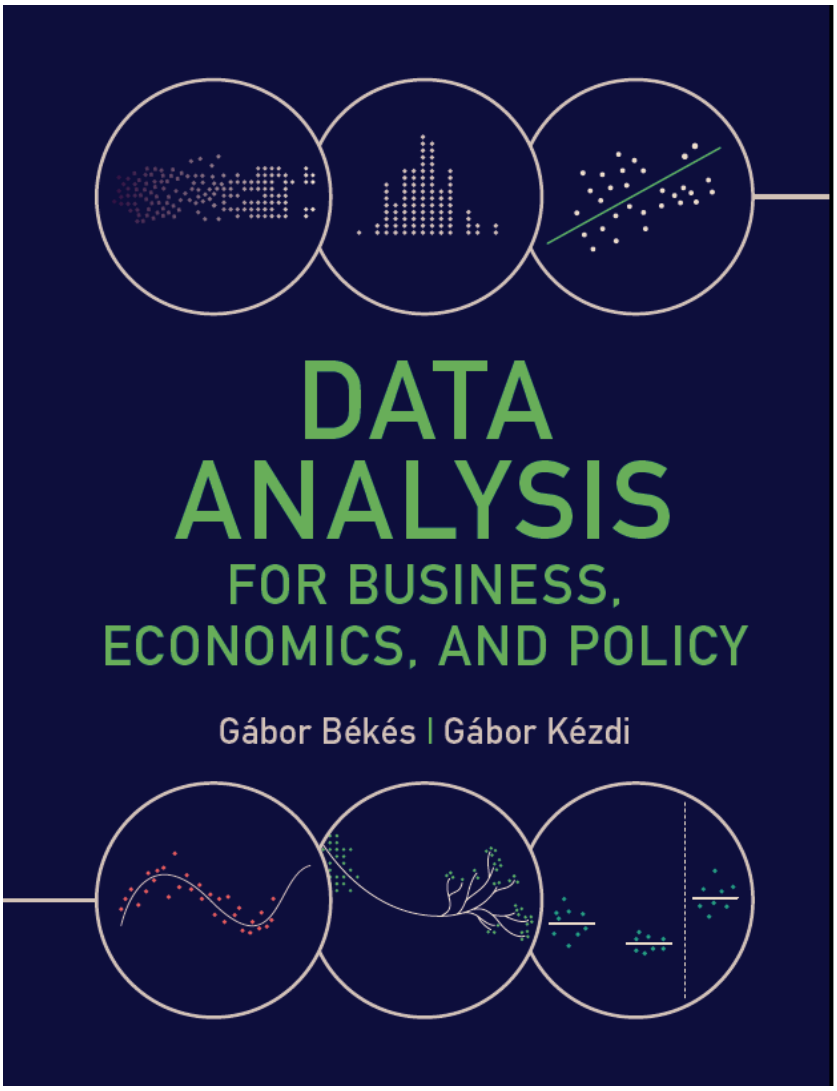
Data Analysis for Business, Economics, and Policy

DRAFT Chapters Text: 2020-02-06. Graphs: 2020-04-30
Under contract with Cambridge University Press (December 2020)
Please do not circulate without author consent!

Gábor Békés (CEU) and **Gábor Kézdi** (U. Michigan)

Wednesday 27th May, 2020



The cover illustrates steps and methods of data analysis. We reorganize messy into tidy datasets, describe our data, and estimate regression to uncover the patterns of association between variables. To make predictions, we fit fluctuations in time series and use machine learning based on decision trees. To learn the causal impact of interventions, we study events.

Part I

Data exploration

Chapter 1

Origins of data

What data is, how to collect it, and how to assess its quality

Motivation

You want to understand whether and by how much online and offline prices differ. To that end you need data on the online and offline prices of the same products. How would you collect such data? In particular, how would you select for which products to collect the data, and how could you make sure that the online and offline prices are for the same products?

The quality of management of companies may be an important determinant of their performance, and it may be affected by a host of important factors, such as ownership or the characteristics of the managers. How would you collect data on the management practices of companies, and how would you measure the quality of those practices? In addition, how would you collect data on other features of the companies?

Part I of our textbook introduces how to think about what kind of data would help answer a question, how to collect such data, and how to start working with data. It also includes chapters that introduce important concepts and tools that are important building blocks of methods that we'll introduce in the rest of the textbook.

We start our textbook by discussing how data is collected, what the most important aspects of data quality are, and how we can assess those aspects. First we introduce data collection methods and data quality because of their prime importance. Data doesn't grow on trees but needs to be collected with a lot of effort, and it's essential to have high quality data to get meaningful answers to our questions. In the end, data quality is determined by how the data was collected. Thus, it's essential for data analysts to understand various data collection methods, how they affect data quality in general, and what the details of the actual collection of their data imply for its quality.

The chapter starts by introducing key concepts of data. It then describes the most important methods of data collection used in business, economics, and policy analysis, such as web scraping, using administrative sources, and conducting surveys. We introduce aspects of data quality, such as validity and reliability of variables and coverage of observations. We discuss how to assess and link data quality to

how the data was collected. We devote a section to Big Data to understand what it is and how it may differ from more traditional data. This chapter also covers sampling, ethical issues, and some good practices in data collection.

This chapter includes three case studies. The case study **Finding a good deal among hotels: data collection** looks at hotel prices in a European city, using data collected from a price comparison website, to help find a good deal: a hotel that is inexpensive relative to its features. It describes the collection of the `hotels-vienna` data. This case study illustrates data collection from online information by web scraping. The second case study, **Comparing online and offline prices: data collection**, describes the `billion-prices` data. The ultimate goal of this case study is comparing online prices and offline prices of the same products, and we'll return to that question later in the textbook. In this chapter we discuss how the data was collected, with an emphasis on what products it covered and how it measured prices. The third case study, **Management quality and firm size: data collection**, is about measuring the quality of management in many organizations in many countries. It describes the `wms-management-survey` data. We'll use this data in subsequent case studies, too. In this chapter we describe this survey, focusing on sampling and the measurement of the abstract concept of management quality. The three case studies illustrate the choices and trade-offs data collection involves, practical issues that may arise during implementation, and how all that may affect data quality.

Learning outcomes. After working through this chapter, you should be able to

- understand the basic aspects of data;
- understand the most important data collection methods;
- assess various aspects of data quality based on how the data was collected;
- understand some of the trade-offs in the design and implementation of data collection;
- carry out a small-scale data collection exercise from the web or through a survey.

1 What is data?

A good definition of data is "factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation" (Merriam-Webster dictionary). According to this definition, information is considered data if its content is based on some measurement ("factual") and if it may be used to support some "reasoning or discussion" either by itself or after structuring, cleaning, and analysis. There is a lot of data out there, and the amount of data, or information that can be turned into data, is growing rapidly. Some of it is easier to get and use for meaningful analysis, some of it requires a lot of work, and some of it may turn out to be useless for answering interesting questions.

