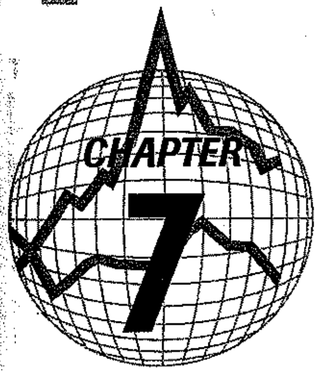
measures of a variable, whereas *attributes* refers to the range of values that a variable can have. For example, the variable "gender" has three attributes (male, female, and unknown), even if none of the respondents state that they are male. A variable has as many values as observations, and as many attributes as the different values that a variable *can* have. However, when these terms are used synonymously, confusion can result.

2. On surveys, missing data may indicate a problem with how the question was phrased, indicating a problem with survey validity.
3. Some analysts feel that the mean should not be used with ordinal-level variables, but it does provide useful information. However, the mean is especially inappropriate for nominal-level variables. For example, we cannot say that the average region is 2.45, on a scale of 1 = Northeast, 2 = South, 3 = Midwest, and 4 = West. When working with nominal variables, we should describe response frequencies, such as "23.7 percent of employees live in the Northeast, 19.8 percent live in the West," and so on.
4. We can also express the formula in the text as (2.3*0.67 + 4.5*0.33) = 3.0.
5. The mode might be useful in the following instance: Consider the number of violent incidents in which all high school students are involved. The mode may well be 0, and the mean 0.2, for example. If only those who have experienced an incident are included, then the mode might be 2 and the mean 4.2, for example. In this case, the mode is used as a precursor to a better understanding of the distribution of the data.
6. Often, we report values to two decimal places, by default. However, on occasion we may wish to report values to more or fewer decimal places. Here, we report the result to three decimal places to avoid rounding errors that would become evident in the next paragraph.



Measures of Dispersion

CHAPTER OBJECTIVES

After reading this chapter, you should be able to

- Make frequency distributions and report results
- Distinguish between a histogram and a stem-and-leaf plot
- Create line charts, pie charts and other visual aids
- Use boxplots for data cleaning
- Identify and address problems of outliers
- Understand the normal curve and standard deviation

This chapter examines ***measures of dispersion,*** which provide information about how the values of a variable are distributed. These measures are the second family of univariate statistics (the first family is *measures of central tendency,* as discussed in Chapter 6). A common measure of dispersion in public and nonprofit management is the frequency distribution. Knowing what percentage of clients or employees score above a specific value is often a prelude to decision making. Such frequency distributions can then be used to make comparisons or create rankings. Frequency distributions often are reported in tabular form, but they are also the basis of many graphs—such as bar charts, pie charts, and line graphs used in presentations.

Measures of dispersion are also used in preliminary data analysis, such as for data cleaning and for generating a first understanding of one's data. The boxplot is a useful tool for data cleaning and is used in this chapter to discuss outliers (that is, unusually large or small values). Analysts need to know how to identify and deal with outliers. An interesting paradox is that knowing the mean or median often invites further questions about the distribution of variables (calling for the use of measures of dispersion), but preliminary analysis of variables' distribution often precedes the analysis of central tendency to ensure that data are clean and appropriate. As a result, analysts commonly first use boxplots for data cleaning, then analyze means and medians, and then examine frequency distributions. This is an important sequence to remember.

This chapter also provides an introduction to the normal distribution, which is relevant for continuous variables. Many continuous variables, such as height and IQ, are normally distributed. Such variables have certain characteristics and terminology with which managers and analysts need to be familiar. In addition, many of the statistical tests discussed in later chapters assume that continuous variables are normally distributed. Those chapters also describe tests for examining this assumption and strategies for dealing with situations when this assumption is not met.

FREQUENCY DISTRIBUTIONS

Frequency distributions describe the range and frequency of a variable's values. They are easy to create and understand, and they often are a prelude to generating data tables and attractive graphics. First we discuss frequency distributions for categorical (nominal or ordinal) variables, and then we look at frequency distributions for continuous variables.

The categories of ordinal or nominal variables often are used in frequency distributions. For example, if a variable is measured on a five-point Likert scale, then we can count, or determine the frequency of, the data elements for each category. Table 7.1 shows a common type of typical frequency distribution; it gives the number of respondents and the frequency of responses (as a percentage) for each item in an employee survey.[1] For example, 81.3 percent [22.8% + 58.5%] of the 969 respondents who answered this question agree or strongly agree that they are satisfied with their job. Likewise, percentages can be determined for other items, and comparisons made across items, too. For example, more respondents agree or strongly agree that they are satisfied with their jobs than agree or strongly agree that each individual is treated with dignity (81.3% versus 42.3%). Frequency distributions are often found in the main bodies of reports and sometimes are used in the statistical appendices of reports, showing the distributions of all survey items.

