Chapter 4
Summarizing Data with Descriptive Statistics

The purpose of acquiring data is ultimately to help us answer a research question we have proposed. Statistics is the toolkit we use to do so. As we said in the opening chapter, there are three basic goals of statistics: summarization, inference, and prediction. This chapter focuses on summarization. Summarization is the process of taking a potentially large volume of data and reducing it to a few quantities that adequately represent the data so that the data can be easily understood. For example, suppose I collected data on heights and weights of a sample of 500 people. I could simply report all of this information, but it would not be clear exactly what height and weight look like in the sample, let alone the general population, nor would it be clear what the relationship is between them. I could, however, summarize both height and weight in terms of their means and standard deviations, and I could compute the correlation between them to show how they are related. We will begin this chapter by discussing methods for summarizing univariate data, that is, data on one variable. Later in the chapter we will discuss basic methods for summarizing bivariate data, that is, data on two variables. Subsequent chapters will discuss methods for examining relationships between variables.

4.1 Summarizing Nominal Level Measures

When summarizing data, we first need to consider the level of measurement of the variable we are intending to summarize. If the variable is measured at the nominal level, there are limited summaries that can be made of it. In particular, some representation of the proportion of observations in each category of the variable is all that can be done. For example, suppose we have collected information on the race of 100 members of a sample as shown in Table 4.1. In these data, there are 84 whites, 12 blacks, and 4 members of other races. We could summarize these data by simply reporting the percentages in each racial category, or we could summarize this information graphically using a bar chart.

S.M. Lynch, *Using Statistics in Social Research: A Concise Approach*,
DOI 10.1007/978-1-4614-8573-5_4,

W	W	B	W	W	W	W	W	W	W
W	B	W	W	W	W	O	W	W	W
W	W	W	W	W	W	B	B	W	O
O	B	W	W	W	W	W	B	W	W
W	W	W	W	W	W	W	W	W	W
W	W	W	W	W	W	W	W	W	W
W	W	W	W	W	B	W	W	W	W
W	W	W	W	W	B	W	W	W	O
B	W	W	W	W	W	W	W	B	W
W	W	W	W	W	W	W	W	B	B

Table 4.1. Values of race for a (contrived) sample of $n = 100$ persons ($W =$ white; $B =$ black; $O =$ other)

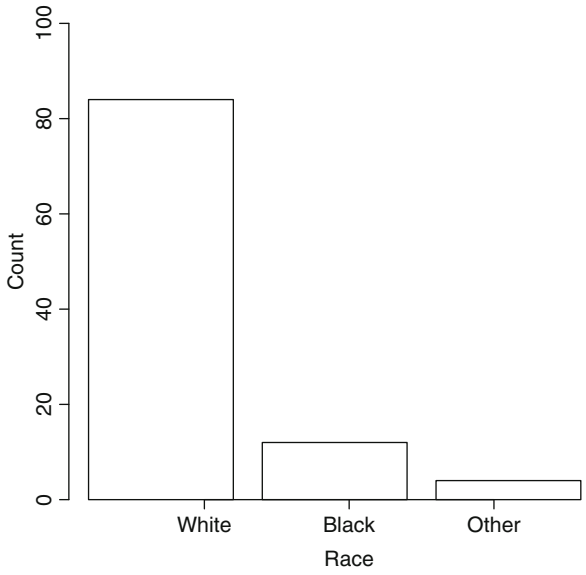

Fig. 4.1. Example of a barchart for the race data shown in Table 4.1

A bar chart provides a visual representation of data, using rectangles of differing heights to represent the proportion (or number) of cases in each category of a variable. Figure 4.1 presents a bar chart for the race data. An important part of constructing a bar chart—or any type of plot or figure, for that matter—is adequate labeling and titling. Notice that, in the figure, the three racial groups are spelled-out as "White," "Black," and "Other." Also notice that the y axis is labeled "Count." We could use "percent" instead, and that is perhaps more common in larger sample sizes. Abbreviations are not used, so the figure is immediately readable without reference to the text.

In addition to the bar chart, pie graphs can also be used to represent the proportion of observations in different categories of a nominal level variable.

Pie graphs, however, are rarely seen in scientific journals, in part because they can be misleading. Pies are always restricted to represent 100 % of a sample (or part of it), and so, in comparing the distribution of one nominal level variable (like religious affiliation) across two groups like whites and nonwhites, we would need to show two pies. These two pies would need to be of different sizes in order to reflect the different sizes of the white and nonwhite samples, and the size difference would have to be reflected in the *area* (not diameter) of the pies. This rule is often forgotten, and the result is that pie graphs often misrepresent the data they are displaying. Therefore, we will not use pie graphs in this book. Such misrepresentation is also common in figures that use symbols (like cows, or houses, etc.) that vary in size in two dimensions to represent different magnitudes. Consequently, we will not use cute symbols in the book either!

4.2 Summarizing Interval and Ratio Level Measures

When the variable we are interested in summarizing is measured at the interval or ratio level (it is numeric), there are considerably more summary measures and graphical displays that we can implement. Usually, we want to start by getting some idea of the center of the data as well as some idea about the spread of the data around its center. In statistics, measures of the center of the data are called "measures of central tendency," while measures of the spread of the data around the center are called "measures of dispersion."

Table 4.2 presents a sample of data on years of schooling for persons ages 80 and above. Our goal is to summarize these data.

