2 Descriptive Statistics

LEARNING OBJECTIVES

In Chapter 1, a fundamental distinction was drawn between descriptive and inferential statistics. We saw that describing data constitutes an important early phase of the scientific method. In this chapter, we will focus upon visual and numerical descriptive summaries of data.

In this chapter, we will begin by describing different types of data and by covering some of the visual approaches that are commonly used to explore and describe data. Following this, numerical measures of description are reviewed. Finally, visual and numerical description is discussed for the special context of spatial data.

2.1 Types of Data

Data may be classified as *nominal*, *ordinal*, *interval*, or *ratio*. Nominal data are observations that have been placed into a set of mutually exclusive and collectively exhaustive categories. Examples of nominal data include soil type and vegetation type. Ordinal data consist of observations that are ranked. Thus it is possible to say that one observation is greater than (or less than) another, but with this much information, it is not possible to say by how much an observation is greater or less than another. It is not uncommon to find ordinal data in almanacs and statistical abstracts; examples include data on the size of cities, by rank.

AQ changed from 10 – OK?

When it is possible to say by how much one observation is greater or less than another, data are either interval or ratio. With interval data, differences in values are identifiable. For example, on the Fahrenheit temperature scale, 44 degrees is 12 degrees warmer than 32 degrees. However, the "zero" is not meaningful on the interval scale, and consequently ratio interpretations are not possible. Thus 44 degrees is not "twice as warm" as 22 degrees. Ratio data, on the other hand, *does*

TABLE 2.1 **Commuting data**

(a) Data on inidividuals

Individual no.	Commuting time (min.)	Individual no.	Commuting time (min.)
1	5	16	42
2	12	17	31
3	14	18	31
4	21	19	26
5	22	20	24
6	36	21	11
7	21	22	19
8	6	23	9
9	77	24	44
10	12	25	21
11	21	26	17
12	16	27	26
13	10	28	21
14	5	29	24
15	11	30	23

(b) ranked commuting times

5, 5, 6, 9, 10, 11, 11, 12, 12, 14, 16, 17, 19, 21, 21, 21, 21, 21, 22, 23, 24, 24, 26, 26, 31, 31, 36, 42, 44, 77

have a meaningful zero. Thus 100 degrees Kelvin *is* twice as warm as 50 degrees Kelvin. Most numerical data are ratio data – indeed it is difficult to think of examples for interval data other than the Fahrenheit scale.

Data may consist of values that are either *discrete* or *continuous*. Discrete variables take on only a finite set of values – examples include the number of sunny days in a year, the annual number of visits by a family to a local public facility, and the monthly number of collisions between automobiles and deer in a region. Continuous variables take on an infinite number of values; examples include temperature and elevation.

2.2 Visual Descriptive Methods

Suppose that we wish to learn something about the commuting behavior of residents in a community. Perhaps we are on a committee that is investigating the potential implementation of a public transit alternative, and we need to know how many minutes, on average, it takes people to get to work by car. We do not have the resources to ask everyone, and so we decide to take a sample of automobile commuters. Let's say we survey $n = 30$ residents, asking them to record their average time it takes to get to work. We receive the responses show in panel (a) of Table 2.1.

We may summarize our data visually by constructing *histograms*, which are vertical bar graphs. To construct a histogram, the data are first grouped into categories. The histogram contains one vertical bar for each category. The height of the bar represents the number of observations in the category (i.e., the frequency),

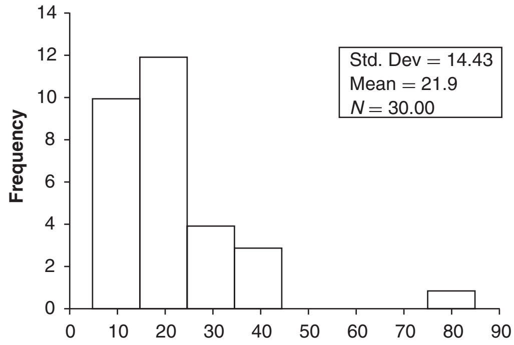

FIGURE 2.1 **Histogram for commuting data**

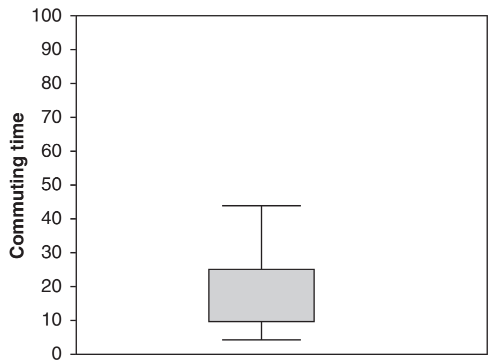

FIGURE 2.2 **Boxplot for commuting data**

and it is common to note the midpoint of the category on the horizontal axis. Figure 2.1 is a histogram for the hypothetical commuting data in Table 2.1, produced by *SPSS for Windows 12.0*. An alternative to the histogram is the *frequency polygon*; it may be drawn by connecting the points formed at the middle of the top of each vertical bar.

Data may also be summarized via *box plots*. Figure 2.2 depicts a box plot for the commuting data. The horizontal line running through the rectangle denotes the median (21), and the lower and upper ends of the rectangle (sometimes called the "hinges") represent the 25^{th} and 75^{th} percentiles, respectively. Velleman and Hoaglin (1981) note that there are two common ways to draw the "whiskers", which extend upward and downward from the hinges. One way is to send the whiskers out to the minimum and maximum values. In this case, the boxplot represents a graphical summary of what is sometimes called a "five-number summary" of the distribution (the minimum, maximum, 25^{th} and 75^{th} percentiles, and the median).

