

# Chapter 5

## Descriptive Statistics

### Learning Objectives

After reading this chapter, you should understand:

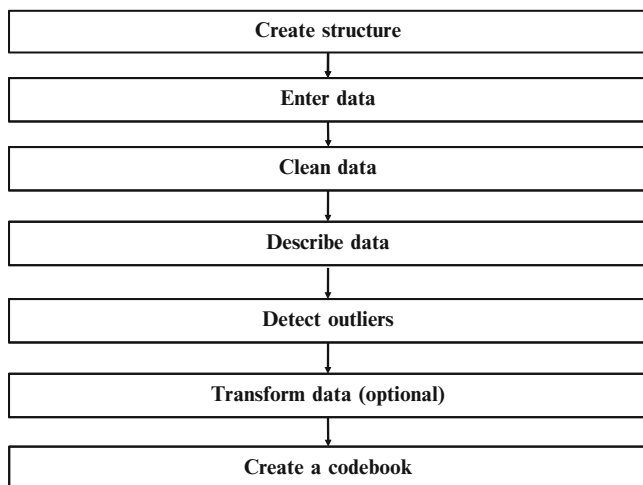
- The workflow involved in a market research study.
- Univariate and bivariate descriptive graphs and statistics.
- How to deal with missing values.
- How to transform data (z-transformation, log transformation, creating dummies, aggregating variables).
- How to identify and deal with outliers.
- What a codebook is.
- The basics of using IBM SPSS Statistics.

**Keywords** Bar chart · Codebook · Correlation · Covariance · Cross tabs · Dummies · Frequency table · Histogram · Line chart · Log transformation · Mean · Median · Mode · Outliers · Pie chart · Scatter plots · SPSS · Standard deviation · Variance · Workflow · z-transformation

This chapter has two purposes: first, we discuss how to keep track of, enter, clean, describe, and transform data. We call these steps the *workflow* of data. Second, we discuss how we can describe data using a software package called IBM SPSS Statistics (abbreviated as SPSS).

### The Workflow of Data

Market research projects involving data become more efficient and effective if a proper *workflow* is in place. A workflow is a strategy to keep track of, enter, clean, describe, and transform data. These data may have been collected through surveys or may be secondary data. Haphazardly entering, cleaning, and analyzing bits of data is not a good strategy, since it increases one's likelihood of making mistakes and makes it hard to replicate results. Moreover, without a good workflow of data, it becomes hard to document the research process and cooperate on projects. For example, how can you outsource the data analysis, if you cannot indicate what the data are about or what specific values mean? Finally, a lack of good workflow



**Fig. 5.1** The workflow of data

increases one's risk of having to duplicate work or even of losing all of your data due to accidents. In Fig. 5.1, we show the steps necessary to create and describe a dataset after the data have been collected.

## Create Structure

The basic idea of setting up a good workflow is that good planning allows the researcher to save time and allows other researchers to do their share of the analysis and/or replicate the research. After the data collection phase, the first step is to save the available data. We recommend keeping track of the dataset by providing data and data-related files in separate directories. We can do this with Windows Explorer or the Apple Mac's Finder. This directory should have subdirectories for at least the data files, the output, syntax, a temporary directory, and a directory with files that are directly related, such as the survey used to collect the data. In Table 5.1, we show an example of a directory structure. Within the main directory, there are five subdirectories, each with distinct files. Notice that in the *Data files* directory, we have the original dataset, two modified datasets (one without missing data and one which includes several transformations of the data) as well as a *.zip* file that contains the original dataset. If the data file is contained in a *.zip* file, it is unlikely to be modified and can be easily opened if the working file is accidentally overwritten or deleted. In the *Output*, *Syntax*, and *Temporary* directories, we provided each file with the suffix *rev1* or *rev2*. We use *rev* (abbreviation of revision), however, you could choose to use another file name, as long as it clearly indicates the revision on which you are working. Finally, in the *Related Files* directory, we have a codebook (more on this later), the survey, two presentations, and two documents containing recommendations.

**Table 5.1** Example of a directory structure for saving market research related files

Directory name	Subdirectory name	Example file names
2010_retailing project	Data files	Retailer.sav
		Retailer.zip
		Retailer rev1.sav
		Retailer rev2.sav
	Output files	Missing data analysis rev2.spv
		Descriptives rev2.spv
		Factor analysis rev2.spv
		Regression analysis rev2.spv
	Syntax files	Missing data analysis.sps
		Descriptives.sps
		Factor analysis.sps
		Regression analysis.sps
	Temporary	Missing data analysis rev1.spv
		Descriptives rev1.spv
		Factor analysis rev1.spv
		Regression analysis rev1.spv
	Related files	Codebook.doc
		Survey.pdf
		Initial findings – presentation to client.ppt
		Final findings – presentation to client.ppt
		Recommendations rev1.doc
		Recommendations rev2.doc

Another aspect to creating structure is to properly set up the variables for your study. Provide clear names for each variable. In SPSS, and most other statistical software programs, you can indicate a name for each variable and a description (in SPSS, this is called a *variable label* and is discussed later). The name of the variable should be short so that it can be read in the dialog boxes. For example, if you have three questions on web browsing enjoyment, three on time pressure during browsing, and several descriptors (age and gender), you could code these as *enjoy1-enjoy3*, *time1-time3*, and *age* and *gender*. The description of the variables should be more informative than the variable name. The description typically includes the original question if the data were collected using surveys. It is also a good idea to indicate what types of data have been collected. For example, you can collect data based on values (*Numeric* in SPSS) or on text (*String* in SPSS). In SPSS, you can indicate the measurement level; nominal data, ordinal data, or scale data (for ratio and interval scaled data). Another point to consider is *coding*. Coding means assigning values to specific questions. When quantitative data are collected, the task is relatively easy; for Likert and semantic differential scales, we use values that correspond with the answers. For example, for a five-point Likert scale, responses can be coded as 1–5 or as 0–4 (with 0 being the most negative and 4 being the most positive response).

Open-ended questions (qualitative data) require more effort; typically, a three-step process is involved. First, we collect all responses. In the second step, we group all responses. Determining the number of groups and to which group a response

belongs is the major challenge in this step. To prevent the process from becoming too subjective, usually two or three market researchers code the responses and discuss any arising differences. The third step is providing a value for each group.

Once a system is set up to keep track of your progress, you need to consider safeguarding your files. Large companies usually have systems for creating *backups* (extra copies of files as a safeguard). If you are working alone or for a small company, you will most probably have to take care of this yourself. You could save your most recent and second most recent version of your file on a separate drive. Always keep two copies and never keep both backups in the same place as theft, fire, or an accident could still mean you'll lose all of your work!

## Enter Data

Capturing the data is the next step for primary data. How do we enter survey or experimental data into a dataset? For large datasets, or datasets created by professional firms, specialized software is often used. For example Epidata (<http://www.epidata.dk>) is frequently used to enter data from paper-based surveys, Entryware's mobile survey (<http://www.techneos.com>) is commonly deployed to enter data from personal intercepts or face-to-face interviewing, while, for example, Voxco's Interviewer CATI is often used for telephone interviewing. The SPSS Data Collection Family (<http://www.spss.com/software/data-collection>) is a suite of different software packages specifically designed to collect and (automatically) enter data collected from online, telephone, and paper-based surveys.

Small firms or individual firms may not have access to such software and may need to enter data manually. You can enter data directly into SPSS. However, a drawback of directly entering data is the risk of making typing errors. Professional software such as Epidata can directly check if values are admissible. For example, if a survey question has only two answer categories such as gender (coded 0/1), Epidata and other packages can directly check if the value entered is 0 or 1, and not any other value. Moreover, if you are collecting very large amounts of data that require multiple typists, specialized software needs to be used. When entering data, check whether a substantial number of surveys or survey questions were left blank and note this.

## Clean Data

Cleaning data is the next step in the workflow. It requires checking for data entry errors, interviewer fraud, and missing data. Datasets often contain wrongly entered data, missing data, blank observations, and other issues that require some decision making by the researcher.

First, one should check if there are any data entry errors. Data entry errors are easy to spot if they occur outside the variable's range. That is, if an item is measured using a 5-point scale, then the lowest value should be 1 (or 0) and the highest 5 (or 4).

Using descriptive statistics (minimum, maximum, and range), we can check if this is indeed true. Data entry errors should always be corrected by going back to the original survey. If we cannot go back (e.g., because the data were collected using face-to-face interviews), we need to delete that particular observation for that particular variable. More subtle errors, for example, incorrectly entering a score of 4 as, say, 3 are difficult to detect using statistics. One way to check for these data entry errors is to randomly select observations and compare the entered responses with the original surveys. You would expect a small number of errors (below 1%). If many typing errors occurred, the dataset should be entered again.


Interviewer fraud is a difficult issue to deal with. It ranges from interviewers “helping” actual respondents provide answers to the falsification of entire surveys. Interviewer fraud is a serious issue and often leads to incorrect results. Fortunately, we can avoid and detect interviewer fraud in several ways. First, never base interviewers’ compensation on the number of completed responses. Once data have been collected, interviewer fraud may be detected in several ways. If multiple interviewers are used, the way in which they select respondents should be similar, if a reasonably large number of responses ( $> 100$ ) is collected per interviewer. We would therefore expect the responses obtained to also be similar. Using the testing techniques discussed in Chap. 6, we can test if this is indeed likely. Furthermore, the persons interviewed can be contacted afterwards to obtain their feedback on survey. If a substantial number of people do not claim to have been interviewed, interviewer fraud is likely. Furthermore, if people have been previously interviewed on a similar subject, we would expect factual variables (education, address, etc.) to change little. Using descriptive statistics, we can check if this is indeed true. If substantial interviewer fraud is suspected, the data should be discarded. Of course, the costs of discarding data are substantial. Therefore, firms should check for interviewer fraud during the data collection process as well.

Missing data are a frequently occurring issue that market researchers have to deal with. There are two levels at which missing data occur, namely at the survey level (entire surveys are missing) and at the item level (respondents have not answered some item). The first issue is called *survey non-response* and the second *item non-response*.

