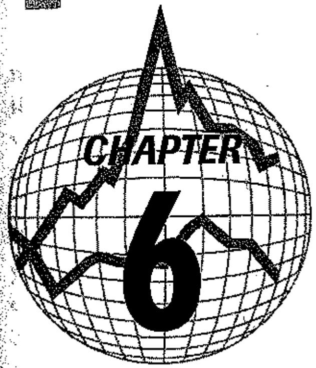
Data cleaning is the process of identifying and removing reporting and recording errors. Errors include mistyped values, errors that arise in the process of uploading, and other implausible values that have been recorded. Data cleaning is aided by univariate analysis, and examples are shown in Chapter 7. It is common practice to assume that unexamined data usually contain various errors that must be identified and removed. Without data cleaning, such errors may have a biasing effect on your results.

Data cleaning usually consists of several activities. First, analysts identify implausible values in their data that they then remove or correct. For example, a variable "age" that has a value of "999" certainly requires further investigation. This might be a coding error or it might be that the value is used to indicate a missing value, in which case analysts should ascertain that "999" is defined in their software program as a missing value for this variable. Analysts can examine the highest and lowest values of their variables and ascertain whether they are plausible. Second, analysts ascertain that their dataset is complete and accurate. To this end, the number of observations (records) in the dataset is verified against the number of records in the source (paper or electronic). In addition, a random sample of records in the statistical software is compared against corresponding records in the original data source; analysts will want to ensure that the data in these records match exactly. Analysts might also compare whether statistics (for example, the mean) are identical between variables in the statistical software program and the original data source. When this not the case, problems with variables or groups of observations may be indicated. Only *after* the analyst has determined that the data are complete and free from data-coding and data-entry errors can data analysis proceed further.

Note

[1] The workbook that accompanies this text includes a manual with examples for data coding and data input into SPSS or any other statistical software program. The workbook also includes an SPSS user's guide.



Central Tendency

CHAPTER OBJECTIVES

After reading this chapter, you should be able to

- Identify three statistics of central tendency
- Calculate the mean, median, and mode
- Know appropriate uses of the mean, median, or mode
- Address problems of missing data
- Know when and how to weight data
- Estimate measures of central tendency from grouped data

The first family of univariate analysis is ***measures of central tendency***, which provide information about the most typical or average value of a variable. Although measures of central tendency are popularly referred to as averages, they are in fact three separate measures: the *mean, median,* and *mode.* Analysts frequently use these types of measure when reporting on, for example, high school violence, housing starts, pollution, and the like. Analysts should always indicate which measure is being used.

Chapters 2 and 3 introduced important research concepts. It is worth reviewing these concepts here, because the discussion that follows illustrates their relevance. Succinctly, ***variables*** are key to research and are defined as empirically observable phenomena that vary. High school violence, housing starts, and pollution are examples of empirical phenomena that vary. Management and policy is very much about changing or shaping variables in some way to make society better off—with a bit less high school violence, more affordable housing, less pollution, and so on. ***Attributes*** are defined as the characteristics of a variable, that is, the specific ways in which a variable can vary. For example, high school violence can be measured as being absent, sporadic, occurring from time to time, or ongoing; these are the attributes of the variable "high school violence." Gender has two attributes, namely, "male" and "female," and so on.

A ***scale*** is defined as the collection of specific attributes (or values) used to measure a specific variable. There are *four* levels of measurement scales: *nominal, ordinal, interval,* and *ratio.* Because many statistics require that variables have certain levels of measurement, managers and analysts must be able to determine the level of measurement for their variables. A ***nominal-level scale*** is one that exhibits no ordering among the categories. Gender is a nominal variable: we cannot say that "male" is more than "female" or vice versa. By contrast, an ***ordinal-level scale*** is one that exhibits order among categories but without exact distances between successive categories. Likert scales, which are common on surveys, are examples of ordinal scales. A typical example of a Likert scale is one with the following response categories: Strongly Agree, Agree, Don't Know, Disagree, and Strongly Disagree. Variables with nominal- and ordinal-level scales are referred to as ***categorical variables.***

Interval- and ***ratio***-level variables are those whose scales exhibit both order *and* distance among categories. For example, someone who earns $75,000 per year makes exactly three times that of someone making $25,000. The *only* difference between interval and ratio scales is that the latter have a true "zero" (for example, height can be zero, but IQ cannot). Variables with nominal- and ordinal-level scales are sometimes referred to as ***continuous variables.*** *Variables, attributes,* and *measurement scales* are of critical importance in statistics. Readers are encouraged to review the more extensive discussions of these concepts found in Chapters 2 and 3.

