

Three

CHAPTER 3

Who, What, When, and How Much? One-Variable Description

Kind of problem	<i>Description</i>
Number of variables	<i>One</i>
Level of measurement	<i>Nominal, ordinal, or interval</i>

Statistical problems involve (1) description, (2) evaluation, or (3) estimation. Statistical problems also involve (1) one variable, (2) two variables, or (3) more than two variables. And each of these variables may have either a (1) nominal, (2) ordinal, or (3) interval (or ratio) level of measurement.

Each combination of purpose, number of variables, and level(s) of measurement is a different type of problem and calls for a different analytical method. This chapter starts with the simplest combination—description problems with one variable. The next two chapters deal with one-variable evaluation and one-variable prediction problems.

Although any event has many characteristics and attributes, sometimes an analyst is only interested in a single variable. In effect, each event is seen with blinders on. One and only one characteristic or attribute is measured. Perhaps an analyst is interested in compliance with a local ordinance establishing a 10 p.m. curfew for minors. Although the people on Main Street can be described by such variables as height, weight, sex, home address, occupation, income, or ethnicity, we are only interested in distinguishing one person from another by age—separating minors (say under the age of 18) from adults.

Description is like a snapshot. People and objects are observed at a particular instant in time. Using the photo the objects can be counted and located with respect to one another. This summary of observed events is called a *distribution*. The descriptive picture is incomplete. It does not include every event, person, or incident that comprises the status quo. It does not include every aspect of every event that we are interested in. Nor does it specify how or why the condition occurred or

28 CHAPTER 3 Who, What, When, and How Much? One-Variable Description

how it might change from one moment to the next. However, with statistical techniques this partial picture is an indicator of the usual or typical state of affairs.

One variable—a particular characteristic or attribute—can classify the observed events into categories or “types.” There are three parts to the complete description of this variable. (1) The analyst needs to *summarize* the data so that observations can be discussed without having to list the measured value for each and every case. (2) The analyst wants to know a typical value—an **average**—a point of *central tendency* among the observed values. (3) The analyst wants to know, *How typical is “typical?”*—the extent to which the data are spread out from the point of central tendency—the **dispersion** of the data. Both central tendency and dispersion need to be reported in order to completely describe a variable.

The methods for summarizing data and describing central tendency and dispersion differ by the variable’s level of measurement. The outline in Exhibit 3–1 will be a road map for the rest of the chapter. If the variable has a nominal level of measurement, the data will be summarized by a frequency table or bar chart, using a statistic called the mode to describe its central tendency and another statistic, the relative frequency of the mode, to describe its dispersion. A frequency table or bar chart can also be used to summarize the data for a variable having an ordinal level of measurement. The central tendency is indicated by the median and the dispersion by the interquartile range. An intervally measured variable is summarized by a **grouped frequency table**, a **one-way scatterplot**, a **box plot**, or a special kind of bar chart called a **histogram**. The central tendency and dispersion of an interval variable are described by the *mean* and the *standard deviation*.

*Exhibit 3–1***One-Variable Description**

Description	Level of measurement		
	<i>Nominal</i>	<i>Ordinal</i>	<i>Interval</i>
Summary of observations	Frequency table Bar chart Dot chart Pie chart	Frequency table Bar chart Dot chart	Grouped frequency table Histogram Box plot One-way scatterplot
Central tendency	Mode	Median	Mean Median
Dispersion	Relative frequency of the mode	Interquartile range	Standard deviation

3.1 *One Variable, Nominal Level of Measurement*

How prevalent is the mayor-council form of city government? The problem requires the description of one variable—one observable characteristic or attribute.

To address this problem, we should begin with the set of diagnostic questions:

What are the units of analysis?

How many units of analysis have been observed?

How many cases (data records) are in the sample?

The individual event that interests the analyst is the adoption of a particular form of government by a city. How often do different outcomes of this event occur? *Cities* are the units of analysis. For this example the related events—defined by size, space, and time—is the set of all cities with more than 5,000 residents in the state of Washington in 2004. These characteristics fix the scope of the inquiry. That is, they define the **statistical population** (cities in Washington)—the statistical **sample** (cities with a population more than 5,000 in 2004)—and the **operational definition** (the legal framework for policymaking). In 2004 there were 106 cities in Washington with a population over 5,000.

What kind of problem is presented?

This is a description problem. We want to know *how often* the mayor-council form of government is observed among the cities in Washington and *how prevalent* the mayor-council form is in relationship to the other possible forms of government.

What are the variables that are being measured for each unit of analysis?

How many variables are involved in the problem: one, two, or more than two?

Each city has a local government. Although the local governments may differ in many ways, we are interested in only characteristic—one variable—the *form of government*.

What is the level of measurement for each variable in the problem: nominal, ordinal, or interval?

Form of government is a **categorical variable**—meaning that there are only a limited number of possible forms of government that each city could choose. Among cities in Washington the values for the variable *form of government* are *mayor-council*, *council-manager*, and *commission*. Each of these designations can be represented by a numeric value. The mayor-council form of government category can be called category 1; the council-manager form of government

30 CHAPTER 3 Who, What, When, and How Much? One-Variable Description

category 2; and the commission form of government category 3. The numbers are only labels—a summary *name* given to each value and its operational definition. The available choices are mutually exclusive—each city has one, and only one, form of government. Thus, this variable has a nominal level of measurement.

A **data file** lists the measured, recorded observations of a set of related events. Exhibit 3–2 presents the data file for the 106 cities in Washington with more than

Exhibit 3–2

Data File: Form of Government in Cities over 5,000 Population in Washington

City/Town	Population	Form of government	
		Category	Numeric code
Seattle	573,000	mayor-council	1
Spokane	198,700	mayor-council	1
Tacoma	198,100	council-manager	2
Vancouver	154,800	council-manager	2
Bellevue	115,500	council-manager	2
Everett	97,500	mayor-council	1
Federal Way	85,800	council-manager	2
Spokane Valley	85,010	council-manager	2
Kent	84,920	mayor-council	1
Yakima	79,480	council-manager	2
Bellingham	72,320	mayor-council	1
Kennewick	60,410	council-manager	2
Lakewood	58,850	council-manager	2
Renton	56,840	mayor-council	1
Shoreline	52,500	council-manager	2
Redmond	47,600	mayor-council	1
Auburn	47,470	mayor-council	1
Kirkland	45,740	council-manager	2
Pasco	44,190	council-manager	2
Richland	43,520	council-manager	2
Olympia	43,330	council-manager	2
Edmonds	39,860	mayor-council	1
Sammamish	38,640	council-manager	2
Puyallup	35,830	council-manager	2
Longview	35,430	council-manager	2
Lynnwood	34,830	mayor-council	1
Bremerton	34,580	mayor-council	1
Lacey	33,180	council-manager	2
Burien	31,040	council-manager	2
Bothell	31,000	council-manager	2
University Place	30,980	council-manager	2

(continued)

*Exhibit 3–2***Data File: Form of Government in Cities over 5,000 Population in Washington
(continued)**

City/Town	Population	Form of government	
		Category	Numeric code
Walla Walla	30,630	council-manager	2
Marysville	29,460	mayor-council	1
Wenatchee	29,320	mayor-council	1
Des Moines	28,960	council-manager	2
Mount Vernon	28,210	mayor-council	1
Pullman	26,590	mayor-council	1
SeaTac	25,140	council-manager	2
Bainbridge Island	22,200	mayor-council	1
Oak Harbor	21,720	mayor-council	1
Mercer Island	21,710	council-manager	2
Mountlake Terrace	20,390	council-manager	2
Mukilteo	19,360	mayor-council	1
Kenmore	19,290	council-manager	2
Port Angeles	18,640	council-manager	2
Maple Valley	17,870	council-manager	2
Tukwila	17,110	mayor-council	1
Issaquah	17,060	mayor-council	1
Ellensburg	16,700	council-manager	2
Covington	16,610	council-manager	2
Aberdeen	16,450	mayor-council	1
Moses Lake	16,340	council-manager	2
Monroe	15,920	mayor-council	1
Anacortes	15,700	mayor-council	1
Camas	15,460	mayor-council	1
Centralia	15,340	council-manager	2
Arlington	14,980	mayor-council	1
Battle Ground	14,960	council-manager	2
Sunnyside	14,710	council-manager	2
Bonney Lake	14,370	mayor-council	1
Mill Creek	14,320	council-manager	2
Tumwater	12,950	mayor-council	1
Lake Forest Park	12,730	mayor-council	1
Kelso	11,820	council-manager	2
Washougal	11,350	mayor-council	1
Enumclaw	11,190	mayor-council	1
Lynden	10,480	mayor-council	1
West Richland	10,210	mayor-council	1
Woodinville	10,140	council-manager	2

