CHAPTER 2

Graphic Display of Data

KEY TERMS

- Tables
  - Table shell
  - Box head
  - Stub
  - Cell
  - Note
  - Source
- Bar charts
  - Grouped bar chart
  - Stacked bar chart
  - 100% component bar chart
- Pie chart
- Histogram
- Frequency polygon
- Line graph
- Scatter diagram

LEARNING OBJECTIVES

At the conclusion of this chapter, you should be able to:

1. Define key terms.
2. Determine which graphic technique is appropriate for the type of information to be conveyed.
3. Determine the appropriate graphic techniques for the various scales of measurement.
4. Outline the essential components of tables.
5. Correctly prepare tables for one, two, three, and/or four variables.
6. Outline the principles for construction of bar charts and pie charts.
7. Differentiate between the following types of bar charts: one-variable bar chart, grouped bar charts, stacked bar charts, and 100% component bar charts.
8. Correctly prepare bar charts and pie charts.

9. Correctly prepare the following types of graphs: histograms, frequency polygons, line graphs, and scatter diagrams.
10. Differentiate between histograms and bar charts.
11. Differentiate between frequency polygons and line graphs.

The purpose of tables, charts, and graphs is to summarize and display data clearly and effectively. They are all means of summarizing quantities of information to the reader. Tables, charts, and graphs offer the opportunity to analyze data sets and to explore, understand, and present distributions, trends, and relationships in the data. The primary purpose of tables, charts, and graphs is to communicate information about the data to the user.

Whether the graphic technique used is a table, chart, or graph, it should

- display the data
- allow the viewer to think about what the data convey
- avoid distortion of the data
- encourage the reader to make comparisons
- reveal data at several levels, from a broad overview to the fine detail
- serve a reasonably clear purpose: description, exploration, tabulation, or decoration
- be closely related to the statistical and verbal descriptions of the data set

CONSTRUCTION OF TABLES

A **table** is an orderly arrangement of values that groups data into rows and columns. Almost any type of quantitative information can be grouped into tables. For example, we can use tables to display frequencies for such vital statistics as morbidity rates or hospital admission and discharge data. Tables are useful for demonstrating patterns and other kinds of relationships. They also serve as a basis for more visual displays of data, such as graphs and charts, where some of the detail may be lost. Because tables generally do not capture the interest of the reader, they should be used sparingly.

Table Shells

Although data cannot be analyzed until they have been collected, it is useful to prepare a **table shell** that shows how the data will be organized and displayed. It also helps one work through the data collection process in advance to ensure that once the data have been collected they can be analyzed in the manner desired. The basic shell for construction of tables appears in Exhibit 2–1. Table shells are tables that are complete except for the data. In summary, a table should be self-explanatory, even if it is taken out of its original context. A table should convey all the information necessary for the reader to understand the data. Check the table to be sure that

- It is a logical unit.
- It is self-explanatory. Ask yourself if the table can stand on its own if photocopied and removed from its context.
- All sources are specified.
- Headings are specific and understandable for every column and row.
- Row and column totals are checked for accuracy.
- Cells are not left blank; enter "0" or "-".
- Categories are mutually exclusive and exhaustive.

Exhibit 2–1 Table Shell

TITLE

Box Head		*Sex*					
		Male		*Female*		*Total*	
	Age	*No.*	*%*	*No.*	*%*	*No.*	*%*
Stub		Row Variable →→→→		→→→→		→→→→	
	<45			Column Variable			
	45–54			↓			
	55–64			↓			
	65–74			↓			
	75+			↓			

Note:

Source: Adapted from *Self-Instructional Manual for Cancer Registries, Book 7: Statistics and Epidemiology for Cancer Registries,* p. 23, United States Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute.

Consideration should also be given to alignment of data in tables. Guidelines for aligning text include the following: align text in a table to the left; text that serves as a column label may be centered; numeric values should be aligned to the right. If the numeric values contain decimals, they should be decimal-aligned. Some word processing and microcomputer statistical software programs have features that assist in the formatting of tables. The essential components of a table are outlined in Exhibit 2–2.

Exhibit 2–2 Essential Components of Tables

TITLE	The title should be as complete as possible and should clearly relate the content of the table. It should answer the following questions: • What are the data? (e.g., counts, percentages) • Who? (e.g., white females with breast cancer; black males with lung cancer) • Where are the data from? (e.g., hospital, state, community) • When? (e.g., year, month) For example: Site distribution by Age and Sex of Cancer Patients upon First Admission to General Hospital
BOX HEAD	The box head contains the captions or column headings. The heading of each column should contain as few words as possible but should explain briefly exactly what the data in the column represent.
STUB	The row captions are known as the stub. Items in the stub should be grouped to facilitate interpretation of data. For example, group ages into five-year intervals.
CELL	The box formed by the intersection of a column and a row.
Optional Items:	
NOTE	Anything in the table that cannot be understood by the reader from the title, box head, or stub should be explained by notes. Notes contain numbers, preliminary or revised numbers, or explanations for any unusual numbers. Definitions, abbreviations, and/or qualifications for captions or cell names should be footnoted. A note usually applies to a specific cell(s) within the table, and a symbol, such as ** or #, can be used to key the cell to the note. If several notes are required, it is better to use small letters than to use symbols for numbers. Note that numbers may be confused with the numbers within the table.
SOURCE	If data from a source outside your research are used, the exact reference to the source should be given. Indicating the source lends authenticity to the data and allows the reader to locate the original source if more information is needed.

Source: Adapted from *Self-Instructional Manual for Cancer Registries, Book 7: Statistics and Epidemiology for Cancer Registries,* p. 24, United States Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute.

A One-Variable Table

The most basic table is a frequency distribution with just one variable. The first column shows the values or categories of the variable represented by the data, such as age or sex. The second column shows the number of persons or events that fall into each category. A third column may be added to show the percentage of persons or events in each category. Because of rounding, column totals for percentages often add up to 99.9% or 100.1%. Even when this occurs, the total given should be 100.0%, with a footnote explaining that the difference is due to rounding. An example of a one-variable table, which displays some hypothetical admissions data, is presented in Table 2–1. The variable is sex, which is divided into two mutually exclusive categories, male and female.

Table 2–1 XYZ Hospital Admissions by Sex

Sex	*Admissions*	*%*
Male	30	60.0%
Female	20	40.0%
Total	50	100.0%

Two- and Three-Variable Tables

We can use tables to display data that have more than one variable. Data can be tabulated to show counts by two or three variables, such as age and sex. A two-variable table that is cross-tabulated is usually called a two-by-two contingency table. In Table 2–2, "lung cancer patients" is classified on two variables, race and sex; race is the row variable and sex is the column variable. Contingency tables, which will be discussed in greater detail in Chapter 9, are often used in calculating measures of association such as chi square.

Table 2–2 XYZ Hospital, Lung Cancer Patients by Race and Sex, 2004

	Sex		
Race	*Male*	*Female*	*Total*
White	316	204	520
Black	35	15	50
Total	351	219	570

Tables 2–3 and 2-4 are examples of data classified on three and four variables. Because tables classifying data on more than two variables can be quite confusing to the reader, they should be avoided if at all possible.

