Chapter 2

Summarizing Data Graphically

A statistical data set consists of a collection of values on one or more variables. The variables can be either numerical or categorical. Numerical variables are further classified as discrete or continuous. These distinctions determine the statistical approaches that are appropriate for summarizing the data. Examples of data include

- • crime rates for large cities across the United States;
- • body temperatures for a randomly chosen sample of adults;
- • the numbers of errors made by cashiers on an 8-hour shift;
- • the gender of individuals purchasing tickets to a concert; and
- • occupations of a sample of fathers and their sons.

One approach to organizing data is through a chart or graph. The type of chart you use depends in part on the way the data are measured — in categories (e.g., occupations) or on a numerical scale (e.g., number of errors). This chapter demonstrates how to examine different types of data through frequency distributions and graphical representations. Section 2.1 describes methods for summarizing categorical data, while Section 2.2 pertains to discrete and continuous numerical variables.

2.1 SUMMARIZING CATEGORICAL DATA

Categorical variables are those that have qualitatively distinct categories as values. For example, gender is a categorical variable with categories "male" and "female." Information on the coding and labeling of categorical data is given in Chapter 1.

Frequencies

One way to display data is in a frequency distribution, which lists the values of a variable (e.g., for the variable occupation: professional, manager, salesperson, etc.) and the corresponding numbers and percentages of participants for each value.

Let us begin by creating a simple frequency distribution of occupations using the "socmob.sav" SPSS data file on the website accompanying this manual. Follow along by using SPSS to open the data file on your computer (using the procedure given in Chapter 1). This data set was used in a study of the effects of family disruption on social mobility. The study collected data on fathers' occupations, their sons' occupations, family structure (intact/not intact), and race.

Notice that the data view lists numbers as the values for all of the variables, even though the variable is a categorical variable. The use of numbers to represent categories was described in Chapter 1. To see the categories each of the values represent, you can examine the contents of the data file (variable labels, variable type, and value labels) by clicking on **Utilities** on the menu bar and clicking on **Variables** from the pull-down menu. You can also click on the **value labels** button on the toolbar, as displayed in Figure 2.1. This will display the value labels (e.g., laborer, manager, professional) in the data editor.

To create a frequency distribution of the father's occupation variable:

1. Click on **Analyze** from the menu bar.
2. Click on **Descriptive Statistics** from the pull-down menu.
3. Click on **Frequencies** from the second pull-down menu to open the Frequencies dialog box (see Fig. 2.2).

Figure 2.1 Toolbar with value labels button activated

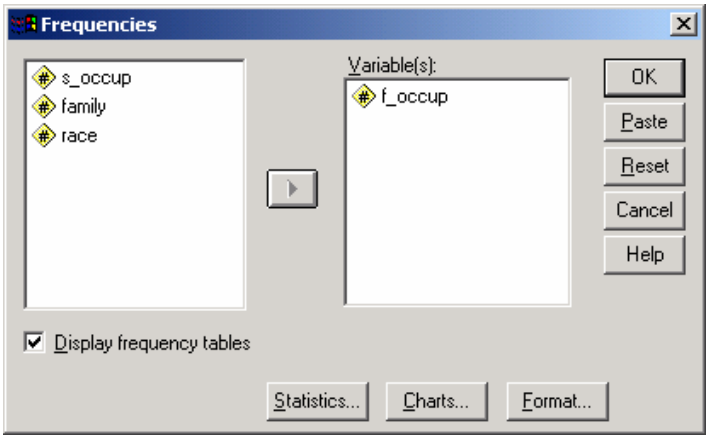

Figure 2.2 Frequencies Dialog Box

4. Click on the label/name of the variable you wish to examine ("f_occup") in the left-hand box.
5. Click on the **right arrow button** to move the variable name into the Variable(s) box.
6. Click on **OK**.

The frequency distribution produced by SPSS is shown in Figure 2.3. This figure shows the content of the output — that which is in the right-hand frame of your Output Viewer.

The "Statistics" table in the output indicates the number of valid and missing values for this variable. There are 1156 valid cases and no missing values. The "father's occupation" table displays the frequency distribution. The occupational categories appear in the left-hand column of this table. The "Frequency" column contains the exact number of cases (e.g., number of fathers) for each of the categories. For example, there are 476 fathers who are laborers and 136 fathers who are professionals.