(continued)

*Exhibit 3–2***Data File: Form of Government in Cities over 5,000 Population in Washington
(continued)**

City/Town	Population	Form of government	
		Category	Numeric code
Cheney	10,070	mayor-council	1
Sedro-Woolley	9,800	mayor-council	1
Ferndale	9,750	mayor-council	1
Edgewood	9,460	council-manager	2
Toppenish	9,000	council-manager	2
Sumner	8,940	mayor-council	1
Newcastle	8,890	council-manager	2
Hoquiam	8,875	mayor-council	1
Port Townsend	8,745	council-manager	2
Shelton	8,735	commission	3
Grandview	8,705	mayor-council	1
Snohomish	8,700	council-manager	2
College Place	8,690	mayor-council	1
East Wenatchee	8,300	mayor-council	1
Port Orchard	8,250	mayor-council	1
Burlington	7,550	mayor-council	1
Poulsbo	7,450	mayor-council	1
Clarkston	7,280	mayor-council	1
Lake Stevens	7,185	mayor-council	1
Chehalis	6,990	council-manager	2
Ephrata	6,930	mayor-council	1
Gig Harbor	6,765	mayor-council	1
Selah	6,740	mayor-council	1
Brier	6,475	mayor-council	1
Normandy Park	6,385	council-manager	2
Snoqualmie	6,345	mayor-council	1
Steilacoom	6,175	mayor-council	1
Othello	6,120	mayor-council	1
Milton	6,100	mayor-council	1
Fircrest	6,080	council-manager	2
Pacific	5,770	mayor-council	1
Union Gap	5,695	mayor-council	1
Duvall	5,595	mayor-council	1
DuPont	5,410	mayor-council	1
Quincy	5,265	mayor-council	1
Liberty Lake	5,255	mayor-council	1
Prosser	5,045	mayor-council	1

Source: Municipal Research and Services Center of Washington, Seattle, Washington
www.mrsc.org/cityprofiles/citylist.aspx

5,000 residents. Note that the data file is structured so that the data are listed sequentially by another variable—*population*. This is a convenient but not essential aspect of the data file. It eases the task of creating the data file and checking the correctness of the observations.

Mayor-council:	An elected council serves as the legislative body with a separately elected head of government.
Council-manager:	The mayor and the council make policy and an appointed manager is responsible for the administration of the city.
Commission:	A board of elected commissioners serves as the legislative body and each commissioner is also responsible for administration of one or more departments.

How prevalent is the mayor-council form of government? The 106 observations are summarized by classifying the cities according to *form of government* and counting the number of cities in each category. The result is a **frequency distribution**—a count of the number of times each value of the variable occurs in a data set. The frequency distribution is shown by means of a **frequency table**. Exhibit 3–3 presents a frequency table for these data. The table lists the *number* of Washington cities having each of the three possible values of *form of government*. The number of observations of each value is also called the **absolute frequency**. Sixty cities in Washington have the *mayor-council* form of government.

Prevalency requires that the absolute frequency for each value be expressed as a proportion of the total number of observations. This proportion is called the **relative frequency**. The frequency table lists the relative frequency of each value. Each relative frequency is found by dividing the absolute frequency count for the value by the total count of observations in the data file. The relative frequencies can be expressed as percentages by multiplying each proportion by 100%. The

Exhibit 3–3

Frequency Table: Form of Government, Washington Cities over 5,000 Population

Value	Form of government	Absolute frequency (number of observations)	Relative frequency	
			Proportion	Percentage
1	Mayor-Council	60	0.566	56.6%
2	Council-Manager	45	0.425	42.5
3	Commission	1	0.009	0.9
Total		106	1.000	100.0%

Source: Municipal Research and Services Center of Washington, Seattle, Washington
<http://www.mrsc.org/cityprofiles/citylist.aspx>

34 CHAPTER 3 Who, What, When, and How Much? One-Variable Description

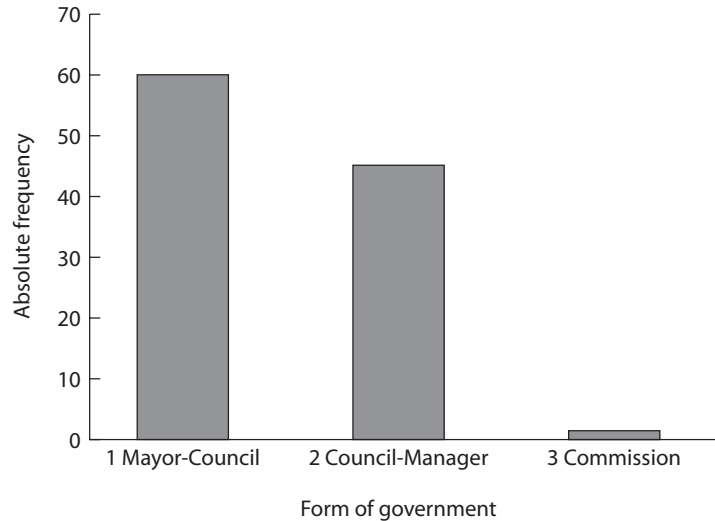
prevalence—the relative frequency—of the mayor-council form of government among cities in Washington is 56.6%.

$$\text{Relative frequency of the mayor-council form of government} = \frac{60}{106} = 0.566$$

A **bar chart** is a common graphic technique for depicting the absolute frequency distribution for one variable having a nominal level of measurement. Each category has a separate bar. The length of each bar against a one-dimensional scale indicates absolute frequency (see Exhibit 3–4). The chart may be oriented so that the bars are either horizontal or vertical. A horizontal bar chart provides more space for labeling the nominal values. The same color or fill pattern should be used throughout. (However, if the frequency of one category is of particular interest, it can be highlighted by a contrasting color or pattern.) The frequency scale should begin at 0 and have no breaks. The overall size of the chart and number of values are the principal determinants of the width and spacing of the bars. In general, the spacing between the bars should be about half the width of the bars. The nominal categories can be arranged in the chart so that the bars are arrayed from largest to smallest. This technique helps us perceive small differences in the absolute frequency between several categories.

Exhibit 3–4

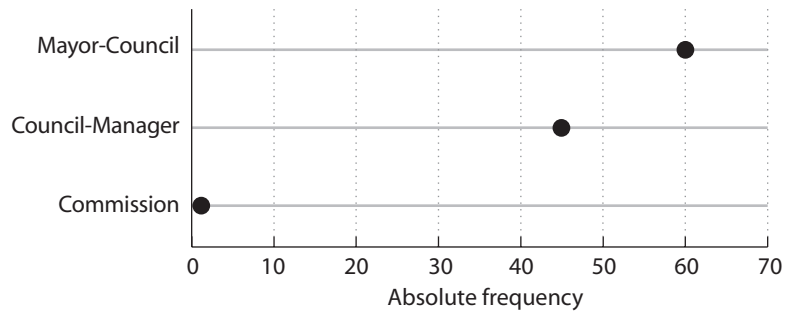
Bar Chart: Form of Government in Cities over 5,000 Population in State of Washington



Source: Municipal Research and Services Center of Washington, Seattle, Washington
<http://www.mrsc.org/cityprofiles/citylist.aspx>

Exhibit 3–5

Dot Chart: Form of Government, Washington Cities over 5,000 Population

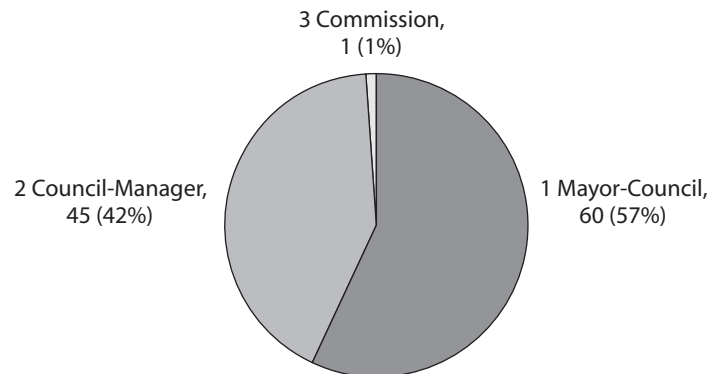


An improved display is the **dot chart**. In this display a bold dot along a faint line indicates the absolute frequency and each data value is given equal visual emphasis (see Exhibit 3–5).

