Descriptive Statistics

5

Keywords

Acquiescence • Aggregation • Bar chart • Bivariate statistics • Box plot • Codebook • Construct score • Correlation • Covariance • Crosstabs • Data entry errors • Dummy variables • Extreme response styles • Frequency table • Histogram • Inconsistent answers • Index • Interquartile range • Interviewer fraud • Item non-response • Line chart • Listwise deletion • Little's MCAR test • Log transformation • Mean • Measures of centrality • Measures of dispersion • Median • Middle response styles • Missing (completely) at random • Missing data • Multiple imputation • Non-random missing • Outliers • Pie chart • Range • Range standardization • Reverse-scaled items • Scale transformation • Scatter plot • Skewed data • Stata • Standard deviation • Standardizing variables • Straight-lining • Survey non-response • Suspicious response patterns • Transforming data • Univariate statistics • Variable respecification • Variance • Workflow • *z*-standardization

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 Stata.

E. Mooi et al., *Market Research*, Springer Texts in Business and Economics,
DOI 10.1007/978-981-10-5218-7_5

5.1 The Workflow of Data

Market research projects involving data become more efficient and effective if they have a proper **workflow of data**, which is a strategy to keep track of the entering, cleaning, describing, and transforming of data. These data may have been collected through surveys or may be secondary data (Chap. 3). Entering, cleaning, and analyzing bits of data haphazardly is not a good strategy, since it increases the likelihood of making mistakes and makes it hard to replicate results. Moreover, without a good data workflow, it becomes hard to document the research process and to 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 a good workflow increases the risk of duplicating work or even losing data. In Fig. 5.1, we show the steps required to create and describe a dataset after the data have been collected. We subsequently discuss each step in greater detail.

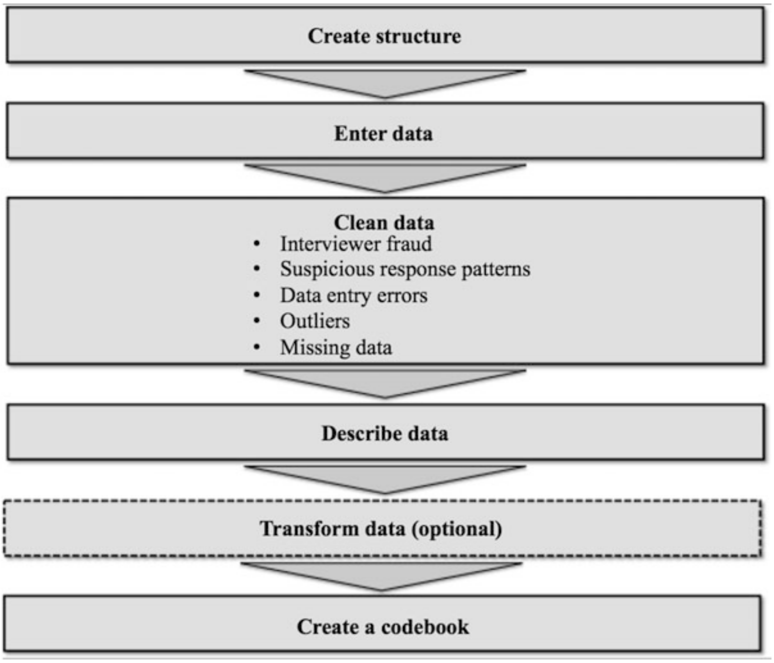


Fig. 5.1 The workflow of data

5.2 Create Structure

The basic idea of setting up a good workflow is that good planning saves the researcher 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 by means of Windows Explorer or macOS Finder. This directory should have subdirectories for at least: (1) the data files, (2) commands, (3) a temporary directory, and (4) related files; that is, a directory with files that are directly related to a project, such as the survey used to collect the data.[1]

In Table 5.1, we show an example of a directory structure. Within the main directory, there are four subdirectories, each with distinct files. Notice that in the **Data files** subdirectory, 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 or other archive file, it is stored and unlikely to be modified, but can be easily opened if the working file is accidentally overwritten or deleted. In the **Data files** subdirectories, we distinguish between two files with the suffix **rev1** and **rev2**. We use **rev** (abbreviation of revision), but you however, can choose another file name

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

Directory name	Subdirectory name	Example file names
Oddjob	Data files	oddjob.dta
		oddjob.zip
		oddjob rev1.dta
		oddjob rev2.dta
	Command files	Missing data analysis.do
		Descriptives.do
		Factor analysis.do
		Regression analysis.do
	Temporary files	Missing data analysis rev1.smcl
		Descriptives rev1.smcl
		Factor analysis rev1.smcl
		Regression analysis rev1.smcl
	Related files	Codebook.docx
		Survey.pdf
		Initial findings–presentation to client.pptx
		Findings–presentation to client.pptx
		Recommendations rev1.docx
		Recommendations rev2.docx

[1] Alternatively, you could also choose one of the many control system versions, including Subversion, Git, and Mecurial, which enable simple branching and project management. These systems work well with version control in centralized and in distributed environments.

as long as it clearly indicates the revision on which you are working. In the **Command files** subdirectory we store all commands that were used to manage our data. These commands may relate to different project phases, including a missing data analysis, descriptives, factor analysis, and other methods used over the course of the project. As the name indicates, **Temporary files** serve as intermediary files that are kept until the final data or command files are established, after which they are removed. Finally, in the **Related Files** subdirectory, we have a codebook (more on this later), the survey, two presentations, and two documents containing recommendations.

Another aspect of creating a structure is setting up the variables for your study properly. This involves making decisions on the following elements:

- variable names,
- variable labels,
- data type, and
- coding of variables.

The variable names should be clear and short so that they can be read in the dialog boxes. For example, if you have three questions on product satisfaction, three on loyalty, and several descriptors (age and gender), you could code these variables as *satisfaction1*, *satisfaction2*, *satisfaction3*, *loyalty1*, *loyalty2*, *loyalty3*, *age*, and *gender*.

In Stata, and most other statistical software programs, you can include *variable labels* that describe what each variable denotes. The description generally includes the original question if the data were collected by means of surveys. Another point to consider is *variable coding*. Coding means assigning values to a variable. When collecting quantitative data, the task is relatively easy; we use values that correspond with the answers for Likert and semantic differential scales (see Chap. 4). For example, when using a 7-point Likert scale, responses can be coded as 1–7 or as 0–6 (with 0 being the most negative and 6 being the most positive response). Open-ended questions (qualitative data) require more effort, usually involving a three-step process. First, we collect all the responses. In the second step, we group these responses. Determining the number of groups and the group to which a response belongs is the major challenge in this step. Two or three market researchers usually code the responses independently to prevent the process from becoming too subjective and thereafter discuss the differences that may arise. The third step is providing a value for each group. Stata can perform such analyses with the help of the additional software package *Wordstat* (https://provalisresearch.com/products/content-analysis-software/wordstat-for-stata/). Please see Krippendorff (2012) for more details about coding qualitative variables.

Once a system has been 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 as a safeguard). If you are working alone or for a small company, you are probably responsible for this. You should save your most recent and second most recent version of your file on a separate drive and have multiple copies of your entire drive! Always keep at least two copies and never keep both backups in the same place, because you could still lose all your work through theft, fire, or an accident! You can use cloud storage services, such as Dropbox, Google Drive, or Microsoft's OneDrive for small projects to prevent loss. Always read the terms of the cloud storage services carefully to determine whether your data's privacy is guaranteed.

5.3 Enter Data

How do we enter survey or experimental data into a dataset? Specialized software is often used for large datasets, or datasets created by professional firms. For example, Epidata (http://www.epidata.dk, freely downloadable) is frequently used to enter data from paper-based surveys, Entryware's mobile survey (http://www.techneos.com) to enter data from personal intercepts or face-to-face interviewing, and Voxco's Interviewer CATI for telephone interviewing. Stata has no dedicated data entry platform. It does, however, facilitate the use of StatTransfer (http://www.stattransfer.com), a software program designed to simplify the transfer of statistical data between many software packages, including Stata, SPSS, Excel, and SAS.

Such software may not be available for smaller projects, in which case data should be entered directly into Stata. A significant drawback of direct data entry is the risk of typing errors, for which Stata cannot check. 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. The software also allows for using multiple typists when very large amounts of data need to be entered simultaneously (note that Epidata can export directly to Stata).

5.4 Clean Data

Cleaning data is the next step in the workflow. It requires checking for:

– interviewer fraud,
– suspicious response patterns,
– data entry errors,
– outliers, and
– missing data.

These issues require researchers to make decisions very carefully. In the following, we discuss each issue in greater detail.

5.4.1 Interviewer Fraud

Interviewer fraud is a difficult and serious issue. It ranges from interviewers "helping" respondents provide answers to entire surveys being falsified. Interviewer fraud often leads to incorrect results. Fortunately, we can avoid and detect interviewer fraud in various ways. First, never base interviewers' compensation on the number of completed responses they submit. Second, check and control for discrepancies in respondent selection and responses. If multiple interviewers were used, each of whom collected a reasonably large number of responses ($n > 100$), a selection of the respondents should be similar. This means that the average responses obtained should also be similar. In Chap. 6 we will discuss techniques to test this. Third, if possible, contact a random number of respondents afterwards for their feedback on the survey. If a substantial number of people claim they were not interviewed, interviewer fraud is likely. Furthermore, if people were previously interviewed on a similar subject, the factual variables collected, such as their gender, should not change (or no more than a trivial percentage), while variables such as a respondent's age and highest education level should only move up. We can check this by means of descriptive statistics. If substantial interviewer fraud is suspected, the data should be discarded. You should check for interviewer fraud during the data collection process to safeguard the quality of data collection and minimize the risk of having to discard the data in the end.

5.4.2 Suspicious Response Patterns

Before analyzing data, we need to identify **suspicious response patterns**. There are two types of response patterns we need to look for:

- straight-lining, and
- inconsistent answers.

Straight-lining occurs when a respondent marks the same response in almost all the items. For example, if a 7-point scale is used to obtain answers and the response pattern is 4 (the middle response), or if the respondent selects only 1s, or only 7s in all the items. A common way of identifying straight-lining is by including one or more **reverse-scaled items** in a survey (see Chap. 4). Reverse-scaled means that the way the question, statement (when using a Likert scale), or word pair (when using a semantic differential scale) is reversed compared to the other items in the set. Box 5.1 shows an example of a four-item scale for measuring consumers' attitude toward the brand (e.g., Sarstedt et al. 2016) with one reverse-scaled item printed in bold. By evaluating the response patterns, we can differentiate between those respondents who are not consistent for the sake of consistency and those who are merely mindlessly consistent. Note, however, that this only applies if respondents do not tick the middle option. Straight-lining is very common, especially in web surveys where respondents generally pay less attention to the answers. Likewise,

Box 5.1 An Example of a Scale with Reverse-Scaled Items (in Bold)

Please rate the *brand* in the advertisement on the following dimensions:

Dislike	1	2	3	4	5	6	7	Like
Unpleasant	1	2	3	4	5	6	7	Pleasant
Good	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Bad**
Expensive	1	2	3	4	5	6	7	Inexpensive
Useless	1	2	3	4	5	6	7	Useful

long surveys and those with many similarly worded items trigger straight-lining (Drolet and Morrison 2001). An alternative is to note potential straight-lined responses and include this as a separate category in the subsequent statistical analyses. This step avoids the need to reduce the sample and indicates the size and direction of any bias.

However, straight-lining can also be the result of *culture-specific response styles*. For example, respondents from different cultures have different tendencies regarding selecting the mid points (**middle response styles**) or the end points of a response scale (**extreme response styles**). Similarly, respondents from different cultures have different tendencies regarding agreeing with statements, regardless of the item content; this tendency is also referred to as **acquiescence** (Baumgartner and Steenkamp 2001). For example, respondents from Spanish-speaking countries tend to show higher extreme response styles and high acquiescence, while East Asian (Japanese and Chinese) respondents show a relatively high level of middle response style. Within Europe, the Greeks stand out as having the highest level of acquiescence and a tendency towards an extreme response style. Harzing (2005) and Johnson et al. (2005) provide reviews of culture effects on response behavior.

Inconsistent answers also need to be addressed before analyzing your data. Many surveys start with one or more screening questions. The purpose of a screening question is to ensure that only individuals who meet the prescribed criteria complete the survey. For example, a survey of mobile phone users may screen for individuals who own an iPhone. If an individual indicates that he/she does not have an iPhone, this respondent should be removed from the dataset.

Surveys often ask the same question with slight variations, especially when reflective measures (see Box 3.1 in Chap. 3) are used. If a respondent gives a different answer to very similar questions, this may raise a red flag and could suggest that the respondent did not read the questions closely, or simply marked answers randomly to complete the survey as quickly as possible.

5.4.3 Data Entry Errors

When data are entered manually, **data entry errors** occur routinely. Fortunately, such errors are easy to spot if they happen outside the variable's range. That is, if an item is measured using a 7-point scale, the lowest value should be 1 (or 0) and the highest 7 (or 6). We can check if this is true by using descriptive statistics (minimum, maximum, and range; see next section). 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 this specific observation for this specific variable.

More subtle errors—for example, incorrectly entering a score of 4 as 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 survey. We do, of course, expect a small number of errors (below 1%). If many data entry errors occur, the dataset should be entered again.

Manual double data entry is another method to detect data entry errors. That is, once the data has been entered manually, a second data checker enters the same data a second time and the two separate entries are compared to ensure they match. Entries that deviate from one another or values that fall outside the expected range of the scales (e.g., 7-point Likert scales should have values that fall within this range) are then indicative of data entry errors (Barchard and Verenikina 2013). Various studies reveal that—although double data entry is more laborious and expensive—it still detects errors better than single data entry (Barchard and Pace 2011; Paulsen et al. 2012).

5.4.4 Outliers

Data often contain **outliers**, which are values situated far from all the other observations that may influence results substantially. For example, if we compare the average income of 20 households, we may find that the incomes range between $20,000 and $100,000, with the average being $45,000. If we considered an additional household with an income of, say, $1 million, this would increase the average substantially.

Malcolm Gladwell's (2008) book "Outliers: The Story of Success" is an entertaining study of how some people became exceptionally successful (outliers).

5.4.4.1 Types of Outliers

Outliers must be interpreted in the context of the study and this interpretation should be based on the types of information they provide. Depending on the source of their uniqueness, outliers can be classified into three categories:

- 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 US dollars, some respondents may just indicate theirs in US dollars (not thousands). Obviously, there is a substantial difference between $30 and $30,000! Moreover, (as discussed before) data entry errors occur frequently. Outliers produced by data collection or entry errors should be deleted, or we need to determine the correct values by, for example, returning to the respondents.
- A second type of outlier occurs because exceptionally high or low values are a part of reality. While such observations can influence results significantly, they are sometimes highly important for researchers, because the characteristics of outliers can be insightful. Think, for example, of extremely successful companies, or users with specific needs long before most of the relevant marketplace also needs them (i.e., lead users). Deleting such outliers is not appropriate, but the impact that they have on the results must be discussed.
- A third type of outlier occurs when *combinations* of values are exceptionally rare. For example, if we look at income and expenditure on holidays, we may find someone who earns $1,000,000 and spends $500,000 of his/her income on holidays. Such combinations are unique and have a very strong impact on the results (particularly the correlations that we discuss later in this chapter). In such situations, the outlier should be retained, unless specific evidence suggests that it is not a valid member of the population under study. It is very useful to flag such outliers and discuss their impact on the results.

5.4.4.2 Detecting Outliers

In a simple form, outliers can be detected using univariate or bivariate graphs and statistics.[2] When searching for outliers, we need to use multiple approaches to ensure that we detect all the observations that can be classified as outliers. In the following, we discuss both routes to outlier detection:

Univariate Detection

The univariate detection of outliers examines the distribution of observations of each variable with the aim of identifying those cases falling outside the range of the "usual" values. In other words, finding outliers means finding observations with very low or very high variable values. This can be achieved by calculating the

[2] There are multivariate techniques that consider three, or more, variables simultaneously in order to detect outliers. See Hair et al. (2010) for an introduction, and Agarwal (2013) for a more detailed methodological discussion.

minimum and maximum value of each variable, as well as the range. Another useful option for detecting outliers is by means of box plots, which are a means of visualizing the distribution of a variable and pinpointing those observations that fall outside the range of the "usual" values. We introduce the above statistics and box plots in greater detail in the *Describe Data* section.

It is important to recognize that there will always be observations with exceptional values in one or more variables. However, we should strive to identify outliers that impact the presented results.

Bivariate Detection

We can also examine pairs of variables to identify observations whose combinations of variables are exceptionally rare. This is done by using a *scatter plot*, which plots all observations in a graph where the x-axis represents the first variable and the y-axis the second (usually *dependent*) variable (see the *Describe Data* section). Observations that fall markedly outside the range of the other observations will show as isolated points in the scatter plot.

A drawback of this approach is the number of scatter plots that we need to draw. For example, with 10 variables, we need to draw 45 scatter plots to map all possible combinations of variables! Consequently, we should limit the analysis to only a few relationships, such as those between a dependent and independent variable in a regression. Scatterplots with large numbers of observations are often problematic when we wish to detect outliers, as there is usually not just one dot, or a few isolated dots, just a cloud of observations where it is difficult to determine a cutoff point.