4.2.1 Measures of Central Tendency

There are three basic statistics that are commonly used to represent the center of a set of data: the mean, the median, and the mode. The mean is commonly called the "average" outside of statistics and is computed as:

12	12	12	17	12	12	14	12	6	17
12	8	12	8	5	9	12	12	12	8
12	16	0	17	7	11	9	12	8	8
11	14	6	12	4	6	9	11	12	0
8	14	7	8	12	8	8	10	12	17
11	12	17	10	12	12	7	12	7	12
13	7	12	12	16	10	10	8	12	6
12	8	5	16	8	0	12	0	12	12
5	8	12	12	12	7	8	14	10	8
7	9	11	12	9	12	6	12	10	12

Table 4.2. Sample of $n = 100$ values of years of schooling for persons ages 80+.

0	0	0	0	4	5	5	5	6	6
6	6	6	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8
8	8	8	8	8	9	9	9	9	9
10	10	10	10	10	10	11	11	11	11
11	12	12	12	12	12	12	12	12	12
12	12	12	12	12	12	12	12	12	12
12	12	12	12	12	12	12	12	12	12
12	12	12	12	12	12	12	13	14	14
14	14	16	16	16	17	17	17	17	17

Table 4.3. Sample of $n = 100$ values of years of schooling for persons ages 80+ sorted to facilitate finding median.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{4.1}$$

where $\bar{x}$ is pronounced "x bar," the expression $\sum_{i=1}^{n}$ means the sum of x across all members of the sample (indexed by i), and n is the sample size (so all members are $x_1, x_2, x_3, \ldots, x_n$). In the years of schooling data, $\bar{x} = 10.12$.

The median is another measure of central tendency; it measures the central value in a set of data. Thus, if we have n sample members, to find the median we would sort the x from the smallest to the largest values and take the $(n+1)/2$th observation as the median if n is odd and take $[x_{n/2} + x_{(n+2)/2}]/2$ (the average of the two centermost observations) as the median if n is even. Here, n is even, and so we need the average of x_{50} and x_{51}.

The median in the years of schooling data is 11—the average of the two centermost observations (see Table 4.3). Notice that the median is not equal to the mean: In this case, the median is greater than the mean.

The third measure of the center of a set of data is the mode. The mode is simply the most frequently occurring value. From Table 4.3 it seems clear that the most frequently occurring value in the data is 12 years of education.

A stem-and-leaf plot is a good way to display the mode and also to visually represent the data to get a sense of how the data are distributed (see Fig. 4.2). In the stem-and-leaf plot, the values to the left of the vertical line are called the stems, while the values to the right are called the leaves. There are several ways to construct a stem-and-leaf plot, and the best way depends on the values in the data. Generally, the values of the stems are the leading digits in the data values; the leaves are the last digit. For example, in the last line of the plot, the 1 represents 10, and each 7 represents a value of 17 in the data. In these data there are five individuals with 17 years of education; hence, there are five "7" leaves. If, as with these data, there are only a few values for the stems (e.g., here, 0 and 1), the stems may be repeated as many times as needed, as long as we evenly divide the stems and leaves. Here, each subsequent stem and leaf row in the figure represents a one-unit increment.

0	000
0	
0	
0	
0	4
0	555
0	66666
0	7777777
0	888888888888888
0	99999
1	000000
1	11111
1	22222222222222222222222222222222222
1	3
1	4444
1	
1	666
1	77777

Fig. 4.2. Stem-and-Leaf plot of education data

Most data sets contain so many observations that it would be unreasonable to treat each leaf as representing a single individual. In such cases, each leaf may be used to represent a number of individuals, as long as this number is consistent throughout the plot.

An alternate way to represent data graphically is via a histogram. A histogram is similar to the stem-and-leaf plot, but imagine turning the stem-and-leaf plot 90 degrees so that the stems are horizontal (see Fig. 4.3). Then, imagine replacing the leaves with rectangles with heights equal to the peaks of the leaves (see Fig. 4.4).

The histogram looks very much like the bar chart presented earlier, but with a few exceptions. First, the ordering of the bars on the x axis is irrelevant in a bar chart, because there is no inherent ordering of the categories of a nominal level variable. In contrast, because a histogram shows the distribution of variables in which the categories are inherently ordered, the bars cannot be rearranged. Second, the bars in a histogram generally touch one another, while the bars in a bar chart do not. This distinction is symbolic: interval and ratio variables are generally assumed to be continuous, meaning that there is no break between adjacent categories of the variable. Put another way, all possible values between observed values are considered to exist, at least theoretically. For example, while we may only observe years of schooling in integer values, fractional values exist in theory.

The assumed continuity of values in numeric measures allows us to use an alternative approach to constructing histograms: we may replace the bars with a