Mayor-council is the most frequently occurring value of *form of government* for cities in Washington. This statistic is called the **mode**. But how typical is it? The relative frequency table indicates the dispersion of the distribution. The relative frequency for *mayor-council* of 56.6% indicates that other forms of government are also important (especially *council-manager* with a relative frequency of

Exhibit 3–6

Pie Chart: Form of Government, Washington Cities over 5,000 Population



Source: Municipal Research and Services Center of Washington, Seattle, Washington
<http://www.mrsc.org/cityprofiles/citylist.aspx>

42.4%). For a variable with a nominal level of measurement, the **relative frequency of the mode** is the summary indicator of dispersion.

A **pie chart** shows the relative frequency distribution of a variable having several nominal categories. A circle depicts the entire sample. The relative frequency of each nominal category is shown as a slice of this figure, as in Exhibit 3–6. The eye is drawn to the nominal category having the largest relative frequency (the mode). In drawing a pie chart each slice should be labeled with its nominal value and relative frequency (rather than using a legend). The number of units of analysis (sample size) should be indicated in the title so that the reader can reproduce the absolute frequency distribution if needed. The pie chart identifies the modal category and conveys a general sense of the dispersion in the distribution. This, and more, information is shown just as easily, and more accurately, in a frequency table or dot chart.

3.2 *One Variable—Ordinal Level of Measurement*

The tabular and graphic summary of data having an ordinal level of measurement is the same as that for a nominal level of measurement—the frequency table and bar chart.

The Insurance Service Organization (ISO) collects data on the capabilities of more than 44,000 fire-response jurisdictions throughout the United States, ranging from rural fire districts to communities of various population sizes including large cities. ISO assigns a Public Protection Classification (PPC) ranking to each jurisdiction. The PPC classification is on an ordinal scale from 1 to 10. Class 1 represents “exemplary public protection,” and Class 10 indicates that the area’s fire-suppression program doesn’t meet ISO’s minimum criteria. The PPC rank affects local fire insurance premiums for businesses and homeowners. Exhibit 3–7 presents the countrywide ISO-PPC frequency distribution and Exhibit 3–8 is a bar chart of this distribution.

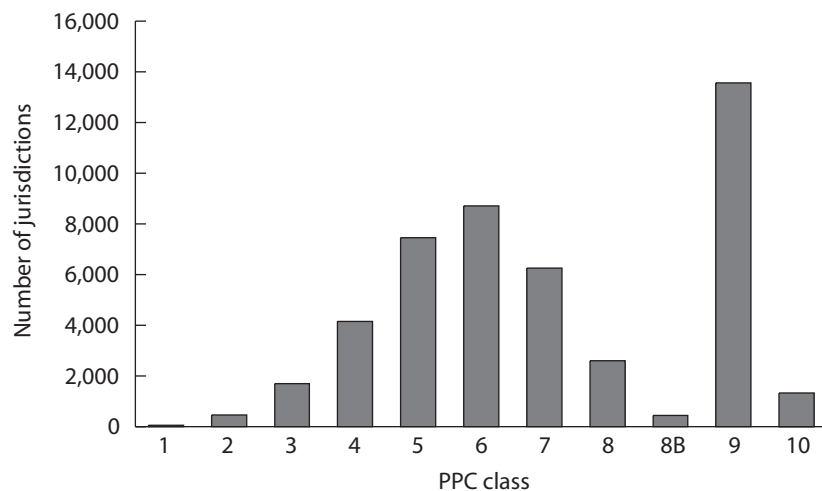
The modal ISO category is Class 9 with a relative frequency of 29.04%. However, a better indicator of central tendency for a variable having an ordinal level of measurement is the *median*. The **median** is the value of the case at the “middle-position” of a frequency distribution. One half of the units of analysis will have values that are greater than the median—and the other half will have values less than the median.

To find the median value of a variable, first sort the cases in ascending order so that the values are arranged from lowest to highest. Then find the middle case. The median is the value taken on by that middle case. (With an even number of units of analysis, there are two middle cases. In this situation, the median is the average of the values taken on by these two middle cases.)

The easiest way to find the median value is to construct a **cumulative frequency distribution**. The cumulative frequency for any value is found by adding its relative frequency to the sum of the relative frequencies for all lower values.

*Exhibit 3–7***Distribution of Insurance Service Organization (ISO) Public Protection Classification (PPC) among U.S. Fire-Response Jurisdictions**

PPC class	Number of jurisdictions	Relative frequency	Cumulative frequency
1	48	0.10%	0.10%
2	453	0.97%	1.07%
3	1,691	3.62%	4.69%
4	4,154	8.90%	13.59%
5	7,460	15.98%	29.57%
6	8,702	18.64%	48.20%
7	6,258	13.40%	61.60%
8	2,601	5.57%	67.17%
8B	441	0.94%	68.12%
9	13,560	29.04%	97.16%
10	1,328	2.84%	100.00%
Total	46,696		

*Exhibit 3–8***Bar Chart: Distribution of Insurance Service Organization (ISO) Public Protection Classification (PPC) among U.S. Fire-Response Jurisdictions**

38 CHAPTER 3 Who, What, When, and How Much? One-Variable Description

Examine the cumulative frequency distribution and find the category that contains the middle case—that is, the 50% mark of the distribution. The median ISO classification is Class 7.

The indicator of dispersion for a variable having an ordinal level of measurement is the **interquartile range**, Q_1 to Q_3 , where Q_1 is the value below which one fourth of the cases fall, with three fourths above, and Q_3 is the value below which the values of three fourths of the cases fall, with the remaining 25% of the cases having higher values. (Q_2 is the value of the middle case, the median.)

One half, 50%, of the cases will fall between the first and third quartiles. But this does not mean that the range from the median value to Q_1 equals the range from the median value to Q_3 —the median is not necessarily at the midpoint of the interquartile range. The analyst should report the values of both the first and third quartiles. These values make it possible to see whether the clustering of cases above the median is different from that below the median.

The cumulative distribution eases identification of the interquartile range. Look for the value taken on by the cases at the 25th percentile and at the 75th percentile in order to specify the interquartile range. The interquartile range, Q_1 to Q_3 , is Class 5 to Class 9.

3.3 One Variable—Interval or Ratio Level of Measurement

The indicator of central tendency for a variable having an interval or ratio level of measurement is the arithmetic **mean**. This is our conventional notion of the average (the mode and median are also averages).

The value of the mean is computed by dividing the sum of the values for all the cases in the data set by the number of cases. The equations in Exhibit 3–9 show this computation. The subscript “ i ” identifies each case. The symbol “ i ” takes on the value 1 for the first case, 2 for the second case, 3 for the third case, and so on up to however many cases there are in the data set—which we will designate by “ N ” if the data set includes the entire population and by “ n ” if the data set is a

Exhibit 3–9

Computation of the Arithmetic Mean

Data	Observation	Size	Mean
Population	x_i	N	$\mu = \frac{\sum x_i}{N}$
Sample	x_i	n	$\bar{x} = \frac{\sum x_i}{n}$

sample. The symbol “ x ” stands for the value of the variable and “ Σ ” (the uppercase Greek letter *sigma*) indicates the arithmetic operation “find the sum.” The resultant value of the mean is represented by “ μ ” (the lowercase Greek letter *mu*) if we are describing a statistical population and by “ \bar{x} ” (pronounced *x*-bar) when our observations are a statistical sample.

Radon is a colorless, odorless radioactive gas produced by the radioactive decay of uranium and radium in soil and rocks. If inhaled in large quantities, it can cause lung cancer. In houses with little ventilation, the concentration of radon can build up to relatively high levels. The U.S. Environmental Protection Agency estimates that from 5,000 to 20,000 lung cancer deaths per year in the United States are due to radon exposure. In 1989–1990 a team from the California Department of Health Services measured residential radon concentrations in Ventura County and northwestern Los Angeles County by installing radon samplers for one year in participating households. Exhibit 3–10 presents data for eight households in one zip code area—Beverly Hills 90210. (Altogether, 862 households participated—among 49 zip code areas.) A radon concentration exceeding 4 picocuries (pCi) per liter is considered to be “high.”

