CHAPTER 3

Descriptive Statistics

3.1 INTRODUCTION

The beginning of an introductory statistics textbook usually contains a few paragraphs placing the subject matter in encyclopedic order, discussing the limitations or wide ramifications of the topic, and tends to the more philosophical rather than the substantive–scientific. Briefly, we consider science to be a study of the world emphasizing qualities of permanence, order, and structure. Such a study involves a drastic reduction of the real world, and often, numerical aspects only are considered. If there is no obvious numerical aspect or ordering, an attempt is made to impose it. For example, quality of medical care is not an immediately numerically scaled phenomenon but a scale is often induced or imposed. Statistics is concerned with the estimation, summarization, and obtaining of reliable numerical characteristics of the world. It will be seen that this is in line with some of the definitions given in the Notes in Chapter 1.

It may be objected that a characteristic such as the gender of a newborn baby is not numerical, but it can be coded (arbitrarily) in a numerical way; for example, 0 = male and 1 = female. Many such characteristics can be *labeled* numerically, and as long as the code, or the dictionary, is known, it is possible to go back and forth.

Consider a set of measurements of head circumferences of term infants born in a particular hospital. We have a quantity of interest—head circumference—which varies from baby to baby, and a collection of actual values of head circumferences.

Definition 3.1. A *variable* is a quantity that may vary from object to object.

Definition 3.2. A *sample* (or data set) is a collection of values of one or more variables. A member of the sample is called an *element*.

We distinguish between a variable and the value of a variable in the same way that the label "title of a book in the library" is distinguished from the title *Gray's Anatomy*. A variable will usually be represented by a capital letter, say, Y, and a value of the variable by a lowercase letter, say, y.

In this chapter we discuss briefly the types of variables typically dealt with in statistics. We then go on to discuss ways of *describing* samples of values of variables, both numerically and graphically. A key concept is that of a *frequency distribution*. Such presentations can be considered part of *descriptive statistics*. Finally, we discuss one of the earliest challenges to statistics, how to *reduce* samples to a few summarizing numbers. This will be considered under the heading of descriptive statistics.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

3.2 TYPES OF VARIABLES

3.2.1 Qualitative (Categorical) Variables

Some examples of qualitative (or categorical) variables and their values are:

1. Color of a person's hair (black, gray, red, . . . , brown)
2. Gender of child (male, female)
3. Province of residence of a Canadian citizen (Newfoundland, Nova Scotia, . . . , British Columbia)
4. Cause of death of newborn (congenital malformation, asphyxia, . . .)

Definition 3.3. A *qualitative variable* has values that are intrinsically nonnumerical (categorical).

As suggested earlier, the values of a qualitative variable can always be put into numerical form. The simplest numerical form is consecutive labeling of the values of the variable. The values of a qualitative variable are also referred to as *outcomes* or *states*.

Note that examples 3 and 4 above are ambiguous. In example 3, what shall we do with Canadian citizens living outside Canada? We could arbitrarily add another "province" with the label "Outside Canada." Example 4 is ambiguous because there may be more than one cause of death. Both of these examples show that it is not always easy to anticipate all the values of a variable. Either the list of values must be changed or the variable must be redefined.

The arithmetic operation associated with the values of qualitative variables is usually that of counting. Counting is perhaps the most elementary—but not necessarily simple—operation that organizes or abstracts characteristics. A *count* is an answer to the question: How many? (Counting assumes that whatever is counted shares some characteristics with the other "objects." Hence it disregards what is unique and reduces the objects under consideration to a common category or class.) Counting leads to statements such as "the number of births in Ontario in 1979 was 121,655."

Qualitative variables can often be ordered or ranked. *Ranking* or *ordering* places a set of objects in a sequence according to a specified scale. In Chapter 2, clinicians ranked interns according to the quality of medical care delivered. The "objects" were the interns and the scale was "quality of medical care delivered." The interns could also be ranked according to their height, from shortest to tallest—the "objects" are again the interns and the scale is "height." The provinces of Canada could be ordered by their population sizes from lowest to highest. Another possible ordering is by the latitudes of, say, the capitals of each province. Even hair color could be ordered by the wavelength of the dominant color. Two points should be noted in connection with ordering or qualitative variables. First, as indicated by the example of the provinces, there is more than one ordering that can be imposed on the outcomes of a variable (i.e., there is no natural ordering); the type of ordering imposed will depend on the nature of the variable and the purpose for which it is studied—if we wanted to study the impact of crowding or pollution in Canadian provinces, we might want to rank them by population size. If we wanted to study rates of melanoma as related to amount of ultraviolet radiation, we might want to rank them by the latitude of the provinces as summarized, say by the latitudes of the capitals or most populous areas. Second, the ordering need not be complete; that is, we may not be able to rank each outcome above or below another. For example, two of the Canadian provinces may have virtually identical populations, so that it is not possible to order them. Such orderings are called *partial*.

3.2.2 Quantitative Variables

Some examples of quantitative variables (with scale of measurement; values) are the following:

1. Height of father ($\frac{1}{2}$ inch units; 0.0, 0.5, 1.0, 1.5, . . . , 99.0, 99.5, 100.0)

2. Number of particles emitted by a radioactive source (counts per minute; 0, 1, 2, 3, . . .)
3. Total body calcium of a patient with osteoporosis (nearest gram; 0, 1, 2, . . . , 9999, 10,000)
4. Survival time of a patient diagnosed with lung cancer (nearest day; 0, 1, 2, . . . , 19,999, 20,000)
5. Apgar score of infant 60 seconds after birth (counts; 0, 1, 2, . . . , 8, 9, 10)
6. Number of children in a family (counts; 0, 1, 2, 3, . . .)

Definition 3.4. A *quantitative variable* has values that are intrinsically numerical.

As illustrated by the examples above, we must specify two aspects of a variable: the scale of measurement and the values the variable can take on. Some quantitative variables have numerical values that are integers, or discrete. Such variables are referred to as *discrete variables*. The variable "number of particles emitted by a radioactive source" is such an example; there are "gaps" between the successive values of this variable. It is not possible to observe 3.5 particles. (It is sometimes a source of amusement when discrete numbers are manipulated to produce values that cannot occur—for example, "the average American family" has 2.125 children). Other quantitative variables have values that are potentially associated with real numbers—such variables are called *continuous variables*. For example, the survival time of a patient diagnosed with lung cancer may be expressed to the nearest day, but this phrase implies that there has been rounding. We could refine the measurement to, say, hours, or even more precisely, to minutes or seconds. The exactness of the values of such a variable is determined by the precision of the measuring instrument as well as the usefulness of extending the value. Usually, a reasonable unit is assumed and it is considered *pedantic* to have a unit that is too refined, or *rough* to have a unit that does not permit distinction between the objects on which the variable is measured. Examples 1, 3, and 4 above deal with continuous variables; those in the other examples are discrete. Note that with quantitative variables there is a natural ordering (e.g., from lowest to highest value) (see Note 3.7 for another taxonomy of data).

In each illustration of qualitative and quantitative variables, we listed all the possible values of a variable. (Sometimes the values could not be listed, usually indicated by inserting three dots ". . . " into the sequence.) This leads to:

Definition 3.5. The *sample space* or *population* is the set of all possible values of a variable.

The definition or listing of the sample space is not a trivial task. In the examples of qualitative variables, we already discussed some ambiguities associated with the definitions of a variable and the sample space associated with the variable. Your definition must be reasonably precise without being "picky." Consider again the variable "province of residence of a Canadian citizen" and the sample space (Newfoundland, Nova Scotia, . . . , British Columbia). Some questions that can be raised include:

1. What about citizens living in the Northwest Territories? (Reasonable question)
2. Are landed immigrants who are not yet citizens to be excluded? (Reasonable question)
3. What time point is intended? Today? January 1, 2000? (Reasonable question)
4. If January 1, 2000 is used, what about citizens who died on that day? Are they to be included? (Becoming somewhat "picky")

3.3 *DESCRIPTIVE* STATISTICS

3.3.1 Tabulations and Frequency Distributions

One of the simplest ways to summarize data is by tabulation. John Graunt, in 1662, published his observations on bills of mortality, excerpts of which can be found in Newman [1956].

Table 3.1 Diseases and Casualties in the City of London 1632

Disease	Casualties
Abortive and stillborn	445
Affrighted	1
Aged	628
Ague	43
⋮	
Crisomes and infants	2268
⋮	
Tissick	34
Vomiting	1
Worms	27
In all	9535

Source: A selection from Graunt's tables; from Newman [1956].

Table 3.1 is a condensation of Graunt's list of 63 diseases and casualties. Several things should be noted about the table. To make up the table, three ingredients are needed: (1) a *collection* of objects (in this case, humans), (2) a *variable* of interest (the cause of death), and (3) the *frequency* of occurrence of each category. These are defined more precisely later. Second, we note that the disease categories are arranged alphabetically (ordering number 1). This may not be too helpful if we want to look at the most common causes of death. Let us rearrange Graunt's table by listing disease categories by greatest frequencies (ordering number 2).

Table 3.2 lists the 10 most common disease categories in Graunt's table and summarizes $8274/9535 = 87\%$ of the data in Table 3.1. From Table 3.2 we see at once that "crisomes" is the most frequent cause of death. (A *crisome* is an infant dying within one month of birth. Gaunt lists the number of "christenings" [births] as 9584, so a crude estimate of neonatal mortality is $2268/9584 \doteq 24\%$. The symbol "$\doteq$" means "approximately equal to.") Finally, we note that data for 1633 almost certainly would not have been identical to that of 1632. However, the number in the category "crisomes" probably would have remained the largest. An example of a statistical question is whether this predominance of "crisomes and infants" has a quality of permanence from one year to the next.

A second example of a tabulation involves keypunching errors made by a data-entry operator. To be entered were 156 lines of data, each line containing data on the number of crib deaths for a particular month in King County, Washington, for the years 1965–1977. Other data on

Table 3.2 Rearrangement of Graunt's Data (Table 3.1) by the 10 Most Common Causes of Death

Disease	Casualties	Disease	Casualties
Crisomes and infants	2268	Bloody flux, scowring, and flux	348
Consumption	1797	Dropsy and swelling	267
Fever	1108	Convulsion	241
Aged	628	Childbed	171
Flocks and smallpox	531		
Teeth	470	Total	8274
Abortive and stillborn	445		

Table 3.3 Number of Keypunching Errors per Line for 156 Consecutive Lines of Data Entered[a]

0	0	1	0	2	0	0	0	1	0	0	0
0	0	0	0	1	0	0	1	2	0	0	1
1	0	0	2	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	1	1	1	1	0	0	0	0	0	0	1
0	1	0	0	1	0	0	0	0	2	0	0
1	0	0	0	2	0	0	0	0	0	0	0
1	0	0	0	1	0	1	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0	0
0	1	0	1	1	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0

[a]Each digit represents the number of errors in a line.

a line consisted of meteorological data as well as the total number of births for that month in King County. Each line required the punching of 47 characters, excluding the spaces. The numbers of errors per line starting with January 1965 and ending with December 1977 are listed in Table 3.3.

One of the problems with this table is its bulk. It is difficult to grasp its significance. You would not transmit this table over the phone to explain to someone the number of errors made. One way to summarize this table is to specify how many times a particular combination of errors occurred. One possibility is the following:

Number of Errors per Line	Number of Lines
0	124
1	27
2	5
3 or more	0

This list is again based on three ingredients: a *collection* of lines of data, a *variable* (the number of errors per line), and the *frequency* with which values of the variable occur. Have we lost something in going to this summary? Yes, we have lost the order in which the observations occurred. That could be important if we wanted to find out whether errors came "in bunches" or whether there was a learning process, so that fewer errors occurred as practice was gained. The original data are already a condensation. The "number of errors per line" does not give information about the location of the errors in the line or the type of error. (For educational purposes, the latter might be very important.)

A difference between the variables of Tables 3.2 and 3.3 is that the variable in the second example was *numerically valued* (i.e., took on numerical values), in contrast with the *categorically valued* variable of the first example. Statisticians typically mean the former when *variable* is used by itself, and we will specify *categorical variable* when appropriate. [As discussed before, a categorical variable can always be made numerical by (as in Table 3.1) arranging the values alphabetically and numbering the observed categories 1, 2, 3, ... This is not biologically meaningful because the ordering is a function of the language used.]

The data of the two examples above were discrete. A different type of variable is represented by the age at death of crib death, or SIDS (sudden infant death syndrome), cases. Table 3.4

Table 3.4 Age at Death (in Days) of 78 Cases of SIDS Occurring in King County, Washington, 1976–1977

225	174	274	164	130	96	102	80	81	148	130	48
68	64	234	24	187	117	42	38	28	53	120	66
176	120	77	79	108	117	96	80	87	85	61	65
68	139	307	185	150	88	108	60	108	95	25	80
143	57	53	90	76	99	29	110	113	67	22	118
47	34	206	104	90	157	80	171	23	92	115	87
42	77	65	45	32	44						

Table 3.5 Frequency Distribution of Age at Death of 78 SIDS Cases Occurring in King County, Washington, 1976–1977

Age Interval (days)	Number of Deaths	Age Interval (days)	Number of Deaths
1–30	6	211–240	1
31–60	13	241–270	0
61–90	23	271–300	1
91–120	18	301–330	1
121–150	7		
151–180	5	Total	78
181–210	3		

displays ages at death in days of 78 cases of SIDS in King County, Washington, during the years 1976–1977. The variable, age at death, is continuous. However, there is rounding to the nearest whole day. Thus, "68 days" could represent 68.438 . . . or 67.8873 . . . , where the three dots indicate an unending decimal sequence.

Again, the table staggers us by its bulk. Unlike the preceding example, it will not be too helpful to list the number of times that a particular value occurs: There are just too many different ages. One way to reduce the bulk is to define intervals of days and count the number of observations that fall in each interval. Table 3.5 displays the data grouped into 30-day intervals (months). Now the data make more sense. We note, for example, that many deaths occur between the ages of 61 and 90 days (two to three months) and that very few deaths occur after 180 days (six months). Somewhat surprisingly, there are relatively few deaths in the first month of life. This age distribution pattern is unique to SIDS.

We again note the three characteristics on which Table 3.5 is based: (1) a *collection* of 78 objects—SIDS cases, (2) a *variable* of interest—age at death, and (3) the *frequency* of occurrence of values falling in specified intervals. We are now ready to define these three characteristics more explicitly.

Definition 3.6. An *empirical frequency distribution* (EFD) of a variable is a listing of the values or ranges of values of the variable together with the frequencies with which these values or ranges of values occur.

The adjective *empirical* emphasizes that an *observed* set of values of a variable is being discussed; if this is obvious, we may use just "frequency distribution" (as in the heading of Table 3.5).

The choice of interval width and interval endpoint is somewhat arbitrary. They are usually chosen for convenience. In Table 3.5, a "natural" width is 30 days (one month) and convenient endpoints are 1 day, 31 days, 61 days, and so on. A good rule is to try to produce between

seven and 10 intervals. To do this, divide the range of the values (*largest to smallest*) by 7, and then adjust to make a simple interval. For example, suppose that the variable is "weight of adult male" (expressed to the nearest kilogram) and the values vary from 54 to 115 kg. The range is $115 - 54 = 61$ kg, suggesting intervals of width $61/7 \doteq 8.7$ kg. This is clearly not a very good width; the closest "natural" width is 10 kg (producing a slightly coarser grid). A reasonable starting point is 50 kg, so that the intervals have endpoints 50 kg, 60 kg, 70 kg, and so on.

To compare several EFDs it is useful to make them comparable with respect to the total number of subjects. To make them comparable, we need:

Definition 3.7. The *size* of a sample is the number of elements in the sample.

Definition 3.8. An *empirical relative frequency distribution* (ERFD) is an empirical frequency distribution where the frequencies have been divided by the sample size.

Equivalently, the relative frequency of the value of a variable is the proportion of times that the value of the variable occurs. (The context often makes it clear that an *empirical* frequency distribution is involved. Similarly, many authors omit the adjective *relative* so that "frequency distribution" is shorthand for "empirical relative frequency distribution.")

To illustrate ERFDs, consider the data in Table 3.6, consisting of systolic blood pressures of three groups of Japanese men: native Japanese, first-generation immigrants to the United States (Issei), and second-generation Japanese in the United States (Nisei). The sample sizes are 2232, 263, and 1561, respectively.

It is difficult to compare these distributions because the sample sizes differ. The *relative* frequencies (proportions) are obtained by dividing each frequency by the corresponding sample size. The ERFD is presented in Table 3.7. For example, the (empirical) relative frequency of native Japanese with systolic blood pressure less than 106 mmHg is $218/2232 = 0.098$.

It is still difficult to make comparisons. One of the purposes of the study was to determine how much variables such as blood pressure were affected by environmental conditions. To see if there is a *shift* in the blood pressures, we could consider the proportion of men with blood pressures less than a specified value and compare the groups that way. Consider, for example, the proportion of men with systolic blood pressures less than or equal to 134 mmHg. For the native Japanese this is (Table 3.7) $0.098 + 0.122 + 0.151 + 0.162 = 0.533$, or 53.3%. For the Issei and Nisei these figures are 0.413 and 0.508, respectively. The latter two figures are somewhat lower than the first, suggesting that there has been a shift to higher systolic

Table 3.6 Empirical Frequency Distribution of Systolic Blood Pressure of Native Japanese and First- and Second-Generation Immigrants to the United States, Males Aged 45–69 Years

Blood Pressure (mmHg)	Native Japanese	Issei	California Nisei
<106	218	4	23
106–114	272	23	132
116–124	337	49	290
126–134	362	33	347
136–144	302	41	346
146–154	261	38	202
156–164	166	23	109
>166	314	52	112
Total	2232	263	1561

Source: Data from Winkelstein et al. [1975].

Table 3.7 Empirical Relative Frequency Distribution of Systolic Blood Pressure of Native Japanese and First- and Second-Generation Immigrants to the United States, Males Aged 45–69 Years

Blood Pressure (mmHg)	Native Japanese	Issei	California Nisei
<106	0.098	0.015	0.015
106–114	0.122	0.087	0.085
116–124	0.151	0.186	0.186
126–134	0.162	0.125	0.222
136–144	0.135	0.156	0.222
146–154	0.117	0.144	0.129
156–164	0.074	0.087	0.070
>166	0.141	0.198	0.072
Total	1.000	0.998	1.001
Sample size	(2232)	(263)	(1561)

Source: Data from Winkelstein et al. [1975].

blood pressure among the immigrants. Whether this shift represents sampling variability or a genuine shift in these groups can be determined by methods developed in the next three chapters.

The concept discussed above is formalized in the empirical cumulative distribution.

Definition 3.9. The *empirical cumulative distribution* (ECD) of a variable is a listing of values of the variable together with the *proportion* of observations less than or equal to that value (cumulative proportion).

Before we construct the ECD for a sample, we need to clear up one problem associated with rounding of values of continuous variables. Consider the age of death of the SIDS cases of Table 3.4. The first age listed is 225 days. Any value between 224.5+ and 225.5− is rounded off to 225 (224.5+ indicates a value greater than 224.5 by some arbitrarily small amount, and similarly, 225.5− indicates a value less than 225.5). Thus, the upper endpoint of the interval 1–30 days in Table 3.5 is $30.4\overline{9}$, or 30.5.

The ECD associated with the data of Table 3.5 is presented in Table 3.8, which contains (1) the age intervals, (2) endpoints of the intervals, (3) EFD, (4) ERFD, and (5) ECD.

Two comments are in order: (1) there is a slight rounding error in the last column because the relative frequencies are rounded to three decimal places—if we had calculated from the frequencies rather than the relative frequencies, this problem would not have occurred; and (2) given the cumulative proportions, the original proportions can be recovered. For example, consider the following endpoints and their cumulative frequencies:

150.5	0.860
180.5	0.924

Subtracting, $0.924 - 0.860 = 0.064$ produces the proportion in the interval 151–180. Mathematically, the ERFD and the ECD are equivalent.

Table 3.8 Frequency Distribution of Age at Death of 78 SIDS Cases Occurring in King County, Washington, 1976–1977

Age Interval (days)	Endpoint of Interval (days)	Number of Deaths	Relative Frequency (Proportion)	Cumulative Proportion
1–30	30.5	6	0.077	0.077
31–60	60.5	13	0.167	0.244
61–90	90.5	23	0.295	0.539
91–120	120.5	18	0.231	0.770
121–150	150.5	7	0.090	0.860
151–180	180.5	5	0.064	0.924
181–210	210.5	3	0.038	0.962
211–240	240.5	1	0.013	0.975
241–270	270.5	0	0.000	0.975
271–300	300.5	1	0.013	0.988
301–330	330.5	1	0.013	1.001
Total		78	1.001	

3.3.2 Graphs

Graphical displays frequently provide very effective descriptions of samples. In this section we discuss some very common ways of doing this and close with some examples that are innovative. Graphs can also be used to enhance certain features of data as well as to distort them. A good discussion can be found in Huff [1993].

One of the most common ways of describing a sample pictorially is to plot on one axis values of the variable and on another axis the frequency of occurrence of a value or a measure related to it. In constructing a *histogram* a number of cut points are chosen and the data are tabulated. The relative frequency of observations in each category is divided by the width of the category to obtain the *probability density*, and a bar is drawn with this height. The area of a bar is proportional to the frequency of occurrence of values in the interval.

The most important choice in drawing a histogram is the number of categories, as quite different visual impressions can be conveyed by different choices. Figure 3.1 shows measurements of albumin, a blood protein, in 418 patients with the liver disease *primary biliary cirrhosis*, using

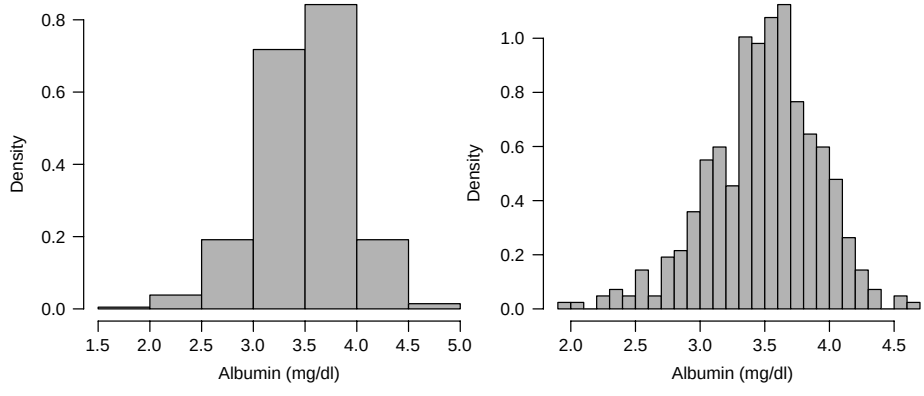


Figure 3.1 Histograms of serum albumin concentration in 418 PBC patients, using two different sets of categories.

data made available on the Web by T. M. Therneau of the Mayo Clinic. With five categories the distribution appears fairly symmetric, with a single peak. With 30 categories there is a definite suggestion of a second, lower peak. Statistical software will usually choose a sensible default number of categories, but it may be worth examining other choices.

The values of a variable are usually plotted on the abscissa (x-axis), the frequencies on the ordinate (y-axis). The ordinate on the left-hand side of Figure 3.1 contains the probability densities for each category. Note that the use of probability density means that the two histograms have similar vertical scales despite having different category widths: As the categories become narrower, the numerator and denominator of the probability density decrease together.

Histograms are sometimes defined so that the y-axis measures absolute or relative frequency rather than the apparently more complicated probability density. Two advantages arise from the use of a probability density rather than a simple count. The first is that the categories need not have the same width: It is possible to use wider categories in parts of the distribution where the data are relatively sparse. The second advantage is that the height of the bars does not depend systematically on the sample size: It is possible to compare on the same graph histograms from two samples of different sizes. It is also possible to compare the histogram to a hypothesized mathematical distribution by drawing the mathematical density function on the same graph (an example is shown in Figure 4.7.

Figure 3.2 displays the empirical cumulative distribution (ECD). This is a *step function* with jumps at the endpoints of the interval. The height of the jump is equal to the relative frequency of the observations in the interval. The ECD is nondecreasing and is bounded above by 1. Figure 3.2 emphasizes the discreteness of data. A *frequency polygon* and *cumulative frequency polygon* are often used with continuous variables to emphasize the continuity of the data. A frequency polygon is obtained by joining the heights of the bars of the histogram at their midpoints. The frequency polygon for the data of Table 3.8 is displayed in Figure 3.3. A question arises: Where is the midpoint of the interval? To calculate the midpoint for the interval 31–60 days, we note

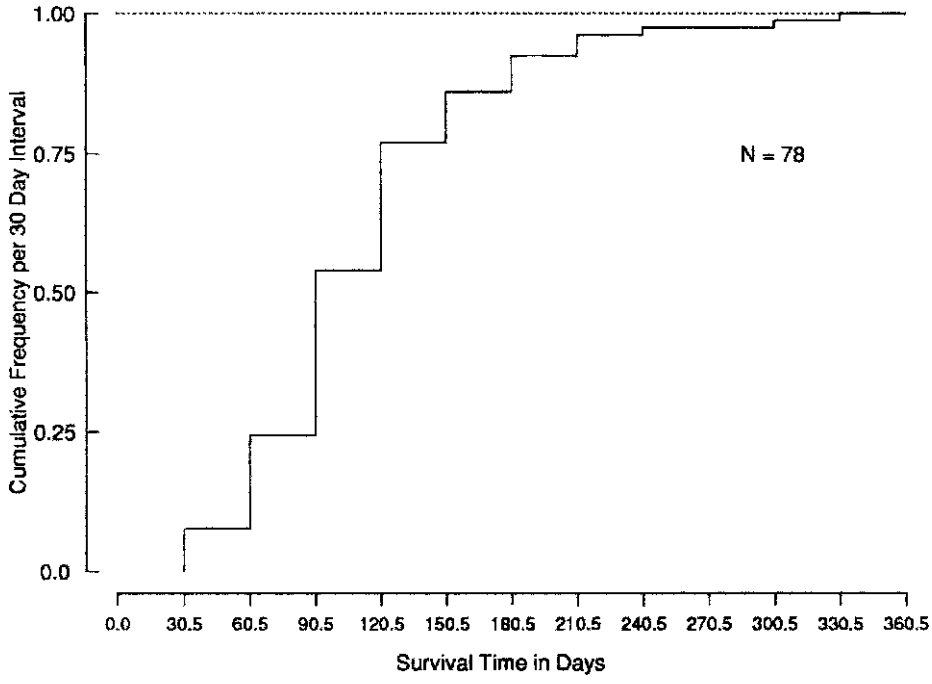


Figure 3.2 Empirical cumulative distribution of SIDS deaths.

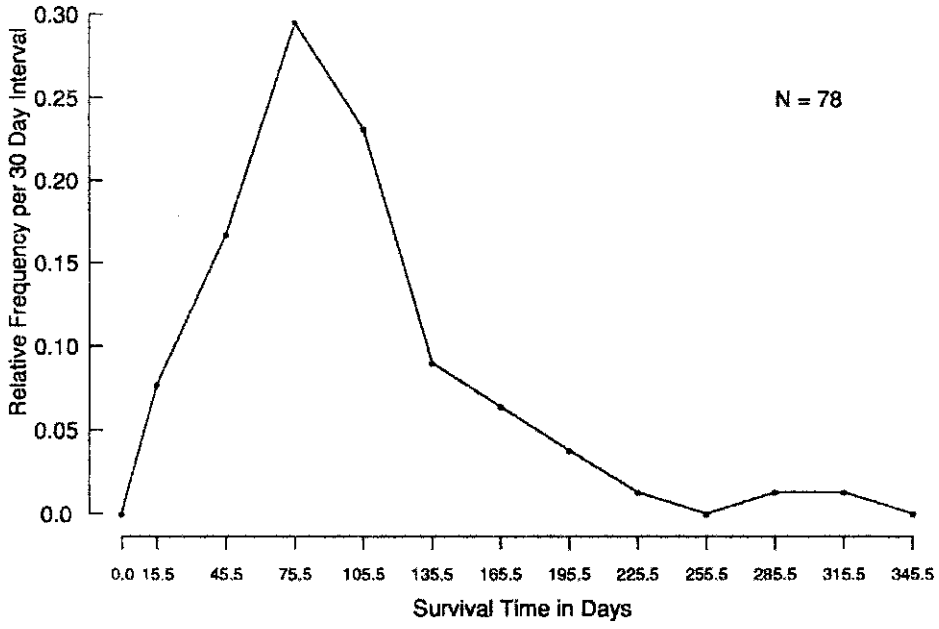


Figure 3.3 Frequency polygon of SIDS deaths.

that the limits of this interval are 30.5–60.5. The midpoint is halfway between these endpoints; hence, $midpoint = (30.5 + 60.5)/2 = 45.5$ days.

All midpoints are spaced in intervals of 30 days, so that the midpoints are 15.5, 45.5, 75.5, and so on. To close the polygon, the midpoints of two additional intervals are needed: one to the left of the first interval (1–30) and one to the right of the last interval observed (301–330), both of these with zero observed frequencies.

A cumulative frequency polygon is constructed by joining the cumulative relative frequencies observed at the endpoints of their respective intervals. Figure 3.4 displays the cumulative relative frequency of the SIDS data of Table 3.8. The curve has the value 0.0 below 0.5 and the value 1.0 to the right of 330.5. Both the histograms and the cumulative frequency graphs implicitly assume that the observations in our interval are evenly distributed over that interval.

One advantage of a cumulative frequency polygon is that the proportion (or percentage) of observations less than a specified value can be read off easily from the graph. For example, from Figure 3.4 it can be seen that 50% of the observations have a value of less than 88 days (this is the median of the sample). See Section 3.4.1 for further discussion.

EFDs can often be graphed in an innovative way to illustrate a point. Consider the data in Figure 3.5, which contains the frequency of births per day as related to phases of the moon. Data were collected by Schwab [1975] on the number of births for two years, grouped by each day of the 29-day lunar cycle, presented here as a circular distribution where the lengths of the sectors are proportional to the frequencies. (There is clearly no evidence supporting the hypothesis that the cycle of the moon influences birth rate.)

Sometimes more than one variable is associated with each of the objects under study. Data arising from such situations are called *multivariate data*. A moment's reflection will convince you that most biomedical data are multivariate in nature. For example, the variable "blood pressure of a patient" is usually expressed by two numbers, systolic and diastolic blood pressure. We often specify age and gender of patients to characterize blood pressure more accurately.

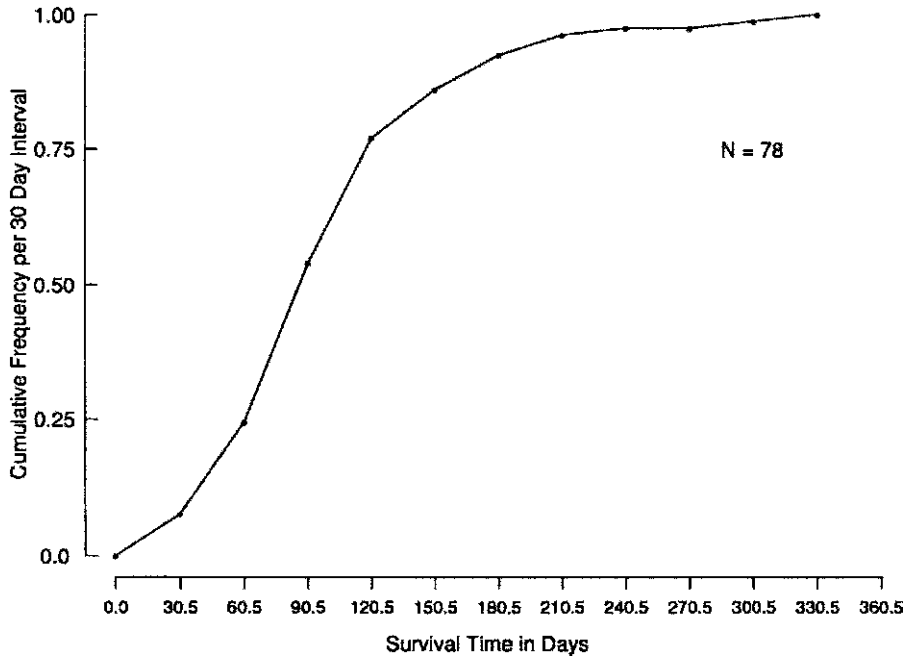


Figure 3.4 Cumulative frequency polygon of SIDS deaths.

