Descriptive Methods

3

Chapter Outline

The Scotsman William Playfair is credited with being the first to publish graphics such as the bar chart, line graph, and pie charts that are commonly used in statistics today (Kennedy 1984). This chapter focuses on the summarization and display of data using the techniques Playfair first published along with several other useful procedures. We will rely on both numerical and pictorial procedures to describe data. We use charts and other procedures because they may capture features in the data that are often overlooked when using summary numerical measures alone. Although the utility of graphical methods has been well established and can be seen in all walks of life, the visual representation of data was not always common practice. According to Galvin Kennedy, the first 50 volumes of the *Journal of the Royal Statistical Society* contain only 14 charts, with the first one appearing in 1841.

3.1 Introduction to Descriptive Methods

The data we use in this section come from the Digitalis Investigation Group (DIG) trial (DIG 1997). The DIG trial was a multicenter trial with 302 clinical centers in the United States and Canada participating. (Its study design features will be discussed in a later chapter.) The purpose of the trial was to examine the safety and efficacy of Digoxin in treating patients with congestive heart failure in sinus rhythm. Subjects were recruited from those who had heart failure with a left ventricular ejection fraction of 0.45 or less and with normal sinus rhythm. The primary endpoint in the trial was to evaluate the effects of Digoxin on mortality from any cause over a three- to five-year period. Basic demographic and physiological data were recoded at the entry to the trial, and outcome related data were recorded during the course of the trial. The data presented in this chapter consists of baseline and outcome variables from 200 patients (100 on Digoxin treatment and 100 on placebo) randomly selected from the multicenter trial dataset.*

*This trial was conducted and supported by the National Heart, Lung, and Blood Institute in cooperation with the study investigators. The NHLBI has employed statistical methods to make components of the full datasets anonymous in order to provide selected data as a teaching resource. Therefore, the data are inappropriate for any publication purposes. The authors would like to thank the NHLBI, study investigators, and study participants for providing the data.

Table 3.1 Digoxin clinical trial data for 40 participants.

ID	Treatment[a]	Age[b]	Race[c]	Sex[d]	Body Mass Index[e]	Serum Creatinine[f]	Systolic Blood Pressure[g]
4995	0	55	1	1	19.435	1.600	150
2312	0	78	2	1	22.503	2.682	104
896	0	50	1	1	27.406	1.300	140
3103	0	60	1	1	29.867	1.091	140
538	1	31	1	1	27.025	1.159	120
1426	0	70	1	1	19.040	1.250	150
4787	1	46	1	1	28.662	1.307	140
5663	0	59	2	1	27.406	1.705	152
1109	0	68	1	2	27.532	1.534	144
666	0	65	1	1	28.058	2.000	120
2705	1	66	1	2	28.762	0.900	150
5668	0	74	1	1	29.024	1.227	116
999	1	47	1	2	30.506	1.386	120
1653	1	63	1	1	28.399	1.100	105
764	1	63	2	2	28.731	0.900	122
3640	0	79	1	1	18.957	2.239	150
1254	1	73	1	1	26.545	1.300	144
2217	1	65	1	1	23.739	1.614	170
4326	0	65	1	1	29.340	1.200	170
5750	1	76	1	1	39.837	1.455	140
6396	0	83	1	1	26.156	1.489	116
2289	0	76	1	1	30.586	1.700	130
1322	1	45	1	2	43.269	0.900	115
4554	1	58	1	2	28.192	1.352	130
6719	1	34	1	1	20.426	1.886	116
1954	1	77	1	1	26.545	1.307	140
5001	1	70	1	1	19.044	1.200	110
1882	0	50	1	1	25.712	1.034	140
5368	1	38	1	1	30.853	0.900	134
787	0	58	2	2	27.369	0.909	100
4375	0	61	1	1	32.079	1.273	128
5753	1	75	1	1	37.590	1.300	138
6745	0	45	1	1	22.850	1.398	130
6646	0	61	1	1	27.718	1.659	128
5407	1	50	1	2	24.176	1.000	130
4181	0	44	2	2	26.370	1.148	124
3403	0	55	1	2	21.790	1.170	130
2439	1	49	1	1	15.204	1.307	140
4055	0	71	1	1	22.229	1.261	100
3641	0	64	1	1	21.228	0.900	130

[a]Treatment group (0 = on placebo; 1 = on Digoxin)
[b]Age in years
[c]Race (1 = White; 2 = Nonwhite)
[d]Sex (1 = Male; 2 = Female)
[e]Body mass index (weight in kilograms/height in meters squared)
[f]Serum creatinine (mg/dL)
[g]Systolic blood pressure (mmHg)

We refer to this working dataset as DIG200 in this book. The DIG200 dataset is reduced to create a smaller dataset including 7 baseline variables from 40 patients referred to as DIG40. Table 3.1 displays the DIG40 dataset. Both data files are available on the supplementary website.

3.2 Tabular and Graphical Presentation of Data

The one- and two-way frequency tables and several types of figures (line graphs, bar charts, histograms, stem-and-leaf plots, scatter plots, and box plots) that aid the description of data are introduced in this section.

3.2.1 Frequency Tables

A *one-way frequency table* shows the results of the tabulation of observations at each level of a variable. In Table 3.2, we show one-way tabulations of sex and race for the 40 patients shown in Table 3.1. Three quarters of the patients are males, and over 87 percent of the patients are whites.

Table 3.2 Frequencies of sex and race for 40 patients in DIG40.

Sex	Number of Patients	Percentage	Race	Number of Patients	Percentage
Male	30	75.0	White	35	87.5
Female	10	25.0	Nonwhite	5	12.5
Total	40	100.0	Total	40	100.0

The variables used in frequency tables may be nominal, ordinal, or continuous. When continuous variables are used in tables, their values are often grouped into categories. For example, age is often categorized into 10-year intervals. Table 3.3 shows the frequencies of age groups for the 40 patients in Table 3.1. More than one half of the patients are 60 and over. Note that the sum of percents should add up to 100 percent, although a small allowance is made for rounding. It is also worth noting that the title of the table should contain sufficient information to allow the reader to understand the table.

Table 3.3 Frequency of age groups for 40 patients in DIG40.

Age Groups	Number of Patients	Percentage
Under 40	3	7.5
40–49	6	15.0
50–59	8	20.0
60–69	11	27.5
70–79	12	30.0
Total	40	100.0

Table 3.4 Cross-tabulation of body mass index and sex for 40 patients in DIG40 with column percentages in parentheses.

	Sex		
Body Mass Index	Male	Female	Total
Under 18.5 (underweight)	1 (3.3%)	0 (0.0%)	1 (2.5%)
18.5–24.9 (normal)	10 (33.3%)	2 (20.0%)	12 (30.0%)
25.0–29.9 (overweight)	14 (46.7%)	6 (60.0%)	20 (50.0%)
30.0 & over (obese)	5 (16.7%)	2 (20.0%)	7 (17.5%)
Total	30	10	40

Two-way frequency tables, formed by the *cross-tabulation* of two variables, are usually more interesting than one-way tables because they show the relationship between the variables. Table 3.4 shows the relationship between sex and body mass index where BMI has been grouped into underweight (BMI < 18.5), normal ($18.5 \leq$ BMI < 25), overweight ($25 \leq$ BMI < 30), and obese (BMI ≥ 30). The body mass index is calculated as weight in kilograms divided by height in meters squared. There are higher percentages of females in the overweight and obese categories than those found for males, but these calculations are based on very small sample sizes.

In forming groups from continuous variables, we should not allow the data to guide us. We should use our knowledge of the subject matter, and not use the data, in determining the groupings. If we use the data to guide us, it is easy to obtain apparent differences that are not real but only artifacts. When we encounter categories with no or few observations, we can reduce the number of categories by combining or collapsing these categories into the adjacent categories. For example, in Table 3.4 the number of obesity levels can be reduced to 3 by combining the underweight and normal categories. There is no need to subdivide the overweight category, even though one-half of observations are in this category. Computer packages can be used to categorize continuous variables (recoding) and to tabulate the data in one- or two-way tables (see **Program Note 3.1** on the website).

There are several ways of displaying the data in a tabular format. In Tables 3.2, 3.3, and 3.4 we showed both numbers and percentages, but it is not necessary to show both in a summary table for presentation in journal articles. Table 3.5 presents basic patient characteristics for 200 patients from the DIG200 data set. Note that the total number (n) relevant to the percentages of each variable is presented at the top of the column and percentages alone are presented, leaving out the frequencies. The frequencies can be calculated from the percentages and the total number.

Table 3.5 Basic patient characteristics at baseline in the Digoxin clinical trial based on 200 patents in DIG200.

Characteristics		Percentage (n = 200)
Sex	Male	73.0
	Female	27.0
Race	White	86.5
	Nonwhite	13.5
Age	Under 40	3.5
	40–49	11.5
	50–59	25.0
	60–69	33.0
	70 & over	26.0
Body mass index	Underweight (< 18.5)	1.5
	Normal (18.5–24.9)	37.5
	Overweight (25–29.9)	42.5
	Obese (≥ 30)	18.5

Other data besides frequencies can be presented in a tabular format. For example, Table 3.6 shows the health expenditures of three nations as a percentage of gross domestic products (GDP) over time (NCHS 2004, Table 115). Health expenditures as a percentage of GDP are increasing much more rapidly in the United States than either Canada or United Kingdom.

3.2.2 Line Graphs

A *line graph* can be used to show the value of a variable over time. The values of the variable are given on the vertical axis, and the horizontal axis is the time variable. Figure 3.1 shows three line graphs for the data shown in Table 3.6. These line graphs also show the rapid increase in health expenditures in the United States compared with those of two other counties with national health plans. The trends are immediately clear in the line graphs, whereas one has to study Table 3.6 before the same trends are recognized.

Table 3.6 Health expenditures as a percentage of gross domestic product over time.

Year	Canada	United Kingdom	United States
1960	5.4	3.9	5.1
1965	5.6	4.1	6.0
1970	7.0	4.5	7.0
1975	7.0	5.5	8.4
1980	7.1	5.6	8.8
1985	8.0	6.0	10.6
1990	9.0	6.0	12.0
1995	9.2	7.0	13.4
2000	9.2	7.3	13.3

Source: National Center for Health Statistics, 2004, Table 115

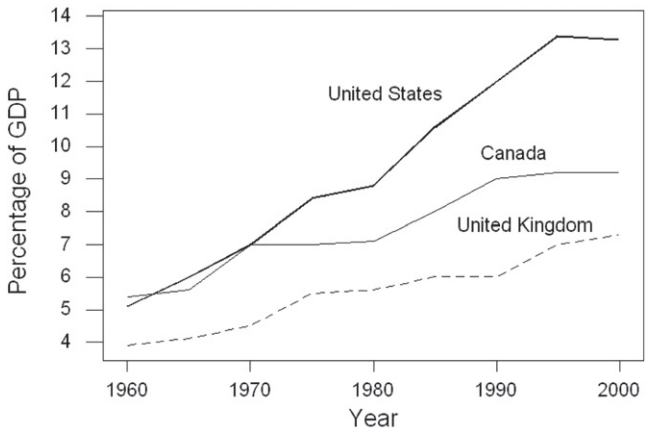


Figure 3.1 Line graph: Health expenditures as percentage of GDP for Canada, United Kingdom, and United States.

It is possible to give different impressions about the data by shortening or lengthening the horizontal and vertical axes or by including only a portion of an axis. In creating and studying line graphs, one must be aware of the scales used for horizontal and vertical axes. For example, with numbers that are extremely variable over time, a logarithmic transformation (discussed later) of the variable on the vertical axis is frequently used to allow the line graph to fit on a page.

Example 3.1

It is well accepted that blood pressure varies from day to day or even minute to minute (Armitage and Rose 1966). We present the following data on systolic blood pressure measurements for three patients taken three times a day over a three-day period in two different ways in Figure 3.2:

	Day 1			Day 2			Day 3		
Patient	8am	2pm	8pm	8am	2pm	8pm	8am	2pm	8pm
1	110	140	100	115	130	110	105	137	105
2	112	138	105	105	133	120	110	128	100
3	105	135	120	110	130	105	115	135	110

In the top graph we show the change in a patient's systolic blood pressure over the three time points for each day without connecting between days. From the line graph, we notice that the individual under study has peaks in his systolic blood pressure, and the peaks occur consistently at the same time point, giving us reason to believe that there may be a circadian rhythm in blood pressure.

Depending on the time of day when the blood pressure is measured, the patient's hypertension status may be defined differently because most cutoff points for stages of hypertension are based on fixed values that ignore the time of day. In the bottom graph the lines are connected between days, with the recognition that the time interval between days is twice as large as the measurement intervals during the day. The general trend shown in the top graph remains, but the consistency between days is less evident. Another measurement at 2am could have established the consistency between days.

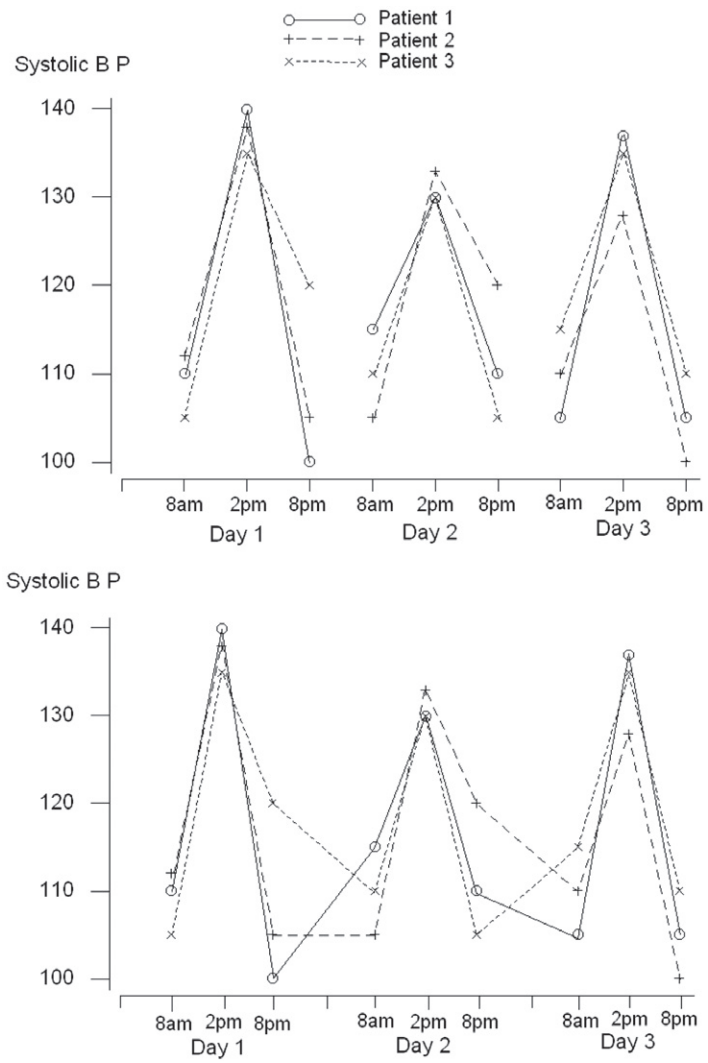


Figure 3.2 Plot of systolic blood pressure taken three times a day over a three-day period for three patients.

Example 3.2

It is possible to represent different variables in the same figure, as Figure 3.3 shows. The right vertical axis is used for lead emissions and the left vertical axis for sulfur oxide emissions. Both pollutants are decreasing, but the decrease in lead emissions is quite dramatic, from approximately 200×10^3 metric tons in 1970 to only about 8×10^3 metric tons in 1988. During this same period, sulfur oxide emissions decreased from about 20×10^6 metric tons to 21×10^6 metric tons. The decrease in the lead emissions is partially related to the use of unleaded gasoline, which was introduced during the 1970s.