An almost universal feature of data is that it rarely comes in a form that can directly help answer our questions. Instead, data analysts need to work a lot with data: structuring, cleaning, and analyzing it. Even after a lot of work, the information and the quality of information contained in the original data determines what conclusions analysts can draw in the end. That's why in this chapter, after introducing the most important elements of data, we focus on data quality and methods of data collection.

Data is most straightforward to analyze if it forms a single **data table**. A data table consists of **observations** and **variables**. Observations are also known as cases. Variables are also called features.

When using the mathematical name for tables, the data table is called the data matrix. A **dataset** is a broader concept that includes, potentially, multiple data tables with different kinds of information to be used in the same analysis. We'll return to working with multiple data tables in Chapter 2.

In a data table, the rows are the observations: each row is a different observation, and whatever is in a row is information about that specific observation. Columns are variables, so that column one is variable one, column two is another variable, etc.

A common file format for data tables is the **csv file** (for "comma separated values"). csv files are text files of a data table, with rows and columns. Rows are separated by end of line signs; columns are separated by a character called a delimiter (often a comma or a semicolon). csv files can be imported in all statistical software.

Variables are identified by names. The data table may have variable names already, and analysts are free to use those names or rename the variables. Personal taste plays a role here: some prefer short names that are easier to work with in code; others prefer long names that are more informative; yet others prefer variable names that refer to something other than their content (such as the question number in a survey questionnaire). It is good practice to include the names of the variables in the first row of a csv data table. The observations start with the second row and go on until the end of the file.

Observations are identified by **identifier** or **ID variables**. An observation is identified by a single ID variable, or by a combination of multiple ID variables. ID variables, or their combinations, should uniquely identify each observation. They may be numeric or text containing letters or other characters. They are usually contained in the first column of data tables.

We use the notation x_i to refer to the value of variable x for observation i, where i typically refers to the position of the observation in the dataset. This way i starts with 1 and goes up to the number of observations in the dataset (often denoted as n or N). In a dataset with n observations, $i = 1, 2, \ldots, n$. (Note that in some programming languages, indexing may start from 0.)

2 Data structures

Observations can have a cross-sectional, time series or a multi-dimensional structure.

Observations in **cross-sectional data**, often abbreviated as **xsec** data, come from the same time, and they refer to different units such as different individuals, families, firms, countries, etc. Ideally, all observations in a cross-sectional dataset are observed at the exact same time. In practice this often means a particular time interval. When that interval is narrow, data analysts treat it as if it were a single point in time.

In most cross-sectional data, the ordering of observations in the dataset does not matter: the first data row may be switched with the second data row, and the information content of the data would be the same. Cross-sectional data has the simplest structure. Therefore we introduce most methods and tools of data analysis using cross-sectional data and turn to other data structures later.

Observations in **time series data** refer to a single unit observed multiple times, such as a shop's monthly sales values. In time series data, there is a natural ordering of the observations, which is typically important for the analysis. A common abbreviation used for time series data is **tseries** data. We shall discuss the specific features of time series data in Chapter 12, where we introduce time series analysis.

Multi-dimensional data, as its name suggests, has more than one dimension. It is also called **panel data**. A common type of panel data has many units, each observed multiple times. Such data is called **longitudinal data**, or cross-section-time-series data, abbreviated as **xt data**. Examples include countries observed repeatedly for several years, data on employees of a firm on a monthly basis, or prices of several company stocks observed on many days.

Multi-dimensional datasets can be represented in table formats in various ways. For xt data, the most convenient format has one observation representing one unit observed at one time (country-year observations, person-month observations, company-day observations) so that one unit (country, employee, company) is represented by multiple observations. In xt data tables, observations are identified by two ID variables: one for the cross-sectional units and one for time. xt data is called **balanced** if all cross-sectional units have observations for the very same time periods. It is called unbalanced if some cross-sectional units are observed more times than others. We shall discuss other specific features of multi-dimensional data in Chapter 24 where we discuss the analysis of panel data in detail.

Another important feature of data is the level of aggregation of observations. Data with information on people may have observations at different levels: age is at the individual level, home location is at the family level, and real estate prices may be available as averages for zip code areas. Data with information on manufacturing firms may have observations at the level of plants, firms as legal entities (possibly with multiple plants), industries with multiple firms, etc. Time-series data on transactions may have observations for each transaction or for transactions aggregated over some time period.

Chapter 2 Section 6 will discuss how to structure data that comes with multiple levels of aggregation and how to prepare such data for analysis. As a guiding principle, the analysis is best done using data aggregated at a level that makes most sense for the decisions examined: if we wish to analyze patterns in customer choices, it is best to use customer-level data; if we are analyzing the effect of firms' decisions, it is best to use firm-level data.

Sometimes data is available at a level of aggregation that is different from the ideal level. If data is too disaggregated (i.e., by establishments within firms when decisions are made at the firm level), we may want to aggregate all variables to the preferred level. If, however, the data is too aggregated (i.e., industry level data when we want firm-level data), there isn't much that can be done. Such data misses potentially important information. Analyzing such data may uncover interesting patterns, but the discrepancy between the ideal level of aggregation and the available level of aggregation may have important consequences for the results and has to be kept in mind throughout the analysis.

Review Box 1.1 ***Structure and elements of data***

- Most datasets are best contained in a data table, or several data tables.
- In a data table, observations are the rows; variables are its columns.
- Notation: x_i refers to the value of variable x for observation i. In a dataset with n observations, $i = 1, 2, \ldots, n$.
- Cross-sectional (xsec) data has information on many units observed at the same time.
- Time series (tseries) data has information on a single unit observed many times.
- Panel data has multiple dimensions – often, many cross-sectional units observed many times (this is also called longitudinal or xt data).

3 A1 Case study – Finding a good deal among hotels: data collection

Introducing the hotels-vienna data

The ultimate goal of our first case study is to use data on all hotels in a city to find good deals: hotels that are underpriced relative to their location and quality. We'll come back to this question and data in subsequent chapters. In the case study of this chapter, our question is how to collect data that we can then use to answer our question.

Comprehensive data on hotel prices is not available ready made, so we have to collect the data ourselves. The data we'll use was collected from a price comparison website using a web scraping algorithm (see more in Section 8).

The `hotels-vienna` data contains information on hotels in one city, Vienna, and one weekday night, November 2017. For each hotel, the data includes information on the name and address of the hotel, the price on the night in focus in US dollars (USD), average customer rating from two sources plus the corresponding number of such ratings, stars of the hotel, distance to the city center, and distance to the main railway station.

