1 Basic Definitions and Concepts

Statistics has its own vocabulary. Many of the terms that comprise statistical nomenclature are familiar: some commonly used in everyday language, with perhaps, somewhat different connotations. Precise definitions are given in this chapter so that no ambiguity will exist when the words are used in subsequent chapters. Specifically, such terms as *discrete* and *continuous variables, frequency distribution, population, sample, mean, median, standard deviation, variance, coefficient of variation (CV), range, accuracy,* and *precision* are introduced and defined. The methods of calculation of different kinds of means, the median, standard deviation, and range are also presented. When studying any discipline, the initial efforts are most important. The first chapters of this book are important in this regard. Although most of the early concepts are relatively simple, a firm grasp of this material is essential for understanding the more difficult material to follow.

1.1 VARIABLES AND VARIATION

Variables are the measurements, the values, which are characteristic of the data collected in experiments. These are the data that will usually be displayed, analyzed, and interpreted in a research report or publication. In statistical terms, these observations are more correctly known as *random variables.* Random variables take on values, or numbers, according to some corresponding probability function. Although we will wait until chapter 3 to discuss the concept of probability, for the present we can think of a random variable as the typical experimental observation that we, as scientists, deal with on a daily basis. Because these measurements may take on different values, repeat measurements observed under apparently identical conditions do not, in general, give the identical results (i.e., they are usually not exactly reproducible). Duplicate determinations of serum concentration of a drug one hour after an injection will not be identical no matter if the duplicates come from (a) the same blood sample or (b) from separate samples from two different persons or (c) from the same person on two different occasions. Variation is an inherent characteristic of experimental observations. To isolate and to identify particular causes of variability require special experimental designs and analysis. Variation in observations is due to a number of causes. For example, an assay will vary depending on

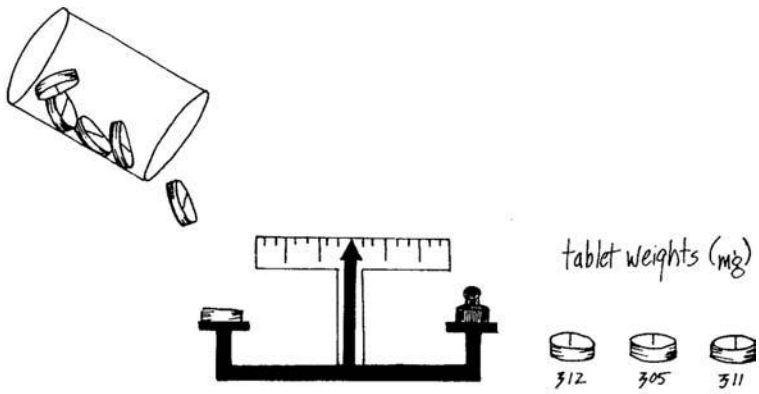
1. the instrument used for the analysis;
2. the analyst performing the assay;
3. the particular sample chosen;
4. unidentified, uncontrollable background error, commonly known as "noise."

This inherent variability in observation and measurement is a principal reason for the need of statistical methodology in experimental design and data analysis. In the absence of variability, scientific experiments would be short and simple: interpretation of experimental results from well-designed experiments would be unambiguous. In fact, without variability, single observations would often be sufficient to define the properties of an object or a system. Since few, if any, processes can be considered absolutely invariant, statistical treatment is often essential for summarizing and defining the nature of data, and for making decisions or inferences based on these variable experimental observations.

1.1.1 Continuous Variables

Experimental data come in many forms.* Probably the most commonly encountered variables are known as *continuous variables.* A continuous variable is one that can take on *any* value within some range or interval (i.e., within a specified lower and upper limit). The limiting factor for the total number of possible observations or results is the sensitivity of the measuring instrument. When weighing tablets or making blood pressure measurements, there are an infinite number of possible values that can be observed if the measurement could be made to an unlimited number of decimal places. However, if the balance, for example, is sensitive only to the nearest milligram, the data will appear as discrete values. For tablets targeted at 1 g and weighed to the nearest milligram, the tablet weights might range from 900 to 1100 mg, a total of 201 possible integral values (900, 901, 902, 903, . . ., 1098, 1099, 1100). For the same tablet weighed on a more sensitive balance, to the nearest 0.1 mg, values from 899.5 to 1100.4 might be possible, a total of 2010 possible values, and so on.

Often, continuous variables cannot be easily measured but can be ranked in order of magnitude. In the assessment of pain in a clinical study of analgesics, a patient can have a continuum of pain. To measure pain on a continuous numerical scale would be difficult. On the other hand, a patient may be able to differentiate slight pain from moderate pain, moderate pain from severe pain, and so on. In analgesic studies, scores are commonly assigned to pain severity, such as no pain = 0, slight pain = 1, moderate pain = 2, and severe pain = 3. Although the scores cannot be thought of as an exact characterization of pain, the value 3 does represent more intense pain than the values 0, 1, or 2. The scoring system above is a representation of a continuous variable by discrete "scores" that can be rationally ordered or ranked from low to high. This is commonly known as a rating scale, and the ranked data are on an ordinal scale. The rating scale is an effort to quantify a continuous, but subjective, variable.



Tablet weights: an example of a variable measurement (a random variable).

1.1.2 Discrete Variables

In contrast to continuous variables, *discrete variables* can take on a countable number of values. These kinds of variables are commonly observed in biological and pharmaceutical experiments and are exemplified by measurements such as the number of anginal episodes in one week or the number of side effects of different kinds after drug treatment. Although not continuous, discrete data often have values associated with them that can be numerically ordered according to their magnitude, as in the examples given earlier of a rating scale for pain and the number of anginal episodes per week.

Discrete data that can be named (nominal), categorized into two or more classes, and counted are called categorical variables, or *attributes*; for example, the attributes may be different

* For a further discussion of different kinds of variables, see section 15.1.

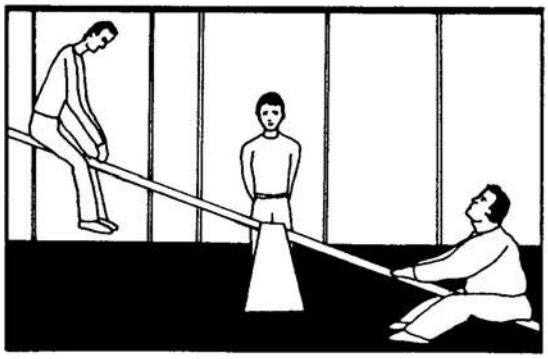
side effects resulting from different drug treatments or the presence or absence of a defect in a finished product. These kinds of data are frequently observed in clinical and pharmaceutical experiments and processes. A finished tablet classified in quality control as "defective" or "not defective" is an example of a categorical or attribute type of variable. In clinical studies, the categorization of a patient by sex (male or female) or race is a classification according to attributes. When calculating ED_{50} or LD_{50}, animals are categorized as "responders" or "nonresponders" to various levels of a therapeutic agent, a categorical response. These examples describe variables that cannot be ordered. A male is not associated with a higher or lower numerical value than a female.

Continuous variables can always be classified into discrete classes where the classes are ordered. For example, patients can be categorized as "underweight," "normal weight," or "overweight" based on criteria such as those listed in Metropolitan Life Insurance tables of "Desirable Weights for Men and Women" [l]. In this example, "overweight" represents a condition that is greater than "underweight."

Thus we can roughly classify data as

1. continuous (blood pressure, weight);
2. discrete, associated with numbers and ordered (number of anginal episodes per week);
3. attributes: categorical, ordered (degree of overweight);
4. attributes: categorical, not ordered (male or female).

Classification by attributes: patients categorized by weight.



Underweight Normal weight Overweight

1.2 FREQUENCY DISTRIBUTIONS AND CUMULATIVE FREQUENCY DISTRIBUTIONS

1.2.1 Frequency Distributions

An important function of statistics is to facilitate the comprehension and meaning of large quantities of data by constructing simple data summaries. The *frequency distribution* is an example of such a data summary, a table or categorization of the frequency[†] of occurrence of variables in various class intervals. Sometimes a frequency distribution of a set of data is simply called a "distribution." For a sampling of continuous data, in general, a frequency distribution is constructed by classifying the observations (variables) into a number of discrete intervals. For categorical data, a frequency distribution is simply a listing of the number of observations in each class or category, such as 20 males and 30 females entered in a clinical study. This procedure results in a more manageable and meaningful presentation of the data.

[†] The frequency is the number of observations in a specified interval or class: for example, tablets weighing between 300 and 310 mg, or the number of patients who are female.

Table 1.1 Serum Cholesterol Changes (mg%) for 156 Patients After Administration of a Drug Tested for Cholesterol-Lowering Effect[a]

17	−12	25	−37	−29	−39
−22	0	−22	−63	34	−31
−64	−12	−49	5	−8	33
−50	−7	16	−11	−38	−17
0	−9	−21	1	2	−30
−32	−34	−14	−18	5	6
24	−6	−49	−8	−49	−37
−25	−12	14	10	−41	−66
−31	35	21	−19	−27	17
−6	−17	−6	1	−28	40
−31	17	−54	−27	−16	16
−44	10	−3	−3	5	6
−19	9	−10	−20	−9	−8
−10	−11	11	−39	19	−32
4	−15	−18	35	6	20
46	24	−27	−19	5	−60
27	23	−22	−1	12	−27
−13	−39	39	−34	−97	−26
38	14	−47	8	26	−15
−62	12	−53	11	21	−47
−54	−11	−5	0	55	34
−69	−11	−44	20	−50	19
0	−25	−24	−4	14	2
−34	16	−23	−71	−58	9
9	2	−2	−58	13	14
17	−13	−22	−3	−17	1

[a]A negative number means a decrease and a positive number means an increase.

Table 1.1 is a tabulation of serum cholesterol changes resulting from the administration of a cholesterol-lowering agent to a group of 156 patients. The data are presented in the order in which results were reported from the clinic.

A frequency distribution derived from the 156 cholesterol values is shown in Table 1.2. This table shows a tabulation of the frequency, or number, of occurrences of values that fall into the various class intervals of "serum cholesterol changes." Clearly, the condensation of the data as shown in the frequency distribution in Table 1.2 allows for a better "feeling" of the experimental results than do the raw data represented by the individual 156 results. For example, one can readily see that most of the patients had a lower cholesterol value in response to the drug (a negative change) and that most of the data lie between −60 and +19 mg%.

When constructing a frequency distribution, two problems must be addressed. The first problem is how many classes or intervals should be constructed, and the second problem is the specification of the width of each interval (i.e., specifying the upper and lower limit of each interval). There are no definitive answers to these questions. The choices depend on the nature

Table 1.2 Frequency Distribution of Serum Cholesterol Changes

Class interval		Frequency
−100 to −81	(−100.5 to −80.5)	1
−80 to −61	(−80.5 to −60.5)	6
−60 to −41	(−60.5 to −40.5)	16
−40 to −21	(−40.5 to −20.5)	31
−20 to −1	(−20.5 to −0.5)	40
+0 to +19	(−0.5 to + 19.5)	43
+20 to +39	(+19.5 to +39.5)	16
+40 to +59	(+39.5 to +59.5)	3

Data taken from Table 1.1.

Table 1.3 Frequency Distribution of Serum Cholesterol Changes Using 16 Class Intervals

Class interval	Frequency
−100 to −91	1
−90 to −81	0
−80 to −71	1
−70 to −61	5
−60 to −51	6
−50 to −41	10
−40 to −31	14
−30 to −21	17
−20 to −11	22
−10 to −1	18
0 to +9	22
−10 to +19	21
+20 to +29	9
+30 to +39	7
+40 to +49	2
+50 to +59	1

of the data and good judgment. The number of intervals chosen should result in a table that considerably improves the readability of the data. The following rules of thumb are useful to help select the intervals for a frequency table:

1. Choose intervals that have significance in relation to the nature of the data. For example, for the cholesterol data, intervals such as 18 to 32 would be cumbersome and confusing. Intervals of width 10 or 20, such as those in Tables 1.2 and 1.3, are more easily comprehended and manipulated arithmetically.
2. Try not to have too many empty intervals (i.e., intervals with no observations). The half of the total number of intervals that contain the least number of observations should contain at least 10% of the data. The intervals with the least number of observations in Table 1.2 are the first two intervals (−100 to −81 and −80 to −61) and the last two intervals (+ 20 to +39 and +40 to +59) (one-half of the eight intervals), which contain 26% or 17% of the 156 observations.
3. Eight to twenty intervals are usually adequate.

Table 1.3 shows the same 156 serum cholesterol changes in a frequency table with 16 intervals. Which table gives you a better feeling for the results of this study, Table 1.2 or Table 1.3? (See also Exercise Problem 3.)

The width of all the intervals, in general, should be the same. This makes the table easy to read and allows for simple computations of statistics such as the mean and standard deviation. The intervals should be mutually exclusive so that no ambiguity exists when classifying values. In Tables 1.2 and 1.3, we have defined the intervals so that a value can be categorized only in one class interval. In this way, we avoid problems that can arise when observations are exactly equal to the boundaries of the class intervals. If the class intervals were defined so as to be continuous, such as −100 to −90, −90 to −80, −80 to −70, and so on, one must define the class to which a borderline value belongs, either the class below or the class above, a priori. For example, a value of −80 might be defined to be in the interval −80 to −70.

Another way to construct the intervals is to have the boundary values have one more "significant figure" than the actual measurements so that none of the values can fall on the boundaries. The extra figure is conveniently chosen as 0.5. In the cholesterol example, measurements were made to the nearest mg%; all values are whole numbers. Therefore, two adjacent values can be no less different than 1 mg%, +10, and +11, for example. The class intervals could then have a decimal of 0.5 at the boundaries, which means that no value can fall exactly on a boundary value. The intervals in parentheses in Table 1.2 were constructed in this manner. This categorization, using an extra figure that is halfway between the two closest possible values,

makes sense from another point of view. After rounding off, a value of +20 can be considered to be between 19.5 and 20.5, and would naturally be placed in the interval 19.5 to 39.5, as shown in Table 1.2.

1.2.2 Stem-and-Leaf Plot

An expeditious and compact way of summarizing and tabulating large amounts of data, by hand, known as the *stem-and-leaf method* [2], is best illustrated with an example. We will use the data from Table 1.1 to demonstrate the procedure.