The numbers in the "Percent" column represent the percentage of the total number of cases that are in each occupational category. These are obtained by dividing each frequency by the total number of cases and multiplying by 100. For example, 11.8% of the sample is comprised of professionals (136/1156 × 100).

Statistics

father's occupation

N	Valid	1156
	Missing	0

father's occupation

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	laborer	476	41.2	41.2	41.2
	craftsperson	204	17.6	17.6	58.8
	salesperson	272	23.5	23.5	82.4
	manager	68	5.9	5.9	88.2
	professional	136	11.8	11.8	100.0
	Total	1156	100.0	100.0	

Figure 2.3 Frequency Distribution of Father's Occupation Variable

The "Valid Percent" column takes into account missing values. In this case, there are no missing values, so the "Percent" and "Valid Percent" columns are the same. The "Cumulative Percent" is a cumulative percentage of the cases for the category and all categories listed before it in the table. For example, 82.4% of all cases in the sample include laborers, craftsmen, and salesmen (41.2% + 17.6% + 23.5%, within rounding error). The cumulative percentages are not meaningful unless the scale of the variable has at least ordinal properties. Ordinal means that the values of the variable are ordered. Numerical variables have ordinal properties, as do ordinal categorical variables (e.g., a variable measuring size, with values equal to small, medium, large, and extra large). The father's occupation variable does not have ordinal properties. That is, being a salesman is not "higher" or "lower" in the list of occupations than is a laborer. The occupations could have been listed in another order without affecting the interpretation of the data.

Frequencies with Missing Data

In this data file, there are no missing cases. Suppose, however, that the families with identification numbers 10128, 10129, 10180, 10343, 10350, 10370, 10434, 10435, 10500, and 10501 were missing information on father's occupation. The frequency distribution for this altered data set would appear as in Figure 2.4. Note that there is an additional row in this distribution chart — the Missing row — which indicates that there are 10 cases for which father's occupation was not

father's occupation

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	laborer	473	40.9	41.3	41.3
	craftsperson	200	17.3	17.5	58.7
	salesperson	269	23.3	23.5	82.2
	manager	68	5.9	5.9	88.1
	professional	136	11.8	11.9	100.0
	Total	1146	99.1	100.0	
Missing	System	10	.9		
Total		1156	100.0		

Figure 2.4 Frequency Distribution for Father's Occupation with Missing Cases

coded. Note also that the Percent and Valid Percent columns now indicate different figures, because of the difference in the denominator used to compute the figures. In the case of laborers, for instance, Percent is computed as $473/1156 \times 100$, and Valid Percent is computed as $473/1146 \times 100$.

Bar Charts

A bar chart is also useful for examining categorical data. In a bar chart, the height of each bar represents the frequency of occurrence for each category of the variable. Let us create a bar chart for the occupation data using an option within the Frequencies procedure.

From the Frequencies dialog box (see steps 1–3 of the Frequencies section):

1. Click on **Charts** to open the Frequencies: Charts dialog box (see Fig. 2.5).
2. Click on **Bar charts** in the Chart Type box.
3. Choose the type of values you want to chart — frequencies or percentages — in the Chart Values box. For this example, we have selected frequencies.
4. Click on **Continue**.
5. Click on **OK** to run the chart procedure.

A bar chart like that in Figure 2.6 should appear in your SPSS Viewer.

The information displayed in this chart is a graphical version of that shown in the frequency distribution in Figure 2.3. The occupation group with the greatest number of people is laborer; the occupation group with the fewest is manager. There are 204 craftspeople; this is determined by looking at the vertical (frequency) axis.