There are often extreme outliers in the data that are far from the mean, and in this case it is not preferable to send whiskers out to these extreme values. Instead,

```
Frequency         Steam & Leaf

  .00                 0 .
 4.00                 0 . 5569
 6.00                 1 . 011224
 3.00                 1 . 679
 9.00                 2 . 111112344
 2.00                 2 . 66
 2.00                 3 . 11
 1.00                 3 . 6
 2.00                 4 . 24
 1.00          Extremes > =77)
Stem width:         10.00
Each leaf:          1 case(s)
```

FIGURE 2.3 **Stem-and-leaf plot for commuting data**

TABLE 2.2

Number of children	**Absolute frequency**
0	100
1	200
2	300
3	100
4+	50
Total	750
6	30
7	8
8	8
9	3
10	1

whiskers are sent out to the outermost observations, which are still within 1.5 times the interquartile range of the hinge. All other observations beyond this are considered outliers, and are shown individually. In the commuting data, 1.5 times the interquartile range is equal to $1.5(14.25) = 21.375$. The whisker extending downward from the lower hinge extends to the minimum value of 5, since this is greater than the lower hinge (11.75) minus 21.375. The whisker extending upward from the upper hinge stops at 44, which is the highest observation less than 47.375 (which in turn is equal to the upper hinge (26) plus 21.375). Note that there is a single outlier – observation 9 – and it has a value of 77 minutes.

A *stem-and-leaf* plot is an alternative way to show how common observations are. It is similar to a histogram tilted onto its side, with the actual digits of each observation's value used in place of bars. John Tukey, the designer of the stem-and-leaf plot, has said, "If we are going to make a mark, it may as well be a meaningful one. The simplest – and most useful – meaningful mark is a digit." (Tukey, 1972, p. 269).

For the commuting data, which have at most two-digit values, the first digit is the "stem", and the second is the "leaf" (see Figure 2.3).

To give another example, consider school district administrators, who often take censuses of the number of school-age children in their district, so that they may

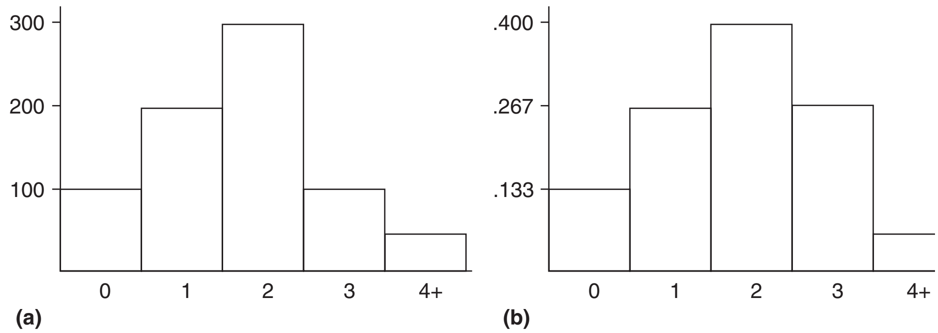

FIGURE 2.4 **Number of children in households: (a) absolute frequency; (b) relative frequency**

form hopefully accurate estimates of future enrollment. Table 2.2 gives the hypothetical responses of 750 households when asked how many school-age children are co-residents.

The absolute frequencies may be translated into relative frequencies by dividing by the total number of observations (in this case, 750). Table 2.3 reveals, for example, that 26.7% of all households surveyed had one child. Note that the sum of the relative frequencies is equal to one. Note also that we can easily construct a frequency polygon using the relative frequencies instead of the absolute frequencies (see Figure 2.4); this frequency polygon in panel (b) has precisely the same shape as that in panel (a); the vertical scale has just been changed by a factor equal to the sample size of 750.

2.3 Measures of Central Tendency

We may continue our numerical descriptive analysis of the data in Table 2.1 by summarizing the information numerically. The *sample mean* commuting time is simply the average of our observations; it is found by adding all of the individual responses and dividing by 30. The sample mean is traditionally denoted by $\bar{x}$; in our example we have $\bar{x} = 21.93$ minutes. In practice, this could sensibly be rounded to 22 minutes. We can use notation to state more formally that the mean is the sum of the observations, divided by the number of observations:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{2.1}$$

where x_i denotes the value of observation i, and where there are n observations. (For a review of mathematical conventions and mathematical notation, see Appendix B.)

The *median* time is defined as the time that splits the ranked list of commuting times in half – half of all respondents have commutes that are longer than the

TABLE 2.3

Number of children	Absolute frequency	Relative frequency
0	100	100/750 = .133
1	200	200/750 = .266
2	300	300/750 = .400
3	100	100/750 = .133
4+	50	50/750 = .067
Total	750	

median, and half have commutes that are shorter. When the number of observations is odd, the median is simply equal to the middle value on a list of the observations, ranked from shortest commute to longest commute. When the number of observations is even, as it is here, we take the median to be the average of the two values in the middle of the ranked list. When the responses are ranked as in panel (b) of Table 2.1, the two in the middle are 21 and 21. The median in this case is equal to 21 minutes. The *mode* is defined as the most frequently occurring value; here the mode is also 21 minutes, since it occurs more frequently (four times) than any other outcome.

Many variables have distributions where a small number of high values cause the mean to be much larger than the median; this is true for income distributions and distance distributions. For example, Rogerson *et al.* (1993) used the US National Survey of Families and Households to study the distance that adult children lived from their parents. For adult children with both parents alive and living together, the mean distance to parents is over 200 miles, and yet the median distance is only 25 miles! Because the mean is not representative of the data in circumstances such as these, it is common to use the median as a measure of central tendency.

Grouped means may be calculated when data are available only for categories. This is achieved by assuming that all of the data within a particular category take on the midpoint value of the category. For example, Table 2.4 portrays some hypothetical data on income (units have been deliberately omitted to keep the example location-free!).