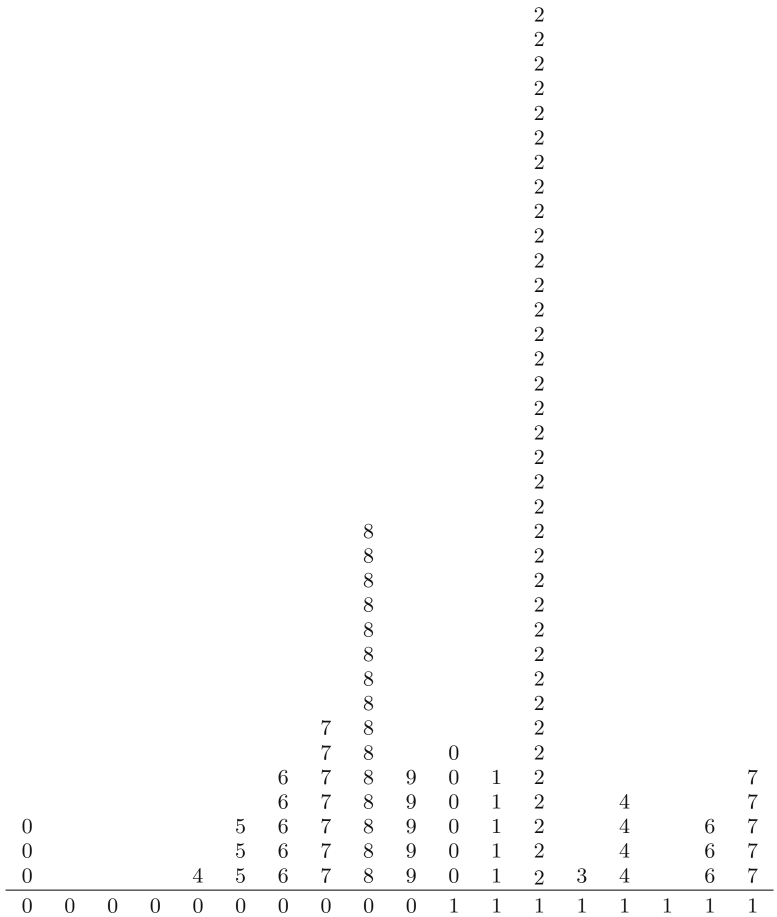

Fig. 4.3. Horizontal version of stem-and-leaf plot for education

smooth curve passing through the height of the bars from one end of the distribution to the other. Figure 4.5 shows this approach.

Returning to the measures of central tendency, with these data, the mean was 10.12, the median was 11, and the mode was 12. Generally speaking, the mean, median, and mode of a variable will not be equal. If the data follow a true normal distribution (a bell curve), then all three values *will* coincide, but this is rarely the case. Instead, sample data are typically asymmetrically distributed around the mean to some extent. The histogram above, for example, shows that the distribution of education is asymmetric, lumpy, and skewed to the left. Skewness is a property that refers to the direction in which the tail of the distribution is longer: here, the bulk of the distribution is clustered around 10–12 years of education, with a long left-hand tail that stretches back to 0. The right tail, in contrast, extends only to 17 years.

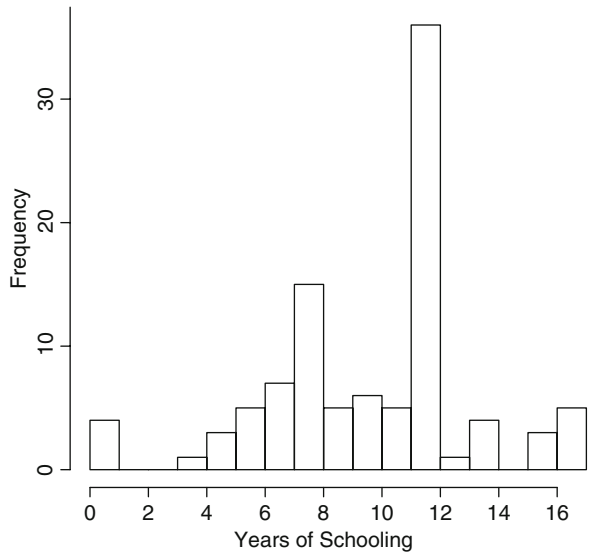

Fig. 4.4. Example of a histogram for the education data.

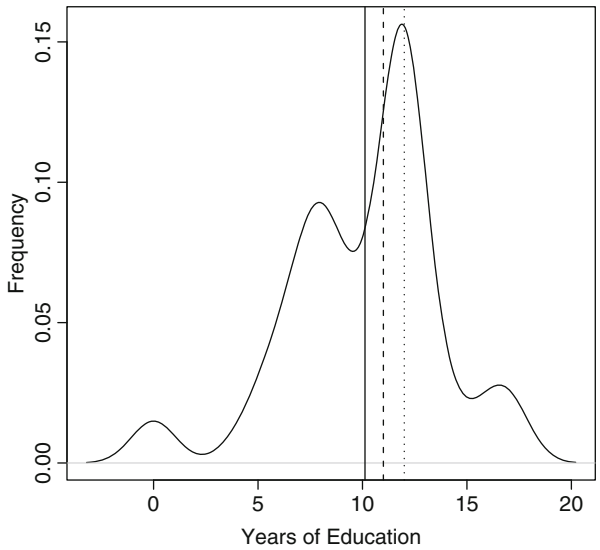

Fig. 4.5. Another approach to plotting a histogram (mean is *solid line*; median is *dotted line*; mode is *dashed line*)

Skewness is typically easy to discern graphically, but it can often be determined from the mean and median as well: the mean is *usually* pulled in the direction of the skew.[1] Thus, in distributions that are known to be highly skewed—like income—the median is usually a better measure of the center of the distribution.

The mode, in comparison to the mean and median, may not even be in the center of the distribution (although it generally is). Some distributions have more than one mode, that is, they may have multiple peaks of roughly equivalent height. In such cases, we call the distribution "multimodal." The distribution for education above is not truly multimodal, but the data do exhibit several peaks: at 0, 8, 12, and 17. The peaks at 0 and 17 are attributable to boundary issues with the response categories. There may be a number of people who had 17 *or more* years of education, but limiting the measure to 17+ years lumps everyone with 17 or more years into one category, making a peak. On the other end of the distribution, there are a number of people in older birth cohorts who received no formal education. The reason for the peak at 12 years is obvious: many people make it through high school but do not go beyond high school. The peak at 8 years is a phenomenon relevant primarily to older cohorts: among older birth cohorts, many people dropped out of school after completing grammar school (primary school) but before entering high school.