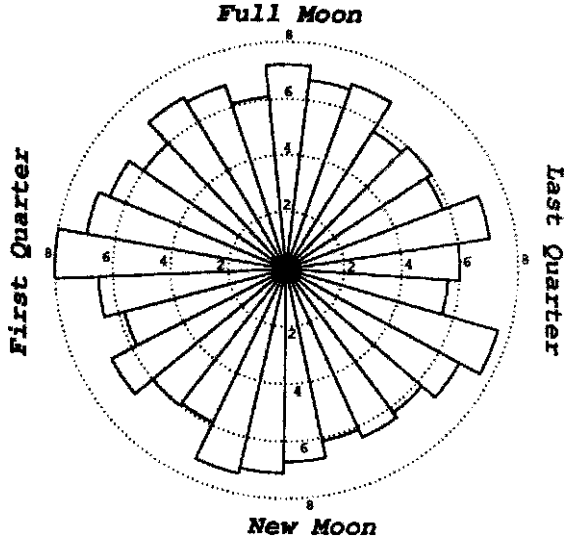


Figure 3.5 Average number of births per day over a 29-day lunar cycle. (Data from Schwab [1975].)

In the multivariate situation, in addition to describing the frequency with which each value of each variable occurs, we may want to study the relationships among the variables. For example, Table 1.2 and Figure 1.1 attempt to assess the relationship between the variables "clinical competence" and "cost of laboratory procedures ordered" of interns. Graphs of multivariate data will be found throughout the book.

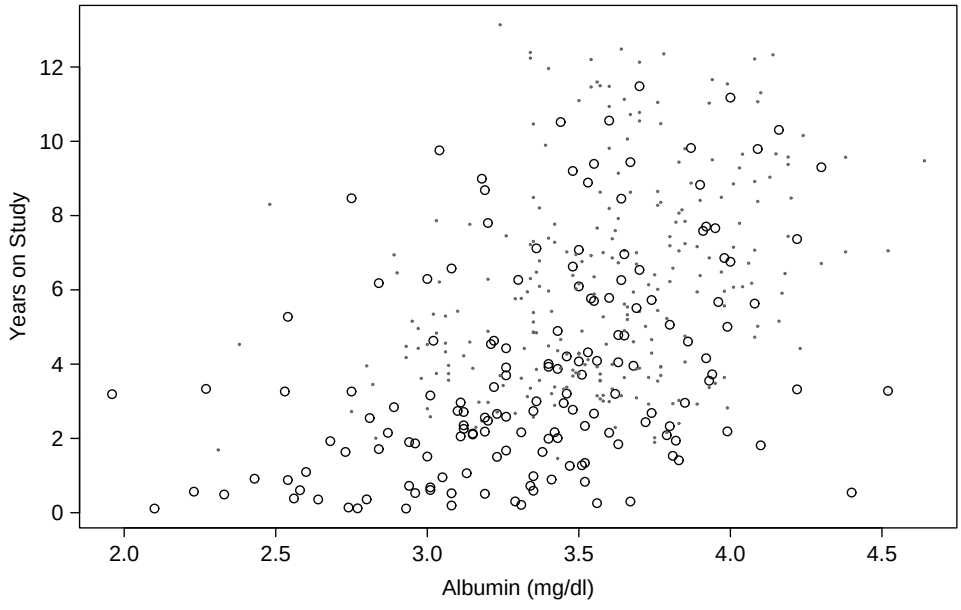


Figure 3.6 Survival time in primary biliary cirrhosis by serum albumin concentrations. Large circles are deaths, small circles are patients alive at last contact. (Data from Fleming and Harrington [1991].)

Here we present a few examples of visually displaying values of several variables at the same time. A simple one relates the serum albumin values from Figure 3.1 to survival time in the 418 patients. We do not know the survival times for everyone, as some were still alive at the end of the study. The statistical analysis of such data occupies an entire chapter of this book, but a simple descriptive graph is possible. Figure 3.6 shows large circles at survival time for patients who died. For those still alive it shows small circles at the last time known alive. For exploratory analysis and presentation these could be indicated by different colors, something that is unfortunately still not feasible for this book.

Another simple multivariate example can be found in our discussion of factor analysis. Figure 14.7 shows a matrix of correlations between variables using shaded circles whose size shows the strength of the relationship and whose shading indicates whether the relationship is positive or negative. Figure 14.7 is particularly interesting, as the graphical display helped us find an error that we missed in the first edition.

A more sophisticated example of multivariate data graphics is the *conditioning plot* [Cleveland, 1993]. This helps you examine how the relationship between two variables depends on a third. Figure 3.7 shows daily data on ozone concentration and sunlight in New York, during the summer of 1973. These should be related monotonically; ozone is produced from other pollutants by chemical reactions driven by sunlight. The four panels show four plots of ozone concentration vs. solar radiation for various ranges of temperature. The shaded bar in the title of each plot indicates the range of temperatures. These ranges overlap, which allows more panels to be shown without the data becoming too sparse. Not every statistical package will produce these coplots with a single function, but it is straightforward to draw them by taking appropriate subsets of your data.

The relationship clearly varies with temperature. At low temperatures there is little relationship, and as the temperature increases the relationship becomes stronger. Ignoring the effect of temperature and simply graphing ozone and solar radiation results in a more confusing relationship (examined in Figure 3.9). In Problem 10 we ask you to explore these data further.

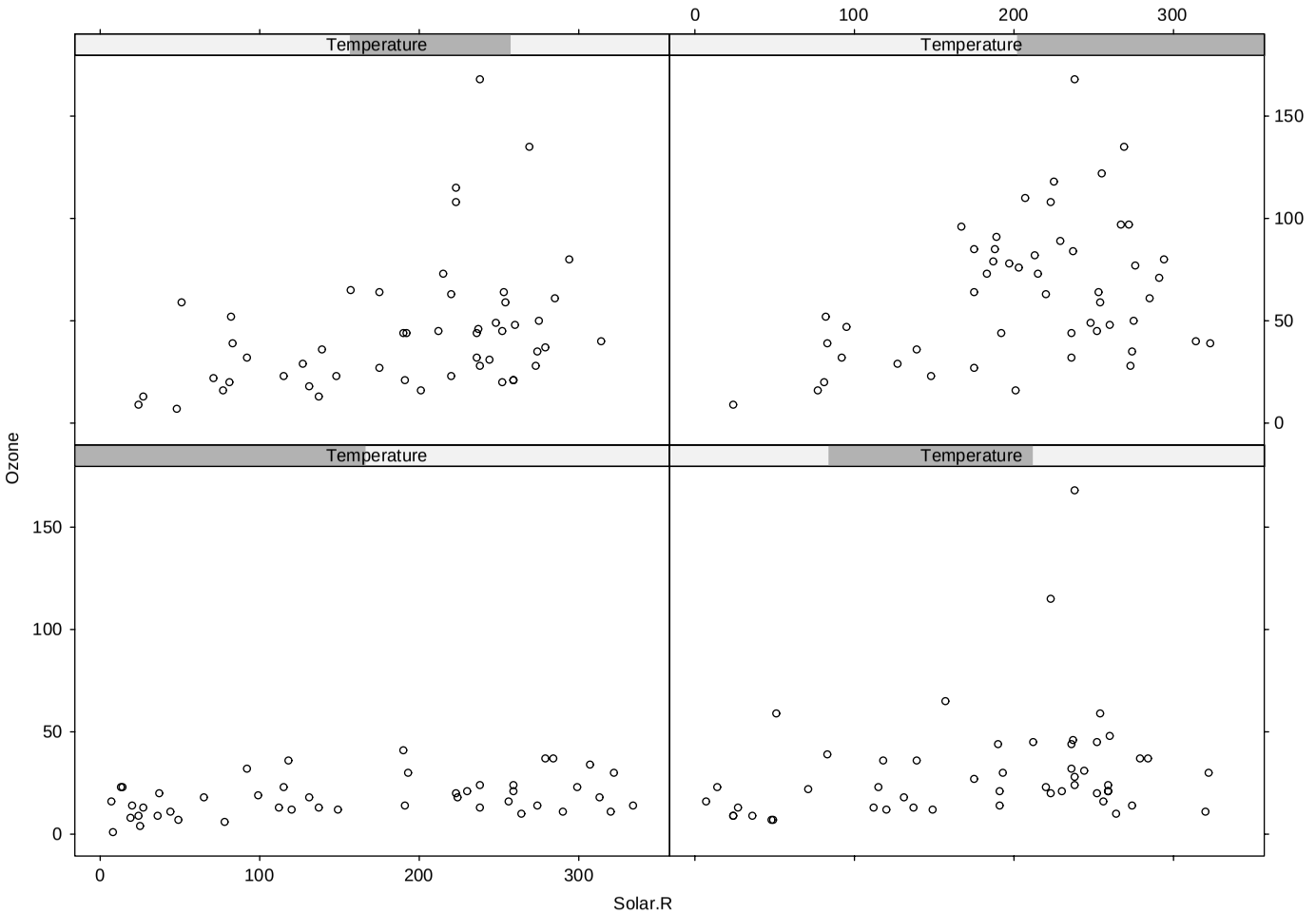


Figure 3.7 Ozone concentration by solar radiation intensity in New York, May–September 1973, conditioned on temperature. (From R Foundation [2002].)

For beautiful books on the visual display of data, see Tufte [1990, 1997, 2001]. A very readable compendium of graphical methods is contained in Moses [1987], and more recent methods are described by Cleveland [1994]. Wilkinson [1999] discusses the structure and taxonomy of graphs.

3.4 DESCRIPTIVE *STATISTICS*

In Section 3.3 our emphasis was on tabular and visual display of data. It is clear that these techniques can be used to great advantage when summarizing and highlighting data. However, even a table or a graph takes up quite a bit of space, cannot be summarized in the mind too easily, and particularly for a graph, represents data with some imprecision. For these and other reasons, numerical characteristics of data are calculated routinely.

Definition 3.10. A *statistic* is a numerical characteristic of a sample.

One of the functions of statistics as a field of study is to describe samples by as few numerical characteristics as possible. Most numerical characteristics can be classified broadly into statistics derived from percentiles of a frequency distribution and statistics derived from moments of a frequency distribution (both approaches are explained below). Roughly speaking, the former approach tends to be associated with a statistical methodology usually termed *nonparametric*, the latter with *parametric* methods. The two classes are used, contrasted, and evaluated throughout the book.

3.4.1 Statistics Derived from Percentiles

A *percentile* has an intuitively simple meaning—for example, the 25th percentile is that value of a variable such that 25% of the observations are less than that value and 75% of the observations are greater. You can supply a similar definition for, say, the 75th percentile. However, when we apply these definitions to a particular sample, we may run into three problems: (1) small sample size, (2) tied values, or (3) nonuniqueness of a percentile. Consider the following sample of four observations:

$$22, 22, 24, 27$$

How can we define the 25th percentile for this sample? There is no value of the variable with this property. But for the 75th percentile, there is an infinite number of values—for example, 24.5, 25, and 26.9378 all satisfy the definition of the 75th percentile. For large samples, these problems disappear and we will define percentiles for small samples in a way that is consistent with the intuitive definition. To find a particular percentile in practice, we would rank the observations from smallest to largest and count until the proportion specified had been reached. For example, to find the 50th percentile of the four numbers above, we want to be somewhere between the second- and third-largest observation (between the values for ranks 2 and 3). Usually, this value is taken to be halfway between the two values. This could be thought of as the value with rank 2.5—call this a *half rank*. Note that

$$2.5 = \left(\frac{50}{100}\right)(1 + \text{sample size})$$

You can verify that the following definition is consistent with your intuitive understanding of percentiles:

Definition 3.11. The Pth *percentile* of a sample of n observations is that value of the variable with rank $(P/100)(1 + n)$. If this rank is not an integer, it is rounded to the nearest half rank.

The following data deal with the aflatoxin levels of raw peanut kernels as described by Quesenberry et al. [1976]. Approximately 560 g of ground meal was divided among 16 centrifuge bottles and analyzed. One sample was lost, so that only 15 readings are available (measurement units are not given). The values were

$$30, 26, 26, 36, 48, 50, 16, 31, 22, 27, 23, 35, 52, 28, 37$$

The 50th percentile is that value with rank $(50/100)(1 + 15) = 8$. The eighth largest (or smallest) observation is 30. The 25th percentile is the observation with rank $(25/100)(1+15) = 4$, and this is 26. Similarly, the 75th percentile is 37. The 10th percentile (or decile) is that value with rank $(10/100)(1 + 15) = 1.6$, so we take the value halfway between the smallest and second-smallest observation, which is $(1/2)(16 + 22) = 19$. The 90th percentile is the value with rank $(90/100)(1 + 15) = 14.4$; this is rounded to the nearest half rank of 14.5. The value with this half rank is $(1/2)(50 + 52) = 51$.

Certain percentile or functions of percentiles have specific names:

Percentile	Name
50	Median
25	Lower quartile
75	Upper quartile

All these statistics tell something about the location of the data. If we want to describe how spread out the values of a sample are, we can use the range of values (largest minus smallest), but a problem is that this statistic is very dependent on the sample size. A better statistic is given by:

Definition 3.12. The *interquartile range* (IQR) is the difference between the 75th and 25th percentiles.

For the aflatoxin example, the interquartile range is $37 - 26 = 11$. Recall the *range* of a set of numbers is the largest value minus the smallest value. The data can be summarized as follows:

Median	30	Measures of location
Minimum	16	
Maximum	52	
Interquartile range	11	Measures of spread
Range	36	

The first three measures describe the location of the data; the last two give a description of their spread. If we were to add 100 to each of the observations, the median, minimum, and maximum would be shifted by 100, but the interquartile range and range would be unaffected.

These data can be summarized graphically by means of a *box plot* (also called a *box-and-whisker plot*). A rectangle with upper and lower edges at the 25th and 75th percentiles is drawn with a line in the rectangle at the median (50th percentile). Lines (whiskers) are drawn from the rectangle (box) to the highest and lowest values that are within $1.5 \times$ IQR of the median; any points more extreme than this are plotted individually. This is Tukey's [1977] definition of the box plot; an alternative definition draws the whiskers from the quartiles to the maximum and minimum.

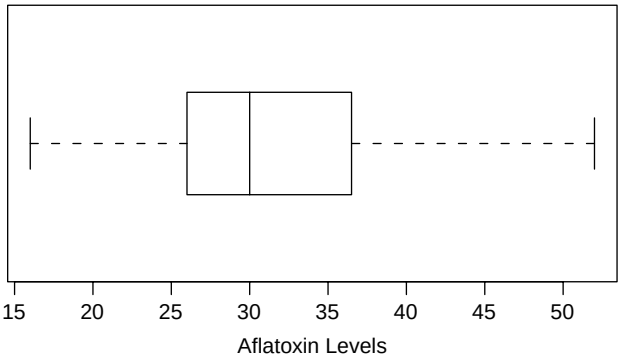


Figure 3.8 Box plot.

The box plot for these data (Figure 3.8) indicates that 50% of the data between the lower and upper quartiles is distributed over a much narrower range than the remaining 50% of the data. There are no extreme values outside the "fences" at median $\pm$ 1.5 $\times$ IQR.

3.4.2 Statistics Derived from Moments

The statistics discussed in Section 3.4.1 dealt primarily with describing the location and the variation of a sample of values of a variable. In this section we introduce another class of statistics, which have a similar purpose. In this class are the ordinary average, or arithmetic mean, and standard deviation. The reason these statistics are said to be derived from *moments* is that they are based on powers or moments of the observations.

Definition 3.13. The *arithmetic mean* of a sample of values of a variable is the average of all the observations.

Consider the aflatoxin data mentioned in Section 3.4.1. The arithmetic mean of the data is

$$\frac{30 + 26 + 26 + \cdots + 28 + 37}{15} = \frac{487}{15} = 32.4\overline{6} \doteq 32.5$$

A reasonable rule is to express the mean with one more significant digit than the observations, hence we round $32.4\overline{6}$—a nonterminating decimal—to 32.5. (See also Note 3.2 on significant digits and rounding.)

Notation. The specification of some of the statistics to be calculated can be simplified by the use of notation. We use a capital letter for the name of a variable and the corresponding lowercase letter for a value. For example, $Y =$ *aflatoxin level* (the name of the variable); $y = 30$ (the value of aflatoxin level for a particular specimen). We use the Greek symbol $\sum$ to mean "sum all the observations." Thus, for the aflatoxin example, $\sum y$ is shorthand for the statement "sum all the aflatoxin levels." Finally, we use the symbol $\overline{y}$ to denote the arithmetic mean of the sample. The arithmetic mean of a sample of n values of a variable can now be written as

$$\overline{y} = \frac{\sum y}{n}$$

For example, $\sum y = 487, n = 15$, and $\overline{y} = 487/15 \doteq 32.5$. Consider now the variable of Table 3.3: the number of keypunching errors per line. Suppose that we want the average

Table 3.9 Calculation of Arithmetic Average from Empirical Frequency and Empirical Relative Frequency Distribution[a]

Number of Errors per Line, y	Number of Lines, f	Proportion of Lines, p	$p \times y$
0	124	0.79487	0.00000
1	27	0.17308	0.17308
2	5	0.03205	0.06410
3	0	0.00000	0.00000
Total	156	1.00000	0.23718

[a]Data from Table 3.3.

number of errors per line. By definition, this is $(0+0+1+0+2+\cdots+0+0+0+0)/156 = 37/156 \doteq 0.2$ error per line. But this is a tedious way to calculate the average. A simpler way utilizes the frequency distribution or relative frequency distribution.

The total number of errors is $(124 \times 0) + (27 \times 1) + (5 \times 2) + (0 \times 3) = 37$; that is, there are 124 lines without errors; 27 lines each of which contains one error, for a total of 27 errors for these types of lines; and 5 lines with two errors, for a total of 10 errors for these types of lines; and finally, no lines with 3 errors (or more). So the arithmetic mean is

$$\overline{y} = \frac{\sum fy}{\sum f} = \frac{\sum fy}{n}$$

since the frequencies, f, add up to n, the sample size. Here, the sum $\sum fy$ is over observed values of y, each value appearing once.

The arithmetic mean can also be calculated from the empirical relative frequencies. We use the following algebraic property:

$$\overline{y} = \frac{\sum fy}{n} = \sum \frac{fy}{n} = \sum \frac{f}{n} y = \sum py$$

The *f/n* are precisely the empirical relative frequencies or proportions, p. The calculations using proportions are given in Table 3.9. The value obtained for the sample mean is the same as before. The formula $\overline{y} = \sum py$ will be used extensively in Chapter 4 when we come to probability distributions. If the values y represent the midpoints of intervals in an empirical frequency distribution, the mean of the grouped data can be calculated in the same way.

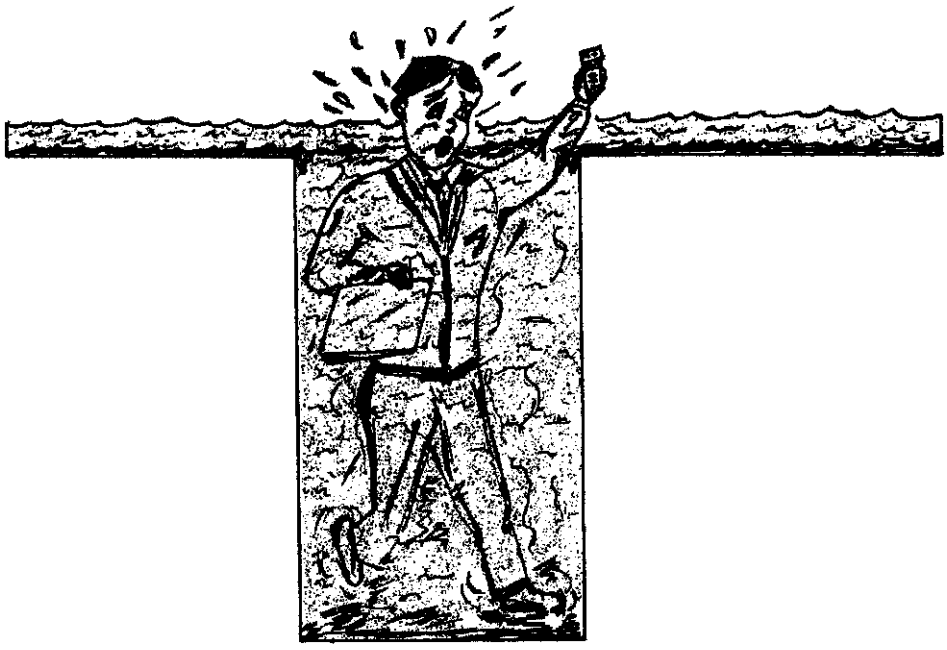
Analogous to the interquartile range there is a measure of spread based on sample moments.

Definition 3.14. The *standard deviation* of a sample of n values of a variable Y is

$$s = \sqrt{\frac{\sum (y - \overline{y})^2}{n - 1}}$$

Roughly, the standard deviation is the square root of the average of the square of the deviations from the sample mean. The reason for dividing by $n - 1$ is explained in Note 3.5. Before giving an example, we note the following properties of the standard deviation:

1. The standard deviation has the same units of measurement as the variable. If the observations are expressed in centimeters, the standard deviation is expressed in centimeters.



Cartoon 3.1 Variation is important: statistician drowning in a river of average depth 10.634 inches.

2. If a constant value is added to each of the observations, the value of the standard deviation is unchanged.
3. If the observations are multiplied by a positive constant value, the standard deviation is multiplied by the same constant value.
4. The following two formulas are sometimes computationally more convenient in calculating the standard deviation by hand:

$$s = \sqrt{\frac{\sum y^2 - n\overline{y}^2}{n-1}} = \sqrt{\frac{\sum y^2 - (\sum y)^2/n}{n-1}}$$

   Rounding errors accumulate more rapidly using these formulas; care should be taken to carry enough significant digits in the computation.
5. The square of the standard deviation is called the *variance*.
6. In many situations the standard deviation can be approximated by

$$s \doteq \frac{\text{interquartile range}}{1.35}$$

7. In many cases it is true that approximately 68% of the observations fall within one standard deviation of the mean; approximately 95% within two standard deviations.

3.4.3 Graphs Based on Estimated Moments

One purpose for drawing a graph of two variables X and Y is to decide how Y changes as X changes. Just as statistics such as the mean help summarize the location of one or two samples,

they can be used to summarize how the location of Y changes with X. A simple way to do this is to divide the data into *bins* and compute the mean or median for each bin.

Example 3.1. Consider the New York air quality data in Figure 3.7. When we plot ozone concentrations against solar radiation without conditioning variables, there is an apparent triangular relationship. We might want a summary of this relationship rather than trying to assess it purely by eye. One simple summary is to compute the mean ozone concentration for various ranges of solar radiation. We compute the mean ozone for days with solar radiation 0–50 langleys, 50–150, 100–200, 150–250, and so on. Plotting these means at the midpoint of the interval and joining the dots gives the dotted line shown in Figure 3.9.

Modern statistical software provides a variety of different *scatter plot smoothers* that perform more sophisticated versions of this calculation. The technical details of these are complicated, but they are conceptually very similar to the local means that we used above. The solid line in Figure 3.9 is a popular scatter plot smoother called *lowess* [Cleveland, 1981].

3.4.4 Other Measures of Location and Spread

There are many other measures of location and spread. In the former category we mention the mode and the geometric mean.

Definition 3.15. The *mode* of a sample of values of a variable Y is that value that occurs most frequently.

The mode is usually calculated for large sets of discrete data. Consider the data in Table 3.10, the distribution of the number of boys per family of eight children. The most frequently occurring value of the variable Y, the number of boys per family of eight children, is 4. There are more families with that number of boys than any other specified number of boys. For data arranged in histograms, the mode is usually associated with the midpoint of the interval having the highest frequency. For example, the mode of the systolic blood pressure of the native Japanese men listed in Table 3.6 is 130 mmHg; the modal value for Issei is 120 mmHg.

Definition 3.16. The *geometric mean* of a sample of nonnegative values of a variable Y is the nth root of the product of the n values, where n is the sample size.

Equivalently, it is the antilogarithm of the arithmetic mean of the logarithms of the values. (See Note 3.1 for a brief discussion of logarithms.)

Consider the following four observations of systolic blood pressure in mmHg:

$$118, 120, 122, 160$$

The arithmetic mean is 130 mmHg, which is larger than the first three values because the 160 mmHg value "pulls" the mean to the right. The geometric mean is $(118 \times 120 \times 122 \times 160)^{1/4} \doteq 128.9$ mmHg. The geometric mean is less affected by the extreme value of 160 mmHg. The median is 121 mmHg. If the value of 160 mmHg is changed to a more extreme value, the mean will be affected the most, the geometric mean somewhat less, and the median not at all.

Two other measures of spread are the average deviation and median absolute deviation (MAD). These are related to the standard deviation in that they are based on a location measure applied to deviations. Where the standard deviation squares the deviations to make them all positive, the average deviation takes the absolute value of the deviations (just drops any minus signs).

Definition 3.17. The *average deviation* of a sample of values of a variable is the arithmetic average of the absolute values of the deviations about the sample mean.

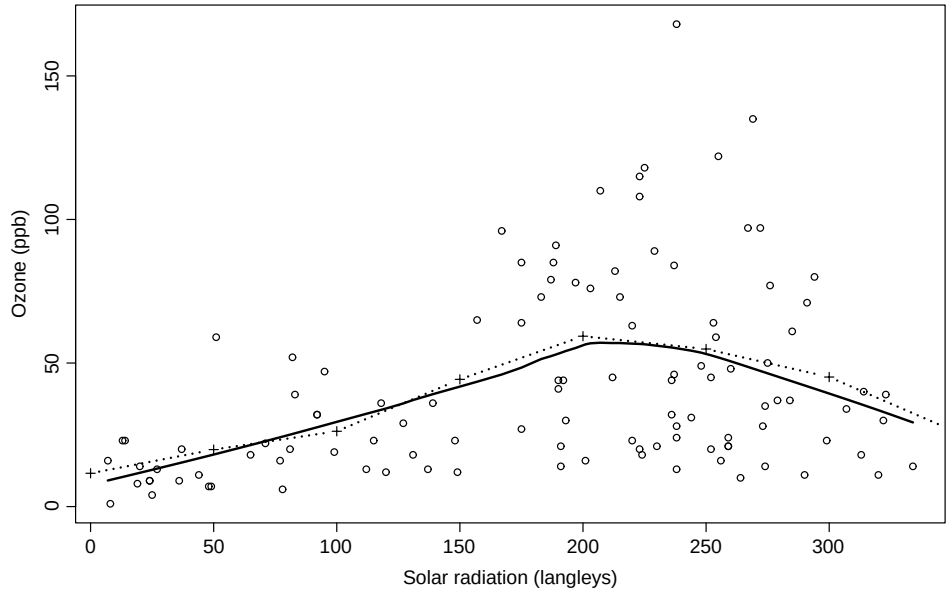


Figure 3.9 Ozone and solar radiation in New York during the summer of 1973, with scatter plot smoothers.

Table 3.10 Number of Boys in Families of Eight Children

Number of Boys per Family of Eight Children	Empirical Frequency (Number of Families)	Empirical Relative Frequency of Families
0	215	0.0040
1	1, 485	0.0277
2	5, 331	0.0993
3	10, 649	0.1984
4	14, 959	0.2787
5	11, 929	0.2222
6	6, 678	0.1244
7	2, 092	0.0390
8	342	0.0064
Total	53, 680	1.0000

Source: Geissler's data reprinted in Fisher [1958].

Using symbols, the average deviation can be written as

$$\text{average deviation} = \frac{\sum |y - \overline{y}|}{n}$$

The median absolute deviation takes the deviations from the median rather than the mean, and takes the median of the absolute values of these deviations.

Definition 3.18. The *median absolute deviation* of a sample of values of a variable is the median of the absolute values of the deviations about the sample median.

Using symbols, the median absolute deviation can be written as

$$\text{MAD} = \text{median}\,\{|y - \text{median}\{y\}|\}$$

The average deviation and the MAD are substantially less affected by extreme values than is the standard deviation.

3.4.5 Which Statistics?

Table 3.11 lists the statistics that have been defined so far, categorized by their use. The question arises: Which statistic should be used for a particular situation? There is no simple answer because the choice depends on the data and the needs of the investigator. Statistics derived from percentiles and those derived from moments can be compared with respect to:

1. *Scientific relevance*. In some cases the scientific question dictates or at least restricts the choice of statistic. Consider a study conducted by the Medicare program being on the effects of exercise on the amount of money expended on medical care. Their interest is in whether exercise affects total costs, or equivalently, whether it affects the arithmetic mean. A researcher studying serum cholesterol levels and the risk of heart disease might be more interested in the proportions of subjects whose cholesterol levels fell in the various categories defined by the National Cholesterol Education Program. In a completely different field, Gould [1996] discusses the absence of batting averages over 0.400 in baseball in recent years and shows that considering a measure of spread rather than a measure of location provides a much clearer explanation

2. *Robustness*. The robustness of a statistic is related to its resistance to being affected by extreme values. In Section 3.4.4 it was shown that the mean—as compared to the median and geometric mean—is most affected by extreme values. The median is said to be more robust. Robustness may be beneficial or harmful, depending on the application: In sampling pollution levels at an industrial site one would be interested in a statistic that was very much affected by extreme values. In comparing cholesterol levels between people on different diets, one might care more about the typical value and not want the results affected by an occasional extreme.

3. *Mathematical simplicity*. The arithmetic mean is more appropriate if the data can be described by a particular mathematical model: the normal or Gaussian frequency distribution, which is the basis for a large part of the theory of statistics. This is described in Chapter 4.

4. *Computational Ease*. Historically, means were easier to compute by hand for moderately large data sets. Concerns such as this vanished with the widespread availability of computers but may reappear with the very large data sets produced by remote sensing or high-throughput genomics. Unfortunately, it is not possible to give general guidelines as to which statistics

Table 3.11 Statistics Defined in This Chapter

Location	Spread
Median	Interquartile range
Percentile	Range
Arithmetic mean	Standard deviation
Geometric mean	Average deviation
Mode	Median absolute deviation

will impose less computational burden. You may need to experiment with your hardware and software if speed or memory limitations become important.

5. *Similarity*. In many samples, the mean and median are not too different. If the empirical frequency distribution of the data is almost symmetrical, the mean and the median tend to be close to each other.

In the absence of specific reasons to chose another statistic, it is suggested that the median and mean be calculated as measures of location and the interquartile range and standard deviation as measures of spread. The other statistics have limited or specialized use. We discuss robustness further in Chapter 8.