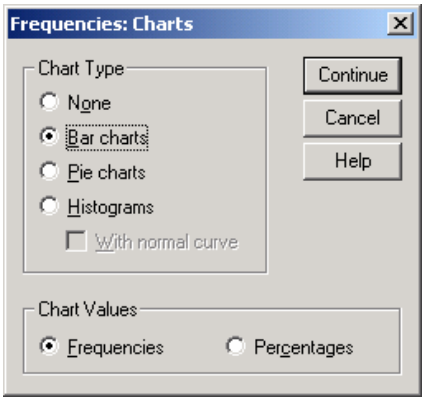

Figure 2.5 Frequencies: Charts Dialog Box

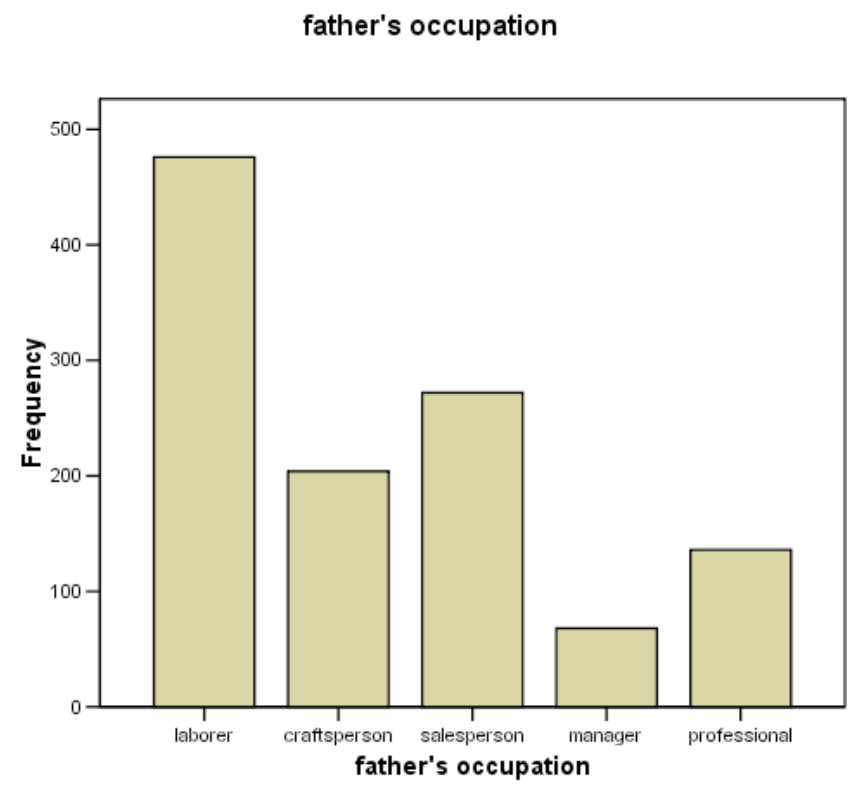

Figure 2.6 Bar Chart of Father's Occupation Variable

2.2 SUMMARIZING NUMERICAL DATA

There are two types of numerical variables — discrete and categorical. The values for discrete variables are counting numbers. For example, an American football game is won by one, two, or three points, not a quantity in between. Continuous variables, on the other hand, do not have such indivisible units. Body temperature, for instance, can be measured to the nearest degree, half-degree, quarter-degree, and so on. For practical purposes in SPSS, there is no difference in summarizing these two types of numerical data.

We shall use the data in the "football.sav"[1] data file to illustrate graphical summaries of numerical data. This file contains data on 250 National Football League (NFL) games from a recent season. One of the variables in this file is "winby," representing the number of points by which the winning team was victorious. We can create a frequency distribution of the winby variable using the same procedures as outlined in Section 2.1. The frequency distribution is in Figure 2.7. We see that 32 games were won by 3 points. This is 12.8% of the 250 games. The cumulative percent column is meaningful with numerical data, and we see that 35.2% of the games were won by 6 or fewer points (or, by "less than a touchdown").

We use histograms instead of bar charts to graphically display numerical data. There are several ways to obtain a histogram in SPSS. One such procedure is identical to the one used in Section 2.1 except you click on the histogram option in the Frequencies: Charts dialog box (Fig. 2.5). An alternative method is to use the Explore procedure, as illustrated below:

1. Click on **Analyze** on the menu bar.
2. Click on **Descriptive Statistics** from the pull-down menu.
3. Click on **Explore** from the pull-down menu. This opens the Explore dialog box as shown in Figure 2.8.
4. Click on the name of the variable ("winby") and click on the **top right arrow button** to move it to the Dependent List box. (In this example, there is no independent, or Factor, variable.)
5. Click on **Plots** in the Display box. This will suppress all statistics in the output. (If you also want SPSS to provide summary statistics, click on **Both.)**
6. Click on the **Plots button** to open the Explore: Plots dialog box.