An ordered series of integers is conveniently chosen (see below) to cover the range of values. The integers consist of the first digit(s) of the data, as appropriate, and are arranged in a vertical column, the "stem." By adding another digit(s) to one of the integers in the stem column (the "leaves"), we can tabulate the data in class intervals as in a frequency table. For the data of Table 1.1, the numbers range from approximately −100 to +60. The stem is conveniently set up as follows:

−10	−7	−4	−1	+1	+4
−9	−6	−3	−0	+2	+5
−8	−5	−2	+0	+3	+6

In this example, the stem is the first digit(s) of the number and the leaf is the last digit. The first value in Table 1.1 is 17. Therefore, we place a 7 (leaf) next to the + 1 in the stem column. The next value in Table 1.1 is −22. We place a 2 (leaf) next to −2 in the stem column, and so on. Continuing this process for each value in Table 1.1 results in the following stem-and-leaf diagram.

−10																					
−9	7																				
−8																					
−7	1																				
−6	4	2	9	3	6	0															
−5	0	4	4	3	8	0	8														
−4	4	9	9	7	4	1	9	7													
−3	2	1	1	4	4	9	7	9	4	8	9	1	0	7	2						
−2	2	5	5	2	1	7	2	4	3	2	7	0	9	7	8	6	7				
−1	9	0	3	2	2	2	7	1	5	1	1	3	4	0	8	1	8	9	9	6	7	7	5
−0	6	7	9	6	6	3	5	2	8	3	1	4	3	8	9	8					
+0	0	4	0	9	0	9	2	5	1	1	8	0	2	5	5	6	5	6	6	2	9	1	
+1	7	7	0	7	4	2	6	6	4	1	0	1	9	2	6	4	3	7	6	9	4		
+2	4	7	4	3	5	1	0	1	0														
+3	8	9	9	5	4	3	4																
+4	6	0																					
+5	5																						
+6																							

This is a list of all the values in Table 1.1. The distribution of this data set is easily visualized with no further manipulation. However, if necessary, one can easily construct a frequency distribution from the configuration of data resulting from the stem-and-leaf tabulation. (Note that all categories in this particular example can contain as many as 10 different numbers except for the −0 category, which can contain only 9 numbers, −1 to −9 inclusive. This "anomaly" occurs because of the presence of both positive and negative values and the value 0. In this example, 0 is arbitrarily assigned a positive value.) In addition to the advantages of this tabulation noted above, the data are in the form of a histogram, which is a common way of graphically displaying data distributions (see chap. 2).

Table 1.4 Frequency Distribution of Tablet Potencies

Potency (mg)	W_i[a]	Frequency X_i[b]
89.5–90.5	1	90
90.5–91.5	0	91
91.5–92.5	2	92
92.5–93.5	1	93
93.5–94.5	5	94
94.5–95.5	1	95
95.5–96.5	2	96
96.5–97.5	7	97
97.5–98.5	10	98
98.5–99.5	8	99
99.5–100.5	13	100
100.5–101.5	17	101
101.5–102.5	13	102
102.5–103.5	9	103
103.5–104.5	0	104
104.5–105.5	0	105
105.5–106.5	5	106
106.5–107.5	4	107
107.5–108.5	0	108
108.5–109.5	0	109
109.5–110.5	2	110
	$\sum W_i = 100$	

[a] W_i is the frequency.
[b] X_i is the midpoint of the interval.

1.2.3 Cumulative Frequency Distributions

A large set of data can be conveniently displayed using a cumulative frequency table or plot. The data are first ordered and, with a large data set, may be arranged in a frequency table with n class intervals. The frequency, often expressed as a proportion (or percentage), of values equal to or less than a given value, X_i, is calculated for each specified value of X_i, where X_i is the upper point of the class interval ($i = 1$ to n). A plot of the cumulative proportion versus X can be used to determine the proportion of values that lie in some interval, that is, between some specified limits. The cumulative distribution for the tablet potencies in Table 1.4 is shown in Table 1.5 and

Table 1.5 Cumulative Frequency Distribution of Tablet Potencies

Potency, X_t(mg)[a]	Cumulative frequency ($\leq X$)	Cumulative proportion
90.5	1	0.01
92.5	3	0.03
93.5	4	0.04
94.5	9	0.09
95.5	10	0.10
96.5	12	0.12
97.5	19	0.19
98.5	29	0.29
99.5	37	0.37
100.5	50	0.50
101.5	67	0.67
102.5	80	0.80
103.5	89	0.89
106.5	94	0.94
107.5	98	0.98
110.5	100	1.00

Data taken from Table 1.4.
[a] X_t is the upper point of the class interval in Table 1.4, excluding null intervals.

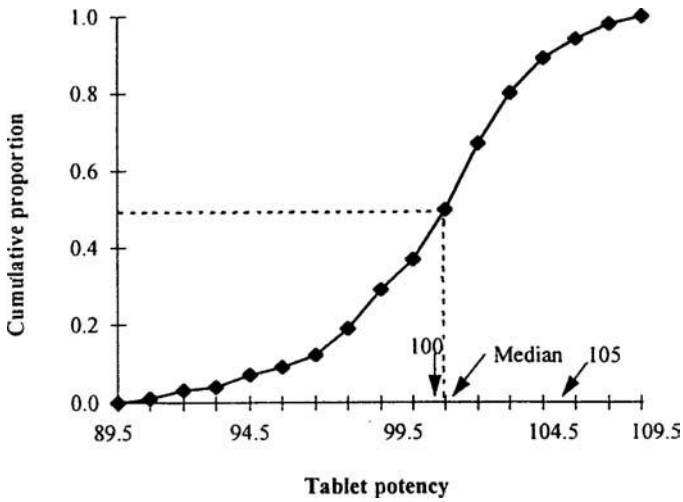


Figure 1.1 Cumulative proportion plot for data in Table 1.5 (tablet potencies).

plotted in Figure 1.1. The cumulative proportion represents the proportion of values less than or equal to X_i (e.g., 29% of the values are less than or equal to 98.5). Also, for example, from an inspection of Figure 1.1, one can estimate the proportion of tablets with potencies between 100 and 105 mg inclusive, equal to approximately 0.48 (0.91 at 105 mg minus 0.43 at 100 mg). (See also Exercise Problem 5.)

The cumulative distribution is a very important concept in statistics. In particular, the application of the cumulative normal distribution, which is concerned with continuous data, will be discussed in chapter 3. A more detailed account of the construction and interpretation of frequency distributions is given in Refs. [3–5].

1.3 SAMPLE AND POPULATION

Understanding the concepts of samples and populations is important when discussing statistical procedures. *Samples* are usually a relatively small number of observations taken from a relatively large *population* or universe. The sample values are the observations, the data, obtained from the population. The population consists of data with some clearly defined characteristic(s). For example, a population may consist of all patients with a particular disease, or tablets from a production batch. The sample in these cases could consist of a selection of patients to participate in a clinical study, or tablets chosen for a weight determination. The sample is only part of the available data. In the usual experimental situation, we make observations on a relatively small sample in order to make inferences about the characteristics of the whole, the population. The totality of available data is the population or universe. When designing an experiment, the population should be clearly defined so that samples chosen are representative of the population. This is important in clinical trials, for example, where inferences to the treatment of disease states are crucial. The exact nature or character of the population is rarely known, and often impossible to ascertain, although we can make assumptions about its properties. Theoretically, a population can be finite or infinite in the number of its elements. For example, a finished package contains a finite number of tablets; all possible tablets made by a particular process, past, present, and future, can be considered infinite in concept. In most of our examples, the population will be considered to be infinite, or at least very large compared to the sample size. Table 1.6 shows some populations and samples, examples that should be familiar to the pharmaceutical scientist.

1.3.1 Population Parameters and Sample Statistics

"Any measurable characteristic of the universe is called a *parameter*" [6]. For example, the average weight of a batch of tablets or the average blood pressure of hypertensive persons in the United States are parameters of the respective populations. Parameters are generally

Table 1.6 Examples of Samples and Populations

Population	Sample
Tablet batch	Twenty tablets taken for content uniformity
Normal males between ages 18 and 65 years available to hospital	Twenty-four subjects selected for a phase I clinical study
Sprague–Dawley weaning rats	100 rats selected to test possible toxic effects of a new drug candidate
Analysts working for company *X*	Three analysts from a company to test a new assay method
Persons with diastolic blood pressure between 105 and 120 mm Hg in the United States	120 patients with diastolic pressure between 105 and 120 mm Hg to enter clinical study to compare two antihypertensive agents
Serum cholesterol levels of one patient	Blood samples drawn once a week for 3 months from a single patient

denoted by Greek letters; for example, the mean of the population is denoted as μ. Note that parameters are characteristic of the population, and are values that are usually unknown to us.

Quantities derived from the sample are called *sample statistics.* Corresponding to the true average weight of a batch of tablets is the average weight for the small sample taken from the population of tablets. We should be very clear about the nature of samples. Emphasis is placed here (and throughout this book) on the variable nature of such sample statistics. A parameter, for example, the mean weight of a batch of tablets, is a fixed value; it does not vary. Sample statistics are variable. Their values depend on the particular sample chosen and the variability of the measurement. The average weight of 10 tablets will differ from sample to sample because

1. we choose 10 different tablets at each sampling;
2. the balance (and our ability to read it) is not exactly reproducible from one weighing to another.

An important part of the statistical process is the characterization of a population by estimating its parameters. The parameters can be estimated by evaluating suitable sample statistics. The reader will probably have little trouble in understanding that the average weight of a sample of tablets (a sample statistic) estimates the true mean weight (a parameter) of the batch. This concept is elucidated and expanded in the remaining sections of this chapter.

1.4 MEASURES DESCRIBING THE CENTER OF DATA DISTRIBUTIONS

1.4.1 The Average

Probably the most familiar statistical term in popular use is the *average*, denoted by $\overline{X}$ (X bar). The average is also commonly known as the *mean* or *arithmetic average.* The average is a summarizing statistic and is a measure of the center of a distribution, particularly meaningful if the data are symmetrically distributed below and above the average. Symbolically, the mean is equal to

$$\frac{\sum_{i=1}^{N} X_i}{N} \tag{1.1}$$

the sum of the observations divided by the number of observations. $\sum_{i=1}^{N} X_i$ is the sum of the N values, each denoted by X_i, (X_1, X_2, ..., X_n), where i can take on the values 1, 2, 3, 4, ..., n.[‡]

[‡] For the most part, when using summation notation in this book, we will not use the full notation, such as $\sum_{i=1}^{N} X_i$, but rather $\sum X$, the i notation being implied, unless otherwise stated.

The average of the values 7, 11, 6, 5, and 4 is

$$\frac{7 + 11 + 6 + 5 + 4}{5} = 6.6.$$

This is an unweighted average, each value contributing equally to the average.

1.4.2 Other Kinds of Averages

When averaging observations, we usually think of giving each observation equal weight. The usual formula for the average ($\sum X_i/N$) gives each value equal weight. If we believe that the values to be averaged do not carry the same weight, then we should use a weighted average. The average of three cholesterol readings 210, 180, and 270 is $(660)/3 = 220$. Suppose that the value of 210 is really the average of two values (200 and 220), we might want to consider giving this value twice as much weight as the other two values, resulting in an average

$$\frac{210 + 210 + 180 + 270}{4} = 217.5$$

or

$$\frac{2 \times 210 + 180 + 270}{2 + 1 + 1} = 217.5.$$

The formula for a weighted average, $\overline{X}_w$ is

$$\frac{\sum W_i X_i}{\sum W_i}, \tag{1.2}$$

where W_i is the weight assigned to the value X_i. The weights for the calculation of a weighted average are often the number of observations associated with the values X_i. This concept is illustrated for the calculation of the average for data categorized in the form of a frequency distribution. Table 1.4 shows a frequency distribution of 100 tablet potencies. The frequency is the number of observations of tablets in a given class interval, as defined previously. The frequency or number of tablets in a "potency" interval is the *weight* used in the computation of the weighted average. The value X associated with the weight is taken as the midpoint of the interval; for example, for the first interval, 89.5 to 90.5, $X_1 = 90$. Applying Eq. (1.2), the weighted average is $\sum W_i X_i / \sum W_i$:

$$\frac{1 \times 90 + 0 \times 91 + 2 \times 92 + 1 \times 93 + 5 \times 94 + \cdots + 4 \times 107 + 2 \times 110}{1 + 0 + 2 + 1 + 5 + \cdots + 4 + 2},$$

which equals $10{,}023/100 = 100.23$ mg.

It is not always obvious when to use a weighted average, and one should have a substantial knowledge of the circumstances and nature of the data in order to make this decision. In the previous example, if the 210 value (the average of two observations) came from one patient and the other values were single observations from two different patients, one may not want to use a weighted average. The reasoning in this example may be that this average is meant to represent the true average cholesterol of these three patients, each with different cholesterol levels. There does not seem to be a good reason to give twice as much weight to the "210" patient because that patient happened to have two readings. This may be more clearly seen if the patient had 100 readings and the other two patients only a single reading. The unweighted average would be very close to the average of the patient with the 100 readings and would not represent the average of the three patients. In this example, the average of three values (one value for each patient) would be a better representation of the average, $(210 + 180 + 270)/3 = 220$.

Table 1.7 Distribution of Particle Size of Powder

Midpoint Sieve size	Log sieve Size (*Y*)	Weight (*W*)	(*WT*) × (*Y*)
10[a]	2.3026	19.260	44.3478
30	3.4012	24.015	81.6797
50	3.1920	22.240	87.0034
70	4.2485	7.525	31.9699
90	4.4998	6.515	29.3163
150[b]	5.0106	*20.445*	*102.4424*
Sum		100.00	376.7595

[a] 10 is for sieve size less than 20, that is, between 0 and 20.
[b] 150 is substituted for >100.

If the four values were obtained from one patient where the 210 average came from one laboratory and the other two values from two different laboratories, the following reasoning might be useful to understand how to treat the data properly. If the different laboratories used the same analytical method that was expected to yield the same result, a weighted average would be appropriate (give twice the weight to the 210 value). If the laboratories have different methods that give different results for the same sample, an unweighted average may be more appropriate.

The distribution of particle size of a powdered blend is often based on the logarithm of the particle size (see sect. 10.1.1). The quantity (weight) of powder in a given interval of particle size may be considered a weighting factor when computing the average particle size. Table 1.7 shows the particle size distribution (frequency distribution) of a powder, where the class intervals are based on the logarithm of the sieve size fractions. The weighted average can be calculated as

$$\overline{X}_{\mathrm{w}} = \frac{\sum \text{weight} \times (\text{log sieve size})}{\sum (\text{weights})}. \tag{1.3}$$

The weight is the percentage of powder found for a given particle size (or interval of sieve sizes). Note that for this example, the sieve size is taken as the midpoint of the untransformed class (sieve size) interval.

