = 4.809 percent, whereas in Table 4.14 it is 1070/3118 = 34.32 percent.

A logical counterpart of the (PV+) value is the *predictive value of a negative test*, or *PV−*.

Predictive value of a negative test.—(PV−) is defined as the proportion of subjects free of the disease among those showing a negative test. For the data of Table 4.13, the (PV−) value is 2079/2096 = 99.19 percent, whereas for Table 4.14, (PV−) = 2079/2249 = 92.44 percent. As is the case for (PV+), the (PV−) value depends on the prevalence of the disease.

The following formulas relate (PV+) and (PV−) to sensitivity, specificity, and prevalence of the disease. We denote sensitivity by the symbol SE, specificity by SP, and prevalence by P; then:

$$(\text{PV}+) = \frac{1}{1 + \dfrac{(1 - SP)\,(1 - P)}{SE \cdot P}} \tag{4.62}$$

$$(\text{PV}-) = \frac{1}{1 + \dfrac{(1 - SE) \cdot P}{SP(1 - P)}} \tag{4.63}$$

As an illustration, the data in Table 4.13 yield:

$$(\text{PV}+) = \frac{1}{1 + \dfrac{(1 - 0.9816)\,(1 - 0.04809)}{(0.8411)\,(0.04809)}} = 0.6978 = 69.78\%$$

$$(\text{PV}-) = \frac{1}{1 + \dfrac{(1 - 0.8411)\,(0.04809)}{(0.9816)\,(1 - 0.04809)}} = 0.9919 = 99.19\%$$

Apart from rounding errors, these values agree with those found by direct inspection of the table.

Interpretation of multiple tests

The practical usefulness of (PV+) and (PV−) is now readily apparent. Suppose that a patient's result by the AFP test is positive and the prevalence

of the disease is 4.809 percent. Then the probability that the patient suffers from hepatocarcinoma is about 70 percent. On the basis of this result, the patient now belongs to a *subgroup* of the total population in which the prevalence of the disease is 70 percent rather than the 4.8 percent applying to the *total* population. Let us assume that a second test is available for the diagnosis of hepatocarcinoma, and that this second test is *independent* of the AFP test. The concept of *independence* of two diagnostic tests is crucial for the correct statistical treatment of this type of problem, but it seems to have received little attention in the literature. Essentially, it means that in the class of patients affected by the disease, the proportion of patients showing a positive result for test B is the same, whether test A was positive or negative. A similar situation must hold for the class of patients free of the disease.

In making inferences from this second test for the patient in question, we can start with a value of *prevalence of the disease* (P) of 70 percent, rather than 4.8 percent, since we know from the result of the AFP test that the patient belongs to the subgroup with this higher prevalence rate. As an illustration, let us assume that the second test has a sensitivity of 65 percent and a specificity of 90 percent and that the second test also is positive for this patient. Then the new (PV+) value is equal to

$$(\mathrm{PV}+) = \frac{1}{1 + \dfrac{(1 - 0.90)\,(1 - 0.70)}{(0.65)\,(0.70)}} = 0.938 = 93.8\%$$

If, on the other hand, the second test turned out to be negative, then the probability that the patient is free of disease would be:

$$(\mathrm{PV}-) = \frac{1}{1 + \dfrac{(1 - 0.65)\,(0.70)}{(0.90)\,(1 - 0.70)}} = 0.524 = 52.4\%$$

In that case, the two tests essentially would have contradicted each other, and no firm diagnosis could be made without further investigations.

A general formula for multiple independent tests

It can easily be shown that the order in which the independent tests are carried out has no effect on the final (PV+) or (PV−) value. In fact, the following general formula can be derived that covers any number of *independent* tests and their possible outcomes.

Denote by $(SE)_i$ and $(SP)_i$ the sensitivity and the specificity of the ith = test, where $i = 1, 2, 3, \ldots, N$. Furthermore, define the symbols A_i and B_i as follows:

$$A_i = \begin{cases} (SE)_i \text{ when the result of test } i \text{ is } + \\ 1 - (SE)_i \text{ when the result of test } i \text{ is } - \end{cases}$$

$$B_i = \begin{cases} 1 - (SP)_i \text{ when the result of test } i \text{ is } + \\ (SP)_i \text{ when the result of test } i \text{ is } - \end{cases}$$

If P is the prevalence rate of the disease *before* administration of any of the tests, and P' is the probability that the subject has the disease *after* administration of the N tests, then:

$$P' = \frac{1}{1 + \frac{(B_1 \cdot B_2 \cdot \ldots \cdot B_N)(1 - P)}{(A_1 \cdot A_2 \cdot \ldots \cdot A_N)P}} \tag{4.64}$$

It is important to keep in mind that Equation 4.64 is valid *only* if all tests are mutually independent in the sense defined above.

Quality Control

The remainder of this chapter deals with the fundamental principles of a quality control and quality assurance program for monitoring and assessing the precision and accuracy of the data being processed within a laboratory.

The definitions of Quality, Quality Assurance, and Quality Control by the American Society for Quality Control (ASQC)[9] apply to either a product or a service, and they are quoted here in their entirety.

1) *Quality.*—"The totality of features and characteristics of a product or service that bear on its ability to satisfy a given need."
2) *Quality assurance.*—"A system of activities whose purpose is to provide assurance that the overall quality-control job is in fact being done effectively. The system involves a continuing evaluation of the adequacy and effectiveness of the overall quality-control program with a view of having corrective measures initiated where necessary. For a specific product or service, this involves verifications, audits, and the evaluation of the quality factors that affect the specification, production, inspection, and use of the product or service."
3) *Quality control.*—"The overall system of activities whose purpose is to provide a quality of product or service that meets the needs of users; also, the use of such a system.

   "The aim of quality control is to provide quality that is satisfactory, adequate, dependable, and economic. The overall system involves integrating the quality aspects of several related steps, including the proper *specification* of what is wanted; *production* to meet the full intent of the specification; *inspection* to determine whether the resulting product or service is in accordance with the specification; and *review of usage* to provide for revision of specification.

   "The term *quality control* is often applied to specific phases in the overall system of activities, as, for example, *process quality control.*"

The Control Chart

According to the ASQC,[9] the control chart is "a graphical chart with control limits and plotted values of some statistical measure for a series of samples or subgroups. A central line is commonly shown."

The results of a laboratory test are plotted on the vertical axis, in units of the test results, versus time, in hours, days, etc., plotted on the horizontal axis. Since each laboratory test should be checked at least once a day, the horizontal scale should be wide enough to cover a minimum of one month of data. The control chart should be considered as a tool to provide a "real-time" analysis and feedback for appropriate action. Thus, it should cover a sufficient period of time to provide sufficient data to study trends, "runs" above and below the central line, and any other manifestation of lack of randomness (see section on detection of lack of randomness).