Survey non-response is very common. In fact, routinely, only 5–25% of all surveys are filled out by respondents. Although higher percentages are possible, they are not the norm for one-shot surveys. Issues such as inaccurate address lists, a lack of interest and time, people confusing market research with selling, and privacy issues, have led response rates to drop over the last decade. Moreover, the amount of market research has increased, leading to respondent fatigue and a further decline in response rates. However, the issue of survey response is best dealt with by properly designing surveys and the survey procedure (see Box 4.4 in Chap. 4 for suggestions). Moreover, survey response issues are only detected once there is a tabulation of all responses. The response percentage of a survey is calculated by dividing the number of usable responses by the total number of surveys sent out (called the *net response*) or by dividing the number of surveys received, including blank returned surveys or those partially completed, by the total number of surveys

sent out (called the *gross response*). Occasionally, researchers send out multiple surveys to the same respondent. If so, it is common to calculate how many respondents have provided at least one response and to divide this by the number of unique respondents to which a survey was sent out. This calculation is also followed for mixed-mode surveys.

Item non-response occurs when respondents do not provide answers to certain questions. This is common and typically 2–10% of questions remain unanswered. However, this number strongly depends on several factors such as the subject matter, the length of the questionnaire and the method of administration. For questions considered sensitive by many people (such as income), non-response can be much higher. The key issue with non-response is if the items are missing completely at random or if they are systematically missing. Items missing completely at random means that the observations answered are a random draw from the population. If data are systematically missing, the observations are not a random draw. For example, if we ask respondents what their incomes are, those with higher incomes are less likely to answer. We might therefore underestimate the overall income. Item non-response with a systematic component to it, nearly always leads to validity issues. How can we check if systematic item non-response issues are present? If we believe that the pattern of missing values is systematic, we could tabulate the occurrence of non-responses against responses for different groups. If we put the variable about which we have concerns on one side of the table, and the number of (non-)response on the other, a table similar to Table 5.2 is produced.

Using a  $\chi^2$ -test (pronounced as *chi-square*; which we discuss under nonparametric tests in the  Web Appendix → Chap. 6), we can test if there is a significant relationship between the respondents’ (non-)response and their income.

You can deal with these missing values in several ways. If the percentage of item non-response is low and the sample large, most researchers choose to completely delete all observations with missing data. However, with a moderate level of item non-response and a large number of items per survey, this leads to high percentages of surveys needing to be deleted (10% or more). In that case, most researchers choose to *impute* observations. Imputation means substituting a missing observation for a likely value. Several imputation procedures exist. However, before deciding on one, always check if imputation makes sense. The key issue is to check if an observation could have been answered. If so, imputation may make sense. If not, imputation should not be carried out. An example is to ask how satisfied someone is with his or her car(s). Even if no answer is provided, a car-owning respondent could have provided an answer and therefore imputation may make sense. However, if the respondent does not own a car, the question could not be answered, and imputation makes no sense.

**Table 5.2** Example of response issues

	Low income	High income
Response	95	80
Non-response	5	20

N = 200

**Table 5.3** Data cleaning issues and how to deal with them

Problem	Action
Data entry errors	Use descriptive statistics (minimum, maximum, range) to check for obvious typing errors and/or compare a subset of surveys to the dataset to check for inconsistencies.
Typing errors	Check descriptive statistics (minimum, maximum, range); use a sample of the surveys and check consistency between the surveys and data.
Interviewer fraud	Check respondents; correlate with previous data if available.
Survey non-response	Before sending out survey use Dillman’s (2008) design principles (see Box 4.4 in Chap. 4).
Item non-response	Check type of non-response: is non-response systematic or random? If small percentage of items is missing: delete cases. Otherwise, use an imputation method such as regression imputation.

A simple imputation procedure is *mean substitution*, which is nothing more than a substitution of a missing value by the average of all observed values of that variable. Mean substitution is simple and easily carried out in most analysis software. However, a drawback of mean substitution is that it reduces correlations (discussed later). We can also apply more complex methods (e.g., EM or regression substitution), which are also included in SPSS (these methods are not discussed here). Hair et al. (2010) provide a basic introduction to imputation. Table 5.3 summarizes the data issues discussed in this section.

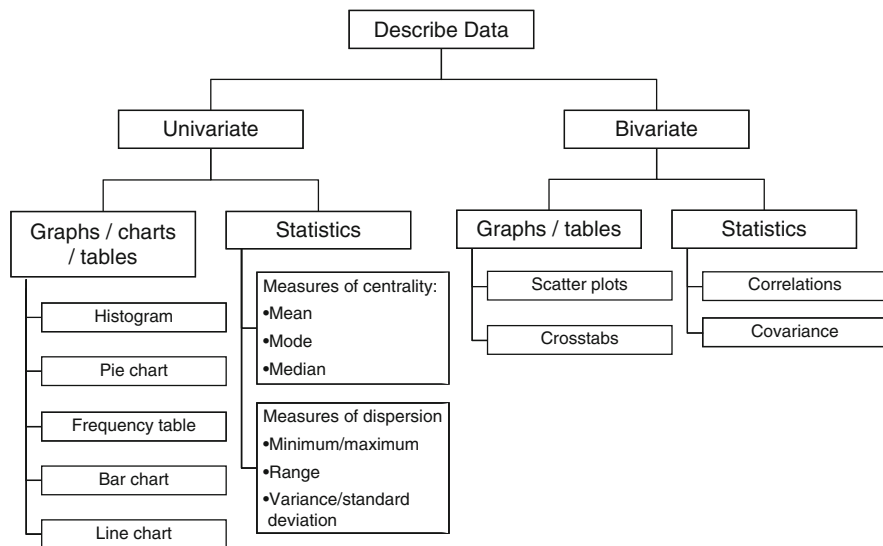
Describe Data

Once we have performed the previous steps, we can turn to the task of describing the data. Data can be described one variable at a time (*univariate descriptives*) or the relationship between two variables can be described (*bivariate descriptives*). We further divide univariate and bivariate descriptives into graphs/charts/tables and statistics.

The choice between graphs/charts/tables and statistics depends on what information you want to convey. Often graphs/charts/tables can tell a reader with a limited background in statistics a great deal. However, graphs/charts/tables can also mislead readers, as we will discuss later in Chap. 10 (Box 10.1). On the other hand, statistics require some background knowledge but have the advantage that they take up little space and are exact. We summarize the different types of descriptive statistics in Fig. 5.2.

Univariate Graphs/Charts/Tables

In the next section, we discuss the most prominent types of univariate graphs/charts/tables: the histogram, pie chart, frequency table, bar chart, and line chart. In Fig. 5.3, these different types of charts are used to show the different sales tax rates in the US. Notice how the charts display the same information in different ways.



**Fig. 5.2** The different types of descriptive statistics

The *histogram* is a graph that shows how frequently a particular variable's values occur. The values a variable can take on are plotted on the  $x$ -axis, where the values are divided into (non-overlapping) classes that are directly adjacent to one another. These classes usually have an equal width. For example, if you create a histogram for the variable *age*, you can use classes of 0–10, 11–20, etc. where the width is equal. However, you could also choose classes such as 0–20, 21–25, 26–30. If you think important information will be left out if too many variables are included in the same class, it is best to use classes with an unequal width or to increase the number of classes. Because histograms use classes, they are helpful for describing interval or ratio scaled data. A histogram can also be used if you want to show the distribution of a variable. If you have a variable with a very large number of categories (more than ten), a histogram becomes difficult to understand. Similarly, if the difference between the least frequently and most frequently occurring values is very high, histograms become difficult to read.

The *pie chart* visualizes how a variable's different values are distributed. Pie charts are easy to understand and work well if the number of values a variable takes on is small (less than 10). Pie charts are particularly useful for displaying percentages of variables, because people interpret the entire pie as being 100%, and can easily see how often a variable's value occurs.

A *frequency table* is a table that includes all possible values of a variable and how often they occur. It is similar to both the histogram and pie chart in that it shows the distribution of a variable's possible values. However, in a frequency table, all values are indicated exactly.

A *bar chart* plots the number of times a particular value occurs in a data set, with the height of the bar representing the number of times a value is observed. Bar



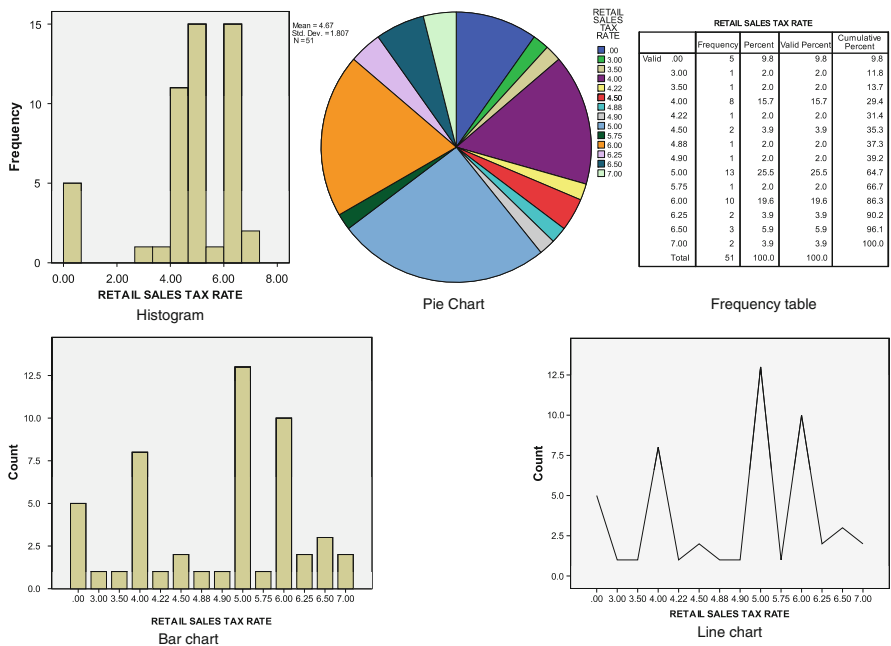


Fig. 5.3 Several univariate graphs, charts, and tables

charts are useful for describing nominal or ordinal data. The bars of bar charts are not directly adjacent, as some blank space is included between each bar. A bar chart is somewhat similar to a histogram, but bar charts can indicate how often the different values of a variable occur (e.g., 20 females, 30 males), can describe different variables next to each other (e.g., average income is 30,000 USD, average age is 39) or show the values of each observation (e.g., observation 1, age = 31, observation 2, age = 25, etc.)