In a three-way classification table, it sometimes becomes quite challenging to arrange the data in a readable format. A multidimensional relationship must be shown in a two-dimensional space. In Table 2–3, we have expanded the classification of the lung cancer data to include not only race/ethnicity and sex but also geographic region. The row categories are first divided by geographic region, and then by race/ethnicity. In Table 2–4, a three-way table, types of cancer are first divided by primary site and then classified by race, and sex.

Table 2–3 Age-Adjusted Rates by Geographic Location, Race/Ethnicity and Sex, Cancer of the Colon and Rectum, 1997–2001

		Sex	
Geographic Region	*Race/Ethnicity*	*Male*	*Female*
San Francisco–Oakland	White	61.4	44.3
	Black	61.7	49.2
	American Indian/Alaska Native	7.4	0.0
	Asian or Pacific Islander	51.6	39.0
	Hispanic	60.9	38.6
Connecticut	White	70.9	52.3
	Black	64.6	51.7
	American Indian/Alaska Native	18.3	60.1
	Asian or Pacific Islander	60.2	6.0
	Hispanic	71.7	52.0
Detroit	White	69.2	48.6
	Black	80.4	60.5
	American Indian/Alaska Native	22.6	12.9
	Asian or Pacific Islander	47.6	27.1

Source: Ries, L.A.G., Eisner, M.P., Kosary, C.L., Hankey, B.F., Miller, B.A., Clegg, L., Mariotto, A., Feuer, E.J., and Edwards, B.K., (eds). SEER Cancer Statistics Review, 1975–2001. National Cancer Institute, Bethesda, MD, http://seer.cancer.gov.

Table 2–4 Number of New Cancer Cases, 1997–2001, by Selected Primary Site, Race and Sex, SEER Geographic Areas

	All Races			*Whites*			*Blacks*		
Site	*Total*	*Male*	*Female*	*Total*	*Male*	*Female*	*Total*	*Male*	*Female*
Oral & Pharynx	18,688	12,452	6,216	14,890	9,917	4,973	1,888	1,329	559
Liver	10,395	7,027	3,368	6,731	4,468	2,263	1,081	759	322
Pancreas	18,790	9,214	9,576	14,999	7,424	7,575	2,116	996	1,120
Lung & Bronchus	105,298	58,192	47,106	86,163	46,518	39,645	11,314	6,815	4,499
Hodgkin's Lymphoma	5,101	2,780	2,321	4,313	2,345	1,968	505	277	228
Colon & Rectum	91,850	46,176	45,674	74,274	37,348	36,926	9,744	4,118	4,626

Source: Ries, L.A.G., Eisner, M.P., Kosary, C.L., Hankey, B.F., Miller, B.A., Clegg, L., Mariotto, A., Feuer, E.J., and Edwards, B.K., (eds). SEER Cancer Statistics Review, 1975–2001. National Cancer Institute, Bethesda, MD, http://seer.cancer.gov.

CHARTS

Graphs and charts of various types are the best means for presenting data for quick visualization of relationships. Graphs and charts emphasize the main points and analyze and clarify relationships between variables that may otherwise remain elusive.

Regardless of the type of graph or chart being prepared, several principles of construction should be followed. First, it is important to avoid distortion of the data. To avoid distortion, the representation of numbers on the graph should be directly proportional to the numerical quantities that are being represented on the graph. It is also important to consider proportion and scale. Graphs should accommodate the eye in that they should emphasize the horizontal. Graphs should be greater in length than they are in height. The three-quarter high rule is a useful guide: the height (*y*-axis) of the graph should be three-fourths the length (*x*-axis) of the graph. A longer horizontal axis helps to point out the causal variable in more detail. Other helpful hints in preparing graphs or charts include spelling out abbreviations in a note so that misunderstandings are avoided; using colors to help clarify groupings that may appear in the graph; and using both upper- and lowercase letters in titles, as the use of all capital letters can be unfriendly to the eyes.

There are many types of charts. We will first discuss the construction of charts for data that fall into categories.

Bar Charts

One-Variable Bar Chart

We can use **bar charts** to display data for one or more variables. Bar charts are appropriate for displaying data that are categorical. The simplest bar chart is the one-variable bar chart. Each category of the variable is represented by a bar. In Figure 2–1, the bar represents one variable—the crude death rate for cancer of the trachea, bronchus, and lung—which is placed in categories by years (1990 through 1998). There is one bar representing the crude death rate for each of the nine years in the bar chart. Guidelines for construction of a bar chart are summarized in Exhibit 2–3.

Figure 2–1 Crude Death Rate by Year, Cancers of the Trachea, Bronchus, and Lung, ICD-9-CM Codes 162.0–162.9, 1990–1998

Source: United States Department of Health and Human Services, Centers for Disease Control and Prevention (CDC), CDC On-Line Database, wonder.cdc.gov.

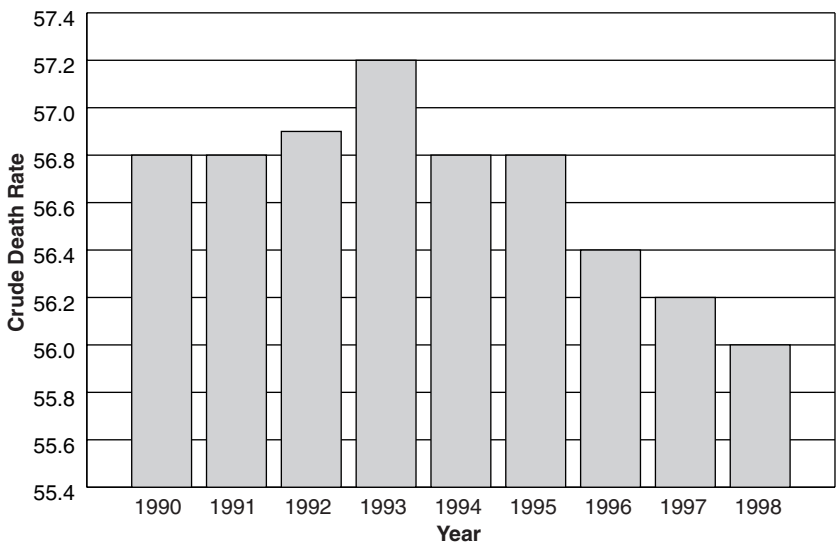


Exhibit 2–3 Guidelines for Constructing a Bar Chart

When constructing a bar chart, keep the following points in mind:

- Arrange the bar categories in a natural order, such as alphabetical order, order by increasing age, or an order that will produce increasing or decreasing bar lengths.
- The bars may be positioned vertically or horizontally.
- The bars should be of the same width.
- The length of the bars should be in proportion to the frequency of the event.
- Avoid using more than three bars (categories) within a group of bars.
- Leave a space between adjacent groups of bars but not between bars within a group.
- Code different variables by differences in bar color, shading, cross-hatching, and so on. Include a legend that interprets your code.

Source: Adapted from *Self-Instructional Manual for Cancer Registries, Book 7: Statistics and Epidemiology for Cancer Registries,* p. 251, United States Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute.

The length or height of each bar is proportional to the number of persons or events in the category. The presentation of the information in this bar chart makes it easy to see at a glance that the crude death rate was the greatest in 1993 and the least in 1998. One can also readily see that the crude death rate has been on the decline since 1995.

