Biostatistics

A Methodology for the Health Sciences

Second Edition

GERALD VAN BELLE
LLOYD D. FISHER
PATRICK J. HEAGERTY
THOMAS LUMLEY

Department of Biostatistics and
Department of Environmental and
Occupational Health Sciences
University of Washington
Seattle, Washington

A JOHN WILEY & SONS, INC., PUBLICATION

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

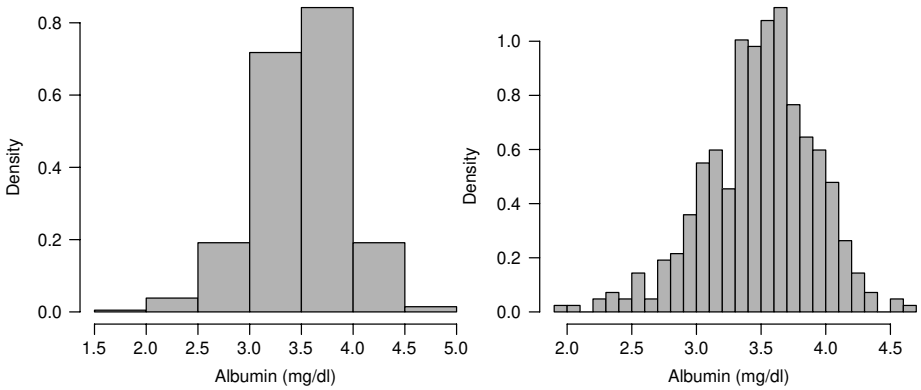

Figure 3.1 Histograms of serum albumin concentration in 418 PBC patients, using two different sets of categories.

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

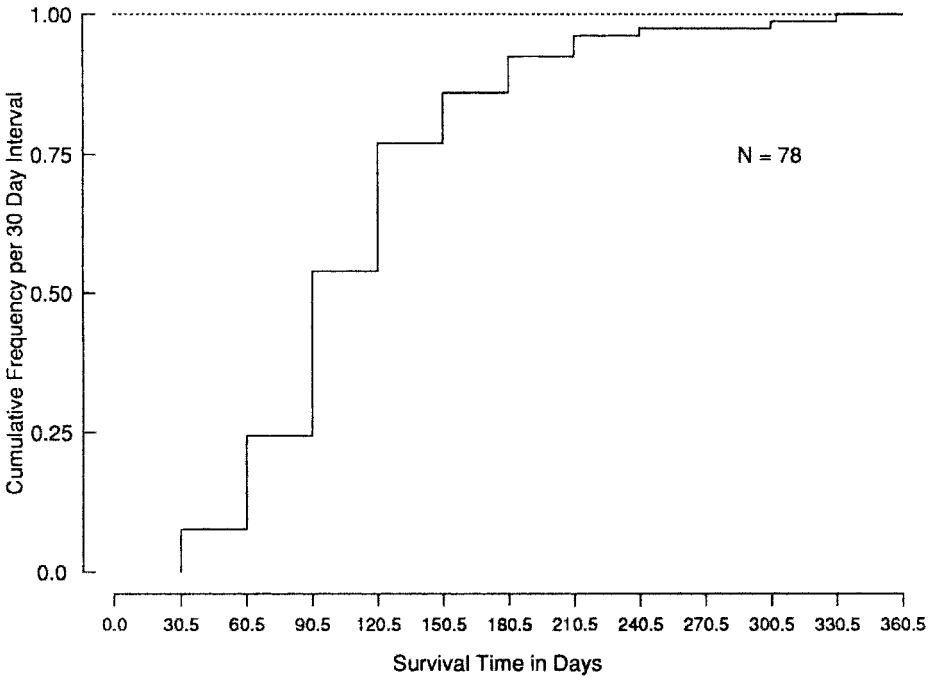

Figure 3.2 Empirical cumulative distribution of SIDS deaths.