The data includes $N = 428$ accommodations in Vienna. Each row refers to a separate accommodation (hotels, hostels and other lodgings). All prices refer to the same weekday night in November 2017, and the data was downloaded at the same time (within one minute). Both are important: the price for different nights may be different, and the price for the same night at the same hotel may change if looked up at a different time. Our dataset has both of these time points fixed. It is therefore a cross-section of hotels – the variables with index i denote individual accommodations, and $i = 1...428$.

The data comes in a single data table, in csv format. The data table has 429 rows: the top row for variable names and 428 hotels. After some data cleaning (to be discussed in Chapter 2, Section 15), the data table has 25 columns corresponding to 25 variables.

The first column is a hotel_id uniquely identifying hotels in the dataset. This is a technical number without actual meaning. We created this variable to replace hotel names, for confidentiality reasons (see more on this in Section 19 in this chapter). Uniqueness of the identifying number is key here: every hotel has a different number. See more about such identifiers in Chapter 02 Section 3.

The second column is a variable that describes the type of the accommodation (i.e., hotel, hostel, or bed-and-breakfast), and the following columns are variables with the name of the city (two versions), distance to the city center, stars of the hotel, average customer rating collected by the price comparison website, the number of ratings used for that average, and price. Other variables contain information regarding the night of stay such as a weekday flag, month, and year, and the size of promotional offer if any. The file VARIABLES.xls has all the information on variables.

Table 1.1 shows what the data table looks like. The variables have short names that are meant to convey their content.

Table 1.1: List of observations

hotel_id	accom_type	country	city	city_actual	dist	stars	rating	price
21894	Apartment	Austria	Vienna	Vienna	2.7	4	4.4	81
21897	Hotel	Austria	Vienna	Vienna	1.7	4	3.9	81
21901	Hotel	Austria	Vienna	Vienna	1.4	4	3.7	85
21902	Hotel	Austria	Vienna	Vienna	1.7	3	4	83
21903	Hotel	Austria	Vienna	Vienna	1.2	4	3.9	82

Note: *List of five observations with key variable values. accom_type is the type of accommodation. city is the city based on the search; city_actual is the municipality.*
Source: `hotels-vienna` dataset. Vienna, For a 2017 November weekday

4 Data quality

Data analysts should know their data. They should know how the data was born, with all details of measurement that may be relevant for their analysis. They should know their data better than their audience. Few things have more devastating consequences for a data analyst's reputation than someone in the audience pointing out serious measurement issues the analyst didn't consider.

Garbage in – garbage out. This summarizes the prime importance of data quality. The results of an analysis cannot be better than the data it uses. If our data is useless to answer our question, the results of our analysis are bound to be useless, no matter how fancy a method we apply to it. Conversely, with excellent data even the simplest methods may deliver very useful results. Sophisticated data analysis may uncover patterns from complicated and messy data but only if the information is there.

Below we list specific aspects of data quality. Good data collection pays attention to these as much as possible. This list should guide data analysts on what they should know about the data they use. This is our checklist. Other people may add more items, define specific items in different ways, or de-emphasize some items. We think that our version includes the most important aspects of data quality organized in a meaningful way. We shall illustrate the use of this list by applying it in the context of the data collection methods and case studies in this book.

Table 1.2: Key aspects of data quality

Aspect	Explanation
Content	The content of a variable is determined by how it was measured, not by what it was meant to measure. As a consequence, just because a variable is given a particular name, it does not necessarily measure that.
Validity	The content of a variable (actual content) should be as close as possible to what it is meant to measure (intended content).
Reliability	Measurement of a variable should be stable, leading to the same value if measured the same way again.
Comparability	A variable should be measured the same way for all observations.
Coverage	Ideally, observations in the collected dataset should include all of those that were intended to be covered (complete coverage). In practice, they may not (incomplete coverage).
Unbiased selection	If coverage is incomplete, the observations that are included should be similar to all observations that were intended to be covered (and, thus, to those that are left uncovered).

We should note that in real life, there are problems with even the highest-quality datasets. But the existence of data problems should not deter someone from using a dataset. Nothing is perfect. It will be our job to understand the possible problems and how they affect our analysis and the conclusions we can draw from our analysis.

The following two case studies illustrate how data collection may affect data quality. In both cases, analysts carried out the data collection with specific questions in mind. After introducing the data collection projects, we shall, in subsequent sections, discuss the data collection in detail and how its various features may affect data quality. Here we start by describing the aim of each project and discussing the most important questions of data quality it had to address.

A final point on quality: as we would expect, high quality data may well be costly to gather. These case study projects were initiated by analysts who wanted answers to questions that required collecting new data. As data analysts, we often find ourselves in such a situation. Whether collecting our own data is feasible depends on its costs, difficulty, and the resources available to us. Collecting data on hotels from a website is relatively inexpensive and simple (especially for someone with the necessary coding skills). Collecting online and offline prices and collecting data on the quality of management practices are expensive and highly complex projects that required teams of experts to work together for many years. It takes a lot of effort, resources, and luck to be able to collect such complex data; but, as these examples show, it's not impossible.

Review Box 1.2 ***Data quality***

Important aspects of data quality include:

- Content of variables: what they truly measure;
- Validity of variables: whether they measure what they are supposed to;
- Reliability of variables: whether they would lead to the same value if measured the same way again;
- Comparability of variables: the extent to which they are measured the same way across different observations;
- Coverage is complete if all observations that were intended to be included are in the data.
- There is no selection bias in data having incomplete coverage if observations in the data are not systematically different from the total.

5 B1 Case study – Comparing online and offline prices: Data collection

Introducing the billion dollar price data

The second case study is about comparing online prices and offline prices of the same products. Potential differences between online and offline prices are interesting for many reasons, including making better purchase choices, understanding the business practices of retailers, and using online data in approximating offline prices for policy analysis.

The main question is how to collect data that would allow us to compare online and offline (i.e., in-store) prices for the very same product. The hard task is to ensure that we capture many products and that they are actually the same product in both sources.

The data was collected as part of the Billion Prices Project (BPP; http://www.thebillionpricesproject.com), an umbrella of multiple projects that collect price data for various purposes using various methods. The online-offline project combines several data collection methods, including data collected from the web and data collected "offline" by visiting physical stores.

BPP is about measuring prices for the same products sold through different channels. The two main issues are identifying products (are they really the same?) and recording their prices. The actual content of the price variable is the price as recorded for the product that was identified. Errors in product identification or in entering the price would lower the validity of the price measures. Recording the prices of two similar products that are not the same would be an issue, and so would be recording the wrong price (e.g., do recorded prices include taxes or temporary sales?).