A *line chart* also describes the different values occurring in a variable but connects each value, thereby giving the impression that that particular variable is continuous. Line charts work well if you want to visualize how a variable changes over time.

### Univariate Statistics

Univariate statistics fall into two groups: those describing centrality and those describing the dispersion of variables. Centrality-based measures include the mean, the mode, and the median. The *mean* (or average) is the sum of each individual observation of a variable divided by the number of observations. The *median* is the value that separates the lowest 50% of cases from the highest 50% of cases. The *mode* is the most frequently occurring value in the dataset.

Nominal data may be summarized using the mode, but not the median nor the mean. The median and also the mode are useful when describing ordinal data. The mean is useful if our data are interval or ratio-scaled.

Each measure of centrality has its own uses. The mean (abbreviated as  $\bar{x}$ ) is most frequently used but is sensitive to unexpected large or small values. It is calculated by the sum of each observation's value divided by the number of observations.  $x_i$  refers to observation  $i$  of variable  $x$  and  $n$  to the total number of observations.

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

Contrary to the mean, neither the mode nor median are sensitive to outliers. As a consequence, the relationship between mean, median and mode provides us with valuable information regarding the distribution of a variable. For example, if the mean is much higher than the median and mode, this suggests that the dataset contains outliers which shift the mean upwards. This is the case when we observe the values 4, 5, 5, and 50, where both median and mode are 5 while the mean is 16. If the mean, median, and mode are more or less the same, the variable is likely to be symmetrically distributed.

For the measures of dispersion, the *minimum* and *maximum* indicate a particular variable's highest and lowest value. The *range* is the difference between the highest value and the lowest value. The *variance* (abbreviated as  $s^2$ ) measures the sum of the squared differences between all of a variable's values and its mean divided by the number of observations minus 1.

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

Variance is one of the most frequently used measures of dispersion. It tells us how strongly observations vary around the mean. By using squared differences between observations and the mean, positive and negative differences cannot cancel each other out. Likewise, values that lie far from the mean increase the variance more strongly than those that are close to the mean.

The *standard deviation* (abbreviated as  $s$ ) is the square root of – and, therefore, a special case of – the variance. Whereas the mean is usually easy to interpret, the variance and standard deviation are less intuitive. A rule of thumb is that in large (normally distributed – this will be discussed in later chapters) datasets, about two-thirds of all observations are between plus and minus one standard deviation away from the mean. Thus, if the standard deviation is 0.70 and the mean 3, 64% of all observations will likely fall between 2.30 and 3.70. Approximately 2 standard deviations plus, or minus the mean, usually indicates that 95% of all observations fall between 1.60 and 4.40.

## Bivariate Graphs/Tables

There is only a small number of bivariate graphs and tables of which the *scatter plot* and the *crosstab* are the most prominent. The scatter plot uses both the  $y$  and  $x$ -axis (and sometimes a third  $z$ -axis) to show how two or three variables relate to one another. Scatter plots are useful to identify outliers and to show the relationship between two variables. Relationships are much harder to visualize when there are three variables. Scatter plots work well if the sample size is not too large (about 100 observations or less). We give an example of a scatter plot later on.

## Bivariate Statistics

While univariate statistics provide insights regarding one variable, bivariate statistics allow for measuring how two variables are associated. Next we will discuss two bivariate measures, the covariance and the correlation.

### Covariance

Two key measures to discuss associations between two variables are the *covariance* and the *correlation*. The covariance is the degree to which variables vary together. It is the sum of the multiplications of the differences between every value of the  $x$  and  $y$  variable and their means.

$$\text{Cov}(x_i, y_i) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

### Correlations

A *correlation* (abbreviated as  $r^2$ ) is a measure of how strongly two variables relate to each other. Correlation coefficients are frequently used to describe data because they are relatively easy to use and provide a great deal of information in just a single value. A (Pearson) correlation coefficient is calculated as follows:

$$r^2 = \frac{\text{Cov}(x_i, y_i)}{s_x \cdot s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

The numerator contains the covariance of  $x$  and  $y$  ( $\text{Cov}(x_i, y_i)$ ) and the denominator contains the product of the standard deviations of  $x$  and  $y$ .<sup>1</sup> Thus, the

---

<sup>1</sup>Note that the terms  $n-1$  in the numerator and denominator cancel each other out and, thus, are not displayed here.

**Table 5.4** Types of descriptive statistics for differently scaled variables

	Nominal	Ordinal	Interval & ratio (Scale in SPSS)
Display distribution	Frequency table	Frequency table	Frequency table
Measure of centrality	Mode	Mode and median	Mode, median, and mean
Measures of dispersion	–	Range	Standard deviation, variance
Measures of association	–	Spearman’s correlation coefficient	Pearson correlation coefficient, covariance

correlation is the covariance divided by the product of the standard deviation. Therefore, the correlation is no longer dependent on the variables’ original measurement as it is the case with the covariance. As such, the calculated value of the correlation coefficient ranges from  $-1$  to  $1$ , where  $-1$  indicates a perfect negative relation (the relationship is perfectly linear) and  $1$  indicates a perfectly positive relationship. A correlation coefficient of  $0$  indicates that there is no correlation.

Cohen (1988) defined standards to describe how “strong” a correlation is. He argued that absolute correlation coefficients below  $0.30$  indicate a weak effect, coefficients between  $0.30$  and  $0.49$  indicate a moderate effect, and values of  $0.50$  and higher indicate a strong effect.

There are several types of correlations that can be calculated but two are most prominent. A Pearson correlation coefficient is the most commonly used and is often simply referred to as correlation. It is appropriate for calculating correlations between two interval or ratio scaled variables. When at least one variable is ordinaly scaled, we use Spearman’s correlation coefficient.

In Table 5.4, we indicate which descriptive statistics are useful for differently scaled variables.

## Detect Outliers

Data often contain *outliers*. Outliers are values that are uniquely different from all the other observations and influence results substantially. For example, if we compare the average amounts different households spend on transport, we may find that some spend as little as  $200$  USD per year, whereas a few households may spend millions (e.g., because they own very expensive cars). If we were to calculate the mean, including those people who spend millions, this would substantially increase the average.

Outliers come in different forms. The first type of outlier is produced by data collection or entry errors. For example, if we ask people to indicate their household income in thousands of USD, some respondents may just indicate theirs in USD (not thousands). Obviously, there is a substantial difference between  $30$  and  $30,000$ ! Moreover, clear data entry errors ( $55$  instead of  $5$ ) often occur. Outliers produced by

data collection or entry errors either need to be deleted or we need to find the correct values, for example, by going back to the respondents.

A second type of outlier occurs because exceptionally high or low values are a part of reality. While such observations can significantly influence results, they are sometimes specifically important for researchers as characteristics of outliers can be insightful. Think for example of companies that are extremely successful or users that face needs long before the bulk of that marketplace encounters them (so called lead users). Malcolm Gladwell's (2008) book "Outliers: The Story of Success" provides entertaining study of what really explains exceptionally successful people (outliers).

A third type of outlier occurs when combinations of variables are exceptionally rare. For example, if we look at income and expenditure on holidays, we may find someone who earns 500,000 USD but spends 300,000 USD of his/her income on holidays. Such combinations are unique and have a very strong impact on particularly correlations.

Finding outliers means finding very low or very high variable values. This can be achieved by calculating the minimum, maximum and range of each variable and through scatter plots (if outliers are a rare combination of variables). The next step is to determine whether we have an explanation for those high or low values. If there is an explanation (e.g., because some exceptionally wealthy people were included in the sample), outliers are typically retained. If the explanation is that it is most probably a typing or data entry error, we always delete these outliers. If there is no clear explanation, the decision to delete or retain outliers is not clear-cut and outliers are generally retained.

## Transform Data

Transforming data is an optional step in the workflow. Researchers transform data for several reasons. The key reasons are that transforming data may be necessary for some analysis techniques, because it may help interpretation or because it is necessary to meet assumptions of techniques discussed in the subsequent chapters. Transforming data means that the original values of a variable are changed consistently by means of a mathematical formula. An example of a simple transformation is coding a variable into two categories. For example, if we have a variable measuring a respondent's income, we may code incomes below 20,000 USD as *low* and those above as *high*. Thus, we use a simple rule to recode our original variable as a new variable.

If we choose to use cluster analysis (see Chap. 9), we often need to *standardize* our variables. Standardizing variables means that we rescale the data so that the mean of the variable is 0 and the standard deviation is 1. This type of standardization is called the *z-transformation* and is applied by subtracting the mean  $\bar{x}$  of every observation  $x_i$  and dividing it by the standard deviation  $s$ . That is:

$$z_i = \frac{(x_i - \bar{x})}{s}$$

The *log transformation* is another type of transformation, which is commonly used if we have *skewed data*. Skewed data arises if we have a variable that is asymmetrically distributed. For example, family income is often a highly skewed variable. The majority of people will have incomes around the median, but a few people will have very high incomes. However, no one will have a negative income leading to a “tail” at the right of the distribution. A histogram will quickly show if data are skewed. Skewed data can be problematic in analyses and it has therefore become common practice to use a log transformation. A log transformation applies a base 10 logarithm to every observation.<sup>2</sup> We should only use the log transformations for skewed variables that have positive numbers. The log cannot be calculated for negative numbers. Another complication is that we cannot calculate the log for the value of 0. A way around this is to add 1 to every observation and then apply a logarithmic transformation. This of course leads to a different interpretation.