Bar charts may be drawn either horizontally or vertically. Figure 2–2 presents the same information that appears in Figure 2–1, but in a horizontal format. Personal preference determines the format used.

It is not uncommon to confuse a bar chart with a histogram. A bar chart is used to display data that fall into groups or categories, whereas histograms are used to illustrate frequency distributions of continuous variables. In a bar chart, the bars that represent the categories of the variables are separated, whereas in a histogram the bars are joined. A **histogram** is used to display the frequency distribution of a continuous variable, such as age. A bar chart is used to display the frequency distribution of a variable that is discrete with noncontinuous categories such as race or sex.

Computer software makes it easy to present bar charts in either two- or three-dimensional form. When bars are presented in three-dimensional form, it is sometimes difficult for the reader to estimate the true height of the bar. In a 3-D bar chart, the back edges of the bar are higher than the front edge, as in Figure 2–3. To make sure that the reader correctly interprets the bar, label the data points at a point on the bars, as shown in Figure 2–3.

Figure 2–2 Crude Death Rate by Year, Cancers of the Trachea, Bronchus, and Lung, ICD-9-CM Codes 162.0–162.9, 1990–1998

Source: United States Department of Health and Human Services, Centers for Disease Control and Prevention (CDC), CDC On-Line Database, wonder.cdc.gov.

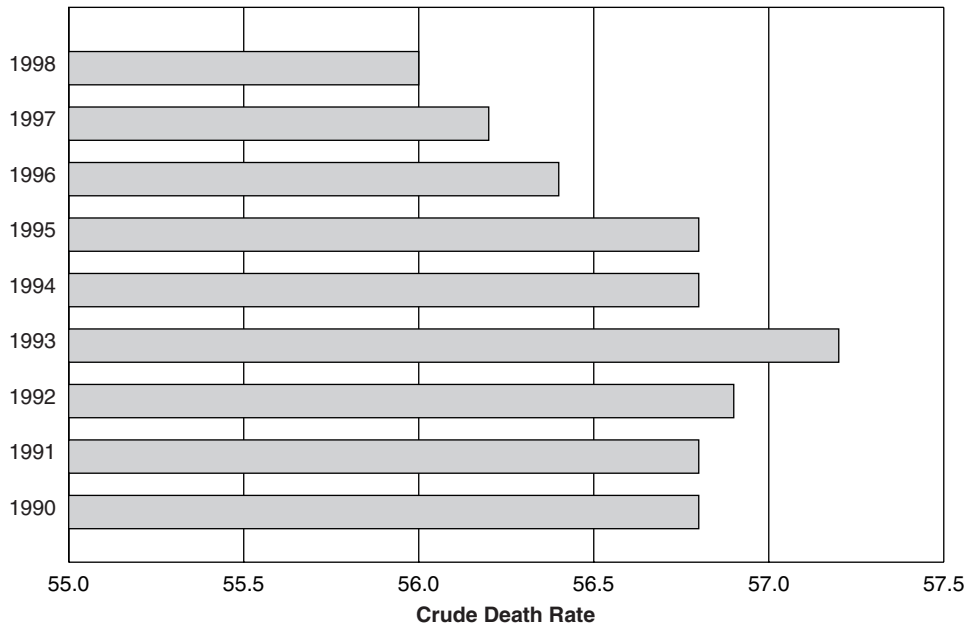
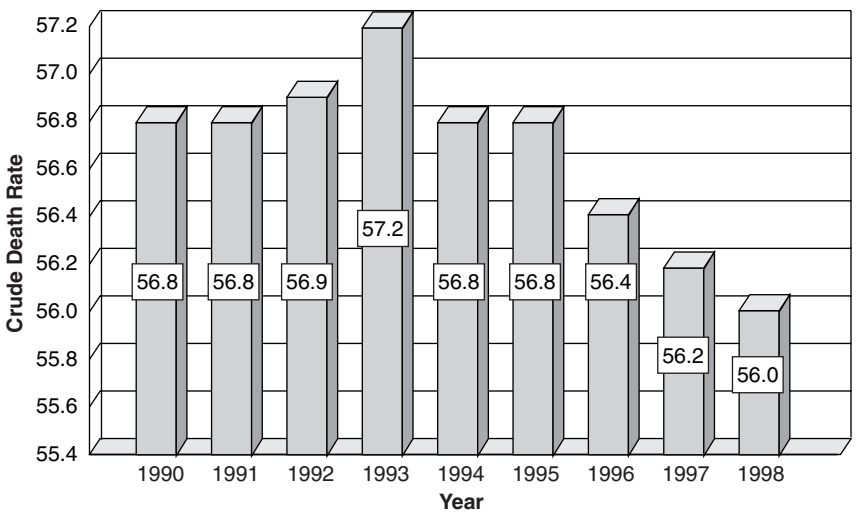


Figure 2–3 Crude Death Rate by Year, Cancers of the Trachea, Bronchus, and Lung, ICD-9-CM Codes 162.0–162.9, 1990–1998

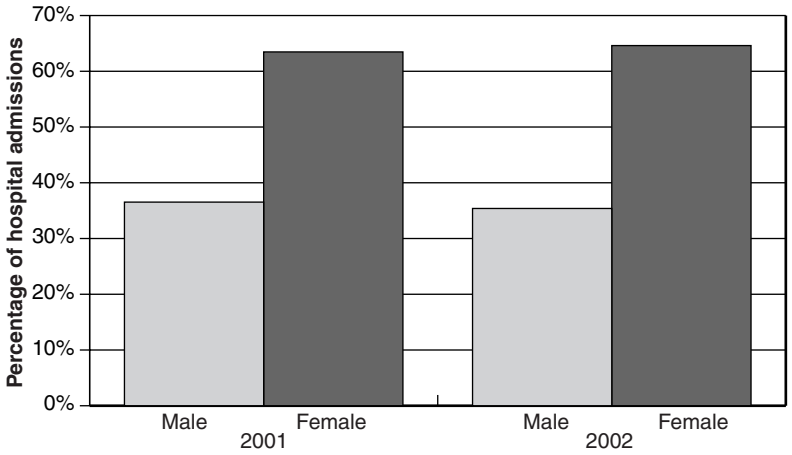
Source: United States Department of Health and Human Services, Centers for Disease Control and Prevention (CDC), CDC On-Line Database, wonder.cdc.gov.



Grouped Bar Charts

A **grouped bar chart** is used to display information from tables containing two or three variables. An example of a grouped bar chart can be demonstrated by the variable sex, which has two categories: male and female. Bars within a group are usually joined; in this case, the grouping is by year. The number of bars within a grouping should be limited to three. There must also be a legend to indicate what categories the bars represent. In viewing the grouped bar chart in Figure 2–4, we can easily see that proportionately, more women than men were admitted to the hospital for the years 2001 and 2002.