5.4.4.3 Dealing with Outliers

In a final step, we need to decide whether to delete or retain outliers, which should be based on whether we have an explanation for their occurrence. If there is an explanation (e.g., because some exceptionally wealthy people were included in the sample), outliers are typically retained, because they are part of the population. However, their impact on the analysis results should be carefully evaluated. That is, one should run an analysis with and without the outliers to assess if they influence the results. If the outliers are due to a data collection or entry error, they should be deleted. If there is no clear explanation, outliers should be retained.

5.4.5 Missing Data

Market researchers often have to deal with **missing data**. There are two levels at which missing data occur:

- Entire surveys are missing (survey non-response).
- Respondents have not answered all the items (item non-response).

Survey non-response (also referred to as *unit non-response*) occurs when entire surveys are missing. Survey non-response is very common and regularly only

5–25% of respondents fill out surveys. Although higher percentages are possible, they are not the norm in one-shot surveys. Issues such as inaccurate address lists, a lack of interest and time, people confusing market research with selling, privacy issues, and respondent fatigue also lead to dropping response rates. The issue of survey response is best solved by designing proper surveys and survey procedures (see Box 4.5 in Chap. 4 for suggestions).

Item non-response occurs when respondents do not provide answers to certain questions. There are different forms of missingness, including people not filling out or refusing to answer questions. Item non-response is common and 2–10% of questions usually remain unanswered. However, this number greatly depends on factors, such as the subject matter, the length of the questionnaire, and the method of administration. Non-response can be much higher in respect of questions that people consider sensitive and varies from country to country. In some countries, for instance, reporting incomes is a sensitive issue.

The key issue with item non-response is the type of pattern that the missing data follow. Do the missing values occur randomly, or is there some type of underlying system?[3] Once we have identified the type of missing data, we need to decide how to treat them. Figure 5.2 illustrates the process of missing data treatment, which we will discuss next.

5.4.5.1 The Three Types of Missing Data: Paradise, Purgatory, and Hell

We generally distinguish between three types of missing data:

– missing completely at random (“paradise”),
– missing at random (“purgatory”), and
– non-random missing (“hell”).

Data are **missing completely at random** (**MCAR**) when the probability of data being missing is unrelated to any other measured variable and is unrelated to the variable with missing values. MCAR data thus occurs when there is no systematic reason for certain data points being missing. For example, MCAR may happen if the Internet server hosting the web survey broke down temporarily for a random reason. Why is MCAR paradise? When data are MCAR, observations with missing data are indistinguishable from those with complete data. If this is the case and little data are missing (typically less than 10% in each variable) listwise deletion can be used. Listwise deletion means that we only analyze complete cases; in most statistical software, such as Stata, this is a default option. Note that this default option in Stata only works when estimating models and only applies to the variables included in the model. When more than 10% of the data are missing, we can use multiple imputation (Eekhout et al. 2014), a more complex approach to missing data treatment that we discuss in the section *Dealing with Missing Data*.

[3] For more information on missing data, see https://www.iriseekhout.com

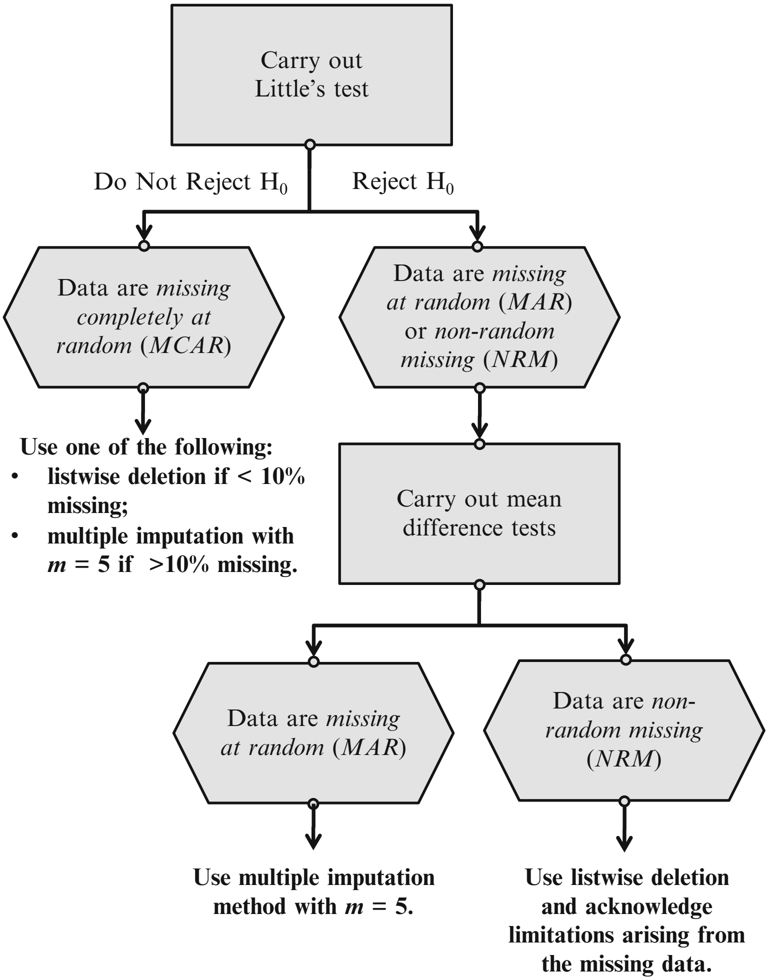

Carry out Little's test
Do Not Reject H_0
Reject H_0
Data are *missing completely at random* (*MCAR*)
Data are *missing at random* (*MAR*) or *non-random missing* (*NRM*)
Use one of the following:
• listwise deletion if < 10% missing;
• multiple imputation with *m* = 5 if >10% missing.
Carry out mean difference tests
Data are *missing at random* (*MAR*)
Data are *non-random missing* (*NRM*)
Use multiple imputation method with *m* = 5.
Use listwise deletion and acknowledge limitations arising from the missing data.

Fig. 5.2 Treating missing data

Unfortunately, data are rarely MCAR. If a missing data point (e.g., x_i) is unrelated to the observed value of x_i, but depends on another observed variable, we consider the data **missing at random** (**MAR**). In this case, the probability that the data point is missing varies from respondent to respondent. The term MAR is unfortunate, because many people confuse it with MCAR; however, the label has stuck. An example of MAR is when women are less likely to reveal their income. That is, the probability of missing data depends on the gender and not on the income. Why is MAR purgatory? When data are MAR, the missing value pattern is not random, but this can be handled by more sophisticated missing data techniques such as multiple imputation techniques. In the ↓ Web Appendix (→ Downloads), we will illustrate how to impute a dataset with missing observations.

Lastly, data are **non-random missing** (**NRM**) when the probability that a data point (e.g., x_i) is missing depends on the variable x and on other unobserved factors. For example, very affluent and poor people are less likely to indicate their income. Thus, the missing income values depend on the income variable, but also on other unobserved factors that inhibit the respondents from reporting their incomes. This is the most severe type of missing data ("hell"), as even sophisticated missing data techniques do not provide satisfactory solutions. Thus, any result based on NRM data should be considered with caution. NRM data can best be prevented by extensive pretesting and consultations with experts to avoid surveys that cause problematic response behavior. For example, we could use income categories instead of querying the respondents' income directly, or we could simply omit the income variable.

A visualization of these three missingness mechanisms can be found under https://iriseekhout.shinyapps.io/MissingMechanisms/

5.4.5.2 Testing for the Type of Missing Data

When dealing with missing data, we must ascertain the missing data's type. If the dataset is small, we can browse through the data for obvious nonresponse patterns. However, missing data patterns become more difficult to identify with an increasing sample size and number of variables. Similarly, when we have few observations, patterns should be difficult to spot. In these cases, we should use one (or both) of the following diagnostic tests to identify missing data patterns:

- Little's MCAR test, and
- mean difference tests.

Little's MCAR test (Little 1998) analyzes the pattern of the missing data by comparing the observed data with the pattern expected if the data were randomly missing. If the test indicates no significant differences between the two patterns, the missing data can be classified as MCAR. Put differently, the null hypothesis is that the data are MCAR. Thus,

- if we do **not** reject the null hypothesis, we assume that the data are MCAR, and
- if we reject the null hypothesis, the data are either MAR or NRM.

If the data cannot be assumed to be MCAR, we need to test whether the missing pattern is caused by another variable in the dataset by using the procedures discussed in Chap. 6.

Looking at group means and their differences can also reveal missing data problems. For example, we can run a two independent samples *t*-test to explore whether there is a significant difference in the mean of a continuous variable (e.g., income) between the group with missing values and the group without missing

Table 5.2 Example of response issues

	Low income	Medium income	High income
Response	65	95	70
Non-response	35	5	30
$N = 300$			

values. In respect of nominal or ordinal variables, we could tabulate the occurrence of non-responses against different groups' responses. If we put the (categorical) variable about which we have concerns in one column of a table (e.g., income category), and the number of (non-)responses in another, we obtain a table similar to Table 5.2.

Using the χ^2*-test* (pronounced as *chi-square*), which we discuss under nonparametric tests in the ⭳ Web Appendix (→ Downloads), we can test if there is a significant relationship between the respondents' (non-)responses in respect of a certain variable and their income. In this example, the test indicates that there is a significant relationship between the respondents' income and the (non-)response behavior in respect of another variable, supporting the assumption that the data are MAR. We illustrate Little's MCAR test, together with the missing data analysis and imputation procedures in this chapter's appendix ⭳ Web Appendix (→ Downloads).

5.4.5.3 Dealing with Missing Data

Research has suggested a broad range of approaches for dealing with missing data. We discuss the listwise deletion and the multiple imputation method.

Listwise deletion uses only those cases with complete responses in respect of all the variables considered in the analysis. If any of the variables used have missing values, the observation is omitted from the computation. If many observations have some missing responses, this decreases the usable sample size substantially and hypotheses are tested with less power (the power of a statistical test is discussed in Chap. 6).

Multiple imputation is a more complex approach to missing data treatment (Rubin 1987; Carpenter and Kenward 2013). It is a simulation-based statistical technique that facilitates inference by replacing missing observations with a set of possible values (as opposed to a single value) representing the uncertainty about the missing data's true value (Schafer 1997). The technique involves three steps. First, the missing values are replaced by a set of plausible values not once, but *m* times (e.g., five times). This procedure yields *m* imputed datasets, each of which reflects the uncertainty about the missing data's correct value (Schafer 1997). Second, each of the imputed *m* datasets are analyzed separately by means of standard data methods. Third and finally, the imputed results from all *m* datasets (with imputed values) are combined into a single multiple-imputation dataset to produce statistical inferences with valid confidence intervals. This is necessary to reflect the uncertainty related to the missing values. According to the literature, deciding on the number of imputations, *m*, can be very challenging, especially when the patterns of the missing data are unclear. As a rule of thumb, an *m* of at least 5 should be

Table 5.3 Data cleaning issues and how to deal with them

Problem	Action
Interviewer fraud	– Check with respondents whether they were interviewed and correlate with previous data if available.
Suspicious response patterns	– Check for straight lining. – Include reverse-scaled items. – Consider removing the cases with straight-lined responses. – Consider cultural differences in response behavior (middle and extreme response styles, acquiescence). – Check for inconsistencies in response behavior.
Data entry errors	– Use descriptive statistics (minimum, maximum, range) to check for obvious data entry errors. – Compare a subset of surveys to the dataset to check for inconsistencies.
Outliers	– Identify outliers by means of univariate descriptive statistics (minimum, maximum, range), box plots, and scatter plots. – Outliers are usually retained unless they: . . . are a result of data entry errors, . . . do not fit the objective of the research, or . . . influence the results severely (but report results with and without outliers for transparency).
Missing data	– Check the type of missing data by running Little's MCAR test and, if necessary, mean differences tests. – When the data are MCAR, use either listwise deletion or the multiple imputation method with an *m* of 5. – When the data are MAR, use the multiple imputation method with an *m* of 5. – When the data are NRM, use listwise deletion and acknowledge the limitations arising from the missing data.

sufficient to obtain valid inferences (Rubin 1987; White et al. 2011). Additional information about the multiple imputation techniques available when using Stata can be found in the ⭳ Web Appendix (→ Downloads).

Now that we have briefly reviewed the most common approaches for handling missing data, there is still one unanswered question: Which one should you use? As shown in Fig. 5.2, if the data are MCAR, listwise deletion is recommended (Graham 2012) when the missingness is less than 10% and multiple imputation when this is greater than 10%. When the data are not MCAR but MAR, listwise deletion yields biased results. You should therefore use the multiple imputation method with an *m* of 5 (White et al. 2011). Finally, when the data are NRM, the multiple imputation method provides inaccurate results. Consequently, you should choose listwise deletion and acknowledge the limitations arising from the missing data. Table 5.3 summarizes the data cleaning issues discussed in this section.

5.5 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 in terms of the relationship between two variables (bivariate descriptives). We further divide univariate and bivariate descriptives into graphs and tables, as well as statistics.

The choice between the two depends on the information we want to convey. Graphs and tables can often tell a non-technical person a great deal. 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. 5.3.

5.5.1 Univariate Graphs and Tables

In this section, we discuss the most common *univariate graphs* and *univariate tables*:

- bar chart,
- histogram,
- box plot,
- pie chart, and the
- frequency table.

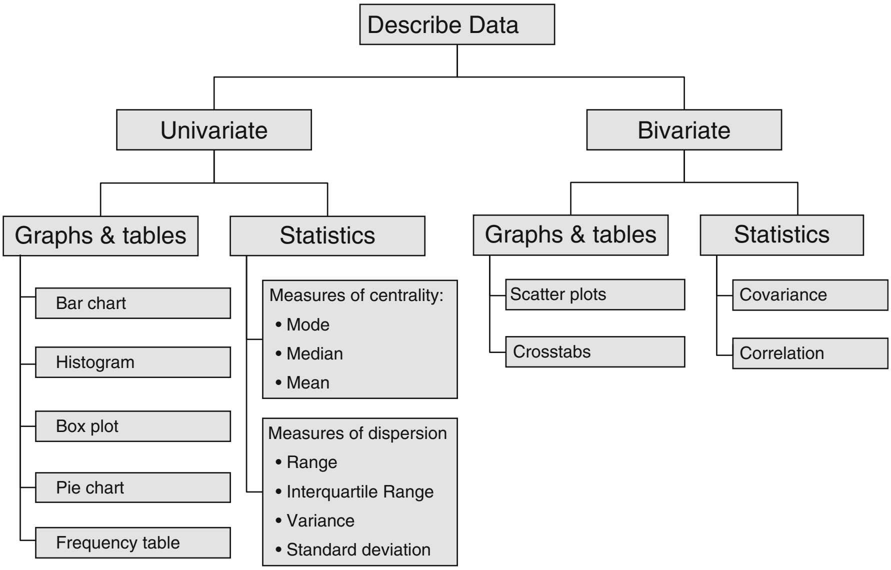
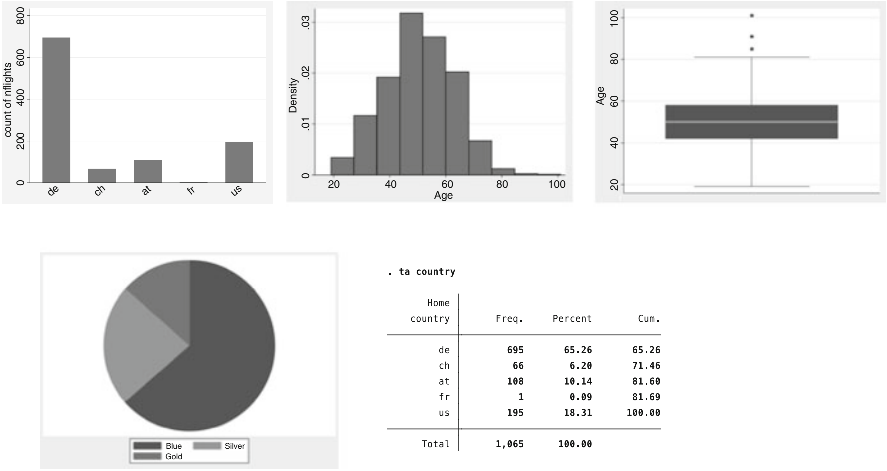


Fig. 5.3 The different types of descriptive statistics



```
. ta country

       Home |
    country |      Freq.     Percent        Cum.
------------+-----------------------------------
         de |        695       65.26       65.26
         ch |         66        6.20       71.46
         at |        108       10.14       81.60
         fr |          1        0.09       81.69
         us |        195       18.31      100.00
------------+-----------------------------------
      Total |      1,065      100.00
```

Fig. 5.4 From *top left* to *bottom right*; the bar chart, histogram, box plot, pie chart, and frequency table

Figure 5.4 draws on these different types of charts and tables to provide information on the characteristics of travelers taken from the Oddjob Airways dataset that we use throughout this book.

A **bar chart** (Fig. 5.4 top left) is a graphical representation of a single categorical variable indicating each category's frequency of occurrence. However, each bar's height can also represent other indices, such as centrality measures or the dispersion of different data groups (see next section). Bar charts are primarily useful for describing nominal or ordinal variables. Histograms should be used for interval or ratio-scaled variables.

A **histogram** (Fig. 5.4 top middle) is a graph that shows how frequently categories made from a continuous variable occur. Differing from the bar chart, the variable categories on the *x*-axis are divided into (non-overlapping) classes of equal width. For example, if you create a histogram for the variable *age*, you can use classes of 21–30, 31–40, etc. A histogram is commonly used to examine the distribution of a variable. For this purpose, a curve following a specific distribution (e.g., normal) is superimposed on the bars to assess the correspondence of the actual distribution to the desired (e.g., normal) distribution. Given that overlaying a normal curve makes most symmetric distributions look more normal then they are, you should be cautious when assessing normality by means of histograms. In Chap. 6 we will discuss several options for checking the normality of data.

