Chapter 14

Quantitative Analysis: Descriptive Statistics

Numeric data collected in a research project can be analyzed quantitatively using statistical tools in two different ways. **Descriptive analysis** refers to statistically describing, aggregating, and presenting the constructs of interest or associations between these constructs. **Inferential analysis** refers to the statistical testing of hypotheses (theory testing). In this chapter, we will examine statistical techniques used for descriptive analysis, and the next chapter will examine statistical techniques for inferential analysis. Much of today's quantitative data analysis is conducted using software programs such as SPSS or SAS. Readers are advised to familiarize themselves with one of these programs for understanding the concepts described in this chapter.

Data Preparation

In research projects, data may be collected from a variety of sources: mail-in surveys, interviews, pretest or posttest experimental data, observational data, and so forth. This data must be converted into a machine-readable, numeric format, such as in a spreadsheet or a text file, so that they can be analyzed by computer programs like SPSS or SAS. Data preparation usually follows the following steps.

Data coding. Coding is the process of converting data into numeric format. A codebook should be created to guide the coding process. A **codebook** is a comprehensive document containing detailed description of each variable in a research study, items or measures for that variable, the format of each item (numeric, text, etc.), the response scale for each item (i.e., whether it is measured on a nominal, ordinal, interval, or ratio scale; whether such scale is a five-point, seven-point, or some other type of scale), and how to code each value into a numeric format. For instance, if we have a measurement item on a seven-point Likert scale with anchors ranging from "strongly disagree" to "strongly agree", we may code that item as 1 for strongly disagree, 4 for neutral, and 7 for strongly agree, with the intermediate anchors in between. Nominal data such as industry type can be coded in numeric form using a coding scheme such as: 1 for manufacturing, 2 for retailing, 3 for financial, 4 for healthcare, and so forth (of course, nominal data cannot be analyzed statistically). Ratio scale data such as age, income, or test scores can be coded as entered by the respondent. Sometimes, data may need to be aggregated into a different form than the format used for data collection. For instance, for measuring a construct such as "benefits of computers," if a survey provided respondents with a checklist of

benefits that they could select from (i.e., they could choose as many of those benefits as they wanted), then the total number of checked items can be used as an aggregate measure of benefits. Note that many other forms of data, such as interview transcripts, cannot be converted into a numeric format for statistical analysis. Coding is especially important for large complex studies involving many variables and measurement items, where the coding process is conducted by different people, to help the coding team code data in a consistent manner, and also to help others understand and interpret the coded data.

Data entry. Coded data can be entered into a spreadsheet, database, text file, or directly into a statistical program like SPSS. Most statistical programs provide a data editor for entering data. However, these programs store data in their own native format (e.g., SPSS stores data as .sav files), which makes it difficult to share that data with other statistical programs. Hence, it is often better to enter data into a spreadsheet or database, where they can be reorganized as needed, shared across programs, and subsets of data can be extracted for analysis. Smaller data sets with less than 65,000 observations and 256 items can be stored in a spreadsheet such as Microsoft Excel, while larger dataset with millions of observations will require a database. Each observation can be entered as one row in the spreadsheet and each measurement item can be represented as one column. The entered data should be frequently checked for accuracy, via occasional spot checks on a set of items or observations, during and after entry. Furthermore, while entering data, the coder should watch out for obvious evidence of bad data, such as the respondent selecting the "strongly agree" response to all items irrespective of content, including reverse-coded items. If so, such data can be entered but should be excluded from subsequent analysis.

Missing values. Missing data is an inevitable part of any empirical data set. Respondents may not answer certain questions if they are ambiguously worded or too sensitive. Such problems should be detected earlier during pretests and corrected before the main data collection process begins. During data entry, some statistical programs automatically treat blank entries as missing values, while others require a specific numeric value such as -1 or 999 to be entered to denote a missing value. During data analysis, the default mode of handling missing values in most software programs is to simply drop the entire observation containing even a single missing value, in a technique called *listwise deletion*. Such deletion can significantly shrink the sample size and make it extremely difficult to detect small effects. Hence, some software programs allow the option of replacing missing values with an estimated value via a process called *imputation*. For instance, if the missing value is one item in a multi-item scale, the imputed value may be the average of the respondent's responses to remaining items on that scale. If the missing value belongs to a single-item scale, many researchers use the average of other respondent's responses to that item as the imputed value. Such imputation may be biased if the missing value is of a systematic nature rather than a random nature. Two methods that can produce relatively unbiased estimates for imputation are the maximum likelihood procedures and multiple imputation methods, both of which are supported in popular software programs such as SPSS and SAS.

Data transformation. Sometimes, it is necessary to transform data values before they can be meaningfully interpreted. For instance, reverse coded items, where items convey the opposite meaning of that of their underlying construct, should be reversed (e.g., in a 1-7 interval scale, 8 minus the observed value will reverse the value) before they can be compared or combined with items that are not reverse coded. Other kinds of transformations may include creating scale measures by adding individual scale items, creating a weighted index from a set

of observed measures, and collapsing multiple values into fewer categories (e.g., collapsing incomes into income ranges).