Table 7.1 Frequency Distribution: Employee Survey

Statement	Mean[a]	Strongly Agree (%)	Agree (%)	Don't Know (%)	Disagree (%)	Strongly Disagree (%)	n
I am satisfied with my job at Seminole County	3.88	22.8	58.5	5.5	10.7	2.5	969
Seminole County is a good place to work compared with other organizations	3.64	15.4	53.6	14.1	13.1	3.8	969
Each individual is treated with dignity	3.02	8.1	34.2	18.6	29.6	9.5	967

[a] 5 = Strongly agree; 4 = Agree; 3 = Don't know; 2 = Disagree; 1 = Strongly disagree.

Table 7.1 also shows the number of observations for each item (n). Because of incidental missing responses, this number may be different for each item. If all items have the same number of observations, then this number can be reported just once, such as at the top or bottom of a table. Table 7.1 also shows the mean of each response. Frequency distribution tables commonly include mean item responses, and this additional information is sometimes used for ordering (ranking) the items, based on their mean responses. Of course, items can also be ordered on the basis of those who agree or strongly agree with an item. Either approach results in the same ordering of the three items in Table 2.2.[2]

Frequency distributions are readily calculated by statistical software programs. However, when variables are continuous, analysts will have to construct their own categories, because otherwise the frequency distribution will likely have a large number of frequency table categories with just one or a few observations. A practical question is how wide each category should be. Although no hard-and-fast rules exist, analysts should avoid recoding in ways that mislead. To avoid perceptions of lying with statistics, a rule of thumb is that categories should be based on *category ranges of equal length,* unless compelling reasons exist that are clearly explained in the text of the report.[3]

Many computer programs produce ***stem-and-leaf plots,*** which assist in category construction.[4] These plots seldom appear in final reports; rather they are tools that aid in analysis, and analysts should know how to read

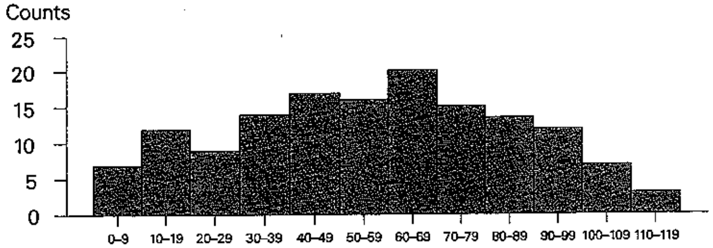
Figure 7.1 Stem-and-Leaf Plot

00	23589
01	0012344589
02	1123367
03	01134456889
04	00111223445567889
05	011233455568899
06	1222344445566778999
07	001235556899
08	01112234444
09	112334556
10	00233
11	035

them. Stem-and-leaf plots show the frequency distribution of continuous variables. The categories are computer generated, based on the values of observations. The stem-and-leaf plot shown in Figure 7.1 was generated for a continuous variable that has 124 observations. In a stem-and-leaf plot, these observations are shown ordered from low to high. The left column is called the stem, and the right column is called the leaves. The values are found by combining each stem with its leaves: the stem is composed of the first figures of any number, and the leaves are the figures added onto it. In this case, the lowest number of the series is 002 followed by 003, 005, 008, 009, 010, 010, 011, 012, and 013, and so on. The five highest values are 103, 103, 110, 113, and 115. The median value is at location [(124 + 1)/2 =] 62.50, and is 058 (the sixty-second and sixty-third values, starting from lowest value 002 and counting forward, are both 58). Based on Figure 7.1, an analyst might decide to construct following four categories, of equal width: 0–29, 30–59, 60–89, and 90–120. Or perhaps more categories might be created, such as in widths of 10: 0–9, 10–19, 20–29, 30–39, and so on. Once an analyst decides which category widths he or she wants to use for recoding, statistical software programs contain procedures for recoding the data values.

Frequencies can also be shown in a ***histogram***, as shown in Figure 7.2. A histogram is similar to a stem-and-leaf plot but differs in that it shows the number of observations in each category. Unlike stem-and-leaf plots, histograms are commonly seen in reports. A histogram is useful because it provides a quick visual representation of the extent and nature of dispersion. Unlike the stem-and-leaf plot, a histogram allows the analyst to see how many observations are present in each category, although it does not show the value of each observation; many analysts use histograms rather than stem-and-leaf plots. Software programs often automatically generate category

Figure 7.2 Histogram



widths (shown in Figure 7.2 in increments of 10, the default used by SPSS for these data), but software users can also define these widths.[5]

Frequency distributions are a staple of analysis. Managers and analysts need to be able to describe them in plain terms that are readily understood by a broad audience. It's not a bad idea to practice making them until you are comfortable doing so. This matter is discussed in Box 7.1.