4.2.2 Measures of Dispersion

In addition to examining measures of the center of distributions of data, we usually would also like to know something about the spread of the data around the center. Some data, like education among younger birth cohorts, are fairly narrowly clustered around the center of the distribution. For example, among younger birth cohorts, most persons graduate from high school, and a large minority graduate college. Very few people do not graduate high school (although we are now seeing a reversal of this trend toward graduation), and most people do not go beyond a 4-year college degree (although there are more and more people who obtain professional degrees that require 2–3 additional years beyond the bachelor's degree). Thus, overall, education data among younger birth cohorts tend to be fairly clustered between 12 and 16 years, with very small tails of the distribution outside that range.

In comparison to education, income inequality has grown dramatically over the last 30 years, and so its distribution is quite dispersed and drastically skewed to the right. The median household income in the US is a little over $50,000, which means that half of the households in the population earn less than that. However, there are certainly households in the US that earn more than $100,000 (about 19 %; an equal

[1] I say *usually* because there are cases in which this rule does not hold. See von Hippel (2005) for more detail. Although we will not discuss skewness further here, a measure of skewness can be computed. Positive values indicate a right-skewed distribution; negative values indicate a left-skewed distribution.

spacing from $0), and indeed, there are a few that earn well more than $1,000,000 per year (only about 1.5 % earn more than $200,000, however).

The purpose of measures of dispersion is to quantify the amount of spread around the center of the distribution. There are four primary measures of dispersion that are used in statistics: the range, the interquartile range, the variance, and the standard deviation. The range is the easiest to calculate: technically it is simply the difference between the maximum and minimum observed values on a variable. In the education data, for example, the minimum was 0 years and the maximum was 17. The range, therefore, is 17. However, the range is often reported as though it were an interval: 0–17.

The range is sensitive to extreme values—consider income. If one person in the population makes $1,000,000,000 per year in income, the range is drastically larger than it would be if that person did not exist and the nearest second made only $1,000,000 per year. Thus, a better rudimentary measure of the spread of a distribution is the interquartile range (IQR).

In order to understand the IQR, we must first define quartiles. More generally, we need to understand quantiles. Quantiles are the cutpoints that divide a (sorted) data distribution into some number of equal-sized subsets. If we are interested in breaking the data into two equal sized groups, the median is the only quantile. If we are interested in breaking the data into three equal sized groups, there are two "tertiles:" the first is the value in the data below which $1/3$ of the data fall; the second is the value below which $2/3$ of the data fall.

We can divide the data into as many quantiles as we would like, but the most commonly-used quantiles are quartiles (four subsets), quintiles (five subsets), deciles (10 subsets), and percentiles (100 subsets). As a rule, for k groups, we need $k-1$ cutpoints to divide the distribution. Thus, for quartiles, we need three cutpoints. The second cutpoint—Q_2—is the median. The first and third quartiles can be found by taking the median of each half of the data produced by dividing the data at the median.

For the education data, the median was 11. The first quartile cutpoint (Q_1) is the median of the lower half of the data, which we can find just as we found the median for the entire data set. Here, $Q_1 = 8$. The third quartile cutpoint is the median of the upper half of the data; its value is $Q_3 = 12$. The interquartile range is computed as $Q_3 - Q_1$ and is thus 4. Notice that, if we added a few more values at either end of the distribution, the IQR would most likely be unaffected. If it were affected, it would probably change only slightly, because the cutpoints are based on the ordering of the data and not so much the magnitude of the values in the data.

With the IQR defined, an additional plot can be defined that is often of use in visualizing the shape of a distribution: the boxplot (sometimes called the box-and-whisker plot). Boxplots require several pieces of information, but once this information is obtained, it is a simple plot to construct. In order to construct a boxplot, we first need the first quartile cutpoint, the median (the second quartile cutpoint), and the third quartile cutpoint. First, a box is drawn, with the first and third cutpoints being the left and right edges of the box, respectively. The median is drawn as a line within the box. After the box is drawn, the "whiskers" are added.

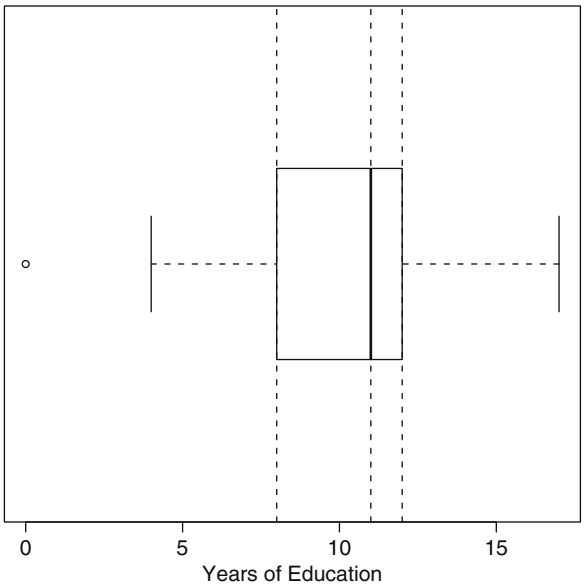

Fig. 4.6. Example of a boxplot for the education data (Q_1, Q_2, and Q_3 are highlighted with *dashed lines*).

The edge of the whiskers is usually computed as the value in the data that is within and closest to 1.5 times the IQR from the first (or third) quartile. Thus, for example, in our data, 1.5 times the IQR is 6. The first quartile cutpoint was 8, and thus, the left whisker would end at the value in the data that is closest to 2 without being beyond 2 in the direction of the tail. In our data, that number would be 4. On the other end of the distribution, the third quartile cutpoint was 12. $12+6=18$, and the largest value in the data that does not exceed 18 is 17. Thus, the values of importance in these data for constructing a boxplot are: 4,8,11,12, and 17. 4 would be the left edge of the lower whisker, 8 is the left edge of the box, 11 is the median (a line in the box), 12 is the right edge of the box, and 17 is the right edge of the upper whisker. Any value that falls outside the whisker—an "outlier"—is generally marked with a symbol like a circle or asterisk. Figure 4.6 shows the boxplot for the education data.