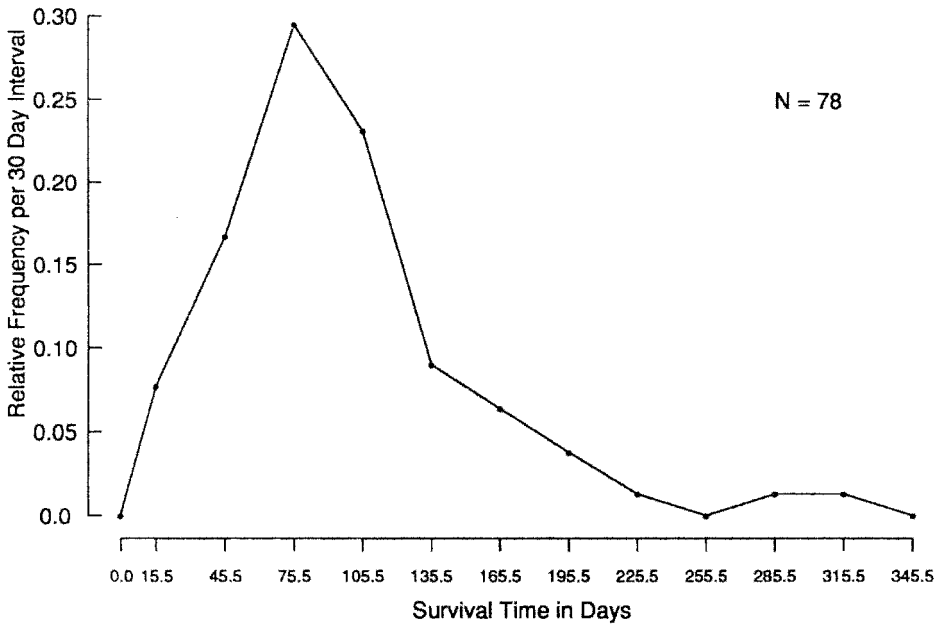

Figure 3.3 Frequency polygon of SIDS deaths.

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.

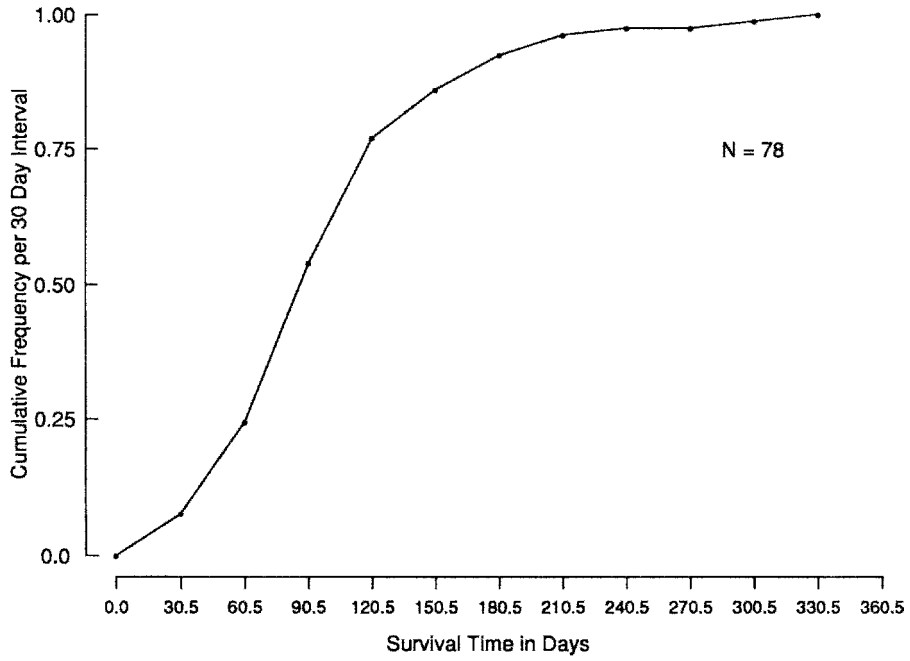

Figure 3.4 Cumulative frequency polygon of SIDS deaths.