Graphical displays often aid in highlighting key results. Tables of frequency distributions, like the one shown in Table 7.1, can be used as the basis of displays that highlight important conclusions. Statistical software programs readily generate these graphs, which can be copied into other programs for further editing and word processing. Managers and analysts need to know how to create attractive graphical displays for their reports and presentations. ***Bar charts*** are graphs that show the frequency of occurrences through stacks (Figure 7.3); they are used with categorical data. Bar chart A shows options for three-dimensional effects and shading. Bar chart B is also called a Pareto chart, an ordering of categories according to their frequency or importance. This is a visually useful way of drawing attention to that which is important, as well as to the unimportance of other categories that, cumulatively, add very little to our understanding of the problem. Sources 5 through 7 seem barely worth debating. As a convention, bar charts are used with categorical variables, and the bars in such charts should not touch each other. Histograms are used with interval- and ratio-level variables, and their bars should touch each other, suggesting that these data are continuous (see Figure 7.2).

Pie charts typically are used to focus on equality: who gets the most (or the least) of what? Pie charts are used with categorical (often nominal) data, and they can be shown in various ways; Figure 7.3 shows a pie chart with a slice that has been pulled out. ***Line charts*** are used with continuous data, *In Greater Depth...*

Box 7.1 Writing It Up . . .

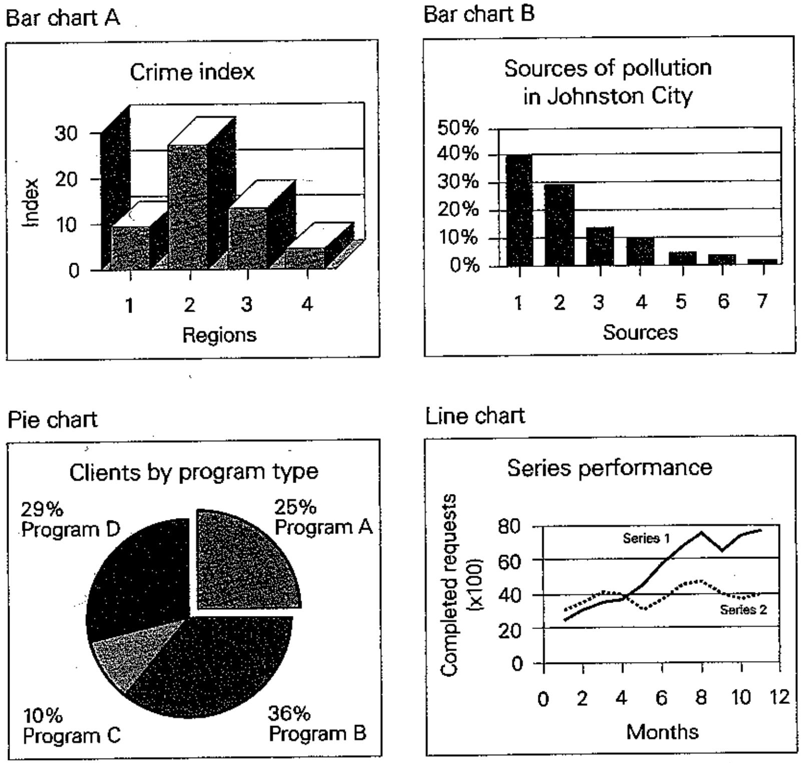
Effective communication of statistical findings is very important. People expect results to be written in a professional manner; otherwise, statistical analysis may well fall on deaf ears. New analysts often want and need to know how to write up their results.

In one sense, writing up statistical results is no different from any other writing. First, you need to know what you want to say. Second, you need to know what you need to say about what you want to say. Finally, you need to know how to say it. Regarding the first point, what you want to say often is either the result of an analysis or an argument that you wish to back up with statistical analysis. For example, you might want to inform an argument about education policy with an analysis of educational attainment. With a histogram or frequency distribution, you probably want to say something about the percentage of observations in one or more categories.

Second, people need to know that your conclusions are based on valid data and analysis. Thus, you will need to tell them what data you used, how they were collected, what they can be used for, and any known shortcomings. You also need to identify statistical analytical methods and show through your analysis that you are conversant with any issues or limitations that are pertinent to specific methods. Sometimes variables can be constructed or analyzed in different ways, and you might want to present a range of possible answers. In the case of a histogram or frequency distribution, in addition to matters of data validity, you might note if a large number of values were missing or if different ways of constructing categories might have led to different interpretations.