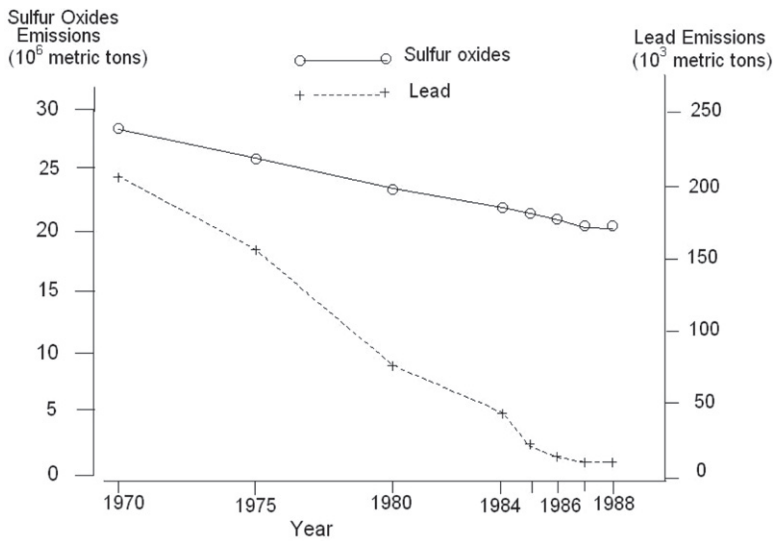


Figure 3.3 Line graph of sulfur oxides and lead emissions in the United States.
Source: National Center for Health Statistics, 1991, Table 64

3.2.3 Bar Charts

A bar chart provides a picture of data that could also be reasonably displayed in tabular format. Bar charts can be created for nominal, ordinal, or continuous data, although they are most frequently used with nominal data. If used with continuous data, the chart could be called a histogram instead of a bar chart. The bar chart can show the number or proportion of people by levels of a nominal or ordinal variable.

Example 3.3

The actual enrollment of individuals in health maintenance organizations (HMOs) in the United States was 9.1 million in 1980, 33.0 million in 1990, and 80.9 million in 2000 (NCHS 2004, Table 134). This information is displayed in Figure 3.4 using a bar chart. The numbers of people enrolled in HMOs in the United States is shown by year (ordinal variable). This bar chart makes it very clear that there has been explosive growth in HMO enrollment. The actual numbers document this growth, but it is more dramatic in the visual presentation.

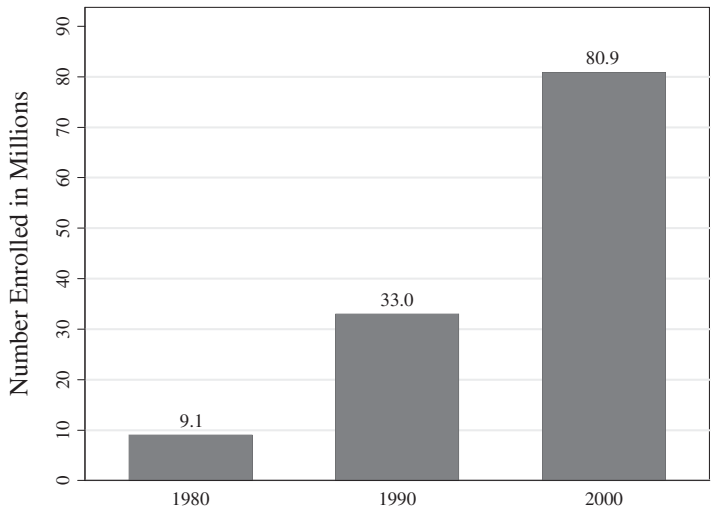


Figure 3.4 Bar chart of the number of persons (in millions) enrolled in Health Maintenance Organizations by year.
Source: National Center for Health Statistics (NCHS), 2004, Table 134

In bar charts, the length of the bar shows the number of observations or the value of the variable of interest for each level of the nominal or ordinal variable. The widths of the bar are the same for all the levels of the nominal or ordinal variable, and the width has no meaning. The levels of the nominal or ordinal variable are usually separated by several spaces that make it easier to view the data. The bars are usually presented vertically, although they could also be presented horizontally.

Bar charts can also be used to present more complicated data. The tabulated data in two- or three-way tables can be presented in bar chart format. For instance, the data in a 2×5 table (e.g., the status of diabetes — yes or no — by five age groups) can be presented by five bars with the length of each bar representing the proportion of people in the age group with diabetes, as shown in Figure 3.5.

When both variables in a two-way table have more than two levels each, we can use a segmented bar chart. Example 3.4 illustrates the presentation of data in a 3×4 table using a segmented bar chart. Data in a three-way table can be presented by a clustered

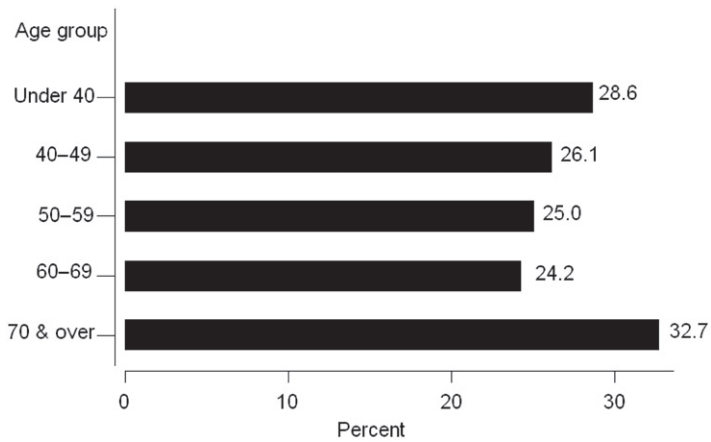


Figure 3.5 Bar chart showing proportion of people in each age group with diabetes, DIG200.

bar chart. Example 3.5 shows a presentation of data in a $2 \times 3 \times 4$ table using a clustered bar chart.

Example 3.4

To examine the relationship between obesity and age, DIG200 data are tabulated in a 3×4 table:

	Age Group (column percent in parentheses)			
Obesity level	**Under 50**	**50–59**	**60–69**	**70 & over**
Normal or underweight (BMI < 25)	11 (36.6)	22 (42.3)	26 (39.4)	19 (36.5)
Overweight (25 ≤ BMI < 30)	11 (36.6)	23 (44.2)	30 (45.5)	21 (40.4)
Obese (BMI ≥ 30)	8 (26.7)	7 (13.5)	10 (15.2)	12 (23.1)
Total	30	52	66	52

The data in this table are presented in Figure 3.6 using two types of segmented bar charts. The first segmented bar chart is based on frequencies (top figure), and the second segmented bar chart is based on percentages (bottom figure). The top figure shows that nearly two-thirds of obese patients are in the 60 and over age groups. The bottom figure shows that the obesity is more prevalent in the under 50 age group.

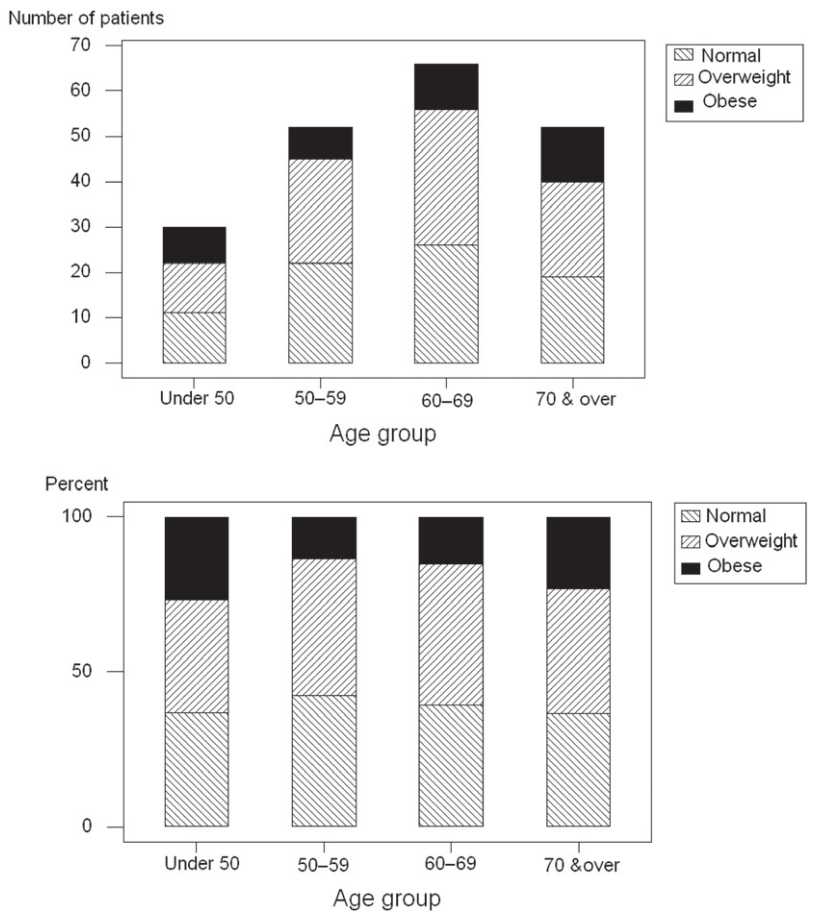


Figure 3.6 Segmented bar charts for levels of obesity by age group, DIG200 (the normal category includes underweight as well as normal)

Example 3.5

To examine how the prevalence of diabetes differs by the level of obesity and age, the DIG200 data are tabulated in a $2 \times 3 \times 4$ table. The results are presented in Figure 3.7 using a clustered bar chart. Three bars depicting the percent of diabetes in three levels of obesity are clustered in each of the age categories. It is interesting to note that the level of obesity is closely associated with the prevalence of diabetes in all age groups except for the 70 and over age group.

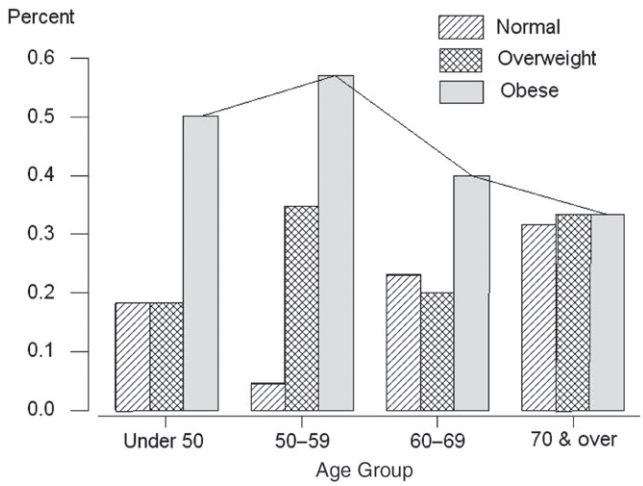


Figure 3.7 Clustered bar charts showing proportion of people in each level of obesity (the normal category includes underweight as well as normal) and age group who have diabetes.

It is often possible for "graphs to conceal more than they reveal" by making comparisons across groups less evident (van Belle 2002). To highlight that individuals categorized as obese have a higher percentage of diabetes across all age categories with the exception of the 70 and over age group, we may introduce a line graph as shown in Figure 3.7. Careful attention should be paid when constructing graphical presentations of data, and possibly several methods should be considered when exploring data in order to find the graph that best captures the data's structure.

Many computer packages are available for creating bar charts (see **Program Note 3.2** on the website).

3.2.4 Histograms

As we said earlier, a histogram is similar to a bar chart but is used with interval/ratio variables. The values are grouped into intervals (often called bins or classes) that are usually of equal width. Rectangles are drawn above each interval, and the area of rectangle represents the number of observations in that interval. If all the intervals are of equal width, then the height of the interval, as well as its area, represents the frequency of the interval. In contrast to bar charts, there are no spaces between the rectangles unless there are no observations in some interval.

Table 3.7 Frequency of individual systolic blood pressures (mmHg): DIG200.

Value	Freq.	Value	Freq.	Value	Freq.	Value	Freq.	Value	Freq.	Value	Freq.
85	**1**	105	1	**116**	**8**	128	3	138	1	**150**	**12**
90	5	106	2	118	5	**130**	**23**	139	2	152	3
95	2	108	2	**120**	**25**	131	1	**140**	**26**	155	1
96	1	**110**	**16**	122	4	132	2	142	1	160	3
100	**14**	112	1	124	4	134	1	144	3	162	1
102	1	114	5	125	3	135	2	145	1	165	1
104	2	115	2	126	1	136	1	148	1	**170**	**5**

We demonstrate here the construction of a histogram for the data on systolic blood pressure values from patients in the DIG200. Before creating the histogram, however, we create a one-way table that will facilitate the creation of the histogram. Table 3.7 gives the frequency of systolic blood pressure values (SBP) for each individual in the DIG200. Note that there are 199 observations because one individual in the placebo group has missing information on her systolic blood pressure.

After inspecting the data, you should notice that a large proportion of the blood pressure values appear to end in zero — 137 out of 199, actually. All the values are even numbers, with the exception of 17 observations, and 15 values that end in 5. This suggests that the person who recorded the blood pressure values may have had a preference for numbers ending in zero. This type of finding is not unusual in blood pressure studies; however, despite this possible digit preference, we are going to create some histograms based on these values shown in Table 3.7.

The following questions must be answered before we can draw the histograms for these data:

1. How many intervals should there be?
2. How large should the intervals be?
3. Where should the intervals be located?

Tarter and Kronmal (1976) discuss these questions in some depth. There are no hard and fast answers to these questions; only guidelines are provided.

The number of intervals is related to the number of observations. Generally 5 to 15 intervals would be used, with a smaller number of intervals used for smaller sample sizes. There is a trade-off between many small intervals, which allow for greater detail with few observations in any category, and a few large intervals, with little detail and many observations in the categories.

One method of determining the number of intervals is suggested by Sturges and elaborated by Scott (1979). The suggested formula is $(\log_2 n + 1)$, where n is the number of observations, to calculate the number of intervals required to construct a histogram. Therefore, the width of the interval can be calculated using the expression $(x_{max} - x_{min})/(\log_2 n + 1)$. Since there are 199 observations in Table 3.7, we need to find the value of $\log_2 199 + 1$. This value is 8.64, and we round it up to 9, meaning that 9 intervals should be used to construct the histogram.

We refer the reader to Appendix A for information on logarithms and how to calculate logarithms with different bases. The graph shown here also gives some feel for the shape of the logarithmic curve with 2 as the base. Briefly, $\log_2 199$ can be calculated dividing

$\log_{10}199$ by $\log_{10}2$, which is 7.64. The base 10 logarithm is available on most calculators or computer software. Alternatively, the value of $\log_2199$ can be read from the graph. The dotted line in the graph shows that the value of $\log_2199$ is about 7.6.

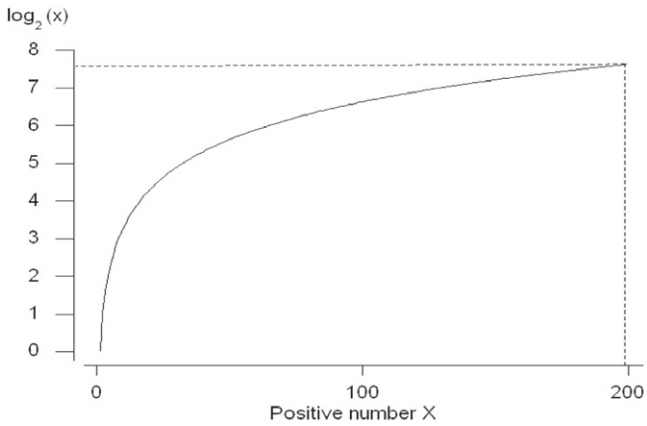


Table 3.8 illustrates the 9 intervals, and the interval width can be calculated using the expression $(x_{max} - x_{min})/(\log_2 n + 1)$. Since $(170 - 85)/8.64 = (85)/8.64 = 9.84$, we round the interval width to 10 mmHg. Notice in Table 3.8 that the notation [85–95) means all values from 85 to 95 but not including 95. Here we use the bracket ([ ) to indicate that the value should be included in the interval, whereas the parenthesis ( )) means up to the value but not including it. We have started the intervals with the value of 85, although we could have also begun the first interval with the value of 80.