Statistical basis for the control chart

General considerations

W. A. Shewhart, in his pioneering work in 1939,[10] developed the principles of the control chart. They can be summarized, as was done by E. I. Grant,[11] as follows: "The measured quantity of a manufactured product is always subject to a certain amount of variation as a result of chance. Some stable 'System of Chance Causes' is inherent in any particular scheme of production and inspection. Variation within this stable pattern is inevitable. The reasons for variation outside this stable pattern may be discovered and corrected." If the words "manufactured product" are changed to "laboratory test," the above statement is directly applicable to the content of this section.

We can think of the "measured quantity" as the concentration of a particular constituent in a patient's sample (for example, the glucose content of a patient's serum). Under the "system of chance causes," this concentration, when measured many times under the same conditions, will fluctuate in such a way as to generate a statistical distribution that can be represented by a mathematical expression. This expression could be the normal distribution, for those continuous variables that are symmetrically distributed about the mean value, or it could be some other suitable mathematical function applicable to asymmetrically or discretely distributed variables (see section on non-normal distributions). Then, applying the known principles of probability, one can find lower and upper limits, known as *control limits*, that will define the limits of variation within "this stable pattern" for a given acceptable tolerance probability. Values outside these control limits will be considered "unusual," and an investigation may be initiated to ascertain the reasons for this occurrence.

Control limits

According to the ASQC,[9] the control limits are the "limits on a control chart that are used as criteria for action or for judging whether a set of data does or does not indicate lack of control."

Probability limits.—If the distribution of the measured quantity is known, then lower and upper limits can be found so that, on the average, a predetermined percentage of the values (e.g., 95 percent, 99 percent) will fall within these limits if the process is under control. The limits will depend on the nature of the probability distribution. They will differ, depending on

whether the distribution of the measured quantity is symmetric, asymmetric to the left or to the right, unimodal or bimodal, discrete or continuous, etc.

The obvious difficulty of finding the correct distribution function for each measured quantity, and of determining the control limits for this distribution, necessitates the use of procedures that are not overly sensitive to the nature of the distribution function.

Three-sigma limits.—The three-sigma limits, most commonly used in industrial practice, are based on the following expression:

Control limits = Average of the measured quantity ± three standard deviations of the measured quantity

The "measured quantity" could be the mean of two or three replicate determinations for a particular chemical test, the range of a set of replicate tests, a proportion defective, a radioactive count, etc.

The range of three standard deviations around the mean, that is, a width of six standard deviations, usually covers a large percentage of the distribution. For normally distributed variables, this range covers 99.7 percent of the distribution (see section on the normal distribution). For non-normally distributed variables, an indication of the percentage coverage can be obtained by the use of two well-known inequalities:

1) *Tchebycheff's Inequality*. For *any* distribution, (discrete or continuous, symmetric or asymmetric, unimodal or bimodal, etc.) with a finite standard deviation, the interval mean ± $K\sigma$ covers a proportion of the population of at least $1 - \frac{1}{K^2}$. Thus for $K = 3$, the coverage will be at least $1 - \frac{1}{9} = \frac{8}{9}$, or roughly 90 percent of the distribution.

2) *Camp-Meidel Inequality*. If the distribution is unimodal, the interval mean ± $K\sigma$ will cover a proportion of at least $1 - \frac{1}{2.25K^2}$ of the population. Thus, for $K = 3$, the coverage will be at least $1 - \frac{1}{20.25}$ or roughly 95 percent of the population.

From the above discussion, it follows that the three-sigma limits cover a proportion of the population that is at least equal to 90 percent for non-normal distributions and is equal to exactly 99.7 percent when the distribution is normal.

Most control charts are based on the mean of several determinations of the same measured equality. By the Central Limit Theorem, (see section on the normal distribution), the larger the sample size, the closer to normality will be the mean of this measured quantity. However, since most clinical tests are based on single or, at best, duplicate determinations, caution should be used in interpreting the amount of coverage given by the control limits for those distributions that are suspected to be skewed, bimodal, etc.

Warning limits.—The warning limits commonly used in practice are defined as:

Warning limits = Average of the measured quantity ± two standard deviations of the measured quantity.

For interpretation of points falling outside the warning and control limits, see the section on the control chart as a management tool.

Variability between and within subgroups

The hypothesis $\sigma_B = 0$

In control charts for variables, the variability is partitioned into two components: within and between subgroups. To this effect, the sequence of measurements is divided into subgroups of n consecutive values each. The variability within subgroups is estimated by first computing the average of the ranges of all subgroups and dividing this average by a factor that depends on n, which can be found in standard statistical tables. As an example, consider the sequence: 10.2, 10.4, 10.1, 10.7, 10.3, 10.3, 10.5, 10.4, 10.0, 9.8, 10.4, 10.9. When divided into subgroups of four, we obtain the arrangement:

Subgroup	Average	Range
10.2, 10.4, 10.1, 10.7	10.350	0.6
10.3, 10.3, 10.5, 10.4	10.375	0.2
10.0, 9.8, 10.4, 10.9	10.275	1.1
Average	10.333	0.63

In this case $n = 4$, and the average range is $\bar{R} = 0.63$.

Generally, n is a small number, often between 2 and 5. Its choice is sometimes arbitrary, dictated only by statistical convenience. More often, and preferably, the choice of n is dictated by the way in which the data were obtained. In the example above, the data may actually consist of three samples, each measured four times. In this case, "within groups" means "within samples," and "between groups" means "between samples."

Another possibility is that there were actually 12 samples, but that the measuring technique requires that they be tested in groups of four. If that is the situation, the relation of between-group to within-group variability depends not only on the sample-to-sample variability but also on the stability of the measuring instrument or technique from one group of four to another group of four. The location of the control limit and the interpretation of the control chart will depend on the nature and the choice of the subgroup.

If the standard deviation *within* subgroups is denoted by σ_W, and the standard deviation *between* subgroups by σ_B, a control chart is sometimes, but by no means always, a test as to whether σ_B exists (is different from zero). If $\sigma_B = 0$, then the variation between the *averages* of subgroups can be *predicted* from σ_W (or, approximately, from $\bar{R}$). The hypothesis $\sigma_B = 0$ can be tested by observing whether the subgroup averages stay within the control limits calculated on the basis of within-subgroup variability. Failure of this event to occur indicates the presence of causes of variability between subgroups. The nature of these causes depends on the criteria used in the selection of the subgroups.

The case $\sigma_B \neq 0$. *Baseline data*

In many applications, the hypothesis $\sigma_B = 0$ is not justified by the physical reality underlying the data. It may, for example, already be known that the subgroups vary from each other by more than can be accounted for by within-subgroup variability. Thus, each subgroup may represent a different

day of testing, and there may be more variability between days than within days. The initial set of data (*baseline* data) is then used primarily to estimate both the within- and the between-components of variability, and control limits are calculated on the basis of both these components (see section on computation of control limits). Data that are obtained subsequent to the baseline period are then evaluated in terms of these control lines. From time to time, the control lines are recalculated using *all* the data obtained up to that time, eliminating, however, those data for which abnormal causes of variability were found.