Dummy coding is a special way of recoding data. *Dummies* are binary variables that indicate if a variable is present or not. For example, we can use dummies to indicate that advertising was used during a particular period (value of the dummy is 1) but not in other periods (value of the dummy is 0). We can use multiple dummies at the same time, for example to indicate if advertising or a special product promotion was used. We can also use multiple dummies to capture categorical variables’ effects. For example, if we want to understand the effects of no, some, and intensive advertising, we can create two dummies, one where the dummy takes the value of 1 if intensive advertising was used (else 0) and one where the value of 1 is used for some advertising (else 0). In this way, we always construct one dummy less than the amount of categories used (in this example, if both dummies take the value zero, this indicates no advertising). We explain dummies in further detail in the Web Appendix (🔗 Web Appendix → Chap. 5)

A frequently used type of recoding is to create *constructs*. As described in Chap. 3, a construct can be defined as a concept which cannot be directly observed, and for which there are multiple referents, but none all-inclusive. The construct is not an individual item that you see in the list, but it is captured by calculating the average of a number of related items. For example, if we want to measure the construct brand trust using three items (“This brand’s product claims are believable,” “This brand delivers what it promises,” and “this brand has a name you can trust”), and one respondent indicated 4, 3, and 6 for these three items, then the construct score for this respondent would be  $(4 + 3 + 6)/3 = 4.33$ .

---

<sup>2</sup>The logarithm is calculated as follows: If  $x=y^b$ , then  $y=\log_b(x)$  where  $x$  is the original variable,  $b$  the logarithm’s base, and  $y$  the exponent. For example,  $\log_{10}$  of 100 is 2 because  $10^2$  is 100. Logarithms cannot be calculated for negative values (such as household debt) and the value of zero.

A special type of data transformation is *aggregation*. Aggregation means that we bring variables measured at a lower level to a higher level. For example, if we know the average of customers' satisfaction and from which stores they buy, we can calculate average satisfaction at the store level. Aggregation only works one way (from lower to higher levels) and is useful if we want to compare groups.

There are also drawbacks to transforming data, such as that we lose information through most transformations. For example, recoding income in USD (measured at the ratio scale) into a "low" and "high" income category will result in an ordinal variable. Thus, in the transformation process, we have lost information. A simple rule of thumb is that information will be lost if we cannot move back from the transformed to the original variable. Another drawback is that transformed data are often more difficult to interpret. For example, log USD or log EUR are much more difficult to interpret and less intuitive.

## Create a Codebook

After all the variables have been organized and cleaned and some initial descriptives have been calculated, a *codebook* is often created. A codebook contains essential details of the data file to allow the data to be shared. In large projects, multiple people usually work on data analysis and entry. Therefore, we need to keep track of the data to minimize errors. Even if just a single researcher is involved, it is still a good idea to create a codebook, as this helps the client use the data and helps if the same data are used later on. On the website accompanying this book ([Web Appendix → Chap. 5](#)), we briefly discuss how a codebook can be created using SPSS. Codebooks usually have the following structure:

*Introduction:* In the introduction, we discuss the goal of the data collection, why the data are useful, who participated, and how the data collection effort was conducted (mail, Internet, etc.).

*Questionnaire(s):* It is common practice to include copies of all types of questionnaires used. Thus, if different questionnaires were used for different respondents (e.g., for French and Italian respondents), a copy of each original questionnaire should be included. Differences in wording may explain the results of the study afterwards, particularly in cross-national studies, even if a back-translation was used (see Chap. 4). These are not the questionnaires received from the respondents themselves but blank copies of each type of questionnaire used. Most codebooks include details of the name of each variable behind the items used. If a dataset was compiled using secondary measures (or a combination of primary and secondary data), the secondary datasets are often briefly discussed (what version was used, when it was accessed, etc.).

*Description of the variables:* This section includes a verbal description of each variable used. It is useful to provide the variable name as used in the data file, a description of what the variable is supposed to measure, and whether the measure has been used previously. You should also describe the measurement level (see Chap. 3).

*Summary statistics:* We include descriptive statistics of each variable in the summary statistics section. The average (only for interval and ratio-scaled data), minimum, and maximum are often calculated. In addition, the number of observations and usable observations (excluding observations with missing values) are included, just like a histogram.

*Datasets:* This last section includes the names of the datasets and sometimes the names of all the revisions of the used datasets. Sometimes, codebooks include the file date to assure that the right files are used.

## Introduction to SPSS

SPSS is a computer package specializing in quantitative data analysis. It is widely used by market researchers. It is powerful, able to deal with large datasets, and relatively easy to use.

In this book, we use version 18 of SPSS, officially called IBM SPSS Statistics. The SPSS screens somewhat confusingly display the name PASW. For simplicity, we refer to version 18 as SPSS. Versions 16 and 17 are similar and can also be used for all examples throughout this book. Versions 16, 17, and 18 are available for Microsoft Windows (XP or higher), the Mac (OS X version 10.5 or higher), and Linux. The differences between these versions are small enough so that all examples in the book should work with all versions.

The regular SPSS package is available at a substantial fee for commercial use. However, large discounts are available for educational use. To obtain these discounts, it is best to go to your university's IT department and enquire if (and where) you can purchase a special student license. You can also download a trial version from [www.spss.com](http://www.spss.com).

In the next sections, we will use the ► sign to indicate that you have to click on something with your mouse. Options, menu items or drop-down lists that you have to look up in dialog boxes are printed in **bold**, just like elements of SPSS output. Variable names, data files or data formats are printed in *italics* to differentiate those from the rest of the text.

## Finding Your Way in SPSS

If you start up SPSS for the first time, it presents a screen similar to Fig. 5.4, unless a previous user has ticked the **Don't show this dialog in the future** box. In that case, you will see a screen similar to Fig. 5.5, but without an active dataset.

In the startup screen, SPSS indicates several options to create or open datasets. The options that you should use are either **Open an existing data source**, under which you can find a list with recently opened data files, or **Type in data**. To open





Fig. 5.4 The start-up screen of SPSS

	Enjoy1	Enjoy2	Enjoy3	Time1	Time2	Time3	Age	Gender	Size_city	Income	Var
1	3	5	3	3	1	4	38	0	5	\$19,747.00	
2	6	6	3	6	0	0	52	1	8	\$19,595.00	
3	0	4	0	5	4	5	59	0	8	\$39,554.00	
4	1	3	0	2	5	1	30	0	8	\$39,708.00	
5	1	6	0	0	0	0	47	1	5	\$22,835.00	
6	0	4	6	0	0	0	44	0	5	\$24,059.00	
7	4	6	4	0	0	0	24	0	8	\$23,360.00	
8	1	6	3	1	1	1	43	1	8	\$26,265.00	
9	5	6	4	0	0	0	50	1	5	\$22,475.00	
10	4	5	1	1	5	1	51	1	5	\$23,288.00	
11	1	3	1	1	1	1	55	1	6	\$39,405.00	
12	2	3	1	4	5	4	16	0	6	\$23,904.00	
13	4	5	4	0	1	1	28	1	7	\$19,589.00	
14	4	5	2	0	0	0	27	0	8	\$21,345.00	
15	4	6	3	0	0	0	39	1	8	\$26,871.00	
16	5	5	3	5	1	1	45	0	1	\$19,872.00	
17	1	5	1	1	0	0	24	0	8	\$19,215.00	
18	2	4	3	2	1	1	30	1	8	\$21,898.00	
19	0	4	2	4	4	1	33	1	8	\$19,271.00	
20	3	5	2	1	1	0	28	0	8	\$22,730.00	
21	5	6	5	0	0	0	31	1	8	\$19,691.00	
22	6	6	4	1	1	1	19	1	8	\$22,476.00	

Fig. 5.5 The SPSS data editor

an unlisted file, simply choose **More Files...** and click **OK** (alternatively, you can click **Cancel**, and then go to ► **File** ► **Open** ► **Data**). Then search for the directory in which the files are kept, click on the file and then on ► **Open**. For the subsequent examples and illustrations, we use a dataset called *retailer.sav* (Web Appendix → Chap. 5). This dataset contains information on how participants feel about the Internet and how they experienced their last Internet session. Moreover, it contains the participant's age, gender, income and the size of city he/she lives in.

SPSS uses two windows, the **SPSS Statistics Data Editor** and the **SPSS Statistics Viewer**. The data editor has two tab fields in the lower left corner. The first is the **Data View** and the second the **Variable View**. Both provide details on the same dataset. The data view shows you the data itself with the variable names in the columns and the observations in the rows. Under the variable view tab, you will see many details of the variables, such as the variable name and type. The **SPSS Statistics Viewer** is a separate window, which opens after you carry out an action in SPSS. The viewer contains the output that you may have produced. If you are used to working with software such as Microsoft Excel, where the data and output are included in a single screen, this may be a little confusing at first. Another aspect of the viewer screen is that it does not change your output once made. Unlike, for example, Microsoft Excel, changing the data after an analysis does not dynamically update the results.

Tip: the notation you will see in this book follows the US style. That is, commas are used to separate thousands (1,000) while decimal points are used to separate whole values from fractions. If you want to change the notation to US style, go to SPSS, then ► **File** ► **New** ► **Syntax** and type in "*SET LOCALE = 'English'*". You also need to type in the last point. Now press enter and type in "*EXECUTE.*" (again, including the point) in the next line. Now run the syntax by choosing **Run** ► **All** in the syntax window. SPSS will then permanently apply the US style notation the next time you use SPSS.

SPSS uses multiple file formats. The *.sav* file format contains data only. SPSS also allows you to open other file formats such as Excel (*.xls* and *.xlsx*) and text files (such as *.txt* and *.dat*). Once these files are open, they can be conveniently saved into SPSS's own *.sav* file format. If you are on SPSS's main screen (see Fig. 5.4) simply go to **File** ► **Open** ► **Files of type** (select the file format) and double click on the file you wish to open. The output produced in SPSS can be saved using the *.spv* file format that is particular to SPSS. To partially remedy this, SPSS provides the option to export the output to Microsoft Word, Excel, or PowerPoint, PDF, HTML, or text. It is also possible to export output as a picture. You can find these export options in the **Statistics Viewer** under **File** ► **Export**.