This is a reasonable approach unless there are some relatively large or small values. In this case, exclude these unusual values from the difference calculation and adjust the minimum and maximum values accordingly. The location of the intervals is also arbitrary. Most researchers either begin the interval with a rounded number or have the midpoint of the interval be a round number. The computer packages create histograms using the procedures similar to the preceding approach with options to accommodate the users' request (see **Program Note 3.3** on the website).

Figure 3.8 displays the histogram for the data in Table 3.8.

Table 3.8 Intervals of histogram suggested by Sturges for the systolic blood pressure data in Table 3.7.

Class (Bin)	Class Width (Bin Width)	Frequency	Relative Frequency	Cumulative Relative Frequency	Cumulative Frequency
1	[85–95)	6	3.02	3.02	6
2	[95–105)	20	10.05	13.07	26
3	[105–115)	27	13.57	26.63	53
4	[115–125)	48	24.12	50.75	101
5	[125–135)	34	17.09	67.84	135
6	[135–145)	36	18.09	85.93	171
7	[145–155)	17	8.54	94.47	188
8	[155–165)	5	2.51	96.98	193
9	[165–175)	6	3.02	100.00	199
	Total	199	100.00		

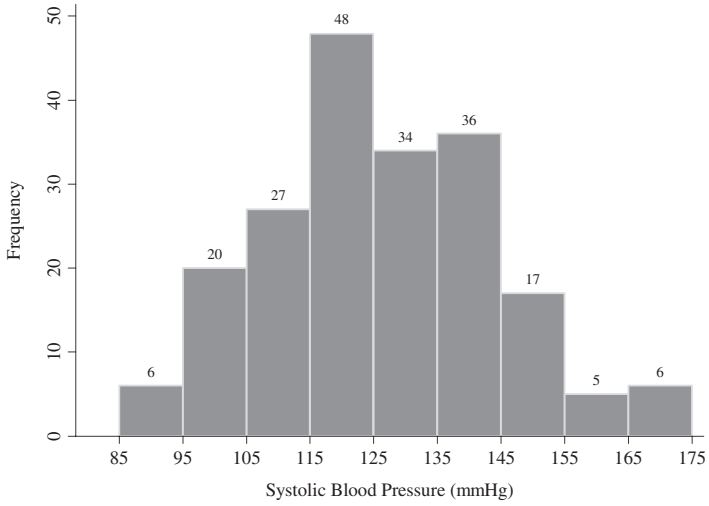


Figure 3.8 Histogram of 199 systolic blood pressure values using 9 intervals of size 10 starting at 85.

Example 3.6

Create histograms to compare the distributions of systolic blood pressures between individuals under 60 years of age and those 60 and over using the DIG200 data set. We begin by displaying the number of observations, the minimum value, and the maximum value for each of the age groups.

Under 60 years of age: $n = 81$, minimum = 90 mmHg, maximum = 170 mmHg
60 years and over: $n = 118$, minimum = 85 mmHg, maximum = 170 mmHg

We use Sturges' rule to determine the number of intervals that should be used to construct each histogram. The suggested number of intervals are:

Under 60 years of age: $(170 - 90)/(\log_2 81 + 1) = 10.9$ or 11 intervals
60 years and over: $(170 - 85)/(\log_2 118 + 1) = 10.8$ or 11 intervals

The same number of intervals is indicated. Even when different numbers of intervals were indicated, it will be better to keep the number of intervals the same for a better comparison.

Figure 3.9 presents two histograms for these groups. The first histogram displays the SBP of patients under 60 years of age and the second histogram for the 60 years and over group.

Notice that in this case the relative frequencies are used rather than frequencies mainly because the histograms are to be compared and the two groups have an unequal number of observations as just shown (i.e., there are 81 patients under 60 years of age and 118 who are 60 years and over). Relative frequencies allow for comparisons between two or more groups even if the groups do not have the same number of subjects. It is obvious, for subjects 60 years and over, that the highest

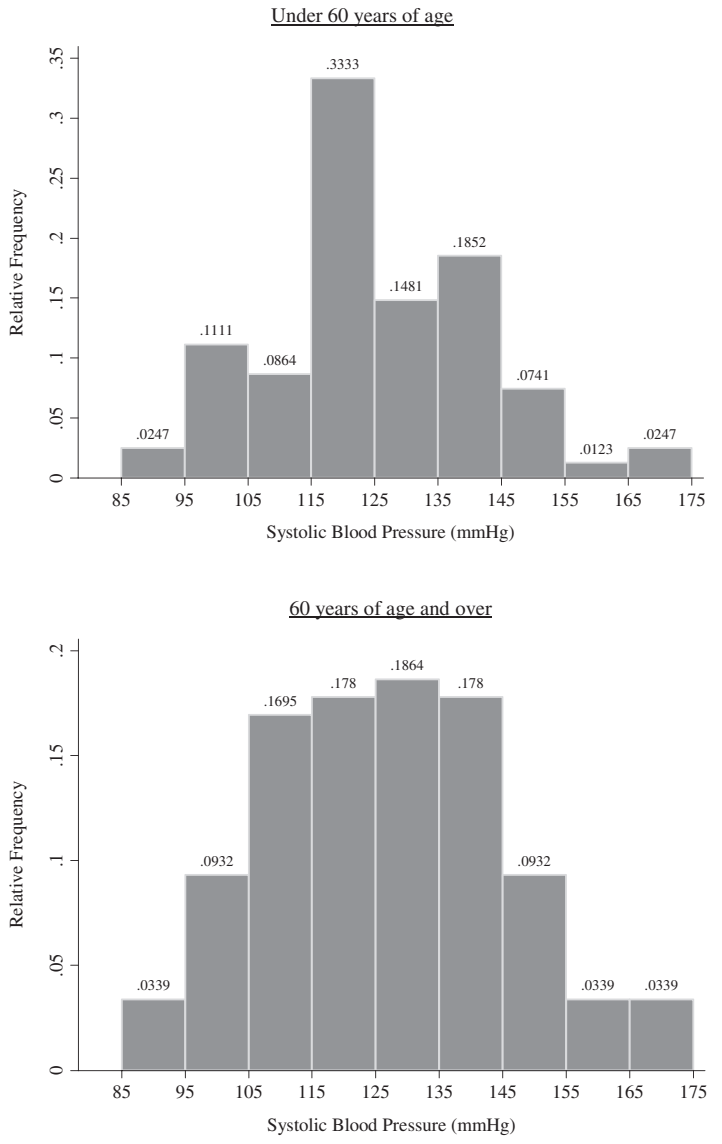


Figure 3.9 Histograms for systolic blood pressure distributions by age group.

percentages of systolic blood pressure readings fall in the intervals between 105 and 145 mmHg. Subjects under 60 years of age have a third of their systolic blood pressure observations in the interval between 115 and 125 mmHg. After comparing the two histograms, it is easy to see that the older age group has a higher concentration of subjects with systolic blood pressure values above 135 mmHg, an observation that was clearly expected.

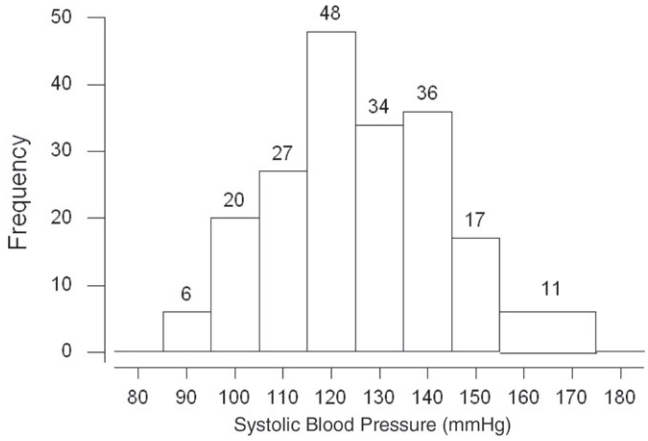


Figure 3.10 Histogram for systolic blood pressure with uneven intervals, DIG200.

It is possible for histograms constructed from the same data to have different shapes. The shapes of the histogram depend on the number of intervals used and how the boundaries are set. These differences in constructing the histogram may lead to different impressions about the data. However, histograms say basically the same thing about the distribution of the sample data even though their shapes are different.

Equal size intervals are used in most histograms. In case the use of unequal size intervals is desired, we must make some adjustments. Since the area of the rectangle for a category in a histogram reflects the frequency of the category, we need to adjust the height of an uneven size interval to keep the area at the same size. For example, assume we are interested in determining the number of subjects with SBP 155 mmHg and higher. We can collapse the last two intervals of the histograms presented in Figure 3.8 into one large interval that is twice as wide as the previous intervals. The histogram with the combined category is presented in Figure 3.10. Note that the frequency for the combined interval is 11, but the height of this interval is 5.5, one-half of the combined frequency. We divided the height by 2 to reflect the fact that the width of this last interval is twice as wide as the other intervals.

3.2.5 Stem-and-Leaf Plots

The *stem-and-leaf plot* looks similar to a histogram except that the stem-and-leaf plot shows the data values instead of using bars to represent the height of an interval. The stem-and-leaf plot is used for a relatively small dataset, while the histogram is used for a large dataset. Considering the systolic blood pressure readings of the 40 patients from the DIG40 data set, the stem contains the tens units and the leaves would be the ones units.

```
 4    10 | 0045
 9    11 | 05666
16    12 | 0002488
(8)   13 | 00000048
16    14 | 000000044
 7    15 | 00002
 2    16 |
 2    17 | 00
```

Notice that a stem-and-leaf plot looks like a histogram except we know the values of all the observations, and histograms don't group data in the same way. The first column shows a cumulative count of all the observations from the top and from the bottom to the interval in which the median value is found. The median is the value such that 50 percent of the values are less than it, and 50 percent are greater than it. The number of observations in the interval containing the median is shown in parentheses. The second column is the stem, and the subsequent columns contain the leaves. For example, in the first row we read a stem of 10 and leaves of 0, 0, 4, and 5. Since the stem represents units of 10 and the leaf unit is 1, these four numbers are 100, 100, 104, and 105. The second row has a stem of 11, and there are 5 leaves referring to the systolic blood pressure values of 110, 115, 116, 116, and 116. Note that the first number in the second row is 9, which is the cumulative count of observations in the first two rows. There are 7 values in the third row, and the cumulative count is now 16. The median is the fourth row, and its value is 130. The method of determining the median is discussed later.

Example 3.7

Here is a stem-and-leaf plot to compare SBP (mmHg) readings of the following males and females in the DIG40 data set:

Males:	100	104	105	110	116	116	116	120	120	128	128	130	130	130	134	138	140
	140	140	140	140	140	140	144	150	150	150	152	170	170				
Females:	100	115	120	122	124	130	130	130	144	150							

Females	Stem	Males
0	10	045
5	11	0666
420	12	0088
000	13	00048
4	14	00000004
0	15	0002
	16	
	17	00

By displaying a two-sided stem-and-leaf plot, a comparison of the distributions of systolic blood pressures between males and females can be made. The comparison shows that female SBPs tend to be lower than male SBPs. The male observations have two extreme values occurring at 170 mmHg even though most of the male readings are concentrated at 140 mmHg.

A nice characteristic of the data that can be seen from histograms or stem-and-leaf plots is whether or not the data are symmetrically distributed. Data are *symmetrically distributed* when the distribution above the median matches the distribution below the median. Data could also come from a skewed or asymmetric distribution. Data from a skewed distribution typically have extreme values in one end of the distribution but no extreme values in the other end of the distribution. When there is a long tail to the right, or to the bottom if the data are presented sideways, data are said to be *positively skewed*. If there are some extremely small values without corresponding extremely large values, the distribution is said to be *negatively skewed*.

Example 3.8

A stem-and-leaf plot for the ages of patients in the DIG40 data set is

3	3 | 148
9	4 | 455679
17	5 | 00055889
(11)	6 | 01133455568
12	7 | 00134566789
1	8 | 3

Notice that the data appears to be slightly asymmetric as the observations below the row containing the median are not grouped as tightly as those above it. In this case, we would consider the distribution of ages to be negatively skewed.

3.2.6 Dot Plots

A *dot plot* displays the distribution of a continuous variable. Consider Example 3.9 following where we want to compare the distribution of the continuous variable, systolic blood pressure, across a nominal variable such as age grouped into two categories — under 60 years of age and 60 years and over. These plots give a visual comparison of the center of the observations as well as providing some idea about how the observations vary. Like stem-and-leaf plots, dot plots are used for a relatively small data set.

Example 3.9

Dot plots comparing SBP across the age groups of "<60" and "≥60" are shown in Figure 3.11.

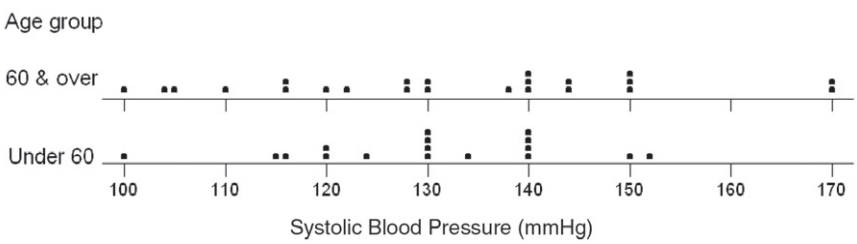


Figure 3.11 Dot plots for systolic blood pressure by age group, DIG40.

The dot plots allow us to see the data in its entirety. From the graphs, we see that the largest systolic blood pressure observation in the 60 and over group is considerably larger than the corresponding largest value in the under 60 years of age group. Also notice that dots are stacked up for observations with the same measurement value. For example, the stacked dots make it clear that there are two observations with the systolic blood pressure reading of 170 mmHg.

3.2.7 Scatter Plots

The two-dimensional *scatter plot* is analogous to the two-way frequency table in that it facilitates the examination of the relation between two variables. Unlike the two-way table, the two-dimensional scatter plot is most effectively used when the variables are continuous. Just as it is possible to have higher dimensional frequency tables, it is possible to have higher dimensional scatter plots, but they become more difficult to comprehend.

The scatter plot pictorially represents the relation between two continuous variables. In a scatter plot, a plotted point represents the values of two variables for an individual. We examine the relationship between serum creatinine levels and systolic blood pressure for 40 patients in the DIG40 data set (Table 3.1) using a scatter plot. Let us look at the top scatter plot in Figure 3.12. Each circle represents a patient's serum creatinine and systolic blood pressure values. For example, the circle in the upper left-hand corner of the plot represents the second patient (ID = 2312) in Table 3.1 with serum creatinine of 2.682 mg/dL and SBP of 104 mmHg. Overall, the scatter plot does not appear to show any relationship at all. There is a positive association between the variables when larger (smaller) values on one variable appear with larger (smaller) values of the other variable.