Histograms plot continuous variables with ranges of the variables grouped into intervals (bins), while bar charts plot nominal and ordinal variables.

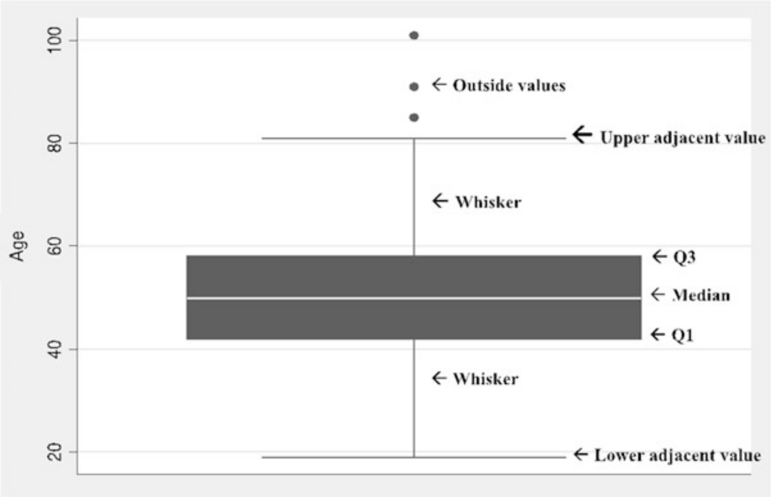


Fig. 5.5 Elements of the box plot

Another way of displaying the distribution of a (continuous) variable is the **box plot** (Fig. 5.4 top right) (also referred to as a **box-and-whisker plot**). The box plot is a graph representing a variable's distribution and consists of elements expressing the dispersion of the data. Note that several elements refer to terminologies discussed in the *Univariate Statistics* section. Figure 5.5 shows a box plot for the variable age based on the Oddjob Airways dataset.

- *Outside values* are observations that fall above the 3rd quartile +1.5 interquartile range.
- The *upper adjacent value* represents observations with the highest value that fall within the 3rd quartile +1.5 interquartile range.
- The upper line extending the box (*whisker*) represents the distance to observations with the highest values that fall within the following range: 3rd quartile + interquartile range. If there are no observations within this range, the line is equal to the maximum value.
- The top and bottom of the box describe the 3rd quartile (top) and 1st quartile (bottom); that is, the box contains the middle 50% of the data, which is equivalent to the interquartile range.
- The solid line inside the box represents the *median*.
- The lower line extending the box (*whisker*) represents the distance to the smallest observation that is within the following range: 1st quartile − interquartile range. If there are no observations within this range, the line is equal to the minimum value.
- The *lower adjacent value* represents observations with lowest values that fall inside the 3rd quartile −1.5 interquartile range.

We can make statements about the dispersion of the data with a box plot. The larger the box, the greater the observations' variability. Furthermore, the box plot helps us identify outliers in the data.

The **pie chart** (i.e., Fig. 5.4 bottom left) visualizes how a variable's different values are distributed. 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 values occur. The limitation of the pie chart is, however, that it is difficult to determine the size of segments that are very similar.

A **frequency table** (i.e., Fig. 5.4 bottom right) is a table that includes all possible values of a variable in absolute terms (i.e., frequency), how often they occur relatively (i.e., percentage), and the percentage of the cumulative frequency, which is the sum of all the frequencies from the minimum value to the category's upper bound (i.e., cumulative frequency). It is similar to 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. Like pie charts, frequency tables are primarily useful if variables are measured on a nominal or ordinal scale.

5.5.2 Univariate Statistics

Univariate statistics fall into two groups: those describing centrality and those describing the dispersion of variables. Box 5.2 at the end of this section shows sample calculation of the statistics used on a small set of values.

5.5.2.1 Measures of Centrality

Measures of centrality (sometimes referred to as *measures of central tendency*) are statistical indices of a "typical" or "average" score. There are two main types of measures of centrality, the median and the mean.[4]

The **median** is the value that occurs in the middle of the set of scores if they are ranked from the smallest to the largest, and it therefore separates the lowest 50% of cases from the highest 50% of cases. For example, if 50% of the products in a market cost less than $1,000, then this is the median price. Identifying the median requires at least ordinal data (i.e., it cannot be used with nominal data).

The most commonly used measure of centrality is the **mean** (also called the *arithmetic mean* or, simply, the *average*). The mean (abbreviated as $\bar{x}$) is the sum of each observation's value divided by the number of observations:

$$\bar{x} = \frac{\text{Sum}(x)}{n} = \frac{1}{n}\sum\nolimits_{i=1}^{n} x_i$$

[4] The mode is another measure. However, unlike the median and mean, it is ill-defined, because it can take on multiple values. Consequently, we do not discuss the mode.

In the above formula, x_i refers to the value of observation i of variable x and n refers to the total number of observations. The mean is only useful for interval or ratio-scaled variables.

Each measure of centrality has its own use. The mean is most frequently used, but is sensitive to very small or large values. Conversely, the median is not sensitive to outliers. Consequently, the relationship between the mean and the median provides us with valuable information about a variable's distribution. If the mean and the median are about the same, the variable is likely to be symmetrically distributed (i.e., the left side of the distribution mirrors the right side). If the mean differs from the median, this suggests that the variable is asymmetrically distributed and/or contains outliers. This is the case when we examine the prices of a set of products valued $500, $530, $530, and $10,000; the median is $530, while the mean is $2,890. This example illustrates why a single measure of centrality can be misleading. We also need to consider the variable's dispersion to gain a more complete picture.

5.5.2.2 Measures of Dispersion

Measures of dispersion provide researchers with information about the variability of the data; that is, how far the values are spread out. We differentiate between four types of measures of dispersion:

- range,
- interquartile range,
- variance, and
- standard deviation.

The **range** is the simplest measure of dispersion. It is the difference between the highest and the lowest value in a dataset and can be used on data measured at least on an ordinal scale. The range is of limited use as a measure of dispersion, because it provides information about extreme values and not necessarily about "typical" values. However, the range is valuable when screening data, as it allows for identifying data entry errors. For example, a range of more than 6 on a 7-point Likert scale would indicate an incorrect data entry.

The **interquartile range** is the difference between the 3rd and 1st quartile. The 1st *quartile* corresponds to the value separating the 25% lowest values from the 75% largest values if the values are ordered sequentially. Correspondingly, the 3rd quartile separates the 75% lowest from the 25% highest values. The interquartile range is particularly important for drawing box plots.

The **variance** (generally abbreviated as s^2) is a common measure of dispersion. The variance is the sum of the squared differences of each value and a variable's mean, divided by the sample size minus 1. The variance is only useful if the data are interval or ratio-scaled:

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

The variance tells us how strongly observations vary around the mean. A low variance indicates that the observations tend to be very close to the mean; a high variance indicates that the observations are spread out. Values far from the mean increase the variance more than those close to the mean.

The most commonly used measure of dispersion is the **standard deviation** (usually abbreviated as s). It is the square root of—and, therefore, a variant of—the variance:

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

The variance and standard deviation provide similar information, but while the variance is expressed on the same scale as the original variable, the standard deviation is standardized. Consequently, the following holds for normally distributed variables (this will be discussed in the following chapters in more detail):

– 66% of all observations are between plus and minus one standard deviation units from the mean,
– 95% of all observations are between plus and minus two standard deviation units from the mean, and
– 99% of all observations are between plus and minus three standard deviation units from the mean.

Thus, if the mean price is $1,000 and the standard deviation is $150, then 66% of all the prices fall between $850 and $1150, 95% fall between $700 and $1300, and 99% of all the observations fall between $550 and $1,450.

5.5.3 Bivariate Graphs and Tables

There are several *bivariate graphs* and *tables*, of which the scatter plot and the crosstab are the most important. Furthermore, several of the graphs, charts, and tables discussed in the context of univariate analysis can be used for bivariate analysis. For example, box plots can be used to display the distribution of a variable in each group (category) of nominal variables.

A **scatter plot** (see Fig. 5.6) uses both the y and x-axis to show how two variables relate to one another. If the observations almost form a straight diagonal line in a

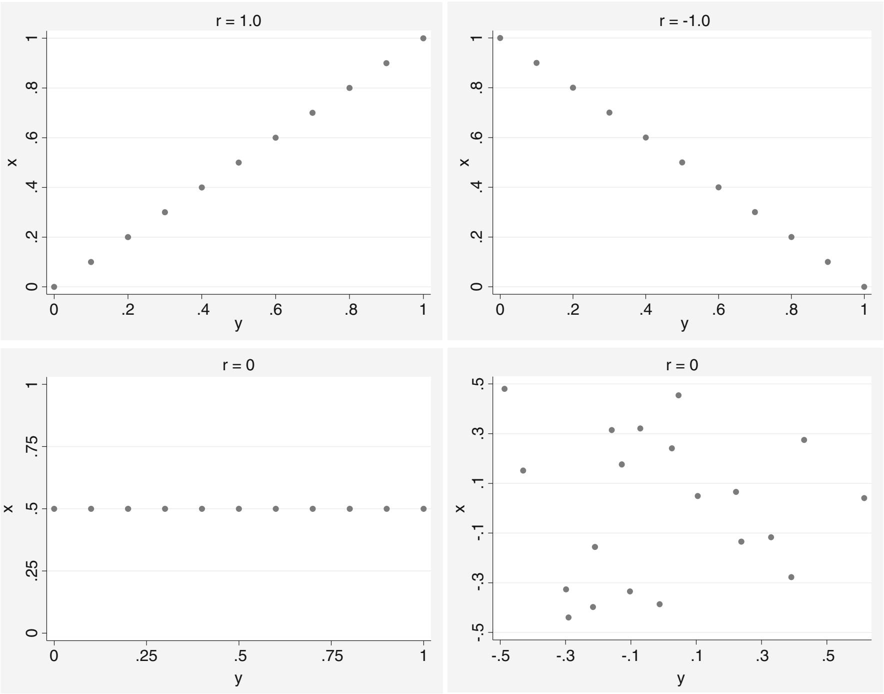


Fig. 5.6 Scatter plots and correlations

scatter plot, the two variables are strongly related.[5] Sometimes, a third variable, corresponding to the color or size (e.g., a *bubble plot*) of the data points, is included, adding another dimension to the plot.

Crosstabs (also referred to as *contingency tables*) are tables in a matrix format that show the frequency distribution of nominal or ordinal variables. They are the equivalent of a scatter plot used to analyze the relationship between two variables. While crosstabs are generally used to show the relationship between two variables, they can also be used for three or more variables, which, however, makes them difficult to grasp. Crosstabs are also part of the χ^2*-test* (pronounced as *chi-square*), which we discuss under nonparametric tests in the ⭳ Web Appendix (→ Downloads).

[5] A similar type of chart is the **line chart**. In a line chart, measurement points are ordered (typically by their x-axis value) and joined with straight line segments.

5.5.4 Bivariate Statistics

Bivariate statistics involve the analysis of two variables to determine the empirical relationship between them. There are two key measures that indicate (linear) associations between two variables; we illustrate their computation in Box 5.2:

- covariance, and
- correlation.

The **covariance** is the degree to which two variables vary together. If the covariance is zero, then two variables do not vary together. The covariance is the sum of the multiplication of the differences between each value of the x_i and y_i variables and their means, divided by the sample size minus 1:

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

The **correlation** (typically abbreviated as r) is a common measure of how strongly two variables relate to each other. The most common type of correlation, the *Pearson's correlation coefficient*, is calculated as follows:

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

The numerator contains the covariance of x_i and y_i ($Cov(x_i, y_i)$), while the denominator contains the product of the standard deviations of x_i and y_i.[6] Thus, the correlation is the covariance divided by the product of the standard deviations. As a result, the correlation is standardized and, unlike the covariance, is no longer dependent on the variables' original measurement. More precisely, the correlation coefficient ranges from -1 to 1, where -1 indicates a perfect negative relationship and 1 indicates the contrary. A correlation coefficient of 0 indicates that there is no relationship, also implying that their covariance is zero.

As a rule of thumb (Cohen 1988), an absolute correlation...

- ... below 0.30 indicates a weak relationship,
- ... between 0.30 and 0.49 indicates a moderate relationship, and
- ... above 0.49 indicates a strong relationship.

[6] Note that the terms $n-1$ in the numerator and denominator cancel each other and are therefore not shown here.

Box 5.2 Sample Calculation of Univariate and Bivariate Statistics
Consider the following list of values for variables x and y, which we treat as ratio-scaled:

x	6	6	7	8	8	8	12	14	14
y	7	6	6	9	8	5	10	9	9

Measures of centrality for x:

$$\text{Median} = 8$$

$$\text{Mean } \bar{x} = \frac{1}{9}\,(6 + 6 + \ldots + 14 + 14) = \frac{83}{9} \approx 9.22$$

Measures of dispersion for x:

$$\begin{aligned} &\text{Minimum} && = 6 \\ &\text{Maximum} && = 14 \\ &\text{Range} && = 14 - 6 = 8 \\ &\text{Interquartile range} && = 6.5 \end{aligned}$$

$$\text{Variance } \left(s^2\right) = \frac{\left[(6 - 9.22)^2 + \ldots + \left(14 - 9.22\right)^2\right]}{9 - 1} = \frac{83.56}{8} \approx 10.44.$$

$$\text{Standard deviation } (s) = \sqrt{s^2} = \sqrt{10.44} \approx 3.23$$

Measures of association between x and y:

$$\begin{aligned} \text{Covariance } (\text{cov}(x, y)) &= \frac{1}{9 - 1}[(6 - 9.22) \cdot (7 - 7.67) + \ldots \\ &\quad +(14 - 9.22) \cdot (9 - 7.67)] = \frac{31.67}{8} \approx 3.96 \\ \text{Correlation } (r) &= \frac{3.96}{3.23 \cdot 1.73} \approx 0.71 \end{aligned}$$

The scatter plots in Fig. 5.6 illustrate several correlations between two variables x and y. If the observations almost form a straight diagonal line in the scatter plot (upper left and right in Fig. 5.6), the two variables have a high (absolute) correlation. If the observations are uniformly distributed in the scatter plot (lower right in Fig. 5.6), or one variable is a constant (lower left in Fig. 5.6), the correlation is zero.

Pearson's correlation coefficient is the most common coefficient and is generally simply referred to as the correlation (Agresti and Finlay 2014). Pearson's correlation is appropriate for calculating correlations between two variables that are both interval or ratio-scaled. However, it can also be used when one variable is

Table 5.4 Types of descriptive statistics for differently scaled variables

	Nominal	Ordinal	Interval & ratio
Univariate graphs & tables			
Bar chart	X	X	
Histogram			X
Box plot			X
Pie chart	X	X	(X)
Frequency table	X	X	(X)
Univariate statistics: Measures of centrality			
Median		X	X
Mean			X
Univariate statistics: Measures of dispersion			
Range		(X)	X
Interquartile range		(X)	X
Variance			X
Standard deviation			X
Bivariate graphs/tables			
Scatter plot			X
Crosstab	X	X	(X)
Bivariate statistics			
Contingency coefficient	X		
Cramer's V	X		
Phi	X		
Spearman's correlation		X	
Kendall's tau		X	
Pearson's correlation			X

interval or ratio-scale and the other is, for example, binary. There are other correlation coefficients for variables measured on lower scale levels. Some examples are:

- *Spearman's correlation coefficient* and *Kendall's tau* when at least one variable for determining the correlation is measured on an ordinal scale.
- *Contingency coefficient*, *Cramer's V*, and *Phi* for variables measured on a nominal scale. These statistical measures are used with crosstabs; we discuss these in the context of nonparametric tests in the ⭳ Web Appendix (→ Downloads).

In Table 5.4, we indicate which descriptive statistics are useful for differently scaled variables. The brackets X indicate that the use of a graph, table, or statistic is potentially useful while (X) indicates that use is possible, but less likely useful, because this typically requires collapsing data into categories, resulting in a loss of information.

5.6 Transform Data (Optional)

Transforming data is an optional step in the workflow. Researchers transform data as certain analysis techniques require this: it might help interpretation or might help meet the assumptions of techniques that will be discussed in subsequent chapters. We distinguish two types of data transformation:

- variable respecification, and
- scale transformation.

5.6.1 Variable Respecification

Variable respecification involves transforming data to create new variables or to modify existing ones. The purpose of respecification is to create variables that are consistent with the study's objective. *Recoding* a continuous variable into a categorical variable is an example of a simple respecification. For example, if we have a variable that measures a respondent's *number of flights (indicating the number of flights per year)*, we could code those flights below 5 as low (=1), between 5 and 10 flights as medium (=2), and everything above 11 as high (=3). Recoding a variable in such a way always results in a loss of information, since the newly created variable contains less detail than the original. While there are situations that require recoding (e.g., we might be asked to give advice based on different income groups where income is continuous), we should generally avoid recoding.

Another example of respecification is swapping the polarity of a question. If you have a variable measured on a 5-point Likert-scale where: 1 = "strongly agree"; 2 = "agree"; 3 = "undecided"; 4 = "disagree"; 5 = "strongly disagree" and you wish to switch the polarity of the values so that value 1 reverses to 5, value 2 becomes 4, and so on.