Figure 2–4 Percentage of Hospital Admissions by Sex, 2001 and 2002



Stacked Bar Charts

In a **stacked bar chart**, bar segments for each data category are stacked like building blocks on top of one another to form a single bar. In a stacked bar chart, the bar represents the total number of cases that occurred in a category; the segments of the bar represent the frequency of cases within the category. As an example, the data that appear in Table 2–5 are presented as a stacked bar chart in Figure 2–5. Each bar in the stacked bar chart represents the total number of cancer cases for a specific primary site; the bar segments represent the number of males and the number of females affected within the total number of cases.

Stacked bar charts should be used with caution, since they are very difficult to interpret. Except for the bottom category, the categories do not rest on a flat baseline. What this means is that where one category of the variable ends, the next begins. Each category rides the bumps of those below it.

From the stacked bar graph in Figure 2–5, it can be readily seen that except for cancer of the colon/rectum and pancreas, men are affected more often than women. But the exact number of cases for women in each category is difficult to determine. Stacked bar charts are deceptive, so they are often used to exaggerate or hide information.

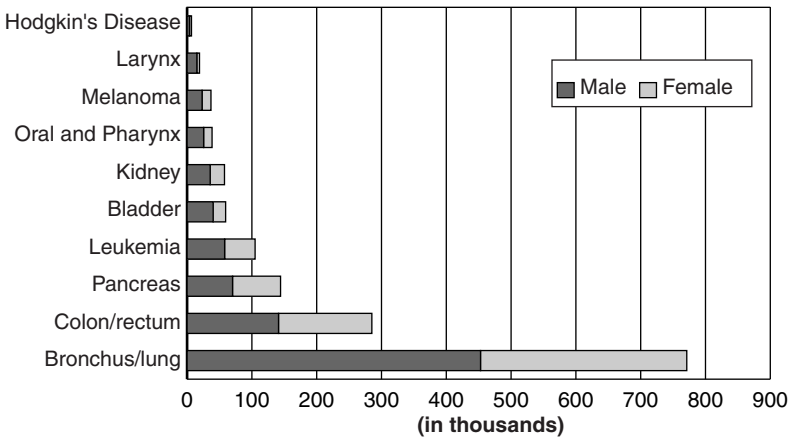
Table 2–5 Number of Deaths by Selected Primary Cancer Sites, 1997–2001

Primary Site	*Male*	*Female*	*Total*
Bronchus/lung	452,846	318,281	771,127
Colon/rectum	141,124	144,007	285,131
Pancreas	70,155	74,069	144,224
Leukemia	57,916	47,037	104,953
Bladder	40,211	19,265	59,476
Kidney	35,649	22,059	57,708
Oral & Pharynx	25,532	13,005	38,537
Melanoma	23,119	13,727	36,846
Larynx	15,069	4,080	19,149
Hodgkin's Disease	3,723	3,044	6,767
Total	865,344	658,574	1,523,918

Source: Ries, L.A.G., Eisner, M.P., Kosary, C.L., Hankey, B.F., Miller, B.A., Clegg, L., Mariotto, A., Feuer, E.J., and Edwards, B.K., (eds). SEER Cancer Statistics Review, 1975–2001. National Cancer Institute, Bethesda, MD, http://seer.cancer.gov.

Figure 2–5 Number of Deaths by Selected Primary Cancer Sites, 1997–2001

Source: SEER Cancer Statistics Review, 1975–2001. National Cancer Institute, Bethesda, MD, http://seer.cancer.gov.

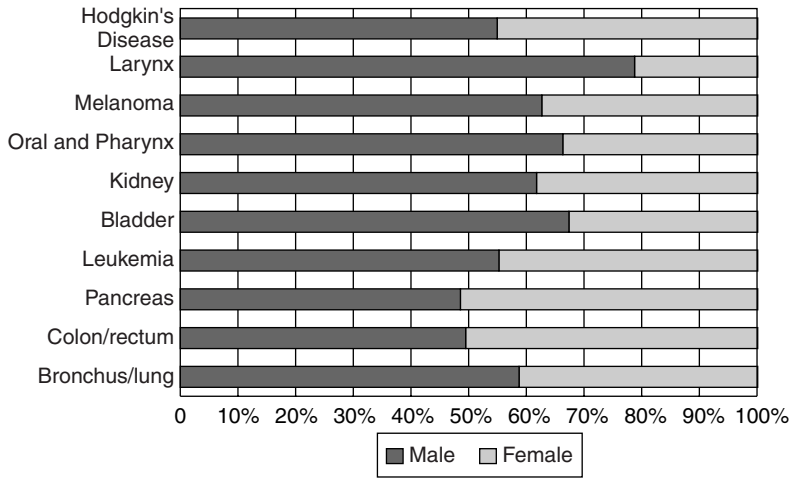


100% Component Bar Charts

The **100% component bar chart** is a variant of the stacked bar chart. In a 100% bar chart, all of the bars are of the same height and show the variable categories as percentages of the total rather than the actual values. Each bar is much like its own pie chart. A set of 100% bar charts can be used instead of multiple pie charts. This is more advantageous because it is easier to make a comparison between bars than between pies. Figure 2–6 presents the same information that appears in Figure 2–5. The stacked bars for each year represent 100% of the various types of cancer cases by sex. Each category of the sex variable is represented in terms of a percentage, in one bar.

Figure 2–6 Percentage of Cancer Deaths by Selected Primary Sites by Sex, 1997–2001

Source: SEER Cancer Statistics Review, 1975–2001. National Cancer Institute, Bethesda, MD, http://seer.cancer.gov.



Pie Charts

A **pie chart** is an easily understood chart in which the sizes of the slices show the proportional contribution of each part of the pie. We can use pie charts to show the component parts of a single group or variable. To calculate the size of each slice of the pie, first determine the proportion of the pie to be represented by each slice. Multiply the proportion by 360—the total number of degrees in a circle. The result will be the size of each slice in degrees.

In the pie chart in Figure 2–7, one slice of the pie, acute myeloid leukemia, represents 34% of the cases, or 34% of the pie. Within the pie chart, this slice of the pie equals 122.4° (360° × 0.34 = 122.4°). All other leukemia types represents 30% of the cases, and the size of its respective slices is 108° (360° × 0.30 = 108°). One category of the pie, chronic lymphoid leukemia, represents 20% of the cases, which is equivalent to 72° (360° × 0.20 = 72°). The remaining slices of the pie represent chronic myeloid leukemia, 9 percent and 32.4° of the pie, and acute lymphoid leukemia, 7 percent and 25.2° of the pie. The sum of the degrees for each slice of the pie is 360° (122.4° + 108° + 72° + 32.4° + 25.2° = 360°). The pie chart in Figure 2–7 demonstrates how the whole pie is divided into segments. By convention, the largest slices of the pie begin at 12 o'clock, as in Figure 2–7. The slices of the pie should be arranged in some logical order. In Figure 2–8, acute lymphoid leukemia appears in the 12 o'clock position. This is an example where pie slices are arranged in alphabetical order rather than according to magnitude.

It is not recommended to use pie charts to compare multiple distributions because they are not optimal for comparing components for more than one group. When components of more than one group are to be compared, a 100% component bar chart should be used.

Figure 2–7 Leukemia Cancer Deaths by Type, 1997–2001, Ordering by Magnitude of Groupings

Source: SEER Cancer Statistics Review, 1975–2001. National Cancer Institute, Bethesda, MD, http://seer.cancer.gov.