From Eq. (1.3), weighted average = 376.7595/100.0 = 3.7676. Since sieve size is in log terms, the antilog of 3.7676 = 43.3 is an estimate of the average particle size. (For *more* advanced methods of estimating the parameters of particle size distributions, see Refs. [7,8].)

The calculation of the variance of a weighted average is dependent on the nature of the weighted average and an experienced statistician should be consulted if necessary (see SAS manual for options). This more advanced concept is discussed further in section 1.5.5.

Two other kinds of averages that are sometimes found in statistical procedures are the geometric and harmonic means. The *geometric mean* is defined as

$$\sqrt[n]{X_1 \cdot X_2 \cdot X_3 \cdots X_n}$$

or the nth root of the product of n observations.

The geometric mean of the numbers 50, 100, and 200 is

$$\sqrt[3]{50 \cdot 100 \cdot 200} = \sqrt[3]{1{,}000{,}000} = 100.$$

If a measurement of population growth shows 50 at time 0, 100 after one day, and 200 after two days, the geometric mean (100) is more meaningful than the arithmetic mean (116.7). The geometric mean is always less than or equal to the arithmetic mean, and is meaningful for data with logarithmic relationships. (See also sect. 15.1.1.) Note that the logarithm of $\sqrt[3]{50 \cdot 100 \cdot 200}$ is equal to $[\log 50 + \log 100 + \log 200]/3$, which is the average of the logarithms

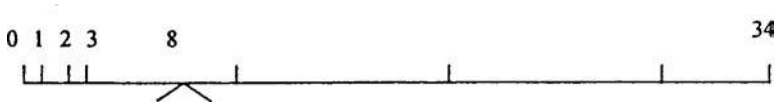


Figure 1.2 Average illustrated as balancing forces.

of the observations. The geometric mean is the antilog of this average (the antilog of the average is 100).

The harmonic mean is the appropriate average following a reciprocal transformation (chap. 10). The harmonic mean is defined as

$$\frac{N}{\sum 1/X_i}.$$

For the three observations 2, 4, and 8 ($N = 3$), the harmonic mean is

$$\frac{3}{1/2 + 1/4 + 1/8} = 3.429.$$

1.4.3 The Median

Although the average is the most often used measure of centrality, the *median* is also a common measure of the center of a data set. When computing the average, very large or very small values can have a significant effect on the magnitude of the average. For example, the average of the numbers 0, 1, 2, 3, and 34 is 8. The arithmetic average acts as the fulcrum of a balanced beam, with weights placed at points corresponding to the individual values, as shown in Figure 1.2. The single value 34 needs four values, 0, 1, 2, and 3, as a counterbalance. Also, the median may be a more appropriate measure of central tendency for skewed distributions such as the log-normal distribution (see sect. 10.1.1).

The *median* represents the center of a data set, without regard for the distance of each point from the center. The median is the value that divides the data in half, half the values being less than and half the values greater than the median value. The median is easily obtained when the data are ranked in order of magnitude. The median of an odd number of *different*[§] observations is the middle value. For $2N + 1$ values, the median is the $(N + 1)$th ordered value. The median of the data 0, 1, 2, 3, and 34 is the third (middle) value, $2(N = 2, 2N + 1 = 5$ values). By convention, the median for an even number of data points is considered to be the average of the two center points. For example, the median of the numbers, 0, 1, 2, and 3 is the average of the center points, 1 and 2, equal to $(1 + 2)/2 = 1.5$. The median is often used as a description of the center of a data set when the data have an asymmetrical distribution. In the presence of either extremely high or extremely low outlying values, the median appears to describe the distribution better than does the average. The median is more stable than the average in the presence of extreme observations. A very large or very small value has the same effect on the calculation of the median as any other value, larger or smaller than the median, respectively. On the other hand, as noted previously, very large and very small values have a significant effect on the magnitude of the mean.

The distribution of individual yearly incomes, which have relatively few very large values (the multimillionaires), serves as a good example of the use of the median as a descriptive statistic. Because of the large influence of these extreme values, the average income is higher than one might expect on an intuitive basis. The median income, which *is* less than the average income, represents a figure that is readily interpreted; that is, one-half of the population earns more (or less) than the median income.

The distribution of particle sizes for bulk powders used in pharmaceutical products *is* often skewed. In these cases, the median is a better descriptor of the centrality of the distribution than

[§] If the median value is not unique, that is, two or more values are equal to the median, the median is calculated by interpolation (3).

is the mean [9]. The median is less efficient than the mean as an estimate of the center of a distribution; that is, the median is more variable [10]. For most of the problems discussed in this book, we will be concerned with the mean rather than the median as a measure of centrality.

An interesting, but not well documented, relationship between the mean and median shows that for positive numbers, the mean must be greater than half the median. This can be proven simply as follows:

Consider $2N + 1$ numbers whose median is "M" and mean is "m." We will choose an odd number of values so that the median is well defined. The mean, m, is the sum of all the numbers divided by $2N + 1$. Of the $2N + 1$ numbers, $N + 1$ is greater than or equal to the median, M. Therefore, m is greater than or equal to $(N + 1)M/(2N + 1)$. But $(N + 1)/(2N + 1) > {}^1/_2$. Therefore, $m > M/2$. Therefore the mean must be greater than half the median.

For example, consider the following extreme example. The data consist of the following values: 1, 1, 1, 999.5, 1000, 10,001,000. The median is 999.5. The mean is 571.8. 571.8 is greater than 999.5/2.

The median is also known as the *50th percentile* of a distribution. To compute percentiles, the data are ranked in order of magnitude, from smallest to largest. The nth percentile denotes a value below which n% of the data are found, and above which $(100 - n)$ % of the data are found. The 10th, 25th, and 75th percentiles represent values below which 10%, 25%, and 75%, respectively, of the data occur. For the tablet potencies shown in Table 1.5, the 10th percentile is 95.5 mg; 10% of the tablets contain less than 95.5 mg and 90% of the tablets contain more than 95.5 mg of drug. The 25th, 50th, and 75th percentiles are also known as the first, second, and third quartiles, respectively.

The *mode* is less often used as the central, or typical, value of a distribution. The mode is the value that occurs with the greatest frequency. For a symmetrical distribution that peaks in the center, such as the normal distribution (see chap. 3), the mode, median, and mean are identical. For data skewed to the right (e.g., incomes), which contain a relatively few very large values, the mean is larger than the median, which is larger than the mode (Fig. 10.1).

1.5 MEASUREMENT OF THE SPREAD OF DATA

The mean (or median) alone gives no insight or information about the spread or range of values that comprise a data set. For example, a mean of five values equal to 10 may comprise the numbers

0, 5, 10, 15, and 20 or 5, 10, 10, 10, and 15.

The mean, coupled with the *standard deviation* or *range*, is a succinct and minimal description of a group of experimental observations or a data distribution. The standard deviation and the range are measures of the spread of the data; the larger the magnitude of the standard deviation or range, the more spread out the data are. A standard deviation of 10 implies a wider range of values than a standard deviation of 3, for example.

1.5.1 Range

The *range*, denoted as R, is the difference between the smallest and the largest values in the data set. For the data in Table 1.1, the range is 152, from −97 to +55 mg%. The range is based on only two values, the smallest and largest, and is more variable than the standard deviation (i.e., it is less stable).

1.5.2 Standard Deviation and Variance

The *standard deviation*, denoted as s.d. or S, is calculated as

$$\sqrt{\frac{\sum (X - \overline{X})^2}{N - 1}}, \tag{1.4}$$

where N is the number of data points (or sample size) and $\sum(X - \overline{X})^2$ is the *sum of squares* of the differences of each value from the mean, $\overline{X}$. The standard deviation is more difficult to calculate than is the range.

Table 1.8 Calculation of the Standard Deviation

X	$\overline{X}$	$X - \overline{X}$	$(X - \overline{X})^2$
101.8	103	−1.2	1.44
103.2	103	0.2	0.04
104.0	103	1.0	1.00
102.5	103	−0.5	0.25
103.5	103	0.5	0.25
$\sum X = 515$			$\sum (X - \overline{X})^2 = 2.98$
s.d. $= \sqrt{\frac{\sum (X - \overline{X})^2}{N-1}} = \sqrt{\frac{2.98}{4}} = 0.86$			

Consider a group of data points: 101.8, 103.2, 104.0, 102.5, and 103.5. The mean is 103.0. Details of the calculation of the standard deviation are shown in Table 1.8. The difference between each value and the mean is calculated: $X - \overline{X}$. These differences are squared, $(X - \overline{X})^2$, and summed. The sum of the squared differences divided by $N - 1$ is calculated, and the square root of this result is the standard deviation.

With the accessibility of electronic calculators and computers, it is rare, nowadays, to hand compute a mean and standard deviation (or any other calculation, for that matter). Nevertheless, when computing the standard deviation by hand (or with the help of a calculator), a well-known shortcut computing formula is recommended. The shortcut is based on the identity

$$\sum (X - \overline{X})^2 = \sum X^2 - \frac{(\sum X)^2}{N}.$$

Therefore,

$$\text{s.d.} = \sqrt{\frac{\sum X^2 - (\sum X)^2 / N}{N - 1}}, \tag{1.5}$$

where $\sum X^2$ is the sum of each value squared and $(\sum X)^2$ is the square of the sum of all the values [$(\sum X)^2/N$ is also known as the *correction term*]. We will apply this important formula, Eq. (1.5), to the data above to illustrate the calculation of the standard deviation. This result will be compared to that obtained by the more time-consuming method of squaring each deviation from the mean (Table 1.8).

$$\sum (X - \overline{X})^2 = 101.8^2 + 103.2^2 + 104.0^2 + 102.5^2 + 103.5^2 - \frac{515^2}{5} = 2.98.$$

The standard deviation is $\sqrt{2.98/4} = 0.86$, as before.

The *variance* is the square of the standard deviation, often represented as S^2. The variance is calculated as

$$S^2 = \frac{\sum (X - \overline{X})^2}{N - 1}. \tag{1.6}$$

In the example of the data in Table 1.8, the variance, S^2, is

$$\frac{2.98}{4} = 0.745.$$

A question that often puzzles new students of statistics is: Why use $N - 1$ rather than N in the denominator in the expression for the standard deviation or variance [Eqs. (1.4) and (1.6)]?

The *variance of the population*, a parameter traditionally denoted as σ^2 (sigma squared), is calculated as[¶]:

$$\sigma^2 = \frac{\sum (X - \overline{X})^2}{N}, \tag{1.7}$$

where N is the number of all possible values in the population. The use of $N - 1$ rather than N in the calculation of the variance of a *sample* (a sample statistic) makes the sample variance an *unbiased estimate* of the population variance. Because the sample variance is variable (a random variable), in any given experiment, S^2 will not be exactly equal to the true population variance, σ^2. However, in the long run, S^2 (calculated with $N - 1$ in the denominator) will equal σ^2, on the average. "On the average" means that if samples of size N were repeatedly randomly selected from the population, and the variance calculated for each sample, the averages of these calculated variance estimates would equal σ^2. Note that the sample variance is an estimate of the true population variance σ^2.

If S^2 estimates σ^2 on the average, the sample variance is an unbiased estimate of the population variance. It can be proven that the sample variance calculated with $N - 1$ in the denominator is an unbiased estimate of σ^2. To try to verify this fact by repeating exactly the same laboratory or clinical experiment (if the population variance were known) would be impractical. However, for explanatory purposes, it is often useful to illustrate certain theorems by showing what would happen upon repeated sampling from the same population. The concept of the unbiased nature of the sample variance can be demonstrated using a population that consists of three values: 0, 1, and 2. The population variance, $\sum(X - \overline{X})^2/3$, is equal to 2/3 [see Eq. (1.7)]. Using the repeated sample approach noted above, samples of size 2 are repeatedly selected at random from this population. The first choice is replaced before selection of the second choice so that each of the three values has an equal chance of being selected on both the first and second selection. (This is known as *sampling with replacement*.) The following possibilities of samples of size 2 are equally likely to be chosen:

0, 1; 1, 0; 0, 2; 2, 0; 1, 2; 2, 1; 1, 1; 2, 2; 0, 0

The *sample variance*[**] of these nine pairs are $[\sum(X - \overline{X})^2/(N-1)]$ 0.5, 0.5, 2, 2, 0.5, 0.5, 0, 0, and 0, respectively. The average of the nine equally likely possible variances is

$$\frac{0.5 + 0.5 + 2 + 2 + 0.5 + 0.5 + 0 + 0 + 0}{9} = \frac{6}{9} = \frac{2}{3},$$

which is exactly equal to the population variance. This demonstrates the unbiased character of the sample variance. The sample standard deviation [Eq. (1.4)] is not an unbiased estimate of the *population standard deviation*, σ, which for a finite population is calculated as

$$\sqrt{\frac{\sum (X - \overline{X})^2}{N}}. \tag{1.8}$$

The observed variance is not dependent on the sample size. The sample variance will equal the true variance "on the average," but the variability of the estimated variance decreases as the sample size increases. The unbiased nature of a *sample estimate* of a *population parameter*, such as the variance or the mean, is a desirable characteristic. $\overline{X}$, the sample estimate of the true population mean, is also an unbiased estimate of the true mean. (The true mean is designated by the Greek letter μ. In general, population parameters are denoted by Greek letters as noted previously.)

[¶] Strictly speaking, this formula is for a population with a finite number of data points.

[**] For samples of size 2, the variance is simply calculated as the square of the difference of the values divided by 2, $d^2/2$. For example, the variance of 0 and 1 is $(1 - 0)^2/2 = 0.5$.

One should be aware that some calculators having a built-in function for calculating the standard deviation use N in the denominator of the formula for the standard deviation. As we have emphasized above, this is correct for the calculation of the population standard deviation (or variance), and will be close to the calculation of the sample standard deviation when N is large.

The value of $N - 1$ is also known as the *degrees of freedom* for the sample (later we will come across situations where degrees of freedom are less than $N - 1$). The concept of degrees of freedom (denoted as d.f.) is very important in statistics, and we will have to know the degrees of freedom for the variance estimates used in statistical tests to be described in subsequent chapters.

Another common misconception is that the standard deviation (or variance) of a sample becomes smaller as the sample size increases. The standard deviation of a sample is an estimate of the true standard deviation. The true standard deviation is a constant and does not change with a change in sample size. However, we can say that the estimate of the true standard deviation as observed in a sample is more reliable and less variable as the sample size increases. But, on the average, the standard deviation of a small or large sample will approximate the true standard deviation. As discussed later in this chapter (sect. 1.5.4), the standard deviation of a mean will decrease with larger sample sizes.