Types of control charts

Depending on the characteristics of the measured quantity, control charts can be classified into three main groups:

1) Control charts for variables (the $\bar{X}$, R Chart). These are used for variables such as clinical chemical determinations, some hematological parameters, etc.
2) Control charts for attributes (the P-Chart). These are used for proportion defective, proportion of occurrence of given disease, etc.
3) Control charts for number of defects per unit (the C-Chart). These may be used for counts, such as the number of cells observed in a given area, radioactive counts, etc.

Preparing a control chart

Objective and choice of variable

The general objectives of a control chart are: (a) to obtain initial estimates for the key parameters, particularly means and standard deviations. These are used to compute the central lines and the control lines for the control charts; (b) to ascertain *when* these parameters have undergone a radical change, either for worse or for better. In the former case, modifications in the control process are indicated; and (c) to determine *when* to look for assignable causes of unusual variations so as to take the necessary steps to correct them or, alternatively, to establish when the process should be left alone.

A daily review of the control chart should indicate whether the resulting product or service is in accordance with specifications. For example, in clinical chemistry, if a control chart based on standard samples shows statistical control for the measurement of a given constituent, then one can proceed with confidence with the determination of this constituent in patient samples. If the chart shows lack of control, an investigation should be started immediately to ascertain the reasons for this irregularity.

No general recommendations can be made here about the types of variables to use for quality control purposes, since they will obviously vary according to the various disciplines of the laboratory. Considerations of this type will be found in the respective specialty chapters of this book. The same statements apply to the types of stable pools or reagents that should be

used, and to the methods of handling these materials in normal laboratory practice.

Selecting a rational subgroup

The generally recommended approach for the selection of a subgroup of data for control purposes (using a single pool of homogeneous material) is that conditions *within* subgroups should be as uniform as possible (same instrument, same reagents, etc.), so if some assignable causes of error are present, they will show up *between* subgroups (see Duncan,[12] p. 347, and Grant,[11] Ch. 6, for further discussions).

When tests on patient samples are performed at regular intervals using standard laboratory equipment, the subgroup becomes automatically defined, since control samples are, or should be, included in each run. Otherwise, tests on control samples should be run at regular intervals during the day in order to detect possible changes in environmental conditions, reagents, calibrations, technicians, etc.

Size and frequency of control sample analyses

A minimum of two replicates should be obtained in each run of the control sample. To account for the possible effects of carryover from other samples, and to have a better indication of the capability of an instrument to reproduce itself under normal conditions within a run, the replicate samples should not be tested back-to-back, but should be separated by patient samples.

As indicated before, the frequency of the runs on control materials is generally tied to the frequency of the tests on patient samples. One general rule is to test the control samples as frequently as possible at the beginning of a control procedure, and to reduce this frequency to a minimum of two or three per day when the results of the control chart show a satisfactory state of control.

Maintaining uniform conditions in laboratory practice

A properly prepared control chart will tend to reflect any change in the precision and accuracy of the results obtained. To avoid wasting time in hunting for unnecessary sources of trouble, care should be taken to maintain laboratory conditions and practices as uniform as possible. These include sampling procedures, dilution techniques, aliquoting methods, storage methods, instrumental techniques, calculating procedures, etc.

Initiating a control chart

When meaningful historical data are not available (as is often the case when a quality control procedure is to be initiated), a plan should be set up to collect a minimum amount of data for each variable to be controlled during an initial *baseline period*.

For a control chart for *variables*, with a minimum of two replicates for each run, data should be collected for a baseline period of at least one month in order to allow sufficient time for the estimation of day-to-day variability.

Means and ranges should be computed for each run and plotted on separate charts. Records should be accurately kept, using standard quality control (QC) forms that are readily available. Any value that appears to be the result of a blunder should be eliminated, and the source of the blunder carefully noted. It is recommended that the number of runs or subgroups be at least 25 for the baseline period.

The same considerations apply to control charts of *proportions and counts*, except that the number of observations for each subgroup is generally larger than the corresponding number used in a control chart of variables. Statistical procedures for determining the sample size, n, for the P-chart or the C-chart can be found in the literature (see Duncan,[12] pp. 345 and 361). In general, n should be large enough to provide a good chance of finding one or more defectives in the sample.

Determining trial control limits

Based on the initial set of data collected during the baseline period, trial control limits can be determined using the procedure outlined in the section on random samples. After plotting these limits in the initial control chart (see section on the case $\sigma_B \neq 0$), points that are outside or very near the limits should be carefully examined, and if some valid reasons are found for their erratic behavior, they should be eliminated and new control limits should be computed. In general, it is better to start with control limits that are relatively narrow in order to better detect future trends, shifts in mean values, and some other types of irregularities. A common experience is that some initial subgroups of data will not be under control but, in general, after some knowledge is gained in the use of the control chart, the process will tend to reach a state of statistical equilibrium. After this time period, one generally has an adequate amount of data to produce realistic estimates of the mean and standard deviations.

Computing control limits

Two variable control charts should be kept, one for the *average value*, and the other for the *range* of individual determinations in each subgroup. In all cases in which a non-zero component for between-subgroups is known to exist, the control limits for the chart of averages will be based on the "total" standard deviation for subgroup averages.

If the subgroups are of size n, and if $\hat{\sigma}_W^2$ and $\hat{\sigma}_B^2$ represent the estimated *components* of variance within subgroups and between subgroups, respectively, then the "total standard deviation" for the averages of subgroups is

$$\hat{\sigma}_T = \sqrt{\hat{\sigma}_B^2 + \frac{\hat{\sigma}_W^2}{n}} \tag{4.65}$$

This quantity can also be obtained by directly calculating the standard deviation of the subgroup averages in the baseline period.

The control chart of the ranges will be used to ascertain whether the variability among individual readings within subgroups is consistent from

subgroup to subgroup. The limits for this chart will be based on the within-subgroup standard deviation.

Calculating the standard deviation

Using the available data for k subgroups, each of size n, we will have the layout shown in Table 4.15. The standard deviation within subgroups can be estimated from

$$S_W \approx \frac{\bar{R}}{d_2} \tag{4.66}$$

where

$$\bar{R} = \frac{\Sigma R_i}{k} \tag{4.67}$$

and the value of d_2 can be obtained from standard control chart tables (see Duncan,[12] p. 927). Values of d_2 for typical sample sizes are given in the following table:

n	d_2
2	1.128
3	1.693
4	2.059
5	2.326

The value of s_W can be accurately determined by pooling the variances from each subgroup (see section on precision and accuracy). However, the above estimate, based on the average range, is sufficiently accurate if the number of subgroups is large enough (say, 25 or more).