The reliability of the price variable also depends on these issues (would a different measurement pick the same product and measure its price the same way?) as well as inherent variability in prices. If

prices change very frequently, any particular measurement would have imperfect reliability. The extent to which the price data are comparable across observations is influenced by the extent to which the products are identified the same way and the prices are recorded the same way.

Coverage of products is an important decision of the price comparison project. Conclusions from any analysis would refer to the kinds of products the data covers.

6 C1 Case study – Management quality and firm performance: Data collection

Introducing the wms-management-survey data

The third case study is about measuring the quality of management in organizations. The quality of management practices are understood to be an important determinant of the success of firms, hospitals, schools, and many other organizations. Yet there is little comparable evidence of such practices across firms, organizations, sectors, or countries.

There are two research questions here: how to collect data on management quality of a firm and how to measure management practices themselves. Similarly to previous case studies, no such dataset existed before the project although management consultancies have had experience in studying management quality at firms they have advised.

The data for this case study is from a large-scale research project aiming to fill this gap. The World Management Survey (WMS; http://worldmanagementsurvey.org) collects data on management practices from many firms and other organizations across various industries and countries. This is a major international survey that combines a traditional survey methodology with other methods.

The most important variables in the WMS are the management practice "scores." Eighteen such scores are in the data, each measuring the quality of management practices in an important area, such as tracking and reviewing performance, the time horizon and breadth of targets, or attracting and retaining human capital. The scores range from 1 through 5, with 1 indicating worst practice and 5 indicating best practice. Importantly, this is the intended content of the variable. The actual content is determined by how it is measured: what information is used to construct the score, where that information comes from, how the scores are constructed from that information, whether there is room for error in that process, etc.

Having a good understanding of the actual content of these measures will inform us about their validity: how close actual content is to intended content. The details of measurement will help us assess their reliability, too: if measured again, would we get the same score or maybe a different one? Similarly, those details would inform us about the extent to which the scores are comparable – i.e., they measure the same thing, across organizations, sectors, and countries.

The goal of the WMS is to measure and compare the quality of management practices across organizations in various sectors and countries. In principle the WMS could have collected data from all organizations in all sectors and countries it targeted. Such complete coverage would have been prohibitively expensive. Instead, the survey covers a sample: a small subset of all organizations. Therefore, we need to assess whether this sample gives a good picture of the management practices of all organizations – or, in other words, if selection is unbiased. For this we need to learn how the organizations covered were selected, a question we'll return to in Section 14 below.

7 How data is born: The big picture

Data can be collected for the purpose of the analysis, or it can be derived from information collected for other purposes.

The structure and content of data purposely collected for the analysis are usually better suited to analysis. Such data is more likely to include variables that are the focus of the analysis, measured in a way that best suits the analysis, and structured in a way that is convenient for the analysis. Frequent methods to collect data include scraping the web for information web scraping or conducting a survey (see Section 8 and Section 11).

Data collected for other purposes can be also very useful to answer our inquiries. Data collected for the purpose of administering, monitoring, or controlling processes in business, public administration, or other environments are called administrative data ("admin" data). If they are related to transactions, they are also called transaction data. Examples include payment, promotion, and training data of employees of a firm; transactions using credit cards issued by a bank; and personal income tax forms submitted in a country.

Admin data usually cover a complete population: all employees in a firm, all customers of a bank, or all tax filers in a country. A special case is Big Data, to be discussed in more detail in Section 17, which may have its specific promises and issues due to its size and other characteristics.

Often, data collected for other purposes is available at low cost for many observations. At the same time, the structure and content of such data are usually further away from the needs of the analysis compared to purposely collected data. This trade-off has consequences that vary across data, methods, and questions to be answered.

Data quality is determined by how the data was born, and data collection affects various aspects of data quality in different ways. For example, validity of the most important variables tends to be higher in purposely collected data, while coverage tends to be more complete in admin data. However, that's not always the case, and even when it is, we shouldn't think in terms of extremes. Instead, it is best to think of these issues as part of a continuum. For example, we rarely have the variables we ideally want even if we collected the data for the purpose of the analysis, and admin data may have variables with high validity for our purposes. Or, purposely collected data may have incomplete coverage but without much selection bias, whereas admin data may be closer to complete coverage but may have severe selection bias for the omitted observations.

However the data was born, its value may increase if it can be used together with information collected elsewhere. Linking data from different sources can result in very valuable datasets. The purpose of linking data is to leverage the advantages of each while compensating for some of their disadvantages. Different datasets may include different variables that may offer excellent opportunities for analysis when combined even if they would be less valuable on their own.

Data may be linked at the level of observations, for the same firms, individuals, or countries. Alternatively, data may be linked at different levels of aggregation: industry-level information linked to firms, zip-code-level information linked to individuals, etc. We shall discuss the technical details of linking data tables in Chapter 02, Section 8. In the end, linkages are rarely perfect: there are usually observations that cannot be linked. Therefore, when working with linked data, data analysts should worry about coverage and selection bias: how many observations are missed by imperfect linking, and whether the included and missing observations are different in important ways.

A promising case of data linkage is a large administrative dataset complemented with data collected

for the purpose of the analysis, perhaps at a smaller scale. The variables in the large but inexpensive data may allow uncovering some important patterns, but they may not be enough to gain a deeper understanding of those patterns. Collecting additional data for a subset of the observations may provide valuable insights at extra cost, but keeping this additional data collection small can keep those costs contained.

For example, gender differences in earnings at a company may be best analyzed by linking two kinds of data. Admin data may provide variables describing current and previous earnings and job titles for all employees. But it may not have information on previous jobs, skill qualifications, or family circumstances, all of which may be relevant for gender differences in what kind of jobs employees have and how much they earn. If we are lucky, we may be able to collect such information through a survey that we administer to all employees, or to some of them (called a sample, see Section 13). To answer some questions, such as the extent of gender differences, analyzing the admin data may suffice. To answer other questions, such as potential drivers of such differences, we may need to analyze the survey data linked to the admin data.

8 Collecting data from existing sources

Data collected from existing sources, for a purpose other than our analysis, may come in many forms. Analysis of such data is called secondary analysis of data. One type of such data is purposely collected to do some other analysis, and we are re-using it for our own purposes. Another type is collected with a general research purpose to facilitate many kinds of data analysis. These kinds of data are usually close to what we would collect for our purposes.

Some international organizations, governments, central banks, and some other organizations collect and store data to be used for analysis. Often, such data is available free of charge. For example, the World Bank collects many time series of government finances, business activity, health, and many others, for all countries. We shall use some of that data in our case studies. Another example is FRED, collected and stored by the U.S. Federal Reserve system, that includes economic time series data on the U.S.A. and some other countries.