Creating a dummy variable is a special way of recoding data. **Dummy variables** (or simply *dummies*) are binary variables that indicate if a certain trait is present or not. For example, we can use a dummy variable to indicate that advertising was used during a period (value of the dummy is 1) or not (value of the dummy is 0). We can also use multiple dummies to capture categorical variables' effects. For example, three levels of *flight intensity* (low, medium, and high) can be represented by two dummy variables: The first takes a value of 1 if the intensity is high (0 else), the second also takes a value of 1 if the intensity is medium (0 else). If both dummies take the value 0, this indicates low flight intensity. We always construct one dummy less than the number of categories. We explain dummies in further detail in the ⭳ Web Appendix (→ Downloads) and more information can be found at http://www.stata.com/support/faqs/data-management/creating-dummy-variables/. Dummies are often used in regression analysis (discussed in Chap. 7).

The creation of *constructs* is a frequently used type of variable respecification. As described in Chap. 3, a construct is a concept that cannot be observed, but can be measured by using multiple items, none of which relate perfectly to the construct. To compute a construct measure, we need to calculate the average (or the sum) of several related items. For example, a traveler's *commitment* to fly with Oddjob Airways can be measured by using the following three items:

- I am very committed to Oddjob Airways.
- My relationship with Oddjob Airways means a lot to me.
- If Oddjob Airways would no longer exist, it would be a true loss for me.

By calculating the average of these three items, we can form a *composite measure* of *commitment*. If one respondent indicated 4, 3, and 4 on the three items' scale, we calculate a **construct score** (also referred to as a composite score) for this person as follows: (4 + 3 + 4)/3 = 3.67. Note that we should take the average over the number of nonmissing responses.[7] In Chap. 8 we discuss more advanced methods of doing this by, for example, creating factor scores.

Similar to creating constructs, we can create an **index** of sets of variables. For example, we can create an index of information search activities, which is the sum of the information that customers require from promotional materials, the Internet, and other sources. This measure of information search activities is also referred to as a composite measure, but, unlike a construct, the items in an index define the trait to be measured.

5.6.2 Scale Transformation

Scale transformation involves changing the variable values to ensure comparability with other variables or to make the data suitable for analysis. Different scales are often used to measure different variables. For example, we may use a 5-point Likert scale for one set of variables and a 7-point Likert scale for a different set of variables in our survey. Owing to the differences in scaling, it would not be meaningful to make comparisons across any respondent's measurement scales. These differences can be corrected by **standardizing variables**.

A popular way of standardizing data is by rescaling these to have a mean of 0 and a variance of 1. This type of standardization is called the ***z*-standardization**. Mathematically, standardized scores z_i (also called *z-scores*) can be obtained by

[7] In Stata, this is best done using the `rowmean` command. For example, `egen commitment = rowmean (com1 com2 com3)`. This command automatically calculates the mean over the number of nonmissing responses.

subtracting the mean $\bar{x}$ of every observation x_i and dividing it by the standard deviation s. That is:

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

Range standardization (r_i) is another standardization technique which scales the data in a specific range. For example, standardizing a set of values to a range of 0 to 1 requires subtracting the minimum value of every observation x_i and then dividing it by the range (the difference between the maximum and minimum value).

$$r_i = \frac{(x_i - x_{min})}{(x_{max} - x_{min})}$$

The range standardization is particularly useful if the mean, variance, and ranges of different variables vary strongly and are used for some forms of cluster analysis (see Chap. 9).

A **log transformation**—another type of transformation—is commonly used if we have skewed data. **Skewed data** occur if we have a variable that is asymmetrically distributed and can be positive or negative. A *positive skew* (also called *right-skewed data* or data skewed to the right) occurs when many observations are concentrated on the left side of the distribution, producing a long right tail. When data are right-skewed, the mean will be higher than the median. A *negative skew* (also called *left-skewed data* or data skewed to the left) is the opposite, meaning that many observations are concentrated on the right of the distribution, producing a long left tail. When data are negatively skewed, the mean will be lower than the median. A histogram will quickly show whether data are skewed. Skewed data can be undesirable in analyses. Log transformations are commonly used to transform data closer to a normal distribution when the data are right-skewed (i.e., the data are non-negative). Taking a natural logarithm will influence the size of the coefficient related to the transformed variable, but will not influence the value of its outcome.[8]

Finally, **aggregation** is a special type of transformation. Aggregation means that we take variables measured at a lower level to a higher level. For example, if we know the average customer's satisfaction with an airline and the distribution channels from which they buy (i.e., the Internet or a travel agent), we can calculate the average satisfaction at the channel level. Aggregation only works from lower to higher levels and is useful if we want to compare groups at a higher level.

[8] 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. Logarithms cannot be calculated for negative values (such as household debt) and for the value of zero. In Stata, you can generate a log-transformed variable by typing: `gen loginc = log(income)`, whereby `loginc` refers to the newly created log-transformed variable and `income` refers to the income variable.

While transforming data is often necessary to ensure comparability between variables or to make the data suitable for analysis, there are also drawbacks to this procedure. Most notably, we may lose information during most transformations. For example, recoding the *ticket price* (measured at the ratio scale) as a "low," "medium," and "high" ticket price will result in an ordinal variable. In the transformation process, we have therefore lost information by going from a ratio to an ordinal scale. Another drawback is that transformed data are often more difficult to interpret. For example, the log (*ticket price*) is far more difficult to interpret and less intuitive than simply using the ticket price.

5.7 Create a Codebook

After all the variables have been organized and cleaned, and some initial descriptive statistics have been calculated, we can create a **codebook,** containing essential details of the data collection and data files, to facilitate sharing. Codebooks usually have the following structure:

Introduction: The introduction discusses 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 the types of questionnaires used. Thus, if different questionnaires were used for different respondents (e.g., for French and Chinese respondents), a copy of each original questionnaire should be included. Differences in wording may afterwards explain the results of the study, particularly those of 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 each variable as comments close to the actual 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 (the version that 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 previously been used. You should also describe the measurement level (see Chap. 3).

Summary statistics: This section includes descriptive statistics of each variable. The average (only for interval and ratio-scaled data), minimum, and maximum are often shown. In addition, the number of observations and usable observations (excluding observations with missing values) are included, just like a histogram (if applicable).

Datasets: This last section includes the names of the datasets and sometimes the names of all the revisions of the used datasets. Codebooks sometimes include the file date to ensure that the right files are used.

5.8 The Oddjob Airways Case Study

The most effective way of learning statistical methods is to apply them to a set of data. Before introducing Stata and how to use it, we present the dataset from a fictitous company called *Oddjob Airways* (but with a real website, http://www.oddjobairways.com) that will guide the examples throughout this book. The dataset *oddjob.dta* (↓ Web Appendix → Downloads) stems from a customer survey of Oddjob Airways. Founded in 1962 by the Korean businessman Toshiyuki Sakata, Oddjob Airways is a small premium airline, mainly operating in Europe, but also offering flights to the US. In an effort to improve its customers' satisfaction, the company's marketing department contacted all the customers who had flown with the airline during the last 12 months and were registered on the company website. A total of 1,065 customers who had received an email with an invitation letter completed the survey online.

The survey resulted in a rich dataset with information about travelers' demographic characteristics, flight behavior, as well as their price/product satisfaction with and expectations in respect of Oddjob Airways. Table 5.5 describes the variables in detail.

5.8.1 Introduction to Stata

Stata is a computer package specializing in quantitative data analysis, and widely used by market researchers. It is powerful, can deal with large datasets, and relatively easy to use. In this book, we use Stata MP4 14.2 (to which we simply refer to as Stata). Prior versions (12 or higher) for Microsoft Windows, Mac or Linux can be used for (almost) all examples throughout the book.

Stata offers a range of versions and packages; your choice of these depends on the size of your dataset and the data processing speed. Stata/SE and Stata/MP are recommended for large datasets. The latter is the fastest and largest version of Stata. Stata/MP can, for example, process large datasets with up to 32,767 variables and 20 billion observations, while Stata/SE can process datasets with the same number of variables, but a maximum of 2.14 billion observations. Stata/IC is recommended

Table 5.5 Variable description and label names of the Oddjob Dataset

Variables	Variable description	Variable name in the dataset
Demographic measures		
Age of the customer	Numerical variable ranging between the ages of 19 and 101.	*Age*
Customer's gender	Dichotomous variable, where 1 = Female; 2 = Male.	*Gender*
Language of customer	Categorical variable, where 1 = German; 2 = English; 3 = French.	*Language*
Home country	Categorical variable, whereby: 1 = Germany (de), 2 = Switzerland (ch); 3 = Austria (at); 4 = France (fr), 5 = the United States (us).	*Country*
Flight behaviour measures		
Flight class	Categorical variable distinguishing between the following categories: 1 = First; 2 = Business; 3 = Economy.	*flight_class*
Latest flight	Categorical variable querying when the customer last flew with Oddjob Airways. Categories are: 1 = Within the last 2 days; 2 = Within the last week; 3 = Within the last month; 4 = Within the last 3 months; 5 = Within the last 6 months; 6 = Within the last 12 months.	*flight_latest*
Flight purpose	Dichotomous variable distinguishing between: 1 = Business; 2 = Leisure.	*flight_purpose*
Flight type	Dichotomous variable, where: 1 = Domestic; 2 = International.	*flight_type*
Number of flights	Numeric variable ranging between 1 and 457 flights per year.	*nflights*
Traveler's status	Categorical variable, where membership status is defined in terms of: 1 = Blue; 2 = Silver; 3 = Gold.	*status*
Perception and satisfaction measures		
Traveler's expectations	23 items reflecting a customer's expectations with the airline: "How high are your expectations that. . ." All items are measured on a continuous scale ranging from 1 very low to 100 very high.	*e1* to *e23*
Traveler's satisfaction	23 items reflecting a customer's satisfaction with Oddjob Airways regarding the features asked in the expectation items (*e1-e23*) on a continuous scale ranging from 1=very unsatisfied to 100=very satisfied.	*s1* to *s23*
Recommendation	Item on whether a customer is likely to recommend the airline to a friend or colleague. This item is measured on an 11-point Likert-scale ranging from 1 very unlikely to 11 very likely.	*nps*
Reputation	One item stating "Oddjob Airways is a reputable airline." This item is measured on a 7-point Likert-scale ranging from 1 fully disagree to 7 fully agree.	*reputation*
Overall price/ performance satisfaction	One item stating "Overall I am satisfied with the price performance ratio of Oddjob Airways." This item is measured on a 7-point Likert-scale ranging from 1 fully disagree to 7 fully agree.	*overall_sat*

(continued)

Table 5.5 (continued)

Variables	Variable description	Variable name in the dataset
General satisfaction	3 items reflecting a customer's overall satisfaction with the airline. All items are measured on a 7-point Likert scale ranging from 1 fully disagree to 7 fully agree.	*sat1* to *sat3*
Loyalty	5 items reflecting a customer's loyalty to the airline. All items are measured on a 7-point Likert-scale ranging from 1 fully disagree to 7 fully agree.	*loy1* to *loy5*
Commitment	3 items reflecting a customer's commitment to fly with the airline. All items are measured on a 7-point Likert-scale ranging from 1 fully disagree to 7 fully agree.	*com1 to com3*

for moderate-sized datasets. This can process datasets with a maximum of 2,047 variables and up to 2.14 billion observations. Small Stata is only available for students and handles small-sized datasets with a maximum of 99 variables and 1,200 observations. To obtain a copy of Stata, check the stata.com website, or ask the IT department, or your local Stata distributor. Special student and faculty prices are available.

In the next sections, we will use the ► sign to indicate that you should click on something. Options, menu items or drop-down lists that you should look up in dialog boxes are printed in **bold**. Variable names, data files or data formats are printed in *italics* to differentiate them from the rest of the text. Finally, Stata commands are indicated in `Courier`.

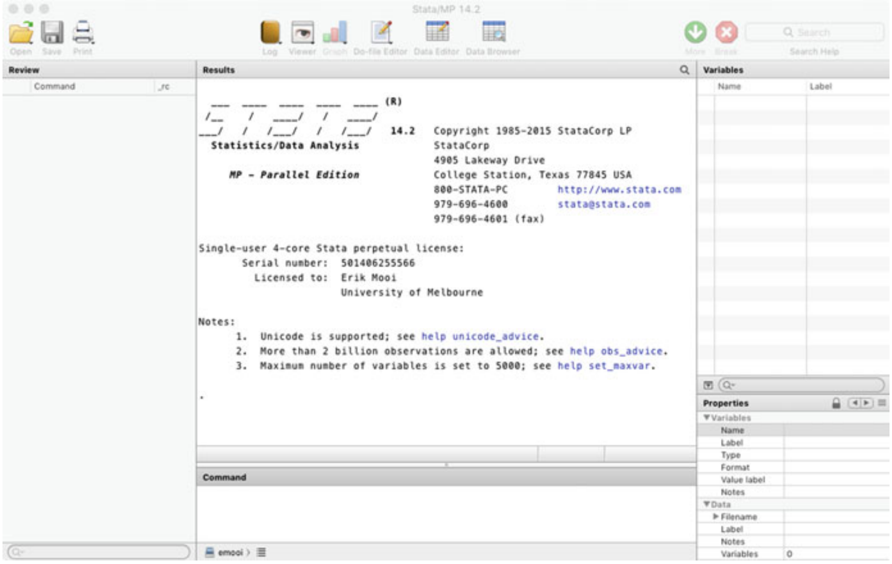
5.8.2 Finding Your Way in Stata

5.8.2.1 The Stata Main/Start Up Window and the Toolbar

If you start up Stata for the first time, it presents a screen as shown in Fig. 5.7. This start up screen is the main Stata window in the Mac version.[9]

The main Stata window in Fig. 5.7 consists of five sub-windows. The first sub-window on the left of the start-up screen is called the **Review** window, which displays the history of commands since starting a session. Successful commands are displayed in black, and those containing errors are displayed in red. If you click on one of the past commands displayed in the **Review** window, it will be automatically copied

[9] If you open Stata in the Windows or Linux operating systems, the toolbar looks a bit different, but is structured along the same lines as discussed in this chapter.



```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   14.2   Copyright 1985-2015 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     MP - Parallel Edition            College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Single-user 4-core Stata perpetual license:
       Serial number:  501406255566
         Licensed to:  Erik Mooi
                       University of Melbourne

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  More than 2 billion observations are allowed; see help obs_advice.
      3.  Maximum number of variables is set to 5000; see help set_maxvar.

.
```

Fig. 5.7 The Stata interface

into the **Command** window, which is located at the bottom of the central screen. The **Review** window stores and displays your commands. It allows you to recall all previous commands, edit, and re-submit them if you wish. The output of your analyses is displayed in the **Results** window, located in the center. The **Variables** window (upper-right) lists all the variables included in the dataset, whereas the **Properties** window (lower-right) displays the features of the selected variable(s).

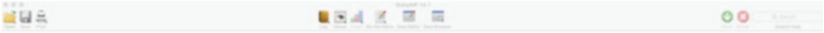
Stata's toolbar is located underneath its menu bar. This toolbar contains a range of shortcuts to frequently used commands, including opening, saving, printing, viewing, and editing your data. An overview of the toolbar icons is shown in Table 5.6.

5.8.2.2 The Menu Bar

Stata's menu bar includes **File**, **Edit**, **Data**, **Graphics**, **Statistics**, **Users**, **Windows**, and **Help**, which will be discussed briefly in this section. The menu options **Graphics** and **Statistics** will, later in this chapter be discussed in greater detail by means of examples.

> You can open Stata's dialog box by simply typing `db`, followed by the operation (i.e., edit, describe, copy, etc.) or technique (regression) that you wish to carry out. For example, if you wish to open the data editor, you could type: `db edit`. Similarly, if you wish to specify your regression model, you could type `db regress`, which will open the dialog window for the regression model.

Table 5.6 Toolbar icons



Stata’s toolbar in detail:

Symbol	Action
Open	Opens dataset by selecting a dataset from a menu.
Save	Saves the active dataset.
Print	Prints contents of a selected window.
Log	Log begins/closes/suspends/resumes the current log. It also allows for viewing the current log if a log file is open.
Viewer	Opens the viewer that provides advice on finding help and how to search through Stata’s online resources.
Graph	Brings the graph window to the front.
Do-file Editor	Opens a new do-file editor or brings a do-file editor to the front.
Data Editor	Opens the data editor (edit) or brings the data editor to the front. It allows you to edit variables.
Data Browser	Opens the data editor (browse) or brings the data editor to the front. It allows you to browse through the variables.
More	This tells Stata to show more output.
Break	This tells Stata to interrupt the current task.
tabulate Search Help	Enables searching for help for Stata commands.

File

Format Types

Stata uses multiple file formats. The *.dta* file format only contains data. Stata can also import other file formats such as Excel (*.xls* and *.xlsx*), SAS, SPSS and can read text files (such as *.txt* and *.dat*). Once these files are open, they can be saved as Stata’s *.dta* file format. Under ► File, you find all the commands that deal with the opening, closing, creating, and saving of the different types of files. In the startup screen, Stata shows several options for creating or opening datasets. The options that you should use are either **Open Recent**, under which you can find a list with recently opened data files, or **New Files**.

You can import different types of files into Stata if you go to ► File ► Import, select the file format, and then select the file you wish to open. The **Example dataset** menu item is particularly useful for learning new statistical methods, as it provides access to all the example datasets mentioned under the titles of the various user manuals for a particular statistical method.