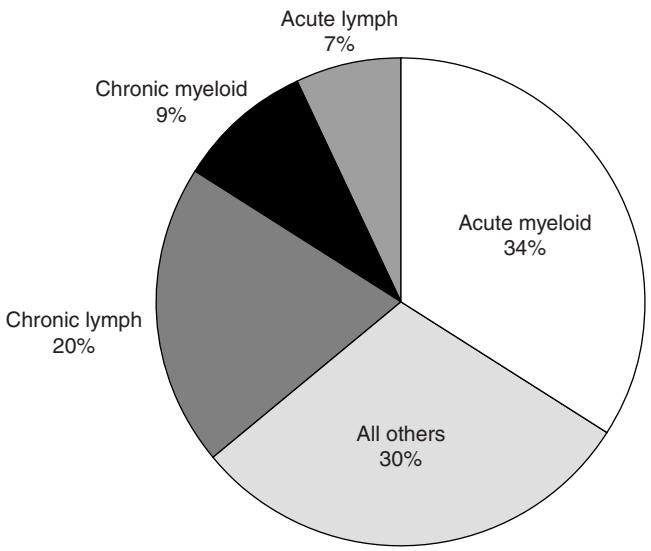
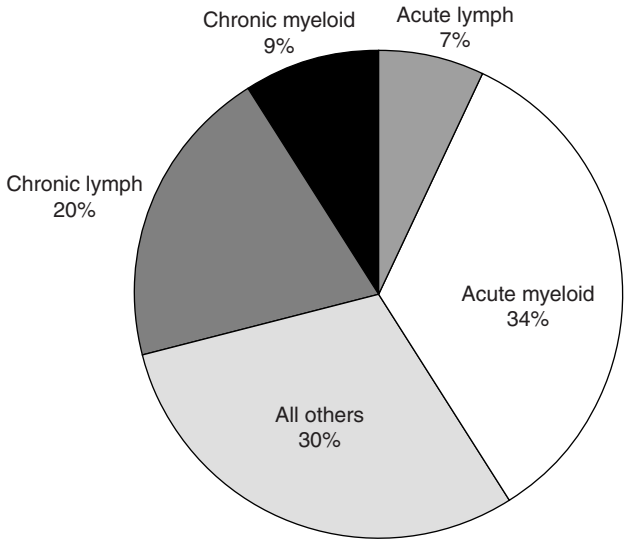


Figure 2–8 Leukemia Cancer Deaths by Type, 1997–2001, Alphabetical Order

Source: SEER Cancer Statistics Review, 1975–2001. National Cancer Institute, Bethesda, MD, http://seer.cancer.gov.



Histograms

Thus far we have discussed graphing data that are in categorical or discrete form. The techniques that will be discussed next are appropriate for data that are continuous in nature.

A histogram is appropriate for displaying a frequency distribution for one continuous variable. The frequency distribution can be presented in either number or percentage form. A histogram consists of a series of bars, each having as its base one class interval and as its height the number (frequency) or percentage of cases in that class. A class interval is a type of category; a class interval can represent one value in a frequency distribution (Figure 2–9) or a group of values in a frequency distribution (Figure 2–10).

In this type of graph, there are no spaces between the bars, since the data points represented are continuous. That is, a data point may fall anywhere in the area covered by the graph. The sum of the heights of the bars represents the total number, or 100% of the cases. When the distribution of the data needs to be emphasized more than the actual values, use a histogram. An example where the class interval represents a single value in a frequency distribution is displayed in Figure 2–9. Each bar of the histogram represents a single age, in contrast to Figure 2–10, where each bar represents an age group.

Figure 2–9 DRG 416, Septicemia, Histogram of Patients Aged 80–89

Source: United States Department of Health and Human Services, Centers for Disease Control and Prevention (CDC), CDC On-Line Database, wonder.cdc.gov.

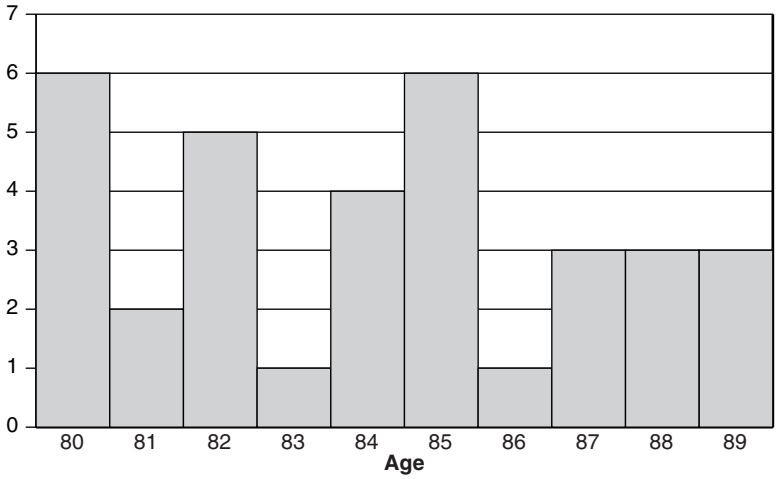
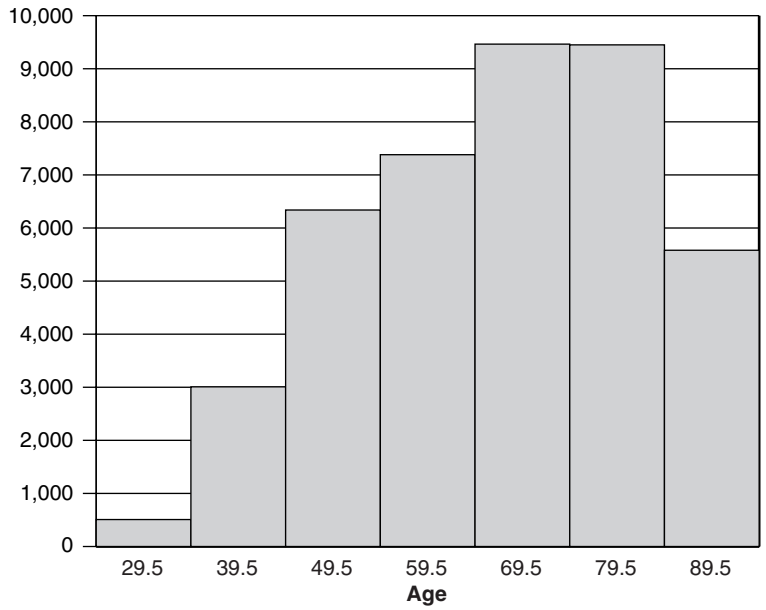


Figure 2–10 Deaths Due to Breast Cancer, ICD-9-CM Codes 174.0–174.9, by Age, 1998

Source: United States Department of Health and Human Services, Centers for Disease Control and Prevention (CDC), CDC On-Line Database, wonder.cdc.gov.



Age at death for breast cancer is the variable represented in the histogram in Figure 2–10. The values at the bottom of the *x*-axis are the midpoints of the class intervals for the following age groups:

Age Group	*Midpoint of Class Interval*
25–34	29.5
35–44	39.5
45–54	49.5
55–64	59.5
65–74	69.5
75–84	79.5
85+	89.5

In the histogram, it is clear that there are two age groups that account for most of the deaths due to breast cancer: 65–74 and 74–84.