1.5.3 Coefficient of Variation

The variability of data may often be better described as a relative variation rather than as an absolute variation, such as that represented by the standard deviation or range. One common way of expressing the variability, which takes into account its relative magnitude, is the ratio of the standard deviation to the mean, s.d./$\overline{X}$. This ratio, often expressed as a percentage, is called the *coefficient of variation*, abbreviated as CV, or RSD, the relative standard deviation. A CV of 0.1 or 10% means that the s.d. is one-tenth of the mean. This way of expressing variability is useful in many situations. It puts the variability in perspective relative to the magnitude of the measurements and allows a comparison of the variability of different kinds of measurements. For example, a group of rats of average weight 100 g and s.d. of 10 g has the same relative variation (CV) as a group of animals with average weight 70 g and s.d. of 7 g. Many measurements have an almost constant CV, the magnitude of the s.d. being proportional to the mean. In biological data, the CV is often between 20% and 50%, and one would not be surprised to see an occasional CV as high as 100% or more. The relatively large CV observed in biological experiments is due mostly to "biological variation," the lack of reproducibility in living material. On the other hand, the variability in chemical and instrumental analyses of drugs is usually relatively small. Thus it is not unusual to find a CV of less than 1% for some analytical procedures.

1.5.4 Standard Deviation of the Mean (Standard Error of the Mean)

The s.d. is a measure of the spread of a group of individual observations, a measure of their variability. In statistical procedures to be discussed in this book, we are more concerned with making inferences about the mean of a distribution rather than with individual values. In these cases, the variability of the mean rather than the variability of individual values is of interest. The sample mean is a random variable, just as the individual values that comprise the mean are variable. Thus, repeated sampling of means from the same population will result in a distribution of means that has its own mean and s.d.

The *standard deviation of the mean*, commonly known as the *standard error of the mean*, is a measure of the variability of the mean. For example, the average potency of the 100 tablets shown in Table 1.4 may have been determined to estimate the average potency of the population, in this case, a production batch. An estimate of the variability of the mean value would be useful. The mean tablet potency is 100.23 mg and the s.d. is 3.687. To compute the s.d. of the mean (also designated as $S_{\overline{X}}$), we might assay several more sets of 100 tablets and calculate the mean potency of each sample. This repeated sampling would result in a group of means, each composed of 100 tablets, with different values, such as the five means shown in Table 1.9. The s.d. of this group of means can be calculated in the same manner as the individual values are calculated

Table 1.9 Means of Potencies of Five Sets of 100 Tablets Selected from a Production Batch

Sample	Mean potency
1	99.84
2	100.23
3	100.50
4	100.96
5	100.07

[Eq. (1.4)]. The s.d. of these five means is 0.431. We can anticipate that the s.d. of the means will be considerably smaller than the s.d. calculated from the 100 individual potencies. This fact is easily comprehended if one conceives of the mean as "averaging out" the extreme individual values that may occur among the individual data. The means of very large samples taken from the same population are very stable, tending to cluster closer together than the individual data, as illustrated in Table 1.9.

Fortunately, we do not have to perform real or simulated sampling experiments, such as weighing five sets of 100 tablets each, to obtain replicate data in order to estimate the s.d. of means. Statistical theory shows that the s.d. of mean values is equal to the s.d. calculated from the individual data divided by $\sqrt{N}$, where N is the sample size[††]:

$$S_{\overline{X}} = \frac{S}{\sqrt{N}}. \tag{1.9}$$

The s.d. of the numbers shown in Table 1.4 is 3.687. Therefore, the s.d. of the mean for the potencies of 100 tablets shown in Table 1.4 is estimated as $S/\sqrt{N} = 3.687/\sqrt{100} = 0.3687$. This theory verifies our intuition; the s.d. of means is smaller than the s.d. of the individual data points. The student should not be confused by the two estimates of the s.d. of the mean illustrated above. In the usual circumstance, the estimate is derived as $S/\sqrt{N}$ (0.3687 in this example). The data in Table 1.3 were used only to illustrate the concept of a s.d. of a mean. In any event, the two estimates are not expected to agree exactly; after all $S_{\overline{X}}$ is also a random variable and only estimates the true value, $\sigma/\sqrt{N}$.

As the sample size increases, the s.d. of the mean becomes smaller and smaller. We can reduce the s.d. of the mean, $S_{\overline{X}}$, to a very small value by increasing N. Thus means of very large samples hardly vary at all. The concept of the s.d. of the mean is important, and the student will find it well worth the extra effort made to understand the meaning and implications of $S_{\overline{X}}$.

1.5.5 Variance of a Weighted Average[‡‡]

The general formula for the variance of a weighted average is

$$S_{\mathrm{w}}^2 = \frac{\left(\sum W_i^2 S_i^2\right)}{\left(\sum W_i\right)^2} \tag{1.10}$$

where S_i^2 is the variance of the ith observation. To compute the variance of the weighted mean, we would need to have an estimate of the variance of each observation.

If the weights of the observations are taken to be $1/S_i^2$ (the reciprocal of the variance, a common situation), then $S_{\mathrm{w}}^2 = 1/\sum(1/S_i^2)$. This formula can be applied to the calculation of the variance of the grand average of a group of i means where the variance of the individual observations is constant, equal to S^2. (We know that the variance of the grand average is S^2/N, where $N = \sum n_i$.) The variance of each mean, S_i^2, is S^2/n_i, where n_i is the number of observations

[††] The variance of a mean, $S_{\overline{X}}^2$, is S^2/N.

[‡‡] This is a more advanced topic.

in group i. In this example, the weights are considered to be the reciprocal of the variance, and $S_w^2 = 1/\sum(n_i/S^2) = S^2/\sum n_i$. Of course, we need to know S^2 (or have an estimate) in order to calculate (or estimate) the variance of the average. An estimate of the variance, S^2, in this example is $\sum n_i(Y_i - \overline{Y}_w)^2/(N-1)$, where the n_i acts as the weights and N is the number of observations.

The following calculation can be used to estimate the variance where a specified number of observations is available as a measure of the weight (as in a set of means). The variance of a set of weighted data can be estimated as follows:

$$\text{estimated variance} = \frac{\sum W_i \left(Y_i - \overline{Y}_w\right)^2}{\sum W_i - 1}, \tag{1.11}$$

where W_i is the weight associated with Y_i, and $\overline{Y}_w$ = weighted average of Y.

A shortcut formula is

$$\frac{\left[\sum \left(W_i Y_i^2\right) - \sum (W_i Y_i)^2 / \sum (W_i)\right]}{\sum W_i - 1}. \tag{1.12}$$

Example:

The diameters of 100 particles were measured with the results shown in Table 1.10.

From Eq. (1.12), the variance is estimated as $[89{,}375 - (2425)^2/100]/99 = 308.8$. s.d. $= \sqrt{308.8} = 17.6$. The s.d. of the mean is $17.6/\sqrt{100} = 1.76$. Note: The weighted average is $2425/100 = 24.25$.

In this example, it makes sense to divide the corrected sum of squares by $(N-1)$, because this sum of squares is computed using data from 100 particles. In some cases, the computation of the variance is not so obvious.

1.6 CODING

From both a practical and a theoretical point of view, it is useful to understand how the mean and s.d. of a group of numbers are affected by certain arithmetic manipulations, particularly adding a constant to, or subtracting a constant from each value; and multiplying or dividing each value by a constant.

Consider the following data to exemplify the results described below:

2, 3, 5, 10

Mean $= \overline{X} = 5$

Variance $= S^2 = 12.67$

Standard deviation $= S = 3.56$

Table 1.10 Data for Calculation of Variance of a Weighted Mean

Diameter (m)	Midpoint	Number of particles = weight	Weight × midpoint	Weight × midpoint^2
Y_i	W_i	$W_i Y_i$	$W_i Y_i^2$	
0–10	5	25	125	625
10–20	15	35	525	7875
30–40	35	15	525	18,375
40–60	50	*25*	*1250*	*62,500*
Sum		100	2425	89,375

1. Addition or subtraction of a constant will cause the mean to be increased or decreased by the constant, but will not change the variance or s.d. For example, adding + 3 to each value results in the following data:

   5, 6, 8, 13
   $\overline{X} = 8$
   $S = 3.56$

   Subtracting 2 from each value results in

   0, 1, 3, 8
   $\overline{X} = 3$
   $S = 3.56$

   This property may be used to advantage when hand calculating the mean and s.d. of very large or cumbersome numbers. Consider the following data:

   1251, 1257, 1253, 1255

   Subtracting 1250 from each value we obtain

   1, 7, 3, 5
   $\overline{X} = 4$
   $S = 2.58$

   To obtain the mean of the original values, add 1250 to the mean obtained above, 4. The s.d. is unchanged. For the original data

   $\overline{X} = 1250 + 4 = 1254$
   $S = 2.58$

   This manipulation is expressed in Eq. (1.13) where X_i represents one of n observations from a population with variance σ^2. C is a constant and $\overline{X}$ is the average of the X_i's.

   $$\text{Average}\,(X_i + C) = \sum \frac{X_i + C}{n} = \overline{X} + C$$

   $$\text{Variance}\,(X_i + C) = \sigma^2 \tag{1.13}$$

2. If the mean of a set of data is $\overline{X}$ and the s.d. is S, multiplying or dividing each value by a constant k results in a new mean of $k\overline{X}$ or $\overline{X}/k$, respectively, and a new s.d. of kS or S/k, respectively. Multiplying each of the original values above by 3 results in

   6, 9, 15, 30
   $\overline{X} = 15\,(3 \times 5)$
   $S = 10.68\,(3 \times 3.56)$

Dividing each value by 2 results in

$$\boxed{1, 1.5, 2.5, 5}$$
$$\overline{X} = 2.5\left(\frac{5}{2}\right)$$
$$S = 1.78\left(\frac{3.56}{2}\right)$$

In general,

$$\text{Average}\,(C \cdot X_i) = C\overline{X}$$
$$\text{Variance}\,(C \cdot X_i) = C^2\sigma^2 \tag{1.14}$$

These results can be used to show that a set of data with mean $\overline{X}$ and s.d. equal to S can be converted to data with a mean of 0 and a s.d. of 1 (as in the "standardization" of normal curves, discussed in sect. 3.4.1). If the mean is subtracted from each value, and this result is divided by S, the resultant data have a mean of 0 and a s.d. of 1. The transformation is

$$\frac{X - \overline{X}}{S}. \tag{1.15}$$

Standard scores are values that have been transformed according to Eq. (1.15) [11]. For the original data, the first value 2 is changed to $(2 - 5)/3.56$ equal to -0.84. The interested reader may verify that transforming the values in this way results in a mean of 0 and a s.d. of 1.

1.7 PRECISION, ACCURACY, AND BIAS

When dealing with variable measurements, the definitions of *precision* and *accuracy*, often obscure and not distinguished in ordinary usage, should be clearly defined from a statistical point of view.

1.7.1 Precision

In the vocabulary of statistics, precision refers to the extent of variability of a group of measurements observed under similar experimental conditions. A precise set of measurements is compact. Observations, relatively close in magnitude, are considered to be precise as reflected by a small s.d. (Note that means are more precisely measured than individual observations according to this definition.) An important, sometimes elusive concept is that a precise set of measurements may have the same mean as an imprecise set. In most experiments with which we will be concerned, the mean and s.d. of the data are independent (i.e., they are unrelated). Figure 1.3 shows the results of two assay methods, each performed in triplicate. Both methods have an average result of 100%, but method II is more precise.

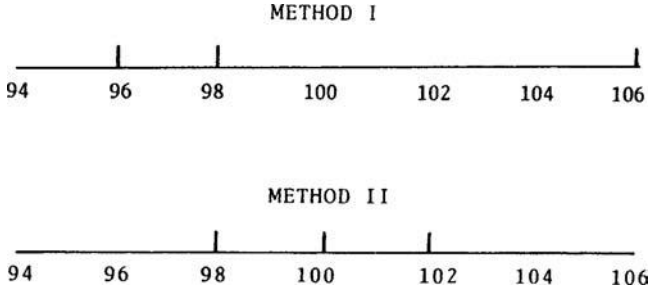


Figure 1.3 Representation of two analytical methods with the same accuracy but different precisions.

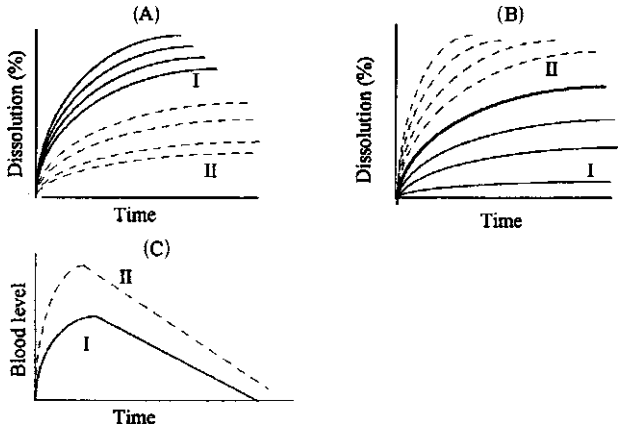


Figure 1.4 In vitro dissolution results for two formulations using two different methods and in vivo blood level versus time results. Methods A and B, in vitro; C, in vivo.

1.7.2 Accuracy

Accuracy refers to the closeness of an individual observation or mean to the true value. The "true" value is the result that would be observed in the absence of error (e.g., the true mean tablet potency or the true drug content of a preparation being assayed). In the example of the assay results shown in Figure 1.3, both methods are apparently equally accurate (or inaccurate).

Figure 1.4 shows the results of two dissolution methods for two formulations of the same drug, each formulation replicated four times by each method. The objective of the in vitro dissolution test is to simulate the in vivo oral absorption of the drug from the two dosage-form modifications. The first dissolution method, A, is very precise but does not give an accurate prediction of the in vivo results. According to the dissolution data for method A, we would expect that formulation I would be more rapidly and extensively absorbed in vivo. The actual in vivo results depicted in Figure 1.4 show the contrary result. The less precise method, method B in this example, is a more *accurate* predictor of the true in vivo results. This example is meant to show that a precise measurement need not be accurate, nor an accurate measurement precise.

Of course, the best circumstance is to have data that are both precise and accurate. If possible, we should make efforts to improve both the accuracy and precision of experimental observations. For example, in drug analysis, advanced electronic instrumentation can greatly increase the accuracy and precision of assay results.