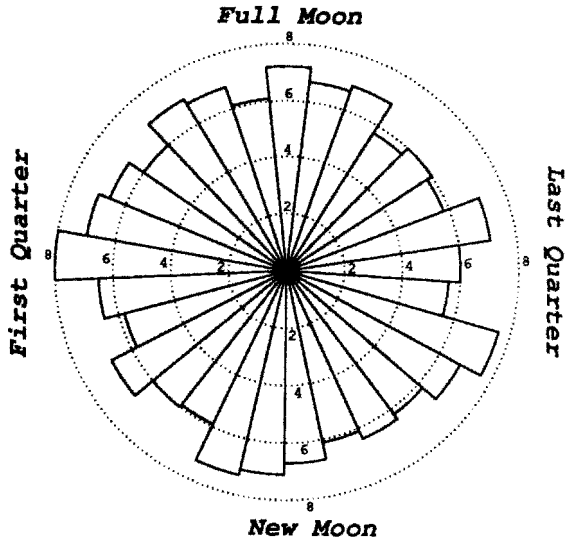

Figure 3.5 Average number of births per day over a 29-day lunar cycle. (Data from Schwab [1975].)

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.

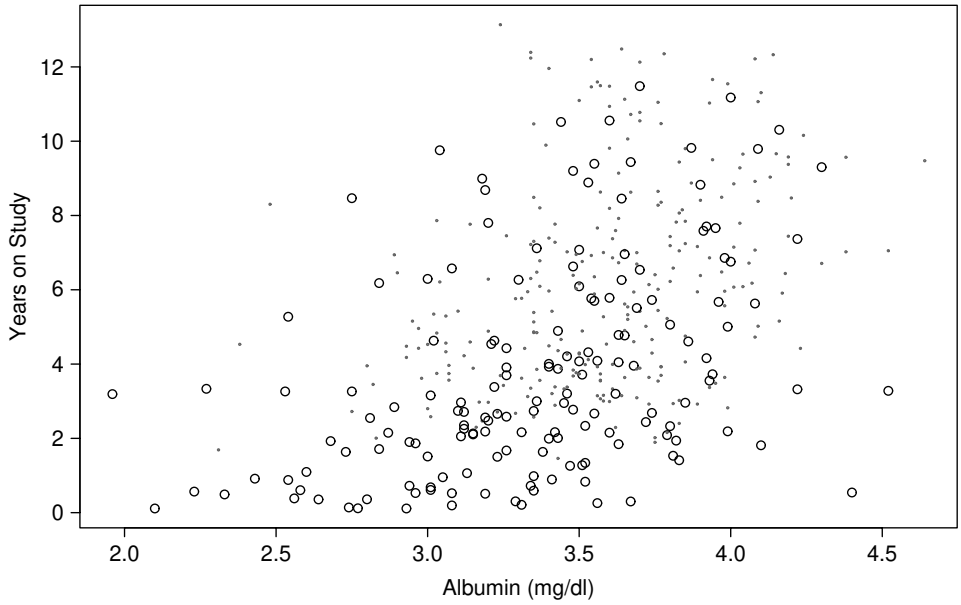

Figure 3.6 Survival time in primary biliary cirrhosis by serum albumin concentrations. Large circles are deaths, small circles are patients alive at last contact. (Data from Fleming and Harrington [1991].)

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.