Frequency Polygons

A **frequency polygon** can be used as an alternative to the histogram. Like a histogram, a frequency polygon is a graph of a frequency distribution. To construct a frequency polygon, simply join the midpoints at the top of each bar in the histogram (4.5, 14.5, 24.5, and so on). The advantage of the frequency polygon over the histogram is that several frequency polygons can be plotted on the same graph for comparison purposes. Frequency polygons also are easy to interpret.

When constructing a frequency polygon, make the *x*-axis longer than the *y*-axis to avoid distorting the data. The frequency of the observations is always placed on the *y*-axis, and the scale of the variables under study is placed on the *x*-axis. Frequency values are plotted at the midpoint of each class interval.

The frequency polygon in Figure 2–11 plots the same data that appear in the histogram in Figure 2–10. Since the *x*-axis represents the total distribution, *the line always starts and ends with zero.*

The frequency polygon tells us that the number of deaths due to breast cancer reaches its peak in the age groups 65 to 74 and 75 to 84. A frequency polygon presents this pattern with more clarity than the histogram.

Line Graphs

A **line graph** is often used to display time trends and survival curves. The *x*-axis shows the unit of time from left to right, and the *y*-axis measures the values of the variable being plotted.

A line graph does not represent a frequency distribution. A line graph consists of a line connecting a series of points on an arithmetic scale. Like all graphs, it should be designed so that it is easy to read. The selection of proper scales, complete and accurate titles, and informative legends is important. If a graph is too long and narrow, either vertically or horizontally, it has an awkward appearance and may exaggerate one aspect of the data. The upward trend for median charges for septicemia patients in the state of Utah is displayed Figure 2–12. The line graph is especially useful when there are a large number of values to

Figure 2–11 Deaths Due to Breast Cancer, ICD-9-CM Codes 174.0–174.9, by Age, 1998

Source: United States Department of Health and Human Services, Centers for Disease Control and Prevention (CDC), CDC On-Line Database, wonder.cdc.gov.

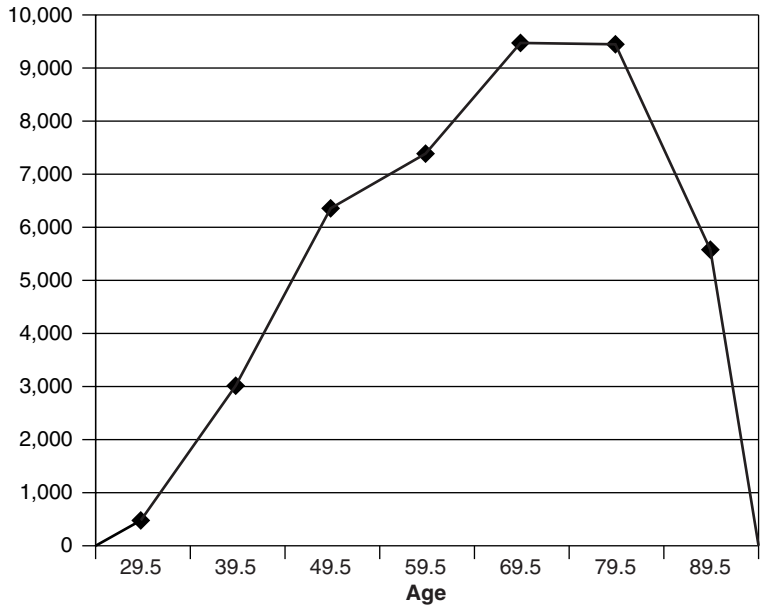
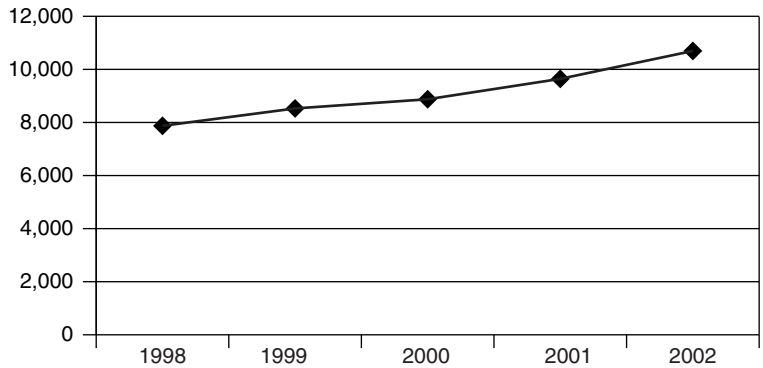


Figure 2–12 DRG 416, Septicemia Age ≥ 17, Median Charges, State of Utah, 1998–2002

Source: Utah Inpatient Hospital Discharge Dataset, Utah Office of Health Care Statistics, www.health.state.ut.us.



be plotted; that is, when you have a continuous variable with an unlimited number of possible points. It also allows the presentation of several sets of data on one graph.

Either actual numbers or percentages may be used on the *y*-axis of the line graph. Use percentages on the *y*-axis when more than one distribution is to be shown on one graph.

A percentage distribution allows comparisons between groups where the actual totals are different.

If more than one set of data is plotted on the same graph, different types of lines (solid or broken) should be used to distinguish between the lines. The number of lines should be kept to a minimum—a line graph can soon become too cluttered. Each line should be identified in a legend or on the graph itself.

There are two kinds of time-trend data: (1) point data, which reflect an instant in time, and (2) period data, which cover an average or total over a specified period of time, such as a one-year or five-year time frame. In point data, the scale marker on the *x*-axis indicates a particular point in time, such as one, two, or three years of survival. On the other hand, in plotting of period data, the horizontal scale lines are used to indicate the interval limits, and the values are plotted at the midpoint at each interval. For example:

Year of Diagnosis	*Midpoint of Interval*
1986–1988	1987
1989–1991	1990
1992–1994	1993
1995–2000	1997.5

Table 2–6 presents an example of point data that are graphed in Figure 2–13. Other examples are presented in Table 2–7 and Figure 2–14.

Figure 2–13 Relative Survival Rates by Year of Diagnosis for Kidney and Renal Pelvis Cancer, 1992–1996

Source: Miller, B.A., Clegg, L., Mariotto, A., Feuer, E.J., and Edwards, B.K., (eds). SEER Cancer Statistics Review, 1975–2001, National Cancer Institute, Bethesda, MD, http://seer.cancer.gov.