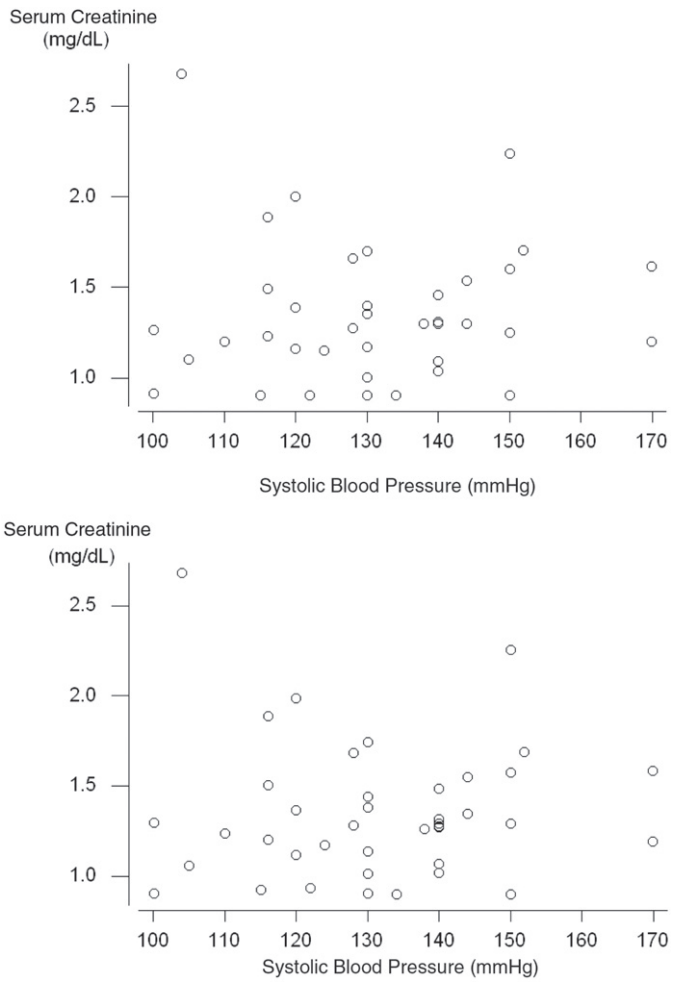


Figure 3.12 Scatter plot of serum creatinine versus systolic blood pressure for 40 patients with and without jittering, DIG40.

The association would be negative if individuals with large values of one variable tended to have small values of the other variable and conversely.

It is possible that several patients have the identical values of both variables. A careful examination of the data in Table 3.1 shows that three patients (ID = 4787, 1954, 2439) have the identical serum creatinine of 1.307 mg/dL and SBP of 140 mmHg. They are represented by one circle in the top scatter plot but by overlapping circles in the bottom scatter plot. In the bottom scatter plot a *jittering* (a very small random value) is added to the values of serum creatinine variable. If the jittering is performed for both variables, then the relative distances between circles could be slightly shifted in one or both directions.

Scatter plots are most effective for small to moderate sample sizes. When there are many variables such as in the DIG40 data set, a scatter plot matrix can be useful in displaying multiple two-way scatter plots (see Figure 3.13). From the plots we can see that there is a tendency for a very slight positive relationship between age and serum creatinine level and a slight negative relationship between serum creatinine and body mass index. There is no visual evidence of a relationship between other variables. Computer packages can be used to create stem-and-leaf plots and scatter plots (see **Program Note 3.4** on the website).

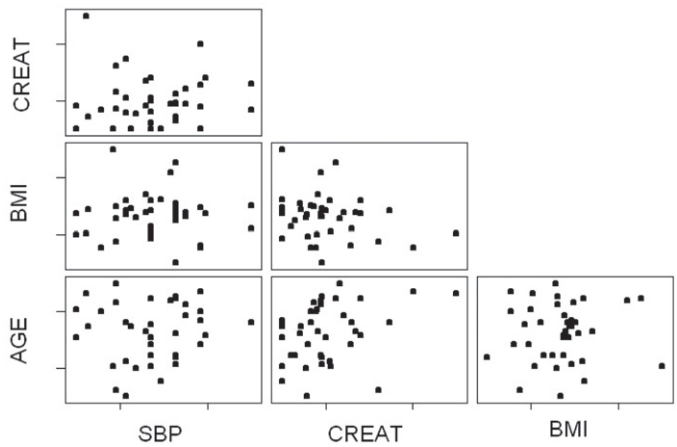


Figure 3.13 Scatter plot matrix examining the interrelationship among systolic blood pressure, creatinine, body mass index, and age, DIG40.

This completes the presentation of the pictorial tools in common use with the exception of the box plot, which is shown later in this chapter. The following material introduces the more frequently used statistics that aid us in describing and summarizing data.

3.3 Measures of Central Tendency

Simple descriptive statistics can be useful in data editing as well as in aiding our understanding of the data. The *minimum* and the *maximum* values of a variable are useful statistics when editing the data. Are the observed minimum and maximum values reasonable or even possible? For the patient's systolic blood pressure readings shown in Table 3.9, the minimum reading is 100 mmHg and the maximum is 170 mmHg. These values are somewhat unusual given that the average systolic blood pressure is approxi-

Table 3.9 Systolic blood pressure reading in ascending order, DIG40.

100	100	104	105	110	115	116	116
116	120	120	120	122	124	128	128
130	130	130	130	130	130	134	138
140	140	140	140	140	140	140	144
144	150	150	150	150	152	170	170

mately 131.4 mmHg, but they are not impossible. We will consider other ways of identifying unusual values in later sections.

3.3.1 Mean, Median, and Mode

In terms of describing data, people usually think of the average value or arithmetic mean. For example, the average systolic blood pressure was useful in determining whether or not the maximum and minimum values were reasonable. There are three frequently used measures of central tendency: the *mean*, the *median*, and the *mode*.

The *sample mean* ($\bar{x}$) is the sum of all the observed values of a variable divided by the number of observations. The *median* is defined to be the middle value — that is, the value such that 50 percent of the observed values fall above it and 50 percent fall below it. It can also be called the 50th percentile, where the *i*th percentile represents the value such that *i* percent of the observations are less than it. The mode is the most frequently occurring value.

Example 3.10

Calculate the mean systolic blood pressure reading using 40 patients in the DIG40 data set presented in Table 3.9.

The average or arithmetic mean is

$$\frac{100+100+104+\cdots+170}{40}=\frac{5256}{40}=131.4\,\text{mmHg}.$$

We can also represent the mean succinctly using symbols. We shall use upper-case X as the symbol for the variable under study — in this case, the SBP for patients in the DIG40 data set. We use lower-case x, with subscripts to distinguish each patient's systolic blood pressure, to represent the observed value of the variable. For example, the first patient's SBP is represented by x_1 and its value is 100 mmHg. The second patient's systolic blood pressure is x_2 and its value is also 100 mmHg. In the same way, x_3 is 104 mmHg, . . . , and x_{40} is 170 mmHg. Then the sum of the SBP can be represented by

$$x_1+x_2+x_3+\cdots+x_{40}=\sum_{i=1}^{40}x_i.$$

The symbol Σ means summation. The value of i beneath Σ gives the subscript of the first x_i to be included in the summation process. The value above Σ gives the subscript

of the last x_i to be included in the summation. The value of i increases in steps of 1 from the beginning value to the ending value. Thus, all the observations with subscripts ranging from the beginning value to the ending value are included in the sum. The formula for the sample mean variable, $\bar{x}$ (pronounced x-bar), is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

or more specifically in the case of this example,

$$\bar{x} = \frac{\sum_{i=1}^{40} x_i}{n} = \frac{(100 + 100 + 104 + \cdots + 170)}{40} = 131.4\,\text{mmHg}.$$

If we have the data for the entire population, not for just a sample of observations from the population, the mean is denoted by the Greek letter μ (pronounced "mu"). Values that come from samples are *statistics*, and values that come from the population are *parameters*. For example, the sample statistic $\bar{x}$ is an estimator of the population parameter μ. The population mean is defined as

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

where N is the population size.

In calculating the median, it is useful to have the data sorted from the lowest to the highest value as that assists in finding the middle value. Table 3.9 shows the sorted systolic blood pressure values for the 40 patients. For a sample of size n, the sample median is the value such that half (n/2) of the sample values are less than it and n/2 are greater than it. When the sample size is odd, the sample median is the [(n + 1)/2]th largest value. For example, the median for a sample of size 33 is thus the 17th largest value. The value 17 comes from (33 + 1)/2. When sample size is even, as in the case of the data on systolic blood pressure readings presented in Table 3.9, there is no observed sample value such that one-half of the sample falls below it and one-half falls above it. By convention, we use the average of the two middle sample values as the median — that is, the average of the (n/2)th and [(n/2) + 1]th largest values.

Example 3.11

Calculate the median systolic blood pressure readings using 40 patients in the DIG40 data set presented in Table 3.9. The data are already sorted in ascending order:

$$x_1 = 100,\ x_2 = 100,\ x_3 = 104,\ \ldots,\ x_{40} = 170.$$

Since we have an even number of patients, identify the (n/2)th observation or the (40/2) = 20th observation and the [(n/2) + 1]th observation or [(40/2) + 1] = 21st observation. Since x_{20} = 130 and x_{21} = 130, the average of these two values is 130.

The mode is the most frequently occurring value. When all the values occur the same number of times, we usually say that there is no unique mode. When two values occur the same number of times and more than any other values, the distribution is said to be bimodal. If there are three values that occur the same number of times and more than any other value, the distribution could be called trimodal. Usually one would not go beyond trimodal in labeling a distribution.

It is not unexpected to have no unique mode when dealing with continuous data, since it is unlikely that two units have exactly the same values of a continuous variable. However, in our data set of systolic blood pressure readings present in Table 3.9, the value of 140 mmHg occurs seven times, more frequently than any other reading, and is thus the mode. Although blood pressure is a continuous variable, the measurer often has a preference for values ending in zero, resulting in multiple observations of some values.

3.3.2 Use of the Measures of Central Tendency

Now that we understand how these three measures of central tendency are defined and found, when are they used? Note that in calculating the mean, we summed the observations. Hence, we can only calculate a mean when we can perform arithmetic operations on the data. We cannot perform meaningful arithmetic operations on nominal data. Therefore, the mean should only be used when we are working with continuous data, although sometimes we find it being used with ordinal data as well. The median does not require us to sum observations, and thus it can be used with continuous and ordinal data, but it also cannot be used with nominal data. The mode can be used with all types of data because it simply says which level of the variable occurs most frequently.

The mean is affected by extreme values, whereas the median is not. Hence, if we are studying a variable such as income that has some extremely large values, that is positively skewed, the mean will reflect these large values and move away from the center of the data. The median is unaffected, and it remains at the center of the data. For data that are symmetrically distributed or approximately so, the mean and median will be the same or very close to each other.

As was just mentioned, the SBP readings ranged from 100 to 170 mmHg for the 40 observations. The sample mean was 131.4 mmHg, and the sample median was 130 mmHg. These two values do not differ very much, since the data set contains observations that are relatively extreme on both the low and high end. However, the two values of 170 mmHg have caused the mean of 131.4 mmHg to be slightly larger than the median of 130 mmHg.

3.3.3 The Geometric Mean

We use another measure of central tendency when the numbers reflect population counts that are extremely variable. For example, in a laboratory setting, the growth in the number of bacteria per area is examined over time. The number of microbes per area does not change by the same amount from one period to the next, but the change is proportional to the number of microbes that were present during the previous time period. Another way of saying this is that the growth is *multiplicative*, not additive. The areas under study may also have used different media, and the microbes may not do well in some of the media, whereas in other media the growth is explosive. Hence, we may have counts in the hundred or thousands for some of the cultures, whereas a few other cultures may have counts in the millions or billions.

The arithmetic mean would not be close to the center of the values in this situation because of the effect of the extremely large values. The median could be used in this situation. However, another measure that is used in these situations is the *geometric mean*. The sample geometric mean for n observations is the nth root of the product of the values — that is,

$$\overline{x}_g = \sqrt[n]{x_1 * x_2 * \cdots x_n}.$$

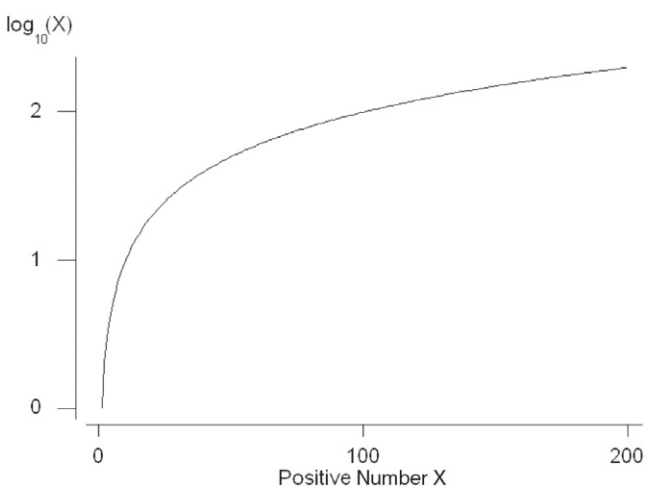
Note that since the nth root is used in its calculation, the geometric mean cannot be used when a value is negative or zero.

This definition of the geometric mean is completely analogous to the arithmetic mean. The arithmetic mean is the value such that if we add it to itself $(n - 1)$ times, it equals the sum of all the observations. It is found by summing the observations and then dividing the sum by n, the sample size. Since in the preceding situation we are dealing with data resulting from a multiplicative process, our measure of central tendency should reflect this. The geometric mean is the value such that if we multiply it by itself $(n - 1)$ times, it equals the product of all the observations. It is found by multiplying the observations and then taking the nth root of the product.

When n is 2, there is little difficulty in finding the geometric mean, since the product of the two observed values is usually not large, and we know that the second root is the square root. However, for larger values of n, the product of the observed values may become very large, and we may lose some accuracy in calculating it, even when a computer is used. Fortunately, there is another way of calculating the product of the observations that does not cause any accuracy to be lost.

We can transform the observations to a logarithmic scale. Use of the logarithmic scale provides for accurate calculation of the geometric mean. After finding the logarithm of the geometric mean, we will transform the value back to the original scale and have the value of the geometric mean. In this section, we shall use logarithms to the base 10, although other bases could be used.

Again, we refer the reader to Appendix A for more information on logarithms and how to perform logarithmic transformation. The following chart shows some idea about the relationship between positive numbers and the corresponding base 10 logarithms.



A key property of the logarithmic transformation is that the level of the mathematical operation performed on the arithmetic scale is reduced a level when the logarithmic scale is used. For example, a product on the arithmetic scale becomes a sum on the logarithmic scale. Therefore, the logarithm of the product of n values is

$$\log(x_1 * x_2 * \cdots x_n) = \sum_{i=1}^{n} \log x_i .$$

In addition, taking the nth root of a product on the arithmetic scale becomes division by n on the logarithmic scale — that is, finding the mean logarithm. In symbols, this is

$$\sqrt[n]{x_1 * x_2 * x_3 * \cdots x_n} = \frac{\sum_{i=1}^{n} \log_{10} x_i}{n} = \overline{\log_{10} x}.$$

We now have the logarithm of the geometric mean, and, to obtain the geometric mean, we must take the antilogarithm of the mean logarithm — that is,

$$\bar{x}_g = \text{antilog}\left(\overline{\log_{10} x}\right).$$

Example 3.12

Suppose that the number of microbes observed from six different areas are the following: 100, 100, 1000, 1000, 10,000, and 1,000,000. The geometric mean is found by taking the logarithm of each observation and then finding the mean logarithm. The corresponding base 10 logarithms are 2, 2, 3, 3, 4, and 6, and their mean is 3.33. The geometric mean is the antilog of 3.33, which is 2154.43. The arithmetic mean of these observations is 168,700, a much larger value than the geometric mean and also much larger than five of the six values. The usual mean does not provide a good measure of central tendency in this case. The value of the median is the average of the two middle values, 1000 and 1000, giving a median of 1000 that is of the same order of magnitude as the geometric mean.