THE MEAN

The ***mean*** (or arithmetic mean) is what most people call "the average," but analysts should use the word *mean* to avoid confusion with other types of averages. Mathematically, the mean is defined as *the sum of a series of observations, divided by the number of observations in the series.* The term is commonly used to describe the central tendency of variables, such as the mean number of crimes, public safety inspections, welfare recipients, abortions, roads under repair, and so on. The mean is appropriate for continuous variables. Mean calculations are essential to most analyses and are used in almost every report.

The following example shows how to calculate the mean. Although computers and hand calculators are typically used to calculate the mean, you should also understand how to do so by hand. Assume that a sample of eight observations of variable x has the following values (or data elements): 20, 20, 67, 70, 71, 80, 90, and 225 ($n = 8$). Obviously, variable x is just a name that could refer to anything, such as the level of violence, educational attainment, arrests, test scores, and the like. A series of values (such as 20, 20, 67, . . .) is also called an *array.* For the above values, the mean is calculated as follows:[1]

$$Mean = \sum_i x_i \,/\, n =$$
$$(20 + 20 + 67 + 70 + 71 + 80 + 90 + 225)/8 = 643/8 = 80.38.$$

This equation is probably not new to you, though the notation might be. As a second example, the mean of 15, 25, and 50 is [(15 + 25 + 50)/3 =] 30. The notation $\sum_i x_i$ means "the sum of all values of (variable) x," as shown above. In our example, this notation is shorthand for $\sum_{i=1}^{8} x_i$, which specifies adding the first eight values of x, in the order shown. In this example, $x_1 =$ 20, $x_2 = 20$, $x_3 = 67$, and so on. Because our variable has only eight values, there is no need for the notation $\sum_{i=1}^{8} x_i$. Also, n is used to indicate that the observations are a sample. If the observations had constituted the entire population, we would have used a different notation: N (or $\sum_i x_i / N$). This is just a matter of notation, which affects neither the definition nor the calculation of the mean.

Getting Started

[illegible]

Calculating the mean is straightforward, indeed, but managers and analysts may encounter some practical issues that, for the most part, concern the data rather than the formula itself. These concerns are relevant to other statistics, too, and hence illustrate important general matters in statistics. First, variables often have missing data. For example, data may be missing for some clients, about some services, or for some years. We do not like to guess the values of these missing data because it is difficult to credibly justify such guesses. The most common approach is to exclude such observations from calculations; if we do not know x_5, then we generally do not guess it either. The incidental exclusion of a few observations from among many (say, hundreds) will usually not bias results in any material way; indeed, most analyses have a few missing observations. However, bias may occur when the proportion of missing data is large. Then analysts need to acknowledge that an extensive amount of data is missing and will need to add an appropriate caveat to their report. Obviously it is best to avoid using variables that have many missing values.[2]

Second, calculations of means usually result in fractions (for example, "the mean number of arrests is 8.52 per officer"). The presence of fractions implies that distances between categories are measured exactly, hence, that variables are continuous (that is, interval or ratio level). However, analysts frequently have ordinal variables, such as responses to survey questions that are based on a five- or seven-point Likert scale (see Box 3.1). Because fractions are not defined for ordinal scales, analysts should avoid writing, "On average, respondents provide stronger support for item A (3.84) than item B (3.23)." Rather, analysts might write, "On average, respondents provide stronger support for item A than item B. For example, whereas 79.8 percent agree or strongly agree that . . . , ouly 65.4 percent agree or strongly agree that. . . ." The latter phrasing is also easier for many readers to understand or relate to. Nonetheless, this recommendation is not always followed; fractional reporting of ordinal-level variables is commonplace in analytical reports and data tables.[3]

Third, caution should be used with time series data (discussed in depth in Chapter 15). Briefly, dollar values should typically first be expressed as constant dollars (that is, adjusted for inflation), before applying the formula to calculate the mean. Today's dollars do not have the same purchasing power as yesterday's dollars; thus, they first need to be made comparable. The mean assumes that data elements are measured in the same units, including the same kind of dollars.