The standard deviation of the k sample averages is:

$$S_{\bar{x}} = \sqrt{\frac{\Sigma(\bar{X}_i - \bar{\bar{X}})^2}{k-1}} \tag{4.68}$$

The between-subgroups standard deviation is given by:

$$S_B = \sqrt{S_{\bar{X}}^2 - \frac{S_W^2}{n}} \tag{4.69}$$

TABLE 4.15. LAYOUT FOR $\bar{X}$, R CONTROL CHARTS

Subgroup	Determinations	Mean	Range
1	$X_{11}, X_{12}, \ldots, X_{1n}$	$\bar{X}_1$	R_1
2	$X_{21}, X_{22}, \ldots, X_{2n}$	$\bar{X}_2$	R_2
3	$X_{31}, X_{32}, \ldots, X_{3n}$	$\bar{X}_3$	R_3
•	•	•	•
•	•	•	•
k	$X_{k1}, X_{k2}, \ldots, X_{kn}$	$\bar{X}_k$	R_k
		$\bar{\bar{X}}$	$\bar{R}$

and the total standard deviation for individual determinations is:

$$S_{T_1} = \sqrt{S_B^2 + S_W^2} \tag{4.70}$$

The total standard deviation for averages of n daily determinations is:

$$S_{T_n} = \sqrt{S_B^2 + \frac{S_W^2}{n}} \tag{4.71}$$

Note that S_{T_n} is identically equal to $S_{\bar{x}}$.

Control limits for the chart of averages

The control limits for the chart of averages are given by:

$$UCL_{\bar{x}} = \bar{\bar{x}} + 3S_{\bar{x}} \tag{4.72}$$

and

$$LCL_{\bar{x}} = \bar{\bar{x}} - 3S_{\bar{x}} \tag{4.73}$$

where UCL = upper control limit; LCL = lower control limit.

The warning limits are:

$$UWL_{\bar{x}} = \bar{\bar{x}} + 2S_{\bar{x}} \tag{4.74}$$

and

$$LWL_{\bar{x}} = \bar{\bar{x}} - 2S_{\bar{x}} \tag{4.75}$$

where UWL = upper warning limit; LWL = lower warning limit.

Control limits for the chart of ranges

Based on the three-sigma limits concept (see section on control limits), the control limits for the chart of ranges are given by $\bar{R} \pm 3\sigma_R$. Using standard control chart notation, these limits are:

$$UCL_R = D_4\bar{R} \tag{4.76}$$

and

$$LCL_R = D_3\bar{R} \tag{4.77}$$

where

$$D_4 = 1 + 3\frac{d_3}{d_2} \tag{4.78}$$

and

$$D_3 = 1 - 3\frac{d_3}{d_2} \tag{4.79}$$

and the values of d_2, d_3, D_3, and D_4 are given in Natrella[3] and Duncan.[12] For $n = 2$, those values are $D_4 = 3.267$, and $D_3 = 0$.

The warning limits for $n = 2$ are:

$$UWL_R = 2.512\,\bar{R}$$

$$LWL_R = 0$$

The numerical value 2.512 is obtained as follows:

$$2.512 = 1 + 2\frac{d_3}{d_2} = 1 + 2\,\frac{(0.853)}{1.128}$$

Examples of average and range ($\bar{X}$ and R) charts

Initial data

The data in Table 4.16 represent 25 daily, duplicate determinations of a cholesterol control, run on a single-channel Autoanalyzer I, 40 per hour. It may appear strange that all 50 values are even. This is due to a stipulation in the protocol that the measured values be read to the nearest even number. The data cover a period of two months, with the analyzer run at a frequency

TABLE 4.16. EXAMPLE OF $\bar{X}$, R CHART: CHOLESTEROL CONTROL RUN

Day	Run 1 X_{i1}	Run 2 X_{i2}	Mean $\bar{X}_i$	Range R_i
1	390	392	391	2
2	392	388	390	4
3	392	388	390	4
4	388	388	388	0
5	378	396	387	18
6	392	392	392	0
7	392	390	391	2
8	398	402	400	4
9	404	406	405	2
10	400	400	400	0
11	402	402	402	0
12	392	406	399	14
13	398	396	397	2
14	380	400	390	20
15	398	402	400	4
16	388	386	387	2
17	402	392	397	10
18	386	390	388	4
19	386	382	384	4
20	390	386	388	4
21	396	390	393	6
22	396	394	395	2
23	384	388	386	4
24	388	382	385	6
25	386	384	385	2
			$\Sigma = 9{,}810$	120

of three days per week. The two daily determinations were randomly located within patient samples. The control consisted of 0.5 ml of sample extracted with 9.5 ml of 99 percent reagent-grade isopropyl alcohol.

Computing trial control limits

From the data in Table 4.16:

$$\bar{\bar{X}} = 9810/25 = 392.4$$

$$\bar{R} = 120/25 = 4.8$$

$$S_{\bar{x}}^2 = \left[\Sigma \bar{X}_i^2 - (\Sigma \bar{X}_i)^2/n\right]/(n-1) = \left[3850320 - (9810)^2/25\right]/24 = 36.5$$

$$S_{\bar{x}} = \sqrt{36.5} = 6.04$$

The control limits for $\bar{X}$ can be computed:

$$UCL_{\bar{x}} = 392.4 + 3(6.04) = 410.5$$

$$LCL_{\bar{x}} = 392.4 - 3(6.04) = 374.3$$

The warning limits for $\bar{X}$ are:

$$UWL_{\bar{x}} = 392.4 + 2(6.04) = 404.5$$

$$LWL_{\bar{x}} = 392.4 - 2(6.04) = 380.3$$

The control limits for R are:

$$UCL_R = (3.367)(4.8) = 15.7$$

$$LCL_R = 0$$

The warning limits of R are:

$$UWL_R = (2.512)(4.8) = 12.1$$

Analysis of data

In Figures 4.4 and 4.5, a graphical representation is shown of the control charts for the mean and range of the daily runs, together with their appropriate control limits.

The means of the daily runs appear to be under control. Only one point, day 9, is above the warning limit, and all points appear to be randomly located around the central line.

The control chart of the range shows two points out of control, days 5 and 14, and one point, day 12, on the upper warning limit.

Let us assume, for the purpose of illustration, that a satisfactory reason was found for those two points to be out of control in the range chart, and that it was decided to recompute new limits for both the $\bar{X}$ and the R charts based on only 23 days of data.