> Stata has an extensive number of user-written programs, which, once installed, act as Stata commands. If you know the name of the program, you can directly type `help`, followed by a keyword, which initiates a keyword search. Alternatively, if you do not know the name of the command, but are looking for a specific method, such as cluster analysis, you can type: `help cluster` to initiate a keyword search for this method. This will open a new window containing information about this technique from the help files, the Stata reference manual, the Stata Journal, the Frequently Asked Questions, references to other articles, and other help files.

Stata *.do* Files

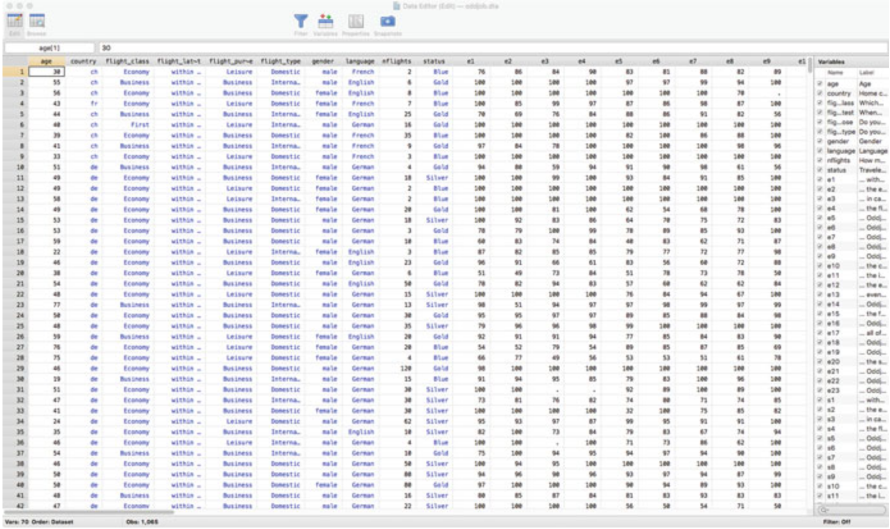
In addition to the menu functionalities, we can run Stata by using its command language, called **Stata command**. Think of this as a programming language that can be directly used (if you feel comfortable with its command) or via dialog boxes. Stata's commands can be saved (as a *.do* file) for later use. This is particularly useful if you undertake the same analyses across different datasets. Think, for example, of standardized marketing reports on daily, weekly, or monthly sales. Note that discussing Stata *.do* files in great detail is beyond the scope of this book, but, as we go through the different chapters, we will show all the Stata's commands that we have used in this book.

Edit

This menu option allows you to copy Stata output and analysis, and paste the content into, for example, a Word document. If you want to copy a table from your output, first select the table and then go to ► Edit ► Copy table. Remember that you need to have an output to make use of this option!

View

The **Data Editor** button is listed first in this menu option and brings the data editor to the front. The data editor looks like a spreadsheet (Fig. 5.8) listing variables in columns and the corresponding observations in rows. String variables are displayed in red, value labels and encoded numeric variables with value labels in blue, and numeric variables in black. The data editor can be used to enter new or amend old data across the rows or the columns of the spreadsheet. Entering the new value on the cell and pressing **Enter** will take you to the next row, while, if you press **Tab**, you are able to work across the rows. Blank columns and rows will be marked as missing, so make sure that you do not skip columns and rows when entering new data. Note that there are different types of variables in the data editor and their



age[1] 30

	age	country	flight_class	flight_lat~t	flight_pur~e	flight_type	gender	language	nflights	status	e1	e2	e3	e4	e5	e6	e7	e8	e9	e1
1	30	ch	Economy	within …	Leisure	Domestic	male	French	2	Blue	76	86	84	90	83	81	88	82	89	
2	55	ch	Business	within …	Business	Interna…	male	English	6	Gold	100	100	100	100	97	97	99	94	100	
3	56	ch	Economy	within …	Business	Domestic	female	English	8	Blue	100	100	100	100	100	100	100	70	.	
4	43	fr	Economy	within …	Leisure	Domestic	female	French	7	Blue	100	85	99	97	87	86	98	87	100	
5	44	ch	Business	within …	Business	Interna…	female	English	25	Gold	70	69	76	84	88	86	91	82	56	
6	40	ch	First	within …	Leisure	Interna…	male	German	16	Gold	100	100	100	100	100	100	100	100	100	
7	39	ch	Economy	within …	Business	Domestic	male	French	35	Blue	100	100	100	100	82	100	86	88	100	
8	41	ch	Business	within …	Business	Interna…	male	French	9	Gold	97	84	78	100	100	100	100	98	96	
9	33	ch	Economy	within …	Leisure	Domestic	male	French	3	Blue	100	100	100	100	100	100	100	100	100	
10	51	de	Business	within …	Business	Interna…	male	German	4	Gold	94	88	59	94	91	90	98	61	56	
11	49	de	Economy	within …	Business	Domestic	female	German	18	Silver	100	100	99	100	93	84	91	85	100	
12	49	de	Economy	within …	Leisure	Domestic	female	German	2	Blue	100	100	100	100	100	100	100	100	100	
13	58	de	Economy	within …	Leisure	Interna…	female	German	2	Blue	100	100	100	100	100	100	100	100	100	
14	49	de	Economy	within …	Business	Domestic	female	German	20	Gold	100	100	81	100	62	54	68	78	100	
15	53	de	Economy	within …	Business	Domestic	male	German	18	Silver	100	92	83	86	64	70	75	72	83	
16	53	de	Economy	within …	Business	Domestic	male	German	3	Gold	78	79	100	99	78	89	85	93	100	
17	59	de	Economy	within …	Business	Domestic	male	German	10	Blue	60	83	74	84	40	83	62	71	87	
18	22	de	Economy	within …	Leisure	Interna…	female	English	3	Blue	87	82	85	85	79	77	72	77	98	
19	46	de	Economy	within …	Business	Domestic	male	English	23	Gold	96	91	66	61	83	56	60	72	88	
20	38	de	Economy	within …	Leisure	Domestic	female	German	6	Blue	51	49	73	84	51	78	73	78	50	
21	54	de	Economy	within …	Business	Domestic	male	English	50	Gold	78	82	94	83	57	60	62	62	84	
22	48	de	Economy	within …	Leisure	Domestic	male	German	15	Silver	100	100	100	100	76	84	94	67	100	
23	77	de	Business	within …	Business	Interna…	male	German	13	Silver	98	51	94	97	97	98	99	97	99	
24	50	de	Economy	within …	Business	Domestic	male	German	30	Gold	95	95	97	97	89	85	88	84	98	
25	48	de	Economy	within …	Business	Domestic	male	German	35	Silver	79	96	96	98	99	100	100	100	100	
26	59	de	Business	within …	Leisure	Domestic	female	English	20	Gold	92	91	91	94	77	85	84	83	90	
27	76	de	Economy	within …	Leisure	Domestic	female	German	20	Blue	54	52	79	54	89	85	87	85	69	
28	75	de	Economy	within …	Leisure	Domestic	female	German	4	Blue	66	77	49	56	53	53	51	61	78	
29	46	de	Economy	within …	Business	Domestic	male	German	120	Gold	98	100	100	100	100	100	100	100	100	
30	19	de	Business	within …	Business	Interna…	male	German	15	Blue	91	94	95	85	79	83	100	96	100	
31	51	de	Economy	within …	Business	Domestic	male	German	30	Silver	100	100	.	.	92	89	100	89	100	
32	47	de	Economy	within …	Business	Interna…	male	German	30	Silver	73	81	76	82	74	88	71	74	85	
33	41	de	Economy	within …	Business	Domestic	female	German	30	Silver	100	100	100	100	32	100	75	85	82	
34	24	de	Economy	within …	Leisure	Domestic	male	German	62	Silver	95	93	97	87	99	95	91	91	100	
35	35	de	Economy	within …	Business	Interna…	male	English	10	Silver	82	100	73	84	79	83	67	74	94	
36	46	de	Economy	within …	Leisure	Interna…	male	German	4	Blue	100	100	.	100	71	73	86	62	100	
37	54	de	Business	within …	Business	Interna…	male	German	10	Gold	75	100	94	95	94	97	94	90	100	
38	46	de	Economy	within …	Business	Domestic	male	German	50	Silver	100	94	95	100	100	100	100	100	100	
39	50	de	Economy	within …	Business	Domestic	male	German	80	Silver	94	96	90	96	93	97	94	87	99	
40	50	de	Economy	within …	Business	Domestic	female	German	80	Gold	97	100	100	100	90	94	89	93	100	
41	48	de	Business	within …	Business	Domestic	male	German	16	Silver	80	85	87	84	81	83	93	83	83	
42	47	de	Economy	within …	Business	Domestic	male	German	22	Silver	100	100	100	100	56	50	54	71	50	

Variables

	Name	Label
☑	age	Age
☑	country	Home c...
☑	flig...lass	Which...
☑	flig...test	When...
☑	flig...ose	Do you...
☑	flig...type	Do you...
☑	gender	Gender
☑	language	Language
☑	nflights	How m...
☑	status	Travele...
☑	e1	... with...
☑	e2	... the e...
☑	e3	... in ca...
☑	e4	... the fl...
☑	e5	... Oddj...
☑	e6	... Oddj...
☑	e7	... Oddj...
☑	e8	... Oddj...
☑	e9	... Oddj...
☑	e10	... the c...
☑	e11	... the i...
☑	e12	... the e...
☑	e13	... even...
☑	e14	... Oddj...
☑	e15	... the f...
☑	e16	... Oddj...
☑	e17	... all of...
☑	e18	... Oddj...
☑	e19	... Oddj...
☑	e20	... the s...
☑	e21	... Oddj...
☑	e22	... Oddj...
☑	e23	... Oddj...
☑	s1	... with...
☑	s2	... the e...
☑	s3	... in ca...
☑	s4	... the fl...
☑	s5	... Oddj...
☑	s6	... Oddj...
☑	s7	... Oddj...
☑	s8	... Oddj...
☑	s9	... Oddj...
☑	s10	... the c...
☑	s11	... the i...

Vars: 70 Order: Dataset Obs: 1,065 Filter: Off

Fig. 5.8 The Stata data editor

properties can be changed in the **Properties** sub-window located at the bottom right of the data editor screen. Another way to bring Data Editor to the front is by typing `edit` in the command window of the Main/Startup Window. Similarly, typing `browse` will open up the data editor (browse) mode, which allows you to navigate through your dataset.

The **View** menu (which is only found on the Mac version) includes other editing options that work in combination with *.do* files (Do-file Editor), graphs (Graph Editor), and structural equation modelling (SEM) (SEM-Builder). These options are beyond the scope of this book and will not be discussed in detail here.

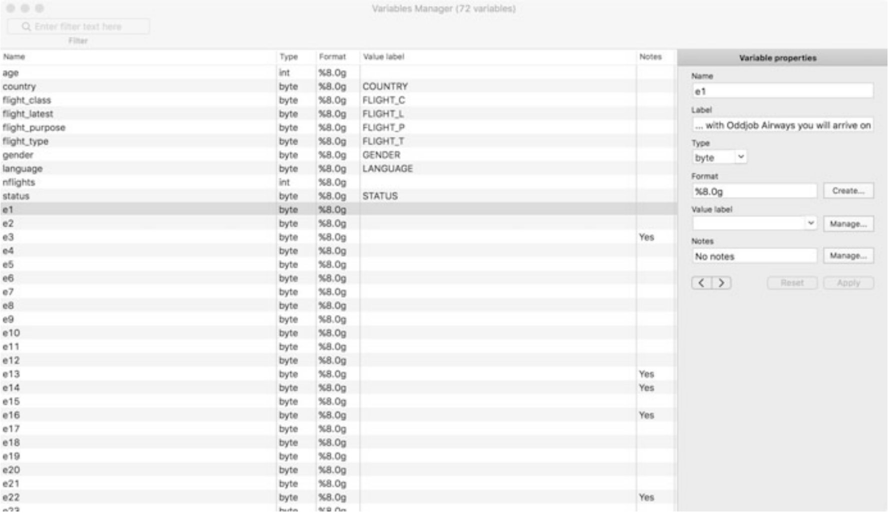
Data

The **Data** menu provides subcommands that allow for summarizing, inspecting, creating, changing or restructuring your dataset.

Under ► Data ► Describe data, you can view or inspect the dataset in a detailed way, including the content, type, and values of the variables included in the dataset. This offers useful information that complements the compact overview that you can obtain from ► Data ► Describe data ► Summary statistics.

Under ► Data ► Data Editor, you can access the different data editor modes (i.e., the edit and browse mode). As discussed above, these options are useful if you want to browse through the specific records in your data (i.e., the browse mode) or wish to change the content of a specific cell (i.e., the edit mode).

By going to ► Data ► Sort, you can sort the data according to the values of one or more specific variable(s). Data are usually sorted according to the respondents' key identifying variable; this a numerical variable that is often called the *id* variable. Depending on the purpose of the analysis, you could, for example, sort



Variables Manager (72 variables)

Name	Type	Format	Value label	Notes
age	int	%8.0g		
country	byte	%8.0g	COUNTRY	
flight_class	byte	%8.0g	FLIGHT_C	
flight_latest	byte	%8.0g	FLIGHT_L	
flight_purpose	byte	%8.0g	FLIGHT_P	
flight_type	byte	%8.0g	FLIGHT_T	
gender	byte	%8.0g	GENDER	
language	byte	%8.0g	LANGUAGE	
nflights	int	%8.0g		
status	byte	%8.0g	STATUS	
e1	byte	%8.0g		
e2	byte	%8.0g		
e3	byte	%8.0g		Yes
e4	byte	%8.0g		
e5	byte	%8.0g		
e6	byte	%8.0g		
e7	byte	%8.0g		
e8	byte	%8.0g		
e9	byte	%8.0g		
e10	byte	%8.0g		
e11	byte	%8.0g		
e12	byte	%8.0g		
e13	byte	%8.0g		Yes
e14	byte	%8.0g		Yes
e15	byte	%8.0g		
e16	byte	%8.0g		Yes
e17	byte	%8.0g		
e18	byte	%8.0g		
e19	byte	%8.0g		
e20	byte	%8.0g		
e21	byte	%8.0g		
e22	byte	%8.0g		Yes

Variable properties

Name: e1

Label: ... with Oddjob Airways you will arrive on

Type: byte

Format: %8.0g

Value label:

Notes: No notes

Fig. 5.9 The variables manager

the data according to the *age* of the respondents, which will sort the observations in an ascending order, listing the youngest respondents first, followed by the older respondents in the subsequent rows. Alternatively, **Advanced sort** in the same dialog box allows the observations to be arranged in descending order.

Under ► Data ► Create or change data, you find several options for creating new variables. The options ► Create new variable and ► Create new variable (extended) allow you to create a new variable from one (or more) existing variables. To change or recode the contents of existing variables, you need to select the option ► Other variable-transformation commands ► Recode categorical variable. This command allows you to recode variable values or to combine sets of values into one value. For example, as shown in Fig. 5.12, you could create a new variable called *age_dummy* that splits the sample into young to middle-aged (say, 40 or less) and old passengers (say, more than 40).

An important element of the Data menu is the **Variables Manager**, which is a tool that allows you to organize and structure the variables in your dataset (see Fig. 5.9). Among others it allows you to:

1. Sort variables: One click on the first column of **Variables Manager** will sort the variables in an ascending order, while a second click will sort them in a descending order. If you want to restore the order of the variable list, right-click on the first column and select the option **Restore column defaults**.
2. Select by filtering through the variable list and to change the variable properties: Enter the first letter or name of a variable in the filter box on the upper-left part of the screen to filter through the list of the included variables. This is especially useful if you want to zoom into a series of variables with similar names. Entering *e1* in the filter box, for example, will search everything that contains the value

e1. It will bring all variables to the front that have *e1* in their root, including *e1*, *e10*, *e11* to *e19*. Once you click on a specific variable (e.g., *e1*), the name, label, type, format, value labels, and notes under **Variable properties** will be populated; their content can then be edited if necessary. The properties of each of these boxes is briefly discussed below:

- **Name**: lists the name of the specified variable. Stata is case sensitive, meaning that the variable *e1* differs from *E1*. Variable names must begin with letters (a to z) or an underscore (_). Subsequent characters can include letters (a to z), numbers (0–9) or an underscore (_). Note that neither spaces nor special characters (e.g., %, &, /) are allowed.
- **Label**: allows for a longer description of the specified variable. This can, for example, be the definition of the variable or the original survey question. Click on the tick box to select the option to attach a label to a new variable and type in the preferred label for this variable.
- **Type**: specifies the output format type of the variable. **String** refers to words and is stored as `str#`, indicating the string's maximum length. String variables are useful if you want to include open-ended answers, email addresses or any other type of information that is not a number. Numbers are stored in Stata as `byte`, `int`, `long`, `float`, or `double`.[10]
- **Format**: describes the display format associated with a specified variable. As described in the Stata manual,[11] formats are denoted by a `%` sign, followed by the number of characters (i.e., width) and the number of digits following the decimal points. For example, `%10.2g` means that our display format (indicated by %) should be 10 characters wide (indicated by the number 10) with two digits (indicated by the number 2 following the decimal point). The `g` format indicates that Stata fills the 10 display characters with as much as it can fit. In addition, Stata has a range of formats for dates and time stamps to control how data are displayed.[12]
- **Value label**: presents a description of the values that a specified variable takes. It allows the researcher to create, edit or drop the label of a specified variable. The value labels for *gender*, for example, can be specified as *female* (for values coded as 1) and *male* (for values coded as 0).