NOTES

3.1 Logarithms

A *logarithm* is an exponent on a base. The base is usually 10 or e (2.71828183 . . .). Logarithms with base 10 are called *common logarithms*; logarithms with base e are called *natural logarithms*. To illustrate these concepts, consider

$$100 = 10^2 = (2.71828183\ldots)^{4.605170\ldots} = e^{4.605170\ldots}$$

That is, the logarithm to the base 10 of 100 is 2, usually written

$$\log_{10}(100) = 2$$

and the logarithm of 100 to the base e is

$$\log_e(100) = 4.605170\ldots$$

The three dots indicate that the number is an unending decimal expansion. Unless otherwise stated, logarithms herein will always be natural logarithms. Other bases are sometimes useful—in particular, the base 2. In determining hemagglutination levels, a series of dilutions of serum are set, each dilution being half of the preceding one. The dilution series may be 1 : 1, 1 : 2, 1 : 4, 1 : 8, 1 : 16, 1 : 32, and so on. The logarithm of the dilution factor using the base 2 is then simply

$$\log_2(1) = 0$$
$$\log_2(2) = 1$$
$$\log_2(4) = 2$$
$$\log_2(8) = 3$$
$$\log_2(16) = 4 \qquad \text{etc.}$$

The following properties of logarithms are the only ones needed in this book. For simplicity, we use the base e, but the operations are valid for any base.

1. Multiplication of numbers is equivalent to adding logarithms ($e^a \times e^b = e^{a+b}$).

2. The logarithm of the reciprocal of a number is the negative of the logarithm of the number ($1/e^a = e^{-a}$).

3. Rule 2 is a special case of this rule: Division of numbers is equivalent to subtracting logarithms ($e^a/e^b = e^{a-b}$).

Most pocket calculators permit rapid calculations of logarithms and antilogarithms. Tables are also available. You should verify that you can still use logarithms by working a few problems both ways.

3.2 Stem-and-Leaf Diagrams

An elegant way of describing data by hand consists of *stem-and-leaf diagrams* (a phrase coined by J. W. Tukey [1977]; see his book for some additional innovative methods of describing data). Consider the aflatoxin data from Section 3.4.1. We can tabulate these data according to their first digit (the "stem") as follows:

Stem (tens)	Leaf (units)	Stem (tens)	Leaf (units)
1	6	4	8
2	6 6 2 7 3 8	5	0 2
3	0 6 1 5 7		

For example, the row 3|06157 is a description of the observations 30, 36, 31, 35, and 37. The most frequently occurring category is the 20s. The smallest value is 16, the largest value, 52.

A nice feature of the stem-and-leaf diagram is that all the values can be recovered (but not in the sequence in which the observations were made). Another useful feature is that a quick ordering of the observations can be obtained by use of a stem-and-leaf diagram. Many statistical packages produce stem-and-leaf plots, but there appears to be little point to this, as the advantages over histograms or empirical frequency distributions apply only to hand computation.

3.3 Color and Graphics

With the wide availability of digital projectors and inexpensive color inkjet printers, there are many more opportunities for statisticians to use color to annotate and extend graphs. Differences in color are processed "preattentively" by the brain—they "pop out" visually without a conscious search. It is still important to choose colors wisely, and many of the reference books we list discuss this issue. Colored points and lines can be bright, intense colors, but large areas should use paler, less intense shades. Choosing colors to represent a quantitative variable is quite difficult, and it is advisable to make use of color schemes chosen by experts, such as those at *http://colorbrewer.org*.

Particular attention should be paid to limitations on the available color range. Color graphs may be photocopied in black and white, and might need to remain legible. LCD projectors may have disappointing color saturation. Ideas and emotions associated with a particular color might vary in different societies. Finally, it is important to remember that about 7% of men (and almost no women) cannot distinguish red and green. The Web appendix contains a number of links on color choice for graphics.

3.4 Significant Digits: Rounding and Approximation

In working with numbers that are used to estimate some quantity, we are soon faced with the question of the number of significant digits to carry or to report. A typical rule is to report the mean of a set of observations to one more place and the standard deviation to two more places than the original observation. But this is merely a guideline—which may be wrong. Following DeLury [1958], we can think of two ways in which approximation to the value of a quantity can arise: (1) through arithmetical operations only, or (2) through measurement. If we express the

mean of the three numbers 140, 150, and 152 as 147.3, we have approximated the exact mean, $147\frac{1}{3}$, so that there is *rounding error*. This error arises purely as the result of the arithmetical operation of division. The rounding error can be calculated exactly: $147.\overline{3} - 147.3 = 0.0\overline{3}$.

But this is not the complete story. If the above three observations are the weights of three teenage boys measured to the nearest pound, the true average weight can vary all the way from $146.8\overline{3}$ to $147.8\overline{3}$ pounds; that is, the recorded weights (140, 150, 152) could vary from the three lowest values (139.5, 149.5, 151.5) to the three highest values (140.5, 150.5, 152.5), producing the two averages above. This type of rounding can be called *measurement rounding*. Knowledge of the measurement operation is required to assess the extent of the measurement rounding error: If the three numbers above represent systolic blood pressure readings in mmHg expressed to the nearest *even* number, you can verify that the actual arithmetic mean of these three observations can vary from 146.33 to 148.33, so that even the third "significant" digit could be in error.

Unfortunately, we are not quite done yet with assessing the extent of an approximation. If the weights of the three boys are a sample from populations of boys and the population mean is to be estimated, we will also have to deal with *sampling variability* (a second aspect of the measurement process), and the effect of sampling variability is likely to be much larger than the effect of rounding error and measurement roundings. Assessing the extent of sampling variability is discussed in Chapter 4.

For the present time, we give you the following guidelines: When calculating by hand, minimize the number of rounding errors in intermediate arithmetical calculations. So, for example, instead of calculating

$$\sum (y - \overline{y})^2$$

in the process of calculating the standard deviation, use the equivalent relationship

$$\sum y^2 - \frac{(\sum y)^2}{n}$$

You should also note that we are more likely to use approximations with the arithmetical operations of division and the taking of square roots, less likely with addition, multiplication, and subtraction. So if you can sequence the calculations with division and square root being last, rounding errors due to arithmetical calculations will have been minimized. Note that the guidelines for a computer would be quite different. Computers will keep a large number of digits for all intermediate results, and guidelines for minimizing errors depend on keeping the size of the rounding errors small rather than the number of occasions of rounding.

The rule stated above is reasonable. In Chapter 4 you will learn a better way of assessing the extent of approximation in measuring a quantity of interest.

3.5 Degrees of Freedom

The concept of degrees of freedom appears again and again in this book. To make the concept clear, we need the idea of a linear constraint on a set of numbers; this is illustrated by several examples. Consider the numbers of girls, X, and the number of boys, Y, in a family. (Note that X and Y are variables.) The numbers X and Y are free to vary and we say that there are two degrees of freedom associated with these variables. However, suppose that the total number of children in a family, as in the example, is specified to be precisely 8. Then, given that the number of girls is 3, the number of boys is fixed—namely, $8 - 3 = 5$. Given the constraint on the total number of children, the two variables X and Y are no longer both free to vary, but fixing one determines the other. That is, now there is only one degree of freedom. The constraint can be expressed as

$$X + Y = 8 \qquad \text{so that} \quad Y = 8 - X$$

Constraints of this type are called *linear constraints*.

Table 3.12 Frequency Distribution of Form and Color of 556 Garden Peas

	Variable 1: Form		
Variable 2: Color	Round	Wrinkled	Total
Yellow	315	101	416
Green	108	32	140
Total	423	133	556

Source: Data from Mendel [1911].

A second example is based on Mendel's work in plant propagation. Mendel [1911] reported the results of many genetic experiments. One data set related two variables: form and color. Table 3.12 summarizes these characteristics for 556 garden peas. Let A, B, C, and D be the numbers of peas as follows:

	Form	
Color	**Round**	**Wrinkled**
Yellow	A	B
Green	C	D

For example, A is the number of peas that are round and yellow. Without restrictions, the numbers A, B, C and D can be any nonnegative integers: There are four degrees of freedom. Suppose now that the total number of peas is fixed at 556 (as in Table 3.12). That is, $A + B + C + D = 556$. Now only three of the numbers are free to vary. Suppose, in addition, that the number of yellows peas is fixed at 416. Now only two numbers can vary; for example, fixing A determines B, and fixing C determines D. Finally, if the numbers of round peas is also fixed, only one number in the table can be chosen. If, instead of the last constraint on the number of round peas, the number of green peas had been fixed, two degrees would have remained since the constraints "number of yellow peas fixed" and "number of green peas fixed" are not independent, given that the total number of peas is fixed.

These results can be summarized in the following rule: Given a set of N quantities and $M(\leq N)$ linear, independent constraints, the number of degrees of freedom associated with the N quantities is $N - M$. It is often, but not always, the case that degrees of freedom can be defined in the same way for nonlinear constraints.

Calculations of averages will almost always involve the number of degrees of freedom associated with a statistic rather than its number of components. For example, the quantity $\sum(y - \overline{y})^2$ used in calculating the standard deviation of a sample of, say, n values of a variable Y has $n - 1$ degrees of freedom associated with it because $\sum(y - \overline{y}) = 0$. That is, the sum of the deviations about the mean is zero.

3.6 Moments

Given a sample of observations $y_1, y_2, \ldots, y_n$ of a variable Y, the rth sample moment about zero, m_r^*, is defined to be

$$m_r^* = \frac{\sum y^r}{n} \qquad \text{for } r = 1, 2, 3, \ldots$$

For example, $m_1^* = \sum y^1/n = \sum y/n = \overline{y}$ is just the arithmetic mean.

The rth sample moment about the mean, m_r, is defined to be

$$m_r = \frac{\sum(y - \overline{y})^r}{n} \qquad \text{for } r = 1, 2, 3, \ldots$$

The value of m_1 is zero (see Problem 3.15). It is clear that m_2 and s^2 (the sample variance) are closely connected. For a large number of observations, m_2 will be approximately equal to s^2. One of the earliest statistical procedures (about 1900) was the *method of moments* of Karl Pearson. The method specified that all estimates derived from a sample should be based on sample moments. Some properties of moments are:

- $m_1 = 0$.
- Odd-numbered moments about the mean of symmetric frequency distributions are equal to zero.
- A unimodal frequency distribution is skewed to the right if the mean is greater than the mode; it is skewed to the left if the mean is less than the mode. For distributions skewed to the right, $m_3 > 0$; for distributions skewed to the left, $m_3 < 0$.

The latter property is used to characterize the *skewness of a distribution*, defined by

$$a_3 = \frac{\sum(y - \overline{y})^3}{[\sum(y - \overline{y})^2]^{3/2}} = \frac{m_3}{(m_2)^{3/2}}$$

The division by $(m_2)^{3/2}$ is to standardize the statistic, which now is unitless. Thus, a set of observations expressed in degrees Fahrenheit will have the same value of a_3 when expressed in degrees Celsius. Values of $a_3 > 0$ indicate positive skewness, skewness to the right, whereas values of $a_3 < 0$ indicate negative skewness. Some typical curves and corresponding values for the skewness statistics are illustrated in Figure 3.10. Note that all but the last two frequency distributions are symmetric; the last figure, with skewness $a_3 = -2.71$, is a mirror image of the penultimate figure, with skewness $a_3 = 2.71$.

The fourth moment about the mean is involved in the characterization of the flatness or peakedness of a distribution, labeled *kurtosis* (degree of archedness); a measure of kurtosis is defined by

$$a_4 = \frac{\sum(y - \overline{y})^4}{[\sum(y - \overline{y})^2]^2} = \frac{m_4}{(m_2)^2}$$

Again, as in the case of a_3, the statistic is unitless. The following terms are used to characterize values of a_4.

$a_4 = 3$	*mesokurtic*: the value for a bell-shaped distribution (Gaussian or normal distribution)
$a_4 < 3$	*leptokurtic*: thin or peaked shape (or "light tails")
$a_4 > 3$	*platykurtic*: flat shape (or "heavy tails")

Values of this statistic associated with particular frequency distribution configurations are illustrated in Figure 3.10. The first figure is similar to a bell-shaped curve and has a value $a_4 = 3.03$, very close to 3. Other frequency distributions have values as indicated. It is meaningful to speak of kurtosis only for symmetric distributions.

3.7 Taxonomy of Data

Social scientists have thought hard about types of data. Table 3.13 summarizes a fairly standard taxonomy of data based on the four scales nominal, ordinal, interval, and ratio. This table is to

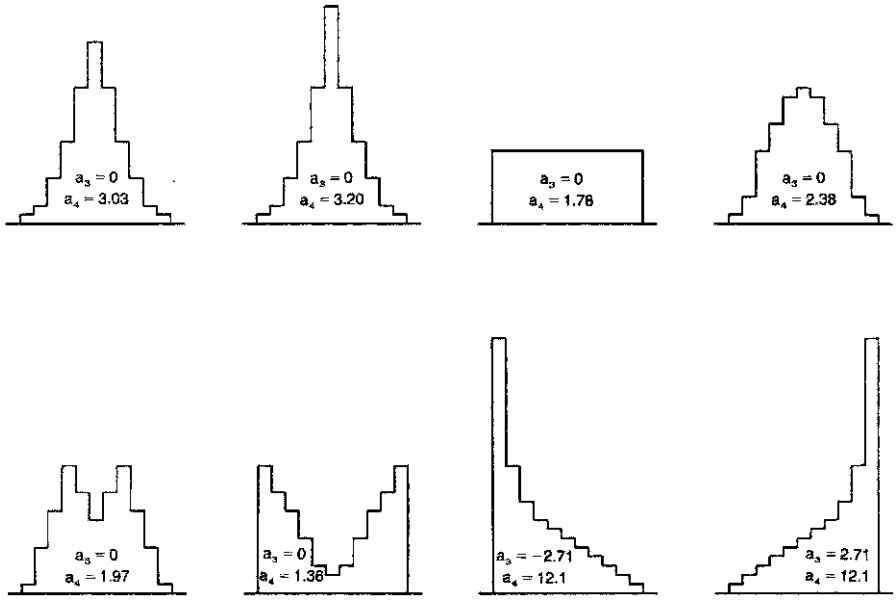


Figure 3.10 Values of skewness (a_3) and kurtosis (a_4) for selected data configurations.

Table 3.13 Standard Taxonomy of Data

Scale	Characteristic Question	Statistic	Statistic to Be Used
Nominal	Do A and B differ?	List of diseases; marital status	Mode
Ordinal	Is A bigger (better) than B?	Quality of teaching (unacceptable/acceptable)	Median
Interval	How much do A and B differ?	Temperatures; dates of birth	Mean
Ratio	How many times is A bigger than B?	Distances; ages; heights	Mean

be used as a guide only. You can be too rigid in applying this scheme (as unfortunately, some academic journals are). Frequently, ordinal data are coded in increasing numerical order and averages are taken. Or, interval and ratio measurements are ranked (i.e., reduced to ordinal status) and averages taken at that point. Even with nominal data, we sometimes calculate averages. For example: coding male $= 0$, female $= 1$ in a class of 100 students, the average is the proportion of females in the class. Most statistical procedures for ordinal data implicitly use a numerical coding scheme, even if this is not made clear to the user. For further discussion, see Luce and Narens [1987], van Belle [2002], and Velleman and Wilkinson [1993].

PROBLEMS

3.1 Characterize the following variables and classify them as qualitative or quantitative. If qualitative, can the variable be ordered? If quantitative, is the variable discrete or continuous? In each case define the values of the variable: (1) race, (2) date of birth, (3) systolic blood pressure, (4) intelligence quotient, (5) Apgar score, (6) white blood count, (7) weight, and (8) quality of medical care.

3.2 For each variable listed in Problem 3.1, define a suitable sample space. For two of the sample spaces so defined, explain how you would draw a sample. What statistics could be used to summarize such a sample?

3.3 Many variables of medical interest are derived from (functions of) several other variables. For example, as a measure of obesity there is the body mass index (BMI), which is given by weight/height^2. Another example is the dose of an anticonvulsant to be administered, usually calculated on the basis of milligram of medicine per kilogram of body weight. What are some assumptions when these types of variables are used? Give two additional examples.

3.4 Every row of 12 observations in Table 3.3 can be summed to form the number of keypunching errors per year of data. Calculate the 13 values for this variable. Make a stem-and-leaf diagram. Calculate the (sample) mean and standard deviation. How do this mean and standard deviation compare with the mean and standard deviation for the number of keypunching errors per line of data?

3.5 The precise specification of the value of a variable is not always easy. Consider the data dealing with keypunching errors in Table 3.3. How is an error defined? A fairly frequent occurrence was the transposition of two digits—for example, a value of "63" might have been entered as "36." Does this represent one or two errors? Sometimes a zero was omitted, changing, for example, 0.0317 to 0.317. Does this represent four errors or one? Consider the list of qualitative variables at the beginning of Section 3.2, and name some problems that you might encounter in defining the values of some of the variables.

3.6 Give three examples of frequency distributions from areas of your own research interest. Be sure to specify (1) what constitutes the sample, (2) the variable of interest, and (3) the frequencies of values or ranges of values of the variables.

3.7 A constant is added to each observation in a set of data (relocation). Describe the effect on the median, lower quartile, range, interquartile range, minimum, mean, variance, and standard deviation. What is the effect on these statistics if each observation is multiplied by a constant (rescaling)? Relocation and rescaling, called linear *transformations*, are frequently used: for example, converting from °C to °F, defined by $°\mathrm{F} = 1.8 \times °\mathrm{C} + 32$. What is the rescaling constant? Give two more examples of rescaling and relocation. An example of nonlinear transformation is going from the radius of a circle to its area: $A = \pi r^2$. Give two more examples of nonlinear transformations.

3.8 Show that the geometric mean is always smaller than the arithmetic mean (unless all the observations are identical). This implies that the mean of the logarithms is not the same as the logarithm of the mean. Is the median of the logarithms equal to the logarithm of the median? What about the interquartile range? How do these results generalize to other nonlinear transformations?

3.9 The data in Table 3.14 deal with the treatment of essential hypertension (*essential* is a technical term meaning that the cause is unknown; a synonym is *idiopathic*). Seventeen patients received treatments *C, A*, and *B*, where C = control period, A = propranolol + phenoxybenzamine, and B = propranolol + phenoxybenzamine + hydrochlorothiazide. Each patient received C first, then either A or B, and finally, B or A. The data consist of the systolic blood pressure in the recumbent position. (Note that in this example blood pressures are not always even-numbered.)

Table 3.14 Treatment Data for Hypertension

	C	*A*	*B*		*C*	*A*	*B*
1	185	148	132	10	180	132	136
2	160	128	120	11	176	140	135
3	190	144	118	12	200	165	144
4	192	158	115	13	188	140	115
5	218	152	148	14	200	140	126
6	200	135	134	15	178	135	140
7	210	150	128	16	180	130	130
8	225	165	140	17	150	122	132
9	190	155	138				

Source: Vlachakis and Mendlowitz [1976].

(a) Construct stem-and-leaf diagrams for each of the three treatments. Can you think of some innovative way of displaying the three diagrams together to highlight the data?

(b) Graph as a single graph the ECDFs for each of treatments *C, A*, and *B*.

(c) Construct box plots for each of treatments *C, A*, and *B*. State your conclusions with respect to the systolic blood pressures associated with the three treatments.

(d) Consider the difference between treatments *A* and *B* for each patient. Construct a box plot for the difference. Compare this result with that of part (b).

(e) Calculate the mean and standard deviation for each of the treatments *C, A*, and *B*.

(f) Consider, again, the difference between treatments *A* and *B* for each patient. Calculate the mean and standard deviation for the difference. Relate the mean to the means obtained in part (d). How many standard deviations is the mean away from zero?

3.10 The New York air quality data used in Figure 3.7 are given in the Web appendix to this chapter. Using these data, draw a simple plot of ozone vs. Solar radiation and compare it to conditioning plots where the subsets are defined by temperature, by wind speed, and by both variables together (i.e., one panel would be high temperature and high wind speed). How does the visual impression depend on the number of panels and the conditioning variables?

3.11 Table 3.15 is a frequency distribution of fasting serum insulin (μU/mL) of males and females in a rural population of Jamaican adults. (Serum insulin levels are expressed as whole numbers, so that "7-" represents the values 7 and 8.) The last frequencies are associated with levels greater than 45. Assume that these represent the levels 45 and 46.

(a) Plot both frequency distributions as histograms.

(b) Plot the relative frequency distributions.

(c) Calculate the ECDF.

(d) Construct box plots for males and females. State your conclusions.

(e) Assume that all the observations are concentrated at the midpoints of the intervals. Calculate the mean and standard deviation for males and females.

(f) The distribution is obviously skewed. Transform the levels for males to logarithms and calculate the mean and standard deviation. The transformation can be carried in at least two ways: (1) consider the observations to be centered at the midpoints,

Table 3.15 Frequency Distribution of Fasting Serum Insulin

Fasting Serum Insulin ($\mu U/mL$)	Males	Females	Fasting Serum Insulin ($\mu U/mL$)	Males	Females
7–	1	3	29–	8	14
9–	9	3	31–	8	11
11–	20	9	33–	4	10
13–	32	21	35–	4	8
15–	32	23	37–	3	7
17–	22	39	39–	1	2
19–	23	39	41–	1	3
21–	19	23	43–	1	1
23–	20	27	≥ 45	6	11
25–	13	23			
27–	8	19	Total	235	296

Source: Data from Florey et al. [1977].

transform the midpoints to logarithms, and group into six to eight intervals; and (2) set up six to eight intervals on the logarithmic scale, transform to the original scale, and estimate by interpolation the number of observations in the interval. What type of mean is the antilogarithm of the logarithmic mean? Compare it with the median and arithmetic mean.

3.12 There has been a long-held belief that births occur more frequently in the "small hours of the morning" than at any other time of day. Sutton [1945] collected the time of birth at the King George V Memorial Hospital, Sydney, for 2654 consecutive births. (*Note:* The total number of observations listed is 2650, not 2654 as stated by Sutton.) The frequency of births by hour in a 24-hour day is listed in Table 3.16.

(a) Sutton states that the data "confirmed the belief ... that more births occur in the small hours of the morning than at any other time in the 24 hours." Develop a graphical display that illustrates this point.

(b) Is there evidence of Sutton's statement: "An interesting point emerging was the relatively small number of births during the meal hours of the staff; this suggested either hastening or holding back of the second stage during meal hours"?

Table 3.16 Frequency of Birth by Hour of Birth

Time	Births	Time	Births	Time	Births
6–7 pm	92	2 am	151	10 am	101
7 pm	102	3 am	110	11 am	107
8 pm	100	4 am	144	12 pm	97
9 pm	101	5–6 am	136	1 pm	93
10 pm	127	6–7 am	117	2 pm	100
11 pm	118	7 am	80	3 pm	93
12 am	97	8 am	125	4 pm	131
1 am	136	9 am	87	5–6 pm	105

(c) The data points in fact represent frequencies of values of a variable that has been divided into intervals. What is the variable?

3.13 At the International Health Exhibition in Britain, in 1884, Francis Galton, a scientist with strong statistical interests, obtained data on the strength of pull. His data for 519 males aged 23 to 26 are listed in Table 3.17. Assume that the smallest and largest categories are spread uniformly over a 10-pound interval.

Table 3.17 Strength of Pull

Pull Strength (lb)	Cases Observed	Pull Strength (lb)	Cases Observed
Under 50	10	Under 90	113
Under 60	42	Under 100	22
Under 70	140	Above 100	24
Under 80	168		
		Total	519

(a) The description of the data is exactly as in Galton [1889]. What are the intervals, assuming that strength of pull is measured to the nearest pound?

(b) Calculate the median and 25th and 75th percentiles.

(c) Graph the ECDF.

(d) Calculate the mean and standard deviation assuming that the observations are centered at the midpoints of the intervals.

(e) Calculate the proportion of observations within one standard deviation of the mean.

3.14 The aflatoxin data cited at the beginning of Section 3.2 were taken from a larger set in the paper by Quesenberry et al. [1976]. The authors state:

> Aflatoxin is a toxic material that can be produced in peanuts by the fungus *Aspergillus flavus*. As a precautionary measure all commercial lots of peanuts in the United States (approximately 20,000 each crop year) are tested for aflatoxin.... Because aflatoxin is often highly concentrated in a small percentage of the kernels, variation among aflatoxin determinations is large.... Estimation of the distribution (of levels) is important. ... About 6200g of raw peanut kernels contaminated with aflatoxin were comminuted (ground up). The ground meal was then divided into 11 subsamples (lots) weighing approximately 560g each. Each subsample was blended with 2800ml methanol-water-hexane solution for two minutes, and the homogenate divided equally among 16 centrifuge bottles. One observation was lost from each of three subsamples leaving eight subsamples with 16 determinations and three subsamples with 15 determinations.

The original data were given to two decimal places; they are shown in Table 3.18 rounded off to the nearest whole number. The data are listed by lot number, with asterisks indicating lost observations.

(a) Make stem-and-leaf diagrams of the data of lots 1, 2, and 10. Make box plots and histograms for these three lots, and discuss differences among these lots with respect to location and spread.

(b) The data are analyzed by means of a MINITAB computer program. The data are entered by columns and the command DESCRIBE is used to give standard

Table 3.18 Aflatoxin Data by Lot Number

1	2	3	4	5	6	7	8	9	10	11
121	95	20	22	30	11	29	34	17	8	53
72	56	20	33	26	19	33	28	18	6	113
118	72	25	23	26	13	37	35	11	7	70
91	59	22	68	36	13	25	33	12	5	100
105	115	25	28	48	12	25	32	25	7	87
151	42	21	27	50	17	36	29	20	7	83
125	99	19	29	16	13	49	32	17	12	83
84	54	24	29	31	18	38	33	9	8	65
138	90	24	52	22	18	29	31	15	9	74
83	92	20	29	27	17	29	32	21	14	112
117	67	12	22	23	16	32	29	17	13	98
91	92	24	29	35	14	40	26	19	11	85
101	100	15	37	52	11	36	37	23	5	82
75	77	15	41	28	15	31	28	17	7	95
137	92	23	24	37	16	32	31	15	4	60
146	66	22	36	*	12	*	32	17	12	*

Table 3.19 MINITAB Analysis of Aflatoxin Data[a]

MTB > desc c1–c11

	N	N*	MEAN	MEDIAN	STDEV	MIN	MAX	Q1	Q3
C1	16	0	109.69	111.00	25.62	72	151	85.75	134.00
C2	16	0	79.25	83.50	20.51	42	115	60.75	94.25
C3	16	0	20.687	21.500	3.860	12	25	19.25	24.00
C4	16	0	33.06	29.00	12.17	22	68	24.75	36.75
C5	15	1	32.47	30.00	10.63	16	52	26.00	37.00
C6	16	0	14.688	14.500	2.651	11	19	12.25	17.00
C7	15	1	33.40	32.00	6.23	25	49	29.00	37.00
C8	16	0	31.375	32.000	2.849	26	37	29.00	33.00
C9	16	0	17.06	17.00	4.19	9	25	15.00	19.75
C10	16	0	8.438	7.500	3.076	4	14	6.25	11.75
C11	15	1	84.00	83.00	17.74	53	113	70.00	98.00

[a]N*, number of missing observations; Q1 and Q3, 25th and 75th percentiles, respectively.

descriptive statistics for each lot. The output from the program (slightly modified) is given in Table 3.19.

(c) Verify that the statistics for lot 1 are correct in the printout.

(d) There is an interesting pattern between the means and their standard deviations. Make a plot of the means vs. standard deviation. Describe the pattern.

(e) One way of describing the pattern between means and standard deviations is to calculate the ratio of the standard deviation to the mean. This ratio is called the *coefficient of variation*. It is usually multiplied by 100 and expressed as the percent coefficient of variation. Calculate the coefficients of variation in percentages for each of the 11 lots, and make a plot of their value with the associated means. Do you see any pattern now? Verify that the average of the coefficients of variation is about 24%. A reasonable number to keep in mind for many biological measurements is that the variability as measured by the standard deviation is about 30% of the mean.

Table 3.20 Plasma Prostaglandin E Levels

Patient Number	Mean Plasma iPGE (pg/mL)	Mean Serum Calcium (ml/dL)
	Patients with Hypercalcemia	
1	500	13.3
2	500	11.2
3	301	13.4
4	272	11.5
5	226	11.4
6	183	11.6
7	183	11.7
8	177	12.1
9	136	12.5
10	118	12.2
11	60	18.0
	Patients without Hypercalcemia	
12	254	10.1
13	172	9.4
14	168	9.3
15	150	8.6
16	148	10.5
17	144	10.3
18	130	10.5
19	121	10.2
20	100	9.7
21	88	9.2

3.15 A paper by Robertson et al. [1976] discusses the level of plasma prostaglandin E (iPGE) in patients with cancer with and without hypercalcemia. The data are given in Table 3.20. Note that the variables are the mean plasma iPGE and mean serum Ca levels—presumably, more than one assay was carried out for each patient's level. The number of such tests for each patient is not indicated, nor is the criterion for the number.

(a) Calculate the mean and standard deviation of plasma iPGE level for patients with hypercalcemia; do the same for patients without hypercalcemia.

(b) Make box plots for plasma iPGE levels for each group. Can you draw any conclusions from these plots? Do they suggest that the two groups differ in plasma iPGE levels?

(c) The article states that normal limits for serum calcium levels are 8.5 to 10.5 mg/dL. It is clear that patients were classified as hypercalcemic if their serum calcium levels exceeded 10.5 mg/dL. Without classifying patients it may be postulated that high plasma iPGE levels tend to be associated with high serum calcium levels. Make a plot of the plasma iPGE and serum calcium levels to determine if there is a suggestion of a pattern relating these two variables.

3.16 Prove or verify the following for the observations $y_1, y_2, \ldots, y_n$.

(a) $\sum 2y = 2\sum y$.

(b) $\sum(y - \overline{y}) = 0$.

(c) By means of an example, show that $\sum y^2 \neq (\sum y)^2$.

(d) If a is a constant, $\sum ay = a \sum y$.
(e) If a is a constant, $\sum (a + y) = na + \sum y$.
(f) $\sum (y/n) = (1/n) \sum y$.
(g) $\sum (a + y)^2 = na^2 + 2a \sum y + \sum y^2$.
(h) $\sum (y - \overline{y})^2 = \sum y^2 - (\sum y)^2/n$.
(i) $\sum (y - \overline{y})^2 = \sum y^2 - n\overline{y}^2$.

3.17 A variable Y is grouped into intervals of width h and represented by the midpoint of the interval. What is the maximum error possible in calculating the mean of all the observations?

3.18 Prove that the two definitions of the geometric mean are equivalent.

3.19 Calculate the average number of boys per family of eight children for the data given in Table 3.10.

3.20 The formula $\overline{Y} = \sum py$ is also valid for observations not arranged in a frequency distribution as follows: If we let $1/N = p$, we get back to the formula $\overline{Y} = \sum py$. Show that this is so for the following four observations: 3, 9, 1, 7.

3.21 Calculate the average systolic blood pressure of native Japanese men using the frequency data of Table 3.6. Verify that the same value is obtained using the relative frequency data of Table 3.7.

3.22 Using the taxonomy of data described in Note 3.6, classify each of the variables in Problem 3.1 according to the scheme described in the note.

REFERENCES

Cleveland, W. S. [1981]. LOWESS: a program for smoothing scatterplots by robust locally weighted regression. *American Statistician*, **35**: 54.

Cleveland, W. S. [1993]. *Visualizing Data*. Hobart Press, Summit, NJ.

Cleveland, W. S. [1994]. *The Elements of Graphing Data*. Hobart Press, Summit, NJ.