The new values are: $\bar{\bar{X}} = 392.7, \bar{R} = 3.57, S_{\bar{x}} = 6.17$, and $n = 23$.

$$UCL_{\bar{x}} = 392.7 + 3(6.17) = 411.2; UWL_{\bar{x}} = 405.0$$

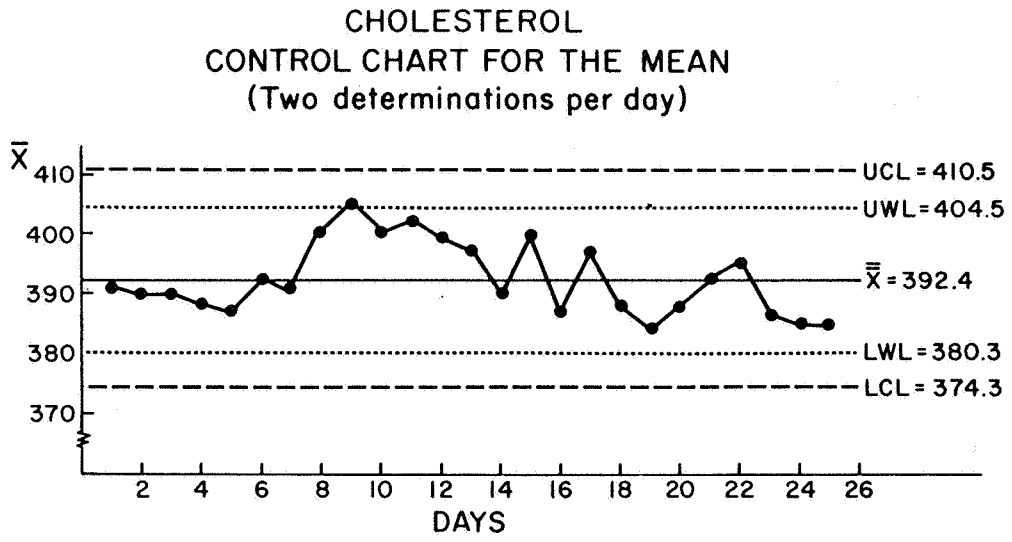

Fig. 4.4. Control chart for the mean, based on 25 daily duplicate determinations of a cholesterol control.

$$LCL_{\bar{x}} = 392.7 - 3(6.17) - 374.2; LWL_{\bar{x}} = 380.4$$

The new limits for the $\bar{X}$ chart are practically the same as the previous limits.

$$UCL_R = (3.267)(3.57) = 11.7; UWL_R = 9.0$$

$$LCL_R = 0 \qquad LWL_R = 0$$

These values establish the final limits, based on the baseline period.

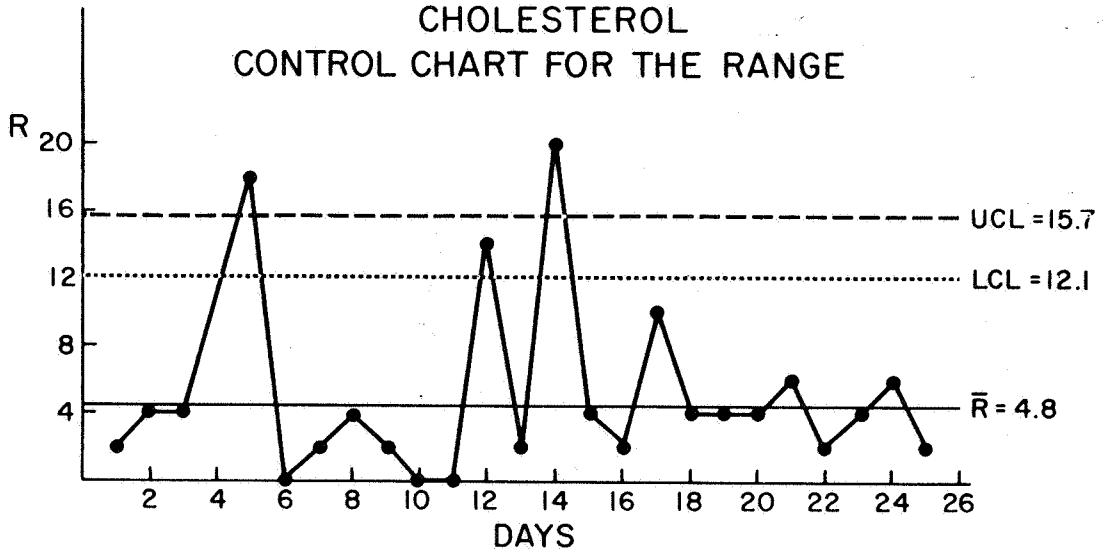

Fig. 4.5. Control chart for the range, based on 25 daily duplicate determinations of a cholesterol control.

Additional data

Nineteen additional points were obtained for days 26 to 44, running through a period of about one-and-a-half months. The values are shown in Table 4.17.

Figures 4.6 and 4.7 show the results of the 19 additional data points plotted against the (corrected) control limits based on the baseline period.

The $\bar{X}$-chart shows two points, days 38 and 39, out of control, about 40 percent of the points near the warning limits, and a definite trend toward large values of $\bar{X}$ after day 30. There is a run of seven points above the central line after day 37 and, in fact, if one considers day 37 to be "above" the central line (the mean of day 37 is 392), the run of points above the central line is of length 12. As indicated in the section on control limits, these considerations are indications of a process out of control.

The R-chart shows one point out of control and two points above the upper warning limit; although the value of $\bar{R}$ based on the 19 additional values, 4.32, is larger than the previous value, $\bar{R} = 3.57$, the difference is not significant.

The new set of points taken by itself produced the following values: $\bar{\bar{X}} = 396.5$, $\bar{R} = 4.32$, and $S_{\bar{x}} = 12.17$, where $n = 19$.

Future control limits

It is generally desirable to have a well-established baseline set so future points can be evaluated with confidence in terms of the baseline central line

TABLE 4.17. ADDITIONAL VALUES FOR CHOLESTEROL CONTROL RUN

Day	Run 1 X_{i1}	Run 2 X_{i2}	Mean $\bar{X}_i$	Range R_i
26	392	392	395	6
27	376	376	376	0
28	390	386	388	4
29	394	384	389	10
30	382	378	380	4
31	384	382	381	2
32	384	388	386	4
33	402	392	397	10
34	390	398	394	8
35	402	402	402	0
36	398	394	396	4
37	390	394	392	4
38	426	428	427	2
39	414	428	421	14
40	402	398	400	4
41	402	400	401	2
42	402	404	403	2
43	400	402	401	2
44	404	404	404	0
			$\Sigma = 7{,}533$	82