3. Keep variables: if you wish to work with a subset of your variables, select all the relevant variables (by dragging the selected variables) and right-click. Then select the option **Keep only selected variables**. Note that this should *only* be done after careful consideration, as dropping relevant variables can ruin the dataset!
4. Drop variables: this is similar to the previous action, but now you only select the variables that you want to drop from the variables list. You can do so by first selecting the variables that you want to drop, then right-click and select the

[10] http://www.stata.com/manuals14/ddatatypes.pdf

[11] http://www.stata.com/manuals14/u.pdf

[12] http://www.stata.com/manuals14/dformat.pdf

option **Drop selected variables**. Here, you also think *very* carefully before dropping variables from the list!

The default missing value in Stata is called *system missing*, which is indicated by a period (**.**) in the dataset. In addition, there are 26 missing values, also called *extended missing values*, ranging from `.a` to `.z`, depending on how the data are stored. These are necessary to understand the type of the missing data. As discussed in the *Missing Data* section, missing values arise for various reasons. In some occasions, such as in panel studies where the same respondent is approached repeatedly, some respondents may refuse to answer a question at each interview leading to missing observations in some interviews. It could also be that the researcher decides to skip a question from the questionnaire. While both situations lead to missing values, their nature differs. The extended missing values in Stata allow the researcher to distinguish between these different reasons by assigning an `.a` to the first situation and `.b` to the second situation, etc.

Graphics

Under ► Graphics, Stata offers a range of graphical options and tools with which to depict data. These vary from graphical options that support distributional graphs, including two-way graphs, charts, histograms, box and contour plots to graphs that support more advanced statistical methods such as time series, panel, regression, survival techniques, etc. We will discuss the application and interpretation of the different plots as we move through the different chapters and statistical techniques in this book.

Statistics

Under ► Statistics, you find numerous analysis procedures, several of which we will discuss in the remainder of the book. For example, under **Summaries, tables, and tests**, you can request univariate and bivariate statistics. The rest are numerous types of regression techniques, as well as a range of other multivariate analysis techniques. We will discuss descriptive statistics in the next section. In Chap. 6, we will describe models that fall within the group ► Linear models and related, while Chap. 7 will discuss techniques in the ► Multivariate analysis group.

User

Under ► User, you find three empty data, graphs, and statistics subdirectories. Stata programmers can use these to add menu options.

Window

The ► Window option enables you to bring the different types of windows to front. You can also zoom in or minimize the screen.

Help

The ► Help function may come in handy if you need further guidance. Under ► Help, you find documentations and references that show you how to use most of the commands included in Stata.

5.9 Data Management in Stata

In this section, we will illustrate the application of some of the most commonly used commands for managing data in Stata. These include the following:

- restrict observations,
- create a new variable from existing variable(s), and
- recode variables.

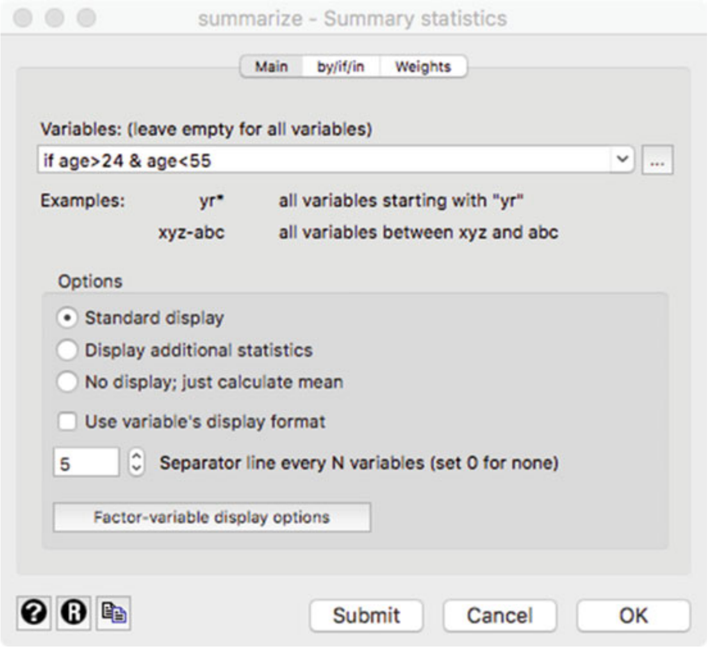
5.9.1 Restrict Observations

You can also use the **Summary Statistics** command to restrict certain observations from the analyses by means of a pre-set condition (e.g., display the summary statistics only for those between 25 and 54 years old). To restrict observations, go to ► Data ► Describe data ► Summary statistics, which opens a screen similar to Fig. 5.10. Under **Variables: (leave empty for all variables)**, enter the condition `if age > 24 & age < 55`. to specify your restriction and then click on **OK**. Stata will now only display the summary statistics for the observations that satisfy this condition. In the Stata command (see below) this restriction appears as an *if* statement.

```
summarize if age>24 & age<55
```

To summarize cases with only valid (non-missing) observations, it is common to add the following rule after the pre-set condition: `& age! = missing ()`. In Stata language, `! =` means "not equal" and `missing()` means "all numerical and string variables included in the dataset." All together, `& age! = missing()`means "if age is not missing from all the included variables in the dataset".

```
summarize if age>24 & age<55 & age !=missing()
```



summarize - Summary statistics

Main | by/if/in | Weights

Variables: (leave empty for all variables)

if age>24 & age<55

Examples: yr* all variables starting with "yr"

xyz-abc all variables between xyz and abc

Options

● Standard display

○ Display additional statistics

○ No display; just calculate mean

□ Use variable's display format

5 Separator line every N variables (set 0 for none)

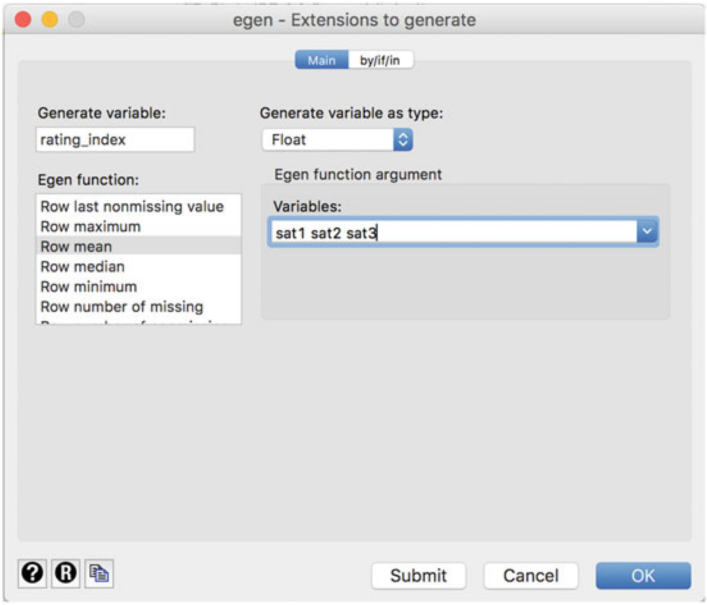
Fig. 5.10 Restrict observations

5.9.2 Create a New Variable from Existing Variable(s)

The **Create new variable (extended)** command enables you to create new variables containing the interquartile range, median, row means, standardized values, and many more different options. We will not discuss all these options in this section, but will demonstrate the power of this tool by creating and index variable from the mean of the following three items related to travelers' satisfaction: *sat1*, *sat2*, and *sat3*. Go to ► Data ► Create or change data ► Create new variable (extended), which will open a dialog box similar to Fig. 5.11.

Next, enter the name of the new variable (e.g., *rating_index*) in the **Generate variable** box on the upper left part of the screen and select **Row Mean** from the **Egen Function** drop-down menu. Under **Generate variable as type**, the variable type **Float** should be automatically selected (i.e., this is Stata's default for numeric variable types). Finally, enter *sat1*, *sat2*, and *sat3* in the **Variables** box and click on **OK**. You have now created a new variable called *rating_index* that appears at the bottom of the variable list. Alternatively, you can type the following command:

```
egen float rating_index = rowmean(sat1 sat2 sat3)
```



Generate variable: rating_index

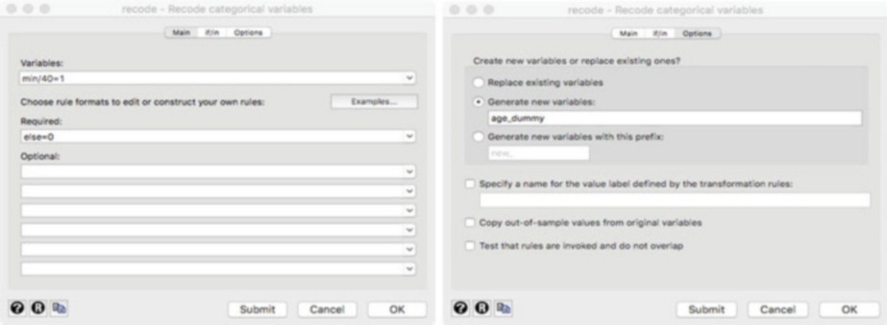
Generate variable as type: Float

Egen function:
- Row last nonmissing value
- Row maximum
- Row mean
- Row median
- Row minimum
- Row number of missing

Egen function argument

Variables: sat1 sat2 sat3

Fig. 5.11 Create new variable(s) extended


recode - Recode categorical variables
Main
if/in
Options
Examples...
Submit
Cancel
OK
recode - Recode categorical variables
Main
if/in
Options
Submit
Cancel
OK

Variables: min/40=1

Choose rule formats to edit or construct your own rules:

Required: else=0

Optional:

Create new variables or replace existing ones?

○ Replace existing variables

● Generate new variables: age_dummy

○ Generate new variables with this prefix: new_

□ Specify a name for the value label defined by the transformation rules:

□ Copy out-of-sample values from original variables

□ Test that rules are invoked and do not overlap

Fig. 5.12 Recode into different variables (Main and Options tabs)

5.9.3 Recode Variables

Recoding (i.e., changing or transforming) the values of an existing variable according to a number of rules is a key data management activity. Numeric variables can be changed by means of the `recode` command. Go to ► Data ► Create or change data ► Other variable-transformation commands ► Recode categorical variables. This will open a dialog box similar to the **Main** Tab left in Fig. 5.12.

Specify the name of the variable that you want to recode (i.e., *age*) under **Variables** in the **Main** tab. Next, specify the values of the new variable under

Required. These are based on the values of the original variable (*age*). The option (min/40 = 1) indicates that all values ranging from the smallest age observations to the age of 40 should be coded as 1. Under **Optional** all other age observations are coded as 0 (else = 0).

Next, click on the **Options** tab (right in Fig. 5.12). In the dialog box that follows, you need to indicate whether you want to: (1) **Replace existing variables**, (2) **Generate new variables**, or (3) **Generate new variables with this prefix**. We always recommend using either the second option or the third. If you were to use the first option, any changes you make to the variable will result in the overwriting of the original variable. Consequently, if you thereafter wish to return to the original data, you will either need to revert to a saved previous version, or need to enter all the data again, because Stata cannot undo these actions! Select the second option (i.e., **Generate new Variables**), enter the name of the new variable (i.e., *age_dummy*), and then click on **OK**. Alternatively, the recoding of this variable can be obtained through the following command:

```
recode age(min/40=1)(else=0),generate(age_dummy)
```

You have now created a new dichotomous variable (i.e., *age_dummy*) located at the bottom of the variables list.

Everyone makes mistakes!

If you are worried that commands may not change the data as desired, type `preserve` in the **Command** window. This keeps a snapshot of the data in the computer's memory. Should you wish to revert, simply type `restore` to go back to where you were. You could also save a copy of the dataset under a new name as a milestone and then later delete it manually. This allows you to back up multiple steps and across work sessions.

5.10 Example

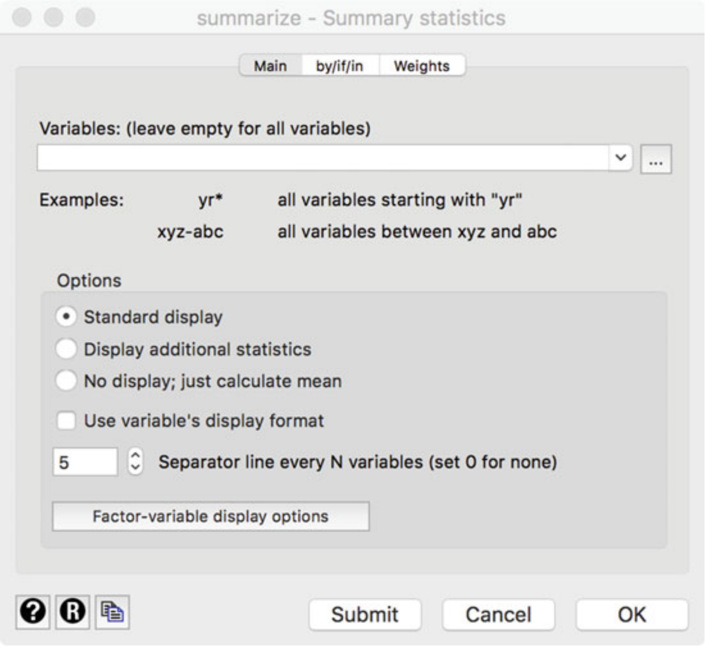
We will now examine the dataset *Oddjob.dta* in closer detail by following all the steps in Fig. 5.1. Cleaning the data generally requires checking for interviewer fraud, suspicious response patterns, data entry errors, outliers, and missing data. Several of these steps rely on statistics and graphs, which we discussed in the context of descriptive statistics (e.g., box plots and scatter plots). Note that missing values strategies and the multiple imputation technique will be described and illustrated by means of *Oddjob.dta* in the ⭳ Web Appendix (→ Downloads).

5.10.1 Clean Data

Since the data were cleaned earlier, we need not check for interviewer fraud or suspicious response patterns. Beside double data entries to detect and minimize errors in the process of data entry, exploratory data analysis is required to spot data entry errors that have been overlooked.

A first step in this procedure is to look at the minimum and maximum values of the relevant variables to detect values that are not plausible (i.e., fall outside the expected range of scale categories). To do so, go to ► Data ► Describe Data ► Summary statistics, which opens a dialog box like the one in Fig. 5.13. The **Variables** box should be left empty to obtain the statistics for all the variables in the dataset. Next, select the **Standard display** option that requires the number of observations, the mean, standard deviation, minimum, and maximum values. Proceed by clicking on **OK**. Alternatively, the dialog box for summary statistics can be brought to front by typing `db summarize` in the **Command** window and clicking on enter.

Table 5.7 shows a partial display of the summary statistics, including the number of observations, means, standard deviation, as well as minimum and maximum values. Under **Obs**, we can see that all the listed variables are observed across all 1,065 respondents, meaning that none of the selected variables suffer from missing observations. Among others, it appears that the *age* of the travelers varies between


summarize - Summary statistics

Main | by/if/in | Weights

Variables: (leave empty for all variables)

Examples: yr* all variables starting with "yr"
xyz-abc all variables between xyz and abc

Options
● Standard display
○ Display additional statistics
○ No display; just calculate mean
□ Use variable's display format
5 Separator line every N variables (set 0 for none)