The mean radon concentration is the sum of the individual observations ($\Sigma x_i = 14.25$) divided by the number of observations ($n = 8$).

$$\bar{x} = \frac{\Sigma x}{n} = \frac{14.25}{8} = 1.78125$$

$$\bar{x} \approx 1.78 \text{ pCi/liter}$$

Exhibit 3–10

Radon Concentration: Eight Residences in Beverly Hills 90210

Household	Radon concentration (picocuries per liter)
1	1.50
2	3.50
3	0.50
4	2.00
5	0.50
6	4.60
7	0.75
8	1.00

Source: Kai-Shen Liu, Yu-Lin Chang, Steven B. Hayward, and Fan-Yen Huang, Survey of Residential Radon Levels in Ventura County and Northwestern Los Angeles County, Indoor Air Quality Program, California Department of Health Services, 2151 Berkeley Way, Berkeley, CA 94704, September 1991.

40 **CHAPTER 3** Who, What, When, and How Much? One-Variable Description

The indicator of dispersion for a variable having an interval or ratio level of measurement is based upon the deviation, or difference, between each observed value and the central tendency value, the mean— $(x_i - \mu)$. Some units of analysis (not necessarily half) will have observed values greater than the mean—resulting in a positive value for the deviation; and some will be less than the mean—and thus have negative values for the deviation. However, the net deviation—the sum of all deviations about the mean—will always be zero. The positive deviations will be balanced by the negative deviations just like bringing a see-saw or weighing scale into balance with different weights at different distances. Many small deviations in one direction are needed to offset a single large deviation in the other. In order to give greater weight to the larger deviations and avoid the problem of the net deviation being zero, each deviation is squared (multiply each value of deviation by itself). The mean of the squared deviations is found by summing and dividing by the number of cases. This is a measure of dispersion called the **variance**—symbolized by “ σ^2 ” (the lowercase Greek letter *sigma*, squared) for a statistical population and s^2 for a statistical sample.

However, another indicator of dispersion is more useful—the **standard deviation**. The standard deviation is the square root of the variance. Its usefulness can be realized by looking at the “units” of variance and standard deviation. If a variable, say height, were measured in feet, then the variance would be represented in feet squared, suggesting area rather than distance. The standard deviation would have units of feet and is conceptually easier to understand as an indicator of dispersion.

The equations in Exhibit 3–11 define the variance and standard deviation for population and sample data. Defining the sample variance with $n - 1$ as the divisor rather than n is a common convention. The reason is that we will want to use the sample variance as an estimate of the population variance. If n were used in the denominator this estimate would be biased. Imagine an infinitely large population from which we draw all possible samples of size n , compute the dispersion of each sample, and then average these values. With n in the denominator the final result

Exhibit 3–11

Computation of the Variance and Standard Deviation

Data	Variance	Standard deviation
Population	$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$	$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$
Sample	$s^2 = \frac{\sum (x_i - \bar{x})^2}{(n - 1)}$	$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(n - 1)}}$

*Exhibit 3–12***Computation of the Sample Standard Deviation: Radon Concentration—Eight Residences in Beverly Hills 90210**

Household	Radon concentration (pCi/liter)	Deviation	Squared deviation
1	1.50	$(1.50 - 1.78125) = -0.28125$	0.079102
2	3.50	$(3.50 - 1.78125) = 1.71875$	2.954102
3	0.50	$(0.50 - 1.78125) = -1.28125$	1.641602
4	2.00	$(2.00 - 1.78125) = 0.21875$	0.047852
5	0.50	$(0.50 - 1.78125) = -1.28125$	1.641602
6	4.60	$(4.60 - 1.78125) = 2.71875$	7.391602
7	0.75	$(0.75 - 1.78125) = -1.03125$	1.063476
8	1.00	$(1.00 - 1.78125) = -0.78125$	0.610352
Mean	1.78125		
Total			15.42969

Standard deviation

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

$$s = \frac{15.42969}{8 - 1}$$

$$s \approx 1.48 \text{ pCi/liter}$$

would be less than the population dispersion. However, if $n - 1$ is used as the divisor the result would exactly equal the population dispersion.¹

Exhibit 3-12 shows the computation of standard deviation for the observed radon concentration in the eight Beverly Hills residences.

The radon example only has eight observations. However, many description problems involve data sets with hundreds or thousands of cases. Consider, for example, how the U.S. Internal Revenue Service might summarize data on the assets of private foundations.

¹CAUTION: Many pocket calculators and computer programs allow us to compute the variance and standard deviation with a simple press of a button or a short command. Be sure to understand exactly what the calculator or computer program does. Is the computational result the population variance or the sample variance? Is the denominator N or $n - 1$?

42 CHAPTER 3 Who, What, When, and How Much? One-Variable Description

Private foundations, organized for charitable purposes, are exempt from income taxes. About 6,300 “operating foundations” conduct their own charitable activities, such as museums, while more than 70,000 “nonoperating foundations” provide charitable support indirectly, principally by grants to charitable organizations.

Constructing a frequency table would not help summarize the data because each foundation has its own unique value for the variable *assets*. Each value of *assets* has a frequency count of one. Thus, all of the foundations would be listed, just as they are in the data table—a cumbersome, confusing, and disorganized presentation. The analyst needs some means of describing a large data set short of listing every case.

One way to present a summary description is to change the level of measurement—from ratio to ordinal—by *grouping* the data. Grouping is accomplished by dividing the scale of measurement—in this case, dollars—into ranges called *classes*. The ranges are selected so that there are no gaps between any of the classes and the “width” of each class range (distance along the scale) is constant. The frequency distribution is described by sorting the cases into the selected classes and then counting the frequency with which each class is observed. There should be a sufficient number of classes so that the cases do not all fall into only two or three groups, yet not so many that small absolute frequency counts of only 0, 1, or 2 are commonplace. The dividing point between one class and the next should be unambiguous (for example, 0 up to 10, 10 up to 20, 20 up to 30, . . . or $0 \leq x < 10$, $10 \leq x < 20$, $20 \leq x < 30$, . . .). Not 0 to 10, 10 to 20, 20 to 30, . . .). The ranges should be chosen so that the class midpoints are easily identified (for example, class midpoint values of 5, 15, 25 are preferable to 2.5, 7.5, 12.5, 17.5, . . .).

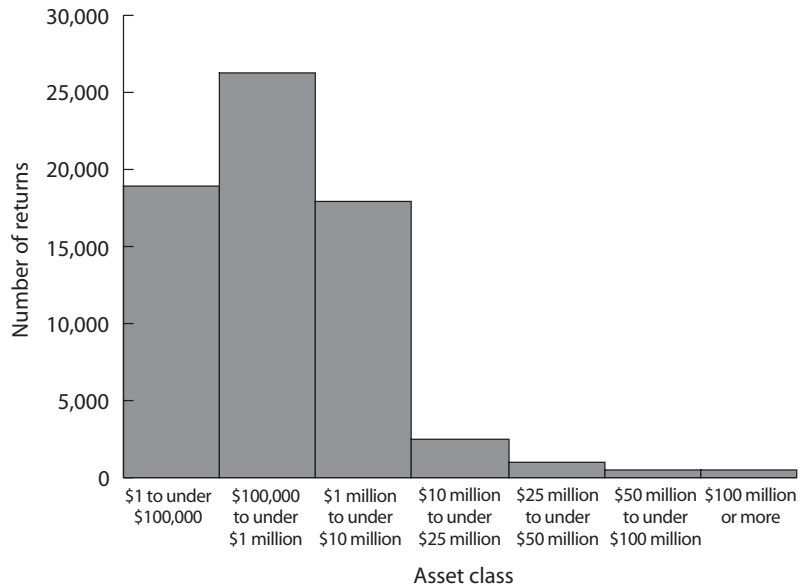
Exhibit 3–13 presents the data on nonoperating private foundations grouped into seven ranges of assets ranging from foundations reporting assets less than \$100,000 to foundations reporting assets exceeding \$100 million. Exhibit 3–14 shows a bar chart of these data. When drawing a bar chart for grouped data, the bars should not be separated—but should be drawn side by side—because the interval or ratio scale for the variable is continuous. This special type of bar chart—for one-variable, grouped data—is called a *histogram*. The range of values for the variable of interest is shown on the horizontal scale. The height of the respective bars indicate the absolute frequency (number of observations) in each group. A histogram shows both the typical and unusual aspects of a frequency distribution. The tallest bar indicates the modal value. The several adjacent tall bars indicate the approximate central tendency of the distribution. The shape of the histogram shows the dispersion of the observations. In Exhibit 3–14 the modal central tendency is the \$100,000 to \$1 million class and there are comparatively few foundations reporting assets more than \$25 million.