In the **SPSS Statistics Data Editor** (Fig. 5.5), you will find the dataset *retailer.sav*. This dataset's variables are included in the columns and their names are indicated at the top of each column. The cases are in the rows, which are numbered from 1 onwards. If you click on the **Variable View** tab, SPSS will show you a

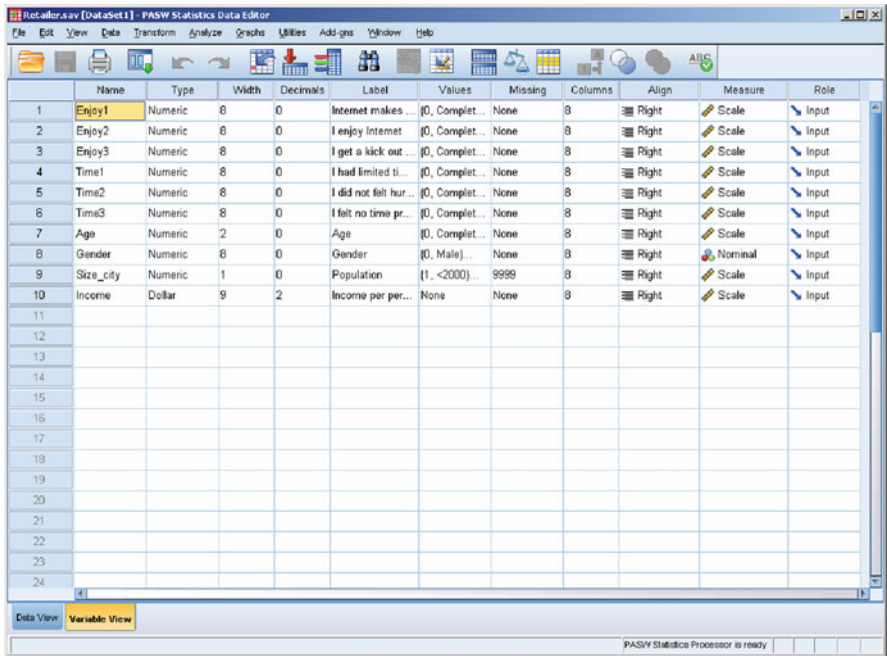


Fig. 5.6 The variable view

screen similar to Fig. 5.6. In the **Variable View**, SPSS provides information on the variables included in your dataset:

- **Name**: here you can indicate the name of the variable. It is best to provide very short names here. Variable names must begin with letters (A to Z) or one of the following special characters (@, # or \$). Subsequent characters can include letters (A to Z), numbers (0–9), a dot (.), and \_, @, #, or \$. Note that neither spaces nor other special characters (e.g., %, &, /) are allowed.
- **Type**: here you can specify what your variable represents. **Numeric** refers to values and **String** refers to words. String is useful if you want to type over open-ended answers provided by respondents. With **Dollar** or **Custom Currency**, you can indicate that your variable represents money. **Width** and **Decimals** indicate the amount of space available for your variables.
- **Labels**: here you can provide a longer description of your variables (called variable labels). This can either be the definition of the variables or the original survey question.
- **Values**: here you can indicate what a certain value represents (called value labels). For example, for gender, which is measured on a nominal scale, 0 could represent “female” and 1 “male.”
- **Missing**: here you can indicate one or more missing value(s). Generally, SPSS deals with missing values in two ways. If you have blanks in your variables (i.e.,

you haven't entered any data for a specific observation), SPSS treats these as system-missing values. These are indicated in SPSS by means of a dot (.). Alternatively, the user (you!) can define user-missing values that are meant to signify a missing observation. By defining user-missing values, we can indicate why specific scores are missing (e.g., the question didn't apply to the respondent or the respondent refused to answer). In SPSS, you should preferably define user-missing values as this at least allows the true missing values to be separated from data that were not recorded (no value was entered). The options under **Missing** provide three options to indicate user-defined missing values. The first option is **No missing values**, which is the default setting and indicates that no values are user-defined missing values. The other two options, **Discrete missing values** and **Range plus one optional discrete missing value**, provide a means to express user-defined missing values. To do so, simply enter values that record that an observation is missing. Each separate value should indicate separate reasons for missing values. For example –999 may record that a respondent could not answer, and –998 that the respondent was unwilling to answer, and –997 might record “other” reasons. Thus, missing values can provide us with valuable information on the respondents' behavior. More importantly, observations with user-missing values (just like with system-missing values) are excluded from data manipulation and analyses. This is of course essential as, for example, including a user-missing value of –999 in descriptive analyses would greatly distort the results. One tip: When picking missing values, take values that would not otherwise occur in the data. For example, for a variable *age*, 1,000 might be an acceptable value, as that response cannot occur. However, the same missing value for a variable *income* might lead to problems, as a respondent might have an income of 1,000. If 1,000 is indicated as (user-defined) missing value, this observation would be excluded from analysis.

- **Columns** and **Align**: these are rarely necessary, so we will skip these.
- **Measure**: here you can specify the measurement level of your variable. SPSS provides you with the option to indicate whether your variable is nominal, ordinal, or whether it is interval or ratio-scaled. The combination of the last two categories is called **Scale** in SPSS.
- The last **Role** option is not necessary for basic analysis.

In SPSS, you also find a number of commands in the menu bar. These include ► File, ► Edit, ► View, ► Data, and ► Transform. Next, we will describe these commands. Thereafter, we will describe ► Analyze ► Graphs to discuss how descriptive statistics and graphs/charts can be produced. The last four commands are ► Utilities, ► Add-ons, ► Windows, and ► Help. You are unlikely to need the first three functions but the help function may come in handy if you need further guidance. Under help, you also find a set of tutorials that can show you how to use most of the commands included in SPSS.

In addition to the menu functionalities, we can run SPSS by using its command language, called SPSS syntax. You can think of it as a programming language that SPSS uses to “translate” those elements on which you have clicked in the menus

into commands that SPSS can understand. Discussing the syntax in great detail is beyond the scope of this book but, as the syntax offers a convenient way to execute analysis, we offer an introduction in the Web appendix (🔗 Web Appendix → Chap. 5). Also, Collier (2010) provides a thorough introduction into this subject. Syntax can be saved (as a *.sps* file) for later use. This is particularly useful if you conduct the same analyses over different datasets. Think, for example, of standardized marketing reports on daily, weekly, or monthly sales.

Under ► File, you find all the commands that deal with the opening and closing of files. Under this command, you will find subcommands that you can use to open different types of files, save files, and create files. If you open a dataset, you will notice that SPSS also opens a new screen. Note that SPSS can open several datasets simultaneously. You can easily switch from dataset to dataset by just clicking on the one which you would like to activate. Active datasets are marked with a green plus sign at the top left of the **Data View** and **Variable View** screen.

Under ► Edit, you will find subcommands to copy and paste data. Moreover, you will find two options to insert cases or variables. If you have constructed a dataset but need to enter additional data, you can use this to add an additional variable and subsequently add data. **Edit** also contains the **Find** subcommand with which you can look for specific cases or observations. This is particularly useful if you want to determine, for example, what the income of 60-year old respondents is (select the *Age* column, then ► Edit ► Find and type in 60). Finally, under ► Edit ► Options, you find a large number of options, including how SPSS formats tables, and where the default file directories are located.

Under ► View, you find several options, of which the **Value Labels** option is the most useful. Value labels are words or short sentences used to indicate what each value represents. For example, *Male* could be the value label used to indicate that responses coded with a 1 correspond to males. SPSS shows value labels in the SPSS Data Editor window if you click on ► View, and then on **Value Labels**.

Under the ► Data command, you will find many subcommands to change or restructure your dataset. The most prominent option is the ► Sort Cases subcommand with which you can sort your data based on the values of a variable. You could, for example, sort your data based on the income of the respondent. The ► Split File subcommand is useful if you want to compare output across different groups, for example, if you want to compare males and females' yearly income. Moreover, we can carry out separate analyses over different groups using the **Split File** command. After clicking on split file, SPSS shows a dialog box similar to Fig. 5.7. All you need to do is to enter the variable that indicates the grouping (select **Compare groups**) and move the grouping variable into **Groups Based on**: SPSS will automatically carry out all subsequent analyses for each subgroup separately. If you want to revert to analyzing the whole dataset, you need to go to ► Data ► Split File, then click on **Analyze all cases, do not create groups**. Attention! It is a common mistake to forget to turn off the split file command. Failing to turn off the command results in all subsequent analyses being carried out over part of the dataset, and not the whole dataset.