Fig. 5.13 Summary statistics dialog box

Table 5.7 A (partial) output of summary statistics

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |      1,065    50.41972    12.27464         19        101
     country |      1,065           2    1.551739          1          5
flight_class |      1,065    2.798122    .4352817          1          3
flight_lat~t |      1,065    3.788732    1.368779          1          6
flight_pur~e |      1,065    1.507042    .5001853          1          2
-------------+---------------------------------------------------------
 flight_type |      1,065    1.476056     .499661          1          2
      gender |      1,065    1.737089    .4404212          1          2
    language |      1,065    1.237559    .4473162          1          3
    nflights |      1,065    13.41878    20.22647          1        457
      status |      1,065    1.498592    .7204373          1          3
-------------+---------------------------------------------------------
```

19 and 101, while the *number of flights* varies between 1 and 457 flights over the past 12 months. Particularly the maximum value in number of flights appears to be implausible. While this observation could represent a flight attendant, it appears more reasonable to consider this observation an outlier, which may need to be eliminated, depending on the type of analysis.

5.10.2 Describe Data

In the next step, we describe the data in more detail, focusing on those statistics and graphs that were not part of the previous step. To do so, we make use of graphs, tables, and descriptive statistics. In Fig. 5.14, we show how you can ask for each previously discussed graph, table, and statistic in Stata.

5.10.2.1 Univariate Graphs and Tables

Bar Charts

To produce a bar chart that plots the *age* of respondents against their *country* of residence, go to ▶ Graphics ▶ Bar chart. This will take you to a dialog box where the **Main** tab is displayed by default (left of Fig. 5.15). Under **Type of data** specify the type of bar chart that you want displayed (**Graph of summary statistics**). Next, under **Orientation**, you should select the option **Vertical,** given that it is Stata's default and indicates the direction in which the bar chart is displayed.

Under **Statistics to plot**, select the variable *age* under **Variables** and indicate that you want to display the mean of this variable for all valid observations by selecting the option **mean**. Next, click on the **Categories** tab (displayed to the right of Fig. 5.15) to indicate how you want to categorize the data in the bar chart. Tick the first box, **Group 1** and select the variable *country* from the drop-down menu under **Grouping Variable.**

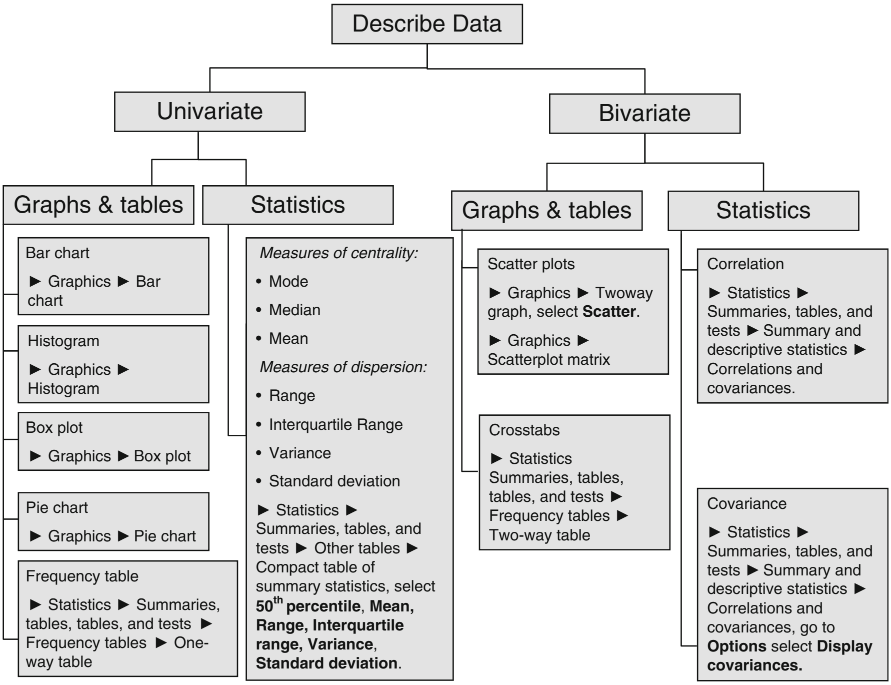
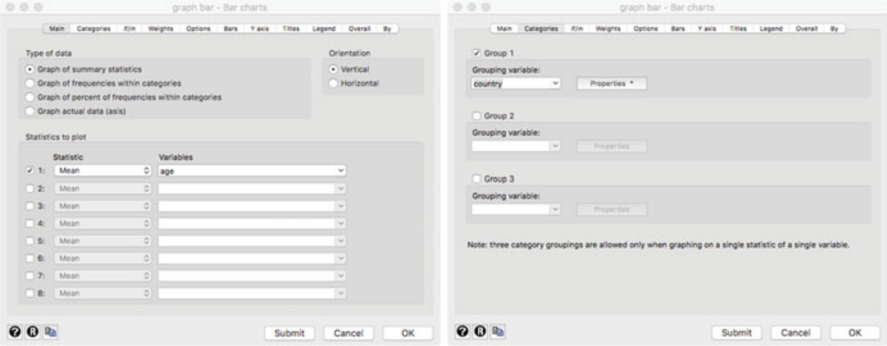

Describe Data
Univariate
Bivariate
Graphs & tables
Statistics
Bar chart
▶ Graphics ▶ Bar chart
Histogram
▶ Graphics ▶ Histogram
Box plot
▶ Graphics ▶ Box plot
Pie chart
▶ Graphics ▶ Pie chart
Frequency table
▶ Statistics ▶ Summaries, tables, tables, and tests ▶ Frequency tables ▶ One-way table
Measures of centrality:
• Mode
• Median
• Mean
Measures of dispersion:
• Range
• Interquartile Range
• Variance
• Standard deviation
▶ Statistics ▶ Summaries, tables, and tests ▶ Other tables ▶ Compact table of summary statistics, select 50th percentile, Mean, Range, Interquartile range, Variance, Standard deviation.
Graphs & tables
Statistics
Scatter plots
▶ Graphics ▶ Twoway graph, select Scatter.
▶ Graphics ▶ Scatterplot matrix
Crosstabs
▶ Statistics Summaries, tables, tables, and tests ▶ Frequency tables ▶ Two-way table
Correlation
▶ Statistics ▶ Summaries, tables, and tests ▶ Summary and descriptive statistics ▶ Correlations and covariances.
Covariance
▶ Statistics ▶ Summaries, tables, and tests ▶ Summary and descriptive statistics ▶ Correlations and covariances, go to Options select Display covariances.

Fig. 5.14 How to ask for graphs, tables, and statistics in Stata



Type of data
● Graph of summary statistics
○ Graph of frequencies within categories
○ Graph of percent of frequencies within categories
○ Graph actual data (asis)

Orientation
● Vertical
○ Horizontal

Statistics to plot

	Statistic	Variables
☑ 1:	Mean	age
□ 2:	Mean	
□ 3:	Mean	
□ 4:	Mean	
□ 5:	Mean	
□ 6:	Mean	
□ 7:	Mean	
□ 8:	Mean	

☑ Group 1
Grouping variable: country Properties *

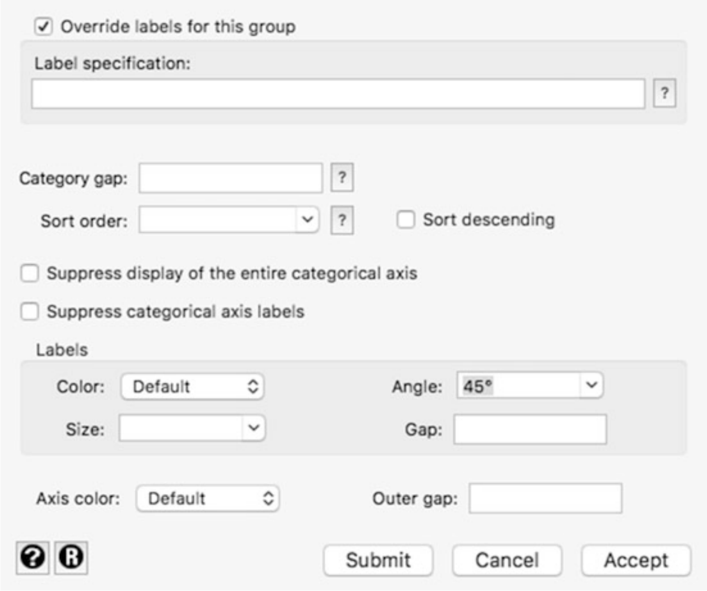
□ Group 2
Grouping variable: Properties

□ Group 3
Grouping variable: Properties

Note: three category groupings are allowed only when graphing on a single statistic of a single variable.

Fig. 5.15 Main dialog box, Bar chart

By default, Stata displays the value labels of the grouping variable horizontally, but you can change this. By clicking on the **Properties** button (see Fig. 5.16), which is next to the **Grouping variable** box, set **Labels** to **Angle: 45**° and then click **Accept**. Stata will then show a graph as in Fig. 5.17.



☑ Override labels for this group

Label specification:

Category gap:

Sort order: □ Sort descending

□ Suppress display of the entire categorical axis

□ Suppress categorical axis labels

Labels

Color: Default Angle: 45°

Size: Gap:

Axis color: Default Outer gap:

Fig. 5.16 Properties dialog box

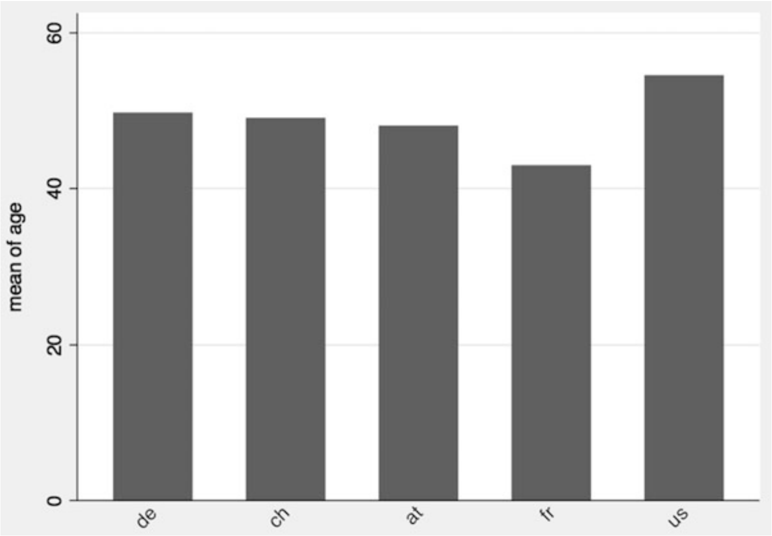


Fig. 5.17 A bar chart

Tip: You can also edit your graph by right-clicking on the graph. Select the option **Start Graph Editor**, which will start the Graph Editing menu. Double click on the *x*-axis and select the option **Label properties** to adjust the angle of the labels in the same way as described above. You can additionally also adjust the range of the values you want to display on the *x*-axis by specifying the desired range. The same applies to adjusting the label properties of the *y*-axis.

Histograms

Histograms are useful for summarizing numerical variables. If you go to ► Graphics ► Histogram, Stata will open a dialog box as shown in Fig. 5.18 on the left. To plot a histogram that summarizes the respondents' *age*, select the relevant variable *age* in the **Variable** drop-down menu and select the **Data are continuous** option. Next, specify **Frequency** under **Y axis** and click on **OK.** Stata will produce a histogram as shown in Fig. 5.18 on the right.

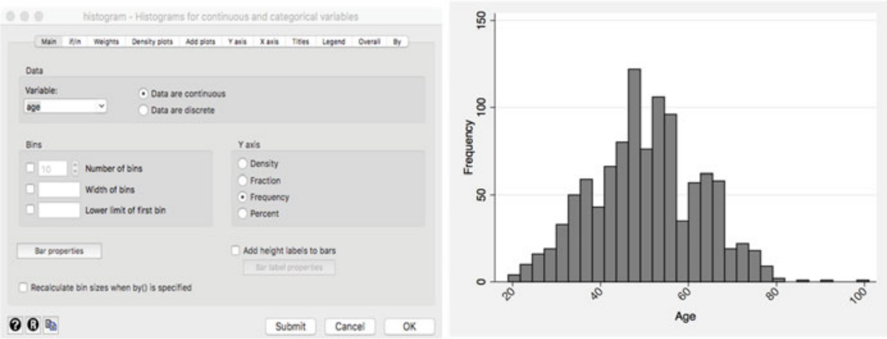
Box Plot

To ask for a box plot, go to ► Graphics ► Box plot, which will open a dialog box similar to the one in Fig. 5.19.

Specify the option **Vertical** under **Orientation** to display the box plot vertically. Next, select the relevant variable *age* in the **Variables** box and then click on **OK**. A box plot like the one in Fig. 5.20 appears.

Pie Charts

Pie charts are useful for displaying categorical or binary variables. Create a pie chart by going to ► Graphics ► Pie chart, which will open a dialog box similar to



histogram - Histograms for continuous and categorical variables

Data

Variable: age

● Data are continuous
○ Data are discrete

Bins

□ 10 Number of bins
□ Width of bins
□ Lower limit of first bin

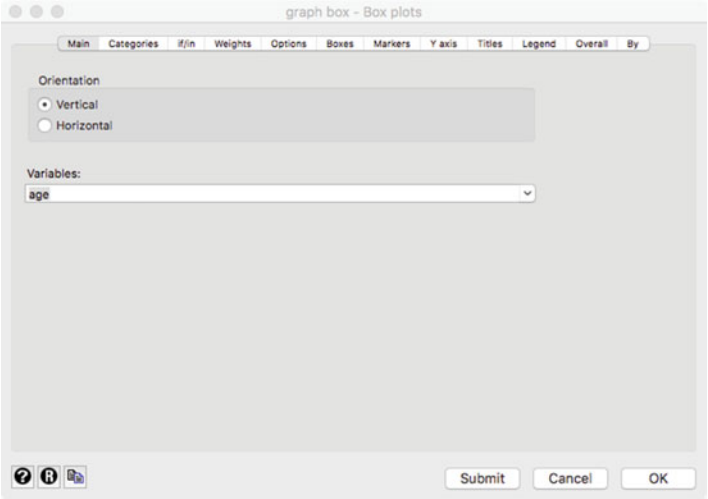
Y axis

○ Density
○ Fraction
● Frequency
○ Percent

□ Add height labels to bars

□ Recalculate bin sizes when by() is specified

Fig. 5.18 Dialog box, histogram and a histogram



Orientation
● Vertical
○ Horizontal

Variables:
age

Fig. 5.19 Box plots graph dialog box

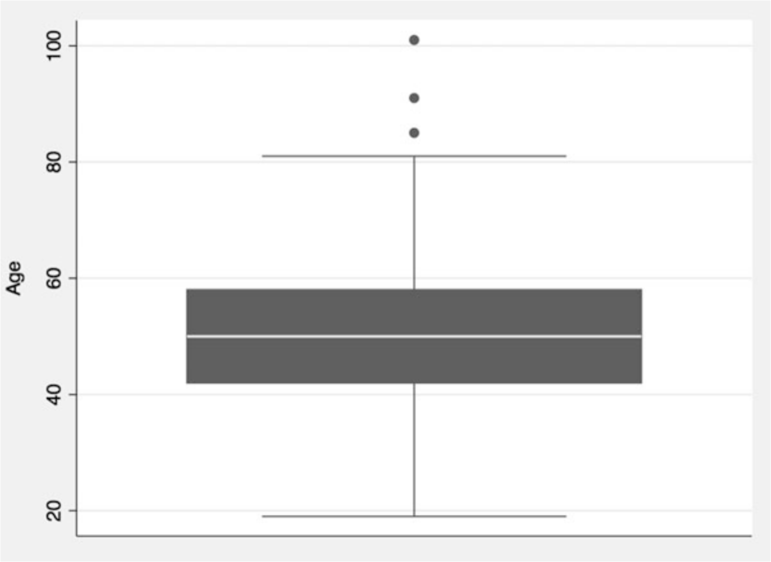
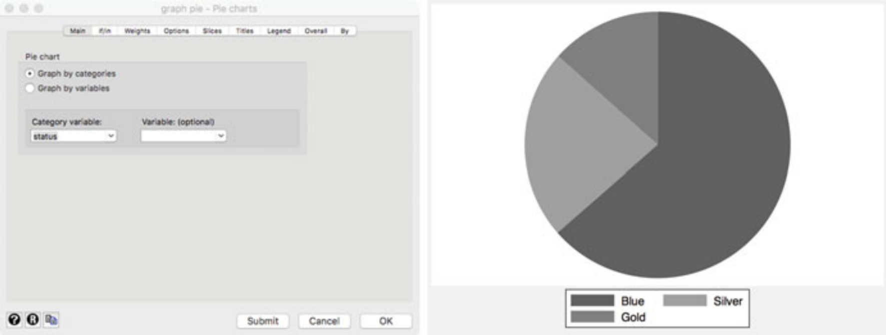


Fig. 5.20 Box plot

that displayed in Fig. 5.21 on the left. In Stata, the default (standard) option is **Graph by Categories**. Plot respondents' membership status (*status*) by selecting *status* under **Category variable** and then clicking on **OK**. Stata will show a pie chart similar to the one on the right in Fig. 5.21.



Pie chart
● Graph by categories
○ Graph by variables

Category variable: status
Variable: (optional)

Fig. 5.21 Pie chart

Table 5.8 Example of a frequency table in Stata

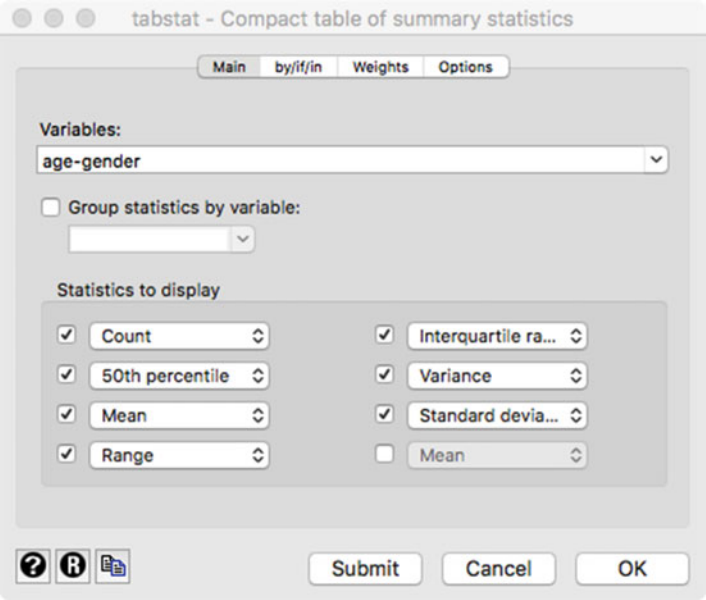
```
tabulate country