The last two measures of dispersion are the variance and the standard deviation. The variance is the average of the squared deviations of the data around the mean, and is denoted s^2:

$$s^2 \equiv var(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}. \tag{4.2}$$

The denominator of this calculation is $n-1$ rather than n for technical reasons, but the $n-1$ ensures that our sample estimate s^2 is an unbiased estimator of

(A) Education (x)	(B) Number of Cases ($f(x)$)	(C) ($\bar{x} = 10.12$) $(x - \bar{x})^2$	(D) $(B) \times (C)$
0	4	102.41	409.66
4	1	37.45	37.45
5	3	26.21	78.64
6	5	16.97	84.87
7	7	9.73	68.14
8	15	4.49	67.42
9	5	1.25	7.27
10	6	.014	.086
11	5	.77	3.87
12	36	3.53	127.24
13	1	8.29	8.29
14	4	15.05	60.22
16	3	34.57	103.72
17	5	47.33	236.67
	$\sum = 100$		$\sum = 1,292.56$

Table 4.4. Variance calculations for education data. Dividing the sum at the bottom right by 99 ($n - 1$) yields 13.06.

the population variance, σ^2. Notice that the numerator sums up the deviations of each observation from the sample mean, squaring this value before summing. Thus, extreme values increase the variance disproportionately compared to values that are closer to the mean. In our data, the variance is 13.06. Table 4.4 shows this computation. The table uses a convenient calculation formula, which simply involves computing the squared deviation from the mean for each unique value in the data and then multiplying by the number of times that value appears. For example, there are four values of 0 in the data; thus, we need only compute $(0 - 10.12)^2$ once and then multiply this value by four. We do not need to directly calculate this deviation four times. This computational formula can be written as:

$$s^2 = \frac{\sum_x f(x) \times (x - \bar{x})^2}{n - 1}. \tag{4.3}$$

The summation symbol, Σ simply has an index of x as shorthand to show that we are summing over all unique values of x in the data.

In addition to the frequency-based formula shown above, another formula can be derived from the original variance formula, and it involves a "trick" that will be important in various formulas later in the book. I show the derivation of this, often-called "computational method," here, which simultaneously illustrates the properties of distributing the summation symbol.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{4.4}$$

$$= \left(\frac{1}{n-1}\right)\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n} \bar{x}^2\right) \tag{4.5}$$

$$= \left(\frac{1}{n-1}\right)\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right) \tag{4.6}$$

$$= \left(\frac{1}{n-1}\right)\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \tag{4.7}$$

Equation 4.5 expands Eq. 4.4 using typical algebraic expansion ("FOIL"). The summation is simply distributed to each term. Equation 4.6 uses a couple of rules. First, multiplicative constants can be move outside of sums: $\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i$. Here, $2\bar{x}$ is a constant, and so it can be moved outside the summation, leaving $2\bar{x}\sum_{i=1}^{n} x_i$. The trick here is to recognize that, since $\bar{x} = \sum_{i=1}^{n} x_i/n$, $\sum_{i=1}^{n} x_i$ therefore is $n\bar{x}$. This leaves the center term as shown in Eq. 4.6. The last term in this equation is $n\bar{x}^2$, because we are summing $\bar{x}^2$ n times. Combining the latter two terms leaves us with Eq. 4.7. Implementing this formula is often easier than implementing the original variance formula, because we do not need to take the deviation of each observation from the mean before squaring it.

The variance is not readily interpretable, because the units are squared. A more informative measure, therefore, is the standard deviation. The standard deviation is just the square root of the variance—$s = \sqrt{s^2}$—but because it is square-rooted, its unit of measurement is the same as the original variable. This is a property that makes the standard deviation a nice measure to help summarize the spread of the data: we can say the data cluster around the mean, give or take $z \times s$ units, where z is the number of standard deviations away from the mean we want our interval to be. In these data, the standard deviation is 3.61, and so we can say that most people have 10.12 years of schooling give or take $z \times 3.61$ years of education. We will be discussing this idea in much greater depth soon, especially the meaning of, and appropriate values for, z.

4.3 Summarizing Bivariate Data

Summarizing variables one at a time is important, but we are often interested in simultaneously summarizing two variables in order to get a stronger feel for what members of a sample look like or in order to get a feel for the relationship between two variables. In terms of getting a better feel for the sample, suppose we had collected information on respondents' sexes (male and female) and races (say white and nonwhite). Univariate summaries of sex and race would be important, but it

Sex	Race		
	White	Black	Total
Male	689	118	807
	85 %	15 %	100 %
	44 %	38 %	
	37 %	6 %	43 %
Female	861	193	1,054
	82 %	18 %	100 %
	56 %	62 %	
	46 %	10 %	57 %
Total	1,550	311	1,861
	100 %	100 %	
	83 %	17 %	100 %

Table 4.5. Cross-tabulation of sex and race for the 2010 GSS.

would be more informative to have the sex-by-race breakdown of the sample. Recall from the previous chapter our discussion of the ecological fallacy: simply knowing that 49 % of the sample is male and 51 % is female, and that 80 % of the sample is white, while 20 % is non-white, does not tell us how many nonwhite males we have in our sample. As we discussed in the previous chapter, we could summarize sex and race simultaneously with a crosstab.