1.7.3 Bias

Accuracy can also be associated with the term *bias*. The notion of bias has been discussed in section 1.4 in relation to the concept of unbiased estimates (e.g., the mean and variance). The meaning of bias in statistics is similar to the everyday definition in terms of "fairness." An accurate measurement, no matter what the precision, can be thought of as unbiased, because an accurate measurement is a "fair" estimate of the true result. A biased estimate is systematically either higher or lower than the true value. A biased estimate can be thought of as giving an "unfair" notion of the true value. For example, when estimating the average result of experimental data, the mean, $\overline{X}$, represents an estimate of the true population parameter, μ, and in this sense is considered accurate and unbiased. An average blood pressure reduction of 10 mm Hg due to an antihypertensive agent, derived from data from a clinical study of 200 patients, can be thought of as an unbiased estimate of the true blood pressure reduction due to the drug, provided that the patients are appropriately selected at "random." The true reduction in this case is the average reduction that would be observed if the antihypertensive effect of the drug were known for all members of the population (e.g., all hypertensive patients). The outcome of a single experiment, such as the 10 mm Hg reduction observed in the 200 patients above, will in all probability not be identical to the true mean reduction. But the mean reduction as observed

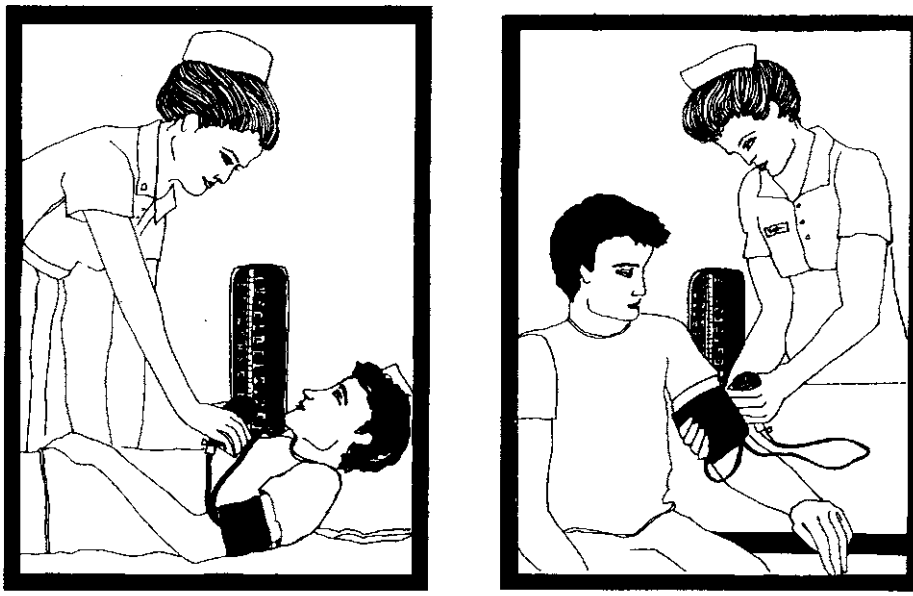


Figure 1.5 Bias in determining the effect of an antihypertensive drug.

in the 200 patients is an accurate and unbiased assessment of the population average. A biased estimate is one which, on the average, does not equal the population parameter. In the example cited above for hypertensives, a biased estimate would result if for all patients one nurse took all the measurements before therapy and another nurse took all measurements during therapy, and each nurse had a different criterion or method for determining blood pressure. See Figure 1.5 for a clarification as to why this procedure leads to a biased estimate of the drug's effectiveness in reducing blood pressure. If the supine position results in higher blood pressure than the sitting position, the results of the study will tend to show a bias in the direction of too large a blood pressure reduction.

The statistical estimates that we usually use, such as the mean and variance, are unbiased estimates. Bias often results from (a) the improper use of experimental design; (b) improper choice of samples; (c) unconscious bias, due to lack of blinding, for example; or (d) improper observation and recording of data, such as that illustrated in Figure 1.5.

1.8 THE QUESTION OF SIGNIFICANT FIGURES

The question of *significant figures* is an important consideration in statistical calculations and presentations. In general, the ordinary rules for retaining significant figures are not applicable to statistical computations. Contrary to the usual rules for retaining significant figures, one should retain as many figures as possible when performing statistical calculations, not rounding off until all computations are complete.

The reason for not rounding off during statistical computations is that untenable answers may result when using computational procedures that involve taking differences between values very close in magnitude if values are rounded off prior to taking differences. This may occur when calculating "sums of squares" (the sum of squared differences from the mean) using the shortcut formula, Eq. (1.4), for the calculation of the variance or s.d. The shortcut formula for $\sum(X - \overline{X})^2$ is $\sum X^2 - (\sum X)^2/N$ that cannot be negative, and will be equal to zero only if all the data have the same value. If the two terms, $\sum X^2$ and $(\sum X)^2/N$, are very similar in magnitude, rounding off before taking their difference may result in a zero or negative difference. This problem is illustrated by calculating the s.d. of the three numbers 1.19, 1.20, and 1.21. If the squares of these numbers are first rounded off to two decimal places, the following calculation

of the s.d. results:

$$S = \sqrt{\frac{\sum (X^2 - \sum X)^2 \Big/ N}{N-1}} = \sqrt{\frac{1.42 + 1.44 + 1.46 - 3.6^2/3}{2}}$$

$$= \sqrt{\frac{4.32 - 4.32}{2}} = 0.$$

The correct s.d. calculated without rounding off is 0.01.

Computers and calculators carry many digits when performing calculations and do not round off further unless instructed to do so. These instruments retain as many digits as their capacity permits through all arithmetic computations. The possibility of rounding off, even considering the large capacity of modern computers, can cause unexpected problems in sophisticated statistical calculations, and must be taken into account in preparing statistical software programs. These problems can usually be overcome by using special programming techniques.

At the completion of the calculations, as many figures as are appropriate to the situation can be presented. Common sense and the usual rules for reporting significant figures should be applied (see Ref. [9] for a detailed discussion of significant figures). Sokal and Rohlf [9] recommend that, if possible, observations should be measured with enough significant figures so that the range of data is between 30 and 300 possible values. This flexible rule results in a relative error of less than 3%. For example, when measuring diastolic blood pressure, the range of values for a particular group of patients might be limited to 60 to 130 mm Hg. Therefore, measurements to the nearest mm Hg would result in approximately 70 possible values, and would be measured with sufficient accuracy according to this rule. If the investigator can make the measurement only in intervals of 2 mm Hg (e.g., 70 and 72 mm Hg can be measured, but not 71 mm Hg), we would have 35 possible data points, which is still within the 30 to 300 suggested by this rule of thumb. Of course, rules should not be taken as "written in stone." All rules should be applied with judgment.

Common sense should be applied when reporting average results. For example, reporting an average blood pressure reduction of 7.42857 for 14 patients treated with an antihypertensive agent would not be appropriate. As noted above, most physicians would say that blood pressure is rarely measured to within 2 mm Hg. Why should one bother to report any decimals at all for the average result? When reporting average results, it is generally good practice to report the average with a precision that is "reasonable" according to the nature of the data. An average of 7.4 mm Hg would probably suffice for this example. If the average were reported as 7 mm Hg, for example, it would appear that too much information is suppressed.

KEY TERMS

Accuracy
Attributes
Average ($\overline{X}$)
Bias
Coding
Coefficient of variation (CV)
Continuous variables
Correction term (CT)
Cumulative distribution
Degrees of freedom (d.f.)
Discrete variables
Frequency distribution
Geometric mean
Harmonic mean
Mean ($\overline{X}$)
Median
Population
Precision
Random variable
Range
Ranking
Rating scale
Sample
Significant figures
Standard deviation (s.d., S)
Standard error of the mean ($S_{\overline{X}}$)
Standard score
Treatment
Unbiased sample
Universe
Variability
Variable
Weighted average

EXERCISES

1. List three experiments whose outcomes will result in each of the following kinds of variables:
   (a) Continuous variables
   (b) Discrete variables
   (c) Ordered variables
   (d) Categorical (attribute) variables
2. What difference in experimental conclusions, if any, would result if the pain scale discussed in section 1.1 were revised as follows no pain = 6, slight pain = 4, moderate pain = 2, and severe pain = 0? (Hint: see sect. 1.6.)
3. (a) Construct a frequency distribution containing 10 class intervals from the data in Table 1.1.
   (b) Construct a cumulative frequency plot based on the frequency distribution from part (a).
4. What is the average result based on the frequency distribution in part (a) of problem 3? Use a weighted-average procedure.
5. From Figure 1.1, what proportion of tablets have potencies between 95 and 105 mg? What proportion of tablets have a potency greater than 105 mg?
6. Calculate the average and standard deviation of (a) the first 20 values in Table 1.1, and (b) the last 20 values in Table 1.1. If these data came from two different clinical investigators, would you think that the differences in these two sets of data can be attributed to differences in clinical sites? Which set, the first or last, is more precise? Explain your answer.
7. What are the median and range of the first 20 values in Table 1.1?
8. (a) If the first value in Table 1.1 were +100 instead of +17, what would be the values of the median and range for the first 20 values?
   (b) Using the first value as 100, calculate the mean, standard deviation, and variance. Compare the results for these first 20 values to the answers obtained in Problem 6.

§§**9. Given the following sample characteristics, describe the population from which the sample may have been derived. The mean is 100, the standard deviation is 50, the median is 75, and the range is 125.

**10. If the population average for the cholesterol reductions shown in Table 1.1 were somehow known to be 0 (the drug does not affect cholesterol levels on the average), would you believe that this sample of 156 patients gives an unbiased estimate of the true average? Describe possible situations in which these data might yield (a) biased results; (b) unbiased results.

**11. Calculate the average standard deviation using the sampling experiment shown in section 1.5.2 for samples of size 2 taken from a population with values of 0, 1, and 2 (with replacement). Compare this result with the population standard deviation. Is the sample standard deviation an unbiased estimate of the population standard deviation?

12. Describe another situation that would result in a biased estimate of blood pressure reduction as discussed in section 1.7.3 (Fig. 1.5).
13. Verify that the standard deviation of the values 1.19, 1.20, and 1.21 is 0.01 (see sect. 1.8). What is the standard deviation of the numbers 2.19, 2.20, and 2.21? Explain the result of the two calculations above.

§§ The double asterisk indicates optional, more difficult problems.

14. For the following blood pressure measurements: 100, 98, 101, 94, 104, 102, 108, 108, calculate (a) the mean, (b) the standard deviation, (c) the variance, (d) the coefficient of variation, (e) the range, and (f) the median.

**15. Calculate the standard deviation of the grouped data in Table 1.2. (Hint : $S^2 = \left[\sum N_i X_i^2 - (\sum N_i X_i)^2/(\sum N_i)\right]/(\sum N_i - 1)$; see Ref. [3]. N_i = frequency per group with midpoint X_i)

16. Compute the arithmetic mean, geometric mean, and harmonic mean of the following set of data. 3, 5, 7, 11, 14, 57

    If these data were observations on the time needed to cure a disease, which mean would you think to be most appropriate?

17. If the weights are 2, 1, 1, 3, 1, and 2 for the numbers 3, 5, 7, 11, 14, and 57 (Exercise 16), compute the weighted average and variance.

REFERENCES

1. Berkow R. The Merck Manual, 14th ed. Rahway, NJ: Merck Sharp & Dohme Research Laboratories, 1982.
2. Tukey J. Exploratory Data Analysis. Reading, MA: Addison-Wesley, 1977.
3. Yule GU, Kendall MG. An Introduction to the Theory of Statistics, 14th ed. London: Charles Griffin, 1965.
4. Sokal RR, Rohlf FJ. Biometry. San Francisco, CA: W.H. Freeman, 1969.
5. Colton T. Statistics in Medicine. Boston, MA: Little, Brown, 1974.
6. Dixon WJ, Massey FJ Jr. Introduction to Statistical Analysis, 3rd ed. New York: McGraw-Hill, 1969.
7. United States Pharmacopeia, 9th Supplement. Rockville, MD: USP Convention, Inc., 1990:3584–3591.
8. Graham SJ, Lawrence RC, Ormsby ED, et al. Particle Size Distribution of Single and Multiple Sprays of Salbutamol Metered-Dose Inhalers (MDIs). Pharm Res 1995; 12:1380.
9. Lachman L, Lieberman HA, Kanig JL. The Theory and Practice of Industrial Pharmacy, 3rd ed. Philadelphia, PA: Lea & Febiger, 1986.
10. Snedecor GW, Cochran WG. Statistical Methods, 8th ed. Ames, IA: Iowa State University Press, 1989.
11. Rothman ED, Ericson WA. Statistics, Methods and Applications. Dubuque, IA: Kendall Hunt, 1983.

2 DATA GRAPHICS

"The preliminary examination of most data is facilitated by the use of diagrams. Diagrams prove nothing, but bring outstanding features readily to the eye; they are therefore no substitute for such critical tests as may be applied to the data, but are valuable in suggesting such tests, and in explaining the conclusions founded upon them." This quote is from Ronald A. Fisher, the father of modern statistical methodology [1]. Tabulation of raw data can be thought of as the initial and least refined way of presenting experimental results. Summary tables, such as frequency distribution tables, are much easier to digest and can be considered a second stage of refinement of data presentation. Summary statistics such as the mean, median, variance, standard deviation, and the range are concise descriptions of the properties of data, but much information is lost in this processing of experimental results. Graphical methods of displaying data are to be encouraged and are important adjuncts to data analysis and presentation. Graphical presentations clarify and also reinforce conclusions based on formal statistical analyses. Finally, the researcher has the opportunity to design aesthetic graphical presentations that command attention. The popular cliché "A picture is worth a thousand words" is especially apropos to statistical presentations. We will discuss some key concepts of the various ways in which data are depicted graphically.

2.1 INTRODUCTION

The diagrams and plots that we will be concerned with in our discussion of statistical methods can be placed broadly into two categories:

1. Descriptive plots are those whose purpose is to transmit information. These include diagrams describing data distributions such as histograms and cumulative distribution plots (see sect. 1.2.3). Bar charts and pie charts are examples of popular modes of communicating survey data or product comparisons.
2. Plots that describe *relationships* between variables usually show an underlying, but unknown analytic relationship between the variables that we wish to describe and understand. These relationships can range from relatively simple to very complex, and may involve only two variables or many variables. One of the simplest relationships, but probably the one with greatest practical application, is the straight-line relationship between two variables, as shown in the Beer's law plot in Figure 2.1. Chapter 7 is devoted to the analysis of data involving variables that have a linear relationship.

When analyzing and depicting data that involve relationships, we are often presented with data in pairs (X, Y pairs). In Figure 2.1, the optical density Y and the concentration X are the data pairs. When considering the relationship of two variables, X and Y, one variable can often be considered the response variable, which is dependent on the selection of the second or causal variable. The response variable Y (optical density in our example) is known as the *dependent* variable. The value of Y depends on the value of the *independent* variable, X (drug concentration). Thus, in the example in Figure 2.1, we think of the value of optical density as being dependent on the concentration of drug.