Fourth, in some cases the mean can be misleading. For example, suppose that most hospital patients stay either one or five days after some type of surgery. Other patients stay other lengths of time, but these lengths are much less frequent. Then we could say that the most common lengths of stay are one and five days. The mean leads us to a single value (say, maybe about three days), which is a poor measure of central tendency in this case because of the two values that occur most often. Though this is not a common situation, awareness of such possibilities may lead analysts to consider the distribution of variables as well (see Chapter 7) and not rely only on measures of central tendency for summarizing and describing their variables.

Finally, in some cases it may be necessary to use a ***weighted mean***, which is defined as a mean for which the observations have been given variable weights. The assignment of weights usually reflects the importance or contribution of each observation relative to other observations. This approach is often taken when measures are based on different population sizes. For example, consider the central tendency of a crime rate of 2.3 percent in city A, with a population of 500,000, and a crime rate of 4.5 percent in city B, with a population of 250,000. Then the mean can be expressed as either [(2.3% + 4.5%)/2 =] 3.4 percent across cities, or as [{(2.3%*500,000) + (4.5%*250,000)}/(500,000 + 250,000) =] 3.0 percent in the region encompassing the populations of cities A and B. The latter is the weighted mean, reflecting the importance of each rate relative to the overall population.[4] Rather than asking which mean is best, it is important to understand the conceptual differences and choose accordingly. Weighted means are also used for adjusting over- and undersampling in surveys. For example, when minorities are undersampled, we might want to weight each of their responses more heavily in order to reflect their actual proportions in the population. Box 6.1 illustrates these calculations. Weighted data can be used to calculate other statistics, too. Each of the eight observations discussed earlier was weighted equally (20 + 20 + 67 + . . .), so using the weighted mean was not necessary.

The issues described here may or may not be germane in every situation. The essential lesson is to be mindful before applying any statistical formula. You need to understand your data and the purposes, definitions, and assumptions of statistical formulas (such as for the mean), and then critically examine summary statistics that result. By taking this approach, you will have increased confidence in the calculated results.

THE MEDIAN

A limitation of the mean is that its usefulness is greatly affected when the data include a few very large or very small values, relative to other values. In the earlier example, if x_8 (the eighth observation in the above sequence, which is 225) had been 950, then the mean for the array would be 171—more than double its initial value! A realistic example of this problem arises when calculating the mean household income in the small hometown of Bill Gates, one of the world's richest people. In that case, the mean is a poor summary statistic of the average household income in that jurisdiction.

The ***median*** is defined as *the middle value in a series (or array) of values.* Its value is, by definition, unaffected by a few very large or small values. *The median should always be used when a few very large or very small values affect estimates of the mean.* Indeed, most summary income statistics of populations report both means and medians because they can be so different. The median is appropriate for both continuous- and ordinal-level variables. The interpretation of the median is that *half of the observations lie above the median, and the other half lie below it.* To find the median, the data must be ordered from low to high and then the value of the middle observation determined. If the number of observations is uneven, the median is the

In Greater Depth...

Box 6.1 Weighting Your Data

Weighted means are easily calculated. The formula for weighted means is $\Sigma w_i x_i / \Sigma w_i$, which means "identify the weights, then multiply each weight with the value of each observation, then add these values and, finally, divide this number by the sum of all weights. Confused? The following example demonstrates this process:

Value	Weight	Weighted Value
2	0.5	1
2	1.0	2
3	2.0	6
3	2.0	6

The unweighted mean is (10/4) = 2.50, and the weighted mean is (15/5.5) = 2.73. Weighted means have many applications, including in survey research. Nonresponse bias is the bias that occurs because survey samples seldom match the population exactly: nonrespondents might have answered differently from respondents. Perhaps the best approach is to conduct a separate survey of nonrespondents and compare their responses with those of the initial respondents. But this method often is expensive and complicated. A second best approach, then, is to compare weighted responses against the actual responses (called unweighted responses). The weighted responses are those that would have been obtained if the sample distribution had perfectly matched that of known population demographics. Typically, census and other sources provide information about age, race, and gender. Consider the following demographics, reported in the workbook *Exercising Essential Statistics*:

Age	Population(%)	Sample(%)
18–45	62.3	62.8
46–65	24.1	26.8
66+	13.6	10.4

Race	Population(%)	Sample(%)
White	81.5	84.3
Nonwhite	18.5	15.7

(continued)

[redacted]

[redacted]

value of the middle observation. If the number of observations is even, the median is the mean of the two observations that are nearest to the middle location of the array. This location is found through visual inspection or the formula $(n + 1)/2$, where n is the number of observations.

