CHAPTER 2

STRATEGIES FOR UNDERSTANDING THE MEANINGS OF DATA

CHAPTER OVERVIEW

This chapter introduces a set of basic procedures and statistical measures for describing data. Data generally consist of an extensive number of measurements or observations that are too numerous or complicated to be understood through simple observation. Therefore, this chapter introduces several techniques including the construction of tables, graphical displays, and basic statistical computations that provide ways to condense and organize information into a set of descriptive measures and visual devices that enhance the understanding of complex data.

TOPICS

LEARNING OUTCOMES

After studying this chapter, the student will

1. understand how data can be appropriately organized and displayed.
2. understand how to reduce data sets into a few useful, descriptive measures.
3. be able to calculate and interpret measures of central tendency, such as the mean, median, and mode.
4. be able to calculate and interpret measures of dispersion, such as the range, variance, and standard deviation.

2.1 INTRODUCTION

In Chapter 1 we stated that the taking of a measurement and the process of counting yield numbers that contain information. The objective of the person applying the tools of statistics to these numbers is to determine the nature of this information. This task is made much easier if the numbers are organized and summarized. When measurements of a random variable are taken on the entities of a population or sample, the resulting values are made available to the researcher or statistician as a mass of unordered data. Measurements that have not been organized, summarized, or otherwise manipulated are called *raw data*. Unless the number of observations is extremely small, it will be unlikely that these raw data will impart much information until they have been put into some kind of order.

In this chapter we learn several techniques for organizing and summarizing data so that we may more easily determine what information they contain. The ultimate in summarization of data is the calculation of a single number that in some way conveys important information about the data from which it was calculated. Such single numbers that are used to describe data are called *descriptive measures*. After studying this chapter you will be able to compute several descriptive measures for both populations and samples of data.

The purpose of this chapter is to equip you with skills that will enable you to manipulate the information—in the form of numbers—that you encounter as a health sciences professional. The better able you are to manipulate such information, the better understanding you will have of the environment and forces that generate the information.

2.2 THE ORDERED ARRAY

A first step in organizing data is the preparation of an ordered array. An *ordered array* is a listing of the values of a collection (either population or sample) in order of magnitude from the smallest value to the largest value. If the number of measurements to be ordered is of any appreciable size, the use of a computer to prepare the ordered array is highly desirable.

An ordered array enables one to determine quickly the value of the smallest measurement, the value of the largest measurement, and other facts about the arrayed data that might be needed in a hurry. We illustrate the construction of an ordered array with the data discussed in Example 1.4.1.

EXAMPLE 2.2.1

Table 1.4.1 contains a list of the ages of subjects who participated in the study on smoking cessation discussed in Example 1.4.1. As can be seen, this unordered table requires considerable searching for us to ascertain such elementary information as the age of the youngest and oldest subjects.

Solution: Table 2.2.1 presents the data of Table 1.4.1 in the form of an ordered array. By referring to Table 2.2.1 we are able to determine quickly the age of the youngest subject (30) and the age of the oldest subject (82). We also readily note that about one-third of the subjects are 50 years of age or younger.

TABLE 2.2.1 Ordered Array of Ages of Subjects from Table 1.4.1

30	34	35	37	37	38	38	38	38	39	39	40	40	42	42
43	43	43	43	43	43	44	44	44	44	44	44	44	45	45
45	46	46	46	46	46	46	47	47	47	47	47	47	48	48
48	48	48	48	48	49	49	49	49	49	49	49	50	50	50
50	50	50	50	50	51	51	51	51	52	52	52	52	52	52
53	53	53	53	53	53	53	53	53	53	53	53	53	53	53
53	53	54	54	54	54	54	54	54	54	54	54	54	55	55
55	56	56	56	56	56	56	57	57	57	57	57	57	57	58
58	59	59	59	59	59	59	60	60	60	60	61	61	61	61
61	61	61	61	61	61	61	62	62	62	62	62	62	62	63
63	64	64	64	64	64	64	65	65	66	66	66	66	66	66
67	68	68	68	69	69	69	70	71	71	71	71	71	71	71
72	73	75	76	77	78	78	78	82						

■

Computer Analysis If additional computations and organization of a data set have to be done by hand, the work may be facilitated by working from an ordered array. If the data are to be analyzed by a computer, it may be undesirable to prepare an ordered array, unless one is needed for reference purposes or for some other use. A computer does not need its user to first construct an ordered array before entering data for the construction of frequency distributions and the performance of other analyses. However, almost all computer statistical packages and spreadsheet programs contain a routine for sorting data in either an ascending or descending order. See Figure 2.2.1, for example.