Third, you need to know how to say things. Generally you should state the facts as they are. It suffices to say, for example, that "[illegible] percent of respondents disagree or strongly disagree with . . ." Leave interpretations to the conclusion and discussion sections, where they should be clearly noted as such. As suggested in Box 1.2, it is useful to find several examples of write-ups that you like and then try to adapt the specific language to your needs. Then, double-check the language to ensure that it is professional (using third person, present tense, and active voice) and that the report looks professional, too.

Figure 7.3 Graphical Displays



partly to avoid displaying a very large number of bars. In Figure 7.3 the lines show averaged occurrences each month. In this figure, two variables are shown in a way that highlights important trend differences.

Getting Started [illegible]

Visual representation is important for analysis and in the communication of study findings to a broader audience. Graphs and charts help draw out differences and inform analytical decisions about differences that matter. For audiences, graphs and charts succinctly summarize and communicate study conclusions, and demonstrate visually why they matter. Analysts are increasingly expected to use visual representations in their reports and oral presentations; they are typically an important and essential part of PowerPoint presentations.

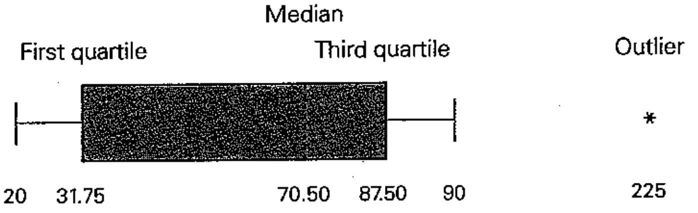
BOXPLOTS

A ***boxplot*** is a graphical device that shows various measures of dispersion. Boxplots are useful for obtaining a quick, visual, preliminary understanding of data. They assist analysts with their data cleaning and are seldom found in final reports. The boxplot is a quick and nicely visual tool, and computer programs readily provide boxplots for any number of variables. Though analysts also use other tools for their data cleaning, the boxplot offers much useful information, such as identifying outliers, which, for example, histograms or frequency distributions do not. Boxplots are appropriate for both continuous- and ordinal-level variables. The following calculations are shown only for purposes of conceptual understanding, given that software programs typically calculate the following values. We use the same array as in Chapter 6, namely, 20, 20, 67, 70, 71, 80, 90, and 225.

Boxplots show statistics that are calculated on the basis of the location of data, such as the median. Figure 7.4 shows the boxplot for our example array. Because our example has eight observations, the median is defined as the value at location [$(n + 1)/2$ =] 4.50 (see Chapter 6), that is, the mean of the values of the fourth and fifth observations, when all observations are ordered from the lowest to the highest values. This value is 70.50. The ***first quartile*** is simply the lowest quartile score (*it is not a range*). That location is defined as half the location of the median, hence, [4.50/2 =] 2.25. The variable value associated with this location (2.25) is defined as the value of the second observation plus one-quarter of the distance between the second and third observations. In our example, that value is calculated as [20 + 0.25*(67 – 20) =] 31.75. The ***third quartile*** is the third quartile score, or location [4.50 + 2.25 =] 6.75. The value is [80 + 0.75*(90 – 80) =] 87.50. Most computer programs also produce a statistic called the ***midspread*** (or ***interquartile range***, IQR). The midspread is defined as the difference between the first quartile and the third quartile, hence, [87.50 – 31.75 =] 55.75. The ***range*** is simply the difference between the highest and lowest values, or [225 – 20 =] 205. Again, even though statistical software programs will calculate these values, you need to have a clear understanding of what these concepts mean.

The boxplot also shows a singular observation with a value of 225 that is labeled "outlier." As mentioned at the beginning of this chapter, ***outliers*** are extremes, or analyst-defined observations with unusual values relative to other values in the data. Outliers may be the result of data-coding errors (which should be either fixed or removed), or they may reflect actual but unusual values in the sample. Such outliers matter because many public decisions are based on average behavior, rather than the unusual behavior of a few.[6] Thus, it makes good sense to distinguish *usual observations from unusual ones.* An important task of data cleaning and preliminary analysis is to identify outliers and to decide whether they should be retained. Our position is that observations flagged as outliers generally should be retained when they are not coding errors, when they are plausible values of the variable in question, and when they do not greatly affect the value of the mean (of continuous variables). However, when outliers are present, their effect on final results should be studied. Analysts should also report any observations (outliers) that have been dropped from the analysis, along with reasons for doing so.