The grouped mean is found by assuming that the ten individuals in the first category have an income of 7500 (the midpoint of the category), the 20 individuals in the second category each have an income of 25,000, the 30 individuals in the next category each have an income of 45,000, and those in the final category each have an income of 77,500. All of these individual values are added, and the

TABLE 2.4

Income	Frequency (Number of individuals)
< 15,000	10
15,000–34,999	20
35,000–54,999	30
55,000–99,999	15

result is divided by the number of individuals. Thus the grouped mean for this example is

$$\frac{10(7{,}500) + 20(25{,}000) + 30(45{,}000) + 15(77{,}500)}{10 + 20 + 30 + 15} = 41{,}167. \qquad (2.2)$$

More formally,

$$\bar{x}_g = \frac{\sum_{i=1}^{G} f_i x_{i,mid}}{\sum_{i=1}^{G} f_i}, \qquad (2.3)$$

where $\bar{x}_i$ denotes the grouped mean, G is the number of groups, f_i is the number of observations in group i, and $x_{i,mid}$ denotes the value of the midpoint of the group.

It is not uncommon to find that the last category is open-ended; instead of the 55,000–99,999 category, it might be more common for data to be reported in a category labeled "55,000 and above". In this case, an educated estimate of the average salary for those in this group should be made. It would also be useful to make a number of such estimates, to see how sensitive the grouped mean was to different choices for the estimate.

2.4 Measures of Variability

We may also summarize the data by characterizing its variability. The commuting data in Table 2.1 range from a low of 5 minutes to a high of 77 minutes. The *range* is the difference between the two values – here it is equal to $77 - 5 = 72$ minutes.

The *interquartile range* is the difference between the 25th and 75th percentiles, and hence can be thought of as the middle half of the data. With n observations, the 25th percentile is represented by observation $(n + 1)/4$, when the data have been ranked from lowest to highest. The 75th percentile is represented by observation $3(n + 1)/4$. These will often not be integers, and interpolation is used, just as it is for the median when there is an even number of observations. For the commuting data, the 25th percentile is represented by observation $(30 + 1)/4 = 7.75$. Interpolation between the 7th and 8th lowest observations requires that we go 3/4 of the way from the 7th lowest observation (which is 11) to the 8th lowest observation (which is 12). This implies that the 25th percentile is 11.75. Similarly, the 75th percentile is represented by observation $3(30 + 1)/4 = 23.25$. Since both the 23rd and 24th observations are both equal to 26, the 75th percentile is equal to 26. The interquartile range is the difference between these two percentiles, or $26 - 11.75 = 14.25$.

The *sample variance* of the data (denoted s^2) may be thought of as the average squared deviation of the observations from the mean. To ensure that the sample variance gives an unbiased estimate of the true, unknown variance of the population from which the sample was drawn (denoted σ^2), s^2 is computed by taking the sum of the squared deviations, and then dividing by $n - 1$, instead of by n. Here the term *unbiased* implies that if we were to repeat this sampling many times, we would find that the average or mean of our many sample variances would be equal to the true variance. Thus the sample variance is found by taking the sum of squared deviations from the mean, and then dividing by $n - 1$:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}. \tag{2.4}$$

An approximate interpretation of the variance is that it represents the average squared deviation of an observation from the mean (it is an approximate interpretation, because $n - 1$, instead of n, is used in the denominator).

In our example, $s^2 = 208.13$. The *sample standard deviation* is equal to the square root of the sample variance; here we have $s = \sqrt{208.13} = 14.43$. Since the sample variance characterizes the average *squared* deviation from the mean, by taking the square root and using the standard deviation, we are putting the measure of variability back on a scale closer to that used for the mean and the original data. It is not quite correct to say that the standard deviation is the average absolute deviation of an observation from the mean, but it is close to correct.

Variances for grouped data are found by assuming that all observations are at the midpoint of their category, and are based on the sum of squared deviations of these midpoint values from the grouped mean:

$$s_g^2 = \frac{\sum_{i=1}^{G} f_i (x_{i,mid} - \bar{x}_g)^2}{(\sum_{i=1}^{G} f_i) - 1}, \tag{2.5}$$

Thus for the data in Table 2.4, the grouped variance is

$$\frac{\begin{array}{c}10(7{,}500 - 41{,}167)^2 + 20(25{,}000 - 41{,}167)^2 + \\ 30(45{,}000 - 41{,}167)^2 + 15(77{,}500 - 41{,}167)^2\end{array}}{(10 + 20 + 30 + 15) - 1} = 4.97356 * 10^8 \tag{2.6}$$

The square root of this, 22,301, is the grouped standard deviation.

2.5 Other Numerical Measures for Describing Data

2.5.1 Coefficient of Variation

Consider the selling price of homes in two communities. In community A, the mean price is 150,000 (units are deliberately omitted, so that the illustration may apply to

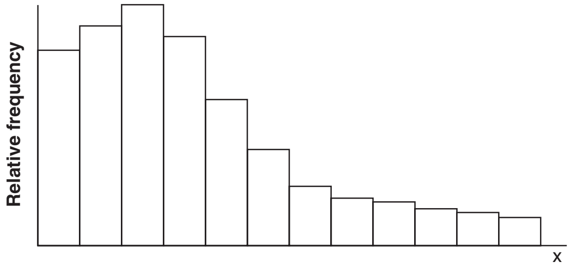

FIGURE 2.5 **A positively skewed distribution**

more than one unit of currency!). The standard deviation is 75,000. In community B, the mean selling price is 80,000, and the standard deviation is 60,000.