If the two variables we are considering are both measured at the nominal level, a crosstab is about as informative as any graphic summary we could produce. We could make a three dimensional bar chart (or an expanded two dimensional one), but the crosstab is probably the most useful summary, especially if row and column percentages are included. Table 4.5 presents a crosstab of race (whites and blacks only) and sex from the 2010 GSS with row, column, and total percentages. The format for crosstabs with percentages may vary. Sometimes only row percentages are displayed; sometimes only column percentages are displayed. In the table, I have given all three in the following order: row, column, total. It is easy to determine which is which in the table by observing which sets of percentages sum to 100 %. For example, the first set of percentages (in the first row only) are 85 and 15 %. These sum to 100 %, and so these are row percentages. They tell us the proportion of observations that are in each column, *for a given row*. Here, 85 % *of males* in the sample are white, while 15 % *of males* are black. Similarly, among females, 82 % are white, while 18 % are black.

The second set of percentages in the cells of the table sum to 100 % by column, and so they are column percentages. The table shows that, *of whites,* 44 % are male and 56 % are female. Among blacks, 38 % are male, while 62 % are female. Finally, the third set of percentages in the cells only sums to 100 % if you sum all of them: $37 + 6 + 46 + 10 = 100$ (after rounding). Thus, these are total percentages; they tell us what proportion of the overall sample falls in a given cell. For example, 6 % of the sample are black males.

How do I know whether I want row or column percentages? The answer depends on my comparison of interest. If I am interested in how genders compare in terms of their racial composition, then I would be interested in the row percentages in this table. Then I could make statements like: the ratio of whites to blacks among males is 85 to 15, but the ratio for females is 82 to 18. Thus, racial diversity is slightly higher for females than for males. This comparison of ratios might be important if, for example, I were studying the impact of imprisonment or mortality on the racial composition of the (noninstitutionalized) population. Men have much higher rates of imprisonment and mortality than women, and so the proportion of men vs. women in a sample needs to be taken into account if we are interested in how imprisonment and mortality affect racial composition. Given that the ratio for men is 85/15 (5.67), while the ratio for women is 82/18 (4.56), it seems that the mortality and imprisonment gap between nonwhite and white men may be larger than the mortality and imprisonment gap between nonwhite and white women (see Western 2009). I say "may be," because other factors, like race-by-gender differences in survey response may account for part of the difference in ratios.

If I were interested in how races compare in terms of their gender composition, then I would be interested in the column percentages here. There is a literature in sociology and economics that discusses a marriage dilemma for black women, especially for highly educated black women. The dilemma is that there are few black men available for marriage to black women, in part because of high mortality among black males, in part because of high imprisonment rates among black males, and in part because of lower average educational attainment and high employment rates among black males relative to black females (see Wilson 1987). Here, let's consider mortality and imprisonment only (since education is not in our crosstab). Evidence for the marriage dilemma can be found in the column percentages: among whites, the male-to-female ratio is 44–56 % (.79), but among blacks, the ratio is 38–62 % (.61). In other words, there are 79 white men for every 100 white women among noninstitutionalized whites, but there are only 61 black men for every 100 black women. Notice that the gender differential is somewhat severe for both races in the survey, because women are simply more likely to be respondents to the GSS interview. However, we might expect the gender difference in response to be somewhat comparable across the two races, but recall that it is possible that there may be race-by-gender differences in survey participation.

If the bivariate data one is interested in summarizing consists of one nominal level variable and one continuous variable, summarizing the data simply requires separating the data by category of the nominal level variable (called "disaggregating" the data) and computing univariate descriptive statistics. These can then be listed together in a table for visual comparison. In terms of graphic representation of such bivariate data, it is straightforward to put multiple histograms (one for each of the categories of the nominal variable) on a single plot.

If the bivariate data you are interested in summarizing consists of two numeric variables, you can either construct a scatterplot, as we will discuss later in the book, or you can categorize one of the variables and follow the strategy discussed in the previous paragraph. However, it is important to keep in mind that treating a numeric

variable as nominal costs information. For example, suppose your data consisted of years of schooling and annual earnings. You could categorize education, say, as less than 12 years of schooling vs. 12 years or more, and then compute summary measures for each education group. Certainly you will find that those with 12+ years of schooling earn more than those with less schooling. But, you lose information about what type of pattern may exist across the entire distribution of schooling: It is also the case that those with 16+ years of schooling earn more than those with 12–15 years.

4.4 What About Ordinal Data?

Thus far, we have considered summarizing nominal and numeric data, but we have not discussed summarizing ordinal data. Technically, because the categories of ordinal measures can be ordered, but the spacing between adjacent categories is not necessarily consistent, measures of central tendency and dispersion are inappropriate for such data. Simply put, the numbers assigned to the categories are arbitrary. Consider the self-rated health item discussed previously in which the outcome categories were: excellent, good, fair, poor. We could assign any value to the "excellent" category, and as long as we assign increasing or decreasing values to the subsequent categories, the order of the categories would be maintained. Thus, we could assign the values 1, 2, 3, and 4 to the categories, but we could just as well assign the values 1, 2.5, 7, and 100. Both maintain the ordering of the categories, and that is all the information an ordinal measure contains. Yet these two alternate approaches to assigning values would yield extremely different values for the measures of central tendency and dispersion we have discussed. On the other hand, if we treat the ordinal data as if it were measured at the nominal level, we lose the information about the ordering of the categories.

There are a number of measures that historically have been used to describe ordinal data and especially relationships between them. However, more often than not, ordinal data are treated as numeric, and we will generally follow that practice in this book. We will assign the lowest value of an ordinal variable either 0 or 1, and assign subsequent values to categories sequentially.