One way to gather information from such providers is to visit their website and download a data table – say, on GDP for countries in a year, or population for countries for many years. Then we import that data table into our software. However, some of these data providers allow direct computer access to their data. Instead of clicking and downloading, data analysts may use an Application Programming Interface, or **API**, to directly load data into a statistical software package. An API is a software intermediary, or an interface, that allows programs, or scripts, to talk to each other. Using an API, data analysts may load these datasets into their statistical software as part of the code they write for that software.

Besides data collected and provided for the purposes of analysis, there is a lot of information out there that can be turned into useful data even though it is not collected with a purpose of analysis. Here we discuss the two most important such sources: information on the web and administrative data. The emergence of Big Data (see Section 17) is due to the availability and use of more and more such information.

Collecting data from the web can yield useful data for many kinds of analysis as more and more economic and social activity takes place on the web or leaves an information trace there. Examples include collecting price data from classified ads or price comparison websites, collecting data on the results of sports tournaments, and collecting data on the frequency of online web search for certain

words.

In principle we can collect data from the web manually, by creating a data table and entering relevant information by hand. Manual data collection may make sense when the number of observations is very small. Whenever possible, though, automated data collection is superior. It involves writing code that collects all relevant data and puts it into an appropriate data table. Collecting data from the web using code is called **web scraping**. Well-written web-scraping code can load and extract data from multiple web pages. Some websites are easier to scrape than others, depending on how they structure and present information. There are many web scraping software solutions available, and there is a lot of help available online. In the end, scraping is writing code so it requires both general experience in coding and learning the specifics of scraping.

Collecting **data from administrative sources** is another important case. As we discussed briefly in Section 7, a lot of information, often in digital format, is collected for administrative purposes in business organizations and governments.

Important advantages of most administrative data include high reliability of the variables they measure, and high, often complete, coverage, which also leads to large size in most cases. Data on employees in a firm tends to include all employees, typically going back in time for many years. Such data may contain many variables with respect to their performance and pay at the firm. Credit card transactions may cover all customers of a financial firm and include all of their transactions, including the identity of the sellers. Individual tax data usually covers all tax filers in a country with all income information entered on the tax form, and perhaps some characteristics of the individual tax filers such as their age and location of residence.

Admin data tends to have two disadvantages, First, typically, it includes few variables and misses many that may be useful for analysis. Second, important variables may have low validity: their content may be quite different from what analysts would want to measure. Employee records in firms contain little information on other jobs, previous employment, or family characteristics. Credit card transaction data has no information on the income, wealth, and other expenditures of the customer. The income variable in individual tax records may be very different from total family income because the tax records of spouses and other family members may not be linked, and some income sources may not be reported.

All advantages and disadvantages of admin data stem from the original purpose of administrative data: facilitating administration in an organization as opposed to lending itself to analysis to answer the specific questions analysts might ask. Data analysts should pay special attention to the content and validity of variables in admin data.

Review Box 1.3 ***Collecting data from existing sources***

- Web scraping is using code to retrieve information available on the web;
- Administrative sources of data collect information for goals other than data analysis;
- Frequent advantages of data collected from existing sources: low costs (especially low marginal costs), large datasets;
- Frequent disadvantages: fewer variables, lower validity of variables.

9 A2 Case Study – Finding a good deal among hotels: data collection

How the data was born

The dataset on hotels in Vienna was collected from a price comparison website, by web scraping. The purpose of the website is not facilitating data analysis. It is to offer a list of hotels with prices and other features to customers and induce them to choose one of their offerings. Customers are requested to enter a date for check-in, a date for check-out, number of visitors, and, optionally, other information on their planned stay.

The price the website lists is the price that customers pay if they choose to make a reservation right away. The website lists all hotels from its registry that have vacancies and meet the criteria. Not all hotels in Vienna may be covered by the website, and not all that are covered may have rooms available for the date in question. When listing hotels, the website shows not only the variables we collected but also photos and other information that our data does not include.

Many of these features have important consequences for the analysis. The dataset was collected on a specific date. The analysis may take time and results will be available on some later date. Prices are valid on the date of data collection; they may change by the time the results of the analysis are available. That may decrease the validity of the price variable for the purpose of the analysis. The data does not contain all information available on the website. Most importantly, it does not use the photos even though those are likely important determinants of the decision. If the analysis is carried out to help find a good price, we should not let the analysis result in a single choice: what looks best in the data may not be the best if prices change or the photos are examined. Instead, it is probably best to produce a shortlist with several candidates and do manual checking of those options to find the best deal from among them.

Coverage of hotels in Vienna is not complete. First, the data contains only hotels with rooms available on the night asked. That is fine if we are analyzing the data to choose a hotel for that night. It may be less fine for other purposes, such as understanding the pricing practices of hotels. Another concern is that not all hotels in the city are covered by the website. That may be an issue as other hotels may offer better deals: that's what selection bias would mean in this case. The only way to check this is to use data from other sources. This case study is restricted to this particular data source. We would need to collect data from other sources to assess how incomplete the coverage is and whether that affects the conclusions of our analysis.

The hotel data looks rather useful and offers a great many options to study hotel features that would affect the room price. It is good quality and rather reliable, and we did manage to capture prices at a given moment. However, it unfortunately does not include detailed descriptions of hotel quality, services, or how the room looks like.

10 B2 Case Study – Comparing online and offline prices: data collection

How the data was born

The BPP online-offline price project collected data, in ten countries, from retailers that sold their products both online and offline (in physical stores). (We'll discuss the selection of stores and products later.) Only those retailers were selected that had a unique product identificaion number both online

and in their stores. The unique identifiers ensured that online and offline price data was collected on the same products.

The project was managed by a team of researchers and research assistants (the “BPP team"). They hired data collectors in each country using online crowdsourcing platforms (Mechanical Turk, Elane) to carry out the offline data collection. Data collectors were responsible for selecting products in the physical stores they were assigned to (more on selection later). They received a mobile phone app designed to scan prices in physical stores in a quick and standardized fashion. With the help of this mobile app, the data collectors scanned the barcode of products, marked whether the product was on promotion or this was a regular price, and took a photo of the price tag. The data entered in the mobile app was synchronized with a central server and was verified by a member of the BPP team. Once the offline price data was collected, the BPP team searched for and entered the online price of the same product.

Online prices were collected on, or shortly after, the day when the corresponding offline prices were entered. Whether taxes were included or not was determined by how prices were presented in offline stores (without sales taxes in the U.S.A.; with value-added tax and, potentially, sales tax in most other countries).

The unique product identifiers ensured that offline and online prices were collected for the exact same products. The manual entry of prices might have left room for error, but the price tag photos ensured that those errors were rare. The data collection harmonized online and offline prices in terms of whether the products were on promotion and whether taxes were included. Shipping costs were not included, and neither were transportation costs for visiting physical stores.