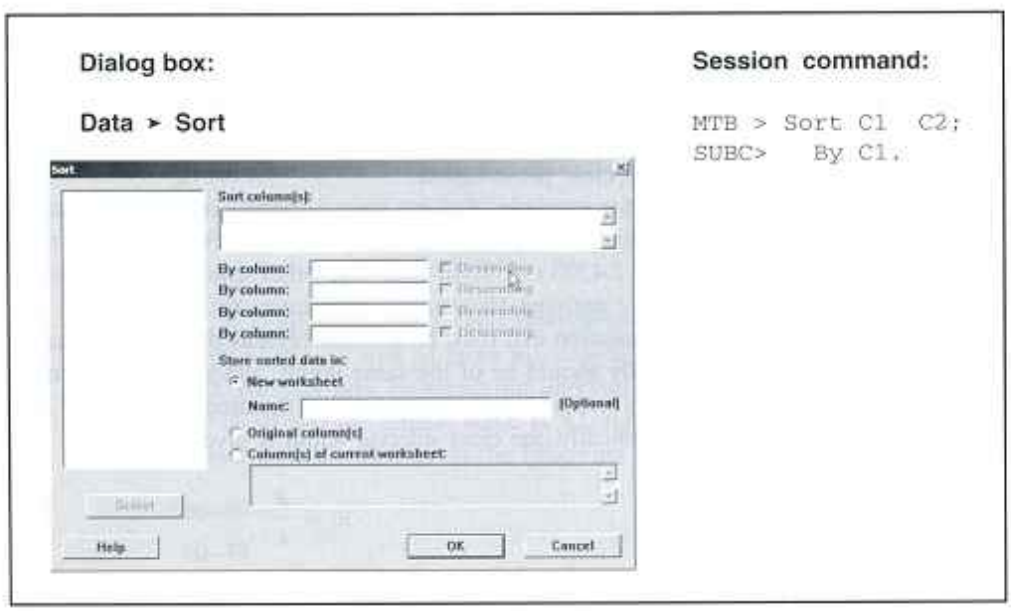

FIGURE 2.2.1 MINITAB dialog box for Example 2.2.1.

2.3 GROUPED DATA: THE FREQUENCY DISTRIBUTION

Although a set of observations can be made more comprehensible and meaningful by means of an ordered array, further useful summarization may be achieved by grouping the data. Before the days of computers one of the main objectives in grouping large data sets was to facilitate the calculation of various descriptive measures such as percentages and averages. Because computers can perform these calculations on large data sets without first grouping the data, the main purpose in grouping data now is summarization. One must bear in mind that data contain information and that summarization is a way of making it easier to determine the nature of this information.

To group a set of observations we select a set of contiguous, nonoverlapping intervals such that each value in the set of observations can be placed in one, and only one, of the intervals. These intervals are usually referred to as *class intervals*.

One of the first considerations when data are to be grouped is how many intervals to include. Too few intervals are undesirable because of the resulting loss of information. On the other hand, if too many intervals are used, the objective of summarization will not be met. The best guide to this, as well as to other decisions to be made in grouping data, is your knowledge of the data. It may be that class intervals have been determined by precedent, as in the case of annual tabulations, when the class intervals of previous years are maintained for comparative purposes. A commonly followed rule of thumb states that there should be no fewer than five intervals and no more than 15. If there are fewer than five intervals, the data have been summarized too much and the information they contain has been lost. If there are more than 15 intervals, the data have not been summarized enough.

Those who need more specific guidance in the matter of deciding how many class intervals to employ may use a formula given by Sturges (1). This formula gives $k = 1 + 3.322(\log_{10} n)$, where k stands for the number of class intervals and n is the number of values in the data set under consideration. The answer obtained by applying *Sturges's rule* should not be regarded as final, but should be considered as a guide only. The number of class intervals specified by the rule should be increased or decreased for convenience and clear presentation.

Suppose, for example, that we have a sample of 275 observations that we want to group. The logarithm to the base 10 of 275 is 2.4393. Applying Sturges's formula gives $k = 1 + 3.322(2.4393) \simeq 9$. In practice, other considerations might cause us to use eight or fewer or perhaps 10 or more class intervals.