The standard deviation is an *absolute* measure of variability; in this example, such variability is clearly lower in community B. However, it is also useful to think in terms of *relative variability*. Relative to its mean, the variability in community B is greater than that in community A. More specifically, the *coefficient of variation* is defined as the ratio of the standard deviation to the mean. Here the coefficient of variation in community A is $75{,}000/150{,}000 = 0.5$; in community B it is $60{,}000/80{,}000 = 0.75$.

2.5.2 Skewness

Skewness measures the degree of asymmetry exhibited by the data. Figure 2.5 reveals that there are more observations below the mean than above it – this is known as positive skewness. Positive skewness can also be detected by comparing the mean and median. When the mean is greater than the median as it is here, the distribution is positively skewed. In contrast, when there are a small number of low observations and a large number of high ones, the data exhibit negative skewness. Skewness is computed by first adding together the cubed deviations from the mean and then dividing by the product of the cubed standard deviation and the number of observations:

$$\text{skewness} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3}. \tag{2.7}$$

The 30 commuting times in Table 2.1 have a positive skewness of 2.06. If skewness equals zero, the histogram is symmetric about the mean.

2.5.3 Kurtosis

Kurtosis measures how peaked the histogram is. Its definition is similar to that for skewness, with the exception that the fourth power is used instead of the third:

$$\text{kurtosis} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4}. \tag{2.8}$$

Data with high degree of peakedness are said to be *leptokurtic*, and have values of kurtosis over 3.0. Flat histograms are *platykurtic*, and have kurtosis values less than 3.0. The kurtosis of the commuting times is equal to 6.43, and hence is relatively peaked.

2.5.4 Standard scores

Since data come from distributions with different means and different degrees of variability, it is common to standardize observations. One way to do this is to transform each observation into a *z-score* by first subtracting the mean of all observations, and then dividing the result by the standard deviation:

$$z = \frac{x - \bar{x}}{s}. \tag{2.9}$$

z-scores may be interpreted as the number of standard deviations an observation is away from the mean. For example, the z-score for individual 1 is $(5 - 21.93)/14.3 = -1.17$. This individual has a commuting time that is 1.17 standard deviations below the mean.

2.6 Descriptive Spatial Statistics

To this point our discussion of descriptive statistics has been general, in the sense that the concepts and methods covered apply to a wide range of data types. In this section we review a number of descriptive statistics that are useful in providing numerical summaries of *spatial* data.

Descriptive measures of spatial data are important in understanding and evaluating such fundamental geographic concepts as *accessibility* and *dispersion*. For example, it is important to locate public facilities so that they are accessible to defined populations. Spatial measures of centrality applied to the location of individuals in the population will result in geographic locations that are in some sense optimal with respect to accessibility to the facility. Similarly, it is important to characterize the dispersion of events around a point. It is useful to summarize the spatial dispersion of individuals around a hazardous waste site. Are individuals with a particular disease less dispersed around the site than are people without the disease? If so, this could indicate that there is increased risk of disease at locations near the site.

2.6.1 Mean Center

The most commonly used measure of central tendency is the *mean center*. For point data, the x- and y-coordinates of the mean center are found by simply finding the mean of the x-coordinates and the mean of the y-coordinates, respectively.

For areal data, one may still find a mean center by attaching weights to, for example, the centroids of each area. To find the center of population for instance, the weights are taken as the number of people living in each subregion. The weighted mean of the *x*-coordinates and *y*-coordinates then provides the location of the mean center. More specifically, when there are *n* subregions,

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i}^{n} w_i}; \quad \bar{y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i}^{n} w_i} \tag{2.10}$$

where the w_i are the weights (e.g., population in region i) and x_i and y_i are the coordinates of the centroid in region i. Conceptually, this is identical to assuming that all individuals living in a particular subregion lived at a prespecified point (such as a centroid) in that subregion.

The mean center has the property that it minimizes the sum of squared distances that individuals must travel (assuming that each person travels to the centralized facility located at the mean center). Although it is easy to calculate, this interpretation is a little unsatisfying – it would be nicer to be able to find a central location that minimizes the sum of distances, rather than the sum of squared distances.

2.6.2 Median Center

The location that minimizes the sum of distances traveled is known as the *median center*. Although its interpretation is more straightforward than that of the mean center, its calculation is more complex. Calculation of the median center is iterative, and one begins by using an initial location (and a convenient starting location is the mean center). Then the new *x*- and *y*-coordinates are updated using the following:

$$x' = \frac{\sum_{i=1}^{n} \frac{w_i x_i}{d_i}}{\sum_{i=1}^{n} \frac{w_i}{d_i}}; \quad y' = \frac{\sum_{i=1}^{n} \frac{w_i y_i}{d_i}}{\sum_{i=1}^{n} \frac{w_i}{d_i}} \tag{2.11}$$

where d_i is the distance from point i to the specified initial location of the median center. This process is then carried out again – new x and y coordinates are again found using these same equations, with the only difference being that d_i is redefined as the distance from point i to the most recently calculated location for the median center. This iterative process is terminated when the newly computed location of the median center does not differ significantly from the previously computed location.

In the application of social physics to spatial interaction, population divided by distance is considered a measure of population "potential" or accessibility. If the *w*'s are defined as populations, then each iteration finds an updated location based upon weighting each point or areal centroid by its accessibility to the current median center. The median center is the fixed point that is "mapped" into itself when weighted by accessibility. Alternatively stated, the median center is an accessibility-weighted mean center, where accessibility is defined in terms of the distances from each point or areal centroid to the median center.

2.6.3 Standard Distance

Aspatial measures of variability such as the variance and standard deviation characterize the amount of dispersion of data points around the mean. Similarly, the spatial variability of locations around a fixed central location may be summarized. The *standard distance* (Bachi 1963) is defined as the square root of the average squared distance of points to the mean center:

$$s_d = \sqrt{\frac{\sum_{i=1}^{n} d_{ic}^2}{n}} \tag{2.12}$$

where d_{ic} is the distance from point i to the mean center.