The geometric mean has also been used in the estimation of population counts — for example, of mosquitos — through the use of capture procedures over several time points or areas. These counts can be quite variable by time or area, and hence, the geometric mean is the preferred measure of central tendency in this situation, too.

These (mean, median, mode, and geometric mean) are the more common measures of central tendency employed in the description of data. The value of central tendency, however, does not completely describe the data. For example, consider the nine systolic blood pressure readings

100 101 102 110 115 124 125 126 135.

Suppose that the four smallest observations were decreased by 10 mmHg and the four largest were increased by 10 mmHg. The values would now be the following:

90 91 92 100 115 134 135 136 145.

The means and medians of the two data sets are the same, 115 mmHg, yet the sets are very different. The sample mean of 115.3 mmHg and the sample median of 115 mmHg capture the essence of the first data set. In the second data set, however, the measures of central tendency are less informative as only one value is close to the mean and median. Therefore, some additional characteristics of the data must be used to provide for a more complete summary and description of the data and to distinguish between dissimilar data sets. The next section deals with this additional characteristic, the variability of the data.

3.4 Measures of Variability

The observations in the preceding second set of data corresponding to the systolic blood pressure of patients varied much more than those in the first set of data, but the means were the same. Hence, to provide for a more complete description of the data, we need to include a measure of its variability. A number of measures or values — the range, the interquartile range, selected percentiles, the variance, the standard deviation, and the coefficient of variation — are used to describe the variability in data.

3.4.1 Range and Percentiles

The *range* is defined as the maximum value minus the minimum value. It is simple to calculate, and it provides some idea of the spread of the data. For the patients under 60 years of age in Table 3.10, the range is the difference between 152 and 100, which is 52. In the second data set pertaining to patients 60 and over, the range is the difference between 170 and 100, which is 70.

This difference in the two ranges points to a dissimilarity between the two data sets. Although the range can be informative, the range has two major deficiencies: (1) It ignores most of the data, since only two observations are used in its definition, and (2) its value depends indirectly on sample size. The range will either remain the same or increase as more observations are added to a data set; it cannot decrease. A better measure of variability would use more of the information in the data by using more of the data points in its definition and would not be so dependent on sample size.

Percentiles, deciles, and quartiles are locations of an ordered data set that divide the data into parts. Quartiles divide the data into four equal parts. The first quartile (q1), or 25th percentile, is located such that 25 percent of the data lie below q1 and 75 percent of the data lie above q1. The second quartile (q2), or 50th percentile or median, is located

Table 3.10 Systolic blood pressure (mmHg) of patients under 60 years and 60 years and over, DIG40.

Under 60 Years					60 Years and Over				
100	115	116	120	120	100	104	105	110	116
124	130	130	130	130	116	120	122	128	128
134	140	140	140	140	130	130	138	140	140
150	152				140	144	144	150	150
					150	170	170		

such that half (50 percent) of the data lie below q2 and the other half (50 percent) of the data lie above q2. The third quartile (q3), or 75th percentile, is located such that 75 percent of the data lie below q3 and 25 percent of the data lie above q3. The *interquartile* range, the difference of the 75th and 25th percentiles (the third and first quartiles), uses more information from the data than does the range. In addition, the interquartile (or semiquartile) range can either increase or decrease as the sample size increases. The interquartile range is a measure of the spread of the middle 50 percent of the values. To find the value of the interquartile range requires that the first and third quartiles be specified, and there are several reasonable ways of calculating them. We shall use the following procedure to calculate the 25th percentile for a sample of size n:

1. If $(n + 1)/4$ is an integer, then the 25th percentile is the value of the $[(n + 1)/4]$th smallest observation.
2. If $(n + 1)/4$ is not an integer, then the 25th percentile is a value between two observations. For example, if n is 22, then $(n + 1)/4$ is $(22 + 1)/4 = 5.75$. The 25th percentile then is a value three-fourths of the way between the 5th and 6th smallest observations. To find it, we sum the 5th smallest observation and 0.75 of the difference between the 6th and 5th smallest observations.

The sample size is 40 for the systolic blood pressure data in Table 3.11. According to our procedure, we begin by sorting the data in ascending order. Next, we calculate $(40 + 1)/4$, which is 10.25. Hence the 25th percentile is a value one-fourth of the way between the 10th and 11th smallest observations. Since the 10th and 11th smallest observations have the same value of 120, the 25th percentile of the first quartile is 120 mmHg. The 75th percentile is found in the same way except that we use $3(n + 1)/4$ in place of $(n + 1)/4$. Since $3(40 + 1)/4$ yields 30.75, the 75th percentile is a value three-fourths of the way between the 30th and 31st observations. Since the 30th and 31st observations have the same value of 140, the 75th percentile, or the third quartile, is 140 mmHg. Hence, the interquartile range is $140 - 120 = 20$. Calculating the interquartile range for systolic blood pressure readings of patients under 60 years of age and 60 years and over gives the values 20 and 28, respectively. The larger interquartile range for the 60 and over age group suggests that there is more variability in the data compared to the systolic blood pressure readings for the younger age group.

The values of five selected percentiles — the 10th, 25th, 50th, 75th, and 90th — when considered together provide good descriptions of the central tendency and the spread of the data. However, when the sample size is very small, the calculation of the extreme percentiles is problematic. For example, when n is 5, it is difficult to determine how the 10th percentile should be calculated. Because of this difficulty, and also because of the instability of the extreme percentiles for small samples, we shall calculate them only when the sample size is reasonably large — say, larger than 30. The next measure of variability to be discussed is the variance, but, before considering it, we discuss the box plot because of its relation to the five percentiles.

Table 3.11 Systolic blood pressure of patients who have had a previous myocardial infarction stratified by the dose level of Digoxon treatment assigned, DIG200.

Low Dose Digoxon Treatment (0.125 mg/dL)					High Dose Digoxon Treatment (0.375 mg/dL)				
140	102	85	160	150	96	118	120	124	140
144	130	130	110	110	120	122	130	140	150

3.4.2 Box Plots

The box and whiskers plot, or just box plot, graphically gives the approximate location of the quartiles, including the median, and extreme values. The advantage of using box plots when exploring data is that several of the characteristics of the data such as outliers, symmetry features, the range, and dispersion of the data can be easily compared between different groups. The lower and upper ends (hinges) of the box mark the 25th and 75th percentiles or the locations of the first and third quartiles, while the solid band indicates the 50th percentile or the median. The whiskers represent the range of values, and the default option used in most statistical packages is to draw the whiskers out to 1.5 or 3 times the interquartile range. If the box plot is presented vertically, the area from the top edge to the bottom edge of the box represents the interquartile range.

From the systolic blood pressure data in Table 3.9, we already found the following information:

$$\begin{aligned}\text{minimum value} &= 100 \text{ mmHg},\\ \text{first quartile} &= 120 \text{ mmHg},\\ \text{median} &= 130 \text{ mmHg},\\ \text{third quartile} &= 140 \text{ mmHg},\\ \text{maximum value} &= 170 \text{ mmHg}.\end{aligned}$$

These values are plotted in a box plot in Figure 3.14.

We can use Figure 3.14 to assess the symmetry of the systolic blood pressure distribution. The box plots and histograms give us an indication of whether or not the data are skewed. For these patients, the distance from the median to the third quartile looks about the same as the corresponding distance to the first quartile. But there is a slightly longer tail to the right than to the left, indicating the distribution is slightly skewed to the right. In the Example 3.13, we go one step further by comparing the systolic blood pressures across all the age groups.

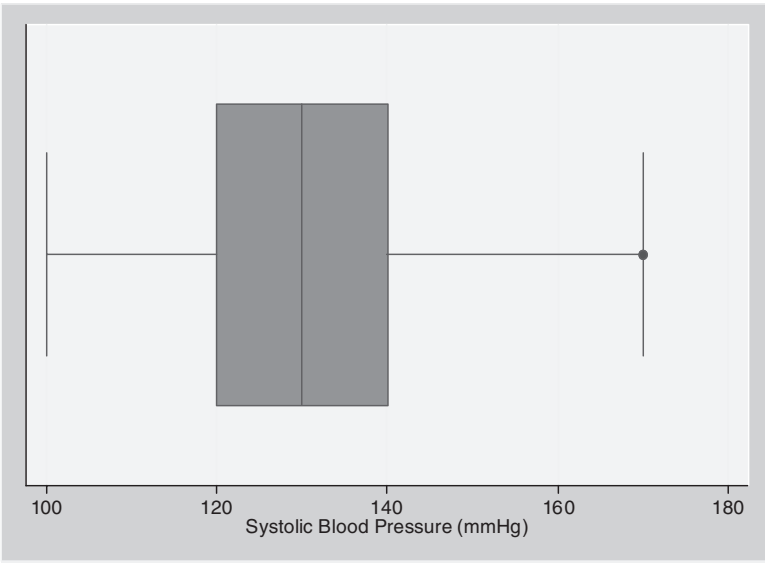


Figure 3.14 Box plot of systolic blood pressure values, DIG40.

Example 3.13

Using the data from Table 3.11, individual box and whisker plots of systolic blood pressure for the two age groups are created in Figure 3.15.

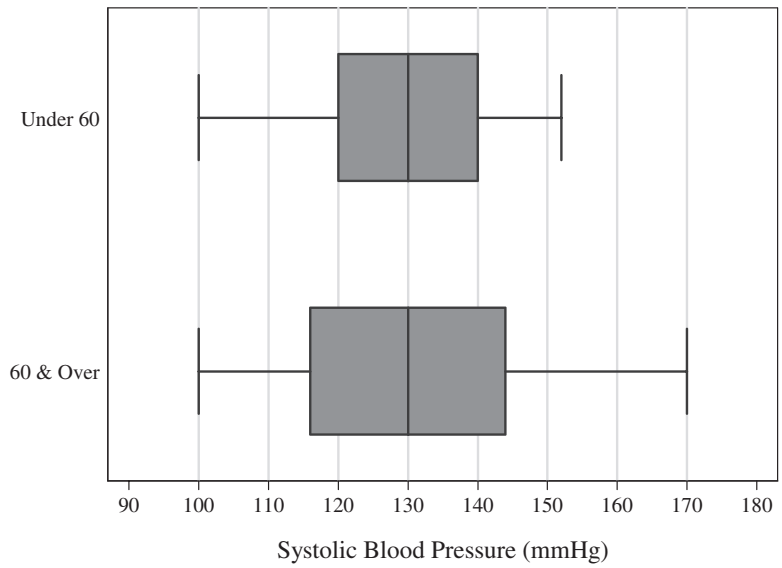


Figure 3.15 Box plots of systolic blood pressure across age groups, DIG40.

By looking at the box and whiskers plots side by side, it's possible to compare the distributions of systolic blood pressures for the two age categories. The medians are identical for both age groups. However, systolic blood pressure readings are more variable for the 60 and over group. This greater variability is shown in the larger width from the first quartile to the third quartile and through the greater range of the 60 and over group.

3.4.3 Variance and Standard Deviation

The *variance* and its square root, the *standard deviation*, are the two most frequently used measures of variability, and both use all the data in their calculations. The variance measures the variability in the data from the mean of the data. The population variance, denoted by σ^2 for a population of size N, is defined as

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}.$$

For a sample of size n, the sample variance s^2, an estimator of σ^2, is defined by

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1},$$

and the sample standard deviation is defined by

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}.$$

The population variance could be interpreted as the average squared difference from the population mean, and the sample variance has almost the same interpretation about the sample mean.

The variance uses the sum of the squared differences from the mean divided by N, whereas the sample variance uses $n - 1$ in its denominator. Why were the squared differences chosen for use instead of the differences themselves? Perhaps the following table will clarify this. In Table 3.12 we find the systolic blood pressure readings for patients on low and high dose Digoxin treatment who have had a previous myocardial infarction.

If we consider only the 10 patients who were on high dose treatment, we can construct the information provided in Table 3.12. The sum of systolic blood pressure minus the mean must be zero since the positive differences cancel the negative differences.

Table 3.12 Differences and squared differences from the mean systolic blood pressure for 10 patients on high dose (0.375 mg/dL) Digoxin treatment who have had a previous myocardial infarction.

	SBP (mmHg)	SBP − mean	(SBP − mean)^2
	96	−30	900
	118	−8	64
	120	−6	36
	124	−2	4
	140	14	196
	120	−6	36
	122	−4	16
	130	4	16
	140	14	196
	150	24	576
Total	1260	0	2040

Additionally, why is $n - 1$ used instead of n in the denominator of the sample variance? It can be shown mathematically that the use of n results in an estimator of the population variance, which on the average slightly underestimates it. The following will give some feel for the use of $n - 1$.

In the formula for the sample variance, the population mean is estimated by the sample mean. This estimation of the population mean reduces the number of independent observations to $n - 1$ instead of n as is shown next. For example, you are told that there are three observations and that two of the values along with the sample mean are known. Can you find the value of the other observation? If you can, this means that there are only two independent observations, not three, once the sample mean is calculated. Suppose that the two values are 6 and 10 and the sample mean is 9. Since the mean of the three observations is 9, this indicates that the sum of the values is 27 and that the unknown value is [27 − (6 + 10)] = 11. In this sample of size three, given knowledge of the sample mean, only two of the observations are independent or free to vary. Hence, once a parameter (in this case the population mean) is estimated from the data, it reduces the number of independent observations (degrees of freedom) by one.

To account for this reduction in the number of independent observations, $n - 1$ is used in the denominator of the sample variance.

For the 10 systolic blood pressure values from patients on high dose Digoxin treatment in Table 3.12, the value of the sample variance is

$$s^2 = \frac{\sum_{i=1}^{10}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{10}(x_i - 126)^2}{10-1} = \frac{2040}{9} = \mathbf{226.7},$$

and the value of the sample standard deviation is

$$s = \sqrt{\frac{\sum_{i=1}^{10}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{10}(x_i - 126)^2}{10-1}} = \mathbf{15.1}.$$

The sample variance and standard deviation for the 10 values from patients on low dose Digoxin treatment in Table 3.11 are 561.4 and 23.7, respectively — much larger values than the corresponding statistics for the 10 values in the high dose group. These statistics reflect the greater variation in the low dose values than in the high dose values.

The variance changes when nonconstant changes are made to all observations in the data. How does the value of the variance change when (1) a constant is added to (subtracted from) all the observations in the data set and (2) all the observations are multiplied (divided) by a constant?

The answer to the first question is that there is no change in the value of the variance, as can be seen from the following. If all the observations are increased by a constant — say, by 10 units — the mean is also increased by the same amount. Therefore, the constants will simply cancel each other out in the squared differences — that is,

$$[(x_i + 10) - (\mu + 10)]^2 = (x_i - \mu)^2$$

and thus there is no change in the sum of the squared differences or in the variance.

When all the observations are multiplied by a constant, the variance is multiplied by the square of the constant as can be seen from the following. If all the observations are multiplied by a constant — say, by 10 — the mean is also multiplied by the same amount. Therefore, in the squared differences we have

$$[(x_i * 10) - (\mu * 10)]^2 = [(x_i - \mu) * 10]^2 = (x_i - \mu)^2 * 10^2$$

and the sum of the squared differences, and thus the variance, is multiplied by the constant squared. This means that the standard deviation is multiplied by the constant. These two properties will be used in Chapter 5.

In later chapters, the variance and the standard deviation are shown to be the most appropriate measures of variation when the data come from a *normal distribution*, as knowledge of them and the mean is all that is necessary to completely describe the data. The normal distribution is the bell-shaped distribution often used in the grading of courses, and it is the most widely used distribution in statistics. The interquartile range and the five percentiles are useful statistics for characterizing the variation in data regardless of the distribution from which the data are selected, but they are not as informative as the mean and variance are when the data come from a normal distribution.