This data collection procedure ensures high validity and high reliability for the price data collected. Comparability of the price data across products and stores is likely high, too. Frequent changes of prices may make measurement result in different prices if carried out at different times, but online and offline data were collected on the same, or very close, dates so the effect of such changes were minimal for the comparison of online and offline prices. Note that the time difference between recording the online and offline price of a product may be informative about the reliability of these variables. Similarly, institutional differences across countries may make the content of price differ (e.g., whether taxes are included), but those differences affect online and offline prices the same way.

11 Surveys

Surveys collect data by asking people questions and recording their answers. Typically, the answers are short and easily transformed into variables, either qualitative (factor) or quantitative (numerical; see 2 Section 1 for information about variable types). The people answering questions are called respondents. The set of questions presented to respondents is called a questionnaire. There are two major kinds of surveys: self-administered surveys and interviews.

In self-administered surveys, respondents are left on their own to answer questions. Typically, the questions are presented to them in writing. Web surveys are the most important example. Respondents see the questions on the screen of a web-connected device, and they tap, click, or enter text to answer the questions. The answers are immediately recorded and transformed into an appropriate data table. This feature is an important advantage of web surveys: there is no need for anyone else to enter the data or put it into an appropriate structure. That means lower costs and less room for error.

Respondents need to be recruited to participate in web surveys just like in any other survey. Apart

from that, however, web surveys have low marginal costs: once the web questionnaire is up and running, having more respondents answer them incurs practically no extra cost. Before the web, self-administered surveys were done on paper, and respondents were most often required to mail the filled out questionnaires. That method entailed higher costs and, arguably, more burden on respondents.

Besides low costs and quick access to data, web surveys have other advantages. They can present questions with visual aids and interactive tools, and they can embed videos and sound. They may be programmed to adapt the list and content of questions based on what respondents entered previously. Web surveys can include checks to ensure cleaner data. For example, they can give alerts to respondents if their answers do not seem to make sense or are inconsistent with information given earlier.

A disadvantage of self-administered questionnaires is that they leave little room for clarifying what the questions are after. This may affect the validity of measurement: respondents may answer questions thinking that some questions are about something different from what was intended. Web surveys have an advantage over traditional paper-based surveys in that they can accommodate explanations or help. However, it is up to the respondents to invoke such help, and there is little room to follow up if the help received is not enough.

Another disadvantage of some self-administered surveys is a high risk of incomplete and biased coverage. Potential respondents are left on their own to decide whether to participate, and those that can and choose to participate may be different from everyone else. People without access to the internet can't answer web surveys. People who can't read well can't answer self-administered surveys that are presented in writing. The coverage issue is more severe in some cases (e.g., surveys of children, of the elderly, or in developing countries) than in others (e.g., surveys of university students). Moreover, when respondents are left on their own, they may be less likely to participate in surveys than when someone is there to talk them into it. With ingenuity and investment these issues may be mitigated (offering web connection, presenting questions in voice, offering compensation). Nevertheless, incomplete and biased coverage need special attention in the case of self-administered surveys.

Interviews are the other main way to conduct surveys besides self-administration. They create a survey situation with two participants: an interviewer and a respondent. During survey interviews, interviewers ask the questions of the survey and nothing else. In a broader sense, interviews may include freer conversations, but here we focus on surveys.

Interviews may be conducted in-person or over the telephone (Skype, etc.). Modern interviews are often done with the help of a laptop, tablet, or other device. Such computer-aided interviews share some advantages with web surveys. They allow for visualization, videos and voice, checks on admissible answers, and consistency across answers. With answers entered into such devices, they can then produce data tables that are ready to use.

An advantage of interviewers is the potential for high validity. Interviewers can, and are often instructed to, help respondents understand questions as they are intended to be understood. Interviewers may also help convince respondents to participate in surveys thus leading to better coverage.

At the same time, comparability of answers may be an issue with interviews. Different interviewers may ask the same survey question in different ways, add different details, may help in different ways, or record answers differently. All this may result in interviewer effects: systematic differences between answers recorded by different interviewers even if the underlying variables have no such differences. It is good practice to mitigate these interviewer effects during data collection by precise instructions to interviewers and thorough training.

The main disadvantage of interviews is their cost. Interviewers need to be compensated for the time they spend recruiting respondents, interviewing them, and, in the case of personal interviews, trav-

eling to meet them. Interviews are thus substantially more expensive than self-administered surveys, especially if they invest in insuring high data quality by using computer-aided techniques and intensive training.

Mixed-mode surveys use different methods within the same survey: telephone for some and web for others; in-person for some and telephone for others, etc. Sometimes the same person is asked to answer different questions in different survey modes. Sometimes different people are asked to participate in different modes. Usually, the idea behind mixed mode surveys is saving on costs while maintaining appropriate coverage. They allow for data to be collected at lower costs for some variables, or some observations, using the more costly survey mode only when needed. Comparability may be an issue in mixed-mode surveys when different people answer the same question in different modes. Extensive research shows that answers to many kinds of questions compare well across survey modes but that some kinds of questions tend to produce less comparable answers.

Review Box 1.4 ***Collecting data by surveys***

- Surveys ask people (respondents) and record their answers;
- In self-administered surveys, such as web surveys, respondents answer questions on their own;
- Interviews (personal, telephone, etc.) involve interviewers as well as respondents;
- Mixed-mode surveys use multiple ways for different respondents or different parts of the survey for the same respondents.

12 C2 Case Study – Management quality and firm size: data collection

How the data was born

The World Management Survey is a telephone interview survey conducted in multiple countries. The interviewers were graduate students in management. All participated in a lengthy and intensive training. An important task for interviewers was to select the respondents within each firm that were knowledgeable about management practices. To make this selection comparable, interviewers had to follow instructions. How to apply those rules in particular situations was practiced during the training sessions.

The key variables of the survey are the 18 management practice scores. The scores were assigned by the interviewers after collecting information on each area. For each of the 18 areas, interviewers were instructed to ask a series of questions. Each interviewer had to read out the exact same questions. Then they recorded the answers and assigned scores based on those answers. The assignment of scores for each area was laid out by the survey in detail with examples (e.g., for tracking performance, score 1 has to be assigned if “measures tracked do not indicate directly if overall business objectives are being met. (...)"). Interviewers practiced applying these rules during the training sessions.

The content of each score is therefore based on information that the interviewers gathered in a standardized way and translated to scores using standardized rules. Their validity, reliability, and comparability is difficult to assess without further investigation. Nevertheless, it is safe to say that they

have substantially higher validity and are more comparable across observations than an alternative measure: asking each respondent to score their own management practices.