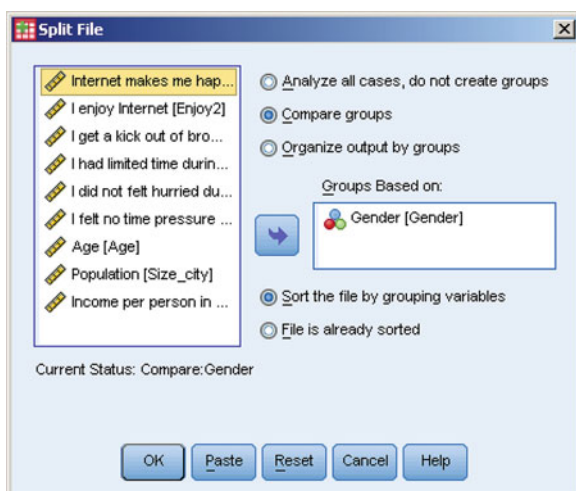


Fig. 5.7 Split file command

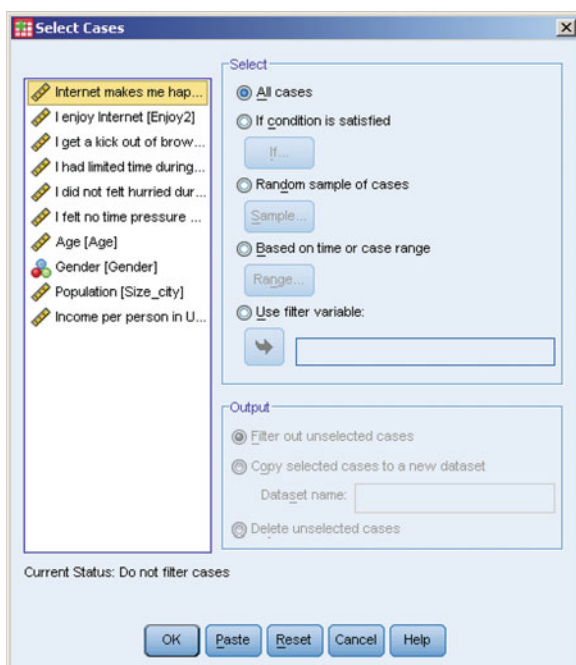


Fig. 5.8 Select cases

Another very useful command is ► Data ► Select cases (SPSS). After clicking on this, SPSS shows a dialog box similar to Fig. 5.8. Here, you can select the cases that you want to analyze. If you go to **If condition is satisfied** and click on **If**, you will see a new screen that you can use to select a subset of the data. For example, you can tell SPSS to analyze females only by entering *gender* = 0 (assuming gender is 0 for females). Back in the data view, SPSS will cross out the observations that will not be included in any subsequent analyses. Remember to turn the selection off if you do not need it by going back to ► Data ► Select Cases and then click on **All cases**.

Under ► Transform, we find several options to create new variables from existing variables. The first subcommand is **Compute variable (SPSS)**. This command allows you to create a new variable from one (or more) existing variables. For example, if you want to create a construct (see Chap. 3), you need to calculate the average of several variables. After you click on ► Transform ► Compute Variable, SPSS shows a dialog box similar to Fig. 5.9.

In this dialog box, you can enter the name of the variable you want to create under **Target Variable** and under **Numeric Expression** you need to indicate how the new variable will be created. In the example, we create a new construct called *Enjoy* which is the average of variables *Enjoy1*, *Enjoy2*, and *Enjoy3*. You can use the buttons in the dialog box to build the formulas that create the new variable (such as the + and – button), or you can use any of the built-in functions that are found under **Function group**. SPSS provides a short explanation of each function. The compute variable command is very useful to transform variables, as we discussed

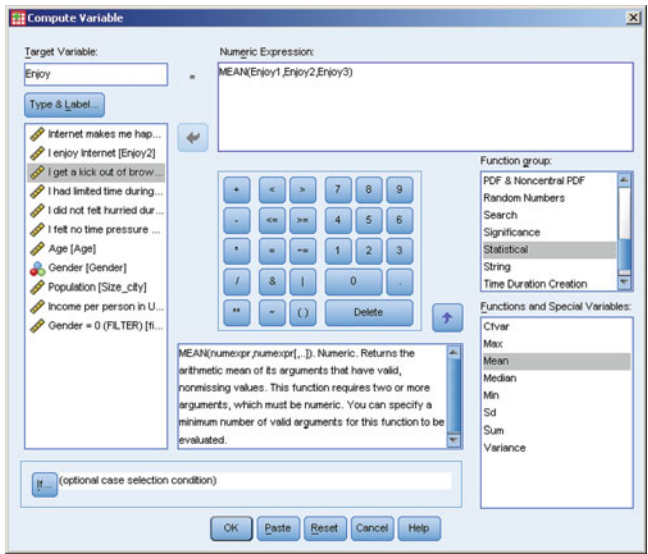
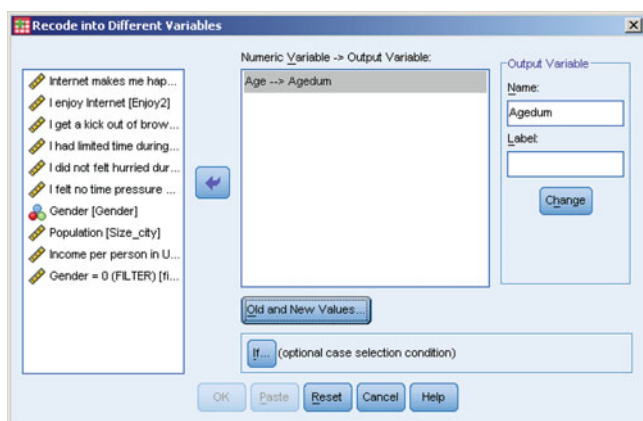


Fig. 5.9 Compute new variables

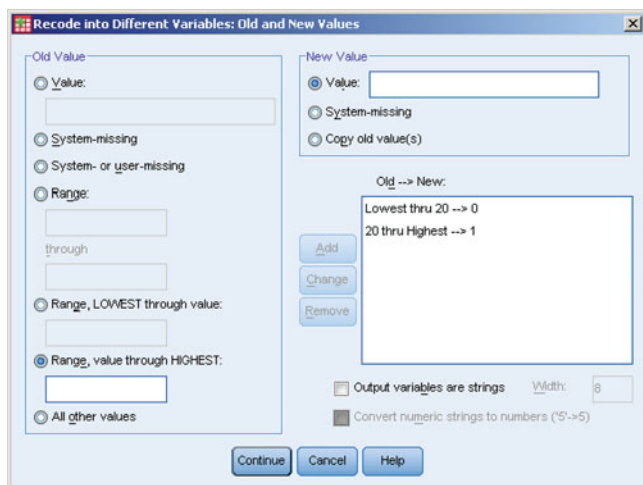
earlier in this chapter. For example, log transformations can be performed with the *Ln* function.

Also included under the ► Transform command are two subcommands to recode variables; ► Recode into Same Variables and ► Recode into Different Variables. We always recommend using the recode into different variables option. If you were to use recode into the same variables, any changes you make to the variable will result in the deletion of the original variable. Thus, if you ever want to go back to the



**Fig. 5.10** Recode into different variables

original data, you either need to have saved a previous version, or enter all data again as SPSS cannot undo these actions! The recode subcommands allow you to change the scaling of a variable. This is useful if you want to create dummies or create











**Fig. 5.11** Recode options



categories of existing variables. For example, if you want to create a dummy variable to compare the younger and older respondents, you could divide the variable age into young (say, younger than 20) and old (say, 20 and older). Figure 5.10 illustrates dialog box for recode into different variables.

In Fig. 5.10, we see several boxes. The first box at the far left shows all variables included in the dataset. This is the input for the recode process. To recode a variable, move it to the middle box by using the arrow located between the left and middle boxes. On the far right, you see an **Output Variable** box in which you should enter the name of the variable and an explanatory label you want to create. Next, click on **Change**. You will then see that in the middle box an arrow is drawn between the original and new variables. The next step is to tell SPSS what values need to be recoded. Do this by clicking on **Old and New Values**. After this, SPSS will present a dialog box similar to Fig. 5.11.

**Table 5.5** Shortcuts symbols

Symbol	Action
	Open dataset
	Save the active dataset
	Recall recently used dialogs
	Undo a user action
	Find
	Split file
	Select cases
	Show value labels

On the left of this dialog box, you should indicate the values of the original variable that you want to recode. On the right, you can indicate the new values for the new variable.

Some of the previous commands are also accessible by means shortcut symbols in **Data View** screen’s menu bar. As these are very convenient, we describe the most frequently used shortcuts in Table 5.5.

## Creating Univariate Graphs, Charts, and Tables in SPSS

In SPSS, all descriptives are included in ► Analyze, and ► Graphs. The ► Graphs command is divided into two subcommands, the **Chart Builder** and **Legacy Dialogs**. The chart builder provides more functionality than the legacy dialogs but is slightly more difficult to use. The legacy dialog provides simple and quick access to the most commonly used graph types, such as the bar chart, pie chart, and scatter plot. Descriptive statistics are included under ► Analyze ► Descriptive Statistics. In Fig. 5.12, we show how to calculate each previously discussed descriptive statistic in SPSS.

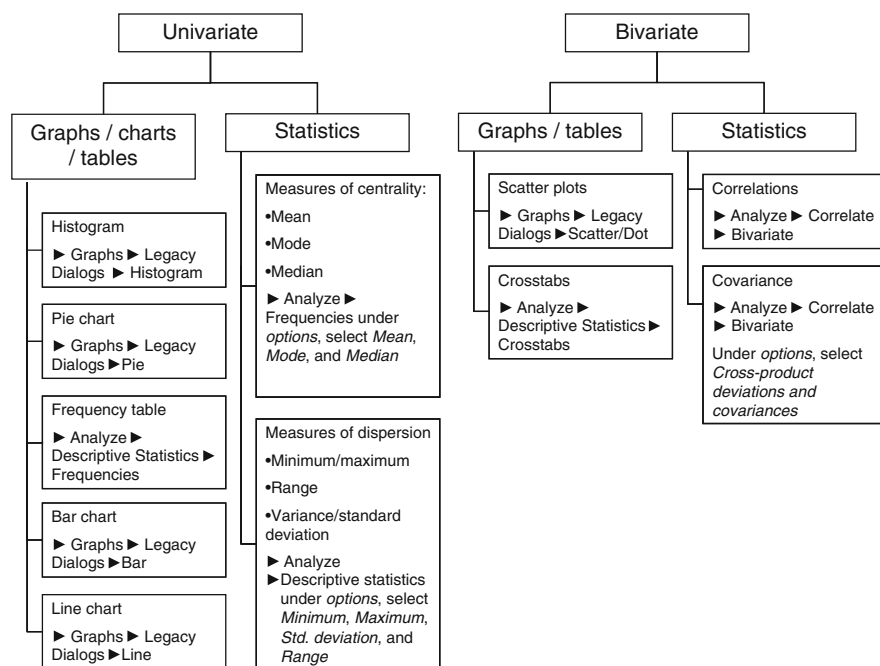


Fig. 5.12 How to calculate descriptive statistics and graphs in SPSS

### Histograms

Histograms are created by clicking on ► Graphs ► Legacy Dialogs and then on ► Histogram. SPSS will then produce a dialog box in which you should enter the variable for which you want to produce a histogram under **Variable**. If you want to create multiple histograms on top of, or next to each other, you can add a variable for which multiple histograms need to be made under **Rows** or **Columns**. If you just move *Internet makes me happy* into the **Variable** box, SPSS will produce a histogram as in Fig. 5.13.