DeLury, D. B. [1958]. Computations with approximate numbers. *Mathematics Teacher*, **51**: 521–530. Reprinted in Ku, H. H. (ed.) [1969]. *Precision Measurement and Calibration*. NBS Special Publication 300. U.S. Government Printing Office, Washington, DC.

Fisher, R. A. [1958]. *Statistical Methods for Research Workers*, 13th ed. Oliver & Boyd, London.

Fleming, T. R. and Harrington, D. P. [1991]. *Counting Processes and Survival Analysis*. John Wiley & Sons, New York.

Florey, C. du V., Milner, R. D. G., and Miall, W. E. [1977]. Serum insulin and blood sugar levels in a rural population of Jamaican adults. *Journal of Chronic Diseases*, **30**: 49–60. Used with permission from Pergamon Press, Inc.

Galton, F. [1889]. *Natural Inheritance*. Macmillan, London.

Gould, S. J. [1996]. *Full House: The Spread of Excellence from Plato to Darwin*. Harmony Books, New York.

Graunt, J. [1662]. Natural and political observations mentioned in a following index and made upon the Bills of Mortality. In Newman, J. R. (ed.) [1956]. *The World of Mathematics*, Vol. 3. Simon & Schuster, New York, pp. 1421–1435.

Huff, D. [1993]. *How to Lie with Statistics*. W. W. Norton, New York.

Luce, R. D. and Narens, L. [1987]. Measurement scales on the continuum. *Science*, **236**: 1527–1532.

Mendel, G. [1911]. *Versuche über Pflanzenhybriden*. Wilhelm Engelmann, Leipzig, p. 18.

Moses, L. E. [1987]. Graphical methods in statistical analysis. *Annual Reviews of Public Health*, **8**: 309–353.

Newman, J. R. (ed.) [1956]. *The World of Mathematics*, Vol. 3. Simon & Schuster, New York, pp. 1421–1435.

Quesenberry, P. D., Whitaker, T. B., and Dickens, J. W. [1976]. On testing normality using several samples: an analysis of peanut aflatoxin data. *Biometrics*, **32**: 753–759. With permission of the Biometric Society.

R Foundation for Statistical Computing [2002]. *R, Version 1.7.0*, Air quality data set. *http://cran.r-project.org*.

Robertson, R. P., Baylink, D. J., Metz, S. A., and Cummings, K. B. [1976]. Plasma prostaglandin E in patients with cancer with and without hypercalcemia. *Journal of Clinical Endocrinology and Metabolism*, **43**: 1330–1335.

Schwab, B. [1975]. Delivery of babies and full moon (letter to the editor). *Canadian Medical Association Journal*, **113**: 489, 493.

Sutton, D. H. [1945]. Gestation period. *Medical Journal of Australia*, Vol. I, **32**: 611–613. Used with permission.

Tufte, E. R. [1990]. *Envisioning Information*. Graphics Press, Cheshire, CT.

Tufte, E. R. [1997]. *Visual Explanations*. Graphics Press, Cheshire, CT.

Tufte, E. R. [2001]. *The Visual Display of Quantitative Information*. 2nd ed. Graphics Press, Cheshire, CT.

Tukey, J. W. [1977]. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

van Belle, G. [2002]. *Statistical Rules of Thumb*. Wiley, New York.

Velleman, P. F. and Wilkinson, L. [1993]. Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician* **46**: 193–197.

Vlachakis, N. D. and Mendlowitz, M. [1976]. Alpha- and beta-adrenergic receptor blocking agents combined with a diuretic in the treatment of essential hypertension. *Journal of Clinical Pharmacology*, **16**: 352–360.

Wilkinson, L. [1999]. *The Grammar of Graphics*. Springer, New York.

Winkelstein, W., Jr., Kagan, A., Kato, H., and Sacks, S. T. [1975]. Epidemiological studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: blood pressure distributions. *American Journal of Epidemiology*, **102**: 502–513.

CHAPTER 4

Statistical Inference: Populations and Samples

4.1 INTRODUCTION

Statistical inference has been defined as "the attempt to reach a conclusion concerning all members of a class from observations of only some of them" [Runes, 1959]. In statistics, "all members of a class" form the *population* or *sample space*, and the subset observed forms a *sample*; we discussed this in Sections 3.1 and 3.2. We now discuss the *process* of obtaining a valid sample from a population; specifically, when is it valid to make a statement about a population on the basis of a sample? One of the assumptions in any scientific investigation is that valid inferences can be made—that the results of a study can apply to a larger population. For example, we can assume that a new therapy developed at the Memorial Sloan–Kettering Cancer Center in New York is applicable to cancer patients in Great Britain. You can easily supply additional examples.

In the next section we note which characteristics of a population are of interest and illustrate this with two examples. In Section 4.3 we introduce probability theory as a way by which we can define valid sampling procedures. In Section 4.4 we apply the theory to a well-known statistical model for a population, the normal frequency distribution, which has practical as well as theoretical interest. One reason for the importance of the normal distribution is given in Section 4.5, which discusses the concept of sampling distribution. In the next three sections we discuss inferences about population means and variances on the basis of a single sample.

4.2 POPULATION AND SAMPLE

4.2.1 Definition and Examples

You should review Chapter 3 for the concepts of *variable*, *sample space* or *population*, and *statistic*.

Definition 4.1. A *parameter* is a numerical characteristic of a population.

Analogous to numerical characteristics of a sample (statistics), we will be interested in numerical characteristics of populations (parameters). The population characteristics are usually unknown because the entire population cannot be enumerated or studied. The problem of

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

statistical inference can then be stated as follows: On the basis of a sample from a population, what can be said about the population from which the sample came? In this section we illustrate the four concepts of population and its corresponding parameters, and sample and its corresponding statistics.

Example 4.1. We illustrate those four concepts with an example from Chapter 3, systolic blood pressure for Japanese men, aged 45–69, living in Japan. The "population" can be considered to be the collection of blood pressures of all Japanese men. The blood pressures are assumed to have been taken under standardized conditions. Clearly, Winkelstein et al. [1975] could not possibly measure all Japanese men, but a subset of 2232 eligible men were chosen. This is the sample. A numerical quantity of interest could be the average systolic blood pressure. This average for the population is a *parameter*; the average for the sample is the *statistic*. Since the total population cannot be measured, the parameter value is unknown. The statistic, the average for the sample, can be calculated. You are probably assuming now that the sample average is a good estimate of the population average. You may be correct. Later in this chapter we specify under what conditions this is true, but note for now that all the elements of inference are present.

Example 4.2. Consider this experimental situation. We want to assess the effectiveness of a new special diet for children with phenylketonuria (PKU). One effect of this condition is that untreated children become mentally retarded. The diet is used with a set of PKU children and their IQs are measured when they reach 4 years of age. What is the population? It is hypothetical in this case: all PKU children who could potentially be treated with the new diet. The variable of interest is the IQ associated with each child. The sample is the set of children actually treated. A parameter could be the median IQ of the hypothetical population; a statistic might be the median IQ of the children in the sample. The question to be answered is whether the median IQ of this treated hypothetical population is the same or comparable to that of non-PKU children.

A sampling situation has the following components: A population of measurement is specified, a sample is taken from the population, and measurements are made. A statistic is calculated which—in some way—makes a statement about the corresponding population parameter. Some practical questions that come up are:

1. Is the population defined unambiguously?
2. Is the variable clearly observable?
3. Is the sample "valid"?
4. Is the sample "big enough"?

The first two questions have been discussed in previous chapters. In this chapter we begin to answer the last two.

Conventionally, parameters are indicated by Greek letters and the estimate of the parameter by the corresponding Roman letter. For example, μ is the population mean, and m is the sample mean. Similarly, the population standard deviation will be indicated by σ and the corresponding sample estimate by s.

4.2.2 Estimation and Hypothesis Testing

Two approaches are commonly used in making statements about population parameters: estimation and hypothesis testing. *Estimation*, as the name suggests, attempts to estimate values of parameters. As discussed before, the sample mean is thought to estimate, in some way, the mean of the population from which the sample was drawn. In Example 4.1 the mean of

the blood pressures is considered an estimate of the corresponding population value. *Hypothesis testing* makes inferences about (population) parameters by supposing that they have certain values, and then testing whether the data observed are consistent with the hypothesis. Example 4.2 illustrates this framework: Is the mean IQ of the population of PKU children treated with the special diet the same as that of the population of non-PKU children? We could hypothesize that it is and determine, in some way, whether the data are inconsistent with this hypothesis.

You could argue that in the second example we are also dealing with estimation. If one could estimate the mean IQ of the treated population, the hypothesis could be dealt with. This is quite true. In Section 4.7 we will see that in many instances hypothesis testing and estimation are but two sides of the same coin.

One additional comment about estimation: A distinction is usually made between point estimate and interval estimate. A sample mean is a *point estimate*. An *interval estimate* is a range of values that is reasonably certain to straddle the value of the parameter of interest.

4.3 VALID INFERENCE THROUGH PROBABILITY THEORY

4.3.1 Precise Specification of Our Ignorance

Everyone "knows" that the probability of heads coming up in the toss of a coin is 1/2 and that the probability of a 3 in the toss of a die is 1/6. More subtly, the probability that a randomly selected patient has systolic blood pressure less than the population median is 1/2, although some may claim, after the measurement is made, that it is either 0 or 1—that is, the systolic blood pressure of the patient is either below the median or greater than or equal to the median.

What do we mean by the phrase "the probability of"? Consider one more situation. We toss a thumbtack on a hard, smooth surface such as a table, if the outcome is $\perp$, we call it "up"; if the outcome is $\top$, we call it "down." What is the probability of "up"? It is clear that in this example we do not know, a priori, the probability of "up"—it depends on the physical characteristics of the thumbtack. How would you *estimate* the probability of "up"? Intuitively, you would toss the thumbtack a large number of times and observe the proportion of times the thumbtack landed "up"—and that is the way we define probability. Mathematically, we define the probability of "up" as the relative frequency of the occurrence of "up" as the number of tosses become indefinitely large. This is an illustration of the *relative frequency* concept of probability. Some of its ingredients are: (1) a trial or experiment has a set of specified outcomes; (2) the outcome of one trial does not influence the outcome of another trial; (3) the trials are identical; and (4) the probability of a specified outcome is the limit of its relative frequency of occurrence as the number of trials becomes indefinitely large.

Probabilities provide a link between a population and samples. A *probability* can be thought of as a numerical statement about what we know and do not know: a precise specification of our ignorance [Fisher, 1956]. In the thumbtack-tossing experiment, we know that the relative frequency of occurrences of "up" will approach some number: the probability of "up." What we do not know is what the outcome will be on the next toss. A probability, then, is a characteristic of a population of outcomes. When we say that the probability of a head in a coin toss is 1/2, we are making a statement about a population of tosses. For alternative interpretations of probability, see Note 4.1. On the basis of the relative frequency interpretation of probability, we deduce that probabilities are numbers between zero and 1 (including zero and 1).

The outcome of a trial such as a coin toss will be denoted by a capital letter; for example, H = "coin toss results in head" and T = "coin toss results in tail." Frequently, the letter can be chosen as a mnemonic for the outcome. The probability of an outcome, O, in a trial will be denoted by $P[O]$. Thus, in the coin-tossing experiment, we have $P[H]$ and $P[T]$ for the probabilities of "head" and "tail," respectively.

4.3.2 Working with Probabilities

Outcomes of trials can be categorized by two criteria: statistical independence and mutual exclusiveness.

Definition 4.2. Two outcomes are *statistically independent* if the probability of their joint occurrence is the product of the probabilities of occurrence of each outcome.

Using notation, let C be one outcome and D be another outcome; $P[C]$ is the probability of occurrence of C, and $P[D]$ is the probability of occurrence of D. Then C and D are statistically independent if

$$P[CD] = P[C]P[D]$$

where $[CD]$ means that both C and D occur.

Statistically independent events are the model for events that "have nothing to do with each other." In other words, the occurrence of one event does not change the probability of the other occurring. Later this is explained in more detail.

Models of independent outcomes are the outcomes of successive tosses of a coin, die, or the spinning of a roulette wheel. For example, suppose that the outcomes of two tosses of a coin are statistically independent. Then the probability of two heads, $P[HH]$, by statistical independence is

$$P[HH] = P[H]P[H] = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Similarly,

$$P[HT] = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$P[TH] = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

and

$$P[TT] = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Note that the outcome HT means "head on toss 1 and tail on toss 2."

You may wonder why we refer to coin tossing and dice throws so much. One reason has been given already: These activities form patterns of probabilistic situations. Second, they can be models for many experimental situations. Suppose that we consider the Winkelstein et al. [1975] study dealing with blood pressures of Japanese men. What is the probability that each of two men has a blood pressure less than the median of the population? We can use the coin-toss model: By definition, half of the population has blood pressure less than the median. The populations can then be thought of as a very large collection of trials each of which has two outcomes: less than the median, and greater than or equal to the median. If the selection of two men can be modeled by the coin-tossing experiment, the probability that both men have blood pressures less than the median is $1/2 \times 1/2 = 1/4$. We now formalize this:

Definition 4.3. Outcomes of a series of repetitions of a trial are a *random sample* of outcomes if the probability of their joint occurrence is the product of the probabilities of each occurring separately. If every possible sample of k outcomes has the same probability of occurrence, the sample is called a *simple random sample*. This is the most common type of random sample.

Suppose that we are dealing with the outcomes of trials. We label the outcomes O_k, where the subscript is used to denote the order in the sequence; O_1 is the outcome specified for the first trial, O_2 is the outcome for the second trial, and so on. Then the outcomes form a random sample if

$$P[O_1 O_2 O_3 \cdots O_k] = P[O_1]P[O_2]P[O_3] \cdots P[O_k].$$

The phrase "a random sample" is therefore not so much a statement about the sample as a statement about the method that produced the sample. The randomness of the sample allows us to make valid statements about the population from which it came. It also allows us to quantify what we know and do not know. (See Note 4.6 for another type of random sampling.)

How can we draw a random sample? For the coin tosses and dice throws, this is fairly obvious. But how do we draw a random sample of Japanese men? Theoretically, we could have their names on slips of paper in a very large barrel. The contents are stirred and slips of paper drawn out—a random sample. Clearly, this is not done in practice. In fact, often, a sample is claimed to be random by default: "There is no reason to believe that it is not random." Thus, college students taking part in a experiment are implicitly assumed to be a "random sample of people." Sometimes this is reasonable; as mentioned earlier, cancer patients treated in New York are considered very similar with respect to cancer to cancer patients in California. There is a gradation in the seriousness of nonrandomness of samples: "Red blood cells from healthy adult volunteers" are apt to be similar in many respects the world over (and dissimilar in others); "diets of teenagers," on the other hand, will vary from region to region.

Obtaining a truly random sample is a difficult task that is rarely carried out successfully. A standard criticism of any study is that the sample of data is not a random sample, so that the inference is not valid. Some problems in sampling were discussed in Chapter 2; here we list a few additional problems:

1. The population or sample space is not defined.
2. Part of the population of interest is not available for study.
3. The population is not identifiable or it changes with time.
4. The sampling procedure is faulty, due to limitations in time, money, and effort.
5. Random allocation of members of a group to two or more treatments does not imply that the group itself is necessarily a random sample.

Most of these problems are present in any study, sometimes in an unexpected way. For example, in an experiment involving rats, the animals were "haphazardly" drawn from a cage for assignment to one treatment, and the remaining rats were given another treatment. "Differences" between the treatments were due to the fact that the more agile and larger animals evaded "haphazard" selection and wound up in the second treatment. For some practical ways of drawing random samples, see Note 4.9.

Now we consider probabilities of mutually exclusive events:

Definition 4.4. Two outcomes are *mutually exclusive* if at most one of them can occur at a time; that is, the outcomes do not overlap.

Using notation, let C be one outcome and D another; then it can be shown (using the relative frequency definition) that $P[C \text{ or } D] = P[C] + P[D]$ if the outcomes are mutually exclusive. Here, the connective "or" is used in its inclusive sense, "either/or, or both."

Some examples of mutually exclusive outcomes are H and T on a coin toss; the race of a person for purposes of a study can be defined as "black," "white," or "other," and each subject can belong to only one category; the method of delivery can be either "vaginal" or by means of a "cesarean section."

Example 4.3. We now illustrate outcomes that are not mutually exclusive. Suppose that the Japanese men in the Winkelstein data are categorized by weight: "reasonable weight" or "overweight," and their blood pressures by "normal" or "high." Suppose that we have the following table:

	Blood Pressure		
Weight	**Normal (N)**	**High (H)**	
Reasonable (R)	0.6	0.1	0.7
Overweight (O)	0.2	0.1	0.3
Total	0.8	0.2	1.0

The entries in the table are the probabilities of outcomes for a person selected randomly from the population, so that, for example, 20% of Japanese men are considered overweight and have normal blood pressure. Consider the outcomes "overweight" and "high blood pressure." What is the probability of the outcome [O or H] (overweight, high blood pressure, or both)? This corresponds to the following data in boldface type:

	N	H	
R	0.6	**0.1**	0.7
O	**0.2**	**0.1**	0.3
Total	0.8	0.2	1.0

$$P[O \text{ or } H] = 0.2 + 0.1 + 0.1 = 0.4$$

But $P[O] + P[H] = 0.2 + 0.3 = 0.5$. Hence, O and H are not mutually exclusive. In terms of calculation, we see that we have added in the outcome $P[OH]$ twice:

	N	H	
R		0.1	
O	0.2	0.1	0.3
Total		0.2	

The correct value is obtained if we subtract $P[OH]$ as follows:

$$\begin{aligned} P[O \text{ or } H] &= P[O] + P[H] - P[OH] \\ &= 0.3 + 0.2 - 0.1 \\ &= 0.4 \end{aligned}$$

This example is an illustration of the addition rule of probabilities.

Definition 4.5. By the *addition rule*, for any two outcomes, the probability of occurrence of either outcome or both is the sum of the probabilities of each occurring minus the probability of their joint occurrence.

Using notation, for any two outcomes C and D,

$$P[C \text{ or } D] = P[C] + [D] - P[CD]$$

Two outcomes, C and D, are mutually exclusive if they cannot occur together. In this case, $P[CD] = 0$ and $P[C \text{ or } D] = P[C] + P[D]$, as stated previously.

We conclude this section by briefly discussing dependent outcomes. The outcomes O and H in Example 4.3 were not mutually exclusive. Were they independent? By Definition 4.2, O and H are statistically independent if $P[OH] = P[O]P[H]$.

From the table, we get $P[OH] = 0.1$, $P[O] = 0.3$, and $P[H] = 0.2$, so that

$$0.1 \neq (0.3)(0.2)$$

Of subjects with reasonable weight, only 1 in 7 has high blood pressure, but among overweight persons, 1 in 3 has high blood pressure. Thus, the probability of high blood pressure in overweight subjects is greater than the probability of high blood pressure in subjects of normal weight. The reverse statement can also be made: 2 of 8 persons with normal blood pressure are overweight; 1 of 2 persons with high blood pressure is overweight.

The statement "of subjects with reasonable weight, only 1 in 7 has high blood pressure" can be stated as a probability: "The probability that a person with reasonable weight has high blood pressure is 1/7." Formally, this is written as

$$P[H|R] = \frac{1}{7}$$

or P[high blood pressure *given* a reasonable weight] $= 1/7$. The probability $P[H|R]$ is called a *conditional* probability. You can verify that $P[H|R] = P[HR]/P[R]$.

Definition 4.6. For any two outcomes C and D, the *conditional probability* of the occurrence of C *given* the occurrence of D, $P[C|D]$, is given by

$$P[C|D] = \frac{P[CD]}{P[D]}$$

For completeness we now state the multiplication rule of probability (which is discussed in more detail in Chapter 6).

Definition 4.7. By the *multiplication rule*, for any two outcomes C and D, the probability of the joint occurrence of C and D, $P[CD]$, is given by

$$P[CD] = P[C]P[D|C]$$

or equivalently,

$$P[CD] = P[D]P[C|D]$$

Example 4.3. [continued] What is the probability that a randomly selected person is overweight and has high blood pressure? In our notation we want $P[OH]$. By the multiplication rule, this probability is

$$P[OH] = P[O]P[H|O]$$

Using Definition 4.6 gives us

$$P[H|O] = \frac{P[OH]}{P[O]} = \frac{0.1}{0.3} = \frac{1}{3}$$

so that

$$P[OH] = 0.3\left(\frac{1}{3}\right) = 0.1$$

Alternatively, we could have calculated $P[OH]$ by

$$P[OH] = P[H]P[O|H]$$

which becomes

$$P[OH] = 0.2\left(\frac{0.1}{0.2}\right) = 0.1$$

We can also state the criterion for statistical independence in terms of conditional probabilities. From Definition 4.2, two outcomes C and D are statistically independent if $P[CD] = P[C]P[D]$ (i.e., the probability of the joint occurrence of C and D is the product of the probability of C and the probability of D). The multiplication rule states that for *any* two outcomes C and D,

$$P[CD] = P[C]P[D|C]$$

Under independence,

$$P[CD] = P[C]P[D]$$

Combining the two, we see that C and D are independent if (and only if) $P[D|C] = P[D]$. In other words, the probability of occurrence of D is not altered by the occurrence of C. This has intuitive appeal.

When do we use the addition rule; when the multiplication rule? Use the addition rule to calculate the probability that either one or both events occur. Use the multiplication rule to calculate the probability of the joint occurrence of two events.

4.3.3 Random Variables and Distributions

Basic to the field of statistics is the concept of a random variable:

Definition 4.8. A *random variable* is a variable associated with a random sample.

The only difference between a *variable* defined in Chapter 3 and a *random variable* is the process that generates the value of the variable. If this process is random, we speak of a random variable. All the examples of variables in Chapter 3 can be interpreted in terms of random variables if the samples are random samples. The empirical relative frequency of occurrence of a value of the variable becomes an estimate of the probability of occurrence of that value. For example, the relative frequencies of the values of the variable "number of boys in families with eight children" in Table 3.12 become estimates of the probabilities of occurrence of these values.

The distinction between discrete and continuous variables carries over to random variables. Also, as with variables, we denote the label of a random variable by capital letters (say $X, Y, V, \ldots$) and a value of the random variable by the corresponding lowercase letter ($x, y, v, \ldots$).

We are interested in describing the probabilities with which values of a random variable occur. For discrete random variables, this is straightforward. For example, let Y be the outcome of the toss of a die. Then Y can take on the values 1, 2, 3, 4, 5, 6, and we write

$$P[Y = 1] = \frac{1}{6}, \quad P[Y = 2] = \frac{1}{6}, \ldots, \quad P[Y = 6] = \frac{1}{6}$$

This leads to the following definition:

Definition 4.9. A *probability function* is a function that for each possible value of a discrete random variable takes on the probability of that value occurring. The function is usually presented as a listing of the values with the probabilities of occurrence of the values.

Consider again the data of Table 3.12, the number of boys in families with eight children. The observed empirical relative frequencies can be considered estimates of probabilities if the 53,680 families are a random sample. The probability distribution is then estimated as shown in Table 4.1. The estimated probability of observing precisely two boys in a family of eight children is 0.0993 or, approximately, 1 in 10. Since the sample is very large, we will treat—in this discussion—the estimated probabilities as if they were the actual probabilities. If Y represents the number of boys in a family with eight children, we write

$$P[Y = 2] = 0.0993$$

What is the probability of two boys or fewer? This can be expressed as

$$P[Y \leq 2] = P[Y = 2 \text{ or } Y = 1 \text{ or } Y = 0]$$

Since these are mutually exclusive outcomes,

$$\begin{aligned} P[Y \leq 2] &= P[Y = 2] + P[Y = 1] + P[Y = 0] \\ &= 0.0993 + 0.0277 + 0.0040 \\ &= 0.1310 \end{aligned}$$

Approximately 13% of families with eight children will have two or fewer boys. A probability function can be represented graphically by a plot of the values of the variable against the probability of the value. The probability function for the Geissler data is presented in Figure 4.1.

How can we describe probabilities associated with continuous random variables? Somewhat paradoxically, the probability of a specified value for a continuous random variable is zero! For example, the probability of finding anyone with height 63.141592654 inches—and not 63.141592653 inches—is virtually zero. If we were to continue the decimal expansion, the probability becomes smaller yet. But we do find people with height, say, 63 inches. When we write 63 inches, however, we do not mean 63.000... inches (and we are almost certain not to find anybody with that height), but we have in mind *an interval* of values of height, anyone with height between 62.500... and 63.500... inches. We could then divide the values of the continuous random variable into intervals, treat the midpoints of the intervals as the values of a discrete variable, and list the probabilities associated with these values. Table 3.7 illustrates this approach with the division of the systolic blood pressure of Japanese men into discrete intervals.

We start with the histogram and the relative frequencies associated with the intervals of values in the histogram. The area under the "curve" is equal to 1 if the width of each interval

Table 4.1 Number of Boys in Eight-Child Families

Number of Boys	Probability	Number of Boys	Probability
0	0.0040	6	0.1244
1	0.0277	7	0.0390
2	0.0993	8	0.0064
3	0.1984		
4	0.2787		
5	0.2222	Total	1.0000

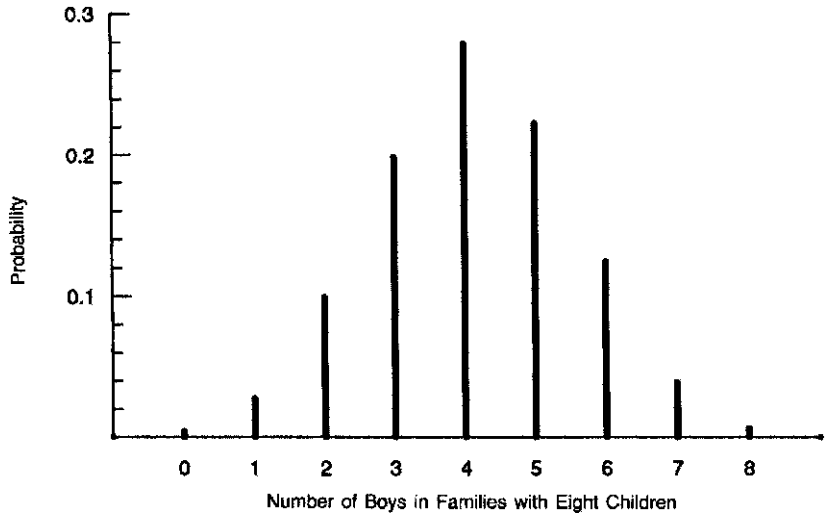


Figure 4.1 Probability function of the random variable "number of boys in families with eight children." (Geissler's date; reprinted in Fisher [1958]; see Table 3.10.)

is 1; or if we normalize (i.e., multiply by a constant so that the area is equal to 1). Suppose now that the interval widths are made smaller and smaller, and simultaneously, the number of cases increased. Normalize so that the area under the curve remains equal to 1; then the curve is assumed to take on a smooth shape. Such shapes are called *probability density functions* or, more briefly, *densities*:

Definition 4.10. *A probability density function* is a curve that specifies, by means of the area under the curve over an interval, the probability that a continuous random variable falls within the interval. The total area under the curve is 1.

Some simple densities are illustrated in Figure 4.2. Figure 4.2(a) and (b) represent uniform densities on the intervals $(-1, 1)$ and $(0, 1)$, respectively. Figure 4.2(c) illustrates a triangular

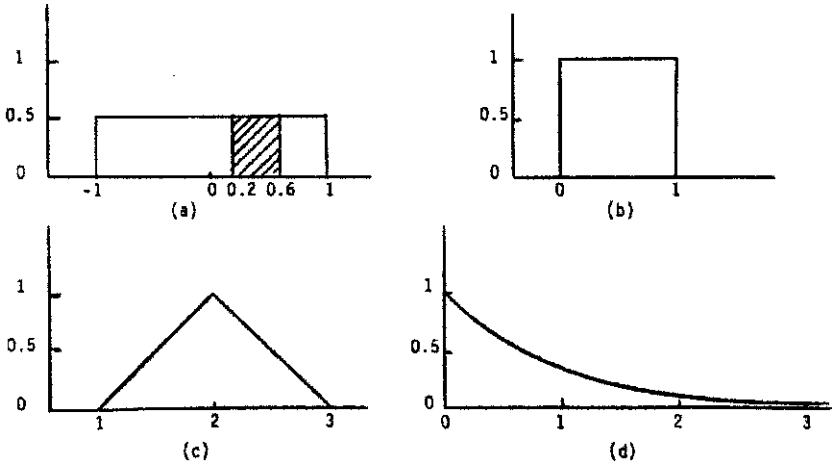


Figure 4.2 Examples of probability density functions. In each case, the area under the curve is equal to 1.

density, and Figure 4.2(d) an exponential density. The latter curve is defined over the entire positive axis. (It requires calculus to show that the area under this curve is 1.) The probability that a continuous random variable takes on a value in a specified interval is equal to the area over the interval. For example, the probability that the random variable in Figure 4.2(a) falls in the interval 0.2–0.6 is equal to the area over the interval. This is, $(0.6 - 0.2)(0.5) = 0.20$, so that we expect 20% of values of this random variable to fall in this interval. One of the most important probability density function is the normal distribution; it is discussed in detail in Section 4.4.

How can we talk about a random sample of observations of a continuous variable? The simplest way is to consider the drawing of an observation as a trial and the probability of observing an arbitrary (but specified) value or smaller of the random variable. Definition 4.3 can then be applied.

Before turning to the normal distribution, we introduce the concept of averages of random variables. In Section 3.4.2, we discussed the average of a discrete variable based on the empirical relative frequency distribution. The average of a discrete variable Y with values $y_1, y_2, \ldots, y_k$ occurring with relative frequencies $p_1, p_2, \ldots, p_k$, respectively, was shown to be

$$\overline{y} = \sum py$$

(We omit the subscripts since it is clear that we are summing over all the values.) Now, if Y is a *random* variable and $p_1, p_2, \ldots, p_k$ are the *probabilities* of occurrence of the values $y_1, y_2, \ldots, y_k$, we give the quantity $\sum py$ a special name:

Definition 4.11. The *expected value of a discrete random variable* Y, denoted by $E(Y)$, is

$$E(Y) = \sum py$$

where $p_1, \ldots, p_k$ are the probabilities of occurrence of the k possible values $y_1, \ldots, y_k$ of Y. The quantity $E(Y)$ is usually denoted by μ.

To calculate the expected value for the data of Table 3.12, the number of boys in families with eight children, we proceed as follows. Let $p_1, p_2, \ldots, p_k$ represent the probabilities $P[Y = 0], P[Y = 1], \ldots, P[Y = 8]$. Then the expected value is

$$\begin{aligned} E(Y) &= p_0 \times 0 + p_1 \times 1 + \cdots + p_8 \times 8 \\ &= (0.0040)(0) + (0.0277)(1) + (0.0993)(2) + \cdots + (0.0064)(8) \\ &= 4.1179 \\ &= 4.12 \text{ boys} \end{aligned}$$

This leads to the statement: "A family with eight children will have an average of 4.12 boys."

Corresponding to the sample variance, s^2, is the variance associated with a discrete random variable:

Definition 4.12. The *variance of a discrete random variable* Y is

$$E(Y - \mu)^2 = \sum p(y - \mu)^2$$

where $p_1, \ldots, p_k$ are the probabilities of occurrence of the k possible values $y_1, \ldots, y_k$ of Y.