4.5 The Abuse and Misuse of Statistics

In this chapter we have defined a number of measures that summarize data. Such measures can, and often are, misused either unintentionally or intentionally, and it is therefore important to be aware how they can be misused or abused. It is also important to describe some adjustments that can be made to these basic measures in order for them to be more useful for accurate presentations of data.

One of the first ways in which descriptive statistics can be misused is to use the wrong measure of central tendency when presenting data with a skewed distribution. As we discussed earlier, the mean can be misleading as a measure of central tendency when the distribution of a variable has a strong skew. For example, suppose there are 10 people in a population, with 9 of them earning $10,000 per year, and one earning $1,000,000 per year. Suppose further that, after 10 years, the person earning $1,000,000 is earning $10,000,000, while the other 9 people are still earning $10,000. One could claim that "average" income increased from $109,000 per year to $1,009,000—a factor of almost 10. Yet, the median didn't change at all. Thus, if one were interested in painting a favorable picture of a policy that generated such an income change, they may choose to report the mean. However, it would clearly be a misleading portrait. For the majority of the population, there was no improvement.

A slightly more complicated approach to presenting misleading information involving dollar amounts is to ignore inflation when reporting change over time. Inflation is a complicated topic, but everyone understands the basic idea: over time, the value of the dollar, in terms of its purchasing power, declines. When I was a child, for example, one could buy a can of soda from a machine for about $.25. In most places today, the cost is now over $1.00. For this reason, in most cases, when researchers report changes in income over time, they adjust the dollar amounts to be in stable units, like 2010 dollars. However, if one is interested in showing that incomes have increased drastically over time, s/he could simply report raw income amounts. For example, in 1970, median individual income was around $7,000 in raw dollars. By 2000, it was around $30,000 in raw dollars. Thus, it appears that income more than quadrupled. However, in fact, after adjusting for inflation, it is easy to see that incomes haven't changed at all.

Although most people would not be fooled by ignoring adjusting income amounts for inflation, many ignore inflation when it comes to other quantities in dollars, like, for example, the price of stamps. I have heard people complain about the cost of first class stamps (and I've been known to complain as well). When I was born, the price of a first class stamp was 8 cents. It is now 46 cents, a factor of almost six times as much. Yet, after adjusting for inflation, 8 cents in my year of birth translates into exactly 46 cents in today's dollars.

Aside from adjusting for inflation, dollars and other quantities that involve counts of things often need additional adjustments. In particular, *denominators* are important. The field of demography—the study of populations—is extremely concerned with adjusting raw numbers using appropriate denominators, because things that can be counted need a context in which to interpret their value (see Preston et al. 2001). For example, saying that 10,000 people per year die from some disease is meaningless unless we know how many people there are in the population. In a population like the U.S. with more than 300 million people, roughly 11,000 people die from stomach cancer each year. Yet, stomach cancer is nowhere near being a leading cause of death (it isn't even in the top 30). An appropriate adjustment to contextualize stomach cancer deaths, therefore, would be to divide by the population size. If you do so, you find that the proportion of the population

that dies from stomach cancer each year is .000035—less than one 100th of 1 %. Given such a small number, a demographer might multiply it by 100,000 and call the result—3.5—the death rate per 100,000 population per year.

While few people attempt to paint stomach cancer as a major health problem, you can certainly find other conditions that produce a comparable number of deaths per year for which advocacy groups may use that number to portray it as a serious problem. The fundamental problem with using such raw numbers is that any large population will necessarily produce large numbers of deaths (or any event) from even small actual rates. Furthermore, comparing two populations of unequal sizes will always make the larger population look much worse (or much better, depending on the measure) if no adjustment is made for population size differences. Sometimes, adjustment also needs to be made to compensate not just for population size, but also age structure. For example, the crude death rate, defined simply as the number of deaths in a year divided by the number of people in the population at mid-year, is nearly twice as high in the U.S. as it is in Mexico. However, this is simply because the U.S. has a much older population: median age in the U.S. is just over 37 years, while median age in Mexico is just over 27 years.

In addition to the failure to use a denominator in presenting statistics, it is common to see advocates use an *improper* denominator to alarm people or to convince them to purchase a product that may not be that important, ultimately. One improper denominator is the clock. We could, for example, say that a person dies from stomach cancer every 48 min in the U.S. Even more startlingly, thinking about all-cause mortality, someone dies every 12 s in the U.S. from *something.* The problem with using time by itself as the denominator is that the clock never changes size, even though populations do. A population half as large as the US, but in which the proportion of deaths due to stomach cancer is exactly the same as in the U.S. (so, 5,000 deaths), will have deaths half as often. In other words, large populations will always look worse than smaller populations when time is the denominator. As a specific example, South Korea has approximately the same number of deaths from stomach cancer each year as the U.S. Thus, we can say South Korea experiences a stomach cancer death every 48 min as well. However, the population of South Korea is about 50 million—one-sixth of the U.S—and so its stomach cancer death rate is over 21 per 100,000. In short, stomach cancer is a major killer in South Korea, but not in the U.S., but using an inappropriate denominator distorts this reality.

Time is an important *part* of the denominator in some cases. For example, when measuring traffic accident fatalities, we may use miles driven *per year* as the denominator. Indeed, if one wanted to compare death rates across modes of transportation (like planes, trains, and automobiles), one would need to use miles traveled per some time unit to make the measures comparable. In demography, person-years are often the denominator for rates. Consider, for example, that, if 10 people die during the course of a year in a population that began with 100 persons, the deaths probably did not occur all at the end of the year. Thus, out of 100 persons who began the year alive, not all of them lived for a full year, and so a natural denominator for calculating the rate of death would involve person-years lived. If we

assumed that, on average, those who died did so in the middle of the year, then there would be 95 person-years lived by the 100 people, and the death rate would be 10/95. Over a 1 year period, the denominator may not be much different, regardless of when the deaths occurred. However, consider how different the results might be if one were interested in some rate over a 5 or 10 year interval and all deaths occurred in the first year versus the last.