Another question that must be decided regards the width of the class intervals. Class intervals generally should be of the same width, although this is sometimes impossible to accomplish. This width may be determined by dividing the range by k, the number of class intervals. Symbolically, the class interval width is given by

$$w = \frac{R}{k} \tag{2.3.1}$$

where R (the range) is the difference between the smallest and the largest observation in the data set. As a rule this procedure yields a width that is inconvenient for use. Again,

we may exercise our good judgment and select a width (usually close to one given by Equation 2.3.1) that is more convenient.

There are other rules of thumb that are helpful in setting up useful class intervals. When the nature of the data makes them appropriate, class interval widths of 5 units, 10 units, and widths that are multiples of 10 tend to make the summarization more comprehensible. When these widths are employed it is generally good practice to have the lower limit of each interval end in a zero or 5. Usually class intervals are ordered from smallest to largest; that is, the first class interval contains the smaller measurements and the last class interval contains the larger measurements. When this is the case, the lower limit of the first class interval should be equal to or smaller than the smallest measurement in the data set, and the upper limit of the last class interval should be equal to or greater than the largest measurement.

Most statistical packages allow users to interactively change the number of class intervals and/or the class widths, so that several visualizations of the data can be obtained quickly. This feature allows users to exercise their judgment in deciding which data display is most appropriate for a given purpose. Let us use the 189 ages shown in Table 1.4.1 and arrayed in Table 2.2.1 to illustrate the construction of a frequency distribution.

EXAMPLE 2.3.1

We wish to know how many class intervals to have in the frequency distribution of the data. We also want to know how wide the intervals should be.

Solution: To get an idea as to the number of class intervals to use, we can apply Sturges's rule to obtain

$$\begin{aligned} k &= 1 + 3.322(\log 189) \\ &= 1 + 3.322(2.2764618) \\ &\approx 9 \end{aligned}$$

Now let us divide the range by 9 to get some idea about the class interval width. We have

$$\frac{R}{k} = \frac{82 - 30}{9} = \frac{52}{9} = 5.778$$

It is apparent that a class interval width of 5 or 10 will be more convenient to use, as well as more meaningful to the reader. Suppose we decide on 10. We may now construct our intervals. Since the smallest value in Table 2.2.1 is 30 and the largest value is 82, we may begin our intervals with 30 and end with 89. This gives the following intervals:

30–39

40–49

50–59

60–69

70–79
80–89

We see that there are six of these intervals, three fewer than the number suggested by Sturges's rule.

It is sometimes useful to refer to the center, called the *midpoint*, of a class interval. The midpoint of a class interval is determined by obtaining the sum of the upper and lower limits of the class interval and dividing by 2. Thus, for example, the midpoint of the class interval 30–39 is found to be $(30 + 39)/2 = 34.5$. ■

When we group data manually, determining the number of values falling into each class interval is merely a matter of looking at the ordered array and counting the number of observations falling in the various intervals. When we do this for our example, we have Table 2.3.1.

A table such as Table 2.3.1 is called a *frequency distribution*. This table shows the way in which the values of the variable are distributed among the specified class intervals. By consulting it, we can determine the frequency of occurrence of values within any one of the class intervals shown.

Relative Frequencies It may be useful at times to know the proportion, rather than the number, of values falling within a particular class interval. We obtain this information by dividing the number of values in the particular class interval by the total number of values. If, in our example, we wish to know the proportion of values between 50 and 59, inclusive, we divide 70 by 189, obtaining .3704. Thus we say that 70 out of 189, or 70/189ths, or .3704, of the values are between 50 and 59. Multiplying .3704 by 100 gives us the percentage of values between 50 and 59. We can say, then, that 37.04 percent of the subjects are between 50 and 59 years of age. We may refer to the proportion of values falling within a class interval as the *relative frequency of occurrence* of values in that interval. In Section 3.2 we shall see that a relative frequency may be interpreted also as the probability of occurrence within the given interval. This probability of occurrence is also called the *experimental probability* or the *empirical probability*.

TABLE 2.3.1 Frequency Distribution of Ages of 189 Subjects Shown in Tables 1.4.1 and 2.2.1

Class Interval	Frequency
30–39	11
40–49	46
50–59	70
60–69	45
70–79	16
80–89	1
Total	189

TABLE 2.3.2 Frequency, Cumulative Frequency, Relative Frequency, and Cumulative Relative Frequency Distributions of the Ages of Subjects Described in Example 1.4.1

Class Interval	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
30–39	11	11	.0582	.0582
40–49	46	57	.2434	.3016
50–59	70	127	.3704	.6720
60–69	45	172	.2381	.9101
70–79	16	188	.0847	.9948
80–89	1	189	.0053	1.0001
Total	189		1.0001	

Note: Frequencies do not add to 1.0000 exactly because of rounding.