Univariate Analysis

Univariate analysis, or analysis of a single variable, refers to a set of statistical techniques that can describe the general properties of one variable. Univariate statistics include: (1) frequency distribution, (2) central tendency, and (3) dispersion. The **frequency distribution** of a variable is a summary of the frequency (or percentages) of individual values or ranges of values for that variable. For instance, we can measure how many times a sample of respondents attend religious services (as a measure of their "religiosity") using a categorical scale: never, once per year, several times per year, about once a month, several times per month, several times per week, and an optional category for "did not answer." If we count the number (or percentage) of observations within each category (except "did not answer" which is really a missing value rather than a category), and display it in the form of a table as shown in Figure 14.1, what we have is a frequency distribution. This distribution can also be depicted in the form of a bar chart, as shown on the right panel of Figure 14.1, with the horizontal axis representing each category of that variable and the vertical axis representing the frequency or percentage of observations within each category.

	Frequency	Percent
Never	85	29.9%
Once per year	23	8.1%
Several times per year	36	12.7%
About once per month	44	15.5%
Several times per month	56	19.7%
Several times per week	21	7.4%
Did not answer	19	6.7%
TOTAL	284	100.0%

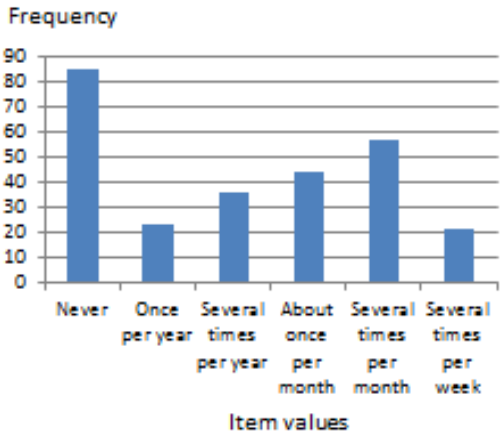

Figure 14.1. Frequency distribution of religiosity

With very large samples where observations are independent and random, the frequency distribution tends to follow a plot that looked like a bell-shaped curve (a smoothed bar chart of the frequency distribution) similar to that shown in Figure 14.2, where most observations are clustered toward the center of the range of values, and fewer and fewer observations toward the extreme ends of the range. Such a curve is called a *normal distribution*.

Central tendency is an estimate of the center of a distribution of values. There are three major estimates of central tendency: mean, median, and mode. The **arithmetic mean** (often simply called the "mean") is the simple average of all values in a given distribution. Consider a set of eight test scores: 15, 22, 21, 18, 36, 15, 25, 15. The arithmetic mean of these values is (15 + 20 + 21 + 20 + 36 + 15 + 25 + 15)/8 = 20.875. Other types of means include *geometric mean* (n^{th} root of the product of n numbers in a distribution) and *harmonic mean* (the reciprocal of the arithmetic means of the reciprocal of each value in a distribution), but these means are not very popular for statistical analysis of social research data.

The second measure of central tendency, the **median**, is the middle value within a range of values in a distribution. This is computed by sorting all values in a distribution in increasing order and selecting the middle value. In case there are two middle values (if there is an even number of values in a distribution), the average of the two middle values represent the median. In the above example, the sorted values are: 15, 15, 15, 18, 22, 21, 25, 36. The two middle values are 18 and 22, and hence the median is (18 + 22)/2 = 20.

Lastly, the **mode** is the most frequently occurring value in a distribution of values. In the previous example, the most frequently occurring value is 15, which is the mode of the above set of test scores. Note that any value that is estimated from a sample, such as mean, median, mode, or any of the later estimates are called a **statistic**.

Dispersion refers to the way values are *spread* around the central tendency, for example, how tightly or how widely are the values clustered around the mean. Two common measures of dispersion are the range and standard deviation. The **range** is the difference between the highest and lowest values in a distribution. The range in our previous example is 36-15 = 21.

The range is particularly sensitive to the presence of outliers. For instance, if the highest value in the above distribution was 85 and the other vales remained the same, the range would be 85-15 = 70. **Standard deviation**, the second measure of dispersion, corrects for such outliers by using a formula that takes into account how close or how far each value from the distribution mean:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n - 1}}$$

where σ is the standard deviation, x_i is the i^{th} observation (or value), μ is the arithmetic mean, n is the total number of observations, and Σ means summation across all observations. The square of the standard deviation is called the **variance** of a distribution. In a normally distributed frequency distribution, it is seen that 68% of the observations lie within one standard deviation of the mean ($\mu \pm 1\ \sigma$), 95% of the observations lie within two standard deviations ($\mu \pm 2\ \sigma$), and 99.7% of the observations lie within three standard deviations ($\mu \pm 3\ \sigma$), as shown in Figure 14.2.

Bivariate Analysis

Bivariate analysis examines how two variables are related to each other. The most common bivariate statistic is the **bivariate correlation** (often, simply called "correlation"), which is a number between -1 and +1 denoting the strength of the relationship between two variables. Let's say that we wish to study how age is related to self-esteem in a sample of 20 respondents, i.e., as age increases, does self-esteem increase, decrease, or remains unchanged. If self-esteem increases, then we have a positive correlation between the two variables, if self-esteem decreases, we have a negative correlation, and if it remains the same, we have a zero correlation. To calculate the value of this correlation, consider the hypothetical dataset shown in Table 14.1.