Figure 7.4 Boxplot



Boxplots also help analysts to calculate cut-off points beyond which any observations are statistically considered as outliers. These cut-off points are called, respectively, the inner and outer fences. The ***inner fence*** is an imaginary value that lies 1.5 times the midspread *below* the first quartile. For the data in our example, the inner fence is [31.75 – (1.5*55.75) =] –51.88. All of our data are greater than the value of the inner fence; thus, our data show no outliers on this lower end. The ***outer fence*** is an imaginary value that lies 1.5 times the midspread *above* the third quartile. It is calculated as [87.5 + (1.5*55.75) =] 171.13. Our data has one observation that exceeds this value, x_8 = 225, which is therefore labeled an outlier.[7] Analysts might consider omitting x_8 from further analysis. Doing so greatly affects the mean, reducing it from 80.38 to 59.71. As expected, omitting x_8 does not much change the median, which goes from 70.5 to 70.0; the loss of the observation merely redefines the location of the middle observation. As discussed earlier, the decision to drop an observation from analysis should be based on argument. If the observation is deemed to be representative of the population from which it is drawn, then it should be retained. If it is thought to be unrepresentative, then it should be excluded. Whatever the decision (retention or deletion), the case must be argued and its impact on further analysis noted.

The boxplot further shows two ***whiskers*** extending out from the first and third quartiles. The end points of these whiskers are *the lowest and highest values of the variables that are not outliers.* These values differ from the lowest and highest values of the data only when there are outliers. Together with the box (shown as the "area" between the first and third quartiles), these whiskers give analysts a quick visual image of the spread of the data. If the whiskers and box are relatively short, then the variable varies little in the sample. If the whiskers and box are long, there is more variation to talk about.

In short, boxplots are a great tool for preliminary data analysis and are easily produced by computer. They help identify outliers and provide valuable information about the distribution. Imagine the analyst who rushes headlong into producing detailed tables and reports, only to redo this work after outliers (and perhaps other problems, too) have been discovered in the data!

STANDARD DEVIATION

When variables are continuous, the question "How widely are the data dispersed around the mean?" is especially salient because continuous variables often have a large number of values that can be narrowly or (very) widely dispersed. By contrast, many ordinal-level variables have only a few different data values (for example, a five-point Likert scale) and thus may have a limited range. The ***normal distribution*** refers to the distribution of a variable that resembles a bell-shaped curve (Figure 7.5). The left and right sides of the curve mirror each other; they are symmetrical. Many variables are normally distributed, such as student scores, IQ scores, average crop yields, or counts of lightning flashes over time. However, in practice, almost no data are *exactly* normally distributed. Many analysts rely on *visual inspection* to determine distribution, supplemented with the statistics described in this section and in subsequent chapters. The sample data are not expected to match a theoretical bell-shaped curve perfectly because, given that they represent a sample, deviations due to chance selection should be expected.[8]

Figure 7.5 — Normal Distribution: Bell-Shaped Curve

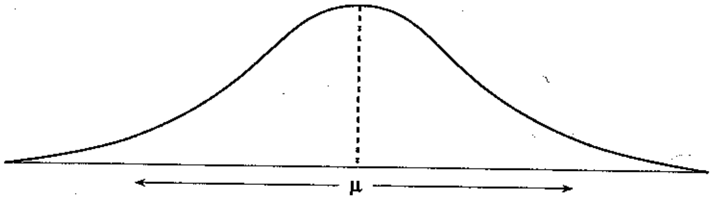
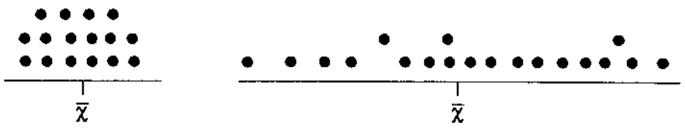


Figure 7.6 — Small and Large Standard Deviation



Small standard deviation

Large standard deviation

The ***standard deviation*** is a measure of dispersion that is calculated based on the values of the data. The standard deviation has the desirable property that, when the data are normally distributed, 68.3 percent of the observations lie within ±1 standard deviation from the mean, and 95.4 percent lie ±2 standard deviations from the mean, 99.7 percent lie ±3 standard deviations from the mean. A key qualifier of the following discussion is that it applies only *when the data are normally distributed.* Although you can use computer programs to calculate the standard deviation, for explanatory purposes we note that the *standard deviation* is defined as[9]

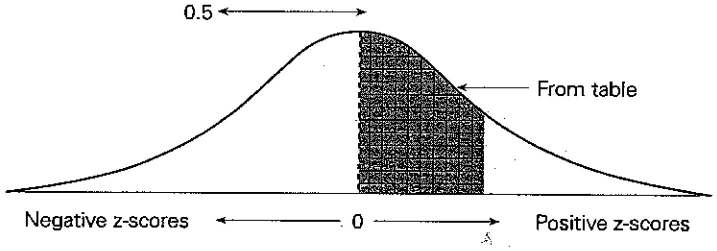
$$s=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$$