Sometimes, denominators can be constructed correctly, but results can be misleading because, in forming ratios to compare two groups (called "relative risk ratios"), the denominators drop out of the calculation. This problem is common in epidemiological work in evaluating causes of disease in which risk ratios are often constructed. For example, suppose you were told that nail biters have ten times the rate of tongue cancer deaths as non-nail biters. My agenda might be to encourage you to stop biting your nails (perhaps I'm selling a product that helps with that), and a factor of 10 seems large. However, tongue cancer has one-fifth the death rate each year as stomach cancer. Thus, if 2,000 people per year die from tongue cancer, then about 1,800 will have been nail biters, and about 180 will have not been nail biters. While that *relative* difference might seem dramatic, the *absolute risk* for both groups is small and possibly not worth considering.

This problem permeates public health promotion campaigns, as well as pharamaceutical advertising. Yet, it occurs in far more settings. Consider the issue of racial/religious profiling by airport security. Many support such profiling on the basis that "Muslims" are more likely than "Christians" to be terrorists. While that may or may not be true, what is the absolute proportion of either population that are terrorists? I would suppose it is far less than 1 % for either group. Thus, is it really reasonable to base a policy on the ratio? Similarly, keep in mind that you are *infinitely* more likely to win the lottery if you buy a lottery ticket than if you don't. Does that therefore merit your buying a ticket?

Each of the aforementioned problems essentially concerns using inappropriate measures, or inappropriately adjusted measures, of central tendency to mislead, either intentionally or otherwise. Measures of dispersion can be equally misleading, most often when they are ignored. Measures of dispersion exist to provide some sense of how much variation exists in the population; focusing exclusively on measures of central tendency can lead to misunderstanding the phenomenon of interest. For example, the average high temperature in Oklahoma is about 70 ° (F), which sounds fairly mild, but it is certainly worth knowing that Oklahoma has several weeks' worth of days over 100 ° and a few months of sub-freezing temperatures each year before deciding to move there with a limited wardrobe.

More seriously, those who have a vested interest in showing that one group is worse (or better) than another on some characteristic often fail to mention variation when reporting group differences. For example, it has been long claimed that there are racial differences in IQ, based on a difference in mean IQ between whites and blacks in the U.S. of about 5–10 points (see Dickens and Flynn 2006). Yet, this focus on the mean ignores that the variation that exists within each group produces such an overlap that the distributions of IQ are more similar than different. Given the

importance of considering both central tendency and dispersion in order to evaluate whether differences between groups are meaningful, we will focus much of our attention on this topic in the remainder of the book.

We could fill up many pages discussing ways in which very basic statistics can be abused; indeed, books have been written on the topic (e.g., Hooke 1983; Huff 1993). The ease with which they can be abused is partially to blame for the claim that you can prove anything with statistics. Hopefully, this section, however, has been sufficient in showing that there are clearly wrong ways to present statistics and has encouraged you to become a more critical consumer of them.

4.6 Conclusions

In this chapter, we discussed several measures used to summarize potentially large amounts of data via only a few numbers. For nominal level data, there are relatively few possible summary measures and graphics. For numeric data, there are more measures. For this type of data, measures are divided into two groups: those that reflect the center of the distribution of data and those that reflect the spread of the data around its center. In addition, we discussed several plots that are incredibly useful for providing immediate, visual summaries of data. We then extended our discussion to summarizing data on pairs of variables simultaneously using cross-tabulations. We will illustrate the summaries repeatedly throughout the remainder of the book.

4.7 Items for Review

Be familiar with the following concepts, terms, and items discussed in the chapter. Be able to perform the calculations and produce the plots listed.

- Goals of statistics (summarization and inference)
- Measures of central tendency (mean, median, mode)
- Measures of dispersion (range, interquartile range, variance, standard deviation)
- Quantiles
- Bar chart
- Stem-and-leaf plot
- Boxplot
- Histogram
- Skewness (right/left)
- Cross-tabulation (row and column percentages)

4.8 Homework

Below is a tiny subset of the GSS data. Use these data for the following questions.

Person	Age (in years)	Sex (1 = male; 2 = female)	Education (in years)	Health (0 = poor. . . 3 = excellent)
1	40	2	13	3
2	29	2	12	1
3	55	1	16	3
4	74	1	12	0
5	67	2	9	2
6	19	1	11	2
7	41	2	12	1
8	24	2	10	3
9	41	2	13	2
10	21	2	12	2
11	39	2	14	2
12	85	2	12	1
13	71	2	10	2
14	18	1	11	2
15	30	2	14	3
16	30	2	10	3
17	41	2	16	2
18	60	2	9	1
19	31	2	12	3
20	56	2	12	2

1. Construct a histogram for health.
2. Construct a stem and leaf plot for education.
3. Construct a boxplot for education.
4. Compute the mean, median, and mode for education.
5. Is the distribution of education symmetric or skewed? If it is skewed, in what direction is the skew?
6. Compute the mean, median, and mode for age.
7. Compute the range, IQR, variance, and s.d. for education.
8. Compute the range, IQR, variance, and s.d. for age.
9. Construct a crosstab for sex and health and interpret.
10. Split the sample into those who are above vs. below the median age. Then, compute the mean and standard deviation of health for both age groups. How does the distribution of health compare across the two age groups?