Although Bachi's measure of standard distance is conceptually appealing as a spatial version of the standard deviation, it is not really necessary to maintain the strict analogy with the standard deviation by taking the square root of the average squared distance. With the aspatial version (i.e., the standard deviation), loosely speaking, the square root "undoes" the squaring and thus the standard deviation may be roughly interpreted as a quantity that is on the same approximate scale as the average absolute deviation of observations from the mean. Squaring and taking square roots is carried out because deviations from the mean may be either positive or negative. But in the spatial version, distances are always positive, and so a more interpretable and natural definition of standard distance would be to simply use the average distance of observations from the mean center (and in practice, the result would usually be fairly similar to that found using the equation above).

2.6.4 Relative Distance

One drawback to the standard distance measure described above is that it is a measure of *absolute dispersion*; it retains the units in which distance is measured. Furthermore, it is affected by the size of the study area. The two panels of Figure 2.6 show situations where the standard distance is identical, but clearly the amount of dispersion about the central location, relative to the study area, is lower in panel (a).

A measure of relative dispersion may be derived by dividing the standard distance by the radius of a circle with area equal to the size of the study area (McGrew and Monroe 1993). This makes the measure of dispersion unitless and standardizes for the size of the study area, thereby facilitating comparison of dispersion in study areas of different sizes.

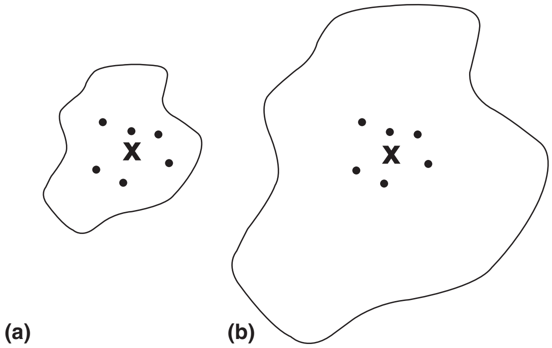

FIGURE 2.6 **Illustration of standard distance. Note that dispersion relative to the study area is lower in panel (b)**

TABLE 2.5

x	y
.8616	.1781
.1277	.4499
.3093	.5080
.4623	.3419
.4657	.3346
.2603	.0378
.6680	.3698
.2705	.1659
.1981	.1372
.8891	.1192

For a circular study area, the relative distance is $s_{d,rel} = s_d/R$ and for a square study area, $s_{d,rel} = s_d\sqrt{\pi/s}$. Note that the maximum relative distance for a circle is 1; for a square the maximum relative distance is $\sqrt{\pi/2} = 1.253$.

2.6.5 Illustration of Spatial Measures of Central Tendency and Dispersion

The descriptive spatial statistics outlined above are now illustrated using the data in Table 2.5. This is a simple set of ten locations; a simple and small dataset has been chosen deliberately, to facilitate the derivation of the quantities by hand, if desired. The study area is assumed to be a square with the length of each side equal to one. In this example we assume implicitly that there are equal weights at each location (or equivalently, that one individual is at each location).

The mean center is (.4513, .2642), and is found simply by taking the mean of each column. The median center is (.4611, .3312). Accuracy to three digits is achieved after 33 iterations. The first few iterations of Equation 2.11 are shown in Table 2.6.

TABLE 2.6

.4512	.2642
.4397	.2934
.4424	.3053
.4465	.3116
.4499	.3159
.4623	.3191
.	.
.	.
.	.
.	.
.4611	.3312

It is interesting to note that the approach to the *y*-coordinate of the median center is monotonic, while the approach to the *x*-coordinate is a damped harmonic.

The sum of squared distances to the mean center is 0.8870; note that this is lower than the sum of squared distances to the median center (0.9328). Similarly, the sum of distances to the median center is 2.655, and this is lower than the sum of distances to the mean center (2.712).

The standard distance is 0.2978 (which is the square root of 0.8870/10); note that this is similar to the average distance of a point from the mean center (2.712/10 = .2712).

Testing the hypothesis that the points are randomly distributed about the center of the square results in a *z*-score of $z = (.2712 - .383)/(\sqrt{.02/10}) = -2.500$ (see Equation 1.1). Since this is less than the critical value of 1.96 associated with a two-sided test using $\alpha = 0.05$, we reject the null hypothesis, and conclude that points are clustered to a greater degree about the center than would be expected under the null hypothesis of random dispersion about the center. A similar conclusion is obtained using Equation 1.2, since

$$z = \sqrt{90(10)}\,\frac{.0887 - .1667}{1} = -2.35. \tag{2.13}$$

2.6.6 Angular Data

Angular data arise in a number of geographical applications; the analysis of wind direction, and the study of the alignment of crystals in bedrock provide two examples. The latter example has been particularly important in the study of continental drift, and in establishing the timing of reversals in earth's magnetic field.

Special considerations arise in the visual and numerical description of angular data. Consider the 146 observations on wind direction given in Table 2.7. A histogram could be constructed, but it is not clear how the horizontal axis should be labeled. A histogram could arbitrarily start with North on the left, as in Figure 2.7a; another possibility is to arrange for the mode to be near the middle of the histogram, as in Figure 2.7b.