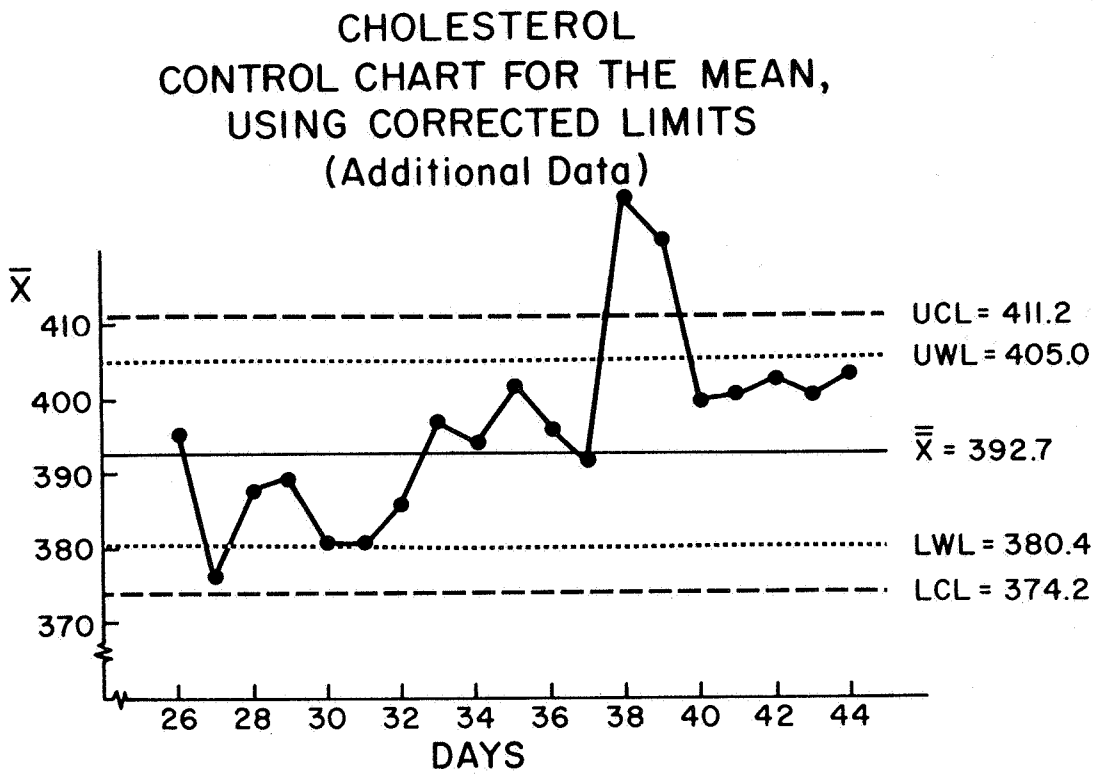

Fig. 4.6. Control chart for the mean, based on 19 additional data points, plotted against the corrected control limits.

and control limits. If, in the example under discussion, the additional set (days 26 to 44) was found to be satisfactorily consistent with the baseline data, then it would be proper to extend the baseline period by this set, i.e., a total of 25 + 19 = 44 points. However, we have already observed a number of shortcomings in the additional set, and the proper action is to search for the causes of these disturbances, i.e., "to bring the process under control." This is of course not a statistical problem.

For the purpose of our discussion, we will assume that an examination of the testing process has revealed faulty procedure starting with day 37. Therefore, we will consider a shortened additional set, of days 26 through 36. The following table gives a comparison of the baseline set (corrected to 23 points as discussed previously) and the shortened additional set (11 points).

	Baseline Set	Additional Set
Number of points, N	23	11
Average, $\bar{\bar{X}}$	392.7	389.5
Average Range, $\bar{R}$	3.57	4.73
Standard deviation, $s_{\bar{x}}$	6.17	8.15

By using the F test, [2–5] it is easily verified that the difference between the two standard deviations is well within the sampling variability that may be expected from estimates derived from samples of 23 and 11 points, respec-

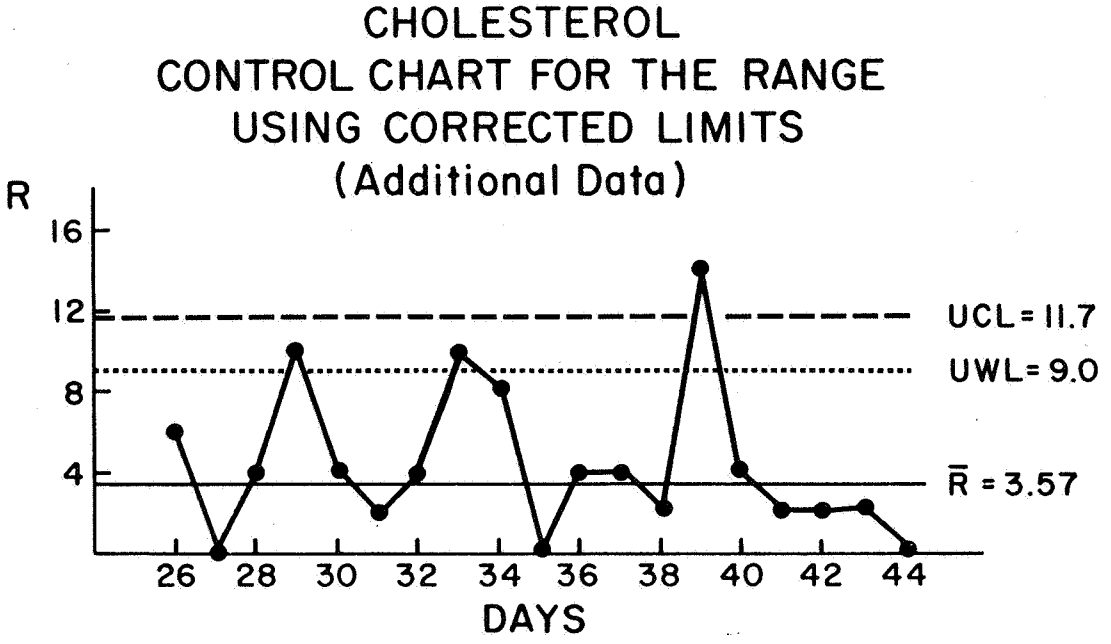

Fig. 4.7. Control chart for the range, based on 19 additional data points, plotted against the corrected control limits.

tively. The difference between the averages, $\bar{\bar{X}}$, is 392.7 − 389.5 = 3.2. A rough test can be made to see whether this difference indicates a real shift between the two sets. The standard error of the difference is approximately $[(6.17)^2/23 + (8.15)^2/11]^{\frac{1}{2}} = 2.77$. Thus the difference, 3.2, is equal to $\frac{3.2}{2.77}$ = 1.15 standard errors, and this is well within sampling errors.

It is therefore not unreasonable in this case to combine the 34 points of both sets to construct a new baseline. This results in the following parameters: $N = 34$, $\bar{\bar{X}} = 391.7$, $\bar{R} = 3.95$, and $s_{\bar{x}} = 6.93$.