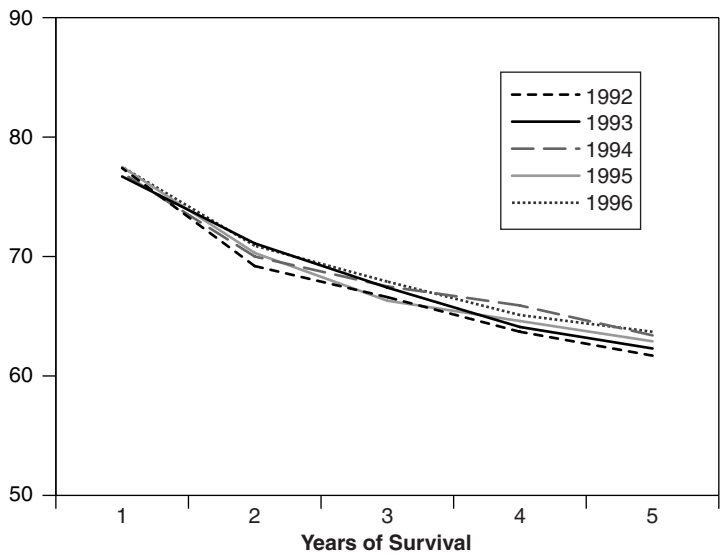


Table 2–6 Survival Rates by Year of Diagnosis, Kidney and Renal Pelvis Cancer, 1992–1996

Years of Survival	*1992*	*1993*	*1994*	*1995*	*1996*
1	77.4	76.7	77.0	77.5	77.5
2	69.2	71.1	70.0	70.3	70.9
3	66.6	67.4	67.5	66.3	67.9
4	63.7	64.1	65.9	64.6	65.1
5	61.7	62.3	63.4	62.9	63.7

Source: Ries, L.A.G., Eisner, M.P., Kosary, C.L., Hankey, B.F., Miller, B.A., Clegg, L., Mariotto, A., Feuer, E.J., and Edwards, B.K., (eds). SEER Cancer Statistics Review, 1975–2001. National Cancer Institute, Bethesda, MD, http://seer.cancer.gov.

Figure 2–14 Five-Year Survival Rates for Kidney and Renal Pelvis Cancer for Patients Diagnosed 1986–1988, 1989–1991, 1992–1994, 1995–2000

Source: Miller, B.A., Clegg, L., Mariotto, A., Feuer, E.J., and Edwards, B.K., (eds). SEER Cancer Statistics Review, 1975–2001, National Cancer Institute, Bethesda, MD, http://seer.cancer.gov.

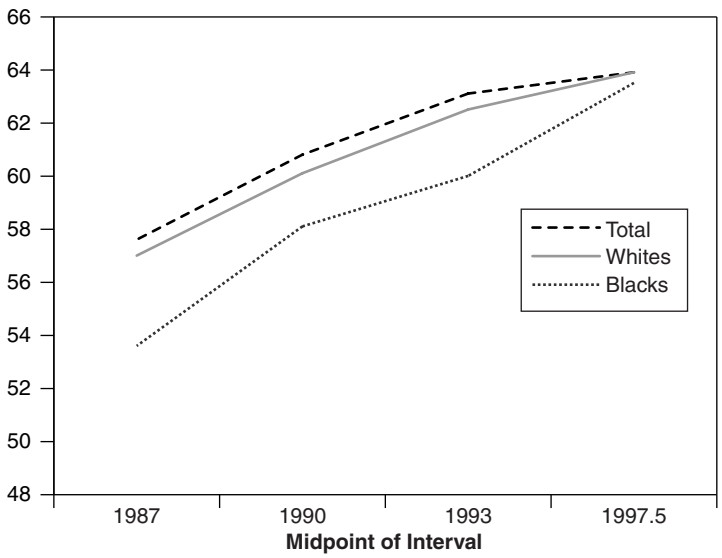


Table 2–7 Five-Year Survival Rates for Kidney and Renal Pelvis Cancer for Patients Diagnosed 1986–1988, 1989–1991, 1992–1994, 1995–2000

		Race		
Year of Diagnosis	*Midpoint of Interval*	*Total*	*Whites*	*Blacks*
1986–1988	1987	57.0	57.6	53.6
1989–1991	1990	60.1	60.8	58.1
1992–1994	1993	62.5	63.1	60.0
1995–2000	1997.5	63.9	63.9	63.5

Source: Ries, L.A.G., Eisner, M.P., Kosary, C.L., Hankey, B.F., Miller, B.A., Clegg, L., Mariotto, A., Feuer, E.J., and Edwards, B.K., (eds). SEER Cancer Statistics Review, 1975–2001. National Cancer Institute, Bethesda, MD, http://seer.cancer.gov.

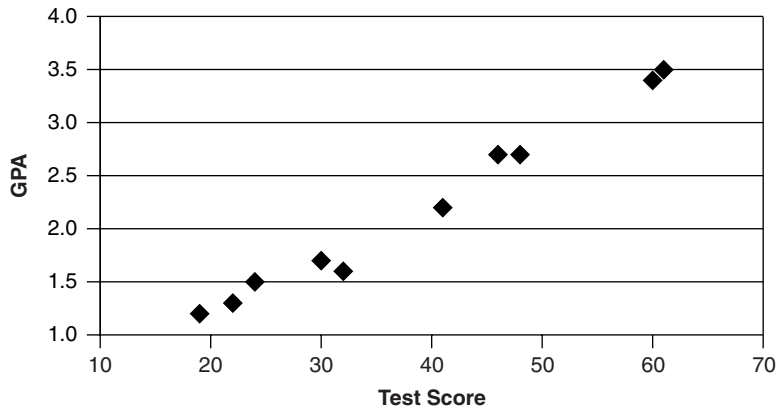
Scatter Diagrams

A **scatter diagram**, or scatter plot, is a graphic technique used to display the relationship between two continuous variables. One variable is plotted on the *x*-axis and the other is plotted on the *y*-axis. To create a scatter diagram, there must be a pair of values for every person, group, or other entity in the data set, one value for each variable. Each pair of values is plotted by placing a point on the graph where the two values intersect. To interpret a scatter diagram, analyze the overall pattern of the plotted points. Plotted points that appear to fall in a straight line indicate a linear relationship between x and y, whereas widely scattered points indicate no relationship between x and y. Table 2–8 presents hypothetical data of test scores and the grade point averages of 10 students. Figure 2–15 is a scatter plot that depicts the relationship between the two variables, test scores (x) and grade point average (y).

Table 2–8 Test Scores and Grade Point Averages of 10 Students

Student	*Test Score (x)*	*GPA (y)*
1	24	1.5
2	61	3.5
3	30	1.7
4	48	2.7
5	60	3.4
6	32	1.6
7	19	1.2
8	22	1.3
9	41	2.2
10	46	2.7

Figure 2–15 Scatter Diagram of Hypothetical Test Scores and Grade Point Average



The scatter diagram in Figure 2–15 indicates a strong linear relationship between the variables test scores and grade point average. Scatter diagrams are used to assist in interpretation of inferential statistics such as correlation and linear regression. We will discuss these topics in Chapter 8.

CONCLUSION

Tables, charts, and graphs are effective methods of summarizing data and displaying data in a clear, concise format. Tables are often used to display data, and they can be used to display data about one or more variables. An advantage of tables is that large amounts of data can be displayed and summarized, as in a four-way table. However, if too much information is included in a table, it can be confusing to the reader.

Bar charts or graphs are often used for displaying categorical data, but they are appropriate for data that are continuous in nature. Bar charts allow for quick visualization of the variable of interest. Relationships between two variables are also easily seen in a bar chart. Bar charts may take the form of a simple bar chart, a grouped bar chart, a stacked bar chart, or a 100% component bar chart. The form selected should be appropriate to the data and easily interpreted by the reader.