TABLE 2.7 **Hypothetical data on wind direction**

Direction	Angular Direction	Frequency
North	0°	10
Northeast	45°	8
East	90°	5
Southeast	135°	6
South	180°	18
Southwest	225°	29
West	270°	42
Northwest	315°	28
Total		146

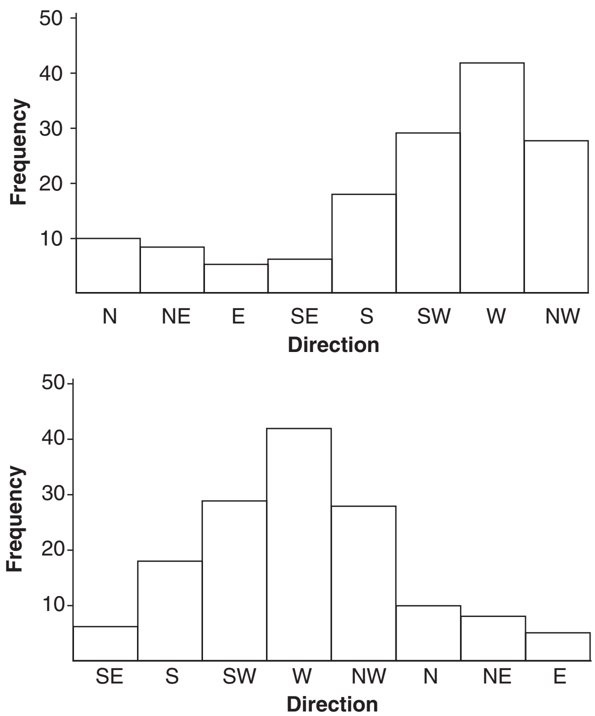

FIGURE 2.7 **Absolute frequencies for directional data**

Neither of the options depicted in Figure 2.7 for constructing a histogram using angular data is ideal, since the observations at the far left of the horizontal axis are similar in direction to the observations on the far right of the horizontal axis. In particular, there is no provision for wrapping the histogram around on itself.

An alternative is the circular histogram (Figure 2.8a). Here bars extend outward in all directions, reflecting the nature of the data. As is the case with more typical histograms, the lengths of the bars are proportional to the frequency. A slight variation of this is the more common rose diagram (Figure 2.8b); here the rectangular bars

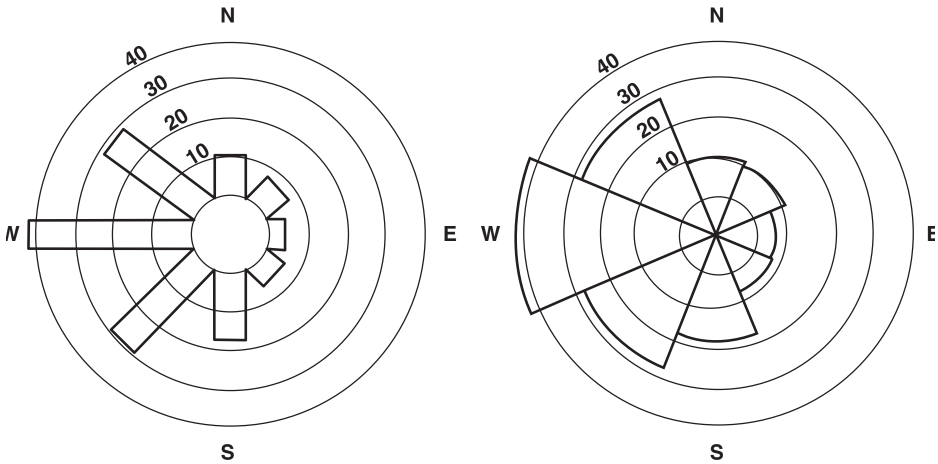

FIGURE 2.8 **Absolute frequencies for directional data: (a) circular histogram; (b) rose diagram**

have been replaced with pie- or wedge-shapes. The rose diagram is effective in portraying visually the nature of angular data.

There are also special considerations that are necessary when considering numerical summaries of angular data. Consider the very simple case where we have two observations – one observation is 1° and the other is 359°. If 0° is taken to be north, both of these observations are very close to north. However, if we take the simple average or mean of 1 and 359, we get $(1 + 359)/2 = 180°$ – due south!. Clearly some other approach is needed, since the average of two observations that are very close to north should not be "south".

Here we describe how to find the mean and variance for angular data.

Mean:

1. Find the sine and cosine of each angular observation.
2. Find the mean of the sines ($\bar{S}$) and the mean of the cosines ($\bar{C}$).
3. Find $\bar{R} = \sqrt{\bar{S}^2 + \bar{C}^2}$
4. The mean angle (say $\bar{\alpha}$) is the angle whose cosine is equal to $\bar{C}/\bar{R}$ and whose sine is equal to $\bar{S}/\bar{R}$. Thus $\alpha = \arccos(\bar{C}/\bar{R})$ and $a = \arcsin(\bar{S}/\bar{R})$.

Variance: A measure of variance for angular data (termed the circular variance) provides an indication of how much variability there is in the data. For example, if all observations consisted of the same angle, the variability and hence the circular variance, should be zero.

The circular variance, designated by $S_{0,}$ is simply equal to $1 - R$. It varies from zero to one. A high value near one indicates that the angular data are dispersed, and come from many different directions. Again, a value near zero implies that the observations are clustered around particular directions.

Readers interested in more detail regarding angular data may find more extensive coverage in Mardia (1970).

EXAMPLE 2.1

Three observations of wind direction yield measurements of 43°, 88°, and 279°. Find the angular mean and the circular variance.

Solution: We start by constructing the following table:

Observation	cosine	sine
43°	0.7314	0.6820
88°	0.0349	0.9994
279°	0.1564	
−0.9877		

The mean of the cosines is equal to $\bar{C} = (.7314 + .0349 + .1564)/3 = 0.3076$. The mean of the sines is equal to $\bar{S} = (.6820 + .9994 - .9877)/3 = 0.2312$. Then $\bar{R} = \sqrt{0.3076^2 + 0.2312^2} = 0.3848$. The mean angle, α, is the angle whose cosine is equal to $\bar{C}/\bar{R} = 0.3076/0.3848 = 0.7994$, and whose sine is equal to $\bar{S}/\bar{R} = 0.2312/0.3848 = 0.6008$.