[1] Appreciation for this and several other data sets used in this manual is expressed to the *Journal of Statistics Education*, http://www.amstat.org/publications/jse/, an international resource for teaching and learning of statistics.

Statistics

WINBY

N	Valid	250
	Missing	0

WINBY

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	9	3.6	3.6	3.6
	2	8	3.2	3.2	6.8
	3	32	12.8	12.8	19.6
	4	17	6.8	6.8	26.4
	5	9	3.6	3.6	30.0
	6	13	5.2	5.2	35.2
	7	23	9.2	9.2	44.4
	8	6	2.4	2.4	46.8
	9	5	2.0	2.0	48.8
	10	20	8.0	8.0	56.8
	11	9	3.6	3.6	60.4
	12	1	.4	.4	60.8
	13	3	1.2	1.2	62.0
	14	17	6.8	6.8	68.8
	15	10	4.0	4.0	72.8
	16	8	3.2	3.2	76.0
	17	4	1.6	1.6	77.6
	18	5	2.0	2.0	79.6
	19	6	2.4	2.4	82.0
	21	4	1.6	1.6	83.6
	22	2	.8	.8	84.4
	23	1	.4	.4	84.8
	24	6	2.4	2.4	87.2
	25	8	3.2	3.2	90.4
	26	3	1.2	1.2	91.6
	27	5	2.0	2.0	93.6
	28	5	2.0	2.0	95.6
	31	3	1.2	1.2	96.8
	32	2	.8	.8	97.6
	34	1	.4	.4	98.0
	35	1	.4	.4	98.4
	36	2	.8	.8	99.2
	38	1	.4	.4	99.6
	43	1	.4	.4	100.0
	Total	250	100.0	100.0	

Figure 2.7 Frequency Distribution of Points by Which Football Games Were Won

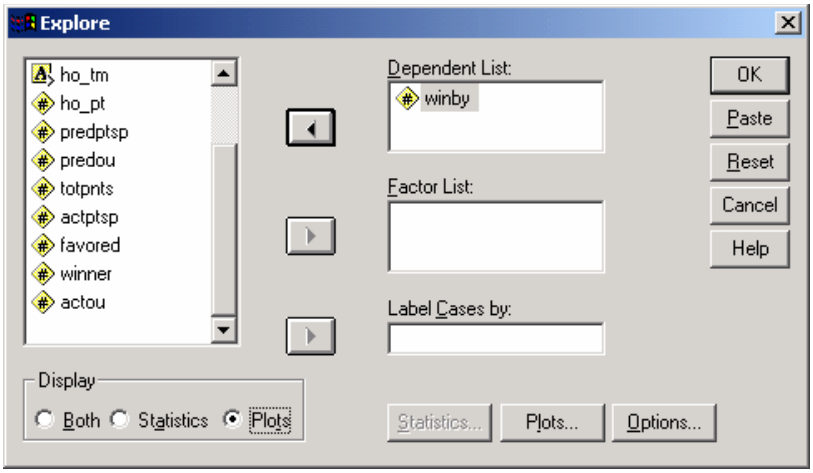

Figure 2.8 Explore Dialog Box

7. Click on **Histogram** in the Descriptive box. (In this example, we are only interested in the histogram, so we also click **None** instead of **Factor levels together** under boxplots and click off the **Stem-and-leaf** option under Descriptive.)
8. Click on **Continue**.
9. Click on **OK** to run the procedure.

The SPSS Viewer will open with the results of the procedure. They are contained in Figure 2.9.

Changing Intervals

The "winby" variable has a range of 42 points (1–43), and the histogram with 22 intervals (selected by SPSS) adequately conveys the nature of the data. It is possible, however, to edit the histogram to change the number of intervals displayed (or the interval width). For example, to change the number of intervals on the x-axis in the above histogram to 10, follow the steps below:

1. In the SPSS Viewer, double click on the histogram to open the SPSS Chart Editor.
2. Click on **Edit** from the menu bar.
3. Click on **Select X Axis** from the pull-down menu to open the Properties dialog box (Fig. 2.10).
4. Click on the **Histogram Options** tab.