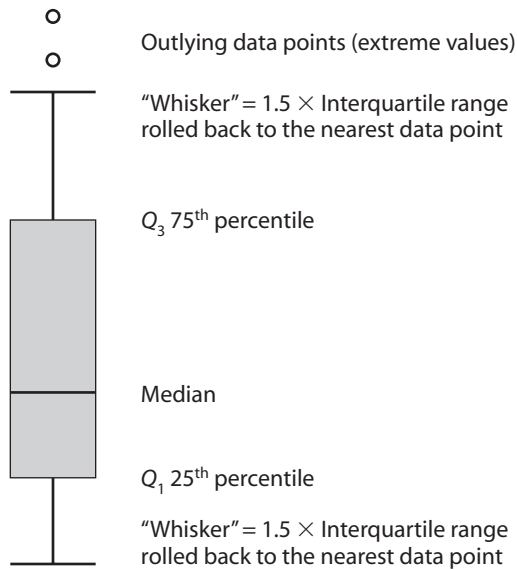
A *box-whisker graph* and a *one-way scatterplot* are two other methods to graphically describe intervally measured data. A box-whisker graph uses a one-dimensional continuous scale. Its general structure is shown in Exhibit 3–15. The central tendency is indicated by the median value of the data, marked by a line. The dispersion of the data is indicated by a rectangle outlining the interquartile

*Exhibit 3–13***Frequency Distribution: Assets of Nonoperating Private Foundations**

Nonoperating foundations	Number of returns
\$1 to under \$100,000	18,900
\$100,000 to under \$1,000,000	26,319
\$1,000,000 to under \$10,000,000	17,869
\$10,000,000 to under \$25,000,000	2,472
\$25,000,000 to under \$50,000,000	1,025
\$50,000,000 to under \$100,000,000	522
\$100,000,000 or more	509
Zero or unreported	2,387
Total	70,004

Source: Internal Revenue Service, Domestic Private Foundations: Number and Selected Financial Data, by Type of Foundation and Size of Fair Market Value of Total Assets, Tax Year 2003
www.irs.gov/taxstats/charitablestats/article/0,,id=96996,00.html

*Exhibit 3–14***Histogram: Assets of Nonoperating Private Foundations**

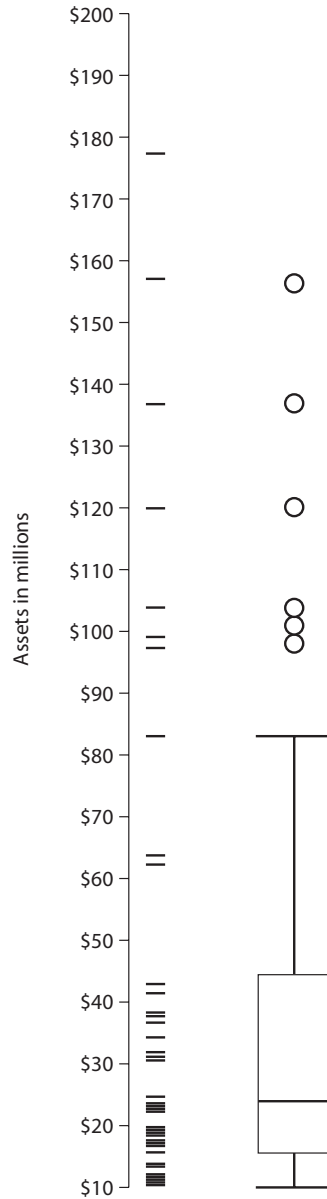
*Exhibit 3–15***Structure of a Box-Whisker Graph**

range. The range of values in the data set is indicated by lines extending from the rectangle to data points that are up to 1.5 times the interquartile range above, and below, the median. Any very extreme outlying data points are marked individually. A one-way scatterplot is drawn by creating a one-dimensional continuous scale and marking each observed value in the data set with a hash mark. The resulting picture provides a general indication of both the central tendency and spread of the data. Exhibit 3–16 displays a one-way scatterplot and a box-whisker graph for nonoperating private foundations reporting more than \$10 million of assets.

The mean of an intervally measured variable should be computed from absolute, ungrouped data. However, sometimes the only data available have already been grouped. We do not have access to the absolute ungrouped values. In effect, the level of measurement has been changed from interval to ordinal. For grouped data, the analyst will have to approximate the mean. Let " m_i " be the midpoint of a class and " f_i " be the frequency of that class—the number of cases that have observed values within that class range. The approximate value for the mean can be found by multiplying each midpoint value by the respective class frequency, finding the sum of these products for all classes in the distribution, and dividing this sum by N , the number of cases. The symbol " \approx " reads "approximately equals."

Exhibit 3–16

One-Way Scatterplot and Box-Whisker Graph: Assets of the Largest Nonoperating Private Foundations (specific data points in this example are hypothetical)



*Exhibit 3–17***Computation of the Mean and Standard Deviation for Grouped Data**

Data	Mean	Standard deviation
Population	$\mu \approx \frac{\sum f_i \cdot m_i}{N}$	$\sigma \approx \sqrt{\frac{\sum f_i \cdot (m_i - \mu)^2}{N}}$
Sample	$\bar{x} \approx \frac{\sum f_i \cdot m_i}{n}$	$s \approx \sqrt{\frac{\sum f_i \cdot (m_i - \bar{x})^2}{n - 1}}$

The standard deviation of grouped data also substitutes the product $f_i \cdot m_i$ for x_i in the defining formulas. Equations for calculating the mean and standard deviation of grouped data are in Exhibit 3–17.

Sometimes the dispersion of a frequency distribution is reported as a ratio—dividing dispersion by the mean or the sample size. This helps the analyst assess the relative precision of estimates that may be derived from a sample.

The **standard error of the mean** is the ratio of a sample standard deviation to the square root of the sample size. Standard error will be addressed again in later chapters. For now it is enough to know that reporting standard error and sample size allows the analyst to calculate the standard deviation.

$$\text{Standard error of the mean} = \frac{s}{\sqrt{n}}$$

The *relative standard error* (also called the *coefficient of variation*) is the ratio of the standard deviation to the mean.

$$\text{Relative standard error (RSE) for population data} = \frac{\sigma}{\mu}$$

$$\text{Relative standard error (RSE) for sample data} = \frac{s}{\bar{x}}$$

The relative standard deviation indicates the degree to which the values of the cases in a frequency distribution deviate from the mean—the higher the ratio, the greater the spread of the distribution.

Homework Exercises

- 3-1** Total scores on the Graduate Management Admission Test (GMAT) can range from 200 to 800. Exhibit HW3–1A presents the percentages of candidates who scored below selected total test scores. In other words, the table provides a cumulative distribution of the relative frequencies.

*Exhibit HW3-1A***Cumulative Distribution of Scores on the Graduate Management Admission Test**

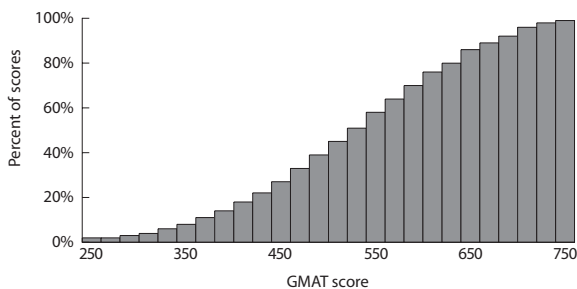
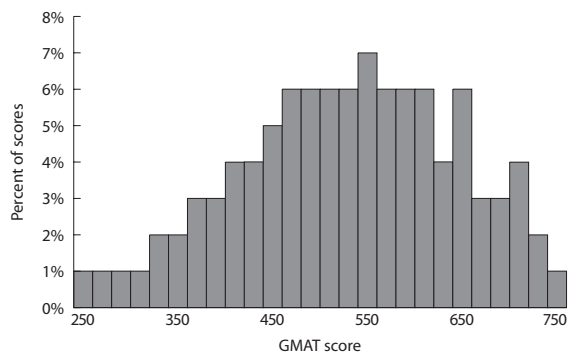
Total score	Percent below
260	2%
280	2%
300	3%
320	4%
340	6%
360	8%
380	11%
400	14%
420	18%
440	22%
460	27%
480	33%
500	39%
520	45%
540	51%
560	58%
580	64%
600	70%
620	76%
640	80%
660	86%
680	89%
700	92%
720	96%
740	98%
760	99%

Source: Graduate Management Admission Council, *Sample GMAT® Score Report*, 2006.
www.mba.com/mba/TaketheGMAT/Tools/SampleScoreReport.htm The mean and standard deviation reported for 622,975 test takers over three years is 526.6 and 117.0, respectively. Calculations in this example differ due to category grouping and rounding of percentages.