Thus, when an observation lies far from its mean, $(x_i-\bar{x})^2$ is large, and when an observation lies close to the mean, $(x_i-\bar{x})^2$ is small. Likewise, when *most* observations are scattered widely around the mean, $\Sigma(x_i-\bar{x})^2$ is large, and when *most* observations are scattered narrowly around the mean, $\Sigma(x_i-\bar{x})^2$ is small. Thus, data that are widely spread around the mean will have a larger standard deviation than data that are closely clustered around the mean. This is shown in Figure 7.6.[10] Computer programs also calculate s^2, called the *variance* of a variable. However, this measure has no particularly important properties and is provided as information only.[11]

For the data shown in the stem-and-leaf plot (see Figure 7.1), the computer will calculate that the mean is 56.39 and the standard deviation is 27.43. Thus, when the data are normally distributed, about two-thirds of the observations will lie between 28.96 and 83.82. About 95 percent lie between the values of 1.53 and 111.25.

The distributional properties of the standard deviation can also be used to determine what percentage of values lie above or below a given value. For example, what is the percentage of caseloads in which female welfare clients have more than 2.6 children, if the mean number of children in such caseloads is 2.20, the standard deviation is 0.44, and the variable is normally

Figure 7.7 Standard Normal Distribution



distributed? To answer such questions, we need to compare our values against a table showing such percentages for the standard normal curve, which is defined as a normal distribution that has a mean of 0 and a standard deviation of 1. Appendix A shows values for areas under the standardized normal distribution. Shown also in Figure 7.7, the area under this curve is 1.00, which means that an observation has a probability of 1.00 of being somewhere under the curve.[12] Note that the areas to the left and right of the midpoint (z = 0) are both 0.50, as the curve is symmetrical. All data can be standardized by using the formula $z = (x_i - \bar{x})/s$, and the resulting values are called ***z-scores*** (or standardized values). Variables whose values have been standardized are called ***standardized variables***, and computers can readily calculate them.

For any given z-score, the question is what percentage of observations have values greater or smaller than it. We plug the above information into the z-score formula and find a standardize value of [(2.60 – 2.20)/0.44 =] 0.91. Appendix A shows areas to the left of the midpoint; for z = 0.91, that area is shown as 0.3186. Thus, 81.86 percent [50% + 31.86%] of cases have a value less than the value that gave rise to this z-score (2.60), and 1 minus 0.8186, or 18.14 percent, have a larger value. Note that negative z-scores indicate a probability less than 0.50. For example, the z-score of caseloads with 1.65 children among female welfare clients in the same sample is [(1.65 – 2.20)/0.44 =] –1.25. The area associated with 1.25 in Appendix A is 0.3944 (only positive values are shown), but negative z-score values indicate areas to the left of the mean, and so [0.5 – 0.3944 =] 10.56 percent of caseloads have fewer than 1.65 children, and [0.5 + 0.3944 =] 89.44 percent of caseloads have more than 1.65 children.

The standard deviation is also used to identify outliers in normally distributed variables. When variables are normally distributed, values that lie more than ±3 standard deviations from the mean are often considered outliers. Thus, for the data of Figure 7.1, values greater than [56.39 + (3*27.3) =] 138.68 or smaller than [56.39 – (3*27.3) =] –25.90 would be considered outliers. No such values are present in Figure 7.1. Of course, outliers can also be identified using boxplots (see the earlier section on boxplots as a method for dealing with outliers).[13]

The standard deviation is also used to calculate the ***confidence interval***, defined as the range within which a statistic is expected to fall on repeated sampling. This is also used to estimate population means from samples. Random samples are often used to estimate population characteristics (see Section II), and statisticians have long worked on the problem that different random samples will yield somewhat different results. A confidence interval also expresses how certain we can be that the real, but unknown population parameter falls within an interval. By convention, the term *parameter* refers to a population characteristic, whereas *statistic* refers to a sample characteristic.

The formula for calculating a 95 percent confidence interval of the mean (that is, the range within which the mean will fall in 95 of 100 samples) is $\bar{x} \pm 1.96 * s / \sqrt{n}$. The measure, $s / \sqrt{n}$, is also called the *standard error of the mean.* The value 1.96 is the z-value corresponding with 95 percent of the area under the standard normal distribution (0.475 on both sides) in Appendix A, and is discussed more fully in Section IV. For the data in Figure 7.1, using this formula, we expect the population mean to lie between [56.39 ± (1.96*27.43) / $\sqrt{124}$ =] 51.56 and 61.22. The formula for calculating a 99 percent confidence interval is $\bar{x} \pm [(2.58 * s)\ s / \sqrt{n}$. See Box 7.2 for further information about confidence intervals.