The quantity $E(Y-\mu)^2$ is usually denoted by σ^2, where σ is the Greek lowercase letter *sigma*. For the example above, we calculate

$$\sigma^2 = (0.0040)(0-4.1179)^2 + (0.0277)(1-4.1179)^2 + \cdots + (0.0064)(1-4.1179)^2$$
$$= 2.0666$$

Several comments about $E(Y-\mu)^2$ can be made:

1. Computationally, it is equivalent to calculating the sample variance using a divisor of n rather than $n-1$, and probabilities rather than relative frequencies.
2. The square root of $\sigma^2(\sigma)$ is called the (population) *standard deviation* of the random variable.
3. It can be shown that $\sum p(y-\mu)^2 = \sum py^2 - \mu^2$. The quantity $\sum py^2$ is called the *second moment about the origin* and can be defined as the average value of the squares of Y or the expected value of Y^2. This can then be written as $E(Y^2)$, so that $E(Y-\mu)^2 = E(Y^2) - E^2(Y) = E(Y^2) - \mu^2$. See Note 4.9 for further development of the algebra of expectations.

What about the mean and variance of a continuous random variable? As before, we could divide the range of the continuous random variable into a number of intervals, calculate the associated probabilities of the variable, assume that the values are concentrated at the midpoints of the intervals, and proceed with Definitions 4.8 and 4.9. This is precisely what is done with one additional step: The intervals are made narrower and narrower. The mean is then the limit of a sequence of means calculated in this way, and similarly the variance. In these few sentences we have crudely summarized the mathematical process known as *integration*. We will only state the results of such processes but will not actually derive or demonstrate them. For the densities presented in Figure 4.2, the following results can be stated:

Figure	Name	μ	σ^2
4.2(a)	Uniform on $(-1, 1)$	0	1/3
4.2(b)	Uniform on (0, 1)	1/2	1/12
4.2(c)	Triangular on (1, 3)	2	1/6
4.2(d)	Exponential	1	1

The first three densities in Figure 4.2 are examples of *symmetric* densities. A symmetric density always has equality of mean and median. The exponential density is not symmetric; it is "skewed to the right." Such a density has a mean that is larger than the median; for Figure 4.2(d), the median is about 0.69.

It is useful at times to state the functional form for the density. If Y is the random variable, then for a value $Y = y$, the height of the density is given by $f(y)$. The densities in Figure 4.2 have the functional forms shown in Table 4.2. The letter e in $f(y) = e^{-y}$ is the base of the natural logarithms. The symbol ∞ stands for positive infinity.

4.4 NORMAL DISTRIBUTIONS

Statistically, a *population* is the set of all possible values of a variable; random selection of objects of the population makes the variable a random variable and the population is described completely (*modeled*) if the probability function or the probability density function is specified.

Table 4.2 Densities in Figure 4.2

Figure	Name of Density	Function	Range of Y
4.2(a)	Uniform on $(-1, 1)$	$f(y) = 0.5$	$(-1, 1)$
		$f(y) = 0$	elsewhere
4.2(b)	Uniform on (0, 1)	$f(y) = 1$	(0, 1)
		$f(y) = 0$	elsewhere
4.2(c)	Triangular on (1,3)	$f(y) = y - 1$	(1, 2)
		$f(y) = 3 - y$	(2, 3)
		$f(y) = 0$	elsewhere
4.2(d)	Exponential	$f(y) = e^{-y}$	$(0, \infty)$
		$f(y) = 0$	elsewhere

A statistical challenge is to find models of populations that use a few parameters (say, two or three), yet have wide applicability to real data. The *normal* or *Gaussian distribution* is one such statistical model.

The term *Gaussian* refers to Carl Friedrich Gauss, who developed and applied this model. The term *normal* appears to have been coined by Francis Galton. It is important to remember that there is nothing normal or abnormal about the normal distribution! A given data set may or may not be modeled adequately by the normal distribution. However, the normal distribution often proves to be a satisfactory model for data sets. The first and most important reason is that it "works," as will be indicated below. Second, there is a mathematical reason suggesting that a Gaussian distribution may adequately represent many data sets—the famous central limit theorem discussed in Section 4.5. Finally, there is a matter of practicality. The statistical theory and methods associated with the normal distribution work in a nice fashion and have many desirable mathematical properties. But no matter how convenient the theory, the assumptions that a data set is modeled adequately by a normal curve should be verified when looking at a particular data set. One such method is presented in Section 4.4.3.

4.4.1 Examples of Data That Might Be Modeled by a Normal Distribution

The first example is taken from a paper by Golubjatnikov et al. [1972]. Figure 4.3 shows serum cholesterol levels of Mexican and Wisconsin children in two different age groups. In each case

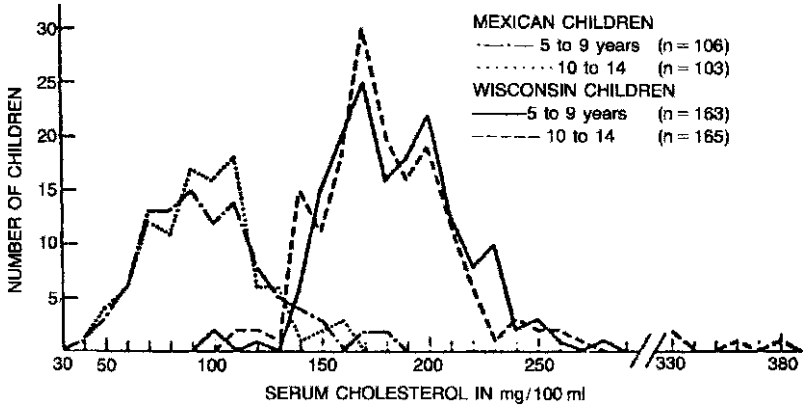


Figure 4.3 Distribution of serum cholesterol levels in Mexican and Wisconsin school children. (Data from Golubjatnikov et al. [1972].)

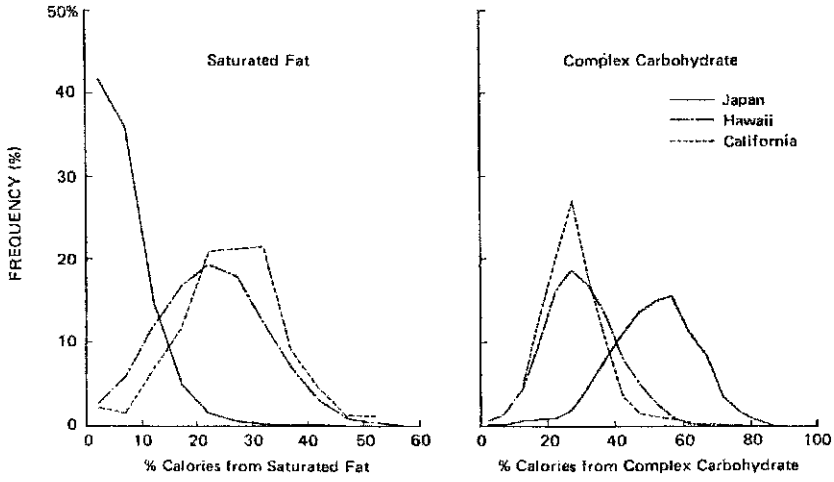


Figure 4.4 Frequency distribution of dietary saturated fat and dietary complex carbohydrate intake. (Data from Kato et al. [1973].)

there is considerable fluctuation in the graphs, probably due to the small numbers of people considered. However, it might be possible to model such data with a normal curve. Note that there seem to be possibly too many values in the right tail to model the data by a normal curve since normal curves are symmetric about their center point.

Figure 4.4 deals with epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii, and California. The curves present the frequency distribution of the percentage of calories from saturated fat and from complex carbohydrate in the three groups of men. Such percentages necessarily lie on the interval from 0 to 100. For the Hawaiian and Californian men with regard to saturated fat, the bell-shaped curve might be a reasonable model. Note, however, that for Japanese men, with a very low percentage of the diet from saturated fat, a bell-shaped curve would obviously be inappropriate.

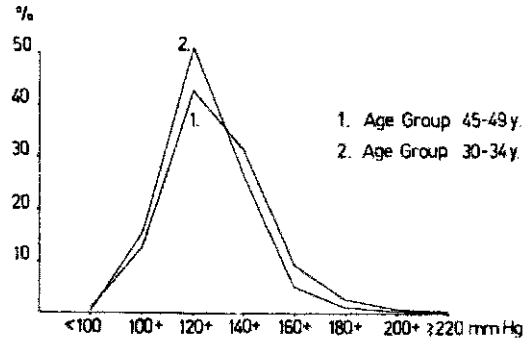
A third example from Kesteloot and van Houte [1973] examines blood pressure measurements on 42,000 members of the Belgian army and territorial police. Figure 4.5 gives two different age groups. Again, particularly in the graphs of the diastolic pressures, it appears that a bell-shaped curve might not be a bad model.

Another example of data that do not appear to be modeled very well by a symmetric bell-shaped curve is from a paper by Hagerup et al. [1972] dealing with serum cholesterol, serum triglyceride, and ABO blood groups in a population of 50-year-old Danish men and women. Figure 4.6 shows the distribution of serum triglycerides. There is a notable asymmetry to the distribution, there being too many values to the right of the peak of the distribution as opposed to the left.

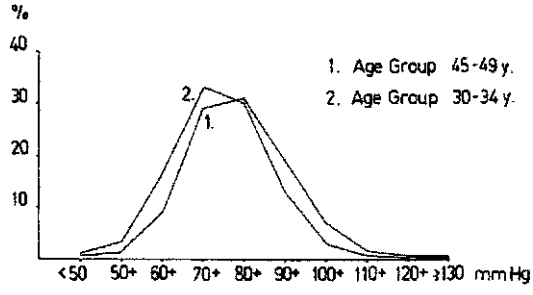
A final example of data that are not normally distributed are the 2-hour plasma glucose levels (mg per 100 mL) in Pima Indians. The data in Figure 4.7 are the plasma glucose levels for male Pima Indians for each decade of age. The data become clearly bimodal (two modes) with increasing decade of age. Note also that the overall curve is shifting to the right with increasing decade: The first mode shifts from approximately 100 mg per 100 mL in the 5- to 14-year decade to about 170 mg per 100 mL in the 65- to 74-year decade.

4.4.2 Calculating Areas under the Normal Curve

A normal distribution is specified completely by its mean, μ, and standard deviation, σ. Figure 4.8 illustrates some normal distributions with specific means and standard deviations. Note that two



Distribution of SBP according to age. (Distribution is slightly skewed towards the higher values.)



Distribution of DBP according to age. (Distribution is slightly skewed towards the higher values.)

Figure 4.5 Distributions of systolic and diastolic blood pressures according to age. (Data from Kesteloot and van Houte [1973].)

normal distributions with the same standard deviation but different means have the same shape and are merely shifted; similarly, two normal distributions with the same means but different standard deviations are centered in the same place but have different shapes. Consequently, μ is called a *location parameter* and σ a *shape parameter*.

The *standard deviation* is the distance from the mean to the point of inflection of the curve. This is the point where a tangent to the curve switches from being over the curve to under the curve.

As with any density, the probability that a normally distributed random variable takes on a value in a specified interval is equal to the area over the interval. So we need to be able to calculate these areas in order to know the desired probabilities. Unfortunately, there is no simple algebraic formula that gives these areas, so tables must be used (see Note 4.15). Fortunately, we need only one table. For any normal distribution, we can calculate areas under its curve using a table for a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ by expressing the variable in the number of standard deviations from the mean. Using algebraic notation, we get the following:

Definition 4.13. For a random variable Y with mean μ and standard deviation σ, the associated *standard score, Z*, is

$$Z = \frac{Y - \mu}{\sigma}$$

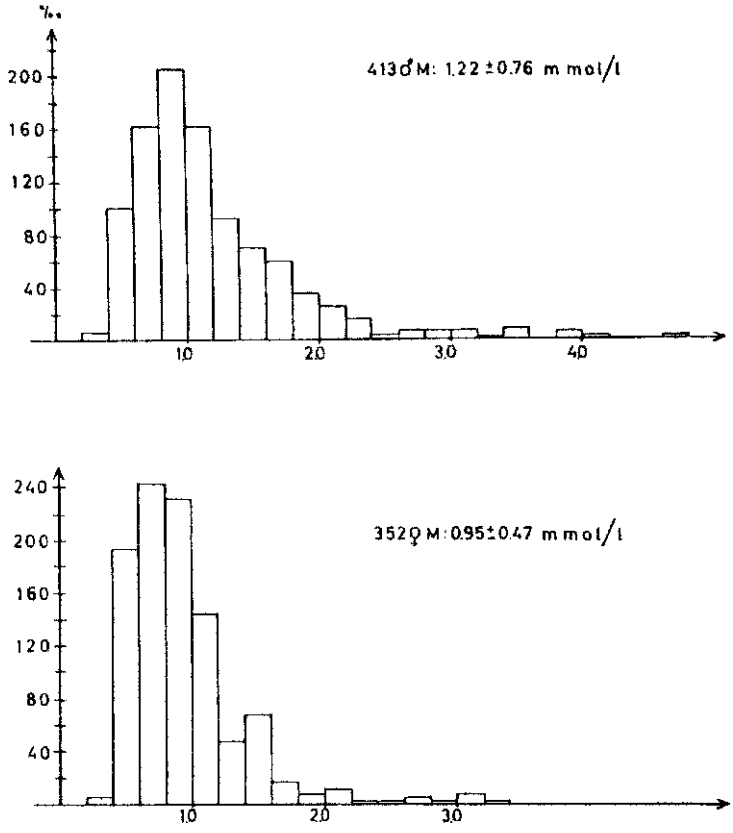


Figure 4.6 Serum triglycerides: 50-year survey in Glostrup. Fasting blood samples were drawn for determination of serum triglyceride by the method of Laurell. (Data from Hagerup et al. [1972].)

Given values for μ and σ, we can go from the "Y scale" to the "Z scale," and vice versa. Algebraically, we can solve for Y and get $Y = \mu + \sigma Z$. This is also the procedure that is used to get from degrees Celsius (°C) to degrees Fahrenheit (°F). The relationship is

$$^{\circ}\mathrm{C} = \frac{^{\circ}\mathrm{F} - 32}{1.8}$$

Similarly,

$$^{\circ}\mathrm{F} = 32 + 1.8 \times {}^{\circ}\mathrm{C}$$

Definition 4.14. A *standard normal distribution* is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

Table A.1 in the Appendix gives standard normal probabilities. The table lists the area to the left of the stated value of the standard normal deviate under the columns headed "cum. dist." For example, the area to the left of $Z = 0.10$ is 0.5398, as shown in Figure 4.9.

In words, 53.98% of normally distributed observations have values less than 0.10 standard deviation above the mean. We use the notation $P[Z \leq 0.10] = 0.5398$, or in general, $P[Z \leq z]$. To indicate a value of Z associated with a specified area, p, to its left, we will use a subscript on the value Z_p. For example, $P[Z \leq z_{0.1}] = 0.10$; that is, we want that value of Z such that

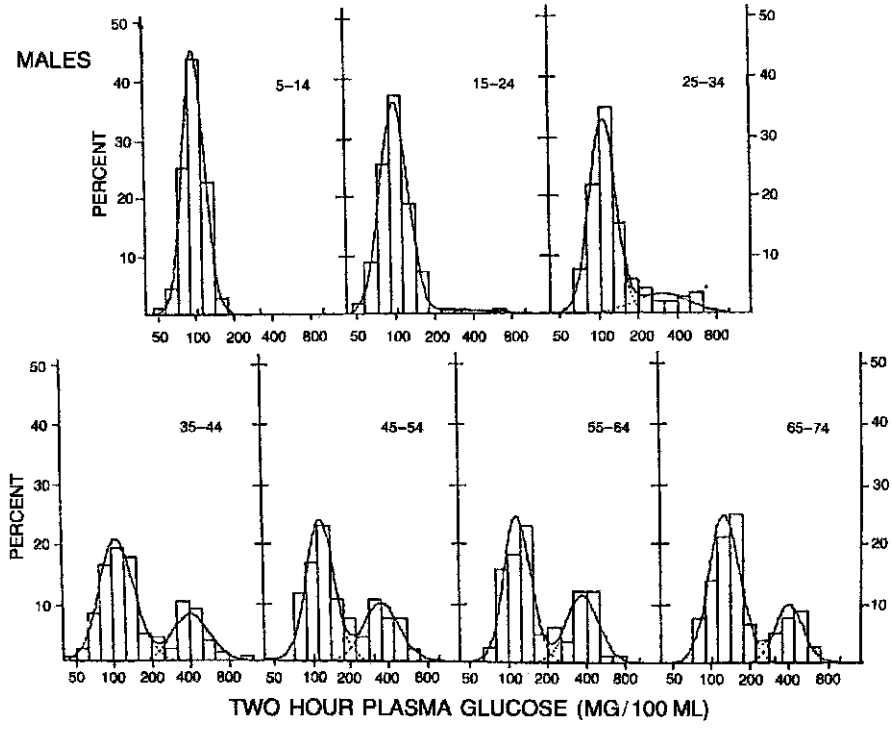


Figure 4.7 Distribution of 2-hour plasma glucose levels (mg/100 mL) in male Pima Indians by decade. (Data from Rushforth et al. [1971].)

0.1 of the area is to its left (call it $z_{0.1}$), or equivalently, such that a proportion 0.1 of Z values are less than or equal to $z_{0.1}$. By symmetry, we note that $z_{1-p} = -z_p$.

Since the total area under the curve is 1, we can get areas in the right-hand tail by subtraction. Formally,

$$P[Z > z] = 1 - P[Z \leq z]$$

In terms of the example above, $P[Z > 0.10] = 1 - 0.5398 = 0.4602$. By symmetry, areas to the left of $Z = 0$ can also be obtained. For example, $P[Z \leq -0.10] = P[Z > 0.10] = 0.4602$. These values are indicated in Figure 4.10.

We now illustrate use of the standard normal table with two word problems. When calculating areas under the normal curve, you will find it helpful to draw a rough normal curve and shade in the required area.

Example 4.4. Suppose that IQ is normally distributed with mean $\mu = 100$ and standard deviation $\sigma = 15$. A person with IQ > 115 has a *high IQ*. What proportion of the population has high IQs? The area required is shown in Figure 4.11. It is clear that IQ = 115 is one standard deviation above the mean, so the statement $P[IQ > 115]$ is equivalent to $P[Z > 1]$. This can be obtained from Table A.1 using the relationship $P[Z > 1] = 1 - P[Z \leq 1] = 1 - 0.8413 = 0.1587$. Thus, 15.87% of the population has a high IQ. By the same token, if an IQ below 85 is labeled *low IQ*, 15.87% of the population has a low IQ.

Example 4.5. Consider the serum cholesterol levels of Wisconsin children as pictured in Figure 4.3. Suppose that the population mean is 175 mg per 100 mL and the population standard

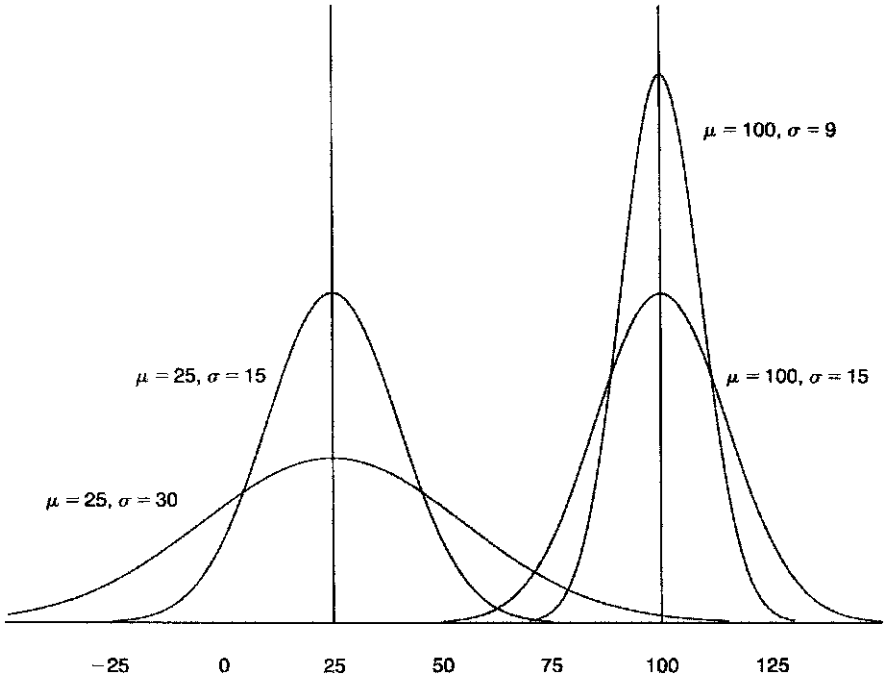


Figure 4.8 Examples of normal distributions.

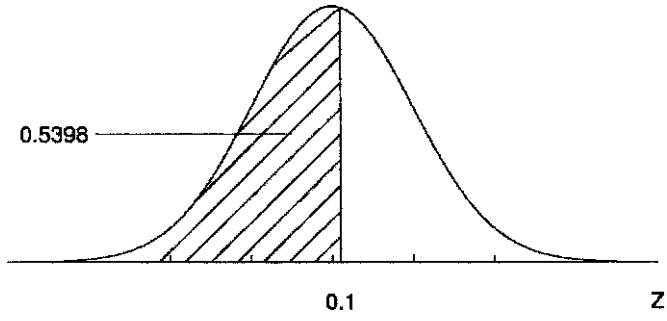


Figure 4.9 Area to the left of $Z = 0.10$ is 0.5398.

deviation is 30 mg per 100 mL. Suppose that a "normal cholesterol value" is taken to be a value within two standard deviations of the mean. What are the *normal limits*, and what proportion of Wisconsin children will be within normal limits?

We want the area within ± 2 standard deviations of the mean (Figure 4.12). This can be expressed as $P[-2 \leq Z \leq +2]$. By symmetry and the property that the area under the normal curve is 1.0, we can express this as

$$P[-2 \leq Z \leq 2] = 1 - 2P[Z > 2]$$

(You should sketch this situation, to convince yourself.) From Table A.1, $P[Z \leq 2] = 0.9772$, so that $P[Z > 2] = 1 - 0.9772 = 0.0228$. (Note that this value is computed for you in the

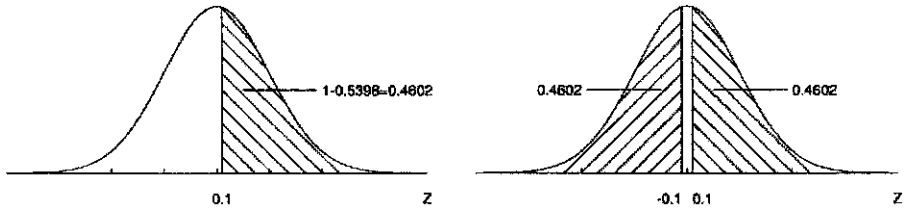


Figure 4.10 $P[Z \leq -0.10] = P[Z > 0.10] = 0.4602$.

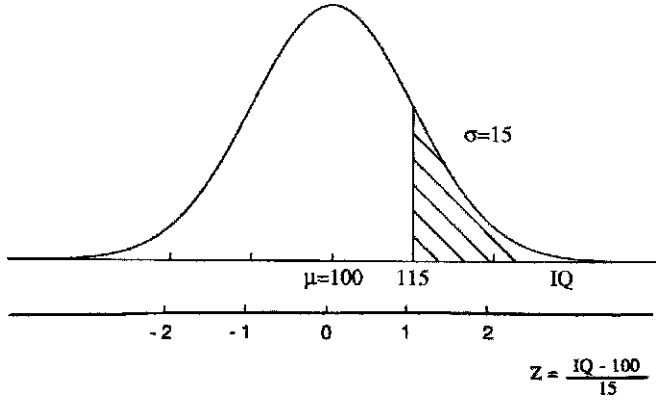


Figure 4.11 Proportion of the population with high IQs.

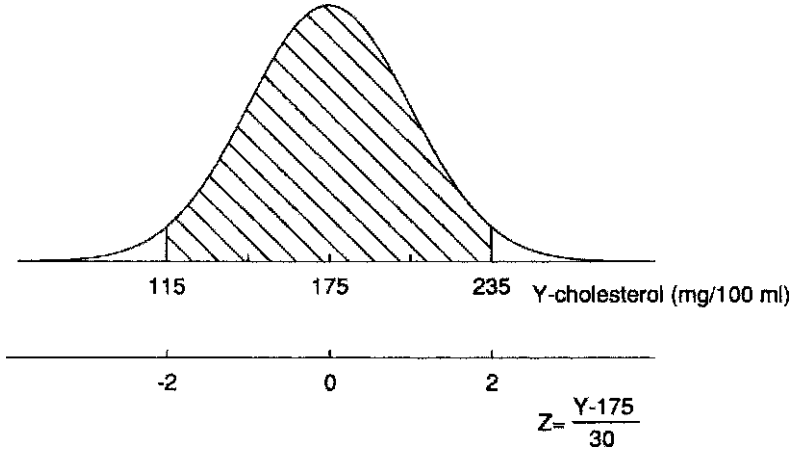


Figure 4.12 Area with ±2 standard deviations of the mean.

column labeled "one-sided.") The desired probability is

$$P[-2 \leq Z \leq 2] = 1 - 2(0.0228)$$
$$= 0.9544$$

In words, 95.44% of the population of Wisconsin schoolchildren have cholesterol values within normal limits.

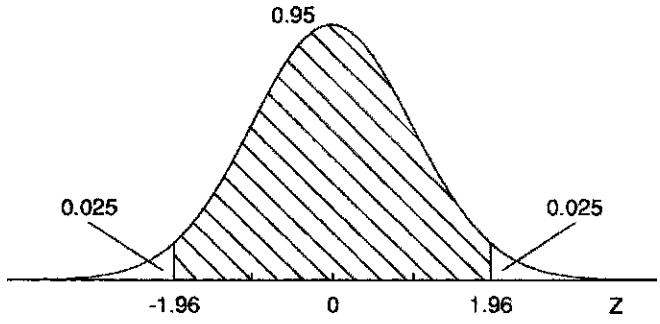


Figure 4.13 Ninety-five percent of normally distributed observations are within ±1.96 standard deviations of the mean.

Suppose that we change the question: Instead of defining normal limits and calculating the proportion within these limits, we define the limits such that, say, 95% of the population has cholesterol values within the stated limits. Before, we went from cholesterol level to Z-value to area; now we want to go from area to Z-value to cholesterol values. In this case, Table A.2 will be useful. Again, we begin with an illustration, Figure 4.13. From Table A.2 we get $P[Z > 1.96] = 0.025$, so that $P[-1.96 \leq Z \leq 1.96] = 0.95$; in words, 95% of normally distributed observations are within ±1.96 standard deviations of the mean. Or, translated to cholesterol values by the formula, $Y = 175 + 30Z$. For $Z = 1.96$, $Y = 175 + (30)(1.96) = 233.8 \doteq 234$, and for $Z = -1.96$, $Y = 175 + (30)(-1.96) = 116.2 \doteq 116$. On the basis of the model, 95% of cholesterol values of Wisconsin children are between 116 and 234 mg per 100 mL. If the mean and standard deviation of cholesterol values of Wisconsin children are 175 and 30 mg per 100 mL, respectively, the 95% limits (116, 234) are called 95% *tolerance limits*.

Often, it is useful to know the range of normal values of a substance (variable) in a normal population. A laboratory test can then be carried out to determine whether a subject's values are high, low, or within normal limits.

Example 4.6. An article by Zervas et al. [1970] provides a list of normal values for more than 150 substances ranging from ammonia to vitamin B_{12}. These values have been reprinted in *The Merck Manual of Diagnosis and Therapy* [Berkow, 1999]. The term *normal values* does not imply that variables are normally distributed (i.e., follow a Gaussian or bell-shaped curve). A paper by Elveback et al. [1970] already indicated that of seven common substances (calcium, phosphate, total protein, albumin, urea, magnesium, and alkaline phosphatase), only albumin values can be summarized adequately by a normal distribution. All the other substances had distributions of values that were skewed. The authors (correctly) conclude that "the distributions of values in healthy persons *cannot* be assumed to be normal." Admittedly, this leaves an unsatisfactory situation: What, then, do we mean by *normal limits*? What proportion of normal values will fall outside the normal limits as the result of random variation? None of these—and other—critical questions can now be answered, because a statistical model is not available. But that appears to be the best we can do at this point; as the authors point out, "good limits are hard to get, and bad limits hard to change."

4.4.3 Quantile–Quantile Plots

How can we know whether the normal distribution model fits a particular set of data? There are many tests for normality, some graphical, some numerical. In this section we discuss a simple graphical test, the *quantile–quantile* (QQ) *plot*. In this approach we plot the quantiles of the data distribution observed against the expected quantiles for the normal distribution. The resulting graph is a version of the cumulative frequency distribution but with distorted axes

chosen so that a normal distribution would give a straight line. In precomputer days, quantile–quantile plots for the normal distribution were obtained by drawing the empirical cumulative frequency distribution on special *normal probability paper*, but it is now possible to obtain quantile–quantile plots for many different distributions from the computer.

A famous book by Galton [1889] contains data on the stature of parents and their adult children. Table 4.3 gives the frequency distributions of heights of 928 adult children. The

Table 4.3 Frequency Distribution of Stature of 928 Adult Children

Endpoint (in.)	Frequency	Cumulative Frequency	Cumulative Percentage
61.7[a]	5	5	0.5
62.2	7	12	1.3
63.2	32	44	4.7
64.2	59	103	11.1
65.2	48	151	16.3
66.2	117	268	28.9
67.2	138	406	43.8
68.2	120	526	56.7
69.2	167	693	74.7
70.2	99	792	85.3
71.2	64	856	92.2
72.2	41	897	96.7
73.2	17	914	98.5
73.7[a]	14	928	100

Source: Galton [1889].

[a] Assumed endpoint.

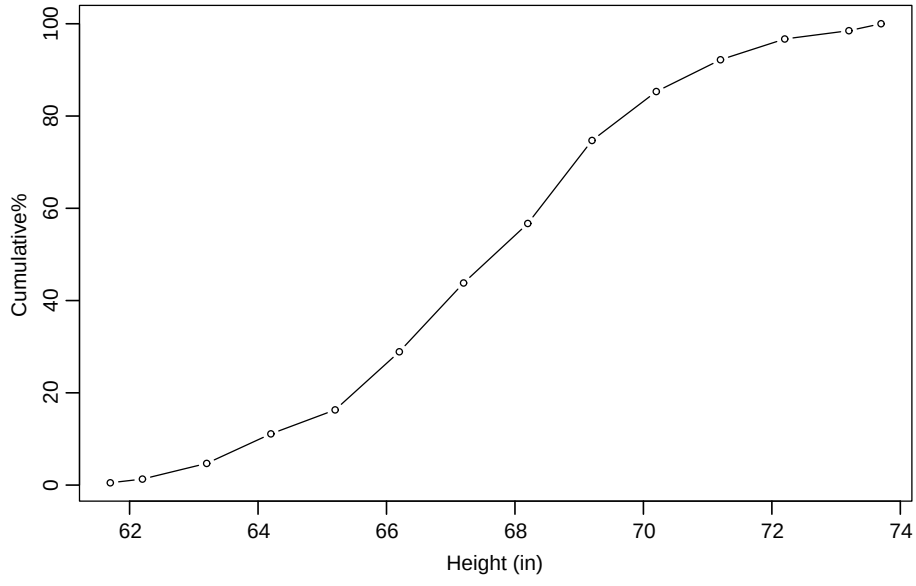


Figure 4.14 Empirical cumulative frequency polygon of heights of 928 adult children. (Data from Galton [1889].)