One last measure of variation is the *coefficient of variation*, defined as 100 percent times the ratio of the standard deviation to the mean. In symbols this is $(\sigma/\mu)100$ percent, and it is estimated by $(s/\bar{x})100$ percent. The coefficient of variation is a relative measure of variation, since in dividing by the mean, it directly takes the magnitude of the values into account. Large values of the coefficient suggest that the data are quite variable.

The coefficient of variation has several uses. One use is to compare the precision of different studies. If another experiment has a much smaller coefficient of variation than that in your study of the same substance, this suggests that there may be room for improvement in your study procedures. Another use is to determine whether or not there is so much variability in the data that the measure of central tendency is of little value. For example, the NCHS does not publish sample means for variables if the estimated coefficient of variation is greater than 30 percent.

Let us calculate the estimated coefficients of variation for our two sets of 10 observations in Table 3.11. For the first set, s was 23.7 and s was 15.1 in the second set. The sample mean was approximately 126 mmHg in both sets. These values lead to coefficients of variation of 18.8 percent (= [23.6946/126.1] 100 percent) and 12.0 percent in sets one and two, respectively. These values reinforce our feeling that the mean provided more useful information in the second set but was of less value in describing the data in the first set.

See **Program Note 3.5** on the website for the use of computer packages for descriptive statistics and box plots.

3.5 Rates and Ratios

Various forms of rates and ratios have been used in describing the health status and the change or growth of population. Rates and ratios are relative numbers that relate some absolute number of events to some other number such as the total population at that time. In this section we examine vital rates and population growth rates.

The rates of diseases and vital rates, which include death rates in general, infant mortality rates, feto-infant, neonatal and postneonatal mortality rates, and birth rates, are frequently used measures in public health. These rates are useful in determining the health status of a population, in monitoring the health status over time, in comparing the health status of populations, and in assessing the impact of policy changes.

For example, the infant mortality rate is often used in comparing the performance of health systems in different countries. In 2000, the United States had an infant mortality rate higher than that of 26 other nations. The U.S. rate was 6.9 infant deaths under 1 year of age per 1000 live births compared to a low rate of 3.2 for Japan. Most of the Western European nations and some Pacific Rim nations or large cities (Japan, Singapore, and Hong Kong) had lower rates than the United States. Canada's health system is often touted as a model for the United States because of its lower cost. How does Canada's infant mortality rate compare to that of the United States? Canada's infant mortality rate in 2000 was 5.3, almost 25 percent lower than the U.S. rate. The progress in reducing infant mortality has been most impressive, as can be seen from the U.S. rate for 1967 of 22.4 shown in Figure 1.1 in Chapter 1 compared to its 2000 rate of 6.9.

As can be seen from the following definition, a rate is basically a relative number multiplied by a constant. A rate is defined as the product of two parts: (1) the number of persons who have experienced the event of interest divided by the population size

and (2) a standard population size. For example, according to the data compiled by the National Center for Health Statistics, there were 4,021,726 live births in an estimated population of 288,369,000 in the United States in 2002. The corresponding birth rate per 100,000 is found by taking (4,021,726/288,369,000) times 100,000, and it equals 13.9 births per 100,000 population. This is considerably lower than the corresponding rate for the United States in 1960 of 23.7 births per 100,000.

However, as is often the case with rates, there is a problem in determining the value of the denominator — that is, the 2002 U.S. population. What is meant by the 2002 population size? Is it as of January 1, July 1, December 31, or some other date? Convention is that the middle of the period (mid-2002) population size is used. An additional problem is that Census data were available for 2000 but not for 2002, which introduces some uncertainty in the value used. In this case, NCHS used an estimate of the 2002 midyear resident population based on the estimates of the U.S. Bureau of Census. The uncertainty in the value of the denominator of the rate should be of little concern given the magnitude of the numbers involved in this situation.

Vital rates are usually expressed per 1000 or per 100,000 population. As was just mentioned, infant mortality rates are expressed per 1000 live births with the exception of feto-infant mortality rates. Feto-infant mortality rates are based on the number of late fetal deaths plus infant deaths under 1 year per 1000 live births. Neonatal mortality rates are based on deaths of infants who were less than 28 days old, and postneonatal rates are based on deaths of infants between 28 and 365 days old. This split of infant deaths is useful because often the neonatal deaths may be due to genetic factors, whereas the postneonatal deaths may have more to do with the environment.

Note that as the infant mortality of 1988 rate example in Chapter 1 showed, the children whose deaths are used in the conventional method of calculating this rate may have been born in 1987, not 1988. Hence, the numerator, the number of deaths, comes from both 1987 and 1988 births, whereas the denominator is based solely on 1988 births. This should cause no problem unless something happened that caused the mortality experience or the number of births to differ greatly between the two years. One way of dealing with this possibility of a difference between the years is to combine several years of data. Often health agencies pool data over three years to provide protection against the instability of small numbers and to reduce the possible, but unlikely, effect of very different birth or mortality experiences across the years.

3.5.1 Crude and Specific Rates

Rates may be either crude or specific. *Crude rates* use the total number of events in their definition, whereas *specific rates* apply to subgroups in the population. For example, there may be age-, gender-, or race-specific birth or death rates. For an age-specific death rate, only the deaths to individuals in the specific age group are used in the numerator, and the denominator is the total number of individuals in the specific age group. Specific rates are used because they supply more information and also allow for more appropriate comparisons of groups.

For example, the crude death rate for the United States in 2002 was 847.3 per 100,000 population, and the age-specific death rates, as shown in Table 3.13, varied from 17.4 for the 5- to 14-year-old group to 14,828.3 for the 85-year-old and over group (NCHS

2004). The age-specific rates provide more information than the crude rates. For the same year the crude death rate for males was 846.6 versus 848.0 for females. There is no appreciable gender difference in the crude death rates. However, the age-specific death rates for males are higher than the female-specific rates in all age groups. Perhaps the lack of a difference between genders in the crude rate is related to differences in the age distributions. The age-specific rates by gender, shown in Table 3.13, provide a better description of the mortality experience than the crude rates. Without knowledge of the age distributions, it is difficult to conclude whether or not the age variable is responsible for the lack of a difference in the crude rates.

Table 3.13 Crude and age-specific death rates for the United States by gender in 2002.

	US Total Population	Male Population	Female Population
All ages, crude	847.3	846.6	848.0
Under 1	695.0	761.5	625.3
1–4	31.2	35.2	27.0
5–14	17.4	20.0	14.7
15–24	81.4	117.3	43.7
25–34	103.6	142.2	64.0
35–44	202.9	257.5	148.8
45–54	430.1	547.5	316.9
55–64	952.4	1,184.0	738.0
65–74	2,314.7	2,855.3	1,864.7
75–84	5,556.9	6,760.5	4,757.9
85 & over	14,828.3	16,254.5	14,209.6

Source: NCHS, 2004, Tables 1, 34, and 35 and page 442

As just shown, one problem with the use of specific rates is that they are not easily summarized. They do provide more information than the crude rate, which gives a single value for a population, but sometimes it is difficult to draw a conclusion based on the examination of the specific rates. However, because of the strong linkage between mortality and age, age often must be taken into account in the comparison of two or more populations. One way of adjusting for age or other variables while avoiding the problem of many specific rates is to use adjusted rates.

3.5.2 Adjusted Rates

Adjusted rates are weighted rates, as will be shown following. There are direct and indirect methods of adjustment; the choice of which method to use depends on what data are available. The direct method requires that we have the specific rates for each study population and a standard population. Table 3.14 provides the age-specific death rates for both male and female populations of the year 2002. The 2000 U.S. population proportions represent the standard population. The standard population provides a referent for purposes of comparison. Starting with 2001, NCHS uses the 2000 U.S. resident population as the standard for age-adjusting death rates. Prior to 2001 the 1940 U.S. population was used as the standard for age-adjusting mortality statistics. The choice of a standard population is subjective. For example, in comparing the rates between states, often the U.S. population would be used as the standard. In comparing counties of a state, the state population often would be used as the standard. For comparing rates over time, the population at a previous time point could be used as the standard. Another alternative might be to pool the populations of the areas or times under study and use the pooled population as the standard.

Table 3.14 Direct method of adjusting the 2002 U.S. male and female death rates using 2000 U.S. population as the standard.

		Male Population		Female Population	
Age	U.S. Population Proportion	Specific Rates[a]	Expected Deaths[a]	Specific Rates[a]	Expected Deaths[a]
Under 1	0.013818	761.5	10.5	625.3	8.6
1–4	0.055317	35.2	1.9	27.0	1.5
5–14	0.145565	20.0	2.9	14.7	2.1
15–24	0.138646	117.3	16.3	43.7	6.1
25–34	0.135573	142.2	19.3	64.0	8.7
35–44	0.162613	257.5	41.9	148.8	24.2
45–54	0.134834	547.5	73.8	316.9	42.7
55–64	0.087247	1,184.0	103.3	738.0	64.4
65–74	0.066037	2,855.3	188.6	1,864.7	123.1
75–84	0.044842	6,760.5	303.1	4,757.9	231.4
85 & over	0.015508	16,254.5	252.1	14,209.6	220.4
Total	1.000000		1013.7[b]		715.2[b]

[a]Per 100,000 population
[b]Age-adjusted death rate per 100,000 population
Source: NCHS, 2004, Tables 1, 34, and 35, and page 442

In performing the age adjustment in Table 3.14, the 2000 U.S. age distribution is used as the standard. The adjustment process consists of applying the male and female age-specific death rates to the standard population's age distribution and then summing the expected number of deaths over the age categories. Another way of saying this is that each age category's death rate is *weighted* by that age category's share of the standard population. The direct age-adjusted death rates for 2002 male and female populations using the U.S. as the 2000 standard population are 1013.7 and 715.2 deaths per 100,000 population, respectively. The male morality rate is about 30 percent higher than the female rate.

The indirect method is an alternative to be used when we do not have the data required for the direct method or when the specific rates may be unstable because they were based on small numbers. The indirect method requires the specific rates for the standard population and the age (or, for example, gender or race) distribution for the population to be adjusted. It is more likely that these data will be available than the age-specific death rates in the population to be adjusted. The first step in calculating the indirect age-adjusted death rate is to multiply the age-specific death rates of the standard population (the U.S.) by the corresponding age distribution of the population to be adjusted. Table 3.15 shows the calculation of indirect age-adjusted rate for American Indian or Alaskan Native male and female populations using the 2000 U.S. age-specific rates as the standard.

The observed crude death rates for American Indian/Alaskan Native male and female populations are 439.6 and 367.7 per 100,000, respectively. The male crude death rate is about 20 percent higher than the female rate. When age is taken into account, the gender difference in mortality may increase, since the average age of the female population is older than that of the male population.

In performing the indirect age standardization, the 2000 U.S. age-specific mortality rates are applied to the age distribution of the male and female populations of American Indian/Alaskan Natives. The expected death rates are created by multiplying the U.S. age-specific death rates by the proportion of people in the corresponding age groups for the male and female American Indian/Alaskan Native populations and then summing these expected numbers of deaths over the age categories. The ratio of the observed to

Table 3.15 Indirect age-adjusted death rates for the 2002 male and female populations of American Indian or Alaska Natives using the 2000 U.S. age-specific death rates as the standard.

		American Indian or Alaskan Native, 2002			
		Male Population		Female Population	
Age	U.S. Age-Specific Rates[a] 2000	Population Proportion	Expected Deaths[a]	Population Proportion	Expected Deaths[a]
All ages, crude	854.0	439.6[a]		367.7[a]	
Under 1	736.7	0.013681	10.1	0.012970	9.6
1–4	32.4	0.065798	2.1	0.063554	2.1
5–14	18.0	0.192182	3.5	0.186122	3.4
15–24	79.9	0.186971	14.9	0.175746	14.0
25–34	101.4	0.154397	15.7	0.144617	14.7
35–44	198.9	0.151792	30.2	0.154345	30.7
45–54	425.6	0.117915	50.2	0.124514	53.0
55–64	992.2	0.065798	65.3	0.070687	70.1
65–74	2,399.1	0.033225	79.7	0.038911	93.4
75–84	5,666.5	0.014332	81.2	0.020752	117.6
85 & over	15,524.4	0.003909	60.7	0.007782	120.8
Total		1.000000	413.6	1.000000	529.4
Standardized mortality ratio (SMR)		439.6/413.6 = 1.063		367.7/529.4 = 0.695	
Indirect age-adjusted death rate		854(1.063) = 907.8[a]		854(0.695) = 593.3[a]	

[a]Per 100,000 population
Source: NCHS, 2004, Tables 1, 34, and 35 and page 442

the expected death rates is the *standardized mortality ratio* (SMR). From Table 3.15, we see that the SMRs for the male and female populations are 1.063 and 0.695, respectively. The male SMR is 53 percent higher than the female SMR and the gender difference is more markedly shown, just as we expected. To find the indirect age-adjusted death rate for American Indian/Alaskan Native populations, we now multiply the crude rate for the standard population (854.0 per 100,000) by the SMRs. Thus, the indirect age-adjusted mortality rates for American Indian/Alaskan Native male and female populations are 907.8 and 593.3 per 100,000, respectively.

Both the direct and indirect age-adjustment methods can be used to adjust for more than one variable; for example, age and gender are often used together. Gender is frequently used because the mortality experiences are often quite different for females and males.

3.6 Measures of Change over Time

To understand the change in the height of a child or the growth of population over time, we may plot the data against time. We look first for an overall pattern and then for deviations from that pattern. For certain phenomena the points follow a straight line, and for other phenomena the points are nonlinear. In this section, we examine two well-known patterns of growth: linear and exponential.

3.6.1 Linear Growth

Linear growth means that a variable increases by a fixed amount at each unit of time. The height of a child or the production of food supply may take this pattern. To describe this pattern, we write a mathematical model for the straight-line growth of variable y.

$$y = a + bt.$$

In this model, b is the increment by which y changes when t increases by one unit and a is the base value of y when $t = 0$.