In determining the frequency of values falling within two or more class intervals, we obtain the sum of the number of values falling within the class intervals of interest. Similarly, if we want to know the relative frequency of occurrence of values falling within two or more class intervals, we add the respective relative frequencies. We may sum, or *cumulate*, the frequencies and relative frequencies to facilitate obtaining information regarding the frequency or relative frequency of values within two or more contiguous class intervals. Table 2.3.2 shows the data of Table 2.3.1 along with the *cumulative frequencies*, the *relative frequencies*, and *cumulative relative frequencies*.

Suppose that we are interested in the relative frequency of values between 50 and 79. We use the cumulative relative frequency column of Table 2.3.2 and subtract .3016 from .9948, obtaining .6932.

We may use a statistical package to obtain a table similar to that shown in Table 2.3.2. Tables obtained from both MINITAB and SPSS software are shown in Figure 2.3.1.

The Histogram We may display a frequency distribution (or a relative frequency distribution) graphically in the form of a *histogram*, which is a special type of bar graph.

When we construct a histogram the values of the variable under consideration are represented by the horizontal axis, while the vertical axis has as its scale the frequency (or relative frequency if desired) of occurrence. Above each class interval on the horizontal axis a rectangular bar, or cell, as it is sometimes called, is erected so that the height corresponds to the respective frequency when the class intervals are of equal width. The cells of a histogram must be joined and, to accomplish this, we must take into account the true boundaries of the class intervals to prevent gaps from occurring between the cells of our graph.

The level of precision observed in reported data that are measured on a continuous scale indicates some order of rounding. The order of rounding reflects either the reporter's personal preference or the limitations of the measuring instrument employed. When a frequency distribution is constructed from the data, the class interval limits usually reflect the degree of precision of the raw data. This has been done in our illustrative example.

Dialog box:

Stat ➤ Tables ➤ Tally Individual Variables

Type *C2* in **Variables.** Check **Counts, Percents, Cumulative counts,** and **Cumulative percents** in **Display.** Click **OK.**

Session command:

```
MTB > Tally C2;
SUBC>   Counts;
SUBC>   CumCounts;
SUBC>   Percents;
SUBC>   CumPercents;
```

Output:

Tally for Discrete Variables: C2

MINITAB Output

C2	Count	CumCnt	Percent	CumPct
0	11	11	5.82	5.82
1	46	57	24.34	30.16
2	70	127	37.04	67.20
3	45	172	23.81	91.01
4	16	188	8.47	99.47
5	1	189	0.53	100.00
N=	189			

SPSS Output

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	30-39	11	5.8	5.8	5.8
	40-49	46	24.3	24.3	30.2
	50-59	70	37.0	37.0	67.2
	60-69	45	23.8	23.8	91.0
	70-79	16	8.5	8.5	99.5
	80-89	1	.5	.5	100.0
	Total	189	100.0	100.0	

FIGURE 2.3.1 Frequency, cumulative frequencies, percent, and cumulative percent distribution of the ages of subjects described in Example 1.4.1 as constructed by MINITAB and SPSS.

We know, however, that some of the values falling in the second class interval, for example, when measured precisely, would probably be a little less than 40 and some would be a little greater than 49. Considering the underlying continuity of our variable, and assuming that the data were rounded to the nearest whole number, we find it convenient to think of 39.5 and 49.5 as the true limits of this second interval. The true limits for each of the class intervals, then, we take to be as shown in Table 2.3.3.

If we construct a graph using these class limits as the base of our rectangles, no gaps will result, and we will have the histogram shown in Figure 2.3.2. We used MINITAB to construct this histogram, as shown in Figure 2.3.3.

We refer to the space enclosed by the boundaries of the histogram as the *area* of the histogram. Each observation is allotted one unit of this area. Since we have 189 observations, the histogram consists of a total of 189 units. Each cell contains a certain proportion of the total area, depending on the frequency. The second cell, for example, contains 46/189 of the area. This, as we have learned, is the relative frequency of occurrence of values between 39.5 and 49.5. From this we see that subareas of the histogram defined by the cells correspond to the frequencies of occurrence of values between the horizontal scale boundaries of the areas. The ratio of a particular subarea to the total area of the histogram is equal to the relative frequency of occurrence of values between the corresponding points on the horizontal axis.