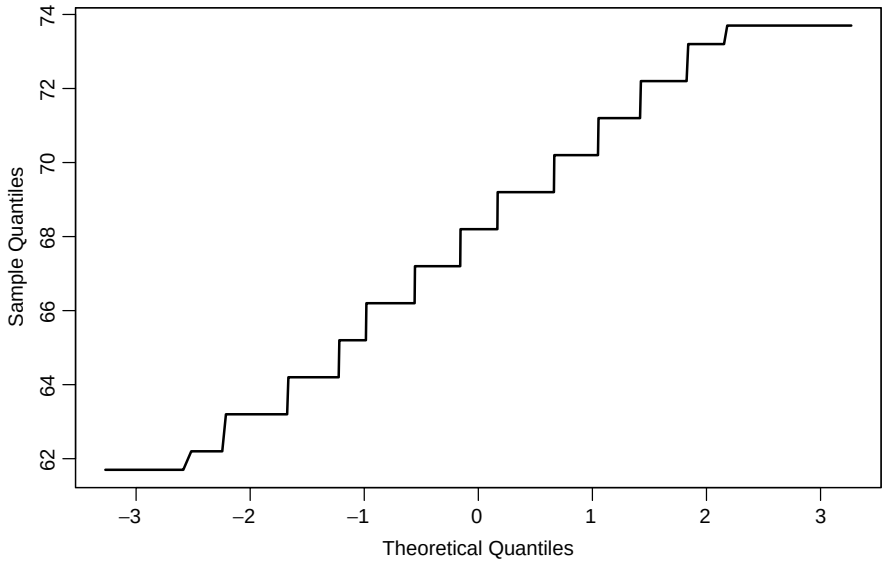


Figure 4.15 Quantile–quantile plot of heights of 928 adult children. (Data from Galton [1889].)

cumulative percentages plotted against the endpoints of the intervals in Figure 4.14 produce the usual sigmoid-shaped curve.

These data are now plotted on normal probability paper in Figure 4.15. The vertical scale has been stretched near 0% and 100% in such a way that data from a normal distribution should fall on a straight line. Clearly, the data are consistent with a normal distribution model.

4.5 SAMPLING DISTRIBUTIONS

4.5.1 Statistics Are Random Variables

Consider a large multicenter collaborative study of the effectiveness of a new cancer therapy. A great deal of care is taken to standardize the treatment from center to center, but it is obvious that the average survival time on the new therapy (or increased survival time if compared to a standard treatment) will vary from center to center. This is an illustration of a basic statistical fact: Sample statistics vary from sample to sample. The key idea is that a statistic associated with a random sample is a random variable. What we want to do in this section is to relate the variability of a statistic based on a random sample to the variability of the random variable on which the sample is based.

Definition 4.15. The probability (density) function of a statistic is called the *sampling distribution of the statistic*.

What are some of the characteristics of the sampling distribution? In this section we state some results about the sample mean. In Section 4.8 some properties of the sampling distribution of the sample variance are discussed.

4.5.2 Properties of Sampling Distribution

Result 4.1. If a random variable Y has population mean μ and population variance σ^2, the sampling distribution of sample means (of samples of size n) has population mean μ and

population variance σ^2/n. Note that this result does not assume normality of the "parent" population.

Definition 4.16. The standard deviation of the sampling distribution is called the *standard error*.

Example 4.7. Suppose that IQ is a random variable with mean $\mu = 100$ and standard deviation $\sigma = 15$. Now consider the average IQ of classes of 25 students. What are the population mean and variance of these class averages? By Result 4.1, the class averages have population mean $\mu = 100$ and population variance $\sigma^2/n = 15^2/25 = 9$. Or, the standard error is $\sqrt{\sigma^2/n} = \sqrt{15^2/25} = \sqrt{9} = 3$.

To summarize:

	Population		
	Mean	**Variance**	**$\sqrt{\text{Variance}}$**
Single observation, Y	100	$15^2 = 225$	$15 = \sigma$
Mean of 25 observations, $\overline{Y}$	100	$15^2/25 = 9$	$3 = \sigma/\sqrt{n}$

The standard error of the sampling distribution of the sample mean $\overline{Y}$ is indicated by $\sigma_{\overline{Y}}$ to distinguish it from the standard deviation, σ, associated with the random variable Y. It is instructive to contemplate the formula for the standard error, $\sigma/\sqrt{n}$. This formula makes clear that a reduction in variability by, say, a factor of 2 requires a fourfold increase in sample size. Consider Example 4.7. How large must a class be to reduce the standard error from 3 to 1.5? We want $\sigma/\sqrt{n} = 1.5$. Given that $\sigma = 15$ and solving for n, we get $n = 100$. This is a fourfold increase in class size, from 25 to 100. In general, if we want to reduce the standard error by a factor of k, we must increase the sample size by a factor of k^2. This suggests that if a study consists of, say, 100 observations and with a great deal of additional effort (out of proportion to the effort of getting the 100 observations) another 10 observations can be obtained, the additional 10 may not be worth the effort.

The standard error based on 100 observations is $\sigma/\sqrt{100}$. The ratio of these standard errors is

$$\frac{\sigma/\sqrt{100}}{\sigma/\sqrt{110}} = \frac{\sqrt{100}}{\sqrt{110}} = 0.95$$

Hence a 10% increase in sample size produces only a 5% increase in precision. Of course, precision is not the only criterion we are interested in; if the 110 observations are randomly selected persons to be interviewed, it may be that the last 10 are very hard to locate or difficult to persuade to take part in the study, and not including them may introduce a serious *bias*. But with respect to *precision* there is not much difference between means based on 100 observations and means based on 110 observations (see Note 4.11).

4.5.3 Central Limit Theorem

Although Result 4.1 gives some characteristics of the sampling distribution, it does not permit us to calculate probabilities, because we do not know the form of the sampling distribution. To be able to do this, we need the following:

Result 4.2. If Y is normally distributed with mean μ and variance σ^2, then $\overline{Y}$, based on a random sample of n observations, is *normally distributed* with mean μ and variance σ^2/n.

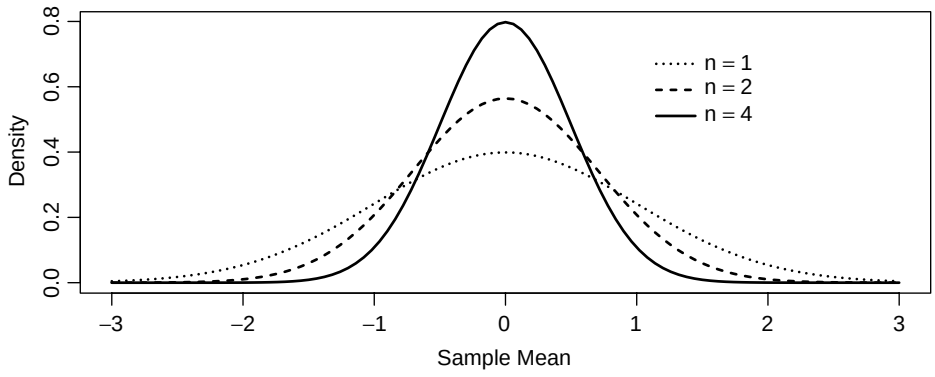


Figure 4.16 Three sampling distributions for means of random samples of size 1, 2, and 4 from a $N(0, 1)$ population.

Result 4.2 basically states that if Y is normally distributed, then $\overline{Y}$, the mean of a random sample, is normally distributed. Result 4.1 then specifies the mean and variance of the sampling distribution. Result 4.2 implies that as the sample size increases, the (normal) distribution of the sample mean becomes more and more "pinched." Figure 4.16 shows three sampling distributions for means of random samples of size 1, 2, and 4.

What is the probability that the average IQ of a class of 25 students exceeds 106? By Result 4.2, $\overline{Y}$, the average of 25 IQs, is normally distributed with mean $\mu = 100$ and standard error $\sigma/\sqrt{n} = 15/\sqrt{25} = 3$. Hence the probability that $\overline{Y} > 106$ can be calculated as

$$\begin{aligned} P[\overline{Y} \geq 106] &= P\left[Z \geq \frac{106 - 100}{3}\right] \\ &= P[Z \geq 2] \\ &= 1 - 0.9772 \\ &= 0.0228 \end{aligned}$$

So approximately 2% of average IQs of classes of 25 students will exceed 106. This can be compared with the probability that a single person's IQ exceeds 106:

$$P[Y > 106] = P\left[Z > \frac{6}{15}\right] = P[Z > 0.4] = 0.3446$$

The final result we want to state is known as the *central limit theorem.*

Result 4.3. If a random variable Y has population mean μ and population variance σ^2, the sample mean $\overline{Y}$, based on n observations, is approximately normally distributed with mean μ and variance σ^2/n, for sufficiently large n.

This is a remarkable result and the most important reason for the central role of the normal distribution in statistics. What this states basically is that means of random samples from *any* distribution (with mean and variance) will tend to be normally distributed as the sample size becomes sufficiently large. How large is "large"? Consider the distributions of Figure 4.2. Samples of six or more from the first three distributions will have means that are virtually normally

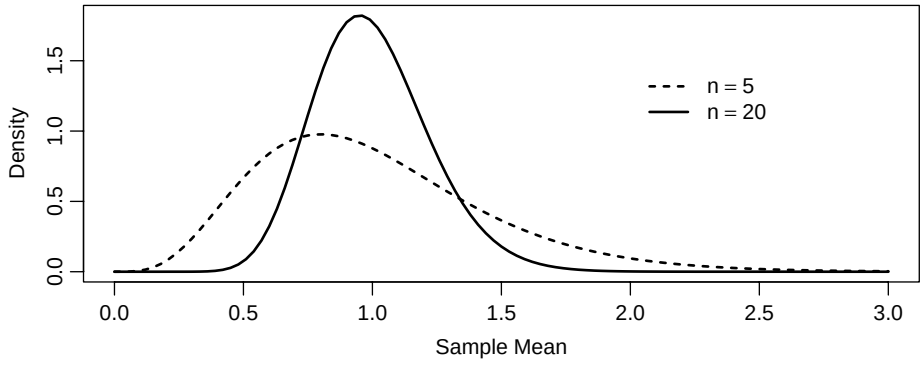


Figure 4.17 Sampling distributions of means of 5 and 20 observations when the parent distribution is exponential.

distributed. The fourth distribution will take somewhat larger samples before approximate normality is obtained; n must be around 25 or 30. Figure 4.17 is a more skewed figure that shows the sampling distributions of means of samples of various sizes drawn from Figure 4.2(d).

The central limit theorem provides some reassurance when we are not certain whether observations are normally distributed. The means of reasonably sized samples will have a distribution that is approximately normal. So inference procedures based on the sample means can often use the normal distribution. But you must be careful not to impute normality to the original observations.

4.6 INFERENCE ABOUT THE MEAN OF A POPULATION

4.6.1 Point and Interval Estimates

In this section we discuss inference about the mean of a population when the population variance is known. The assumption may seem artificial, but sometimes this situation will occur. For example, it may be that a new treatment alters the level of a response variable but not its variability, so that the variability can be assumed to be known from previous experiments. (In Section 4.8 we discuss a method for comparing the variability of an experiment with previous established variability; in Chapter 5 the problem of inference when both population mean and variance are unknown is considered.)

To put the problem more formally, we have a random variable Y with unknown population mean μ. A random sample of size n is taken and inferences about μ are to be made on the basis of the sample. We assume that the population variance is known; denote it by σ^2. Normality will also be assumed; even when the population is not normal, we may be able to appeal to the central limit theorem.

A "natural" estimate of the population mean μ is the sample mean $\overline{Y}$. It is a natural estimate of μ because we know that $\overline{Y}$ is normally distributed with the same mean, μ, and variance σ^2/n. Even if Y is not normal, $\overline{Y}$ is approximately normal on the basis of the central limit theorem. The statistic $\overline{Y}$ is called a *point estimate* since we estimate the parameter μ by a single value or point.

Now the question arises: How precise is the estimate? How can we distinguish between two samples of, say, 25 and 100 observations? Both may give the same—or approximately the same—sample mean, but we know that the mean based on the 100 observations is more accurate, that is, has a smaller standard error. One possible way of summarizing this information is to give the sample mean and its standard error. This would be useful for *comparing* two samples. But this does not seem to be a useful approach in considering one sample and its information about

the parameter. To use the information in the sample, we set up an *interval* estimate as follows: Consider the quantity $\mu \pm (1.96)\sigma/\sqrt{n}$. It describes the spread of sample means; in particular, 95% of means of samples of size n will fall in the interval $[\mu - 1.96\sigma/\sqrt{n}, \mu + 1.96\sigma/\sqrt{n}]$. The interval has the property that as n increases, the width decreases (refer to Section 4.5 for further discussion). Suppose that we now replace μ by its point estimate, $\overline{Y}$. How can we interpret the resulting interval? Since the sample mean, $\overline{Y}$, varies from sample to sample, it cannot mean that 95% of the sample means will fall in the interval for a specific sample mean. The interpretation is that the probability is 0.95 that the interval *straddles* the population mean. Such an interval is referred to as a *95% confidence interval* for the population mean, μ. We now formalize this definition.

Definition 4.17. A $100(1-\alpha)\%$ *confidence interval* for the mean μ of a normal population (with variance known) based on a random sample of size n is

$$\overline{Y} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $z_{1-\alpha/2}$ is the value of the standard normal deviate such that $100(1-\alpha)\%$ of the area falls within $\pm z_{1-\alpha/2}$.

Strictly speaking, we should write

$$\left(\overline{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

but by symmetry, $z_{\alpha/2} = -z_{1-\alpha/2}$, so that it is quicker to use the expression above.

Example 4.8. In Section 3.3.1 we discussed the age at death of 78 cases of crib death (SIDS) occurring in King County, Washington, in 1976–1977. Birth certificates were obtained for these cases and birthweights were tabulated. Let Y = birthweight in grams. Then, for these 78 cases, $\overline{Y} = 2993.6 = 2994$ g. From a listing of all the birthweights, it is known that the standard deviation of birthweight is about 800 g (i.e., $\sigma = 800$ g). A 95% confidence interval for the mean birthweight of SIDS cases is calculated to be

$$2994 \pm (1.96)\left(\frac{800}{\sqrt{78}}\right) \quad \text{or} \quad 2994 \pm (1.96)(90.6) \quad \text{or} \quad 2994 \pm 178$$

producing a lower limit of 2816 g and an upper limit of 3172 g. Thus, on the basis of these data, we are 95% confident that we have straddled the population mean, μ, of birthweight of SIDS infants by the interval (2816, 3172).

Suppose that we had wanted to be more confident: say, a level of 99%. The value of Z now becomes 2.58 (from Table A.2), and the corresponding limits are $2994 \pm (2.58)(800/\sqrt{78})$, or (2760, 3228). The width of the 99% confidence interval is greater than that of the 95% confidence interval (468 g vs. 356 g), the price we paid for being more sure that we have straddled the population mean.

Several comments should be made about confidence intervals:

1. Since the population mean μ is fixed, it is not correct to say that the probability is $1-\alpha$ that μ is in the confidence interval *once it is computed*; that probability is zero or 1. Either the mean is in the interval and the probability is equal to 1, or the mean is not in the interval and the probability is zero.

2. We can increase our confidence that the interval straddles the population mean by decreasing α, hence increasing $Z_{1-\alpha/2}$. We can take values from Table A.2 to construct the following confidence levels:

Confidence Level	Z-Value
90%	1.64
95%	1.96
99%	2.58
99.9%	3.29

The effect of increasing the confidence level will be to increase the width of the confidence interval.

3. To decrease the width of the confidence interval, we can either decrease the confidence level or increase the sample size. The width of the interval is $2z_{1-\alpha/2}\sigma/\sqrt{n}$. For a fixed confidence level the width is essentially a function of $\sigma/\sqrt{n}$, the standard error of the mean. To decrease the width by a factor of, say, 2, the sample size must be increased by a factor of 4, analogous to the discussion in Section 4.5.2.
4. Confidence levels are usually taken to be 95% or 99%. These levels are a matter of convention; there are no theoretical reasons for choosing these values. A rough rule to keep in mind is that a 95% confidence interval is defined by the sample mean ± 2 standard errors (*not* standard deviations).

4.6.2 Hypothesis Testing

In estimation, we start with a sample statistic and make a statement about the population parameter: A confidence interval makes a probabilistic statement about straddling the population parameter. In hypothesis testing, we start by assuming a value for a parameter, and a probability statement is made about the value of the corresponding statistic. In this section, as in Section 4.6.1, we assume that the population variance is known and that we want to make inferences about the mean of a normal population on the basis of a sample mean. The basic strategy in hypothesis testing is to measure how far an observed statistic is from a hypothesized value of the parameter. If the distance is "great" (Figure 4.18) we would argue that the hypothesized parameter value is inconsistent with the data and we would be inclined to reject the hypothesis (we could be wrong, of course; rare events do happen).

To interpret the distance, we must take into account the basic variability (σ^2) of the observations and the size of the sample (n) on which the statistic is based. As a rough rule of thumb that is explained below, if the observed value of the statistic is more than two standard errors from the hypothesized parameter value, we question the truth of the hypothesis.

To continue Example 4.8, the mean birthweight of the 78 SIDS cases was 2994 g. The standard deviation σ_0 was assumed to be 800 g, and the standard error $\sigma/\sqrt{n} = 800/\sqrt{78} = 90.6$ g. One question that comes up in the study of SIDS is whether SIDS cases tend to have a different birthweight than the general population. For the general population, the average birthweight is about 3300 g. Is the *sample* mean value of 2994 g consistent with this value? Figure 4.19 shows that the distance between the two values is 306 g. The standard error is 90.6,

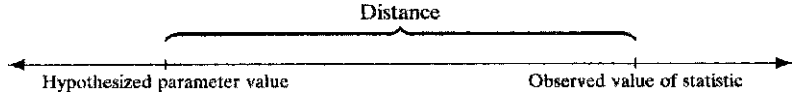


Figure 4.18 Great distance from a hypothesized value of a parameter.

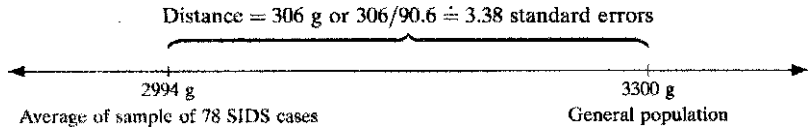


Figure 4.19 Distance between the two values is 306 g.

so the observed value is 306/90.6 = 3.38 standard errors from the hypothesized population mean. By the rule we stated, the distance is so great that we would conclude that the mean of the *sample* of SIDS births is inconsistent with the mean value in the general population. Hence, we would conclude that the SIDS births come from a population with mean birthweight somewhat less than that of the general population. (This raises more questions, of course: Are the gestational ages comparable? What about the racial composition? and so on.) The best estimate we have of the mean birthweight of the population of SIDS cases is the sample mean: in this case, 2994 g, about 300 g lower than that for the normal population.

Before introducing some standard hypothesis testing terminology, two additional points should be made:

1. We have expressed "distance" in terms of number of standard errors from the hypothesized parameter value. Equivalently, we can associate a tail probability with the observed value of the statistic. For the sampling situation described above, we know that the sample mean $\overline{Y}$ is normally distributed with standard error $\sigma/\sqrt{n}$. As Figure 4.20 indicates, the farther away the observed value of the statistic is from the hypothesized parameter value, the smaller the area (probability) in the tail. This tail probability is usually called the *p-value*. For example (using Table A.2), the area to the right of 1.96 standard errors is 0.025; the area to the right of 2.58 standard errors is 0.005. Conversely, if we specify the area, the number of standard errors will be determined.
2. Suppose that we planned before doing the statistical test that we would not question the hypothesized parameter value if the observed value of the statistic fell within, say, two standard errors of the parameter value. We could divide the sample space for the statistic (i.e., the real line) into three regions as shown in Figure 4.21. These regions could have been set up before the value of the statistic was observed. All that needs to be determined then is in which region the observed value of the statistic falls to determine if it is consistent with the hypothesized value.

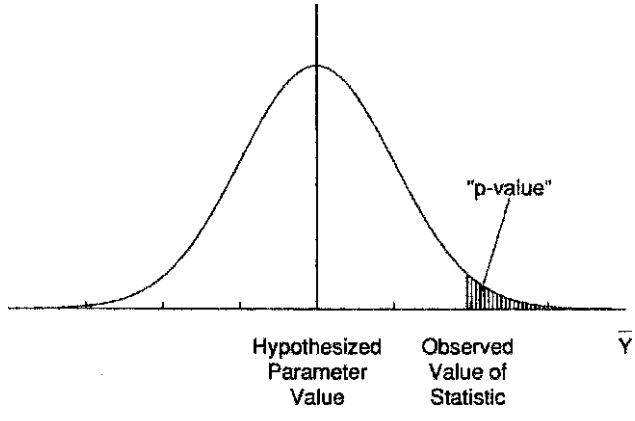


Figure 4.20 The farther away the observed value of a statistic from the hypothesized value of a parameter, the smaller the area in the tail.

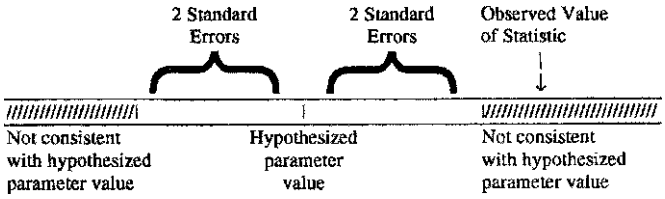


Figure 4.21 Sample space for the statistic.

We now formalize some of these concepts:

Definition 4.18. A *null hypothesis* specifies a hypothesized real value, or values, for a parameter (see Note 4.15 for further discussion).

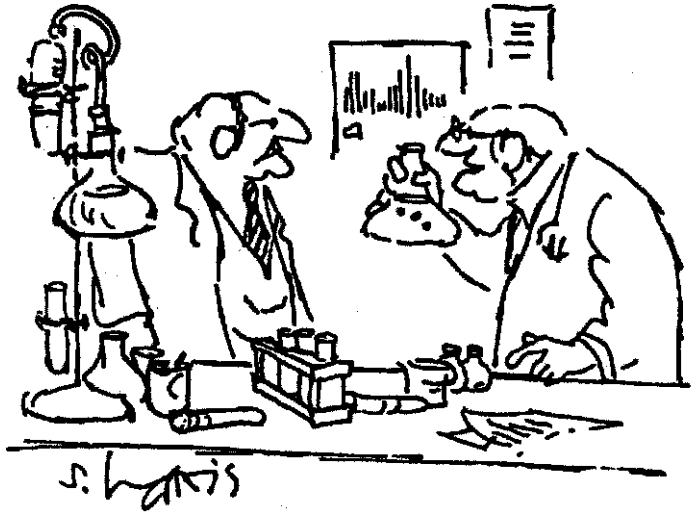
Definition 4.19. The *rejection region* consists of the set of values of a statistic for which the null hypothesis is rejected. The values of the boundaries of the region are called the *critical values*.

Definition 4.20. A *Type I error* occurs when the null hypothesis is rejected when, in fact, it is true. The *significance level* is the probability of a Type I error when the null hypothesis is true.

Definition 4.21. An *alternative hypothesis* specifies a real value or range of values for a parameter that will be considered when the null hypothesis is rejected.

Definition 4.22. A *Type II error* occurs when the null hypothesis is not rejected when it is false.

Definition 4.23. The *power of a test* is the probability of rejecting the null hypothesis when it is false.



"It may very well bring about immortality, but it will take forever to test it."

© 1976 by Sidney Harris — *American Scientist* Magazine

Cartoon 4.1 Testing some hypotheses can be tricky. (From *American Scientist*, March–April 1976.)

Definition 4.24. The *p-value* in a hypothesis testing situation is that value of p, $0 \leq p \leq 1$, such that for $\alpha > p$ the test rejects the null hypothesis at significance level α, and for $\alpha < p$ the test does not reject the null hypothesis. Intuitively, the p-value is the probability under the null hypothesis of observing a value as unlikely or more unlikely than the value of the test statistic. The p-value is a measure of the distance from the observed statistic to the value of the parameter specified by the null hypothesis.

Notation

1. The null hypothesis is denoted by H_0 the alternative hypothesis by H_A.
2. The probability of a Type I error is denoted by α, the probability of a Type II error by β. The power is then

$$\begin{aligned} \text{power} &= 1 - \text{probability of Type II error} \\ &= 1 - \beta \end{aligned}$$

Continuing Example 4.8, we can think of our assessment of the birthweight of SIDS babies as a type of decision problem illustrated in the following layout:

	State of Nature SIDS Birthweights	
Decision SIDS Birthweights	**Same as Normal**	**Not the Same**
Same as normal	Correct $(1 - \alpha)$	Type II error (β)
Not the same	Type I error (α)	Correct $(1 - \beta)$

This illustrates the two types of errors that can be made depending on our decision and the *state of nature*. The null hypothesis for this example can be written as

$$H_0 : \mu = 3300 \text{ g}$$

and the alternative hypothesis written as

$$H_A : \mu \neq 3300 \text{ g}$$

Suppose that we want to reject the null hypothesis when the sample mean $\overline{Y}$ is more than two standard errors from the H_0 value of 3300 g. The standard error is 90.6 g. The rejection region is then determined by $3300 \pm (2)(90.6)$ or 3300 ± 181.

We can then set up the hypothesis-testing framework as indicated in Figure 4.22. The rejection region consists of values to the left of 3119 g (i.e., $\mu - 2\sigma/\sqrt{n}$) and to the right of 3481 g (i.e., $\mu + 2\sigma/\sqrt{n}$). The observed value of the statistic, $\overline{Y} = 2994$ g, falls in the rejection region, and we therefore reject the null hypothesis that SIDS cases have the same mean birthweight as normal children. On the basis of the sample value observed, we conclude that SIDS babies tend to weigh less than normal babies.

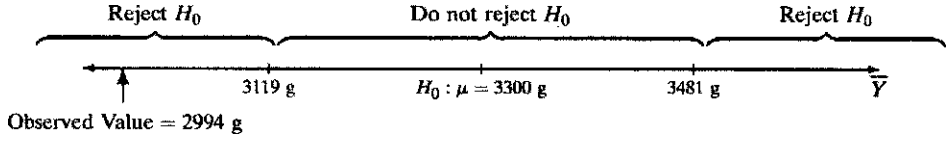


Figure 4.22 Hypothesis-testing framework for birthweight assessment.

The probability of a Type I error is the probability that the mean of a sample of 78 observations from a population with mean 3300 g is less than 3119 g or greater than 3481 g:

$$P[3119 \le \overline{Y} \le 3481] = P\left[\frac{3119 - 3300}{90.6} \le Z \le \frac{3481 - 3300}{90.6}\right]$$
$$= P[-2 \le Z \le +2]$$

where Z is a standard normal deviate.
From Table A.1,

$$P[Z \le 2] = 0.9772$$

so that

$$1 - P[-2 \le Z \le 2] = (2)(0.0228) = 0.0456$$

the probability of a Type I error. The probability is 0.0455 from the two-sided p-value of Table A.1. The difference relates to rounding.

The probability of a Type II error can be computed when a value for the parameter under the alternative hypothesis is specified. Suppose that for these data the alternative hypothesis is

$$H_A : \mu = 3000 \text{ g}$$

this value being suggested from previous studies. To calculate the probability of a Type II error—and the power—we assume that $\overline{Y}$, the mean of the 78 observations, comes from a normal distribution with mean 3000 g and standard error as before, 90.6 g. As Figure 4.23 indicates, the probability of a Type II error is the area over the interval (3119, 3481). This can be calculated as

$$P[\text{Type II error}] = P[3119 \le \overline{Y} \le 3481]$$
$$= P\left[\frac{3119 - 3000}{90.6} \le Z \le \frac{3481 - 3000}{90.6}\right]$$
$$\doteq P[1.31 \le Z \le 5.31]$$
$$\doteq 1 - 0.905$$
$$\doteq 0.095$$

So $\beta = 0.095$ and the power is $1 - \beta = 0.905$. Again, these calculations can be made before any data are collected, and they say that if the SIDS population mean birthweight were 3000 g and the normal population birthweight 3300 g, the probability is 0.905 that a mean from a sample of 78 observations will be declared significantly different from 3300 g.

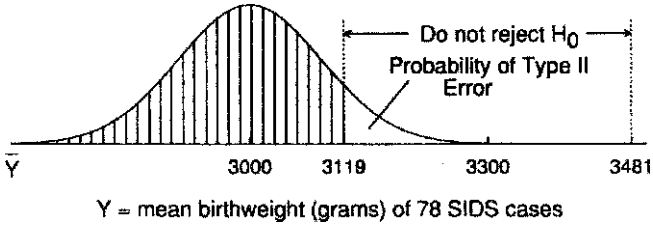


Figure 4.23 Probability of a Type II error.

Let us summarize the analysis of this example:

Hypothesis-testing setup (no data taken)

- $H_0 : \mu = 3300$ g
- $H_A : \mu = 3000$ g
- $\sigma = 800$ g (known)
- $n = 78$
- rejection region: ± 2 standard errors from 3000 g
- $\alpha = 0.0456$
- $\beta = 0.095$
- $1 - \beta = 0.905$

Observe: $\overline{Y} = 2994$

Conclusion: Reject H_0

The value of α is usually specified beforehand: The most common value is 0.05, somewhat less common values are 0.01 or 0.001. Corresponding to the confidence level in interval estimation, we have the *significance level* in hypothesis testing. The significance level is often expressed as a percentage and defined to be $100\alpha\%$. Thus, for $\alpha = 0.05$, the hypothesis test is carried out at the 5%, or 0.05, significance level.

The use of a single symbol β for the probability of a Type II error is standard but a bit misleading. We expect β to stand for one number in the same way that α stands for one number. In fact, β is a function whose argument is the assumed true value of the parameter being tested. For example, in the context of $H_A : \mu = 3000$ g, β is a function of μ and could be written $\beta(\mu)$. It follows that the power is also a function of the true parameter: power $= 1 - \beta(\mu)$. Thus one must specify a value of μ to compute the power.

We finish this introduction to hypothesis testing with a discussion of the one- and two-tailed test. These are related to the choice of the rejection region. Even if α is specified, there is an infinity of rejection regions such that the area over the region is equal to α. Usually, only two types of regions are considered as shown in Figure 4.24. A *two-tailed test* is associated with a

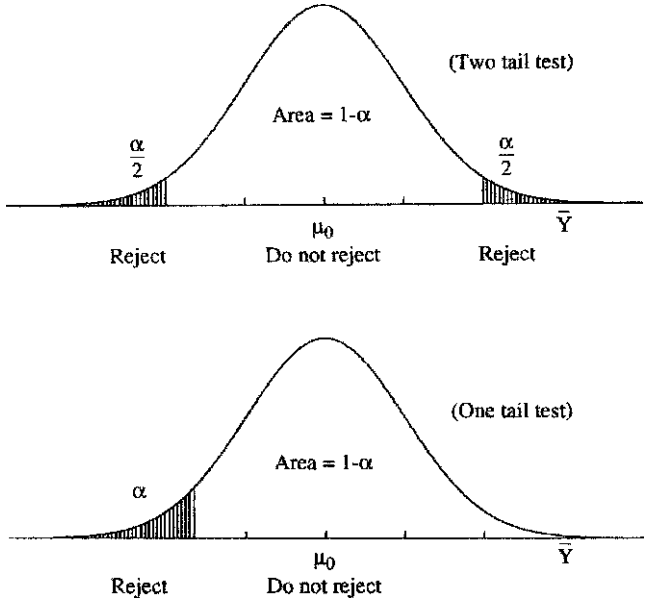


Figure 4.24 Two types of regions considered in hypothesis testing.