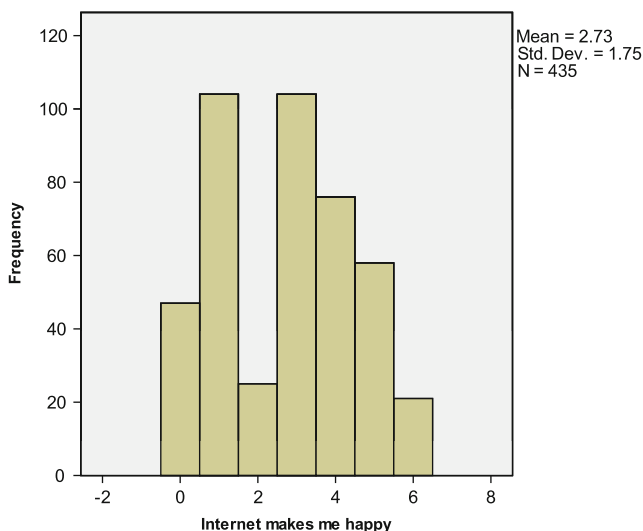


Fig. 5.13 A simple histogram

## Pie Charts

Pie charts can be made by clicking on ► Graphs ► Legacy Dialogs ► Pie. SPSS will then ask you to make one of three choices. For a standard pie chart, the first option (**Summaries for groups of cases**) is best. In the subsequent dialog box, all you need to do is enter the variable from which you want to make a pie chart in the box, titled **Define Slices by**. If we move *Gender* into the **Define Slices by** box, and click on **OK**, SPSS will show a pie chart similar to Fig. 5.14.

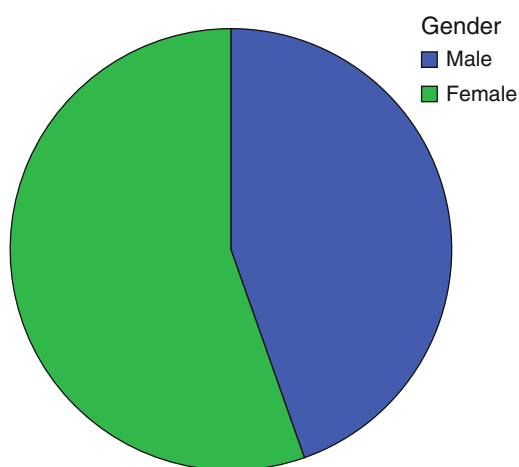


Fig. 5.14 A pie chart

Frequency Tables

A frequency table can be made by clicking on ► Analyze ► Descriptive Statistics ► Frequencies. You can enter as many variables as you want, as SPSS will make separate frequency tables for each variable. Then click on **OK**. If we use *Internet makes me happy*, SPSS will produce tables similar to Table 5.6. The first table includes just the number of observations, while the second includes the actual frequency table.

Table 5.6 A frequency table

Statistics

Internet makes me happy

N	Valid	435
	Missing	0

Internet makes me happy

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Completely disagree	47	10.8	10.8	10.8
	Disagree	104	23.9	23.9	34.7
	Somewhat disagree	25	5.7	5.7	40.5
	Neutral	104	23.9	23.9	64.4
	Somewhat agree	76	17.5	17.5	81.8
	Agree	58	13.3	13.3	95.2
	Completely agree	21	4.8	4.8	100.0
Total		435	100.0	100.0	

Bar Charts

Bar charts are made by clicking on ► Graphs ► Legacy Dialogs ► Bar. Next, SPSS will ask you to choose a type of bar chart (**Simple**, **Clustered**, or **Stacked**) and what

the data in the bar chart represent (**Summaries for groups of cases**, **Summaries of separate variables**, or **Values of individual cases**). For a simple bar chart, click on **Define** and then move the variable for which you want to create a bar chart in the **Category Axis** box. Under **Bars Represent**, you can indicate what the bars are. Choosing for **N of cases** or **% of Cases** results in a graph similar to a histogram, while choosing **Other statistic (e.g., mean)** can be helpful to summarize a variable's average values across different groups. The groups then need to be identified in the **Category Axis** and that which is summarized in the **Variable** box. For example, go to ► **Graphs** ► **Legacy Dialogs** ► **Bar**, then click on **Simple** and **Define**. Subsequently, enter *Internet makes me happy* under **Category axis**. SPSS will then show a graph similar to Fig. 5.15.

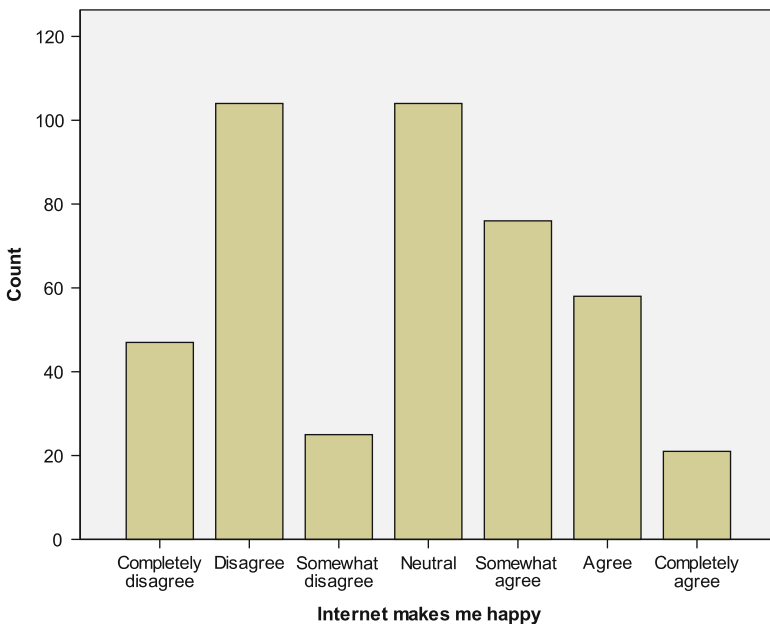
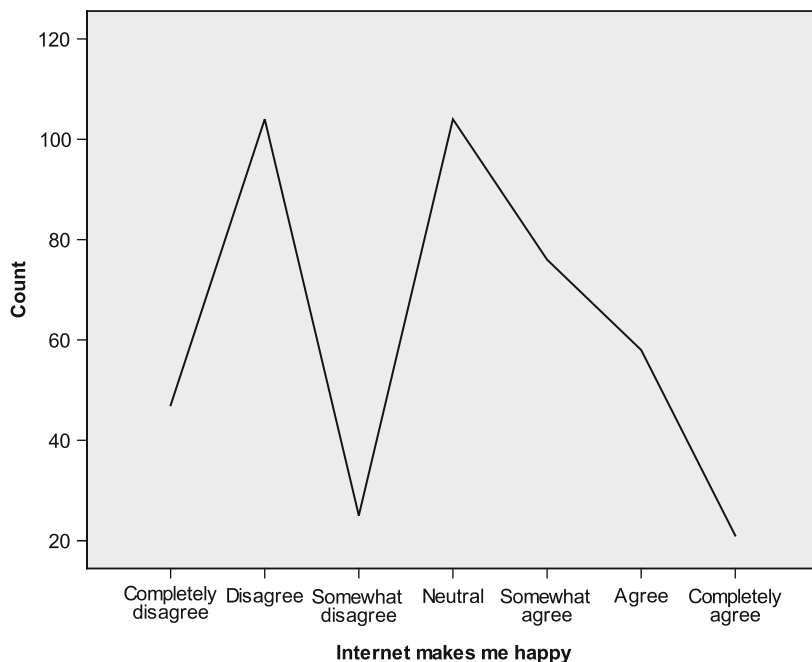


Fig. 5.15 A bar chart

### Line Charts

Line charts are made by clicking on ► **Graphs** ► **Legacy Dialogs** ► **Line**. SPSS will then ask you what type of line chart you want: a **Simple** line chart, a **Multiple** line chart or a line indicating a variable's values for different groups called **Drop-line**. You also need to indicate what the data need to represent: **Summaries for groups of cases** (used in most cases), **Summaries of separate variables**, or **Values of individual cases**. As an example, go to ► **Graphs** ► **Legacy Dialogs** ► **Line**, then

click on **Simple** and **Define**. Enter *Internet makes me happy* into the **Category Axis** box and click on **OK**. The resulting graph will be similar to Fig. 5.16.



**Fig. 5.16** A line chart

### *Calculating Univariate Descriptive Statistics in SPSS*

The mean, mode, and median are easily calculated in SPSS by going to ► Analyze ► Descriptive Statistics ► Frequencies. If you just need these three descriptives, uncheck the **Display Frequency tables** box. Under **Statistics**, you can then check **Mean, Median, and Mode**. Clicking on **Continue** and then **OK** will show you the requested statistics in the output window. Notice that if you ask SPSS to calculate descriptives, it will also display **Valid N (listwise)**. The value indicated behind this, are the number of observations for which we have full information available (i.e., no missing values).

You can also create a z-transformation under descriptive statistics. Go to ► Analyze ► Descriptive statistics ► Descriptives. Then click on **Save standardized values as variables**. For every variable you enter in the **Variable(s)** box, SPSS will create a new variable (including the prefix *z*), standardize the variable, and save it. In the **Label** column under **Variable View** it will even show the original value label preceded by *Zscore:* to indicate that you have transformed that particular variable.

Measures of dispersion are most easily created by going to Analyze ► Descriptive Statistics ► Descriptives. Enter the variables for which you want to calculate descriptives into the **Variable(s)** box. Under **Options**, you can check **Std. deviation**, **Minimum**, **Maximum**, and **Range**. You can also check **Mean** but no options are included to calculate the mode or median (for this, you have to go back to ► Analyze ► Descriptive Statistics ► Frequencies).