Pie charts are useful for displaying the parts of a whole. For example, we could display the proportion of patients admitted by third-party payer, or the proportion of burn patients admitted by severity of burn. Pie charts should be used to display proportions of one nominal-level variable; pie charts are not appropriate for comparing distributions of two or more variables.

Histograms and frequency polygons are used to display the frequency distribution of one continuous variable. Histograms and frequency polygons represent 100% of the cases in a frequency distribution; the shape of the distribution can be easily seen in these two types of graphs.

Line graphs are used to display trends in data; they are also used in survival analysis. A line graph consists of a line connecting a series of points on an arithmetic scale. To avoid distortion in the data, the graph should not be too long or too narrow. When constructing bar graphs and line graphs, the three-quarters-high rule should be used as a guide to avoid data distortion. Either actual numbers or percentages may be displayed in a line graph.

ADDITIONAL RESOURCES

Al-Assaf, A.F., and Schmele, J.A., eds. 1993. *The textbook of total quality management.* Delray Beach, FL: St. Lucie Press. 124–125.

Jones, G.E. 1995. *How to lie with charts.* San Francisco: Cybex, Inc.

Longo, D., and Bohr, D., eds. 1991. *Quantitative methods in quality management.* Chicago: American Hospital Association. 10–11.

Ries, L.A.G., Eisner, M.P., Kosary, C.L., Hankey, B.F., Clegg, L., Mariotto, A., Feuer, E.J., and Edwards, B.K., eds. *SEER Cancer Statistics Review, 1975–2001.* National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr.

SEER's Web-based Training Modules. U.S. National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) Program. http://training.seer.cancer.gov.

Self instructional manual for cancer registries, Book 7: Statistics and epidemiology for cancer registries. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, NIH Publication No. 94-3766. 1994.

Surveillance, Epidemiology, and End Results (SEER) program (www.seer.cancer.gov). *SEER Stat Databases: Incidence.* National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2004, based on the November 2003 submission.

U.S. Department of Health and Human Services, Public Health Service. 1992. *Principles of epidemiology: An introduction to applied epidemiology and biostatistics.* Atlanta, GA: USDHHS.

Utah Hospital Discharge Data, Public Dataset, www.health.state.ut.us.

Appendix 2–A

Exercises for Solving Problems

KNOWLEDGE QUESTIONS

1. Define the key terms listed at the beginning of this chapter.
2. The purpose of a table, chart, or graph is to communicate information to the data user. What questions should be considered to accomplish this objective?
3. What questions should be answered in the title of a table, chart, or graph?
4. What points should be considered when constructing a bar chart?
5. Describe the differences between a stacked bar chart and a 100% component bar chart.
6. Differentiate between a bar chart and a histogram.

MULTIPLE CHOICE

1. You want to graph the average length of stay by sex and service for the month of April. The best choice is to use a:
   a. bar graph
   b. histogram
   c. line graph
   d. pie chart
2. You want to graph the number of deaths due to prostate cancer for the years 1998 to 2002. The best choice is to use a:
   a. frequency polygon
   b. histogram
   c. line graph
   d. pie chart

3. A pie chart may be used to display the:
   a. average length of stay by year
   b. percentage of discharges by third-party payer
   c. number of discharges per year and third-party payer
   d. number of patients discharged by sex and service

4. A histogram may be used to display:
   a. discharges by age
   b. discharges by third-party payer
   c. discharges by service
   d. discharges by sex

5. You want to display the number of discharges by sex and service for 1999. The best choice is to use a:
   a. bar chart
   b. cluster line graph
   c. histogram
   d. line graph

PROBLEMS

Prepare the appropriate charts and graphs for the following problems. Include a title for each and identify the data source when indicated.

1. The admissions data in Table 2–A–1 compare actual admissions by hospital service with the budgeted number of hospital admissions for the month of January for Critical Care Hospital. Using computer graphic software, construct a bar chart that compares budgeted admissions with actual admissions. Write a short summary of the results.

Table 2–A–1 Admissions Report for January

Hospital Service	*Budgeted Admissions*	*Actual Admissions*
Medicine	769	728
Surgery	583	578
OB/GYN	440	402
Psychiatry	99	113
Physical Medicine and Rehab	57	48
Other Adult	178	191
Newborn	312	294

2. Using the data in Table 2–A–2, prepare a pie chart for January patient days by service for Critical Care Hospital.

Table 2–A–2 Patient Days by Service

Hospital Service	*Patient Days*
Medicine	4,436
Surgery	4,036
OB/GYN	1,170
Psychiatry	1,223
Physical Medicine and Rehab	1,318
Other Adult	688
Newborn	1,633

3. Table 2–A–3 contains length-of-stay data by service for the month of January for Critical Care Hospital. Construct a stacked bar chart that compares actual average length of stay with the budgeted average length of stay.

Table 2–A–3 Average Length of Stay (ALOS) by Service

Hospital Service	*Budgeted ALOS*	*Actual ALOS*
Medicine	6.39	6.09
Surgery	7.23	6.98
OB/GYN	3.22	2.91
Psychiatry	11.56	10.82
Physical Medicine and Rehab	22.98	27.46
Other Adult	3.93	3.60
Newborn	4.97	5.55

4. Organize the following statistics for the month of January into a table.

Critical Care Cancer Research Institute Statistics for January

Discharges		Discharge Service Days	
Medicine	198	Medicine	1,313
Surgery	152	Surgery	947
Gynecology	74	Gynecology	328
Otolaryngology	48	Otolaryngology	290
Average Length of Stay			
Medicine	6.6		
Surgery	6.2		
Gynecology	4.4		
Otolaryngology	6.0		

5. Exhibit 2–A–1 displays the lengths of stay for 80 patients at the Critical Care Cancer Research Institute. Construct a histogram of these data.

Exhibit 2–A–1 Lengths of Stay for 80 Patients

1	2	3	5	6	10	13	16
1	2	3	5	6	10	13	17
1	2	3	5	6	10	13	17
1	2	4	5	6	10	14	17
1	2	4	5	8	11	14	18
1	2	4	5	8	11	14	19
1	2	4	5	8	11	14	19
1	2	4	6	8	11	15	20
2	3	4	6	8	12	15	20
2	3	5	6	9	12	16	20

6. The average charges for malignant neoplasms of the large intestine and colon and average charges for malignant neoplasms of the trachea, bronchus, and lung appear in Table 2–A–4. Prepare a line graph that compares average charges for both malignancies.

Table 2–A–4 Average Total Charges, Cancers of the Colon and Lung, State of Utah, 1995–2002

Year	*Cancer of the Trachea, Bronchus, and Lung*	*Cancer of the Colon and Rectum*
1995	$13,364	$15,275
1996	$15,403	$14,622
1997	$14,958	$16,511
1998	$15,232	$16,659
1999	$16,331	$19,633
2000	$19,618	$20,424
2001	$21,216	$21,646
2002	$22,256	$22,362

Source: Utah Inpatient Hospital Discharge Dataset, Utah Office of Health Care Statistics, www.health.state.ut.us.

7. Review the data in Table 2–4. Determine the percentage of total male and female cancer cases for each site. Prepare a bar chart to display your results.