- a. Draw a cumulative relative frequency histogram.
- b. Draw a relative frequency histogram.
- c. What is the mean value?
- d. What is the median value?
- e. What is (are) the modal value(s)?
- f. What is the standard deviation?

48 CHAPTER 3 Who, What, When, and How Much? One-Variable Description

- a. Cumulative relative frequency histogram. See Exhibit HW3–1B.
- b. Relative frequency histogram. See Exhibit HW3–1C.
- c. See Exhibit HW3–1D. Mean $\mu \approx 524$.
- d. Median ≈ 540 .
- e. Mode ≈ 550 .
- f. Standard deviation ≈ 114 .

*Exhibit HW3–1B***Cumulative Relative Frequency Histogram***Exhibit HW3–1C***Relative Frequency Histogram**

*Exhibit HW3-1D***Calculations for the Mean and Standard Deviation of the Graduate Management Admission Test Score Distribution**

Score	Percent below	Class interval	Class midpoint	Relative frequency	$f \cdot m$	$f \cdot (m - \mu)^2$
260	2%	< 260	250	0.01	2.5	721.9969
280	2%	260–279	270	0.01	2.7	618.5169
300	3%	280–299	290	0.01	2.9	523.0369
320	4%	300–319	310	0.01	3.1	435.5569
340	6%	320–339	330	0.02	6.6	712.1538
360	8%	340–359	350	0.02	7.0	569.1938
380	11%	360–379	370	0.03	11.1	663.3507
400	14%	380–399	390	0.03	11.7	496.9107
420	18%	400–419	410	0.04	16.4	472.6276
440	22%	420–439	430	0.04	17.2	314.7076
460	27%	440–459	450	0.05	22.5	235.9845
480	33%	460–479	470	0.06	28.2	142.3014
500	39%	480–499	490	0.06	29.4	49.4214
520	45%	500–519	510	0.06	30.6	4.5414
540	51%	520–539	530	0.06	31.8	7.6614
560	58%	540–559	550	0.07	38.5	68.5783
580	64%	560–579	570	0.06	34.2	157.9014
600	70%	580–599	590	0.06	35.4	305.0214
620	76%	600–619	610	0.06	36.6	500.1414
640	80%	620–639	630	0.04	25.2	495.5076
660	86%	640–659	650	0.06	39.0	1034.381
680	89%	660–679	670	0.03	20.1	686.7507
700	92%	680–699	690	0.03	20.7	880.3107
720	96%	700–719	710	0.04	28.4	1463.828
740	98%	720–739	730	0.02	14.6	892.9538
760	99%	740–800	750	0.01	7.5	534.9969
Total				0.99	523.9	12988.33

50 CHAPTER 3 Who, What, When, and How Much? One-Variable Description

- 3-2** A planner wants to draw a histogram of the age structure of the population in Visalia. Group the values of the variable “age” into ten-year intervals. What level of measurement does the variable “age” have before and after grouping?

Let age be symbolized by x :

$0 \geq x < 10$ The first group reads “ x is greater than or equal to 0 and less than 10.”

$0 \geq x < 20$

$0 \geq x < 30$

$0 \geq x < 40$

$0 \geq x < 50$

$0 \geq x < 60$

$0 \geq x < 70$

$x > 70$

Age is a variable with an interval level of measurement. After grouping, the variable has an ordinal level of measurement.

-
- 3-3** Suppose that an analyst is interested in how fast vehicles are traveling on Highway 101. The posted speed limit is 55 mph. What is the matter with the following grouping of the variable “speed?”

under 5

5–17

18–20

21–29

31–39

40–45

46–49

50–59

59–65

This is a real mess.

- Groups of unequal size.
- Discontinuous groups.
- Some possible values of speed excluded.
- Overlapping groups.
- Low range of speeds covered in too much detail while likely range of speeds not covered in enough detail—very high speeds not covered at all.

-
- 3-4** A local public radio station is soliciting contributions from its audience during Pledge Week. The announcer states that the average contribution of its members is

\$100. Do you think that this average is the mean, median, or mode of the contributions? Why?

The TV station, since it is seeking contributions, reports the highest average. In this case it would be the mean. The mean would be larger than the median due to the relatively few but substantial contributions made by wealthy donors. The median, a smaller number, would be more representative of what its viewing (and donating) audience contributes—50% of the contributors giving higher and 50% giving lower amounts than the median value reported. The mode is the most frequently occurring contribution amount—for example \$35.

- 3-5** Exhibit HW3-5A shows expenditures for special education as a percentage of total school operating budget for 327 Massachusetts school districts for one fiscal year.¹
- What is the average special education expenditure (as a percentage of school operating expenditures)?
 - What is the standard deviation?
 - Draw a histogram of special education expenditures.

Exhibit HW3-5A

Special Education Expenditures as a Percentage of Total Operating Budget: Massachusetts School Districts FY2001

Special education expenditure percentage	Number of school districts
< 5%	11
5 to 10%	22
10 to 15%	56
15 to 20%	170
20 to 25%	60
25% or more	8
Total	327

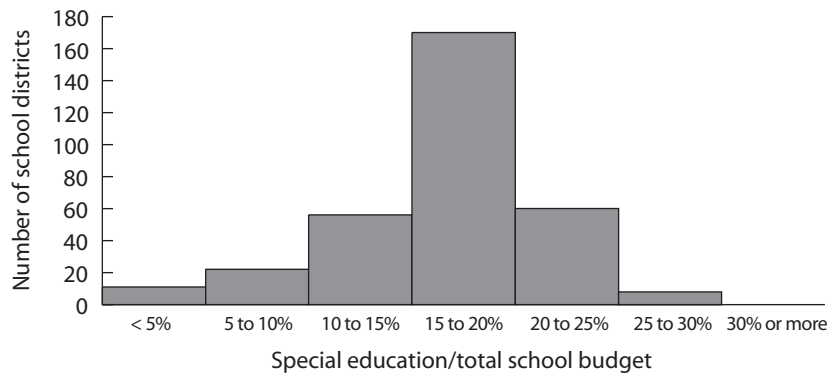
¹Special Education Expenditures as a Percentage of Total School Budget, Preliminary FY01, Massachusetts Department of Education, finance1.doe.mass.edu/education/spedexp01.html

52 CHAPTER 3 Who, What, When, and How Much? One-Variable Description

- a. Mean \approx 16.6%.
- b. Standard deviation \approx 5.0%.
- c. See Exhibit HW3–5B.

*Exhibit HW3–5B***Histogram: Special Education Expenditures**

Special Education Expenditures as Percentage of Total Operating Budget:
Massachusetts School Districts FY2001



3-6 “Congregate housing for the elderly” is defined by the government as age-segregated housing built specifically for the elderly (62 years and over) which provides, at the very least, an on-site meal program. What are the characteristics of congregate facilities beyond provision of a meal program? To answer this question, 27 congregate facilities were asked to list their services (see Exhibit HW3–6A).

- a. In what ways is this table different from other frequency tables?
- b. How many variables are presented in the table? What are the values of each variable?
- c. Can a histogram be drawn?
- d. Write a brief statement, based on these data, that summarizes the characteristics of congregate facilities for the elderly.