2.2 THE HISTOGRAM

The *histogram*, sometimes known as a *bar graph*, is one of the most popular ways of presenting and summarizing data. All of us have seen bar graphs, not only in scientific reports but also in advertisements and other kinds of presentations illustrating the distribution of scientific data.

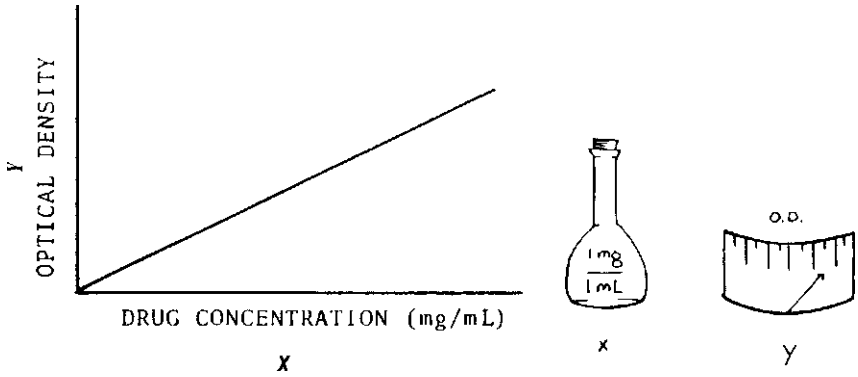


Figure 2.1 Beer's law plot illustrating a linear relationship between two variables.

The histogram can be considered as a visual presentation of a frequency table. The frequency, or proportion, of observations in each class interval is plotted as a bar, or rectangle, where the area of the bar is proportional to the frequency (or proportion) of observations in a given interval. An example of a histogram is shown in figure 2.2, where the data from the frequency table in Table 1.2 have been used as the data source. As is the case with frequency tables, class intervals for histograms should be of equal width. When the intervals are of equal width, the height of the bar is proportional to the frequency of observations in the interval. If the intervals are not of equal width, the histogram is not easily or obviously interpreted, as shown in Figure 2.2(B).

The choice of intervals for a histogram depends on the nature of the data, the distribution of the data, and the purpose of the presentation. In general, rules of thumb similar to that used

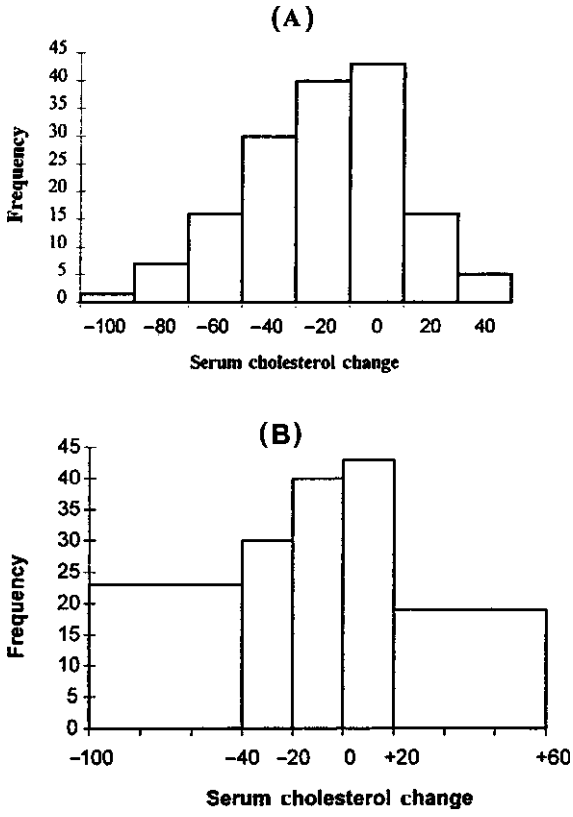


Figure 2.2 Histogram of data derived from Table 1.2.

for frequency distribution tables (sect. 1.2) can be used. Eight to twenty equally spaced intervals usually are sufficient to give a good picture of the data distribution.

2.3 CONSTRUCTION AND LABELING OF GRAPHS

Proper *construction and labeling* of graphs are crucial elements in graphical data representation. The design and actual construction of graphs are not in themselves difficult. The preparation of a *good* graph, however, requires careful thought and competent technical skills. One needs not only a knowledge of statistical principles, but also, in particular, computer and drafting competency. There are no firm rules for preparing good graphical presentations. Mostly, we rely on experience and a few guidelines. Both books and research papers have addressed the need for a more scientific guide to optimal graphics that, after all, is measured by how well the graph communicates the intended messages(s) to the individuals who are intended to read and interpret the graphs. Still, no rules will cover all situations. One must be clear that no matter how well a graph or chart is conceived, if the draftsmanship and execution is poor, the graph will fail to achieve its purpose.

A "good" graph or chart should be as *simple* as possible, yet clearly transmit its intended message. Superfluous notation, confusing lines or curves, and inappropriate draftsmanship (lettering, etc.) that can distract the reader are signs of a poorly constructed graph. The books *Statistical Graphics*, by Schmid [2], and *The Visual Display of Quantitative Information* by Tufte [3] are recommended for those who wish to study examples of good and poor renderings of graphic presentations. For example, Schmid notes that visual contrast should be intentionally used to emphasize important characteristics of the graph. Here, we will present a few examples to illustrate the recommendations for good graphic presentation as well as examples of graphs that are not prepared well or fail to illustrate the facts fairly.

Figure 2.3 shows the results of a clinical study that was designed to compare an active drug to a placebo for the treatment of hypertension. This graph was constructed from the X, Y pairs, *time* and *blood pressure*, respectively. Each point on the graph (+ , ■) is the average blood pressure for either drug or placebo at some point in time subsequent to the initiation of the study.

Proper construction and labeling of the typical rectilinear graph should include the following considerations:

1. A *title* should be given. The title should be brief and to the point, enabling the reader to understand the purpose of the graph without having to resort to reading the text. The title can be placed below or above the graph as in Figure 2.3.
2. The *axes* should be *clearly* delineated and *labeled.* In general, the zero (0) points of both axes should be clearly indicated. The ordinate (the Y axis) is usually labeled with the description parallel to the Y axis. Both the ordinate and abscissa (X axis) should be each appropriately

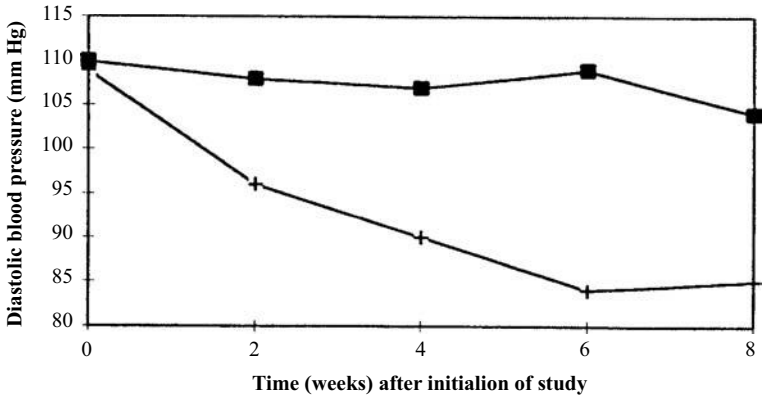


Figure 2.3 Blood pressure as a function of time in a clinical study comparing drug and placebo with a regimen of one tablet per day. ■, placebo (average of 45 patients); +, drug (average of 50 patients).

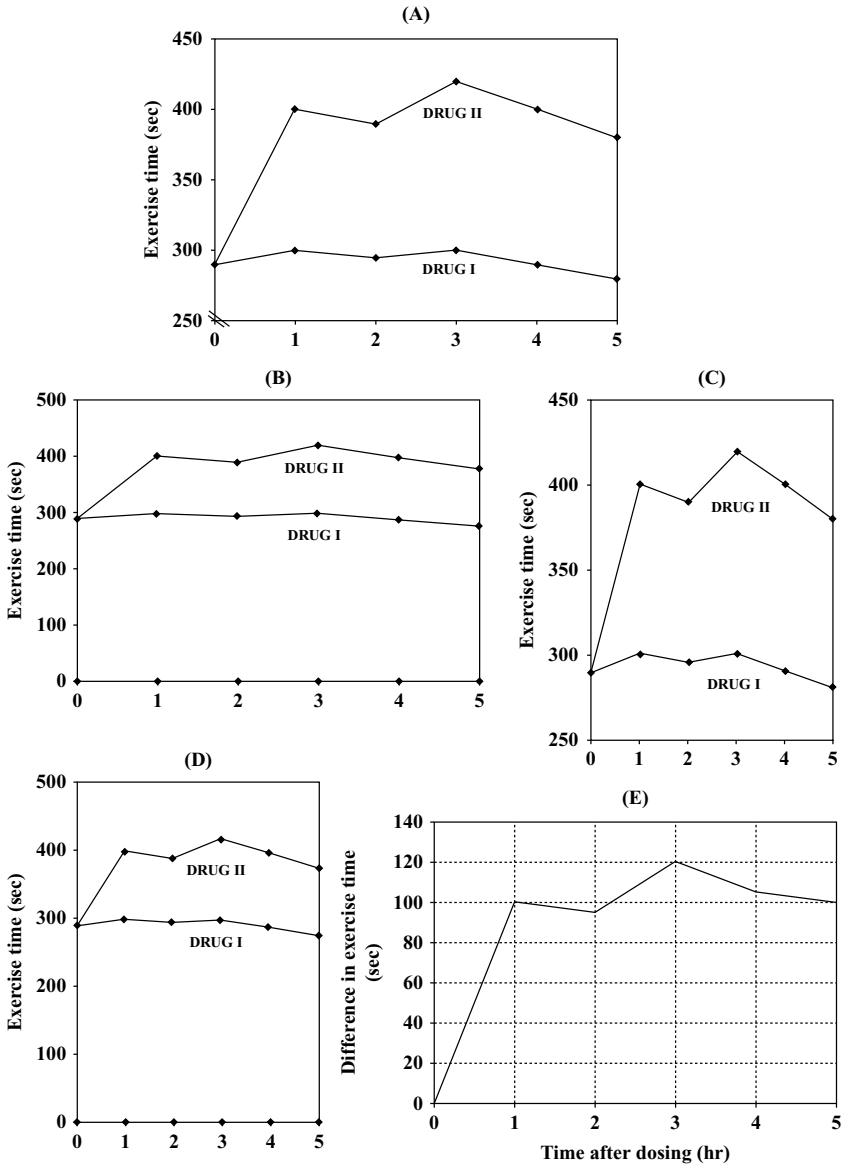


Figure 2.4 Various graphs of the same data presented in different ways. Exercise time at various time intervals after administration of single doses of two nitrate products. ♦ = Drug I, ■ = Drug II.

labeled and subdivided in units of equal width (of course, the *X* and *Y* axes almost always have different subdivisions). In the example in Figure 2.3, note the units of mm Hg and weeks for the ordinate and abscissa, respectively. Grid lines may be added [Fig. 2.4(E)] but, if used, should be kept to a minimum, not be prominent and should not interfere with the interpretation of the figure.

3. The numerical *values* assigned to the axes should be *appropriately spaced* so as to nicely cover the extent of the graph. This can easily be accomplished by trial and error and a little manipulation. The scales and proportions should be constructed to present a fair picture of the results and should not be exaggerated so to prejudice the interpretation. Sometimes, it may be necessary to skip or omit some of the data to achieve this objective. In these cases, the use of a "broken line" is recommended to clearly indicate the range of data not included in the graph (Fig. 2.4).

4. If appropriate, a *key* explaining the symbols used in the graph should be used. For example, at the bottom of Figure 2.3, the key defines ■ as the symbol for placebo and + for drug. In many cases, labeling the curves directly on the graph (Fig. 2.4) results in more clarity.
5. In situations where the graph is derived from laboratory data, inclusion of the *source* of the data (name, laboratory notebook number, and page number, for example) is recommended.

Usually graphs should stand on their own, independent of the main body of the text.

Examples of various ways of plotting data, derived from a study of exercise time at various time intervals after administration of a single dose of two long-acting nitrate products to anginal patients, are shown in Figures 2.4(A) to 2.4(E). All of these plots are accurate representations of the experimental results, but each gives the reader a different impression. It would be wrong to expand or contract the axes of the graph, or otherwise distort the graph, in order to convey an incorrect impression to the reader. Most scientists are well aware of how data can be manipulated to give different impressions. If obvious deception is intended, the experimental results will not be taken seriously.

When examining the various plots in Figure 2.4, one could not say which plot best represents the meaning of the experimental results without knowledge of the experimental details, in particular the objective of the experiment, the implications of the experimental outcome, and the message that is *meant* to be conveyed. For example, if an improvement of exercise time of 120 seconds for one drug compared to the other is considered to be significant from a medical point of view, the graphs labeled A, C, and E in Figure 2.4 would all seem appropriate in conveying this message. The graphs labeled B and D show this difference less clearly. On the other hand, if 120 seconds is considered to be of little medical significance, B and D might be a better representation of the data.

Note that in plot A of Figure 2.4, the ordinate (exercise time) is broken, indicating that some values have been skipped. This is not meant to be deceptive, but is intentionally done to better show the differences between the two drugs. As long as the zero point and the break in the axis are clearly indicated, and the message is not distorted, such a procedure is entirely acceptable.

Figures 2.4(B) and 2.5 are exaggerated examples of plots that may be considered not to reflect accurately the significance of the experimental results. In Figure 2.4(B), the clinically significant difference of approximately 120 seconds is made to look very small, tending to diminish drug differences in the viewer's mind. Also, fluctuations in the hourly results appear to be less than the data truly suggest. In Figure 2.5, a difference of 5 seconds in exercise time between the two drugs appears very large. Care should be taken when constructing (as well as reading) graphs so that experimental conclusions come through clear and true.

6. If more than one curve appears on the same graph, a convenient way to differentiate the curves is to use different symbols for the experimental points (e.g., ○, ×, Δ, □, +) and, if necessary, connecting the points in different ways (e.g., —.—.—., ... , –.–.–.–). A key or label is used, which is helpful in distinguishing the various curves, as shown in Figures 2.3 to 2.6. Other ways of differentiating curves include different kinds of crosshatching and use of different colors.