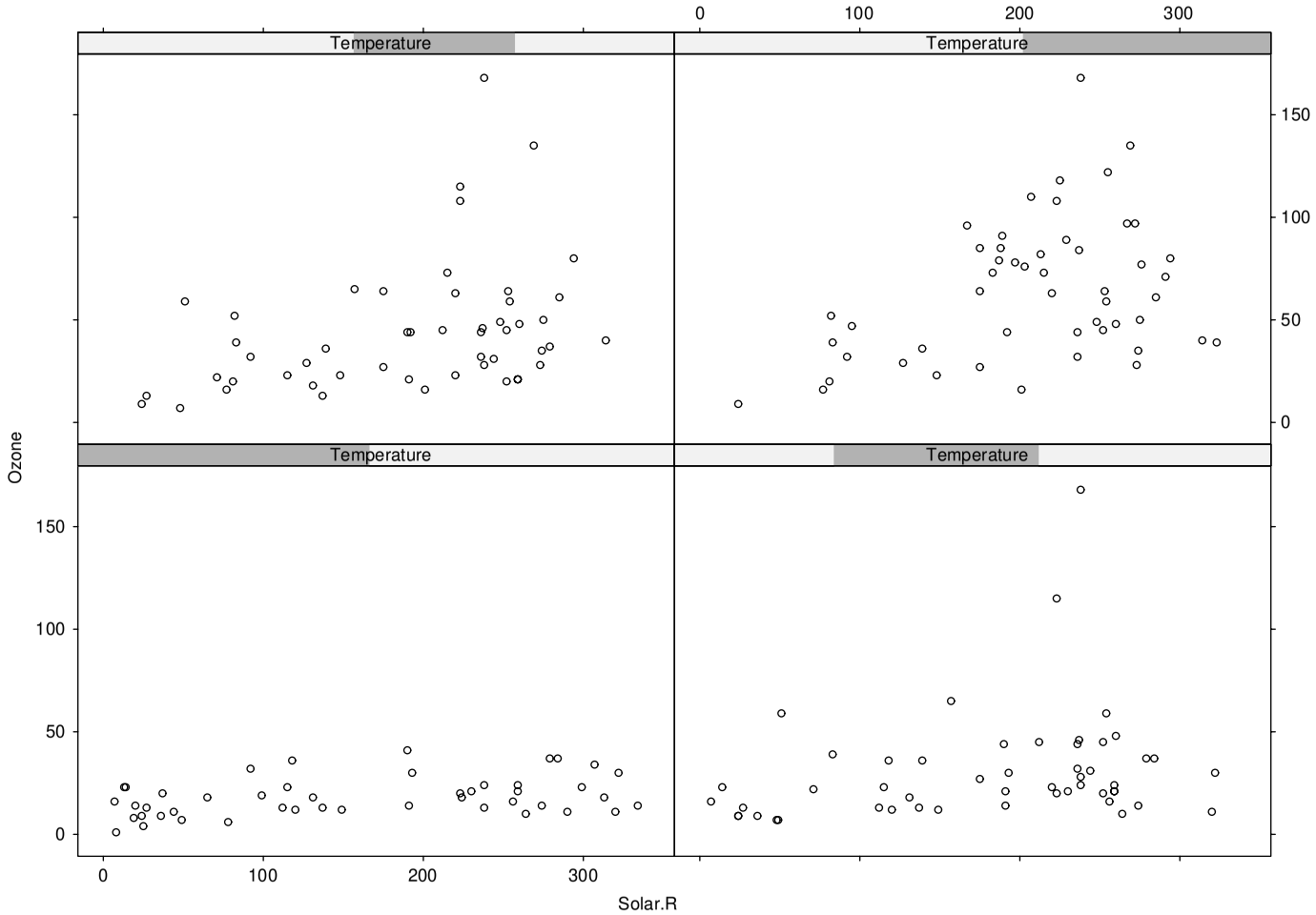

Figure 3.7 Ozone concentration by solar radiation intensity in New York, May–September 1973, conditioned on temperature. (From R Foundation [2002].)

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.

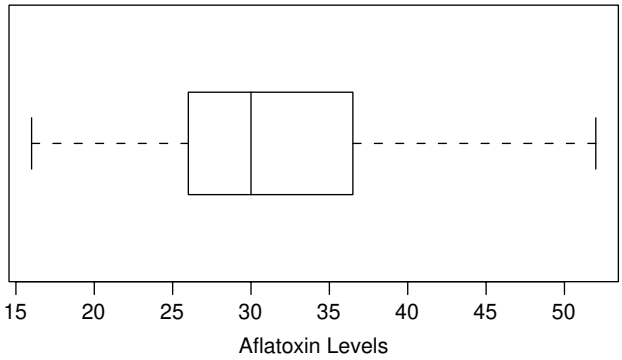

Figure 3.8 Box plot.