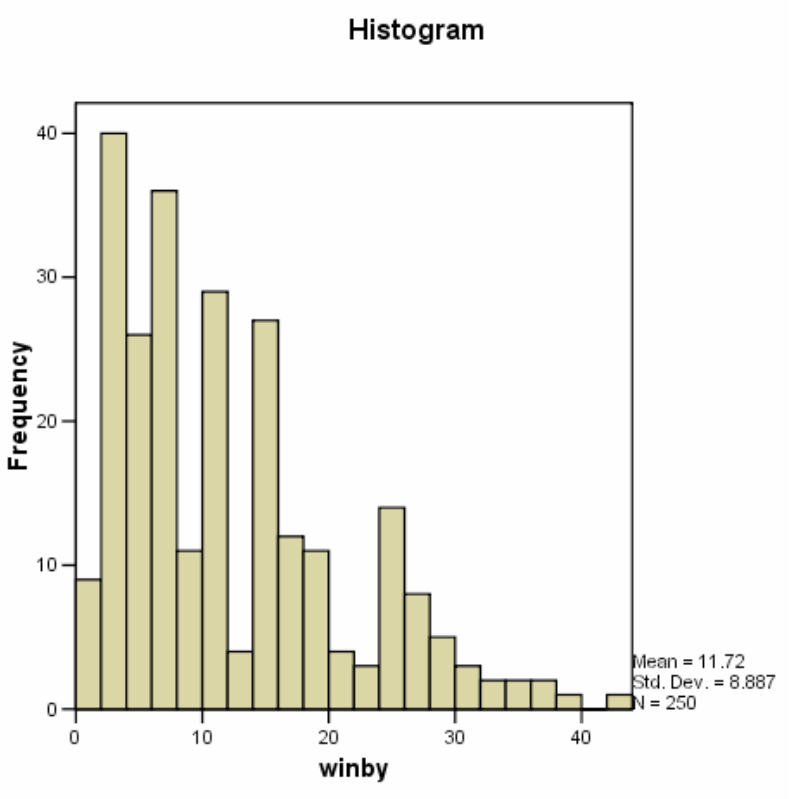

Figure 2.9 Output of Explore Procedure Creating Histogram of Number of Points by Which Football Games Were Won

5. In the Bin Sizes section, select **Custom**.
6. Change the number of intervals to 10.
7. Click on **Apply** to redraw the graph.
8. The cancel button changes to a close button. Click on **Close** to close the Properties dialog box.
9. Click the **X** in the right corner of the Chart Editor window to close it. Note that the histogram now has 10 intervals, but it is still positively skewed in shape (Fig. 2.11).