       Home |
    country |      Freq.     Percent        Cum.
------------+-----------------------------------
         de |        695       65.26       65.26
         ch |         66        6.20       71.46
         at |        108       10.14       81.60
         fr |          1        0.09       81.69
         us |        195       18.31      100.00
------------+-----------------------------------
      Total |      1,065      100.00
```

Frequency Tables

We can produce a frequency table by clicking on ► Statistics ► Summaries, tables, and tests ► Frequency tables ► One-way table. Select the variable *country* under **Categorical variable** and then click on **OK**. This operation will produce Table 5.8, which displays the value of each country with the corresponding absolute number of observations (i.e., **Freq.**), the relative values (i.e., **Percent**), as well as the cumulative relative values (i.e., **Cum.**). It shows that **65.25 percent** of our sample consists of travelers who reside in Germany, followed by travelers from the United States (**18.31 percent**), Austria (**10.14 percent**), Switzerland (**6.20 percent**), and, finally, France (**0.09 percent**).

5.10.2.2 Univariate Statistics

Another useful way of summarizing your data is through the **Tabstat** option, which you can find under ► Statistics ► Summaries, tables, and tests ► Other tables ► Compact table of summary statistics. Selecting this menu option opens a dialog box similar to that in Fig. 5.22. In the **Variables** box, select the relevant variables for which you would like to display summary statistics. These variables range from *age* to *gender*. Next, under **Statistics to display** tick the blank boxes and specify the



Variables:
age-gender

□ Group statistics by variable:

Statistics to display

☑ Count	☑ Interquartile ra...
☑ 50th percentile	☑ Variance
☑ Mean	☑ Standard devia...
☑ Range	□ Mean

Fig. 5.22 Dialog box for univariate statistics

Table 5.9 Example of a summary table using the tabstat option

```
tabstat age-gender, statistics (count p50 mean range iqr var sd)

   stats |       age   country  flight~s  flight~t  fligh~se  fligh~pe    gender
---------+----------------------------------------------------------------------
       N |      1065      1065      1065      1065      1065      1065      1065
     p50 |        50         1         3         4         2         1         2
    mean |  50.41972         2  2.798122  3.788732  1.507042  1.476056  1.737089
   range |        82         4         2         5         1         1         1
     iqr |        16         2         0         2         1         1         1
variance |  150.6667  2.407895  .1894702  1.873557  .2501853  .2496611  .1939708
      sd |  12.27464  1.551739  .4352817  1.368779  .5001853   .499661  .4404212
--------------------------------------------------------------------------------
```

descriptive statistics to be displayed from the drop-down menu. In this example, we specify the number of nonmissing observations (**Count** in Stata), the median (**50th percentile** in Stata), the **Mean**, the **Range**, the **Interquartile range**, the **Variance**, and the **Standard deviation**.

Clicking on **OK** will display an output similar to that in Table 5.9, which includes the number of nonmissing observations (**N**), median (**p50**), mean (**mean**), range (**range**), interquartile range (**iqr**), variance (**var**), and standard deviation (**sd**).

Note that `tabstat` arranges the table like a dataset, with the variables as columns. For a more a typical table, go to the **Options** tab (see Fig. 5.22) and change the value of the **Use as columns** from **Variables** to **Statistics**.

5.10.2.3 Bivariate Graphs and Tables

Scatter Plots and Matrix Scatter Plots

Matrix scatter plots can be easily displayed in Stata by going to ► Graphics ► Twoway graph (scatter, line, etc.) or Graphics ► Scatterplot matrix. These two separate graphs differ in the following way:

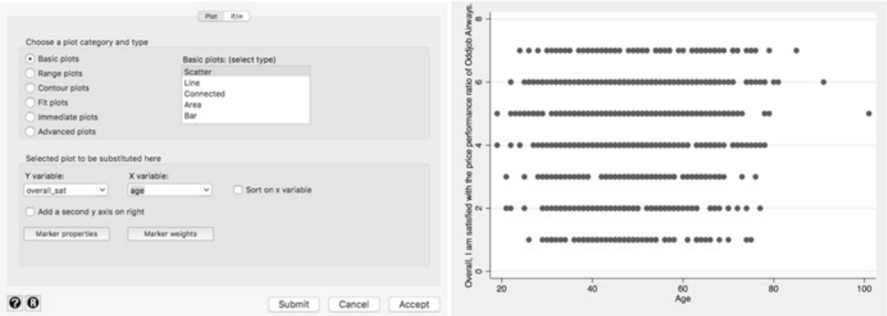
1. **Twoway graph (scatter, line, etc.)**: plots two variables against each other, but you can display multiple scatter plots at a time on the same graph. Produce a twoway scatter plot by going to ► Graphics ► Twoway graph (scatter, line, etc.) and clicking on **Create**. In the dialog box that opens (on the left-hand side of Fig. 5.23), select **Scatter**, enter the outcome variable overall satisfaction (*overall_sat*) in the **Y variable** box, and *age* in the **X variable** box, click on **Accept**, and then on **OK**. The right-hand side of Fig. 5.23 shows the resulting scatter plot.

2. **Scatterplot matrix**: creates scatter plots for multiple variables simultaneously.

Going to Graphics ► Scatterplot matrix opens a dialog box similar to the one shown on the left of Fig. 5.24. Select *nflights*, *overall_sat*, and *age* under **Variables**. To change the density of the scatter matrix, click on the button **Marker Properties**. This opens the dialog window on the right of Fig. 5.24. Select the option **Point** under **Symbol**, click on **Accept**, and then on **OK.**

This will produce the matrix scatter plot shown in Fig. 5.25. The graph shows distinct bands, because the overall satisfaction variable takes on 7 values (1 to 7).

The first scatter plot in the first row reveals the relationship between the *number of flights* and *overall price satisfaction*. Next to this on the first row, the relationship between the *number of flights* and *age* is displayed. In the second row, the relationship between the *overall price satisfaction* and *number of flights* is shown and so on. In the same row, to the right, the relationship



Choose a plot category and type

● Basic plots
○ Range plots
○ Contour plots
○ Fit plots
○ Immediate plots
○ Advanced plots

Basic plots: (select type) Scatter

Selected plot to be substituted here

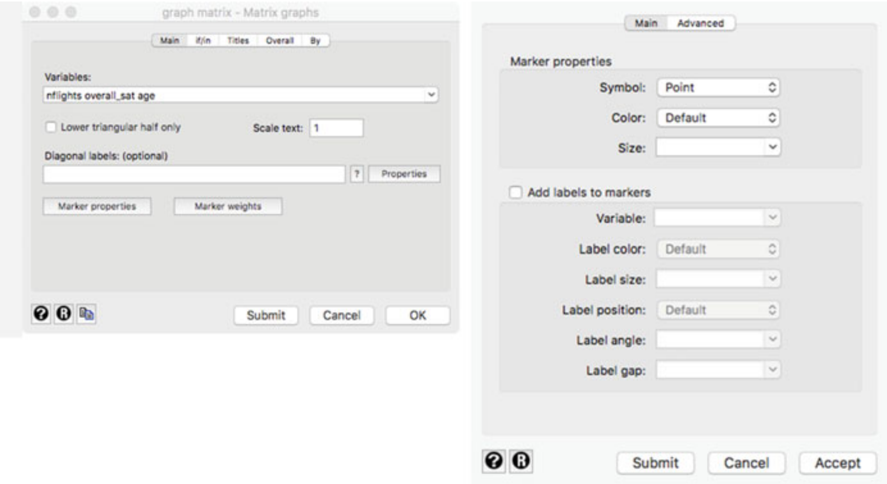
Y variable: overall_sat

X variable: age

□ Sort on x variable

□ Add a second y axis on right

Fig. 5.23 Dialog box and scatter plot



graph matrix - Matrix graphs

Variables: nflights overall_sat age

□ Lower triangular half only

Scale text: 1

Diagonal labels: (optional)

Marker properties

Symbol: Point

Color: Default

Size:

□ Add labels to markers

Variable:

Label color: Default

Label size:

Label position: Default

Label angle:

Label gap:

Fig. 5.24 Dialog box scatter plot matrix and the marker properties box

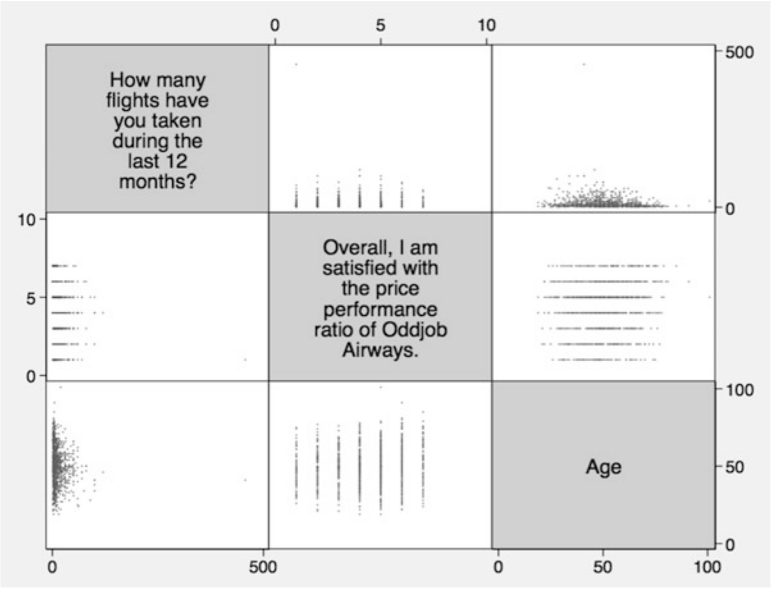


Fig. 5.25 Matrix scatter plot

between the *overall price satisfaction* and *age* is displayed. Finally, the last row displays the relationships between *age* and *number of flights* (first scatter plot), as well as the relationship between *age* and *overall price satisfaction* (second scatter plot).

Table 5.10 Example of a crosstab

```
tabulate country gender, column row

+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
| column percentage |
+-------------------+

      Home |        Gender
   country |    female       male |     Total
-----------+----------------------+----------
        de |       180        515 |       695 
           |     25.90      74.10 |    100.00 
           |     64.29      65.61 |     65.26 
-----------+----------------------+----------
        ch |        17         49 |        66 
           |     25.76      74.24 |    100.00 
           |      6.07       6.24 |      6.20 
-----------+----------------------+----------
        at |        25         83 |       108 
           |     23.15      76.85 |    100.00 
           |      8.93      10.57 |     10.14 
-----------+----------------------+----------
        fr |         1          0 |         1 
           |    100.00       0.00 |    100.00 
           |      0.36       0.00 |      0.09 
-----------+----------------------+----------
        us |        57        138 |       195 
           |     29.23      70.77 |    100.00 
           |     20.36      17.58 |     18.31 
-----------+----------------------+----------
     Total |       280        785 |     1,065 
           |     26.29      73.71 |    100.00 
           |    100.00     100.00 |    100.00 
```

Cross Tabulation

Cross tabulations are useful for understanding the relationship between two variables scaled on a nominal or ordinal scale. To create a crosstab, go to ▶ Statistics ▶ Summaries, tables, and tests ▶ Frequency tables ▶ Two-way table with measures of association. It is important that you specify which variable goes in the column and which in the rows. Choose *country* under **Row variable** and *gender* under **Column variable**. Next, select the boxes **Within-column relative frequencies** and **Within-row relative frequencies** under **Cell Contents** to display the column and row percentages. Click on **OK** and Stata produces a table similar to the one in Table 5.10.

5.10.2.4 Bivariate Statistics: Correlation and Covariance

In Stata, we can calculate bivariate correlations by going to ▶ Statistics ▶ Summaries, tables, and tests ▶ Summary and descriptive statistics ▶ Correlations and covariances. In the dialog box that opens, select the variables to be considered in the analysis. For example, enter *nflights*, *age*, and *overall_sat* in the **Variables** box. When you click on **OK**, Stata will produce a correlation matrix like the one in Table 5.11.

Table 5.11 Correlation matrix produced in Stata

```
correlate nflights age overall_sat
(obs=1,065)

             | nflights      age overal~t
-------------+---------------------------
    nflights |   1.0000
         age |  -0.1158   1.0000
 overall_sat |  -0.1710   0.1207   1.0000
```

Table 5.12 Covariance matrix produced in Stata

```
correlate nflights age overall_sat, covariance
(obs=1,065)

             | nflights      age overal~t
-------------+---------------------------
    nflights |   409.11
         age | -28.7408  150.667
 overall_sat | -5.62173  2.40806  2.64118
```

The correlation matrix in Table 5.11 shows the correlation between each pairwise combination of three variables. For example, the correlation between *nflights* and *age* is −**0.1158**, which is negative and rather weak. Conversely, the relationship between *age* and *overall_sat* is positive (**0.1207**), but still rather weak.

Alternatively, a covariance matrix as in Table 5.12, can be obtained as follows. Go to ► Statistics ► Summaries, tables, and tests ► Summary and descriptive statistics ► Correlations and covariances and ticking the box **Display covariances** in the **Options** tab. This will produce the following covariance matrix.

5.11 Cadbury and the UK Chocolate Market (Case Study)

The UK chocolate market is expected to be £6.46 billion in 2019. Six subcategories of chocolates are used to identify the different chocolate segments: boxed chocolate, molded bars, seasonal chocolate, count lines, straight lines, and "other."

To understand the UK chocolate market for molded chocolate bars, we have a dataset *(chocolate.dta)* that includes a large supermarket's weekly sales of 100g molded chocolate bars from January 2016 onwards. This data file can be downloaded from the book's ⭳ Web Appendix (→ Downloads). This file contains a set of variables. Once you have opened the dataset, you will see the set of variables we discuss in the main Stata window under *Variables* (see Fig. 5.7).

The first variable is *week*, indicating the week of the year and starts with Week 1 of January 2016. The last observation for 2016 ends with observation 52, but the variable continues to count onwards for 16 weeks in 2017.[13] The next variable is

[13] Note an ordinary year has 52 weeks and 1 day, while a leap year has 52 weeks and 2 days. This is because 1 week comprises part of 2016 and part of 2017.

sales, which indicates the weekly sales of 100g Cadbury bars in £. Next, four price variables are included, *price1-price4*, which indicate the price of Cadbury, Nestlé, Guylian, and Milka in £. Next, *advertising1-advertising4* indicate the amount of £ 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 by means of 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 with providing descriptive statistics for a client by means of this 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. In addition, create a few more 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 for describing the usage of advertising? Which statistics are appropriate for describing point-of-purchase advertising?

5.12 Review 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 5 years? What types of descriptive statistics work best to determine the market shares of five different types of insurance providers? Should we use just one or multiple descriptive statistics?
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 codebook of the Household Income and Labour Dynamics in Australia (HILDA) Survey at: http://melbourneinstitute.unimelb.edu.au/hilda/for-data-users/user-manuals. Is this codebook clear? What do you think of its structure?

5.13 Further Readings

https://www.stata.com/manuals13/mi.pdf

This manual provides a hands-on application of multiple imputation in Stata.

http://www.stata.com/manuals13/u.pdf

There is a detailed description of Stata's properties coupled with hands-on examples in Stata's manual.

http://www.ats.ucla.edu/stat/mult_pkg/whatstat/default.htm

The following link provides a range of general guidelines regarding the type of statistical analyses of and tips about the application of various methods using Stata.

https://www.iriseekhout.com/promotie/thesis/

General strategies on how to deal with missing data.

https://eagereyes.org/pie-charts

This blog provides a good description of the advantages and disadvantages related to pie charts.

This book provides an introduction to multivariate analysis and easy to follow discussions of fundamental statistical concepts.

Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2013). *Multivariate data analysis. A global perspective* (7th ed.). Upper Saddle River: Prentice-Hall.

This book teaches you how to make high-quality graphs in Stata.

Mitchel, M.N. *A Visual Guide to Stata Graphics*. (2008). Stata Press: StataCorp LP.

This book teaches readers how to decipher the meaning of symbols, tables, and figures included in research reports in order to improve their ability to critically assess such reports.

Huck, W.S. (2014). Reading statistics and research (6th ed.). Harlow: Pearson.

SticiGui at http://www.stat.berkeley.edu/~stark/SticiGui/Text/correlation.htm.

These websites interactively demonstrate how strong the correlations between different datasets are.

References

Agarwal, C. C. (2013). *Outlier analysis*. New York: Springer.

Agresti, A., & Finlay, B. (2014). *Statistical methods for the social sciences* (4th ed.). London: Pearson.

Barchard, K. A., & Pace, L. A. (2011). Preventing human error: The impact of data entry methods on data accuracy and statistical results. *Computers in Human Behavior, 27*(5), 1834–1839.

Barchard, K. A., & Verenikina, Y. (2013). Improving data accuracy: Electing the best data checking technique. *Computers in Human Behavior, 29*(50), 1917–1912.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143–156.

Carpenter, J., & Kenward, M. (2013). *Multiple imputation and its application*. New York: Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.

Drolet, A. L., & Morrison, D. G. (2001). Do we really need multiple-item measures in service research? *Journal of Service Research, 3*(3), 196–204.

Eekhout, I., de Vet, H. C. W., Twisk, J. W. R., Brand, J. P. L., de Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of Clinical Epidemiology, 67*(3), 335–342.

Gladwell, M. (2008). *Outliers: The story of success*. New York: Little, Brown, and Company.

Graham, J. W. (2012). *Missing data: Analysis and design*. Berlin et al.: Springer.

Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis. A global perspective* (7^{th} ed.). Upper Saddle River: Pearson.

Harzing, A. W. (2005). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross Cultural Management, 6*(2), 243–266.

Johnson, T., Kulesa, P., Lic, I., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles. Evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36*(2), 264–277.

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Thousand Oaks: Sage.

Little, R. J. A. (1998). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*(404), 1198–1202.

Paulsen, A., Overgaard, S., & Lauritsen, J. M. (2012). Quality of data entry using single entry, double entry and automated forms processing – An example based on a study of patient-reported outcomes. *PloS One, 7*(4), e35087.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Sarstedt, M., Diamantopoulos, A., Salzberger, T., & Baumgartner, P. (2016). Selecting single items to measure doubly-concrete constructs: A cautionary tale. *Journal of Business Research, 69*(8), 3159–3167.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30*(4), 377–399.