13 Sampling

Sometimes data is collected on all possible observations, attempting complete coverage. This makes sense when we are target few observations (e.g., employees of a medium-sized firm) or the marginal cost of data collection is negligible (as with web scraping). Often, though, finding more observations may be costly (e.g., recruiting respondents for surveys), and collecting data on new observations may also have high costs (e.g., additional personal interviews). In such cases it makes sense to gather information on only a subset of all potential observations. Data collection here is preceded by **sampling**: selecting observations for which data should be collected. The set of observations on which data is collected in the end is called the **sample**. The larger set of observations from which a sample is selected is called the **population** or universe.

Samples have to represent the population. A sample is **representative** if the distribution of all variables in the sample are the same as, or very close to, their corresponding distribution in the population. (The distribution of variables is the frequency of their values, e.g., fraction female, percent with income within a certain range. Chapter 3 Section 2 will discuss distributions in more detail.) A representative sample of products in a supermarket has the same distribution of prices, sales, frequency of purchase etc. as all products in the supermarket. A representative sample of transactions in a financial institution has the same distribution of value, volume, etc. as when all transactions are considered. A representative sample of workers in an economy has the same distribution of demographic characteristics, skills, wages, etc. as all workers in the economy. Representative samples do not cover all observations in the population, but they are free from selection bias.

Whether a sample is representative is impossible to tell directly. We don't know the value of all variables for all observations in the population, otherwise we would not need to collect data from a sample in the first place. There are two ways of assessing whether a sample is representative: evaluating the data collection process and, if possible, benchmarking the few variables for which we know the distribution in the population.

Benchmarking looks at variables for which we know something in the population. We can benchmark our sample by comparing the distribution of those variables in the sample to those in the population. One example could be comparing the share of various industries in a sample of companies to the share of industries published by the government, based on data that includes the population of companies.

If this kind of benchmarking reveals substantial differences then the sample is not representative. If it shows similar distributions then the sample is representative for the variable, or variables, used in the comparison. It may or may not be representative for other variables. A successful benchmarking is necessary but not sufficient for a sample to be representative.

The other way to establish whether a sample is representative, besides benchmarking, is evaluating the sampling process. That means understanding how exactly the observations were selected, what rules were supposed to be followed, and to what extent those rules were followed. To understand what good sampling methods are, the next section introduces the concept of random sampling, argues why it leads to representative samples, and provides examples of random samples.

14 Random sampling

Random sampling is the process that most likely leads to representative samples. With the simplest ideal random sampling, all observations in the population have the same chance of being selected into the sample. In practice that chance can vary. Which observations are selected is determined by a random rule. For the purpose of getting representative samples, selection rules are random if they are not related to the distribution of the variables in the data. Textbook examples of random rules include throwing dice or drawing balls from urns.

In practice, most random samples are selected with the help of random numbers generated by computers. These numbers are parts of a sequence of numbers that is built into the computer. The sequence produces numbers without a recognizable pattern. Where the sequence starts is either specified by someone or determined by the date and time the process is launched. In a sense these numbers are not truly random as they always come up the same if started from the same point, in contrast with repeatedly throwing dice or drawing balls from urns. Nevertheless, that is not a real concern here because this selection rule is unrelated to the distribution of variables in any real-life data.

Other methods of random sampling include fixed rules that are unrelated to the distribution of variables in the data. Good examples are selecting people with odd-numbered birth dates (a 50% sample), or people with birthdays on the 15th of every month (approx. 3% sample). Again, these rules may not be viewed as “truly random" in a stricter sense, but that’s not a concern for representation as long as the rules are not related to the variables used in the analysis.

In contrast, non-random sampling methods may lead to selection bias. Non-random sampling methods are related to key variables. In other words, they have a higher or lower likelihood of selecting observations that are different in some key variables. As a result, the selected observations tend to be systematically different from the population they are drawn from.

Consider two examples of non-random sampling methods. Selecting people from the first half of an alphabetic order is likely to lead to selection bias because people with different names may belong to different groups of society. Selecting the most recently established 10% of firms is surely not random for many reasons. One reason is called survivor bias: newly established firms include those that would fail within a short time after their establishment while such firms are not present among older firms. The practice questions will invite you to evaluate particular sampling methods and come up with other good and not-so-good methods.

Random sampling works very well if the sample is large enough. In small samples, it is possible that by chance, we pick observations that are not representative for the population. Consider for instance whether samples represent the age distribution of the population of a city. By picking a sample of two people, the share of young and older people may very well be different from their shares in the entire population. Thus, there is a considerable chance that this sample ends up being not representative even though it’s a random sample. However, in a random sample of a thousand people, the share of young and old people is likely to be very similar to their shares in the population, leading to a representative sample. The larger the sample, the larger the chance of picking a representative sample.

An important, although not necessarily intuitive, fact is that it is the size of the sample that matters and not its size as a proportion of the entire population. A sample of five thousand observations may equally well represent populations of fifty thousand, ten million, or three hundred million.

Quite naturally, the larger the random sample, the better. But real life raises other considerations such as costs and time of data collection. How large a random sample is large enough depends on many things. We shall return to this question when we first discuss inference from samples to populations,

in Chapter 5.

Random sampling is the best method of producing representative samples. True, it is not bullet-proof, with a tiny chance a sample may be way off. But that tiny chance is really tiny, especially for large samples. Nevertheless, the fact that it is not literally bullet-proof makes some people uncomfortable when they first encounter it. In fact, it took a lot of evidence to convince most data users of the merits of random sampling.

In practice, sampling often starts with a sampling frame: the list of all observations from which the sample is to be drawn. Incomplete and biased coverage may arise at this stage: the sampling frame may not include the entire population or may include observations that are not part of the population to be represented.

Ideally, data is collected from all members of a sample. Often, however, that is not possible. Surveys need respondents who are willing to participate and answer the questions. The fraction of people that were successfully contacted and who answered the questionnaire is called the response rate. A low response rate increases the chance of selection bias. That is, of course, not necessarily true: a sample with an 80% response rate may be more biased than another sample with a 40% response rate. It is good practice to report response rates with the data description and, if possible, to benchmark variables available in the population.

Review Box 1.5 ***Basic concepts of sampling***

- The set of all observations relevant for the analysis is called the population.
- The subset for which data is collected is called a sample
- A representative sample has very similar distributions of all variables to that of the population.
- Benchmarking statistics available both in the sample and the population helps in assessing the representative nature of samples.
- Random sampling means selecting observations by a rule that is unrelated to any variable in the data.
- Random sampling is the best way to get a representative sample.
- Incomplete sampling frames and non-response are frequent issues; whether they affect the representative nature of the sample needs to be assessed.