The above formula has cousiderable utility for managers and analysts. It provides a measure of how accurate the estimate of the mean is without having to consider or draw other samples; it is based ouly on the single sample at hand. It also prevents the manager and analyst from having to stake his or her conclusions on a single number, such as 56.39. If the mean of data taken from some other sample or data in the next period is, say, 52.59, this value would still within the realm of current findings.

Getting Started [illegible]

Finally, computers also calculate two measures that assist in determining whether data are normally distributed. These measures are used in conjunction with visual inspection efforts. ***Skewness*** is a measure of whether the peak is centered in the middle of the distribution. A positive value indicates that the peak is "off" to the left, and a negative value suggests that it is off to the right. ***Kurtosis*** is a measure of the extent to which data are concentrated in the peak versus the tail. A positive value indicates that data are concentrated in the peak; a negative value indicates that data are concentrated in the tail (giving the curve a "fat tail"). Values of skewness and kurtosis have little inherent meaning, other

In Greater Depth...

Box 7.2 Confidence Intervals

Pollsters typically use the term *sampling error* to indicate a 95 percent confidence interval. Thus, if survey results report that 79.5 percent of program clients are satisfied with a program, with a sampling error of, say, 3 percent, then the analyst is stating that he or she is 95 percent certain that between 76.5 percent and 82.5 percent of all clients are satisfied with the program. The probability of the estimate falling within the confidence interval, here 95 percent, is also called the *confidence level.*

The formula for calculating the 95 percent confidence interval for a *proportion, p,* such as from *categorical* variables, and based on a large sample ($n > 100$), is as follows:

$$p \pm 1.96^* \sqrt{[p(1-p)/n]}.$$

Sampling errors, such as those shown in Table 5.2, are usually calculated for $p = .500$, which produces the largest possible confidence interval. For example, the 95 percent confidence interval for $n = 124$ and $p = .500$ is $.5 \pm .088$, and for $p = .795$, it is $.795 \pm .071$.

The formula for calculating confidence intervals in small samples ($n < 100$) of continuous and normally distributed variables is analogous, but it uses the so-called "t-distribution" (discussed further in Chapter 11) to determine the constant. Appendix C shows that this value for a 95 percent confidence interval increases from 1.96 for a large sample to 2.086 for $n = 20$. For our sample, with mean = 56.39 and $s = 27.43$, we can be 95 percent certain that the population mean lies between $56.39 \pm [2.086^*(27.43 / \sqrt{20})] = 43.60$ and 69.18. This larger interval than that shown in the text, or n = 124, reflects less certainty in our estimates because we have fewer observations. The theoretical underpinnings of confidence intervals involve inference and hypothesis testing (see Section IV).[14]

than that large values indicate greater asymmetry. A rule of thumb is that the ratio (absolute value) of skewness to its standard error, and of kurtosis to its standard error, should be less than two (these statistics are calculated by the computer). Large ratios indicate departure from symmetry. The respective ratios of skewness and kurtosis of our data, as calculated by the computer, are $| -0.06/0.22 |$ and $| -0.73/0.43 |$, which are both well below 2.0. Thus, our data are well centered; the tail is a little fat but not enough to cause us to worry about the normality of our data.[15]

Many computer programs can also superimpose a curve over a histogram to help tell analysts if their data are normally distributed. If this curve looks close to a perfect bell-shaped curve, the data are considered normally distributed. Many statistical tests discussed in Chapter 11 and beyond assume that variables are normally distributed; we return to this important matter in subsequent chapters.

SUMMARY

Measures of dispersion provide important information about the distribution of a variable's values, and they also help with data cleaning. Frequency distributions show the percentage of observations (for example, clients or employees) that score in or above a certain category. Frequency distributions often are reported in tabular form, and are very common in reports. They are also used to generate graphs that are used in reports and presentations; such graphs are essential in helping analysts to communicate their findings to a broader audience.

The boxplot is a useful tool for data cleaning. In particular, it helps to detect outliers, or extreme values. Outliers generally should be retained when they are not coding errors, when they are plausible values of the variable in question, and when they do not greatly affect the value of the mean. Analysts commonly first use boxplots for data cleaning, then analyze means and medians, and finally examine frequency distributions.

Finally, this chapter also examines the normal distribution. Many variables are normally distributed. When variables are normally distributed, the standard deviation is a measure of the spread of the data value around the mean, and confidence intervals can be calculated that express the range within which a statistic is expected to fall. Such information provides an appropriate sense of the accuracy of calculated means.