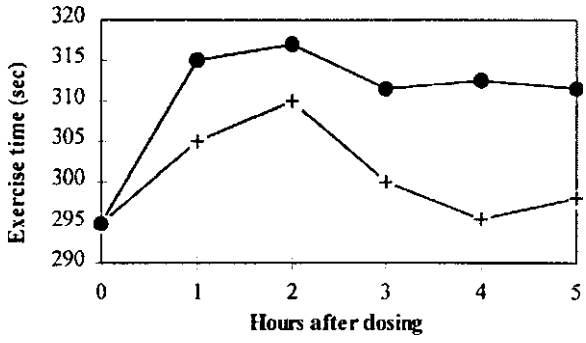


Figure 2.5 Exercise time at various time intervals after administration of two nitrate products. •, product I; +, product II.

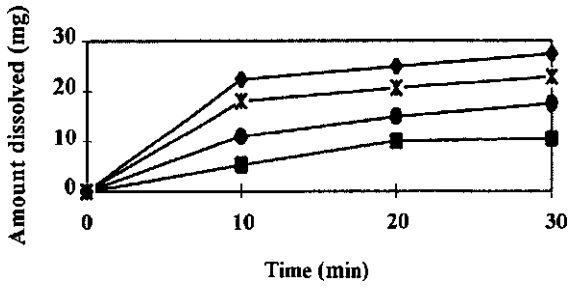


Figure 2.6 Plot of dissolution of four successive batches of a commercial tablet product. ♦ = batch I, • = batch II, × = batch 3, ♦ = batch 4.

7. One should take care not to place too many curves on the same graph, as this can result in confusion. There are no specific rules in this regard. The decision depends on the nature of the data, and how the data look when they are plotted. The curves graphed in Figure 2.7 are cluttered and confusing. The curves should be presented differently or separated into two or more graphs. Figure 2.8 is a clearer depiction of the dissolution results of the five formulations shown in Figure 2.7.
8. The *standard deviation* may be indicated on graphs as shown in Figure 2.9. However, when the standard deviation is indicated on a graph (or in a table, for that matter), it should be made clear whether the variation described in the graph is an indication of the standard deviation (S) or the standard deviation of the mean ($S_{\bar{x}}$). The standard deviation of the mean, if appropriate, is often preferable to the standard deviation not only because the values on the graph are mean values, but also because $S_{\bar{x}}$ is smaller than the s.d., and therefore less cluttering. *Overlapping* standard deviations, as shown in Figure 2.10, should be avoided, as this representation of the experimental results is usually more confusing than clarifying.
9. The manner in which the points on a graph should be connected is not always obvious. Should the individual points be connected by straight lines, or should a smooth curve that approximates the points be drawn through the data? (See Fig. 2.11.) If the graphs represent functional relationships, the data should probably be connected by a smooth curve. For example, the blood level versus time data shown in Figure 2.11 are described most accurately by a smooth curve. Although, theoretically, the points should not be connected by straight lines as shown in Figure 2.11(A), such graphs are often depicted this way. Connecting the individual points with straight lines may be considered acceptable if one recognizes that this representation is meant to clarify the graphical presentation, or is done for some other appropriate reason. In the blood-level example, the area under the curve is proportional to the amount of drug absorbed. The area is often computed by the trapezoidal rule [4], and depiction of the data as shown in Figure 2.11(A) makes it easier to visualize and perform such calculations.

Figure 2.12 shows another example in which connecting points by straight lines is convenient but may *not* be a good representation of the experimental outcome. The straight line connecting the blood pressure at zero time (before drug administration) to the blood pressure after two weeks of drug administration suggests a gradual decrease (a linear decrease) in blood

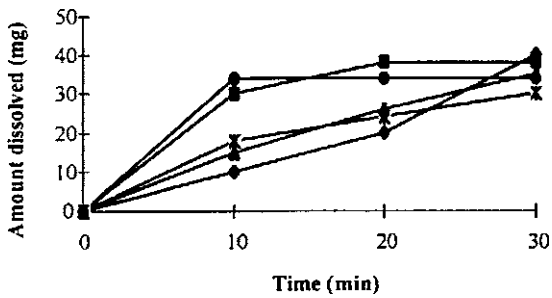


Figure 2.7 Plot of dissolution time of five different commercial formulations of the same drug. • = product A, ■ = product B, × = product C, ▲ = product D, ♦ = product E.

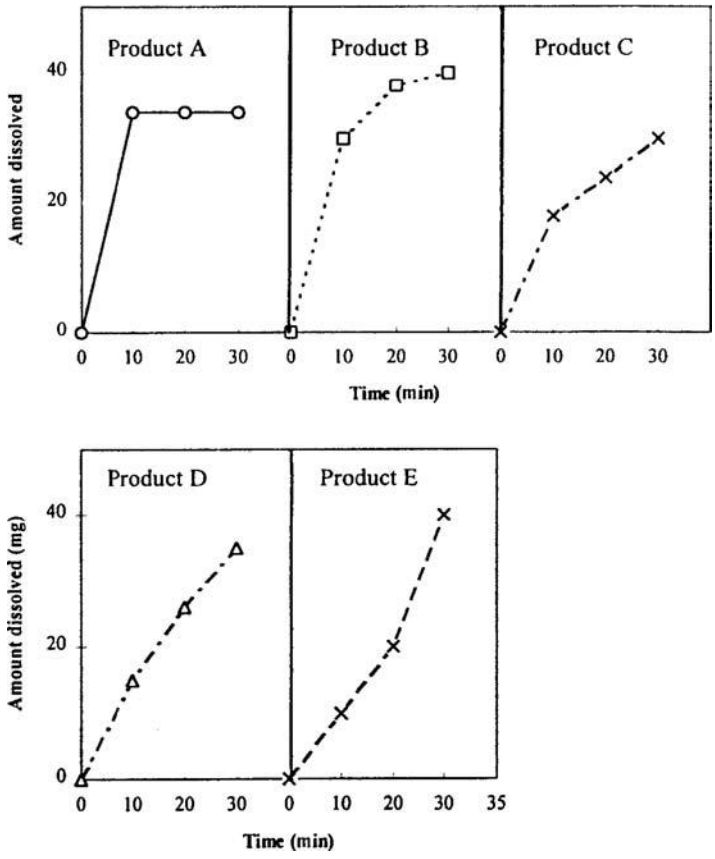


Figure 2.8 Individual plots of dissolution of the five formulations shown in Fig. 2.7.

pressure over the two-week period. In fact, no measurements were made during the initial two-week interval. The 10-mm Hg decrease observed after two weeks of therapy may have occurred before the two-week reading (e.g., in one week, as indicated by the dashed line in Fig. 2.12). One should be careful to ensure that graphs constructed in such a manner are not misinterpreted.

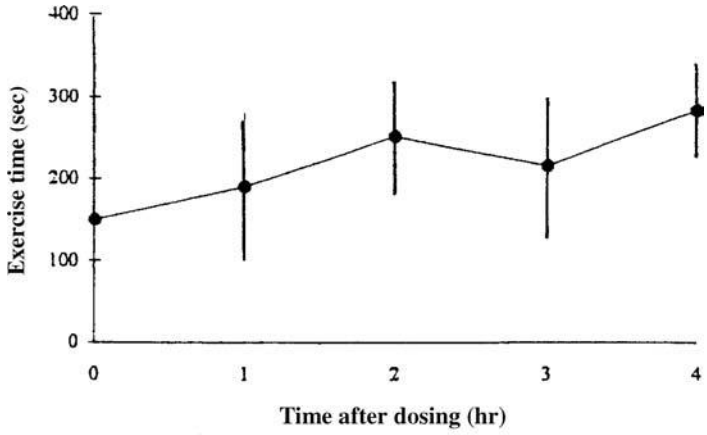


Figure 2.9 Plot of exercise time as a function of time for an antianginal drug showing mean values and standard error of the mean.

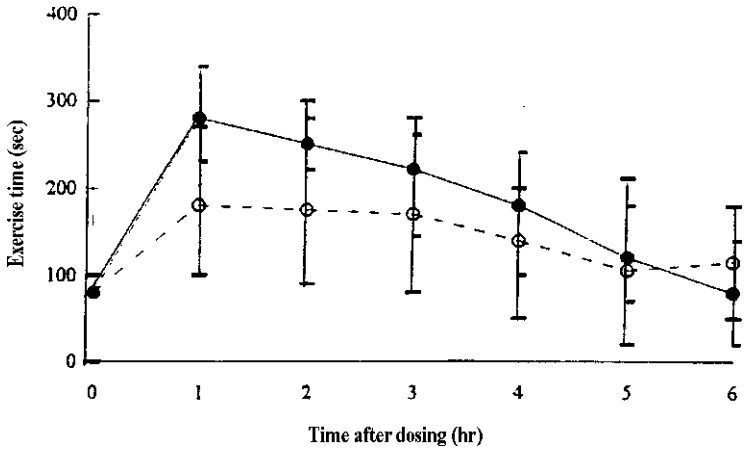


Figure 2.10 Graph comparing two antianginal drugs that is confusing and cluttered because of the overlapping standard deviations. •, drug A; o, drug B.

2.4 SCATTER PLOTS (CORRELATION DIAGRAMS)

Although the applications *of correlation* will be presented in some detail in chapter 7, we will introduce the notion of *scatter plots* (also called correlation diagrams or scatter diagrams) at this time. This type of plot or diagram is commonly used when presenting results of experiments. A typical scatter plot is illustrated in Figure 2.13. Data are collected in pairs (X and Y) with the objective of demonstrating a trend or relationship (or lack of relationship) between the X and Y variables. Usually, we are interested in showing a linear relationship between the variables (i.e., a straight line). For example, one may be interested in demonstrating a relationship (or correlation) between time to 80% *dissolution* of various tablet formulations of a particular drug

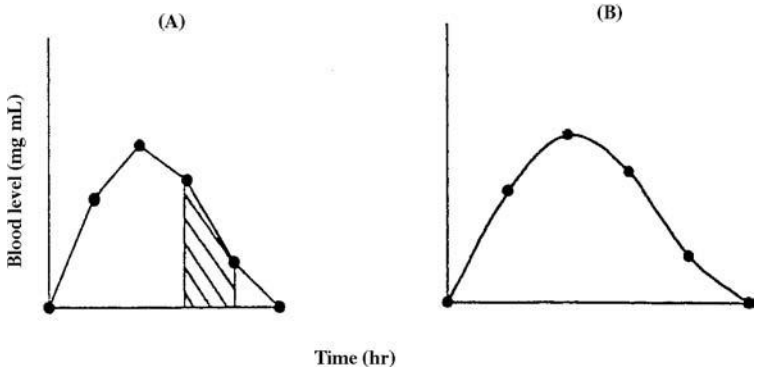


Figure 2.11 Plot of blood level versus time data illustrating two ways of drawing the curves.

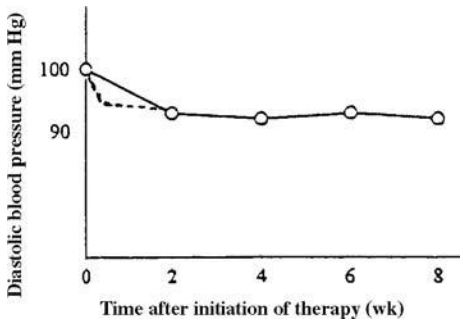


Figure 2.12 Graph of blood pressure reduction with time of antihypertensive drug illustrating possible misinterpretation that may occur when points are connected by straight lines.

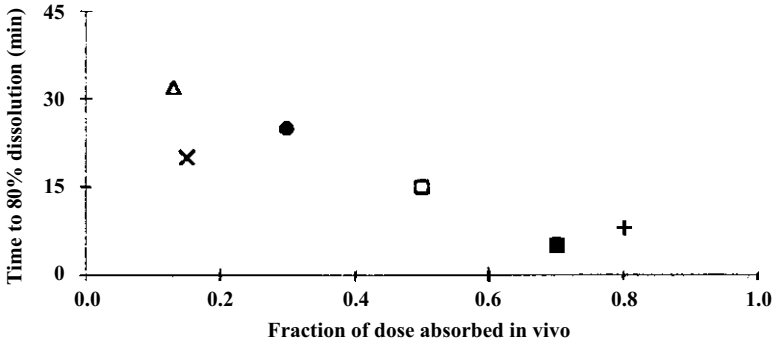


Figure 2.13 Scatter plot showing the correlation of dissolution time and in vivo absorption of six tablet formulations. △, formulation A; ×, formulation B; •, formulation C; □, formulation D; ■, formulation E; +, formulation F.

and the *fraction of the dose absorbed* when human subjects take the various tablets. The data plotted in Figure 2.13 show pictorially that as dissolution increases (i.e., the time to 80% dissolution decreases) in vivo absorption increases. Scatter plots involve data pairs, X and Y, both of which are variable. In this example, *dissolution time* and *fraction absorbed* are both random variables.

2.5 SEMILOGARITHMIC PLOTS

Several important kinds of experiments in the pharmaceutical sciences result in data such that the *logarithm* of the response (Y) is linearly related to an independent variable, X. The semilogarithmic plot is useful when the response (Y) is best depicted as proportional changes relative to changes in X, or when the spread of Y is very large and cannot be easily depicted on a rectilinear scale. Semilog graph paper has the usual equal interval scale on the X axis and the logarithmic scale on the Y axis. In the logarithmic scale, equal intervals represent ratios. For example, the distance between 1 and 10 will exactly equal the distance between 10 and 100 on a logarithmic scale. In particular, first-order kinetic processes, often apparent in drug degradation and pharmacokinetic systems, show a linear relationship when log C is plotted versus time. First-order processes can be expressed by the following equation:

$$\log\ C\ =\ \log\ C_0 - \frac{kt}{2.3} \tag{2.1}$$

where C is the concentration at time t, C_0 the concentration at time 0, k the first-order rate constant, t the time, and log represents logarithm to the base 10.

Table 2.1 shows blood-level data obtained after an intravenous injection of a drug described by a one-compartment model [3].

Figure 2.14 shows two ways of plotting the data in Table 2.1 to demonstrate the linearity of the log C versus t relationship.

1. Figure 2.14(A) shows a plot of log C versus time. The resulting straight line is a consequence of the relationship of log concentration and time as shown in Eq. 2.1. This is an equation of a straight line with the Y intercept equal to log C_0 and a slope equal to $-k/2.3$. Straight-line relationships are discussed in more detail in chapter 8.