The new control limits are:

For $\bar{X}$:	$UCL = 412.5$	$UWL = 405.6$
	$LCL = 370.9$	$LWL = 377.8$
For R:	$UCL = 12.9$	$UWL = 9.9$
	$LCL = 0$	$LWL = 0$

Using these new parameters, it can be noted that the points corresponding to days 37 through 44 may indicate a potential source of trouble in the measuring process.

Control chart for individual determinations

It is possible, although not recommended, to construct charts for individual readings. Extreme caution should be used in the interpretation of points out of control for this type of chart, since individual variations may not follow a normal distribution. When a distribution is fairly skewed, then a transformation (see section on transformation of scale) would be applied before the chart is constructed.

The steps to follow are:

1) Use a moving range of two successive determinations;
2) Compute $\bar{R} = \frac{\Sigma R_i}{k}$;
3) Determine the control limits for $\bar{X}$:

$$\bar{X} \pm 3\frac{\bar{R}}{d_2}$$

For $n = 2$, $d_2 = 1.128$, and hence the control limits are:

$$\bar{X} \pm 2.66\,\bar{R}$$

4) The upper control limit for R is $D_4\bar{R} = 3.267\,\bar{R}$. The lower control limit is equal to zero.

Other types of control charts

Control charts can also be constructed based on the average standard deviation, $\bar{\sigma}$, of several subgroups of sample data, or on "standard" values of σ, called σ' in the quality control literature (See Duncan,[12] Chap. 20).

Control chart for attributes—the P-chart

The fraction defective chart is generally used for quality characteristics that are considered attributes and are not necessarily quantitative in nature. To use this chart, it is only necessary to count the number of entities that have a well-defined property, such as being defective, have a certain type of disease, or have a glucose content greater than a given value, and translate this number into a proportion. The data used in this chart are easy to handle, and the cost of collection is normally not very high. In some instances, the P-chart can do the job of several average and range charts, since the classification of a "defective" element may depend on several quantitative characteristics, each of which would require an individual set of average and range charts for analysis.

The sample size for each subgroup will depend on the value of the proportion P being estimated. A small value of P will require a fairly large sample size in order to have a reasonable probability of finding one or more "defectives" in the sample (See Duncan[12]). In general, a value of n between 25 and 30 is considered adequate for the calculation of a sample proportion.

Control limits and warning limits

Since the standard deviation of a proportion is directly related to the value of the proportion, an estimate p of P is all that is needed for the calculation of the central line and of the control limits.

The central line is located at the value $\bar{p}$. The three-sigma control limits are:

$$UCL_p = \bar{p} + 3\sqrt{\frac{\bar{p}\,\bar{q}}{n}} \tag{4.80}$$

$$LCL_p = \bar{p} - 3\sqrt{\frac{\bar{p}\,\bar{q}}{n}} \tag{4.81}$$

where $\bar{q} = 1 - \bar{p}$. The estimate $\bar{p}$ is obtained as follows:

Let the data be represented by the table:

Sample Number	Size	Number of Elements Having a Certain Characteristic	Proportion
1	n	X_1	p_1
2	n	X_2	p_2
3	n	X_3	p_3
.	.	.	.
.	.	.	.
.	.	.	.
k	n	X_n	p_k
	Total	ΣX_i	Σp_i

where $p_i = \frac{X_i}{n}$

Average proportion:

$$\bar{p} = \frac{\Sigma p_i}{k} \tag{4.82}$$

The warning limits are:

$$UWL_p = \bar{p} + 2\sqrt{\frac{\bar{p}\,\bar{q}}{n}} \tag{4.83}$$

$$LWL_p = \bar{p} - 2\sqrt{\frac{\bar{p}\,\bar{q}}{n}} \tag{4.84}$$

When the sample size does not remain constant from subgroup to subgroup, the recommended procedure is to compute control limits using the average sample size. However, when a point falls near the control limits thus calculated, then the actual limits for this point, using its own sample size, should be estimated before a conclusion is reached about its state of control.

Control charts for number of defects per unit—the C-chart

In some instances, it is more convenient to maintain control charts for the number of defects per unit, where the unit may be a single article or a subgroup of a given size. The "number of defects" may be, for instance, the number of tumor cells in an area of a specified size, the number of radioactive counts in a specified period of time, etc. In all these instances, the probability of occurrence of a single event (e.g., an individual defect) is very

small, but the unit is large enough to make the average number of occurrences (number of defects) a measurable number.

The Poisson distribution

It can be shown that, when the probability P of an event is very small but the sample size n is large, then the distribution of the number of occurrences c of this event tends to follow a Poisson distribution with parameter $nP = c'$. The mean and standard deviation of c are:

$$E(c) = c' \tag{4.85}$$

$$\sigma_c = \sqrt{c'} \tag{4.86}$$

The random variable c represents the number of defects per unit, the number of radioactive counts in a given period of time, the number of bacteria in a specified volume of liquid, etc.

Control limits.—The upper and lower limits are given by:

$$UCL_c = \bar{c} + 3\sqrt{\bar{c}} \tag{4.87}$$

$$LCL_c = \bar{c} - 3\sqrt{\bar{c}} \tag{4.88}$$

Here $\bar{c}$ is the average number of defects, or counts, obtained using a sufficiently large number, k, of units. $\bar{c}$ is a sample estimate of the unknown, or theoretical value c'.

The warning limits are:

$$UWL_c = \bar{c} + 2\sqrt{\bar{c}} \tag{4.89}$$

$$LWL_c = \bar{c} - 2\sqrt{\bar{c}} \tag{4.90}$$

Detecting lack of randomness

If a process is in a state of statistical control, the observations plotted in the control chart should randomly fall above and below the central line, with most of them falling within the control limits. However, even if all the points fall within the upper and lower control limits, there might still exist patterns of nonrandomness that require action, lest they lead eventually to points outside the control limits. Procedures for detecting such patterns will be discussed.

Rules based on the theory of runs

The most frequent test used to detect a lack of randomness is based on the theory of runs. A run may be defined as a succession of observations of the same type. The length of a run is the number of observations in a given run. For example, if the observations are classified as a or b, depending on whether they fall above or below the mean, then one set of observations may look like:

a a a b a b b b a a b

Here we have six runs, of length 3, 1, 1, 3, 2, 1, respectively.

Another criterion for the definition of a run would be the property of increase or decrease of successive observations. Such runs are called "runs *up* and *down*." For example, the sequence 2, 1.7, 2.2, 2.5, 2.8, 2.0, 1.8, 2.6, 2.5, has three runs *down* and two runs *up*. In order of occurrence, the *lengths* of the runs are 1, 3, 2, 1, 1.