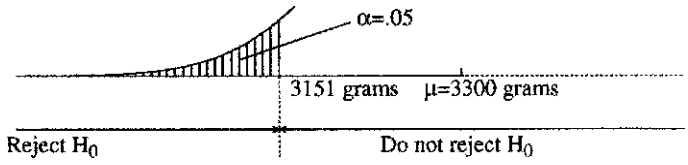


Figure 4.25 Start of the rejection region in a one-tailed test.

rejection region that extends both to the left and to the right of the hypothesized parameter value. A *one-tailed test* is associated with a region to one side of the parameter value. The alternative hypothesis determines the type of test to be carried out. Consider again the birthweight of SIDS cases. Suppose we know that if the mean birthweight of these cases is not the same as that of normal infants (3300 g), it must be less; it is not possible for it to be more. In that case, if the null hypothesis is false, we would expect the sample mean to be below 3300 g, and we would reject the null hypothesis for values of $\overline{Y}$ below 3300 g. We could then write the null hypothesis and alternative hypothesis as follows:

$$H_0 : \mu = 3300 \text{ g}$$
$$H_A : \mu < 3300 \text{ g}$$

We would want to carry out a one-tailed test in this case by setting up a rejection region to the left of the parameter value. Suppose that we want to test at the 0.05 level, and we only want to reject for values of $\overline{Y}$ below 3300 g. From Table A.2 we see that we must locate the start of the rejection region 1.64 standard errors to the left of $\mu = 3300$ g, as shown in Figure 4.25. The value is $3300 - (1.64)(800/\sqrt{78})$ or $3300 - (1.64)(90.6) = 3151$ g.

Suppose that we want a two-tailed test at the 0.05 level. The Z-value (Table A.2) is now 1.96, which distributes 0.025 in the left tail and 0.025 in the right tail. The corresponding values for the critical region are $3300 \pm (1.96)(90.6)$ or (3122, 3478), producing a region very similar to the region calculated earlier.

The question is: When should you do a one-tailed test and when a two-tailed test? As was stated, the alternative hypothesis determines this. An alternative hypothesis of the form $H_A : \mu \neq \mu_0$ is called *two-sided* and will require a two-tailed test. Similarly, the alternative $H_A : \mu < \mu_0$ is called one-sided and will lead to a one-tailed test. So should the alternative hypothesis be one- or two-sided? The experimental situation will determine this. For example, if nothing is known about the effect of a proposed therapy, the alternative hypothesis should be made two-sided. However, if it is suspected that a new therapy will do nothing or increase a response level, and if there is no reason to distinguish between no effect and a decrease in the response level, the test should be one-tailed. The general rule is: The more specific you can make the experiment, the greater the power of the test (see Fleiss et al. [2003, Sec. 2.4]). (See Problem 4.33 to convince yourself that the power of a one-tailed test is greater *if* the alternative hypothesis specifies the situation correctly.)

4.7 CONFIDENCE INTERVALS VS. TESTS OF HYPOTHESES

You may have noticed that there is a very close connection between the confidence intervals and the tests of hypotheses that we have constructed. In both approaches we have used the standard normal distribution and the quantity α.

In *confidence intervals* we:

1. Specify the confidence level $(1 - \alpha)$.

2. Read $z_{1-\alpha/2}$ from a standard normal table.
3. Calculate $\overline{Y} \pm z_{1-\alpha/2}\sigma/\sqrt{n}$.

In *hypothesis testing* we:

1. Specify the null hypothesis ($H_0 : \mu = \mu_0$).
2. Specify α, the probability of a Type I error.
3. Read $z_{1-\alpha/2}$ from a standard normal table.
4. Calculate $\mu_0 \pm z_{1-\alpha/2}\sigma/\sqrt{n}$.
5. Observe $\overline{Y}$; reject or accept H_0.

The two approaches can be represented pictorially as shown in Figure 4.26. It is easy to verify that if the confidence interval does not straddle μ_0 (as is the case in the figure), $\overline{Y}$ will fall in the rejection region, and vice versa. Will this always be the case? The answer is "yes." When we are dealing with inference about the value of a parameter, the two approaches will give the same answer. To show the equivalence algebraically, we start with the key inequality

$$P\left[-z_{1-\alpha/2} \leq \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right] = 1 - \alpha$$

If we solve the inequality for $\overline{Y}$, we get

$$P\left[\mu - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}} \leq \overline{Y} \leq \mu + \frac{z_{1-\alpha/2}}{\sqrt{n}}\right] = 1 - \alpha$$

Given a value $\mu = \mu_0$, the statement produces a region ($\mu_0 \pm z_{1-\alpha/2}\sigma/\sqrt{n}$) within which $100(1-\alpha)\%$ of sample means fall. If we solve the inequality for μ, we get

$$P\left[\overline{Y} - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \overline{Y} + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

This is a confidence interval for the population mean μ. In Chapter 5 we examine this approach in more detail and present a general methodology.

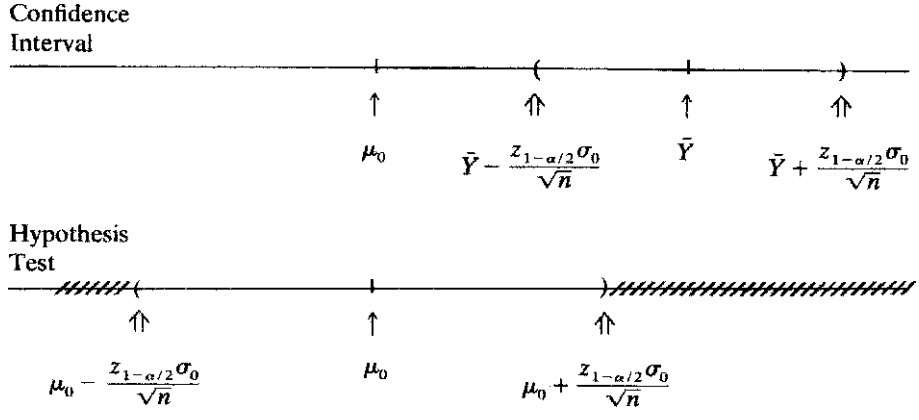


Figure 4.26 Confidence intervals vs. tests of hypothesis.

If confidence intervals and hypothesis testing are but two sides of the same coin, why bother with both? The answer is (to continue the analogy) that the two sides of the coin are not the same; there is different information. The confidence interval approach emphasizes the precision of the estimate by means of the width of the interval and provides a point estimate for the parameter, regardless of any hypothesis. The hypothesis-testing approach deals with the consistency of observed (new) data with the hypothesized parameter value. It gives a probability of observing the value of the statistic or a more extreme value. In addition, it will provide a method for estimating sample sizes. Finally, by means of power calculations, we can decide beforehand whether a proposed study is feasible; that is, what is the probability that the study will demonstrate a difference if a (specified) difference exists?

You should become familiar with both approaches to statistical inference. Do not use one to the exclusion of another. In some research fields, hypothesis testing has been elevated to the only "proper" way of doing inference; all scientific questions have to be put into a hypothesis-testing framework. This is absurd and stultifying, particularly in pilot studies or investigations into uncharted fields. On the other hand, not to consider *possible* outcomes of an experiment and the chance of picking up differences is also unbalanced. Many times it will be useful to specify very carefully what is known about the parameter(s) of interest *and* to specify, in perhaps a crude way, alternative values or ranges of values for these parameters. If it is a matter of emphasis, you should stress hypothesis testing before carrying out a study and estimation after the study has been done.

4.8 INFERENCE ABOUT THE VARIANCE OF A POPULATION

4.8.1 Distribution of the Sample Variance

In previous sections we assumed that the population variance of a normal distribution was known. In this section we want to make inferences about the population variance on the basis of a sample variance. In making inferences about the population mean, we needed to know the sampling distribution of the sample mean. Similarly, we need to know the sampling distribution of the sample variance in order to make inferences about the population variance; analogous to the statement that for a normal random variable, Y, with sample mean $\overline{Y}$, the quantity

$$\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$$

has a normal distribution with mean 0 and variance 1. We now state a result about the quantity $(n-1)s^2/\sigma^2$. The basic information is contained in the following statement:

Result 4.4. If a random variable Y is normally distributed with mean μ and variance σ^2, then for a random sample of size n the quantity $(n-1)s^2/\sigma^2$ has a chi-square distribution with $n-1$ degrees of freedom.

Each distribution is indexed by $n-1$ degrees of freedom. Recall that the sample variance is calculated by dividing $\sum(y-\overline{y})^2$ by $n-1$, the degrees of freedom.

The chi-square distribution is skewed; the amount of skewness decreases as the degrees of freedom increases. Since $(n-1)s^2/\sigma^2$ can never be negative, the sample space for the chi-square distribution is the nonnegative part of the real line. Several chi-square distributions are shown in Figure 4.27. The mean of a chi-square distribution is equal to the degrees of freedom, and

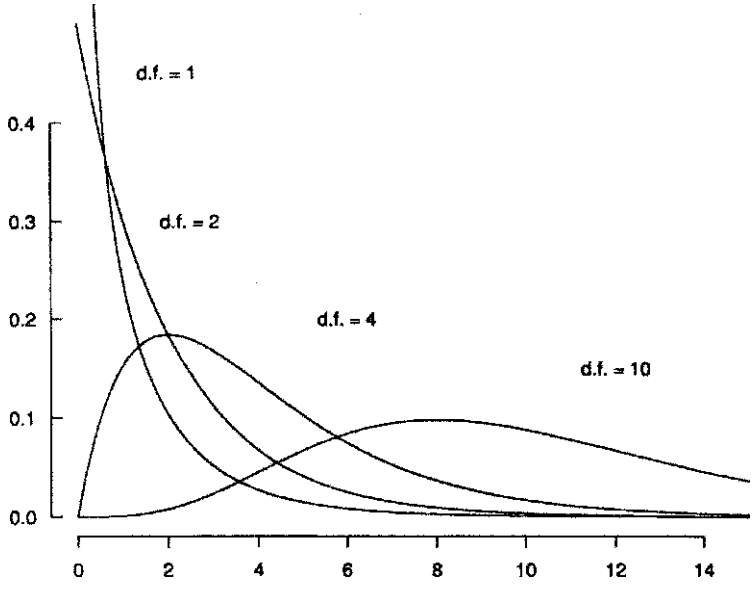


Figure 4.27 Chi-square distributions.

the variance is twice the degrees of the freedom. Formally,

$$E\left[\frac{(n-1)s^2}{\sigma^2}\right] = n - 1 \tag{1}$$

$$\text{var}\left[\frac{(n-1)s^2}{\sigma^2}\right] = 2(n-1) \tag{2}$$

It may seem somewhat strange to talk about the variance of the sample variance, but under repeated sampling the sample variance will vary from sample to sample, and the chi-square distribution describes this variation if the observations are from a normal distribution.

Unlike the normal distribution, a tabulation of the chi-square distribution requires a separate listing for each degree of freedom. In Table A.3, a tabulation is presented of percentiles of the chi-square distribution. For example, 95% of chi-square random variables with 10 degrees of freedom have values less than or equal to 18.31. Note that the median (50th percentile) is very close to the degrees of freedom when the number of the degrees of freedom is 10 or more.

The symbol for a chi-square random variable is χ^2, the Greek lowercase letter chi, to the power of 2. So we usually write $\chi^2 = (n-1)s^2/\sigma^2$. The degrees of freedom are usually indicated by the Greek lowercase letter ν (nu). Hence, χ^2_ν is a symbol for a chi-square random variable with ν degrees of freedom. It is not possible to maintain the notation of using a capital letter for a variable and the corresponding lowercase letter for the value of the variable.

4.8.2 Inference about a Population Variance

We begin with hypothesis testing. We have a sample of size n from a normal distribution, the sample variance s^2 has been calculated, and we want to know whether the value of s^2 observed is consistent with a hypothesized population value σ_0^2, perhaps known from previous research. Consider the quantity

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

If s^2 is very close to σ^2, the ratio s^2/σ^2 is close to 1; if s^2 differs very much from σ^2, the ratio is either very large or very close to 0: This implies that $\chi^2 = (n-1)s^2/\sigma^2$ is either very large or very small, and we would want to reject the null hypothesis. This procedure is analogous to a hypothesis test about a population mean; we measured the distance of the observed sample mean from the hypothesized value in units of standard errors; in this case we measure the "distance" in units of the hypothesized variance.

Example 4.9. The SIDS cases discussed in Section 3.3.1 were assumed to come from a normal population with variance $\sigma^2 = (800)^2$. To check this assumption, the variance, s^2, is calculated for the first 11 cases occurring in 1969. The birthweights (in grams) were

3374, 3515, 3572, 2977, 4111, 1899, 3544, 3912, 3515, 3232, 3289

The sample variance is calculated to be

$$s^2 = (574.3126 \text{ g})^2$$

The observed value of the chi-square quantity is

$$\chi^2 = \frac{(11-1)(574.3126)^2}{(800)^2}$$
$$= 5.15 \text{ with 10 degrees of freedom}$$

Figure 4.14 illustrates the chi-square distribution with 10 degrees of freedom. The 2.5th and 97.5th percentiles are 3.25 and 20.48 (see Table A.3). Hence, 95% of chi-square values will fall between 3.25 and 20.48.

If we follow the usual procedure of setting our significance level at $\alpha = 0.05$, we will not reject the null hypothesis that $\sigma^2 = (800 \text{ g})^2$, since the observed value $\chi^2 = 5.15$ is less extreme than 3.25. Hence, there is not sufficient evidence for using a value of σ^2 not equal to 800 g.

As an alternative to setting up the rejection regions formally, we could have noted, using Table A.3, that the observed value of $\chi^2 = 5.15$ is between the 5th and 50th percentiles, and therefore the corresponding two-sided p-value is greater than 0.10.

A $100(1-\alpha)\%$ confidence interval is constructed using the approach of Section 4.7. The key inequality is

$$P[\chi^2_{\alpha/2} \le \chi^2 \le \chi^2_{1-\alpha/2}] = 1-\alpha$$

The degrees of freedom are not indicated but assumed to be $n-1$. The values $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are chi-square values such that $1-\alpha$ of the area is between them. (In Figure 4.14, these values are 3.25 and 20.48 for $1-\alpha = 0.95$.)

The quantity χ^2 is now replaced by its equivalent, $(n-1)s^2/\sigma^2$, so that

$$P\left[\chi^2_{\alpha/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{1-\alpha/2}\right] = 1-\alpha$$

If we solve for σ^2, we obtain a $100(1-\alpha)\%$ confidence interval for the population variance. A little algebra shows that this is

$$P\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right] = 1-\alpha$$

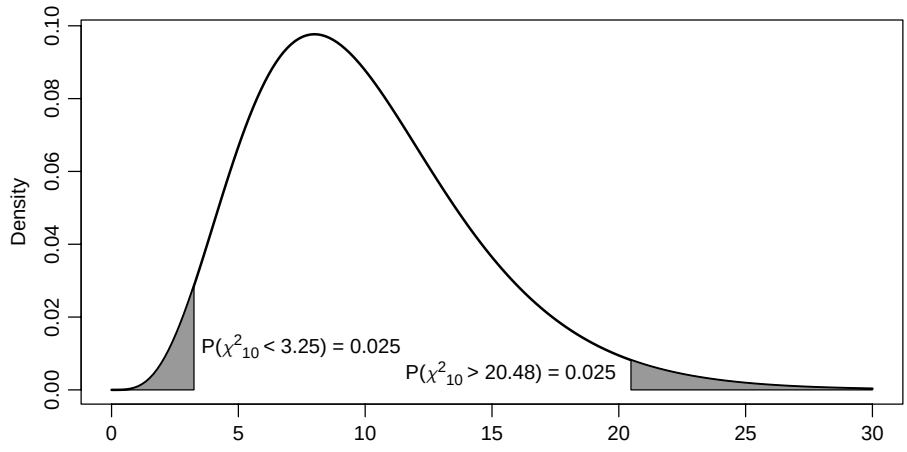


Figure 4.28 Chi-square distribution with 10 degrees of freedom.

Given an observed value of s^2, the confidence interval required can now be calculated.

To continue our example, the variance for the 11 SIDS cases above is $s^2 = (574.3126 \text{ g})^2$. For $1 - \alpha = 0.95$, the values of χ^2 are (see Figure 4.28)

$$\chi^2_{0.025} = 3.25, \qquad \chi^2_{0.975} = 20.48$$

We can write the key inequality then as

$$P[3.25 \leq \chi^2 \leq 20.48] = 0.95$$

The 95% confidence interval for σ^2 can then be calculated:

$$\frac{(10)(574.3126)^2}{20.48} \leq \sigma^2 \leq \frac{(10)(574.3126)^2}{3.25}$$

and simplifying yields

$$161{,}052 \leq \sigma^2 \leq 1{,}014{,}877$$

The corresponding values for the population standard deviation are

$$\text{lower 95\% limit for } \sigma = \sqrt{161,052} = 401 \text{ g}$$

$$\text{upper 95\% limit for } \sigma = \sqrt{1{,}014{,}877} = 1007 \text{ g}$$

These are rather wide limits. Note that they include the null hypothesis value of $\sigma = 800$ g. Thus, the confidence interval approach leads to the same conclusion as the hypothesis-testing approach.

NOTES

4.1 Definition of Probability

The relative frequency definition of probability was advanced by von Mises, Fisher, and others (see Hacking [1965]). A radically different view is held by the *personal* or *subjective school*,

exemplified in the work of De Finetti, Savage, and Savage. According to this school, probability reflects subjective belief and knowledge that can be quantified in terms of betting behavior. Savage [1968] states: "My probability for the event A under circumstances H is the amount of money I am indifferent to betting on A in an elementary gambling situation." What does Savage mean? Consider the thumbtack experiment discussed in Section 4.3.1. Let the event A be that the thumbtack in a single toss falls $\perp$. The other possible outcome is $\top$; call this event B. You are to bet a dollars on A and b dollars on B, such that you are indifferent to betting either on A or on B (you must bet). You clearly would not want to put all your money on A; then you would prefer outcome A. There is a split, then, in the total amount, $a + b$, to be bet so that you are indifferent to either outcome A or B. Then *your* probability of A, $P[A]$, is

$$P[A] = \frac{b}{a+b}$$

If the total amount to be bet is 1 unit, you would split it $1 - P$, P, where $0 \leq P \leq 1$, so that

$$P[A] = \frac{P}{1-P+P} = P$$

The bet is a device to link quantitative preferences for amounts b and a of money, which are assumed to be well understood, to preferences for degrees of certainty, which we are trying to quantify. Note that Savage is very careful to require the estimate of the probability to be made under as specified circumstances. (If the thumbtack could land, say, $\top$ on a soft surface, you would clearly want to modify your probability.) Note also that betting behavior is a *definition* of personal probability rather than a guide for action. In practice, one would typically work out personal probabilities by comparison to events for which the probabilities were already established (Do I think this event is more or less likely than a coin falling heads?) rather than by considering sequences of bets.

This definition of probability is also called *personal probability*. An advantage of this view is that it can discuss more situations than the relative frequency definition, for example: the probability (rather, *my* probability) of life on Mars, or my probability that a cure for cancer will be found. You should not identify personal probability with the irrational or whimsical. Personal probabilities do utilize empirical evidence, such as the behavior of a tossed coin. In particular, if you have good reason to believe that the relative frequency of an event is P, your personal probability will also be P. It is possible to show that any self-consistent system for choosing between uncertain outcomes corresponds to a set of personal probabilities.

Although different individuals will have different personal probabilities for an event, the way in which those probabilities are updated by evidence is the same. It is possible to develop statistical analyses that summarize data in terms of how it should change one's personal probabilities. In simple analyses these *Bayesian methods* are more difficult to use than those based on relative frequencies, but the situation is reversed for some complex models. The use of Bayesian statistics is growing in scientific and clinical research, but it is still not supported by most standard software. An introductory discussion of Bayesian statistics is given by Berry [1996], and more advanced books on practical data analysis include Gelman et al. [1995] and Carlin and Louis [2000]. There are other views of probability. For a survey, see the books by Hacking [1965] and Barnett [1999] and references therein.

4.2 Probability Inequalities

For the normal distribution, approximately 68% of observations are within one standard deviation of the mean, and 95% of observations are within two standard deviations of the mean. If the distribution is not normal, a weaker statement can be made: The proportion of observations

within K standard deviations of the mean is greater than or equal to $(1 - 1/K^2)$; notationally, for a variable Y,

$$P\left[-K \leq \frac{Y - E(Y)}{\sigma} \leq K\right] \leq 1 - \frac{1}{K^2}$$

where K is the number of standard deviations from the mean. This is a version of *Chebyshev's inequality*. For example, this inequality states that at least 75% of the observations fall within two standard deviations of the mean (compared to 95% for the normal distribution). This is not nearly as stringent as the first result stated, but it is more general. If the variable Y can take on only positive values and the mean of Y is μ, the following inequality holds:

$$P[Y \leq y] \leq 1 - \frac{\mu}{y}$$

This inequality is known as the *Markov inequality*.

4.3 Inference vs. Decision

The hypothesis tests discussed in Sections 4.6 and 4.7 can be thought of as decisions that are made with respect to a value of a parameter (or *state of nature*). There is a controversy in statistics as to whether the process of inference is equivalent to a decision process. It seems that a "decision" is sometimes not possible in a field of science. For example, it is not possible at this point to decide whether better control of insulin levels will reduce the risk of neuropathy in diabetes mellitus. In this case and others, the types of inferences we can make are more tenuous and cannot really be called decisions. For an interesting discussion, see Moore [2001]. This is an excellent book covering a variety of statistical topics ranging from ethical issues in experimentation to formal statistical reasoning.

4.4 Representative Samples

A random sample from a population was defined in terms of repeated independent trials or drawings of observations. We want to make a distinction between a random and a representative sample. A random sample has been defined in terms of repeated independent sampling from a population. However (see Section 4.3.2), cancer patients treated in New York are clearly not a random sample of all cancer patients in the world or even in the United States. They will differ from cancer patients in, for instance, Great Britain in many ways. Yet we do frequently make the assumption that if a cancer treatment worked in New York, patients in Great Britain can also benefit. The experiment in New York has wider applicability. We consider that with respect to the outcome of interest in the New York cancer study (e.g., increased survival time), the New York patients, although not a random sample, constitute a representative sample. That is, the survival times are a random sample from the population of survival times.

It is easier to disprove randomness than representativeness. A measure of scientific judgment is involved in determining the latter. For an interesting discussion of the use of the word *representative*, see the papers by Kruskal and Mosteller [1979a–c].

4.5 Multivariate Populations

Usually, we study more than one variable. The Winkelstein et al. [1975] study (see Example 4.1) measured diastolic and systolic blood pressures, height, weight, and cholesterol levels. In the study suggested in Example 4.2, in addition to IQ, we would measure physiological and psychological variables to obtain a more complete picture of the effect of the diet. For completeness we therefore define a *multivariate population* as the set of all possible values of a specified set of variables (measured on the objects of interest). A second category of topics then comes up:

relationships among the variables. Words such as *association* and *correlation* come up in this context. A discussion of these topics begins in Chapter 9.

4.6 Sampling without Replacement

We want to select two patients *at random* from a group of four patients. The same patient cannot be chosen twice. How can this be done? One procedure is to write each name on a slip of paper, put the four slips of paper in a hat, stir the slips of paper, and—without looking—draw out two slips. The patients whose names are on the two slips are then selected. This is known as *sampling without replacement*. (For the procedure to be *fair*, we require that the slips of paper be indistinguishable and well mixed.) The events "outcome on first draw" and "outcome on second draw" are clearly not independent. If patient A is selected in the first draw, she is no longer available for the second draw. Let the patients be labeled *A, B, C*, and *D*. Let the symbol *AB* mean "patient A is selected in the first draw and patient B in the second draw." Write down all the possible outcomes; there are 12 of them as follows:

$$\begin{array}{cccc} AB & BA & CA & DA \\ AC & BC & CB & DB \\ AD & BD & CD & DC \end{array}$$

We define the selection of two patients to be random if each of the 12 outcomes is equally likely, that is, the probability that a particular pair is chosen is 1/12. This definition has intuitive appeal: We could have prepared 12 slips of paper each with one of the 12 pairs recorded and drawnout one slip of paper. If the slip of paper is drawn randomly, the probability is 1/12 that a particular slip will be selected.

One further comment. Suppose that we only want to know which two patients have been selected (i.e., we are not interested in the order). For example, what is the probability that patients C and D are selected? This can happen in two ways: CD or DC. These events are mutually exclusive, so that the required probability is $P[CD \text{ or } DC] = P[CD] + P[DC] = 1/12 + 1/12 = 1/6$.

4.7 Pitfalls in Sampling

It is very important to define the population of interest carefully. Two illustrations of rather subtle pitfalls are Berkson's fallacy and length-biased sampling. *Berkson's fallacy* is discussed in Murphy [1979] as follows: In many studies, hospital records are reviewed or sampled to determine relationships between diseases and/or exposures. Suppose that a review of hospital records is made with respect to two diseases, A and B, which are so severe that they always lead to hospitalization. Let their frequencies in the population at large be p_1 and p_2. Then, assuming independence, the probability of the joint occurrence of the two diseases is $p_1 p_2$. Suppose now that a healthy proportion p_3 of subjects (H) never go to the hospital; that is, $P[H] = p_3$. Now write $\overline{H}$ as that part of the population that will enter a hospital at some time; then $P[\overline{H}] = 1 - p_3$. By the rule of conditional probability, $P[A|\overline{H}] = P[A\overline{H}]/P[\overline{H}] = p_1/(1-p_3)$. Similarly, $P[B|\overline{H}] = p_2/(1-p_3)$ and $P[AB|\overline{H}] = p_1 p_2/(1-p_3)$, and this is not equal to $P[A|\overline{H}]P[B|\overline{H}] = [p_1/(1-p_3)][p_2/(1-p_3)]$, which must be true in order for the two diseases to be unrelated in the hospital population. Now, you can show that $P[AB|\overline{H}] < P[AB]$, and, quoting Murphy:

> The hospital observer will find that they occur together less commonly than would be expected if they were independent. This is known as Berkson's fallacy. It has been a source of embarrassment to many an elegant theory. Thus, cirrhosis of the liver and common cancer are both reasons for admission to the hospital. *A priori*, we would expect them to be less commonly associated in the hospital than in the population at large. In fact, they have been found to be negatively correlated.

Table 4.4 Expected Composition of Visit-Based Sample in a Hypothetical Population

	Type of Patient		
Variable	Hypertensive	Other	Total
Number of patients	200	800	1000
Visits per patient per year	12	1	13
Visits contributed	2400	800	3200
Expected number of patients in a 3% sample of visits	72	24	96
Expected percent of sample	75	25	100

Source: Shepard and Neutra [1977].

(Murphy's book contains an elegant, readable exposition of probability in medicine; it will be worth your while to read it.)

A second pitfall deals with the area of *length-biased sampling*. This means that for a particular sampling scheme, some objects in the population may be more likely to be selected than others. A paper by Shepard and Neutra [1977] illustrates this phenomenon in sampling medical visits. Our discussion is based on that paper. The problem arises when we want to make a statement about a population of patients that can only be identified by a sample of patient visits. Therefore, frequent visitors will be more likely to be selected. Consider the data in Table 4.4, which illustrates that although hypertensive patients make up 20% of the total patient population, a sample based on visits would consist of 75% hypertensive patients and 25% other.

There are other areas, particularly screening procedures in chronic diseases, that are at risk for this type of problem. See Shepard and Neutra [1977] for suggested solutions as well as references to other papers.

4.8 Other Sampling Schemes

In this chapter (and almost all the remainder of the book) we are assuming *simple random sampling*, that is, sampling where every unit in the population is equally likely to end up in the sample, and sampling of different units is independent. A sufficiently large simple random sample will always be representative of the population. This intuitively plausible result is made precise in the mathematical result that the empirical cumulative distribution of the sample approaches the true cumulative distribution of the population as the sample size increases.

There are some important cases where other random sampling strategies are used, trading increased mathematical complexity for lower costs in obtaining the sample. The main techniques are as follows:

1. *Stratified sampling*. Suppose that we sampled 100 births to study low birthweight. We would expect to see about one set of twins on average, but might be unlucky and not sample any. As twins are much more likely to have low birthweight, we would prefer a sampling scheme that fixed the number of twins we observed.
2. *Unequal probability sampling*. In conjunction with stratified sampling, we might want to increase the number of twin births that we examined to more than the 1/90 in the population. We might decide to sample 10 twin births rather than just one.
3. *Cluster sampling*. In a large national survey requiring face-to-face interviews or clinical tests, it is not feasible to use a simple random sample, as this would mean that nearly every person sampled would live in a different town or city. Instead, a number of cities or counties might be sampled and simple random sampling used within the selected geographic regions.

4. *Two-phase sampling*. It is sometimes useful to take a large initial sample and then take a smaller subsample to measure more expensive or difficult variables. The probability of being included in the subsample can then depend on the values of variables measured at the first stage. For example, consider a study of genetic influences on lung cancer. Lung cancer is rare, so it would be sensible to use a stratified (case–control) sampling scheme where an equal number of people with and without lung cancer was sampled. In addition, lung cancer is extremely rare in nonsmokers. If a first-stage sample asked about smoking status it would be possible to ensure that the more expensive genetic information was obtained for a sufficient number of nonsmoker cancer cases as well as smokers with cancer.

These sampling schemes have two important features in common. The sampling scheme is fully known in advance, and the sampling is random (even if not with equal probabilities). These features mean that a valid statistical analysis of the results is possible. Although the sample is not representative of the population, it is unrepresentative in ways that are fully under the control of the analyst. Complex probability samples such as these require different analyses from simple random samples, and not all statistical software will analyze them correctly. The section on Survey Methods of the American Statistical Association maintains a list of statistical software that analyzes complex probability samples. It is linked from the Web appendix to this chapter. There are many books discussing both the statistical analysis of complex surveys and practical considerations involved in sampling, including Levy and Lemeshow [1999], Lehtonen and Pahkinen [1995], and Lohr [1999]. Similar, but more complex issues arise in environmental and ecological sampling, where measurement locations are sampled from a region.

4.9 How to Draw a Random Sample

In Note 4.6 we discussed drawing a random sample without replacement. How can we draw samples with replacement? Simply, of course, the slips could be put back in the hat. However, in some situations we cannot collect the total population to be sampled from, due to its size, for example. One way to sample populations is to use a table of random numbers. Often, these numbers are really *pseudorandom*: They have been generated by a computer. Use of such a table can be illustrated by the following problem: A random sample of 100 patient charts is to be drawn from a hospital record room containing 45,850 charts. Assume that the charts are numbered in some fashion from 1 to 45,850. (It is not necessary that they be numbered consecutively or that the numbers start with 1 and end with 45,850. All that is required is that there is some unique way of numbering each chart.) We enter the random number table randomly by selecting a page and a column on the page at random. Suppose that the first five-digit numbers are

$$06812, \quad 16134, \quad 15195, \quad 84169, \quad \text{and} \quad 41316$$

The first three charts chosen would be chart 06812, 16134, and 15195, in that order. Now what do we do with the 84169? We can skip it and simply go to 41316, realizing that if we follow this procedure, we will have to throw out approximately half of the numbers selected.

A second example: A group of 40 animals is to be assigned at random to one of four treatments A, B, C, and D, with an equal number in each of the treatments. Again, enter the random number table randomly. The first 10-digit numbers between 1 and 40 will be the numbers of the animals assigned to treatment A, the second set of 10-digit numbers to treatment B, the third set to treatment C, and the remaining animals are assigned to treatment D. If a random number reappears in a subsequent treatment, it can simply be omitted. (Why is this reasonable?)