*Exhibit HW3–6A***Service Availability on 27 Sample Sites of Congregate Housing for the Elderly**

Service	Number of sites with service	Percentage of sites with service
Meal service	27	100%
Recreational	27	100
Social	24	89
Educational	24	89
Security	24	89
Transportation	23	85
Commercial	22	81
Medical	21	78
Housekeeping	18	67
Linen	16	59
Protective	16	59

- a.** This is a bit tricky. At first glance it appears to be a conventional representation of a frequency distribution for one variable, “service,” with 11 values, “meal service, . . . , protective.” If each service is examined one at a time, at each of the 27 facilities, and asked the question, “What kind of service is this?”, the analyst would indeed get the frequencies shown—27 of the services provide meals, 16 provide linen. Each instance of providing a service would be a separate case (“event” or “unit of analysis”). The variable would have a nominal level of measurement with 11 possible values. But this is not what was done. Instead, each facility was asked, “Which of the following eleven services do you provide?” Each facility is a separate case, and multiple responses are allowed to the inquiry.
- b.** This means that each type of service is a separate variable (there are 11 variables) and the possible values for each variable are “yes” or “no.” The relative frequency of the affirmative responses are listed in the table.
- c.** No. The variables have a nominal level of measurement. Strictly speaking, a bar chart, illustrating the frequency distribution, could be drawn for each variable—but it would be redundant (yes and no categories) and not useful in addressing the basic question in the problem—part d. A bar chart could be drawn showing the relative frequency of the 11 different services among the 27 sites.
- d.** “The sample congregate housing sites provide activity and entertainment oriented services more frequently than services to the residents’ physical needs.”

54 CHAPTER 3 Who, What, When, and How Much? One-Variable Description

- 3-7** Interview 10 people who are registered to vote. Ask each person about his or her political affiliation. Record the responses as 1. Democratic; 2. Republican; 3. Independent; 4. Other; 5. No Response or Decline to State
- What is the variable in this problem?
 - What is its level of measurement?
 - How many possible values can the variable take on? What are they?
 - Can a modal value be computed? A median value? A mean value?
 - Set up a table for properly recording these data.
-
- The one variable in this problem is “political affiliation.”
 - It is a nominal level of measurement.
 - For each observation the variable can take on any one of five possible values: 1 meaning Democratic; 2 meaning Republican; 3 meaning Independent; 4 meaning Other; 5 meaning Decline to State or No Response.
 - Because the variable has a nominal level of measurement a modal value—the most frequently occurring value—can be found. Median and mean values cannot be found.
 - See Exhibit HW3–7A.

*Exhibit HW3–7A***Hypothetical Data File**

Observation number	Political affiliation 1 = Democratic 2 = Republican 3 = Independent 4 = Other 5 = No Response or Decline to State
1	1
2	3
3	2
4	4
5	1
6	2
7	2
8	5
9	1
10	2

- 3-8** The frequency table shown in Exhibit HW3–8A is part of a study done by an emergency medical services planning agency in order to “delineate the magnitude and character of critical trauma care.”
- What is the population being studied?
 - What variable is being observed?
 - How many observations are there?
 - Draw a histogram of the frequency distribution. (You must exclude cases for which the observation of age is missing.)
 - Another way to display the frequency distribution of a variable inherently having an interval level of measurement is to draw a frequency polygon. The figure is drawn by connecting the points describing the midpoint and frequency of each class with straight lines. Each straight line cuts off a triangular section of one of the histogram bars and seems to add on another triangular area of equal size. Thus the area under the frequency polygon (probability) equals the area of the histogram bars. The tails of the frequency polygon are tied off (brought down to zero) by following the rule of equal areas (a constant interval between midpoint values). Superimpose a frequency polygon on the histogram drawn in part d of this question.

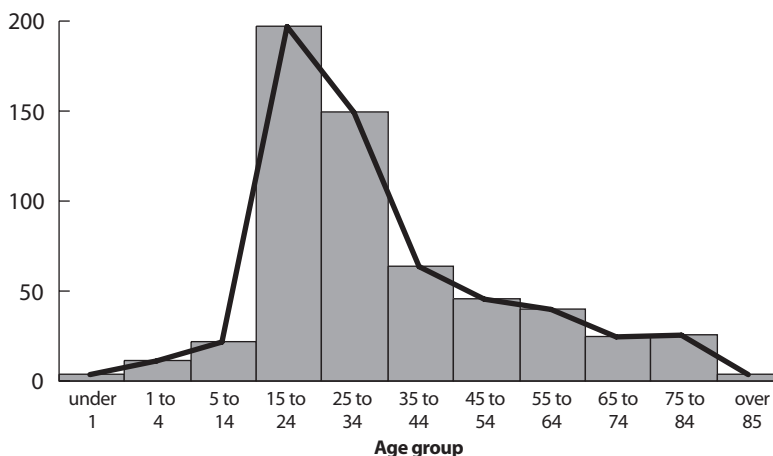
Exhibit HW3–8A

Number of Trauma Cases by Age in Emergency Services Area 6

Age group	Number of cases	EMS-6 population percentages
Under 1	4	2.1%
1–4	11	7.8%
5–14	22	16.3%
15–24	197	16.4%
25–34	150	15.9%
35–44	64	10.9%
45–54	46	9.2%
55–64	40	9.5%
65–74	25	7.1%
75–84	26	3.6%
Over 85	4	1.0%
Unknown	38	
Total	627	99.8%

56 CHAPTER 3 Who, What, When, and How Much? One-Variable Description

- f. Find the modal, median, and mean values of the frequency distribution.
- g. Find the relative frequency of each group. Compare these to the population relative frequencies shown in the table. Write a brief statement summarizing this comparison. What additional information is needed in order to determine whether this comparison is typical of the rest of the United States or not?
- h. Find the standard deviation of the frequency distribution.
-
- a. All trauma cases in emergency services area 6.
- b. Age.
- c. 627.
- d,e. See Exhibit HW3–8B.
- f. Modal age group: 15 to 24; median age group: 25 to 34; mean age \approx 34.2.
- g. See Exhibit 3–8C. The distribution of trauma victims by age does not match the age distribution of the general population. The incidence of trauma among people less than 15 years old is less than half of its relative frequency in the general population. The incidence of trauma among people from 15 to 35 years old is about double its relative proportion in the general population. Further study would require a similar frequency distribution for trauma cases by age group for the total United States and the age distribution of the U.S. population.
- h. Standard deviation \approx 18.9.

*Exhibit HW3–8B***Frequency Polygon: Trauma Cases by Age**

*Exhibit HW3–8C***Number of Trauma Cases by Age: Emergency Services Area 6**

Age group	Absolute frequency	Relative frequency
		(<i>n</i> = 589)
Under 1	4	0.7%
1–4	11	1.9%
5–14	22	3.7%
15–24	197	33.4%
25–34	150	25.5%
35–44	64	10.9%
45–54	46	7.8%
55–64	40	6.8%
65–74	25	4.2%
75–84	26	4.4%
over 85	4	0.7%
Unknown	38	—
Total	627	100.0%

Source: Trauma Study Part I: July 1, 1980–June 30, 1981, Alpine, Mother Lode, San Joaquin Emergency Medical Services Agency, Modesto, California, May 1983.

- 3-9** Analysts surveyed cities to obtain information about personnel practices in their police departments. They received responses from 1,250 cities as summarized in Exhibit HW3–9A. What was the average population of cities in the survey?

*Exhibit HW3–9A***Responses to Survey Regarding Police Personnel**

Population group	Number of cities reporting
Over 1,000,000	4
500,000–1,000,000	13
250,000–499,999	27
100,000–249,999	78
50,000–99,999	171
25,000–49,999	310
10,000–24,999	647
Total	1250

58 CHAPTER 3 Who, What, When, and How Much? One-Variable Description

3-10 Suppose an analyst is provided the data in Exhibit HW3-10A for a large school district.

- a. How many teachers does the school district employ?
- b. How many schools are there in the school district?
- c. How many variables are presented in the table?
- d. What are the values of the variables?
- e. What is the unit of analysis?
- f. How many observations are in the table?

Using the modal value of the distribution in Exhibit HW3-10A, one colleague guesses that there are about 50 teachers per school. It is anticipated that a salary negotiation settlement will bring a teacher's salary "on the average" up to \$50,000 per year. Therefore, your colleague projects that the school district budget for teachers salaries will increase to \$100 million.

- g. Develop another budget estimate using the above data.
- h. What is the difference between the budget estimates?

Exhibit HW3-10A

School District Data

Number of teachers	Number of schools
0-19	3
20-39	12
40-59	20
60-79	5

- a. Do not know from the data given.
- b. 40 schools.
- c. One variable—number of teachers.

- d. Four groups: 0–19, 20–39, 40–59, 60–79.
- e. Each school is one unit of analysis.
- f. Forty.
- g. The first budget estimate was derived by multiplying:

$$(40)(50)(50,000) = \$100 \text{ million}$$

The estimated mean number of teachers per school site is 43.5 (using the formula for estimating the mean of grouped data). Thus a revised budget estimate is

$$(40)(43.5)(50,000) = \$87 \text{ million}$$

- h. The difference in the budget estimates is \$13 million—no small change even for a school district of this size.