Using a calculator or table we find that the mean angle is 37°. The circular variance is equal to $1 - \bar{R} = 1 - .3848 = .6152$. This value is closer to one than to zero, indicating a tendency for high variability – that is, the angles are relatively dispersed and are coming from different directions.

2.7 Descriptive Statistics in SPSS for Windows 12.0

2.7.1 Data Input

After starting *SPSS*, data are input for the variable or variables of interest. Each column represents a variable. For the commuting example set out in Table 2.1, the 30 observations were entered into the first column of the spreadsheet. Alternatively, respondent ID could have been entered into the first column (i.e., the sequence of integers, from 1 to 30), and the commuting times would then have been entered in the second column. The order that the data are entered into a column is unimportant.

2.7.2 Descriptive Analysis

2.7.2.1 Simple descriptive statistics Once the data are entered, click on Analyze (or Statistics, in older versions of *SPSS for Windows*). Then click on Descriptive

Statistics. Then click on Explore. A split box will appear on the screen; move the variable or variables of interest from the left box to the box on the right that is headed "Dependent List" by highlighting the variable(s) and clicking on the arrow. Then click on OK.

2.7.2.2 Other Options Options for producing other related statistics and graphs are available. To produce a histogram for instance, before clicking OK above, click on Plots, and you can then check a box to produce a histogram. Then click on Continue and OK.

2.7.2.3 Results Table 2.8 displays results of the output. In addition to this table, boxplots (Figure 2.2), stem and leaf displays (Figure 2.3) and, optionally, histograms (Figure 2.1) are also produced.

AQ supply Table 2.8

Table 2.8 Missing

EXERCISES

1. The 236 values that appear below are the 1990 median household incomes (in dollars) for the 236 census tracts of Buffalo, New York.

   (a) For the first 19 tracts, find the mean, median, range, interquartile range, standard deviation, variance, skewness, and kurtosis using only a calculator (though you may want to check your results using a statistical software package). In addition, construct a stem-and-leaf plot, a box plot, and a histogram for these 19 observations.
   (b) Use a statistical software package to repeat part (a), this time using all 236 observations.
   (c) Comment on your results. In particular, what does it mean to find the mean of a set of medians? How do the observations that have a value of 0 affect the results? Should they be included? How might the results differ if a different geographic scale was chosen?

   22342, 19919, 8187, 15875, 17994, 30765, 31347, 27282, 29310, 23720, 22033, 11706, 15625, 6173, 15694, 7924, 10433, 13274, 17803, 20583, 21897, 14531, 19048, 19850, 19734, 18205, 13984, 8738, 10299, 10678, 8685, 13455, 14821, 23722, 8740, 12325, 10717, 21447, 11250, 16016, 11509, 11395, 19721, 23231, 21293, 24375, 19510, 14926, 22490, 21383, 25060, 22664, 8671, 31566, 26931, 0, 24965, 34656, 24493, 21764, 25843, 32708, 22188, 19909, 33675, 15608, 15857, 18649, 21880, 17250, 16569, 14991, 0, 8643, 22801, 39708, 17096, 20647, 30712, 19304, 24116, 17500, 19106, 17517, 12525, 13936, 7495, 10232, 6891, 16888, 42274, 43033, 43500, 22257, 22931, 31918, 29072, 31948, 36229, 33860, 32586, 32606, 31453, 32939, 30072, 32185, 35664, 27578, 23861, 18374, 26563, 30726, 33614, 30373, 28347, 37786, 48987, 56318, 49641, 85742, 43229, 53116, 44335, 30184, 36744, 39698, 0, 21987, 66358, 46587, 26934, 27292, 31558, 36944, 43750, 49408, 37354, 31010, 35709, 32913, 25594, 25612, 28980, 28800, 28634, 18958, 26515, 24779, 21667, 24660, 29375, 29063, 30996, 45645, 39312, 34287, 35533, 27647, 24342, 22402, 28967, 39083, 28649, 23881, 31071, 27412, 27943, 34500, 19792, 41447, 35833, 41957, 14333, 12778, 20000, 19656, 22302, 33475, 26580, 0, 24588, 31496, 30179, 33694, 36193, 41921, 35819, 39304, 38844, 37443, 47873, 41410, 34186, 36798, 38508, 38382, 37029, 48472, 38837, 40548, 35165, 39404, 34281, 24615, 34904, 21964, 42617, 58682, 41875, 40370, 24511, 31008, 16250, 29600, 38205, 35536, 35386, 36250, 31341, 33790, 31987, 42113, 37500, 33841, 37877, 35650, 28556, 27048, 27736, 30269, 32699, 28988, 22083, 27446, 76306, 19333

(Continued)

(Continued)

2. Ten migration distances corresponding to the distances moved by recent migrants are observed (in miles): 43, 6, 7, 11, 122, 41, 21, 17, 1, 3. Find the mean and standard deviation, and then convert all observations into z-scores.
3. By hand, find the mean, median, and standard deviation of the following variables for subset A of the RSSI dataset: RSSI, slope, altitude, and distance to the nearest cell tower.
4. Using *SPSS* or *Excel*, answer the following questions using the full RSSI dataset:

   (a) Make histograms of (i) RSSI values for those observations that are more than 3000 meters from the nearest cell tower, and (ii) those observations that are less than 3000 meters from the nearest cell tower. Comment on the differences.
   (b) What percentage of RSSI observations are within 2 km of the nearest cell tower?
   (c) What percentage of RSSI observations are made at elevations greater than 400 meters?
   (d) Find the mean, median, and standard deviation for RSSI, slope, altitude, and distance to the nearest cell tower.