4.10 Algebra of Expectations

In Section 4.3.3 we discuss random variables, distributions, and expectations of random variables. We defined $E(Y) = \sum py$ for a discrete random variable. A similar definition, involving

integrals rather than sums, can be made for continuous random variables. We will now state some rules for working with expectations.

1. If a is a constant, $E(aY) = aE(Y)$.
2. If a and b are constants, $E(aY + b) = aE(Y) + b$.
3. If X and Y are two random variables, $E(X + Y) = E(X) + E(Y)$.
4. If a and b are constants, $E(aX + bY) = E(aX) + E(bY) = aE(X) + bE(Y)$.

You can demonstrate the first three rules by using some simple numbers and calculating their average. For example, let $y_1 = 2$, $y_2 = 4$, and $y_3 = 12$. The average is

$$E(Y) = \frac{1}{3} \times 2 + \frac{1}{3} \times 4 + \frac{1}{3} \times 12 = 6$$

Two additional comments:

1. The second formula makes sense. Suppose that we measure temperature in °C. The average is calculated for a series of readings. The average can be transformed to °F by the formula

$$\text{average in } ^\circ\text{F} = \frac{9}{5} \times \text{average in } ^\circ\text{C} + 32$$

   An alternative approach consists of transforming each original reading to °F and then taking the average. It is intuitive that the two approaches should provide the same answer.
2. It is not true that $E(Y^2) = [E(Y)]^2$. Again, a small example will verify this. Use the same three values ($y_1 = 2$, $y_2 = 4$, and $y_3 = 12$). By definition,

$$E(Y^2) = \frac{2^2 + 4^2 + 12^2}{3} = \frac{4 + 16 + 144}{3} = \frac{164}{3} = 54.\overline{6}$$

   but

$$[E(Y)]^2 = 6^2 = 36$$

   Can you think of a special case where the equation $E(Y^2) = [E(Y)]^2$ is true?

4.11 Bias, Precision, and Accuracy

Using the algebra of expectations, we define a statistic T to be a biased estimate of a parameter τ if $E(T) \neq \tau$. Two typical types of bias are $E(T) = \tau + a$, where a is a constant, called *location bias*; and $E(T) = b\tau$, where b is a positive constant, called *scale bias*. A simple example involves the sample variance, s^2. A more "natural" estimate of σ^2 might be

$$s_*^2 = \frac{\sum (y - \overline{y})^2}{n}$$

This statistic differs from the usual sample variance in division by n rather than $n - 1$. It can be shown (you can try it) that

$$E(s_*^2) = \frac{n - 1}{n}\sigma^2$$

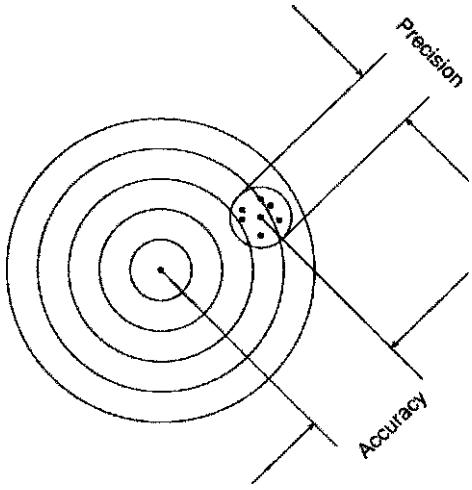


Figure 4.29 Accuracy involves the concept of bias.

Hence, s_*^2 is a biased estimate of σ^2. The statistic s_*^2 can be made unbiased by multiplying s_*^2 by $n/(n-1)$ (see rule 1 in Note 4.10); that is,

$$E\left[\frac{n}{n-1}s_*^2\right] = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2$$

But $n/(n-1)s_*^2 = s^2$, so s^2 rather than s_*^2 is an unbiased estimate of σ^2. We can now discuss precision and accuracy. *Precision* refers to the degree of closeness to each other of a set of values of a variable; *accuracy* refers to the degree of closeness of these values to the quantity (parameter) being measured. Thus, precision is an internal characteristic of a set of data, while accuracy relates the set to an external standard. For example, a thermometer that consistently reads a temperature 5 degrees too high may be very precise but will not be very accurate. A second example of the distribution of hits on a target illustrates these two concepts. Figure 4.29 shows that accuracy involves the concept of bias. Together with Note 4.10, we can now make these concepts more precise. For simplicity we will refer only to location bias.

Suppose that a statistic T estimates a quantity τ in a biased way; $E[T] = \tau + a$. The variance in this case is defined to be $E[T - E(T)]^2$. What is the quantity $E[T-\tau]^2$? This can be written as

$$E[T-\tau]^2 = E[T-(\tau+a)+a]^2 = E[T-E[T]+a]^2$$

$$\underset{\text{(mean square error)}}{E[T-\tau]^2} = \underset{\text{(variance)}}{E[T-E[T]]^2} + \underset{\text{(bias)}}{a^2}$$

The quantity $E[T-\tau]^2$ is called the *mean square error*. If the statistic is unbiased (i.e., $a = 0$), the mean square error is equal to the variance (σ^2).

4.12 Use of the Word Parameter

We have defined *parameter* as a numerical characteristic of a population of values of a variable. One of the basic tasks of statistics is to estimate values of the unknown parameter on the basis of a sample of values of a variable. There are two other uses of this word. Many clinical scientists use *parameter* for *variable*, as in: "We measured the following three parameters: blood pressure,

amount of plaque, and degree of patient satisfaction." You should be aware of this pernicious use and strive valiantly to eradicate it from scientific writing. However, we are not sanguine about its ultimate success. A second incorrect use confuses *parameter* and *perimeter*, as in: "The parameters of the study did not allow us to include patients under 12 years of age." A better choice would have been to use the word *limitations*.

4.13 Significant Digits (continued)

This note continues the discussion of significant digits in Note 3.4. We discussed approximations to a quantity due to arithmetical operations, measurement rounding, and finally, sampling variability. Consider the data on SIDS cases of Example 4.11. The mean birthweight of the 78 cases was 2994 g. The probability was 95% that the interval 2994 ± 178 straddles the unknown quantity of interest: the mean birthweight of the population of SIDS cases. This interval turned out to be 2816–3172 g, although the last digits in the two numbers are not very useful. In this case we have carried enough places so that the rule mentioned in Note 3.4 is not applicable. The biggest source of approximation turns out to be due to sampling. The approximations introduced by the arithmetical operation is minimal; you can verify that if we had carried more places in the intermediate calculations, the final confidence interval would have been 2816–3171 g.

4.14 A Matter of Notation

What do we mean by 18 ± 2.6? In many journals you will find this notation. What does it mean? Is it mean plus or minus the standard deviation, or mean plus or minus the standard error? You may have to read a paper carefully to find out. Both meanings are used and thus need to be specified clearly.

4.15 Formula for the Normal Distribution

The formula for the normal probability density function for a normal random variable Y with mean μ and variance σ^2 is

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]$$

Here, $\pi = 3.14159\ldots$, and e is the base of the natural logarithm, $e = 2.71828\ldots$. A standard normal distribution has $\mu = 0$ and $\sigma = 1$. The formula for the standard normal random variable, Z, is

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$$

Although most statistical packages will do this for you, the heights of the curve can easily be calculated using a hand calculator. By symmetry, only one half of the range of values has to be computed [i.e., $f(z) = f(-z)$]. For completeness in Table 4.5 we give enough points to enable you to graph $f(z)$. Given any normal variable y with mean μ and variance σ^2, you can calculate $f(y)$ by using the relationships

$$Z = \frac{Y-\mu}{\sigma}$$

and plotting the corresponding heights:

$$f(y) = \frac{1}{\sigma} f(z)$$

where Z is defined by the relationship above. For example, suppose that we want to graph the curve for IQ, where we assume that IQ is normal with mean $\mu = 100$ and standard deviation

Table 4.5 Heights of the Standard Normal Curve

z	f(z)	z	f(z)	z	f(z)	z	f(z)	z	f(z)
0.0	0.3989	0.5	0.3521	1.0	0.2420	1.5	0.1295	2.0	0.0540
0.1	0.3970	0.6	0.3332	1.1	0.2179	1.6	0.1109	2.1	0.0440
0.2	0.3910	0.7	0.3123	1.2	0.1942	1.7	0.0940	2.2	0.0355
0.3	0.3814	0.8	0.2897	1.3	0.1714	1.8	0.0790	2.3	0.0283
0.4	0.3683	0.9	0.2661	1.4	0.1497	1.9	0.0656	2.4	0.0224

$\sigma = 15$. What is the height of the curve for an IQ of 109? In this case, $Z = (109-100)/15 = 0.60$ and $f(\text{IQ}) = (1/15)f(z) = (1/15)(0.3332) = 0.0222$. The height for an IQ of 91 is the same.

4.16 Null Hypothesis and Alternative Hypothesis

How do you decide which of two hypotheses is the null and which is the alternative? Sometimes the advice is to make the null hypothesis the hypothesis of "indifference." This is not helpful; indifference is a poor scientific attitude. We have three suggestions: (1) In many situations there is a prevailing view of the science that is accepted; it will continue to be accepted unless "definitive" evidence to the contrary is produced. In this instance the prevailing view would be made operational in the null hypothesis. The null hypothesis is often the "straw man" that we wish to reject. (Philosophers of science tell us that we never prove things conclusively; we can only disprove theories.) (2) An excellent guide is *Occam's razor*, which states: Do not multiply hypotheses beyond necessity. Thus, in comparing a new treatment with a standard treatment, the simpler hypothesis is that the treatments have the same effect. To postulate that the treatments are different requires an additional operation. (3) Frequently, the null hypothesis is one that allows you to calculate the p-value. Thus, if two treatments are assumed the same, we can calculate a p-value for the result observed. If we hypothesize that they are not the same, then we cannot compute a p-value without further specification.

PROBLEMS

4.1 Give examples of populations with the number of elements finite, virtually infinite, potentially infinite, and infinite. Define a sample from each population.

4.2 Give an example from a study in a research area of interest to you that clearly assumes that results are applicable to, as yet, untested subjects.

4.3 Illustrate the concepts of *population, sample, parameter*, and *statistic* by two examples from a research area of your choice.

4.4 In light of the material discussed in this chapter, now review the definitions of statistics presented at the end of Chapter 1, especially the definition by Fisher.

4.5 In Section 4.3.1, probabilities are defined as long-run relative frequencies. How would you interpret the probabilities in the following situations?

(a) The probability of a genetic defect in a child born to a mother over 40 years of age.
(b) The probability of you, the reader, dying of leukemia.
(c) The probability of life on Mars.
(d) The probability of rain tomorrow. What does the meteorologist mean?

4.6 Take a thumbtack and throw it onto a hard surface such as a tabletop. It can come to rest in two ways; label them as follows:

$$\perp = \text{up} = U$$

$$\top = \text{down} = D$$

(a) Guess the probability of U. Record your answer.

(b) Now toss the thumbtack 100 times and calculate the proportion of times the outcome is U. How does this agree with your guess? The observed proportion is an estimate of the probability of U. (Note the implied distinction between *guess* and *estimate*.)

(c) In a class situation, split the class in half. Let each member of the first half of the class toss a thumbtack 10 times and record the outcomes as a histogram: (i) the number of times that U occurs in 10 tosses; and (ii) the proportion of times that U occurs in 10 tosses. Each member of the second half of the class will toss a thumbtack 50 times. Record the outcomes in the same way. Compare the histograms. What conclusions do you draw?

4.7 The estimation of probabilities and the proper combination of probabilities present great difficulties, even to experts. The best we can do in this book is warn you and point you to some references. A good starting point is the paper by Tversky and Kahneman [1974] reprinted in Kahneman et al. [1982]. They categorize the various errors that people make in assessing and working with probabilities. Two examples from this book will test your intuition:

(a) In tossing a coin six times, is the sequence HTHHTT more likely than the sequence HHHHHH? Give your "first impression" answer, then calculate the probability of occurrence of each of the two sequences using the rules stated in the chapter.

(b) The following is taken directly from the book:

> A certain town is served by two hospitals. In the larger hospital, about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50%, sometimes lower. For a period of one year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days? The larger hospital, the smaller hospital, [or were they] about the same (that is, within 5% of each other)?

Which of the rules and results stated in this chapter have guided your answer?

4.8 This problem deals with the *gambler's fallacy*, which states, roughly, that if an event has not happened for a long time, it is "bound to come up." For example, the probability of a head on the fifth toss of a coin is assumed to be greater if the preceding four tosses all resulted in tails than if the preceding four tosses were all heads. This is incorrect.

(a) What statistical property associated with coin tosses is violated by the fallacy?

(b) Give some examples of the occurrence of the fallacy from your own area of research.

(c) Why do you suppose that the fallacy is so ingrained in people?

4.9 Human blood can be classified by the ABO blood grouping system. The four groups are A, B, AB, or O, depending on whether antigens labeled *A* and *B* are present on red blood cells. Hence, the AB blood group is one where both *A* and *B* antigens are present; the O group has none of the antigens present. For three U.S. populations, the following distributions exist:

	Blood Group				
	A	**B**	**AB**	**O**	**Total**
Caucasian	0.44	0.08	0.03	0.45	1.00
American black	0.27	0.20	0.04	0.49	1.00
Chinese	0.22	0.25	0.06	0.47	1.00

For simplicity, consider only the population of American blacks in the following question. The table shows that for a person selected randomly from this population, $P[\text{A}] = 0.27$, $P[\text{B}] = 0.20$, $P[\text{AB}] = 0.04$, and $P[\text{O}] = 0.49$.

(a) Calculate the probability that a person is *not* of blood group A.

(b) Calculate the probability that a person is either A *or* O. Are these mutually exclusive events?

(c) What is the probability that a person carries *A* antigens?

(d) What is the probability that in a marriage both husband and wife are of blood group O? What rule of probability did you use? (What assumption did you need to make?)

4.10 This problem continues with the discussion of ABO blood groups of Problem 4.9. We now consider the black and Caucasian population of the United States. Approximately 20% of the U.S. population is black. This produces the following two-way classification of race and blood type:

	Blood Group				
	A	**B**	**AB**	**O**	**Total**
Caucasian	0.352	0.064	0.024	0.360	0.80
American black	0.054	0.040	0.008	0.098	0.20
Total	0.406	0.104	0.032	0.458	1.00

This table specifies, for example, that the probability is 0.352 that a person selected at random is both Caucasian and blood group A.

(a) Are the events "blood group A" and "Caucasian race" statistically independent?

(b) Are the events "blood group A" and "Caucasian race" mutually exclusive?

(c) Assuming statistical independence, what is the expected probability of the event "blood group A and Caucasian race"?

(d) What is the conditional probability of "blood group A" given that the race is Caucasian?

4.11 The distribution of the Rh factor in a Caucasian population is as follows:

Rh Positive (Rh^+, Rh^+)	Rh Positive (Rh^+, Rh^-)	Rh Negative
0.35	0.48	0.17

Rh^- subjects have two Rh^- genes, while Rh^+ subjects have two Rh^+ genes or one Rh^+ gene and one Rh^- gene. A potential problem occurs when a Rh^+ male mates with an Rh^- female.

(a) Assuming random mating with respect to the Rh factor, what is the probability of an Rh^- female mating with an Rh^+ male?

(b) Since each person contributes one gene to an offspring, what is the probability of Rh incompatibility given such a mating? (Incompatibility occurs when the fetus is Rh^+ and the mother is Rh^-.)

(c) What is the probability of incompatibility in a population of such matings?

4.12 The following data for 20- to 25-year-old white males list four primary causes of death together with a catchall fifth category, and the probability of death within five years:

Cause	Probability
Suicide	0.00126
Homicide	0.00063
Auto accident	0.00581
Leukemia	0.00023
All other causes	0.00788

(a) What is the probability of a white male aged 20 to 25 years dying from *any* cause of death? Which rule did you use to determine this?

(b) Out of 10,000 white males in the 20 to 25 age group, how many deaths would you expect in the next five years? How many for each cause?

(c) Suppose that an insurance company sells insurance to 10,000 white male drivers in the 20 to 25 age bracket. Suppose also that each driver is insured for $100,000 for accidental death. What annual rate would the insurance company have to charge to break even? (Assume a fatal accident rate of 0.00581.) List some reasons why your estimate will be too low or too high.

(d) Given that a white male aged 20 to 25 years has died, what is the most likely cause of death? Assume nothing else is known. Can you explain your statement?

4.13 If $Y \sim N(0,1)$, find

(a) $P[Y \leq 2]$

(b) $P[Y \leq -1]$

(c) $P[Y > 1.645]$

(d) $P[0.4 < Y \leq 1]$

(e) $P[Y \leq -1.96 \text{ or } Y \geq 1.96] = P[|Y| \geq 1.96]$

4.14 If $Y \sim N$ (2,4), find

(a) $P[Y \leq 2]$
(b) $P[Y \leq 0]$
(c) $P[1 \leq Y < 3]$
(d) $P[0.66 < Y \leq 2.54]$

4.15 From the paper by Winkelstein et al. [1975], glucose data for the 45 to 49 age group of California Nisei as presented by percentile are:

Percentile	90	80	70	60	50	40	30	20	10
Glucose (mg/100 mL)	218	193	176	161	148	138	128	116	104

(a) Plot these data on normal probability paper connecting the data points by straight lines. Do the data seem normal?
(b) Estimate the mean and standard deviation from the plot.
(c) Calculate the median and the interquartile range.

4.16 In a sample of size 1000 from a normal distribution, the sample mean $\overline{Y}$ was 15, and the sample variance s^2 was 100.

(a) How many values do you expect to find between 5 and 45?
(b) How many values less than 5 or greater than 45 do you expect to find?

4.17 Plot the data of Table 3.8 on probability paper. Do you think that age at death for these SIDS cases is normally distributed? Can you think of an a priori reason why this variable, age at death, is not likely to be normally distributed? Also make a QQ plot.

4.18 Plot the aflatoxin data of Section 3.2 on normal probability paper by graphing the cumulative proportions against the individual ordered values. Ignoring the last two points on the graph, draw a straight line through the remaining points and estimate the median. On the basis of the graph, would you consider the last three points in the data set *outliers*? Do you expect the arithmetic mean to be larger or smaller than the median? Why?

4.19 Plot the data of Table 3.12 (number of boys per family of eight children) on normal probability paper. Consider the endpoints of the intervals to be 0.5, 1.5, . . . , 8.5. What is your conclusion about the normality of this variable? Estimate the mean and the standard deviation from the graph and compare it with the calculated values of 4.12 and 1.44, respectively.

4.20 The random variable Y has a normal distribution with mean 1.0 and variance 9.0. Samples of size 9 are taken and the sample means, $\overline{Y}$, are calculated.

(a) What is the sampling distribution of $\overline{Y}$?
(b) Calculate $P[1 < \overline{Y} \leq 2.85]$.
(c) Let $W = 4\overline{Y}$. What is the sampling distribution of W?

4.21 The sample mean and standard deviation of a set of temperature observations are 6.1°F and 3.0°F, respectively.

(a) What will be the sample mean and standard deviation of the observations expressed in °C?

(b) Suppose that the original observations are distributed with population mean μ°F and standard deviation σ°F. Suppose also that the sample mean of 6.1°F is based on 25 observations. What is the approximate sampling distribution of the mean? What are its parameters?

4.22 The frequency distributions in Figure 3.10 were based on the following eight sets of frequencies in Table 4.6.

Table 4.6 Sets of Frequencies for Figure 3.10

	Graph Number							
Y	1	2	3	4	5	6	7	8
−1	1	1	8	1	1	14	28	10
−2	2	2	8	3	5	11	14	24
−3	5	5	8	8	9	9	10	14
−4	10	9	8	11	14	6	8	10
−5	16	15	8	14	11	3	7	9
−6	20	24	8	15	8	2	6	7
−7	16	15	8	14	11	3	5	6
−8	10	9	8	11	14	6	4	4
−9	5	5	8	8	9	9	3	2
−10	2	2	8	3	5	11	2	1
−11	1	1	8	1	1	14	1	1
Total	88	88	88	88	88	88	88	88
a_4	3.03	3.20	1.78	2.38	1.97	1.36	12.1	5.78

(The numbers are used to label the graph for purposes of this exercise.) Obtain the probability plots associated with graphs 1 and 6.

4.23 Suppose that the height of male freshmen is normally distributed with mean 69 inches and standard deviation 3 inches. Suppose also (contrary to fact) that such subjects apply and are accepted at a college without regard to their physical stature.

(a) What is the probability that a randomly selected (male) freshman is 6 feet 6 inches (78 inches) or more?

(b) How many such men do you expect to see in a college freshman class of 1000 men?

(c) What is the probability that this class has at least one man who is 78 inches or more tall?

4.24 A normal distribution (e.g., IQ) has mean $\mu = 100$ and standard deviation $\sigma = 15$. Give limits within which 95% of the following would lie:

(a) Individual observations

(b) Means of 4 observations

(c) Means of 16 observations

(d) Means of 100 observations

(e) Plot the width of the interval as a function of the sample size. Join the points with an appropriate freehand line.

(f) Using the graph constructed for part (e), estimate the width of the 95% interval for means of 36 observations.

4.25 If the standard error is the measure of the precision of a sample mean, how many observations must be taken to double the precision of a mean of 10 observations?

4.26 The duration of gestation in healthy humans is approximately 280 days with a standard deviation of 10 days.

(a) What proportion of (healthy) pregnant women will be more than one week "overdue"? Two weeks?

(b) The gestation periods for a set of four women suffering from a particular condition are 240, 250, 265, and 280 days. Is this evidence that a shorter gestation period is associated with the condition?

(c) Is the sample variance consistent with the population variance of $10^2 = 100$? (We assume normality.)

(d) In view of part (c), do you want to reconsider the answer to part (b)? Why or why not?

4.27 The mean height of adult men is approximately 69 inches; the mean height of adult women is approximately 65 inches. The variance of height for both is 4^2 inches. Assume that husband–wife pairs occur without relation to height, and that heights are approximately normally distributed.

(a) What is the sampling distribution of the mean height of a couple? What are its parameters? (The variance of two statistically independent variables is the sum of the variances.)

(b) What proportion of couples is expected to have a mean height that exceeds 70 inches?

(c) In a collection of 200 couples, how many average heights would be expected to exceed 70 inches?

***(d)** In what proportion of couples do you expect the wife to be taller than the husband?

4.28 A pharmaceutical firm claims that a new analgesic drug relieves mild pain under standard conditions for 3 hours, with a standard deviation 1 hour. Sixteen patients are tested under the same conditions and have an average pain relief time of 2.5 hours. The hypothesis that the population mean of this sample is actually 3 hours is to be tested against the hypothesis that the population mean is in fact less than 3 hours; $\alpha = 0.05$.

(a) What is an appropriate test?

(b) Set up the appropriate critical region.

(c) State your conclusion.

(d) Suppose that the sample size is doubled. State precisely how the region where the null hypothesis is not rejected is changed.

***4.29** For Y, from a normal distribution with mean μ and variance σ^2, the variance of $\overline{Y}$, based on n observations, is σ^2/n. It can be shown that the sample median $\tilde{Y}$ in this situation has a variance of approximately $1.57\sigma^2/n$. Assume that the standard error of $\tilde{Y}$ equal to the standard error of $\overline{Y}$ is desired, based on $n = 10$; 20, 50, and 100 observations. Calculate the corresponding sample sizes needed for the median.

***4.30** To determine the strength of a digitalis preparation, a continuous intrajugular perfusion of a tincture is made and the dose required to kill an animal is observed. The lethal dose varies from animal to animal such that its logarithm is normally distributed. One cubic centimeter of the tincture kills 10% of all animals, 2 cm^3 kills 75%. Determine the mean and standard deviation of the distribution of the logarithm of the lethal dose.

4.31 There were 48 SIDS cases in King County, Washington, during the years 1974 and 1975. The birthweights (in grams) of these 48 cases were:

2466	3941	2807	3118	2098	3175	3515
3317	3742	3062	3033	2353	2013	3515
3260	2892	1616	4423	3572	2750	2807
2807	3005	3374	2722	2495	3459	3374
1984	2495	3062	3005	2608	2353	4394
3232	2013	2551	2977	3118	2637	1503
2438	2722	2863	2013	3232	2863	

(a) Calculate the sample mean and standard deviation for this set.

(b) Construct a 95% confidence interval for the population mean birthweight assuming that the population standard deviation is 800 g. Does this confidence interval include the mean birthweight of 3300 g for normal children?

(c) Calculate the p-value of the sample mean observed, assuming that the population mean is 3300 g and the population standard deviation is 800 g. Do the results of this part and part (b) agree?

(d) Is the sample standard deviation consistent with a population standard deviation of 800? Carry out a hypothesis test comparing the sample variance with population variance $(800)^2$. The critical values for a chi-square variable with 47 degrees of freedom are as follows:

$$\chi^2_{0.025} = 29.96, \qquad \chi^2_{0.975} = 67.82$$

(e) Set up a 95% confidence interval for the population standard deviation. Do this by first constructing a 95% confidence interval for the population variance and then taking square roots.

4.32 In a sample of 100 patients who had been hospitalized recently, the average cost of hospitalization was \$5000, the median cost was \$4000, and the modal cost was \$2500.

(a) What was the total cost of hospitalization for all 100 patients? Which statistic did you use? Why?

(b) List one practical use for *each* of the three statistics.

(c) Considering the ordering of the values of the statistics, what can you say about the distribution of the raw data? Will it be skewed or symmetric? If skewed, which way will the skewness be?

4.33 For Example 4.8, as discussed in Section 4.6.2:

(a) Calculate the probability of a Type II error and the power if α is fixed at 0.05.

(b) Calculate the power associated with a one-tailed test.

(c) What is the price paid for the increased power in part (b)?

4.34 The theory of hypothesis testing can be used to determine statistical characteristics of laboratory tests, keeping in mind the provision mentioned in connection with Example 4.6. Suppose that albumin has a normal (Gaussian) distribution in a healthy population with mean $\mu = 3.75$ mg per 100 mL and $\sigma = 0.50$ mg per 100 mL. The normal range of values will be defined as $\mu \pm 1.96\sigma$, so that values outside these limits will be classified as "abnormal." Patients with advanced chronic liver disease have reduced albumin levels; suppose that the mean for patients from this population is 2.5 mg per 100 mL and the standard deviation is the same as that of the normal population.

(a) What are the critical values for the rejection region? (Here we work with an individual patient, $n = 1$.)

(b) What proportion of patients with advanced chronic liver disease (ACLD) will have "normal" albumin test levels?

(c) What is the probability that a patient with ACLD will be classified correctly on a test of albumin level?

(d) Give an interpretation of Type I error, Type II error, and power for this example.

(e) Suppose we consider only low albumin levels to be "abnormal." We want the same Type I error as above. What is the critical value now?

(f) In part (e), what is the associated power?

4.35 This problem illustrates the power of probability theory.

(a) Two SIDS infants are selected at random from a population of SIDS infants. We note their birthweights. What is the probability that both birthweights are (1) below the population median; (2) above the population median; (3) straddle the population median? The last interval is a nonparametric confidence interval.

(b) Do the same as in part (a) for four SIDS infants. Do you see the pattern?

(c) How many infants are needed to have interval 3 in part (a) have probability greater than 0.95?

REFERENCES

Barnett, V. [1999]. *Comparative Statistical Inference*. Wiley, Chichester, West Sussex, England.

Berkow, R. (ed.) [1999]. *The Merck Manual of Diagnosis and Therapy*, 17th ed. Merck, Rahway, NJ.

Berry, D. A. [1996]. *Statistics: A Bayesian Perspective*. Duxbury Press, North Scituate, MA.

Carlin, B. P., and Louis, T. A. [2000]. *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed. CRC Press, Boca Raton, FL.

Elveback, L. R., Guillier, L., and Keating, F. R., Jr. [1970]. Health, normality and the ghost of Gauss. *Journal of the American Medical Association*, **211**: 69–75.

Fisher, R. A. [1956]. *Statistical Methods and Scientific Inference*. Oliver & Boyd, London.

Fleiss, J. L., Levin, B., and Park, M. C. [2003]. *Statistical Methods for Rates and Proportions*, 3rd ed. Wiley, New York.

Galton, F. [1889]. *Natural Inheritance*. Macmillan, London.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. A. [1995]. *Bayesian Data Analysis.* CRC Press, Boca Raton, FL.

Golubjatnikov, R., Paskey, T., and Inhorn, S. L. [1972]. Serum cholesterol levels of Mexican and Wisconsin school children. *American Journal of Epidemiology*, **96**: 36–39.

Hacking, I. [1965]. *Logic of Statistical Inference*. Cambridge University Press, London.

Hagerup, L., Hansen, P. F., and Skov, F. [1972]. Serum cholesterol, serum-triglyceride and ABO blood groups in a population of 50-year-old Danish men and women. *American Journal of Epidemiology*, **95**: 99–103.

Kahneman, D., Slovic, P., and Tversky, A. (eds.) [1982]. *Judgement under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge.

Kato, H., Tillotson, J., Nichaman, M. Z., Rhoads, G. G., and Hamilton, H. B. [1973]. Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: serum lipids and diet. *American Journal of Epidemiology*, **97**: 372–385.

Kesteloot, H., and van Houte, O. [1973]. An epidemiologic study of blood pressure in a large male population. *American Journal of Epidemiology*, **99**: 14–29.

Kruskal, W., and Mosteller, F. [1979a]. Representative sampling I: non-scientific literature. *International Statistical Review*, **47**: 13–24.

Kruskal, W., and Mosteller, F. [1979b]. Representative sampling II: scientific literature excluding statistics. *International Statistical Review*, **47**: 111–127.

Kruskal, W., and Mosteller, F. [1979c]. Representative sampling III: the current statistical literature. *International Statistical Review*, **47**: 245–265.

Lehtonen, R., and Pahkinen, E. J. [1995]. *Practical Methods for Design and Analysis of Complex Surveys*. Wiley, New York.

Levy, P. S., and Lemeshow S. [1999]. *Sampling of Populations: Methods and Applications*, 3rd Ed. Wiley, New York.

Lohr, S. [1999]. *Sample: Design and Analysis*. Duxbury Press, Pacific Grove, CA.

Moore, D. S. [2001]. *Statistics: Concepts and Controversies*, 5th ed. W. H. Freeman, New York.

Murphy, E. A. [1979]. *Biostatistics in Medicine*. Johns Hopkins University Press, Baltimore.

Runes, D. D. [1959]. *Dictionary of Philosophy*. Littlefield, Adams, Ames, IA.

Rushforth, N. B., Bennet, P. H., Steinberg, A. G., Burch, T. A., and Miller, M. [1971]. Diabetes in the Pima Indians: evidence of bimodality in glucose tolerance distribution. *Diabetes*, **20**: 756–765. Copyright © 1971 by the American Diabetic Association.

Savage, I. R. [1968]. *Statistics: Uncertainty and Behavior*. Houghton Mifflin, Boston.

Shepard, D. S., and Neutra, R. [1977]. Pitfalls in sampling medical visits. *American Journal of Public Health*, **67**: 743–750. Copyright © by the American Public Health Association.

Tversky, A., and Kahneman, D. [1974]. Judgment under uncertainty: heuristics and biases. *Science*, **185**: 1124–1131. Copyright © by the AAAS.

Winkelstein, W. Jr., Kazan, A., Kato, H., and Sachs, S. T. [1975]. Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: blood pressure distributions. *American Journal of Epidemiology*, **102**: 502–513.

Zervas, M., Hamacher, H., Holmes, O., and Rieder, S. V. [1970]. Normal laboratory values. *New England Journal of Medicine*, **283**: 1276–1285.