Example 3.14

The stature-for-age growth chart of U.S. boys is shown in Figure 3.16 (NCHS) 2006. The growth pattern exhibits a roughly linear trend between ages 2 to 15 years. For a typical child (50th percentile) a is about 34 inches (at age 2) and b is roughly 2.5

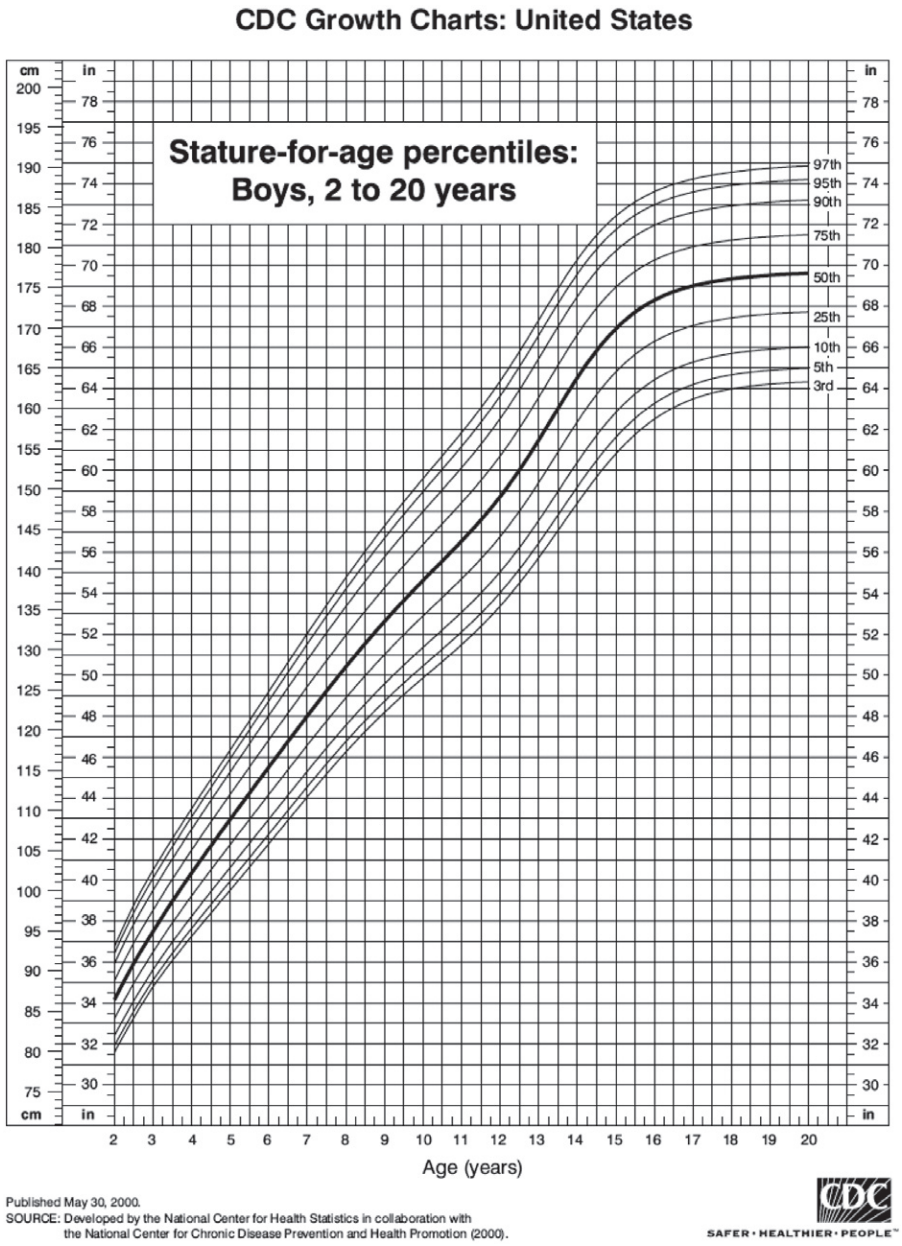


Figure 3.16 Growth chart (stature-for-age) for U.S. boys, 2 to 20 years of age.

inches. From this we can tell that the stature of a 12-year-old boy would be about 59 inches [= 34 + 2.5(10)], and the chart also shows this value. The chart also shows that the stature of boys varies more as they grow older.

We will explore this linear growth model further in Chapter 13. Because no straight line usually passes exactly through all data points, we need to find a line that fits the points as well as possible. We will learn how to estimate the best fitting line from the data.

3.6.2 Geometric Growth

The population size of a community usually does not follow the linear growth model. The change in the population size over time in an area can simply be described as the number of people added or reduced between two time points. For comparison purposes, we can express the change as percent of the base population. If the time period is the same, the percent of change can be compared between populations. The percent of change from time 0 to time t in the population P is calculated by

$$\frac{P_t - P_0}{P_0}(100) = \left(\frac{P_t}{P_0} - 1\right)(100).$$

For example, the U.S. population increased from 248,709,873 in 1990 to 281,421,906 in 2000, showing a 13.15 percent increase over a 10-year period.

Percent change indicates a degree of change, but it is not yet a "rate of change." Like other vital rates, a rate of change should express change as a relative change in population size *per year*. We need to convert the percent change into an annual rate. But we cannot simply take one-tenth of the percent change (arithmetic mean) as an annual growth rate. Equal degrees of growth do not produce equal successive absolute increments because they follow the principle of compounded interest. In other words, a constant rate of growth produces larger and larger absolute increments, simply because the base of total population steadily becomes larger. Therefore, the linear growth model would not apply to population growth.

If a population is growing at an annual rate of r, then the population at time 1 would be the base plus an incremental change — that is, $(a + ar)$ or $a(1 + r)$. If the population is subject to the same constant growth rate, the population at time t will be

$$y = a(1 + r)^t.$$

Example 3.15

The geometric growth model fits well to the growth of money deposited at a bank with the interest added at the end of each year. Suppose $1000 is deposited and earns interest at an annual rate of 10 percent for 10 years. The amount in the account (y) at each anniversary date can be calculated by $y = 1000(1 + 0.1)^t$, where t ranges from 1 to 10. Figure 3.17 shows the results. The money grew more than 100 percent because the interest was compounded annually.

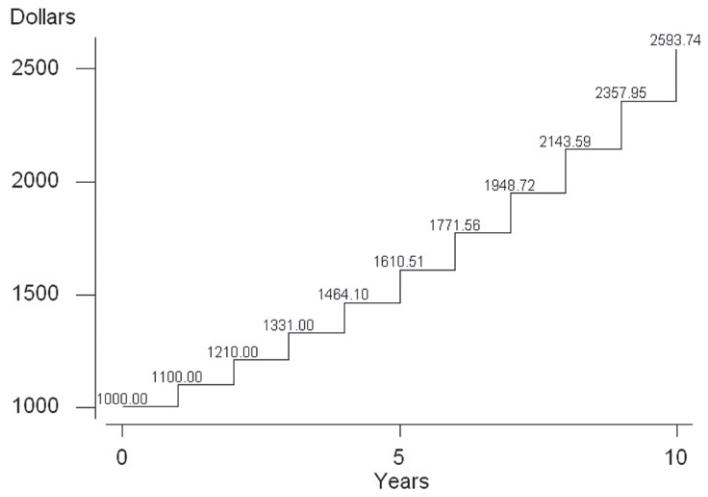


Figure 3.17 Account value over time for $1000 earning an annual interest rate of 10 percent.

If one wants to have the $1000 to be tripled over the 10-year period, then what level of annual interest rate would be required? We can solve $3000 = 1000(1 + r)^{10}$ for r as follows:

$$r+1=\exp\left(\frac{\ln(3)}{10}\right)=\exp\left(\frac{1.09861}{10}\right)=1.1161.$$

One needs to find a bank that offers an annual interest rate of 11.6 percent.

3.6.3 Exponential Growth

We know that population is changing continuously as births and deaths occur throughout the year. We want to find a model that describes the growth as a continuous process. This new model is the exponential growth model and it has the following form:

$$y = ae^{rt}$$

where r is annual growth rate, e is a mathematic constant approximately equal to 2.71828, and a is the population at $t = 0$. Figure 3.18 graphically shows the exponential growth of a population of 10,000 at an annual growth rate of 5 percent over a 30-year period.

Relating to the bank interest rate example, this model assumes that the interest is compounded continuously.

Example 3.16

The U.S. population grew from 248,709,873 in 1990 to 281,421,906 in 2000. What would be an annual growth rate over the 10-year period? We can solve the following equation for r as follows:

$$281421906 = 248709873e^{10r}$$

$$\ln(281421906/248709873) = 10r$$

$$r = \left(\frac{\ln(281421906/248709873)}{10}\right) = \left(\frac{0.1236}{10}\right) = 0.01236.$$

The U.S. population grew at the annual rate of 1.24 percent.

Using the growth rate computed, we could project future size of population. Let us project the U.S. population in 2009 assuming the rate of growth remains constant.

$$y = 281421906(e^{9(0.01236)}) = 292050102$$

Over 10 million people would be added to the U.S. population in 9 years. This type of projection is acceptable for a short time period, but it should not be used for a long-range projection.

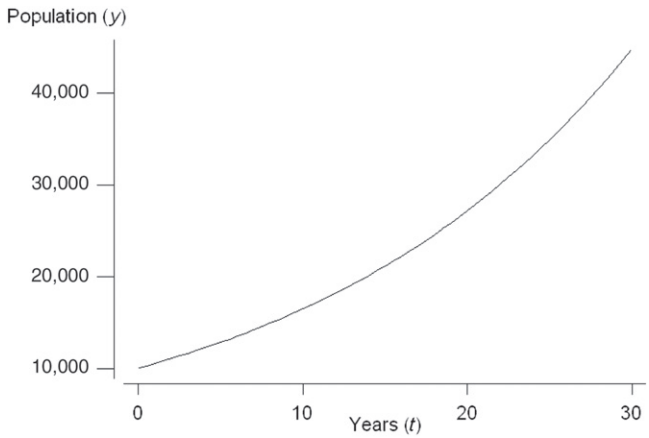


Figure 3.18 Increase of population of 10,000 at an annual rate of increase of 5 percent.

Example 3.17

(population doubling time): How long would it take to double the 2000 U.S. population assuming the annual growth rate remains constant? To answer this question, we solve the following equation for *t*.

$$2a = ae^{rt}, \text{ where } r = 0.01236$$

$$2 = e^{0.01236t}$$

$$\ln(2) = 0.01236t$$

$$t = \frac{\ln(2)}{0.01236} = \frac{0.69315}{0.01236} = 56.09.$$

The U.S. population will double in 56 years or in 2056.

Doubling means that $y/a = 2$ and natural logarithm of 2 is 0.69315. The solution suggests that if a population is increasing at an annual rate of 1 percent, then the population size will double in about 70 years. The time required to triple the population can be obtained by using ln(3). Similarly, the time required to increase the population by 50 percent can be obtained by using ln(1.5).

3.7 Correlation Coefficients

Earlier in the chapter, we presented a scatter plot of serum creatinine level and systolic blood pressure for 40 patients in the DIG40 data set, and we concluded that there was no appreciable association between serum creatinine and systolic blood pressure. Although this statement is informative, it is imprecise. To be more precise, a numerical value that reflects the strength of the association is needed. Correlation coefficients are statistics that reflect the strength of association.

3.7.1 Pearson Correlation Coefficient

The most widely used measure of association between two variables, X and Y, is the *Pearson correlation coefficient* denoted by ρ (rho) for the population and by r for the sample. This measure is named after Karl Pearson, a leading British statistician of the late 19th and early 20th centuries, for his role in the development of the formula for the correlation coefficient.

We want the correlation coefficient to be large, approaching +1 as a limit, as the values of the X, Y pair show an increasing tendency to be large or small together. When the values of the X, Y pair tend to be opposite in magnitude — that is, a large value of X with a small value of Y, or vice versa — the measure should be large negatively, approaching −1 as the limit. If there is no overall tendency of the values of the X, Y pair, the measure should be close to 0.

By large or small, we mean in relation to its mean. Because of the preceding requirements for the correlation coefficient, one simple function that may be of interest here is the product of $(x_i - \bar{x})$ with $(y_i - \bar{y})$. Let us focus on the sign of the differences, temporarily ignoring the magnitude. The possibilities are as follows:

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
+	+	+
−	−	+
+	−	−
−	+	−

The product of the differences does what we want; that is, it is positive when the X, Y pairs are large or small together and negative when one variable is large and the other variable is small. By summing the product of the differences over all the sample pairs, the sum should give some indication whether there is a positive, negative, or no association in the data. If all the products are positive (negative), the sum will be a large positive (negative) value. If there is no overall tendency, the positive terms in the sum will tend to cancel out with the negative terms in the sum, driving the value of the sum toward 0.

However, the value of the sum of the products depends on the magnitude of the data. Since we want the maximum value of our measure to be 1, we must do something to remove the dependence of the measure on the magnitude of the values of the variables. If we divide the measure by something reflecting the variability in the X and Y variables, this should remove this dependence. The actual formula for r, reflecting these ideas, is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})*(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 * \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Dividing the numerator and denominator of this formula by $n - 1$ enables us to rewrite the formula in terms of familiar statistics — that is,

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})*(y_i - \bar{y})/(n-1)}{\sqrt{s_x^2 * s_y^2}}.$$

In this version, we used the formula for the sample variance — that is, $s_x^2 = \Sigma(x_i - \bar{x})^2 / (n-1)$. The sample variance can also be expressed as $s_x^2 = \Sigma(x_i - \bar{x})(x_i - \bar{x})/n-1$. Hence, the sample variance could also be said to measure how X varies with itself. The numerator looks very similar to this, and it measures how the variables X and Y covary.

The denominator, $\sqrt{s_x^2 * s_y^2}$, standardizes r so that it varies from –1 to +1. For example, if $Y = X$, then the numerator becomes $\Sigma(x_i - \bar{x})^2 / n-1$ — that is, s_x^2, which is the same as the denominator, and their ratio is +1.

For the data shown in Figure 3.12 the correlation coefficient turns out to be 0.025, confirming our earlier statement of a very slight positive relationship between serum creatinine and systolic blood pressure.

Example 3.18

We consider the following data on diastolic and systolic blood pressure readings for 12 adults.

Systolic blood pressure:	120	118	130	140	140	128	140	140	120	128	124	135
Diastolic blood pressure:	60	60	68	90	80	75	94	80	60	80	70	85

We first use a scatter plot of systolic blood pressure versus diastolic blood pressure (shown in Figure 3.19) to get a feel for the data. The jittering is added in the plot to

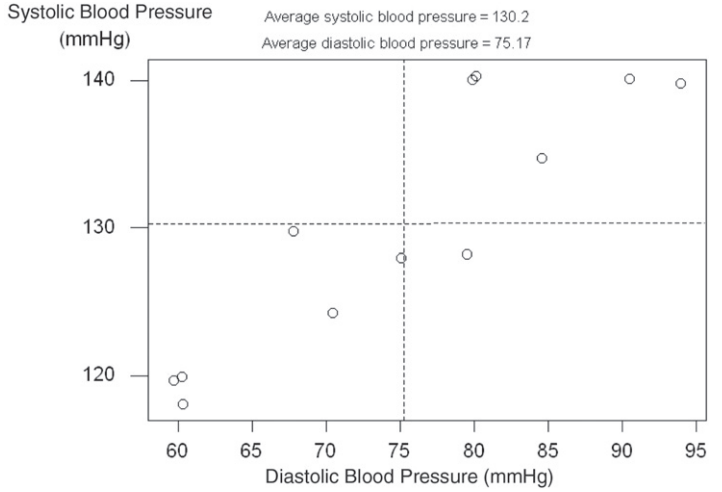


Figure 3.19 Scatter plot of systolic blood pressure versus diastolic blood pressure.

show the identical values for four adults. By adding vertical and horizontal lines showing the mean diastolic and mean systolic blood pressures, we can partition the scatter plot into four quadrants. Because most of the data cluster in the upper right and lower left quadrants, we expect that there will be a very strong correlation between these two variables.

The calculated correlation coefficient is 0.894, showing a strong positive association.

The correlation coefficient is not a general purpose measure of association, but it measures linear association — that is, the tendency of the (x_i, y_i) pairs to lie on a straight line. The following example demonstrates this point.

Example 3.19

For this example we consider the following values of Y and X:

Y:	4	1	0	1	4
X:	–2	–1	0	1	2

The sample mean of Y is 2, and the sample mean of X is 0. The pieces required to calculate r are

Y	X	$(Y-2)$	*	$(X-0)$ = product	$(Y-2)^2$	$(X-0)^2$
4	–2	2	*	–2 = –4	4	4
1	–1	–1	*	–1 = 1	1	1
0	0	–2	*	0 = 0	4	0
1	1	–1	*	1 = –1	1	1
4	2	2	*	2 = 4	4	4
Total 10	0	0		0 0	14	10

The estimated Pearson correlation coefficient, r, is then $0/\sqrt{14*10} = 0$. There is no linear association between Y and X. However, note that the first column (values of Y) and the last column (X^2) are the same. Hence, there is a perfect quadratic (squared) relation between Y and X that was not found by the Pearson correlation coefficient. The scatter plot in Figure 3.20 graphically shows this relationship. Connecting these points gives the parabola shape associated with a quadratic relationship.