5. Using *SPSS* or *Excel*, answer the following questions using the full Tyne and Wear dataset:

   (a) Provide descriptive information on house prices, number of bedrooms, number of bathrooms, floor area, and date built. For each, provide the mean, median, standard deviation, and skewness. Also for each, provide a boxplot.
   (b) What percentage of homes have garages?
   (c) What percentage of homes were built during each of the following time periods: pre-war, inter-war, and post-war?

6. Make histograms of (a) RSSI values for those observations that are more than 3000 meters from the nearest cell tower, and (b) those observations that are less than 3000 meters from the nearest cell tower. Comment on the differences.
7. For the observations in subset A of the RSSI dataset:

   (a) find the mean center
   (b) find the standard distance

8. Given $a = 3$, $b = 4$, and

(Continued)

Observation	x	y
1	3	2
2	5	4
3	7	6
4	2	8
5	1	10

Find the following:

(a) $\sum y_i$
(b) $\sum y_i^2$
(c) $\sum ax_i + by_i$
(d) $\prod x_i$
(e) $2\sum_{i=2}^{3} y_i$
(f) $3x_2 + y_4$
(g) $32!/(30!)$
(h) $\sum_k x_k y_k$

9. Let $a = 5$, $x_1 = 6$, $x_2 = 7$, $x_3 = 8$, $x_4 = 10$, $x_5 = 11$, $y_1 = 3$, $y_2 = 5$, $y_3 = 6$, $y_4 = 14$, and $y_5 = 12$. Find the following:

(a) $\sum x_i$
(b) $\sum x_i y_i$
(c) $\sum(x_i + ay_i)$
(d) $\sum_{i=1}^{3} y_i^2$
(e) $\sum_{i=1}^{i=n} a$
(f) $\sum_k 2(y_k - 3)$
(g) $\sum_{i=1}^{i=5}(x_i - \bar{x})$

10. Find $8!/3!$.

11. Find $\binom{10}{5}$.

(Continued)

(Continued)

12. Use the following table of commuting to determine the number of commuters leaving and entering each zone. Also find the total number of commuters. For each answer, also give the correct notation, assuming y_{ij} denotes the number of commuters who leave origin *i* to go to destination zone *j*.

Destination zone				
Origin zone	1	2	3	4
1	32	25	14	10
2	14	33	19	9
3	15	27	39	20
4	10	12	20	40

13. The following data represent stream link lengths in a river network (given in meters):

    100, 426, 322, 466, 112, 155, 388, 1155, 234, 324, 556, 221, 18, 133, 177, 441.

    Find the mean and standard deviation of the link lengths.
14. For the following annual rainfall data, find the grouped mean and the grouped variance.

Rainfall	Number of Years Observed
< 20″	5
20–29.9″	10
30–39.9″	12
40–49.9″	11
50–59″	3
> 60″	2

    Notice that assumptions must be made about the "midpoints" of the open-ended age groups. In this example, use 15" and 65" as the midpoints of the first and last rainfall groups, respectively.

15. A square grid is placed over a city map. What is the Euclidean distance between two places located at (1,3) and (3,6)?
16. In the example above, what is the Manhattan distance?
17. Use a rose diagram to portray the following angular data:

(Continued)

Direction	Frequency	Direction	Frequency
N	43	S	60
NE	12	SW	70
E	23	W	75
SE	45	NW	65

18. Draw frequency distributions that have (a) positive skewness, (b) negative skewness, (c) low kurtosis and no skewness, and (d) high kurtosis and no skewness.
19. What is the coefficient of variation among the following commuting times: 23, 43, 42, 7, 23, 11, 23, 55?
20. A square grid is placed over a city map. There are residential areas at the coordinates (0,1), (2,3) and (5,6). The respective populations of the three areas are 2500, 2000, and 3000. A centralized facility is being considered for either the point (4,4) or the point (4,5). Which of the two points is the better location for a centralized facility, given that we wish to minimize the total Manhattan distance traveled by the population to the facility? Justify your answer by giving the total Manhattan distance traveled by the population to each of the two possible locations.
21. What do high and low values of kurtosis imply for the shape of a frequency distribution? Draw a diagram to illustrate your answer.
22. Is the following data on incomes positively or negatively skewed? You do not need to calculate skewness, but you should justify your answer.

    Data in thousands: 45, 43, 32, 23, 45, 43, 47, 39, 21, 90, 230.

23. (a) Find the weighted mean center of population, where cities' population and coordinates are given as follows:

City	x	y	Population
A	3.3	4.3	34,000
B	1.1	3.4	6,500
C	5.5	1.2	8,000
D	3.7	2.4	5,000
E	1.1	1.1	1,500

(b) Find the unweighted mean center, and comment on the differences between your two answers.
(c) Find the distances of each city to the weighted mean center of population.
(d) Find the standard distance (weighted) for the five cities.
(e) Find the relative standard distance by assuming that the study area is a rectangle with coordinates of (0,0) in the southwest and (6,6) in the northeast.

(Continued)

(Continued)

(f) Repeat part (e), this time assuming that the coordinates of the rectangle range from (0,0) in the southwest to (8,8) in the northeast.

24. Find the angular mean and the circular variance for the following sample of nine angular measurements: 43°, 45°, 52°, 61°, 75°, 88°, 88°, 279°, and 357°.
25. A public facility is to be located as closely as possible to the mean center of five residential areas. Given the following data:

x-coord.	y-coord.	Population
3	7	40
2	2	10
1	1	50
6	2	30
2	5	20

(a) Find the mean center. What is the aggregate travel distance? Find the aggregate travel distance for each of the four grid points surrounding the mean center. In which direction(s) is the median center likely to be located?
(b) Find the standard distance.
(c) Assuming that the study area is bounded by (0,0) in the southwest and by (6,7) in the northeast, find the relative dispersion.