-
- 3-11** Explain the differences in structure and content between a *data file* and a *frequency table*. Provide an example.

A data file lists each observation recorded for a set of related events, case by case. Each line (observation number) represents a separate and unique case (event). The body of the table contains the measured values for each case with respect to the variables of interest. A frequency table lists the possible values that could be measured for a particular variable and the corresponding number of times (number of cases) each value appeared in the data file for that one variable.

-
- 3-12** Key words used to describe public policies are important in the development of public opinion and perception. Connecticut officials were interested in developing a baseline of public awareness before launching a new “clean energy” campaign. They were particularly interested in the different connotations of “clean” and “renewable.” One group was asked, Have you ever heard of “renewable energy”? Another group was asked, Have you ever heard of “clean energy”? Respondents who replied yes were then asked to explain what the term meant to them or to provide examples. Exhibit HW3–12A provides a few responses to the survey. These data are summarized in Exhibit HW3–12B.

Exhibit HW3–12A

Multiple responses to a survey question

Have heard of “renewable energy”		Have heard of “clean energy”	
Respondent	Response	Respondent	Response
1	1, 3, 4	1	2, 4
2	2, 5	2	2, 5, 7
3	9	3	1, 6, 9
4	11, 14	4	2, 14
5	1, 3, 6	5	13, 16
6	2, 4	6	17
7	1, 9		
8	8, 12, 16		

Multiple responses allowed

- | | |
|-------------------------------------|------------------------------------|
| 1 Reusable/recyclable/won't deplete | 10 Energy efficiency |
| 2 Less polluting/cleaner air | 11 From natural resources |
| 3 Solar energy/photovoltaics | 12 Biomass; burning organic matter |
| 4 Wind power/windmills/wind farms | 13 Costs more/less |
| 5 Better for environment | 14 Hybrid/electric cars |
| 6 Less fossil fuel/oil | 15 Hydrogen |
| 7 Renewable/clean energy | 16 Other |
| 8 Water power/hydroelectricity | 17 Don't know |
| 9 Recycling | |

Source: http://ctinnovations.com/communities/files/CCEF_survey_report_May_18_2005_Final.pdf

Exhibit HW3–12B

Frequency Distribution: Survey Responses

Definition/Example	Renewable energy	Clean energy
1 Reusable/recyclable/won't deplete	3	1
2 Less polluting/cleaner air	2	3
3 Solar energy/photovoltaics	2	0
4 Wind power/windmills/wind farms	2	1
5 Better for environment	1	1
6 Less fossil fuel/oil	1	1
7 Renewable/clean energy	0	1
8 Water power/hydroelectricity	1	0
9 Recycling	2	1
10 Energy efficiency	0	0
11 From natural resources	1	0
12 Biomass; burning organic matter	1	0
13 Costs more/less	0	1
14 Hybrid/electric cars	1	1
15 Hydrogen	0	0
16 Other	1	1
17 Don't know	0	1

- 3-13** Grade point average is one output assessment educational indicator. Some data for students in an MPA program are shown in Exhibit HW3–13A. Compute the appropriate measures of central tendency and dispersion for this frequency distribution.

Exhibit HW3–13A

MPA Grade Point Averages

Grade point average	Number of students
3.5000–4.0000	502
3.0000–3.4999	227
2.5000–2.9999	28
2.0000–2.4999	11
1.5000–1.9999	1
1.0000–1.4999	2
Total	771

Mean \approx 3.54, standard deviation \approx 0.34.

- 3-14** Consider the following statement from a research report:

Brush species make up much of the fuel load in forested wildlands. Basic physical and chemical characteristics of these species influence ease of ignition, rate of fire spread, burning time, and fire intensity. Quantitative knowledge in the variations in brush characteristics is essential to progress in fire control and effective use of fire in wildland management.²

Five shrub species common to northern California brush fuels were studied. One of the variables measured was ash content. “The amount of ash or minerals in vegetation can affect how well the material burns. . . . Another study showed that the moisture content of living vegetation with high ash content tends to be higher than that with low ash content.” The report included the following statement: “Ash content of the woody material [as opposed to plant foliage] was low and did not vary consistently among species or size class of material. Average ash content of the live woody material was 1.6 percent with a coefficient of variation of 25.2 percent. Ash content of the dead material was only slightly lower, averaging 1.4 percent. Variability of the dead fuel ash content was greater, however, with a coefficient of variation of 37.1 percent. The greater variation probably resulted

²Countryman, Clive M., Physical Characteristics of Some Northern California Brush Fuels, Gen. Tech. Rept. PSW-61, Berkeley, CA: Pacific Southwest Forest and Range Experiment Station, Forest Service, U.S. Department of Agriculture, 1982.

62 CHAPTER 3 Who, What, When, and How Much? One-Variable Description

from differences in the amount of weathering among individual samples [cases] rather than from variation among species or size of material.” What are the standard deviations of the ash content of the live and dead material that the researcher measured?

The standard deviation for the ash content of the live material is 0.40 percent. The standard deviation for the ash content of the dead material is 0.52 percent.

- 3-15** Each civil service employee of the U.S. Government is classified according to grade on a general schedule of employment. The grades indicate increasing levels of expertise, responsibility, or authority—and commensurately higher salaries. A frequency table showing the number of employees by grade in 2004 is shown in Exhibit HW3–15A.
- Summarize these data with a bar chart.
 - What is the average GS grade?

Exhibit HW3–15A

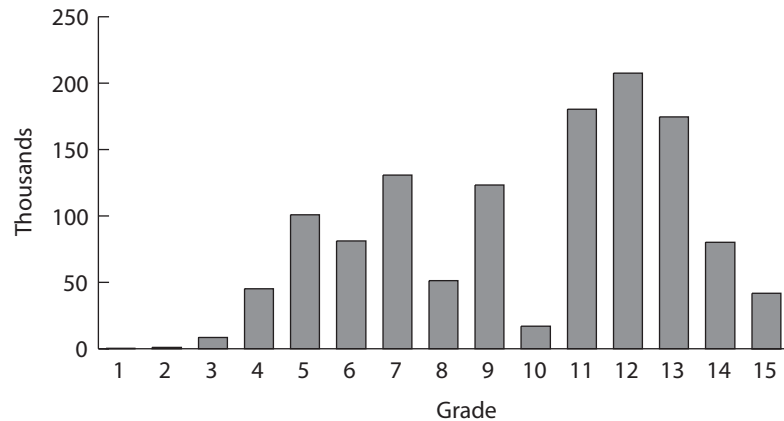
Full-Time Civilian General Schedule Employment by Grade (2004)

GS grade	Number of employees
1	270
2	955
3	8,445
4	45,327
5	100,984
6	81,255
7	130,828
8	51,413
9	123,437
10	16,975
11	180,333
12	207,566
13	174,575
14	80,205
15	41,845
Total	1,244,413

- c. To what extent do civil service employees have GS classifications different from the average value?
-
- a. See Exhibit HW3–15B Bar Chart: Federal Employment by GS Grade.
- b. The median GS classification in 2004 was 11 (in 1991 it was 9).
- c. The interquartile range is 7 to 12.

Exhibit HW3–15B

Bar Chart: Federal Employment by GS Grade (2004)



- 3-16** How accessible are hospital-provided abortions? To answer this question, activists in one state conducted a survey of every general hospital. Volunteers called each facility. The caller first asked if the hospital had a gynecology department. If it did, the caller asked that the call be transferred to it. The caller stated that she was pregnant and trying to get an abortion, and then asked whether the hospital would perform the procedure. Of the 362 hospitals contacted, 261 (72.1%) replied that they did not provide abortions. Eighteen (5%) did provide abortions. Seventy-two hospitals (19.9%) had restricted abortion policies such as requiring a woman to have a prior relationship with a doctor who will do the procedure and who has surgery rights at the hospital. Eleven hospitals (3%) did not provide clear answers about their abortion accessibility. Present these results in an appropriate table and graph.

64 CHAPTER 3 Who, What, When, and How Much? One-Variable Description*Exhibit HW3–16A***Abortion Accessibility in Hospitals**

Hospital policy	Number	Relative frequency
Do not provide abortions	261	72.1%
Restricted provision of abortions	72	19.9%
Provide abortions	18	5.0%
Unknown	11	3.0%
Total	362	100.0%

*Exhibit HW3–16B***Bar Chart: Abortion Accessibility in Hospitals**