In our earlier example, variable x has an even (eight) number of observations, which have already been arrayed from low to high. The two middle values are 70 and 71 (at locations 4 and 5); so the median is 70.50 (at location [(8 + 1)/2 =] 4.50). If a ninth observation is added to variable x, for example, $x_9 = 275$, the median becomes 71. Note that these estimates are unaffected by the values of the highest or lowest values. If $x_9 = 875$, the median is still 71, because the value of this variable does not affect the value of the middle observation in the series. Note that having few very large or very small values can also be caused by data entry errors, for example, coding x_8 as 1,225, when it should be 225. This property of the mean is yet another reason for taking data cleaning seriously.

Examples of the median are common in demographic studies of income, in which a few individuals or households typically have very large incomes. Other examples include studies of average jail time served by inmates (some people serve very long sentences!), average wait times for tax returns or class registration, and average jury awards (some people have received huge sums). A rule of thumb is that when the mean and median are considerably different, analysts should report both. For example, the U.S. Census reports both the mean and median incomes of U.S. households; in 1999 these were, respectively, $56,644 and $41,994, which are considerably different. When the mean and median are similar, it suffices to report only the mean. The measure of what constitutes a "considerable difference" is a judgment call informed by the magnitude of the difference and the study's context. Of course, the earlier cautions about missing data, fractional reporting, time series data, and weighted samples apply when calculating medians, too.

Finally, sometimes analysts only have access to already published, tabulated data tables, rather than having the actual observations as the basis for their calculations. Then, the data have already been grouped, such as by age or income categories. Census data often come in this format, for example. The appendix to this chapter describes how to calculate measures of central tendency for data that have already been grouped.

THE MODE

The ***mode*** is defined as *the most frequent (typical) value(s) of a variable*. The mode is appropriate for all levels of variable measurement. In our example, the mode of variable *x* is the value 20; it occurs twice in the array. Another example is that the mode of people living in households is two. Perhaps the mode of assaults on school grounds is five annually. The mode is used infrequently, but an advantage of the mode is that it can also be used with *nominal*-level data, which is not possible for calculating the mean and median.[5] However, when the mode is used as a measure of central tendency for nominal-level data, managers frequently turn to ***measures of dispersion,*** discussed in Chapter 7, to express the frequency with which the mode occurs. For example, a manager who is analyzing choices of clients or respondents among a range of program options (a nominal variable) will state the number or percentage of clients or respondents who most often chose the most popular program option (that is, the mode).

SUMMARY

Descriptive statistics of the average are commonly used by public managers and analysts. After managers and analysts have verified the accuracy of their data, they may wish to calculate measures of central tendency. There are three such measures: the mean, median, and mode. The most commonly used of these measures is the mean, defined as the sum of a series of observations, divided by the number of observations in the series. Weighted means reflect the importance or contribution of each observation relative to other observations or can be used to account for over- or undersampling. Caution should be exercised in using means with ordinal-level variables.

The median is defined as the middle value in a series (or array) of values. The median should always be used when a few very large or very small values affect estimates of the mean. The mode is defined as the most frequent (or typical) value(s) of a variable. Analysts should always indicate which measure is being used, rather than referring to any of these measures simply as the "average."

The mean is appropriate for continuous-level variables, whereas the median is appropriate for both continuous- and ordinal-level variables. The mode is appropriate for all levels of measurement. Managers and analysts should be mindful when variables are missing a great deal of data or involve time series data.