KEY TERMS

Bar charts (p. 115)
Boxplot (p. 118)
Confidence interval (p. 123)
First quartile (p. 118)
Frequency distributions (p. 112)
Histogram (p. 114)
Inner fence (p. 119)
Interquartile range (p. 118)
Kurtosis (p. 123)
Line charts (p. 115)
Measures of dispersion (p. 111)
Midspread (p. 118)
Normal distribution (p. 120)
Outer fence (p. 119)
Outliers (p. 118)
Pie charts (p. 115)
Range (p. 118)
Skewness (p. 123)
Standard deviation (p. 120)
Standardized variables (p. 122)
Stem-and-leaf plots (p. 113)
Third quartile (p. 118)
Whiskers (p. 119)
Z-scores (p. 122)

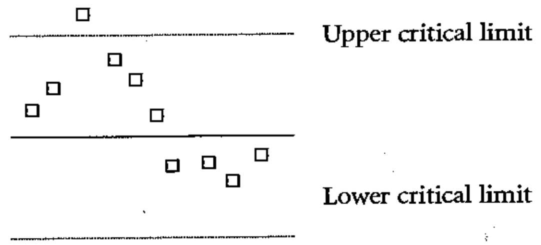
Notes

1. These data are from the Employees Attitudes dataset, which can be found on the CD that accompanies the workbook, *Exercising Essential Statistics.*
2. Both approaches are found in practice. As noted in Chapter 6, analysts should be mindful that fractional values (means) are not defined for ordinal variables. The write-up should report the percentages of data values in relevant categories.
3. Table 6.1 (see appendix to Chapter 6) shows how data for a continuous variable might be recoded into five categories. For example, categories for the data in Table 6.1 could have been created with unequal intervals (such as category 1 with range 1–14, category 2 with range 15–17, and category 3 with range 18–25) that would lead to a different conclusion.
4. For example, in SPSS, stem-and-leaf plots are produced by the "Explore" routine (Analyze → Descriptive Statistics → Explore).
5. In SPSS, category widths are often defined after the default histogram is generated; the editing options allow category widths (called "bin sizes") to be changed.
6. For example, decisions about educational programs are based on the mean attainment of students in those programs, expressed as "on average, enrolled students improved their XYZ ability by x percent more than students who did not enroll in the programs." Of course, irregular behavior draws attention in its own right, and public laws are often passed to address that. In addition, a lot can be learned from case studies of infrequent behavior.
7. Some computer programs, including SPSS, distinguish between outliers and extreme values. Then, outliers are defined as observations that lie 1.5 times the midspread from the first and third quartiles, whereas extreme values are observations that lie 3 times the midspread from the first and third quartiles.
8. But if the sample is consistent with a bell-shaped curve and if we had an infinite number of drawings, the sample would eventually look normal. This matter is taken up in Chapter 11.
9. To calculate the standard deviation for a population, divide by *N*.
10. Standard deviations can also be calculated from grouped data, Chapter 6 (Appendix), using the following revised formula:

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

where the *i*'s indicate the group categories. Consider an example. Referring to the data in Table 6.1, we first calculate the estimated group means of each category and then subtract these values from the previously calculated group mean. The estimated category means are 3, 8, 13, 18, and 23, respectively. Subtracting the value of the overall mean of 15.1, we get −12.1, −7.1, −2.1, 2.9, and 7.9. Then we take the squared difference of each, which is 146.4, 50.4, 4.4, 8.4, and 62.4, and weight each of these values by the number of observations in each category. Thus, the value for the first category is [12*146.4 =] 1,756.8, and subsequent values are 252.0, 79.2, 302.4, and 873.6. We add these numbers, get 3,264, and divide by 85, which is 38.4, and then take the square root, which is 6.2.

11. However, in Chapter 12 we will see that *variance* is used to explain the amount of variation in a variable.
12. A curve in which the area is 1.00 is also called a *density curve.*
13. Because boxplots do not assume variables to be normally distributed, they have broader use as a tool for data cleaning. However, boxplots cannot be used for constructing confidence intervals, and so both tools have their place.
14. Standard deviations also underlie the development of control charts, which are used in assessing production and service delivery processes. Control charts help managers determine the likelihood that unusually high or low performance is caused by chance. The upper and lower critical limits of control charts are defined as $UCL = \bar{x} + 3(s / \sqrt{n})$ and $LCL = \bar{x} - 3(s / \sqrt{n})$:



15. Another rule of thumb is that both of the following measures should be less than ±1.96: *skewness*/$\sqrt{6/n}$ and *kurtosis*/$\sqrt{24/n}$. In our case, the respective values are −0.27 and 1.66.