Table 2.1 Blood Levels After Intravenous Injection of Drug

Time after injection, *t* (hr)	Blood level, C (μg/mL)	Log blood level
0	20	1.301
1	10	1.000
2	5	0.699
3	2.5	0.398
4	1.25	0.097

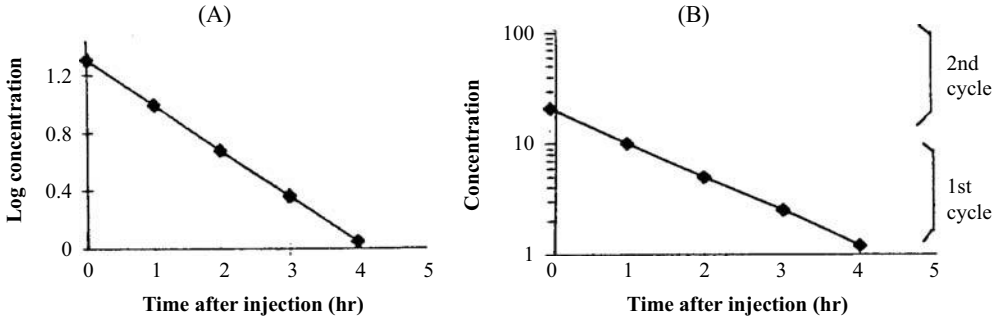


Figure 2.14 Linearizing plots of data from Table 2.1. (Plot A) log *C* versus time; (plot B) semilog plot.

2. Figure 2.14(B) shows a more convenient way of plotting the data of Table 2.1, making use of *semilog graph paper.* This paper has a logarithmic scale on the *Y* axis and the usual arithmetic, linear scale on the *X* axis. The logarithmic scale is constructed so that the spacing corresponds to the logarithms of the numbers on the *Y* axis. For example, the distance between 1 and 2 is the same as that between 2 and 4. (Log 2−log 1) is equal to (log 4−log 2). The semilog graph paper depicted in Figure 2.14(B) is two-cycle paper. The *Y* (log) axis has been repeated two times. The decimal point for the numbers on the *Y* axis is accommodated to the data. In our example, the data range from 1.25 to 20 and the *Y* axis is adjusted accordingly, as shown in Figure 2.14(B). The data may be plotted directly on this paper without the need to look up the logarithms of the concentration values.

2.6 OTHER DESCRIPTIVE FIGURES

Most of the discussion in this chapter has been concerned with plots that show relationships between variables such as blood pressure changes following two or more treatments, or drug decomposition as a function of time. Often occasions arise in which graphical presentations are better made using other more pictorial techniques. These approaches include the popular bar and pie charts. Schmid [2] differentiates bar charts into two categories: (a) *column charts* in which there is a vertical orientation and (b) *bar charts* in which the bars are horizontal. In general, the bar charts are more appropriate for comparison of categorical variables, whereas the column chart is used for data showing relationships such as comparisons of drug effect over time.

Bar charts are very simple but effective visual displays. They are usually used to compare some experimental outcome or other relevant data where the length of the bar represents the magnitude. There are many variations of the simple bar chart [2]; an example is shown in Figure 2.15. In Figure 2.15(A), patients are categorized as having a good, fair, or poor response. Forty percent of the patients had a good response, 35% had a fair response, and 25% had a poor response.

Figure 2.15(B) shows bars in pairs to emphasize the comparative nature of two treatments. It is clear from this diagram that Treatment *X* is superior to Treatment *Y*. Figure 2.15(C) is another way of displaying the results shown in Figure 2.15(B). Which chart do you think better sends the message of the results of this comparative study, Figure 2.15(B) or 2.15(C)? One should be aware that the results correspond only to the length of the bar. If the order in which the bars are presented is not obvious, displaying bars in order of magnitude is recommended. In the example in Figure 2.15, the order is based on the nature of the results, "Good," "Fair," and "Poor." Everything else in the design of these charts is superfluous and the otherwise principal objective is to prepare an aesthetic presentation that emphasizes but does not exaggerate the results. For example, the use of graphic techniques such as shading, crosshatching, and color, tastefully executed, can enhance the presentation.

Column charts are prepared in a similar way to bar charts. As noted above, whether or not a bar or column chart is best to display data is not always clear. Data trends over time usually are best shown using columns. Figure 2.16 shows the comparison of exercise time for two drugs using a column chart. This is the same data used to prepare Figure 2.4(A) (also, see Exercise Problem 8 at the end of this chapter).

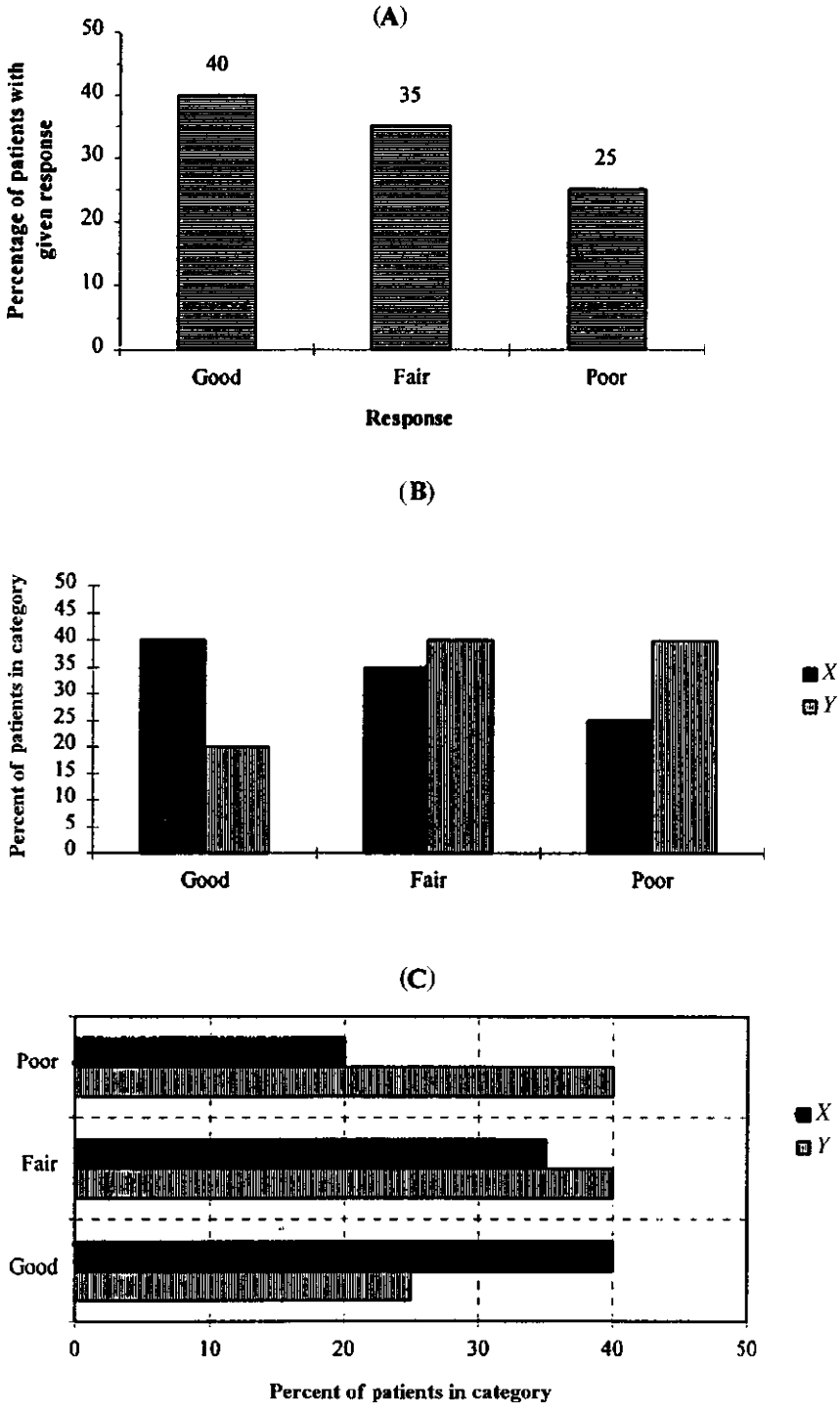


Figure 2.15 Graphical representation of patient responses to drug therapy.

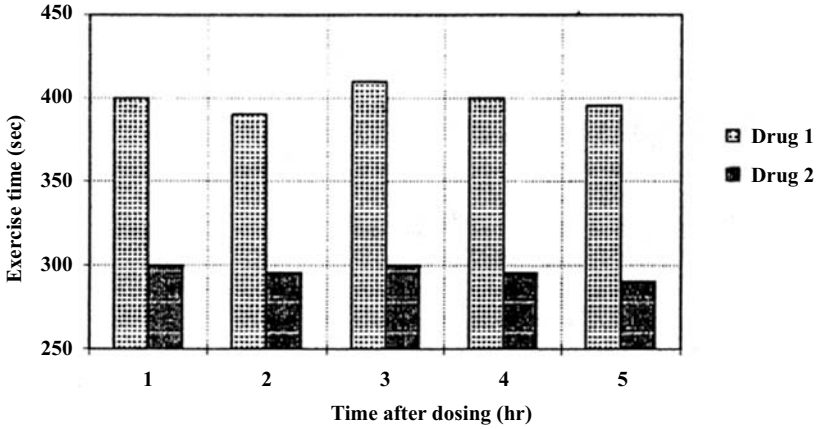


Figure 2.16 Exercise time for two drugs in the form of a column chart using data of Figure 2.4.

Pie charts are popular ways of presenting categorical data. Although the principles used in the construction of these charts are relatively simple, thought and care are necessary to convey the correct message. For example, dividing the circle into too many categories can be confusing and misleading. As a rule of thumb, no more than six sectors should be used. Another problem with pie charts is that it is not always easy to differentiate two segments that are reasonably close in size, whereas in the bar graph, values close in size are easily differentiated, since length is the critical feature.

The circle (or pie) represents 100%, or *all* of the results. Each segment (or slice of pie) has an area proportional to the area of the circle, representative of the contribution due to the particular segment. In the example shown in Figure 2.17(A), the pie represents the anti-inflammatory drug market. The slices are proportions of the market accounted for by major drugs in this therapeutic class. These charts are frequently used for business and economic descriptions, but can be applied to the presentation of scientific data in appropriate circumstances. Figure 2.17(B) shows the proportion of patients with good, fair, and poor responses to a drug in a clinical trial (see also Fig. 2.15).

Of course, we have not exhausted all possible ways of presenting data graphically. We have introduced the cumulative plot in section 1.2.3. Other kinds of plots are the stick diagram (analogous to the histogram) and frequency polygon [5]. The number of ways in which data can be presented is limited only by our own ingenuity. An elegant pictorial presentation of data can "make" a report or government submission. On the other hand, poor presentation of data can detract from an otherwise good report. The book *Statistical Graphics* by Calvin Schmid is recommended for those who wish detailed information on the presentation of graphs and charts.

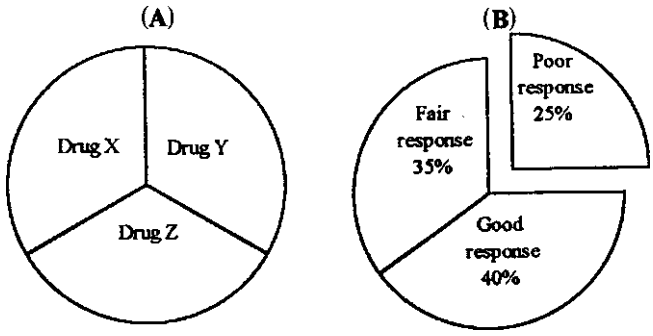


Figure 2.17 Examples of pie charts.

KEY TERMS

Bar charts
Bar graphs
Column charts
Correlation
Data pairs
Dependent variables
Histogram
Independent variables
Key
Pie charts
Scatter plots
Semilog plots

EXERCISES

1. Plot the following data, preparing and labeling the graph according to the guidelines outlined in this chapter. These data are the result of preparing various modifications of a formulation and observing the effect of the modifications on tablet hardness.

Formulation modification		
Starch (%)	**Lactose (%)**	**Tablet hardness (kg)**
10	5	8.3
10	10	9.1
10	15	9.6
10	20	10.2
5	5	9.1
5	10	9.4
5	15	9.8
5	20	10.4

(Hint: Plot these data on a single graph where the *Y* axis is tablet hardness and the *X* axis is lactose concentration. There will be two curves, one at 10% starch and the other at 5% starch.)

2. Prepare a histogram from the data of Table 1.3. Compare this histogram to that shown in Figure 2.2(A). Which do you think is a better representation of the data distribution?

3. Plot the following data and label the graph appropriately.

Patient	***X*: response to product A**	***Y*: response to product B**
1	2.5	3.8
2	3.6	2.4
3	8.9	4.7
4	6.4	5.9
5	9.5	2.1
6	7.4	5.0
7	1.0	8.5
8	4.7	7.8

What conclusion(s) can you draw from this plot if the responses are pain relief scores, where a high score means more relief?

4. A batch of tables was shown to have 70% with no defects, 15% slightly chipped, 10% discolored, and 5% dirty. Construct a pie chart from these data.

5. The following data from a dose–response experiment, a measure of physical activity, are the responses of five animals at each of three doses.

Dose (mg)	Responses
1	8, 12, 9, 14, 6
2	16, 20, 12, 15, 17
4	20, 17, 25, 27, 16

Plot the individual data points and the average at each dose versus (a) dose, (b) log dose.

6. The concentration of drug in solution was measured as a function of time.

Time (weeks)	Concentration
0	100
4	95
8	91
26	68
52	43

(a) Plot concentration versus time.
(b) Plot log concentration versus time.

7. Plot the following data and label the axes appropriately.

Patient	*X*: Cholesterol (mg%)	*Y*: Triglycerides (mg%)
1	180	80
2	240	180
3	200	70
4	300	200
5	360	240
6	240	200
Tablet	*X*: Tablet potency (mg)	*Y*: Tablet weight (mg)
1	5	300
2	6	300
3	4	280
4	5	295
5	6	320
6	4	290

8. Which figure do you think best represents the results of the exercise time study. Figure 2.16 or Figure 2.4(A)? If the presentation were to be used in a popular nontechnical journal read by laymen and physicians, which figure would you recommend?

REFERENCES

1. Fisher RA. Statistical Methods for Research Workers, 13th ed. New York, Hafner, 1963.
2. Schmid CF. Statistical Graphics. New York: Wiley, 1983.
3. Tufte ER. The Visual Display of Quantitative Data. Chelshire, CT: Graphics Press, 1983.
4. Gibaldi M, Perrier D. Pharmacokinetics, 2nd ed. New York: Marcel Dekker, 1982.
5. Dixon WJ, Massey FJ Jr. Introduction to Statistical Analysis, 3rd ed. New York: McGraw-Hill, 1969.