Returning to runs above and below the central value, it is possible through use of the theory of probability, and assuming that the probability is one-half that an observation will fall above the central line (and, consequently, one-half that it will fall below the central line), to determine the probability distribution of the lengths of runs. Tables are available for several of these distributions (See Duncan,[12] Chap. 6). Some rules of thumb based on the theory of runs that are very useful in pointing out some lack of randomness are:

1) A run of length 7 or more. This run may be up or down, above or below the central line in the control chart. (For runs above or below the median, the probability of a run of length 7 is 0.015.)
2) A run of two or three points outside the warning limits.
3) Ten out of 11 successive points on the same side of the central line.

Distribution of points around the central line

When a sufficient number of observations is available, the pattern of distribution of points around the central line should be carefully examined. In particular, if the points tend to cluster near the warning or control limits, or if they show characteristics of bimodality, or if they show a pronounced skewness either to the left or the right, then the assumption of normality will not be satisfied and some transformation of scale may be necessary.

Interpreting patterns of variation in a control chart

Indication of lack of control

A process is out of control when one or more points falls outside the control limits of either the $\bar{x}$ or the R-chart, for control of variables, or outside the limits of the P-chart, for control of attributes.

Points outside the control limits of the R-chart tend to indicate an increase in magnitude of the within-group standard deviation. An increase in variability may be an indication of a faulty instrument which eventually may cause a point to be out of control in the $\bar{x}$-chart.

When two or more points are in the vicinity of the warning limits, more tests should be performed on the control samples to detect any possible reasons for out-of-control conditions.

Various rules are available in the literature about the procedures to follow when control values are outside the limits (see, for example, Haven[13]).

Patterns of variation

By examining the $\bar{x}$ and R-charts over a sufficient period of time, it may be possible to characterize some patterns that will be worth investigating in order to eliminate sources of future troubles.

Some of these patterns are shown in Figure 4.8.

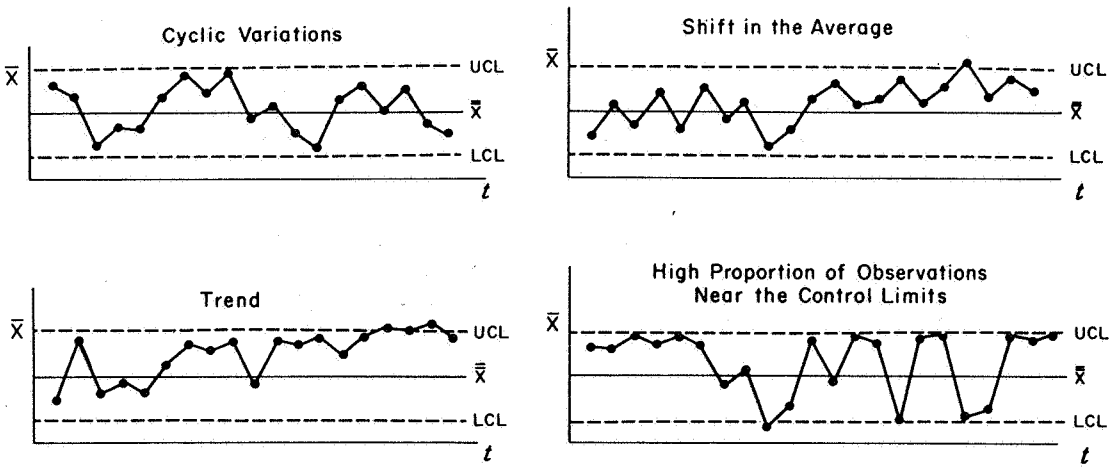

Fig. 4.8. Four patterns of variation in an $\bar{X}$-chart.

The control chart as a management tool

As indicated in the ASQC definition of quality assurance, ". . . The system involves a *continuing evaluation* of the adequacy and effectiveness of the overall quality-control program with a view of *having corrective measures* initiated where necessary . . ."[9]

The key words, "continuing evaluation" and "having corrective measure initiated," indicate the essence of a quality control program. It is important that the results of the control chart be subjected to a daily analysis in order to detect not only the out-of-control points but also any other manifestation of lack of randomness as shown by a time sequence of daily observations. It is always better and more economical to prevent a disaster than to take drastic measures to cure one. Since each test method should be subjected to quality control, the control charts should be prominently displayed at the location where the test is performed, not only to facilitate the logging of results as soon as they are obtained but also to give the technician responsible for the test an easy graphical representation of the time sequence of events. In addition, preprinted forms containing the relevant classification should be available for easy recording of information such as names, dates, time of day, reagent lot number, etc.

When all the pertinent data provided by the control charts are available, the supervisor, or section manager, should have all the meaningful information required to take corrective measures as soon as a source of trouble has been detected. Monthly or periodic review of the results, as performed by a central organization with the aid of existing computer programs, is important to provide the laboratory director with an important management tool, since the output of these programs may include such items as costs, inter- and intra-laboratory averages, historical trends, etc. However, as pointed out by Walter Shewhart[10] and other practitioners of quality control, the most important use of the control chart occurs where the worker is, and it should be continuously evaluated at that location as soon as a new point is displayed on the chart.

References

1. CHEMICAL RUBBER PUBLISHING COMPANY. 1974. Handbook of chemistry and physics. 55th ed. Cleveland, Ohio.
2. MANDEL, J. 1964. The statistical analysis of experimental data. Interscience–Wiley, New York.
3. NATRELLA, M. G. 1963. Experimental statistics. Natl. Bur. Stand. Handb. 91, Washington, D.C.
4. DAVIES, O. L., and P. GOLDSMITH, eds. 1972. Statistical methods in research and production. Oliver & Boyd, Hafner, New York.
5. SNEDECOR, G. W., and W. G. COCHRAN. 1972. Statistical methods. Iowa State Univ. Press, Ames.
6. PROSCHAN, F. 1969. Confidence and tolerance intervals for the normal distribution. *In* H. H. Ku, ed., Precision measurement and calibration, statistical concepts and procedures. Natl. Bur. Stand. Tech. Publ. 300, vol. 1. Washington, D.C.
7. MANDEL, J. 1971. Repeatability and reproducibility. Mater. Res. Stand. 11(8): 8–16.
8. GALEN, R. S., and S. R. GAMBINO. 1975. Beyond normality, the predictive value and efficiency of medical diagnoses. Wiley, New York.
9. AMERICAN SOCIETY FOR QUALITY CONTROL, STATISTICAL TECHNICAL COMMITTEE. 1973. Glossary and tables for statistical quality control. Milwaukee, WI.
10. SHEWHART, W. A. 1931. Economic control of manufactured product. Van Nostrand, New York.
11. GRANT, E. L., and R. S. LEAVENWORTH. 1972. Statistical quality control. McGraw–Hill, New York.
12. DUNCAN, A. J. 1974. Quality control and industrial statistics. Richard D. Irwin, Homewood, IL.
13. HAVEN, G. T. 1974. Outline for quality control decisions. The Pathologist 28: 373–378.

U.S. GPO: 1986-491-070/40047