KEY TERMS

(includes bolded terms in the Section III introduction)

Attributes (see also Chapter 3) (p. 100)
Bivariate analysis (see section introduction) (p. 97)
Categorical variables (see also Chapter 3) (p. 100)
Continuous variables (see also Chapter 3) (p. 100)
Data cleaning (see section introduction) (p. 98)
Data coding (see section introduction) (p. 97)
Data input (see section introduction) (p. 97)
Descriptive statistics (see section introduction) (p. 96)
Grouped data (see appendix to this chapter) (p. 108)
Interval-level scales (see also Chapter 3) (p. 100)
Mean (p. 100)
Measures of central tendency (p. 99)
Measures of dispersion (see also Chapter 7) (p. 106)
Median (p. 103)
Mode (p. 106)
Nominal-level scale (see also Chapter 3) (p. 100)
Ordinal-level scale (see also Chapter 3) (p. 100)
Ratio-level scales (see also Chapter 3) (p. 100)
Scale (see also Chapter 3) (p. 100)
Univariate analysis (see section introduction) (p. 96)
Variables (see also Chapter 2) (p. 100)
Weighted mean (p. 102)

APPENDIX

Using Grouped Data

The calculations described in this chapter have assumed that the analyst has data for each observation. This is the assumption used in statistical software programs. However, analysts sometimes have published data only in tabular format, or in a similar format. ***Grouped data*** refers to observations that have already been grouped in different categories. An example is shown in Table 6.1. The column labeled "Interval of variable *x*" could be almost anything, such as the groupings of city sizes, students' test scores, motorists' speeds through toll booths with electronic collection, or regional water quality ratings. The ranges show the values of each category. Ranges are sometimes shown as footnotes to tables, which then show only categories and frequencies. The "Frequency" column counts occurrences. For example, there are 12 cities in category 1, 5 cities in category 2, and so on. The column *Cumulative frequency* shows the running total of frequencies of each category and may be absent from some grouped data tables.

Calculations of means and medians of grouped data are *best-guess estimates* and should be used *only when individual observations are unavailable.* Unfortunately, few computer programs can read tabular formats of grouped data, in which case calculations must be done by hand. *Note that your ability to make these calculations will not affect your understanding of other material in this book.*

The *mean of grouped data* is calculated in two steps. First, the mean of the categories is calculated using the formula $\sum_i w_i r_i \,/\, \sum_i w_i$, where *r* is the row number and *w* is the number of observations in each row. Applying the data shown in Table 6.1, we find that the weighted mean of categories is [{(12*1) + (5*2) + (18*3) + (36*4) + (14*5)}/(12 + 5 + 18 + 36 + 14) = 290/85 =] 3.412.[6]

Second, the *variable value* associated with this group mean value is determined. This requires interpolation in the following manner: The mean of the grouped data, 3.412, lies somewhere between categories 3 and 4. The estimate of the average variable value associated with category 3 is defined as the midpoint of its range, or [(11 + 15)/2 =] 13, and the midpoint of the value associated with category 4 is [(16 + 20)/2 =] 18. Then the variable value associated with the category location of 3.412 (which is 3.000 + 0.412) is defined as the midpoint estimate of the range associated with category 3 (that is, 13) *plus* 0.412 of the difference of these category midpoints, or [18 – 13 =] 5. Hence, the estimated value of the variable mean is [13 + (0.412*5) =] 15.06 (with rounding). An equivalent expression is that 3.412 "lies 41.2 percent from category 3 toward category 4," which is shown graphically below:

Variable value:	13	←2.06→	|	18
Category value:	3		3.412	4

The *median of grouped data* is estimated in an analogous way. The sample has a total of 85 observations; the median is defined by the value of the forty-third [(85 + 1)/2] observation when values are ordered. Examining the cumulative frequencies, we find that the median falls somewhere between the third and fourth categories:

Variable frequency:	12	17	35		43	71	85
Category value:	1	2	3	←0.222→	|	4	5

The value of the forty-third observation lies [(43 – 35)/(71 – 35) =] 0.222 from category 3 toward category 4, with a category value of 3.222. Using the same method of interpolation described for the group mean, we calculate the corresponding variable value of category location 3.222 as [13 + 0.222*(18 – 13) =] 14.11. Note the difference between the estimated group mean and median. The linear interpolation used for calculating grouped means and medians assumes defined distances between categories, hence, a continuous level of variable measurement.

The *mode of the grouped data* is the most frequent observation. This is category 4, which has a midpoint value of 18. The mode of these grouped data is thus 18.

Table 6.1 Illustration of Grouped Data

Category	Interval of variable x	Frequency	Cumulative frequency
1	1–5	12	12
2	6–10	5	17
3	11–15	18	35
4	16–20	36	71
5	21–25	14	85

Notes

1. Additional examples can be found in the workbook. Also, as a matter of nomenclature, we distinguish between attributes (introduced in Chapter 2) and values. The term *value* refers to the actual, observed responses or

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.