Thus, even if r is 0, it does not mean that the two variables are unrelated; it means that there is no linear relation between the two variables. The use of a scatterplot first, followed by the calculation of r, may find the existence of a nonlinear association that could be missed when r alone is used.

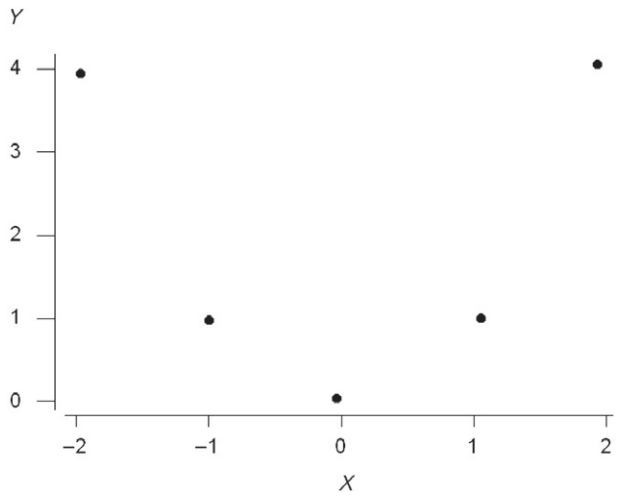


Figure 3.20 Scatter plot for the data in Example 3.20.

3.7.2 Spearman Rank Correlation Coefficient

The Pearson correlation coefficient was designed to be used jointly with normally distributed variables. However, it is used, sometimes incorrectly, with all types of data in practice. Instead of using the Pearson correlation coefficient with nonnormally distributed variables, it may be better to use a modification suggested by Spearman, an influential British psychometrician, in 1904. Spearman suggested ranking the values of Y and also ranking the values of X. These ranks are then used instead of the actual values of Y and X in the formula for the sample Pearson correlation coefficient. The result of this calculation is the sample Spearman rank correlation coefficient, denoted by r_s. In addition to being used with nonnormal continuous data, the Spearman rank correlation coefficient can also be used with ordinal data.

When ranking the data, ties (two or more subjects having exactly the same value of a variable) are likely to occur. In case of ties, the tied observations receive the same average rank. For example, if three observations of X are tied for the third smallest value, the ranks involved are 3, 4, and 5. The average of these three ranks is 4, and that is the rank that each of the three observations would be assigned. The occurrence of ties causes no problem in the calculation of the Spearman correlation coefficient when the Pearson formula is used with the ranks.

Example 3.20

Let us calculate the Spearman rank correlation coefficient for the data used in Example 3.19. The values of systolic and diastolic blood pressure values and their respective rankings are shown here. Note that there are several ties in ranking and average rankings are given.

SBP		DBP	
Value	Rank	Value	Rank
120	2.5	60	2
118	1.0	60	2
130	7.0	68	4
140	10.5	90	11
140	10.5	80	8
128	5.5	75	6
140	10.5	94	12
140	10.5	80	8
120	2.5	60	2
128	5.5	80	8
124	4.0	70	5
135	8.0	85	10

The calculated r_s is 0.866, slightly less than the Pearson correlation coefficient of 0.894.

See **Program Note 3.6** on the website for calculation of Pearson and Spearman correlation coefficients.

Conclusion

In this chapter we presented tables, graphs, and plots, as well as a few key statistics. The pictures and statistics together enable one to describe single variables and the relationship between two variables for the sample data. Although the description of the sample data and the provision of estimates of the population parameters are important, sometimes we wish to go beyond that — for example, to give a range of likely values for the population parameters or to determine whether or not it is likely that two populations under study have the same mean. Doing this requires the use of probability distributions, a topic covered in a subsequent chapter.

EXERCISES

3.1 Create a bar chart of the following data on serum cholesterol for non-Hispanic whites based on Table II-42 in *Nutrition Monitoring in the United States* (Life Sciences Research Office 1989)

Gender	Age	N	Mean Serum Cholesterol (mg/dL)[a]
Male	40–49	572	223.5
	50–59	575	228.9
	60–69	1354	226.2
	70–74	427	215.8
Female	40–49	615	218.5
	50–59	649	243.6
	60–69	1487	249.0
	70–74	533	248.3

[a]These data are from the Second National Health and Nutrition Examination Survey of noninstitutionalized persons conducted during the 1976–1980 period (NCHS 1981)

A high value of serum cholesterol is thought to be a risk factor for heart disease. The National Cholesterol Education Program (NCEP) of the National Institutes of Health in 1987 stated that the recommended value for serum cholesterol is below 200 mg/dl, and a value between 200 and 240 is considered to be the borderline. A value above 240 may indicate a problem, and NCEP recommended that a lipoprotein analysis should be performed. Based on these data, it appears that many non-Hispanic whites have serum cholesterol values that are too high, particularly women. The medical literature is also finally beginning to recognize that homocysteine is a very important risk factor for heart disease, even among people with normal levels of serum cholesterol (http://www.quackwatch.org/03HealthPromotion/homocysteine.html).

a. Give some possible reasons why non-Hispanic white males have higher mortality from heart and cerebrovascular diseases when it appears from these data that non-Hispanic white females should have the higher rates.
b. Provide a possible explanation why the serum cholesterol values for older males are lower than for the younger males and the reverse is true for females.

3.2 Create line graphs for the following expenditures for the Food Stamps Program in New York State during the 1980s.

Year	Actual Expenditures (in millions of dollars)	Inflation-adjusted Expenditures[a]
1980	745.3	745.3
1981	901.2	814.1
1982	835.7	717.3
1983	930.9	766.8
1984	904.4	709.3
1985	939.4	712.2
1986	926.5	685.3
1987	901.8	638.7
1988	909.1	613.4
1989	964.7	616.4

[a]Expenditures adjusted for inflation using the consumer price index for the Northeast Region with 1980 as the base.
Source: Division of Nutritional Sciences, 1992

What, if any, tendencies in the expenditures (both actual and inflation-adjusted) do you see? Which expenditures data do you think should be used in describing the New York State Food Stamps Program? Explain your choice.

3.3 Use line graphs to represent the short-stay hospital occupancy rates shown here.

	Hospital Ownership			
Year	Federal	Nonprofit	Proprietary	State/Local
1960	82.5	76.6	65.4	71.6
1970	77.5	80.1	72.2	73.2
1975	77.6	77.4	65.9	69.7
1980	77.8	78.2	65.2	70.7
1985	74.3	67.2	52.1	62.8
1989	71.0	68.8	51.7	64.8

Source: NCHS, 1992

Discuss the trends, if any, in these data.

3.4 Based on DIG200 data on the Web, explore prevalence of hypertension (variable name: HYPERTEN) by age, sex, and race, using appropriate descriptive tools you learned in Chapter 3. Present and discuss your findings, offering possible explanations for your findings and suggesting ways to conduct further study on the subject.

3.5 The following data on hazardous government jobs appeared as a bar chart in the "USA SNAPSHOTS" section of *USA Today* on April 30, 1992. The variable shown was the number of assaults suffered by federal officers based on 1990 FBI figures. The least number of assaults suffered were by the Internal Revenue Service (three assaults), the Bureau of Indian Affairs (five assaults), and the Postal Inspectors (six assaults). The most assaults were suffered by the Immigration and Naturalization Service with 409, followed by U.S. attorneys with 269 and the Bureau of Prisons with 185 assaults. What additional information do you need to conclude anything about which federal officers have the more hazardous — from the perspective of assaults — jobs?

3.6 A study was performed to determine which of three drugs was more effective in the treatment of a health problem. The responses of subjects who received each of three drugs (A, B, and C) were provided by Cochran (1955). The following shows the pattern of response for the 46 subjects:

Response to			
A	**B**	**C**	**Frequency**
yes	yes	yes	6
yes	yes	no	16
yes	no	yes	2
yes	no	no	4
no	yes	yes	2
no	yes	no	4
no	no	yes	6
no	no	no	6
	Total		46

a. Give an example of a type of health problem that would be appropriate for this study.
b. Create a two-way frequency table showing the relationship between drugs A and C. Does it appear that the responses to these drugs are related?
c. Create a bar chart that shows the number of subjects with a favorable response by drug.

3.7 Using the data shown in Table 3.1, calculate the coefficient of variation for body mass index. Do you think that any measure of central tendency adequately describes these data? Explain your answer.

3.8 Lee (1980) presented survival times in months from diagnosis for 71 patients with either acute myeloblastic leukemia (AML) or acute lymphoblastic leukemia (ALL).

AML patients:
18 31 31 31 36 01 09 39 20 04 45 36 12 08 01 15 24 02 33 29 07 00 01 02 12 09 01 01 09 05 27 01 13 01 05 01 03 04 01 18 01 02 01 08 03 04 14 03 13 13 01

ALL patients:
16 25 01 22 12 12 74 01 16 09 21 09 64 35 01 07 03 01 01 22

a. Calculate the sample mean and median for both AML and ALL patients separately. Which measure do you believe is more appropriate to use with these data? Explain.
b. Create box plots, histograms, and stem-and-leaf plots to show the distributions of the survival times for AML and ALL patients. Which type of figure is more informative for these data? Which type of patient has the longer survival time after diagnosis?
c. Give examples of additional variables that are needed in order to interpret appropriately these survival times.

3.9 Is it possible to calculate the mean occupancy rate for the short-stay hospitals in 1960 given the data provided in Exercise 3.3? If it is, calculate it. If not, state why it cannot be calculated.

3.10 Provide an appropriate summarization of the following data on the results of inspections of food establishments (e.g., food processing plants, food warehouses, and grocery stores) conducted by the Division of Food Inspection Services of the New York State Department of Agriculture and Markets.

	Number Inspected		Approximate Number Failed	
Year	Upstate	NYC & LI[a]	Upstate	NYC & LI
1980	19,599	23,676	2,548	5,209
1982	17,183	22,767	3,093	6,830
1984	13,731	18,677	2,746	6,350
1986	10,915	15,948	2,292	6,379
1988	13,614	15,070	3,267	6,179
1990	12,609	16,285	3,026	6,677

[a]New York City and Long Island
Source: Division of Nutritional Sciences, 1992

Do you think that there were more or fewer cases of foodborne illness in New York State in 1990 than in 1980?

3.11 Diagnosis Related Groups (DRGs) are used in the payment for the health care of Medicare-funded patients. In the creation of the DRGs, suppose that the lengths of stay for 50 patients in one of the proposed groups were the following:

1 1 2 2 2	2 2 2 3 3	3 4 4 4 5	5 5 5 6 6
6 7 7 8 8	8 9 9 10 12	13 15 15 17 17	18 19 19 20 23
26 29 31 34 36	43 49 52 67 96		

Calculate the mean, standard deviation, coefficient of variation, and the five key percentiles for these data. Are these data skewed? Do the patients in this DRG appear to have homogeneous lengths of stay? Which measures, if any, should be used in the description of these data? Explain your answer.

3.12 The following data represent bacteria counts measured in water with levels of 0, 1, and 3% sodium chloride. The counts are the number per milliliter.

Level of Sodium Chloride	Counts
0%	10^7, 10^6, 10^8, 10^9, 10^8, 10^{10}
1%	10^4, 10^4, 10^5, 10^6
3%	10^3, 10^4, 10^4, 10^3, 10^5

a. Calculate the mean and the coefficient of variation for these data.
b. Calculate the median and the geometric mean.
c. Comment on which measure of central tendency is appropriate for these data.

3.13 Of the estimated 1,488,939 male residents of Harris County, Texas, in 1986, there were 8,672 deaths. Of the 1,453,611 female residents, there were 6,913 deaths. The estimated 1986 U.S. population was approximately 48.7 percent male and 51.3 percent female.
a. Calculate the crude death rate and the sex-specific death rates for Harris County in 1986.
b. Do you believe that a sex-adjusted death rate will be very different from the crude death rate? Provide the reason for your belief.
c. Calculate a sex-adjusted death rate for Harris County in 1986.

3.14 The Pearson correlation coefficient between age and creatinine for the data in Table 3.1 was 0.319. This suggests a modest linear relation between these two variables.
a. Create a scatter plot of protein per age and creatinine. Is there a linear relationship? Are there any observations that clearly deviate from the linear trend?
b. Calculate the Pearson correlation coefficient, ignoring the one or two observations that are considered to be outliers. Which measure of correlation do you think best characterizes the strength of the relation?

3.15 The U.S. population (in 1000) in 1980 and 2000 are shown below by ethnic groups:

Ethnic Groups	1980	2000
White	194,811	229,086
Black/African American	25,531	36,594
American Indian/Alaskan Native	1,420	2,984
Asian/Pacific Islander	3,729	11,757

Source: U.S. Bureau of the Census, 2000

a. Calculate the annual growth rate, and show which group grew the fastest.
b. Project the population in 2015 by ethnic group, assuming the growth rate remains constant over time.
c. When will the 2000 population be doubled if the growth rate remains constant?

REFERENCES

Armitage, P., and G. A. Rose. "The Variability of Measurements of Casual Blood Pressure." *Clinical Science* 30:325–335, 1966.

Cochran, W. G. "The Comparison of Percentages in Matched Samples," *Biometrika* 37:356–366, 1955.

The Digitalis Investigative Group. "The Effect of Digoxin on Mortality and Morbidity in Patients with Heart Failure." *New England Journal of Medicine* 336:525–533, 1997.

Division of Nutritional Sciences, Cornell University in Cooperation with the Nutrition Surveillance Program of the New York State Department of Health. *New York State Nutrition: State of the State, 1992*. NYSDH, New York, 1992, Tables 2.5 and 3.2.

Kennedy, G. *Invitation to Statistics*. Blackwell, 1984, pp. 43–47.

Lee, E. T. *Statistical Methods for Survival Data Analysis*. Belmont, CA: Wadsworth, 1980.

Life Sciences Research Office, Federation of American Societies for Experimental Biology. *Nutrition Monitoring in the United States: An Update Report on Nutrition Monitoring*. DHHS Publ. No. (PHS) 89-1255. U.S. Department of Agriculture and the U.S. Department of Health and Human Services, Public Health Service, Washington, U.S. Government Printing Office, 1989.

National Center for Health Statistics. McDowell, A., A. Engel, J. T. Massey, and K. Maurer. "Plan and Operation of the Second National Health and Nutrition Examination Survey, 1976–80," *Vital and Health Statistics*, Ser. 1, No. 15, DHHS Publ. No. (PHS) 81-1317. Public Health Service, Washington, U.S. Government Printing Office, 1981.

National Center for Health Statistics. *Health, United States, 1990.* Hyattsville, MD: DHHS Pub. No. 91-1232. Public Health Service, 1991.

National Center for Health Statistics. *Health, United States, 1991 and Prevention Profile*. DHHS Publ. No. 92-1232. Public Health Service, Hyattsville, MD, 1992.

National Center for Health Statistics. *Health, United States, 2004 with Chartbook on Trends in the Health of Americans*. Hyattsville, MD: DHHS Pub. No. 2004-1232, 2004.

National Center for Health Statistics. www.cdc.gov/nchs/data/nhanes/growthcharts/set1/chart07.pdf, 2006.

Scott, D. W. "On Optimal and Data-Based Histograms." *Biometrika* 66:605–610, 1979.

Tarter, M. E. and R. A. Kronmal, "An Introduction to the Implementation and Theory of Nonparametric Density Estimation." *American Statistician* 30:105–112, 1976.

U.S. Bureau of the Census. *The 2000 Census of Population and Housing*. Summary Tape File 1A.

van Belle, G. *Statistical Rules of Thumb*. John Wiley, 2002, pp. 162–167.