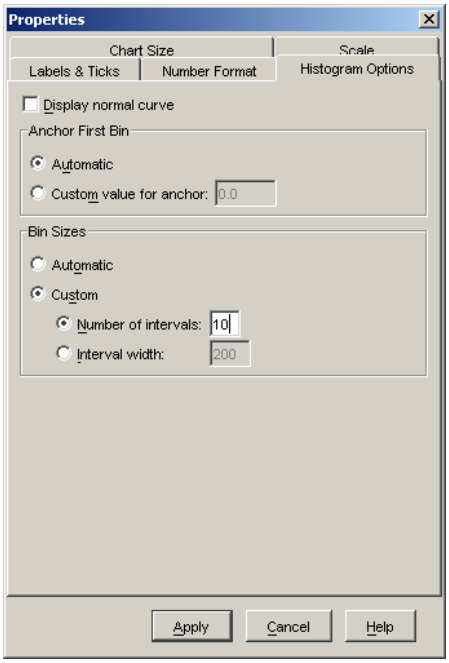

Figure 2.10 Histogram Options Tab of Properties Dialog Box

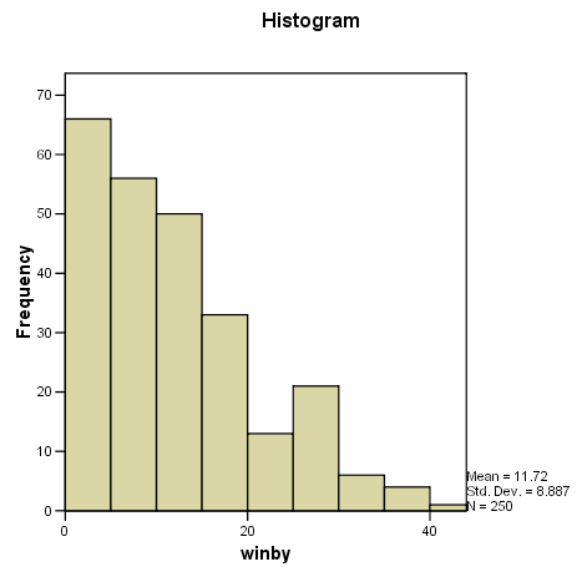

Figure 2.11 Histogram of Points by Which Games Were Won (10 Intervals)

Stem-and-Leaf Plot

Another way to display numerical data is with a stem-and-leaf plot. This display divides data into intervals. The value for each observation is split into two parts — the stem and the leaf. The stem represents the interval and the leaf represents the last digit(s) of the actual data points.

To direct SPSS to produce a stem-and-leaf plot of "winby," follow the steps 1–6 given in Section 2.2, plus:

1. Click on **Stem-and-leaf** in the Descriptive box of the Explore: Plots dialog box.
2. Click on **Continue**.
3. Click on **OK** to run the procedure.

The stem-and-leaf plot will appear in the SPSS Viewer, as shown in Figure 2.12.

```
WINBY Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     9.00        0 . 111111111
    40.00        0 . 2222222233333333333333333333333333333333
    26.00        0 . 44444444444444444555555555
    36.00        0 . 666666666666777777777777777777777777
    11.00        0 . 88888899999
    29.00        1 . 00000000000000000000111111111
     4.00        1 . 2333
    27.00        1 . 444444444444444445555555555
    12.00        1 . 666666667777
    11.00        1 . 88888999999
     4.00        2 . 1111
     3.00        2 . 223
    14.00        2 . 44444455555555
     8.00        2 . 66677777
     5.00        2 . 88888
     3.00        3 . 111
     2.00        3 . 22
     6.00 Extremes    (>=34)

 Stem width:        10
 Each leaf:       1 case(s)
```

Figure 2.12 Stem-and-Leaf Display of Points by Which Games Were Won

The stem-and-leaf plot is similar to a histogram. In this case, the stem represents the 10's digit, and the leaf represents the 1's digit. The points displayed on the plot range from 1 [(0 × 10) + 1] to 32 [(3 × 10) + 2] points. The frequency column gives the number of observations in each interval. Notice that there are five sections of stem with each of the values 0, 1, and 2. The frequency column shows that 9 games were won by 1 point, 40 games were won by 2 or 3 points, and so on. Six of the games were won by 34 or more points, and are considered "extremes."

Chapter Exercises

2.1 Open the "bodytemp.sav" data file containing gender, body temperature, and pulse rate of 130 adults. These data were collected in part to examine whether normal body temperature is 98.6° Fahrenheit. Use the data set to conduct to following analyses and answer the following questions:

- **a.** Create a frequency distribution of body temperature. How many adults in the sample have a normal body temperature of 98.6° Fahrenheit?
- **b.** What percent of adults have a temperature less than 98.6° Fahrenheit?
- **c.** What is the lowest and highest temperature?
- **d.** Create a histogram of the temperature; adjust the graph so that there are 12 intervals.

2.2 Open the "fire.sav" data file, which contains demographic and performance data on 28 firefighter candidates, and use SPSS to answer the following:

- **a.** How many male firefighter applicants are there in the sample? What percentage of the total number of applicants is this?
- **b.** What percent of applicants are members of a minority group?
- **c.** For females only, what percentage of females had a time of less than 18 seconds on stair climb task? (Hint: you need to use the Select If command detailed in Chapter 1.)
- **d.** Repeat part c. for males only.
- **e.** Create a stem-and-leaf plot for the written test score.
- **f.** How many applicants had a score between 85 and 89 on the written test?

2.3 Open the 'titanic.sav" data file, which contains data on 2201 passengers on the Titanic. The variables are: gender, age category, class, and survival. Use SPSS to conduct the following analyses:

a. Create a bar chart of the class variable. Which class level had the most passengers?

b. Were there more first-class or second-class passengers?

c. How many passengers survived?