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.

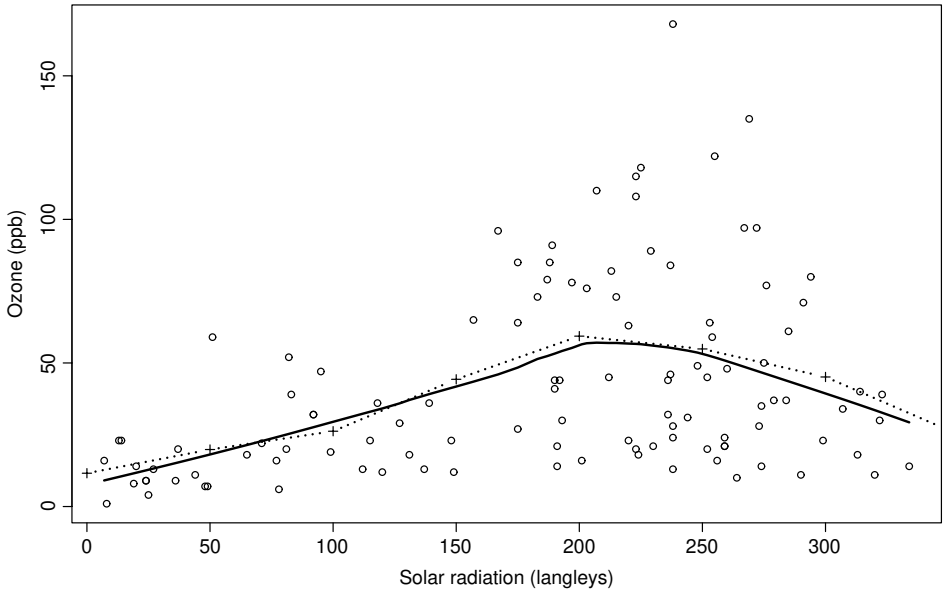

Figure 3.9 Ozone and solar radiation in New York during the summer of 1973, with scatter plot smoothers.

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

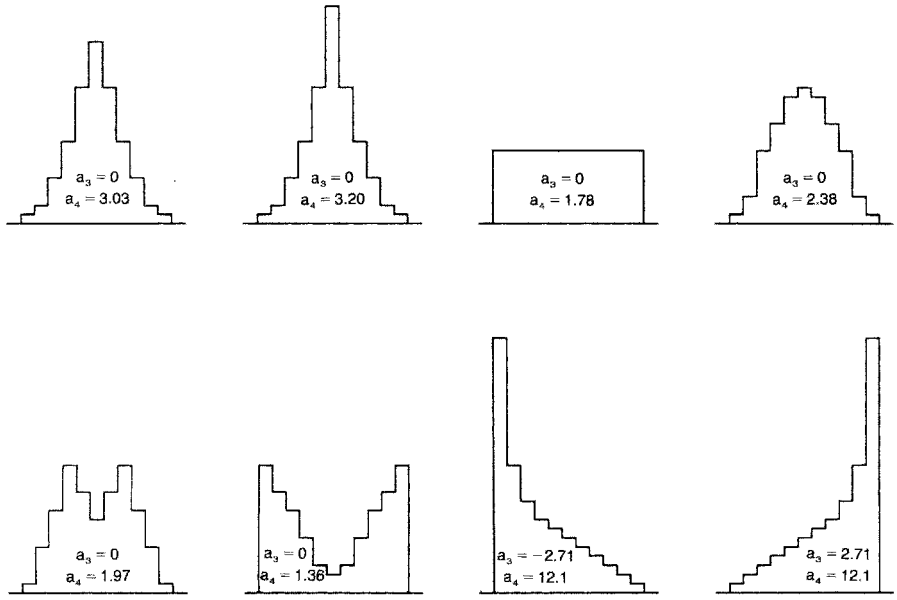

Figure 3.10 Values of skewness (a_3) and kurtosis (a_4) for selected data configurations.

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 $°F = 1.8 \times °C + 32$. 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.