DESCRIPTIVE STATISTICS

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 meas-

urement, 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 smok-

ing 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.

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 with-

out 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 inter-

vals 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 + 3.3221log 10 n2, 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 + 3.32212.43932 M 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 =

R

k

(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 com-

prehensible. 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 measure-

ment 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 dis-

play 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

k = 1 + 3.3221log 1892

= 1 + 3.32212.27646182

L 9

Now let us divide the range by 9 to get some idea about the class

interval width. We have

R

82 - 30

52

=

=

= 5.778

k

9

9

It is apparent that a class interval width of 5 or 10 will be more con-

venient 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 130 + 392>2 = 34.5.

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 infor-

mation by dividing the number of values in the particular class interval by the total num-

ber 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 inter-

val. 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.

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 fre-

quencies, 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 hori-

zontal 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 fre-

quency 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.

The Frequency Polygon A frequency distribution can be portrayed graphi-

cally in yet another way by means of a frequency polygon, which is a special kind of

line graph. To draw a frequency polygon we first place a dot above the midpoint of each

class interval represented on the horizontal axis of a graph like the one shown in Figure

2.3.2. The height of a given dot above the horizontal axis corresponds to the frequency

of the relevant class interval. Connecting the dots by straight lines produces the frequency

polygon. Figure 2.3.4 is the frequency polygon for the age data in Table 2.2.1.

Note that the polygon is brought down to the horizontal axis at the ends at points

that would be the midpoints if there were an additional cell at each end of the corre-

sponding histogram. This allows for the total area to be enclosed. The total area under

the frequency polygon is equal to the area under the histogram. Figure 2.3.5 shows the

frequency polygon of Figure 2.3.4 superimposed on the histogram of Figure 2.3.2. This

figure allows you to see, for the same set of data, the relationship between the two

graphic forms.

Stem-and-Leaf Displays Another graphical device that is useful for represent-

ing quantitative data sets is the stem-and-leaf display. A stem-and-leaf display bears a

strong resemblance to a histogram and serves the same purpose. A properly constructed

stem-and-leaf display, like a histogram, provides information regarding the range of the

data set, shows the location of the highest concentration of measurements, and reveals the

presence or absence of symmetry. An advantage of the stem-and-leaf display over the his-

togram is the fact that it preserves the information contained in the individual measure-

ments. Such information is lost when measurements are assigned to the class intervals of

a histogram. As will become apparent, another advantage of stem-and-leaf displays is the

fact that they can be constructed during the tallying process, so the intermediate step of

preparing an ordered array is eliminated.

To construct a stem-and-leaf display we partition each measurement into two parts.

The first part is called the stem, and the second part is called the leaf. The stem consists

of one or more of the initial digits of the measurement, and the leaf is composed of one

or more of the remaining digits. All partitioned numbers are shown together in a single

display; the stems form an ordered column with the smallest stem at the top and the largest

at the bottom. We include in the stem column all stems within the range of the data even

when a measurement with that stem is not in the data set. The rows of the display con-

tain the leaves, ordered and listed to the right of their respective stems. When leaves con-

sist of more than one digit, all digits after the first may be deleted. Decimals when pres-

ent in the original data are omitted in the stem-and-leaf display. The stems are separated

from their leaves by a vertical line. Thus we see that a stem-and-leaf display is also an

ordered array of the data.

Stem-and-leaf displays are most effective with relatively small data sets. As a rule

they are not suitable for use in annual reports or other communications aimed at the gen-

eral public. They are primarily of value in helping researchers and decision makers under-

stand the nature of their data. Histograms are more appropriate for externally circulated

publications. The following example illustrates the construction of a stem-and-leaf

display.