## Creating Bivariate Graphs and Tables in SPSS

### Scatter Plots

Scatter plots are made easily by going to ► Graphs ► Legacy Dialogs ► Scatter/Dot. Subsequently, SPSS shows five different options: a **Simple Scatter** plot, a **Matrix Scatter** plot (showing different scatter plots for different groups), a **Simple Dot** graph, an **Overlay Scatter** plot (showing different groups within one scatter plot), and the **3-D Scatter** plot (for relationships between three variables). As an example, go to ► Graphs ► Legacy Dialogs ► Scatter/Dot. Then click on **Simple Scatter** and **Define**. Enter *Internet makes me happy* under **Y Axis** and *Income per person in USD* under **X Axis**. The resulting graph will look like Fig. 5.17.

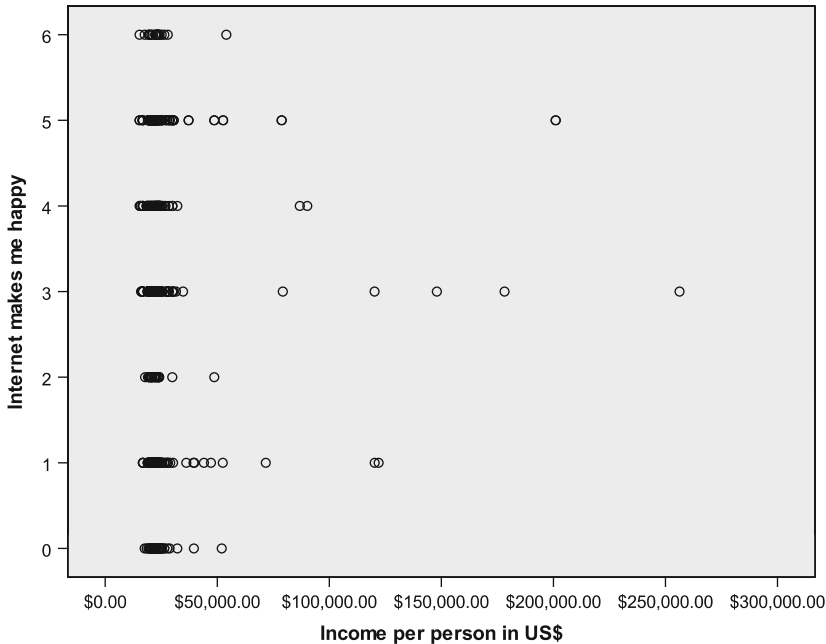


Fig. 5.17 A simple scatter plot

Crosstabs

Crosstabs are useful to describe data, particularly if your variables are nominally or ordinally scaled. Crosstabs are made by going to ► Analyze ► Descriptive Statistics ► Crosstabs. You can enter multiple variables under **Row(s)** and **Column(s)**, but crosstabs are easiest to interpret if you enter just one variable under **Row(s)** and one under **Column(s)**. Try making a crosstab by going to ► Analyze ► Descriptive Statistics ► Crosstabs. Under **Row(s)** you can enter *Internet makes me happy* and *Gender* under **Column(s)**. If you then click on **OK**, SPSS produces a table (see Table 5.7).

Table 5.7 Example of a crosstab

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Internet makes me happy * Gender	435	100.0%	0	.0%	435	100.0%

Internet makes me happy \* Gender Crosstabulation

Count

		Gender		Total
		Male	Female	
Internet makes me happy	Completely disagree	24	23	47
	Disagree	46	58	104
	Somewhat disagree	11	14	25
	Neutral	54	50	104
	Somewhat agree	33	43	76
	Agree	19	39	58
	Completely agree	7	14	21
Total		194	241	435



Calculating Bivariate Descriptive Statistics in SPSS

Correlations

In SPSS, we can calculate correlations by going to ► Analyze ► Correlate ► Bivariate. In the dialog box that pops up, we can select whether we want SPSS to calculate Pearson’s or Spearman’s correlation coefficient. Table 5.8 is an example of a correlation matrix produced in SPSS.

Table 5.8 Correlation matrix produced in SPSS

Correlations		Internet makes me happy	I enjoy Internet
Internet makes me happy	Pearson Correlation	1	.466 **
	Sig. (2-tailed)		.000
	N	435	435
I enjoy Internet	Pearson Correlation	.466 **	1
	Sig. (2-tailed)	.000	
	N	435	435

\*\*. Correlation is significant at the 0.01 level (2-tailed).

The correlation matrix in Table 5.8 shows the correlation between two variables labeled as *Internet makes me happy* and *I enjoy Internet*. The correlation is 0.466. SPSS also indicates the number of observations used to calculate each correlation (435 observations) and indicates the significance, indicated by *Sig. (2-tailed)*. We discuss what this means in Chap. 6.

Case Study

The UK chocolate market achieved annual sales of 6.40 billion GBP in 2008. Six sub-categories of chocolates are used to identify the different chocolate segments: boxed chocolate, molded bars, seasonal chocolate, countlines, straightlines, and “other”. To understand the UK chocolate market for molded chocolate bars, we have a dataset (*chocolate.sav*) that includes a large supermarket’s weekly sales of 100 g

## Chocolate Confectionery Industry Insights



<http://www.globalbusinessinsights.com/content/rbcg0125m.pdf>

molded chocolate bars from January 2009 onwards. This data file can be downloaded on the book's website (☞ Web Appendix → Chap. 5). This file contains a set of variables. Once you have opened the dataset and clicked on **Variable View**, you will see the set of variables we discuss next.

The first variable is *week*, indicating the week of the year and starts with Week 1 of January 2009. The last observation for 2009 ends with observation 52, but the variable continues to count onwards for 16 weeks in 2010. The next variable is *sales*, which indicates the weekly sales of 100 g Cadbury bars in GBP. Next, four price variables are included, *price1-price4*, which indicate the price of Cadbury, Nestlé, Guylian, and Milka in GBP. Next, *advertising1-advertising4* indicate the amount of GBP the supermarket spent on advertising each product during that week. A subsequent block of variables, *pop1-pop4*, indicate whether the products were promoted in the supermarket using a point of purchase advertising. This variable is measured as yes/no. Variables *promo1-promo4* indicate whether the product was put at the end of the supermarket aisle – where it is more noticeable. Lastly, *temperature* indicates the weekly average temperature in degrees Celsius.

You have been tasked to provide descriptive statistics for a client, using this available dataset. To help you with this task, the client has prepared a number of questions:

1. Do Cadbury's chocolate sales vary substantially across different weeks? When are Cadbury's sales at their highest? Please create an appropriate graph to illustrate any patterns.
2. Please tabulate point-of-purchase advertising for Cadbury against point-of-purchase advertising for Nestlé. Also create a few further crosstabs. What are the implications of these crosstabs?
3. How do Cadbury's sales relate to the price of Cadbury? What is the strength of the relationship?
4. Which descriptive statistics are appropriate to describe the usage of advertising? Which are appropriate to describe point-of-purchase advertising?

## Questions

1. Imagine you are given a dataset on car sales in different regions and are asked to calculate descriptive statistics. How would you set up the analysis procedure?
2. What summary statistics could best be used to describe the change in profits over the last five years? What types of descriptives work best to determine the market shares of five different types of insurance providers? Should we use just one or multiple descriptives?
3. What information do we need to determine if a case is an outlier? What are the benefits and drawbacks of deleting outliers?
4. Download the US census codebook for 2007 at <http://usa.ipums.org/usa/code-books/DataDict2007.pdf>. Is this codebook clear? What do you think of its structure?

## Further Readings

Cohen J, Cohen P, West SG, Aiken LS (2003) Applied multiple regression/correlation analysis for the behavioral sciences, 3rd edn. Lawrence Erlbaum Associates, Mahwah, NJ

*This is the seminal book on correlation (and regression) analysis which provides a detailed overview into this field. It is aimed at experienced researchers.*

Field A (2009) Discovering statistics using SPSS, 3rd edn. Sage, London

*In Chap. 6 of his book, Andy Field provides a very thorough introduction into the principles of correlation analysis from an application-oriented point of view.*

Hair JF Jr, Black WC, Babin BJ, Anderson RE (2010) Multivariate data analysis. A global perspective, 7th edn. Prentice-Hall, Upper Saddle River, NJ

*A widely used book on multivariate data analysis. Chapter 2 of this book discusses missing data issues.*

Levesque R, Programming and Data Management for IBM SPSS Statistics 18. A Guide for PASW Statistics and SAS Users. Chicago, SPSS, Inc.

*An advanced book demonstrating how to manage data in SPSS. Can be downloaded for free at <http://www.spss.com/sites/dm-book/resources/PASW-Statistics-18-DM-Book.pdf>*

SPSS (PASW) On-Line Training Workshop at <http://calcnnet.mth.cmich.edu/org/spss/index.htm>

*An excellent website on which you can find movies describing different aspects of SPSS.*

SPSS Learning Modules at <http://www.ats.ucla.edu/stat/spss/modules/>

*On this website further data organization issues are discussed. This is useful if you already have a dataset, but cannot use it because of the way the data are organized.*

SticiGui at <http://www.stat.berkeley.edu/~stark/Java/Html/Correlation.htm>

*This website interactively demonstrates how strong correlations are for different datasets.*

## References

- Cohen J (1988) Statistical power analysis for the behavioral sciences. 2nd edn. Lawrence Erlbaum Associates, Hillsdale, NJ
- Collier J (2010) Using SPSS syntax: a beginner's guide. Sage, Thousand Oaks, CA
- Dillman DA (2008) Internet, mail, and mixed-mode surveys: the tailored design method. Wiley, Hoboken, NJ
- Gladwell M (2008) Outliers: the story of success. Little, Brown, and Company, New York, NY
- Hair JF Jr, Black WC, Babin BJ, Anderson RE (2010) Multivariate data analysis. Pearson, Upper Saddle River, NJ