15 B3 Case Study – Comparing online and offline prices: data collection

The process of sampling

The BPP online-offline prices project carried out data collection in ten countries. In each country, it selected the largest retailers that sold their products both online and offline and were in the top twenty in terms of market share in their respective countries. The set of all products in these stores is the sampling frame. Sampling of products was done in the physical stores by the data collectors. The

number of products to include was kept small to ensure a smooth process (e.g., to avoid conflicts with store personnel).

The sample of products selected by data collectors may not have been representative of all products in the store. For example, products in more eye-catching displays may have been more likely to be sampled. At the same time, we have no reason to suspect that the online-offline price difference of these products were different from the rest. Thus, the sample of products may very well be representative of the online-offline price differences of the entire population, even though it may not be representative of the products themselves.

This case study asked how to collect data so that we could compare online and offline prices. With careful sampling, and a massive effort to ensure that the very same products are compared, the project did a good job in data collection. A key potential shortcoming is related to external validity: the products collected may not be fully representative of the consumption basket.

16 C3 Case Study – Management quality and firm size: data collection

The process of sampling

The WMS carried out surveys in the U.S.A. and three large European countries. The starting point for the sampling frame was a list of all firms maintained by data providers Compustat in the U.S.A. and the Bureau van Dijk's Orbis/Amadeus in the European countries. The sampling frame was then adjusted by keeping only firms in the targeted sectors and were of medium size (50 to 10,000 employees). The survey took a random sample. The response rate was 54%. This is considered a relatively high rate provided that participation was voluntary and respondents received no compensation.

The data collectors benchmarked many variables and concluded that there were two deviations: response rate was smaller in one country than the rest, and it was usually higher in larger firms. The distribution of other variables were similar in the sampling frame and the sample.

The project aimed to collect systematic data on management quality that could be compared across firms and countries. The way it was set up created a unified data collection process across countries. Systematic checks were introduced to avoid bias in collecting answers. Once again a key possible shortcoming is related to external validity: firms that answered the survey may not fully represent the economy of the country.

17 Big Data

Data, or less structured information that can turned into data, became ubiquitous in the twenty-first century as websites, apps, machines, sensors and other sources, collect and store it in increasing and unfathomable amounts. The resulting information from each source is often massive to an unprecedented scale. For example, scanned barcodes at retail stores can lead to millions of observations for a retail chain in a single day. For another example, Twitter, a social media site, generates 500 million tweets per day that can be turned into data to analyze many questions. The commonly used term for such data is **Big Data**. It is a fairly fluid concept, and there are several definitions around. Nevertheless, most data analysts agree that Big Data provides unique opportunities but also specific challenges that often need extra care and specific tools.

A frequently used definition of Big Data is **the four Vs**: volume (scale of data), variety (different forms), velocity (real time data collection), and veracity (accuracy). The fourth v is actually a question of data quality that we think is better left out from the definition. Rephrasing the first three of the Vs, Big Data refers to data that is

- massive: contains many observations and/or variables;
- complex: does not fit into a single data table;
- continuous: often automatically and continuously collected and stored.

Big Data is massive. Sometimes that means datasets whose size is beyond the ability of typical hardware and software to store, manage, and analyze it. A simpler, and more popular version, is that Big Data is data that we cannot store on our laptop computer.

Big data often has a complex structure, making it rather difficult to convert into data tables. A variety of new types of data appeared that require special analytical tools. For example, networks have observations that are linked to each other, which may be stored in various forms. Maps are multi-dimensional objects with spatial relationship between observations. Text, pictures, or video content may be transformed into data, but its structure is often complex. In particular, text mining has become an important source of information both for social scientists and business data scientists.

Big Data is often automatically and continuously collected. Apps and sensors collect data as part of their routine, continuously updating and storing information. As part of the functioning of social media or the operation of machines such as airplane turbines, all the data is stored.

Big Data almost always arises from existing information as opposed to being collected purposely by analysts. That existing information typically comes from administrative sources, transactions, and as other by-products of day-to-day operations. Thus Big Data may be thought of as admin data with additional characteristics. As a result, it tends to share the advantages and disadvantages of admin data.

A main advantage of Big Data, shared with other kinds of admin data, is that it can be collected from existing sources, which often leads to low costs. The volume, complexity, or continuous updating of information may make data Big Data collection more challenging, but there are usually appropriate methods and new technology to help there.

Coverage of Big Data is often high, sometimes complete. Complete coverage is a major advantage. When coverage is incomplete, though, the left-out observations are typically different from the included ones, leading to selection bias. In other words, Big Data with incomplete coverage is rarely a random sample of the population we'd like to cover. To re-iterate a point we made earlier with respect to non-coverage: higher coverage does not necessarily mean better representation. In fact a relatively small random sample may be more representative of a population than Big Data that covers a large but selected subset of, say, 80%. Thus, if its coverage is incomplete, Big Data may be susceptible to selection bias – something that data analysts need to address.

Another common feature of Big Data, shared with admin data, is that it may very well miss important variables, and the ones included may not have high validity for the purpose of our analysis. That can be a major disadvantage. At the same time, because of the automatic process of information gathering, the variables tend to be measured with high reliability and in comparable ways.

The specific characteristics of Big Data have additional implications for data management and analysis as well as data quality.

The massive size of Big Data can offer new opportunities, but it typically requires advanced technical solutions to collect, structure, store and work with the data. Size may have implications for what data analysis methods are best to use and how to interpret their results. We'll discuss these implications as we go along. Here we just note that sometimes when the data is big because there are many observations, all analysis can be done on a random sample of the observations. Sometimes, going from a massive number of observations (say, billions or trillions) to a large but manageable number of observations (say, a few millions) can make the analysis possible without making any difference to the results. This is not always an option, but when it is, it's one to consider.

We'll see an example for such massive data when we analyze the effect of a merger between two airlines on prices in Chapter 22. The original data is all tickets issued for routes in the U.S.A. A 10% sample of the data is made available for public use (without personal information). Even this sample data is available in multiple data files as it is too large to store on a laptop computer. We'll have to select parts of the data and aggregate it to apply the methods we'll cover in the textbook.

If Big Data is of a complex nature, this has consequences for the management and structuring of the data. Sometimes, with thorough re-structuring, even complex data can be transformed into a single data table that we can work with. For example, connections in a network may be represented in panel data, or features of texts may be summarized by variables, such as the frequency of specific words, that can be stored as a data table. Other times, though, complexity calls for methods that are beyond what we cover in this textbook and may also be beyond the traditional toolkit of statistical analysis. For example, the features of routes on maps or the sound of the human voice are less straightforward to transform into data tables.

With Big Data that is continuously collected and updated, the process of data work and analysis is different from the more traditional way of doing data analysis that we'll focus on in this textbook. For example, instead of having a final data table ready for analysis, such cases require constant updating of the data and with it, updating all analysis as new data comes in. This approach implies some fundamental differences to data analysis that are beyond the scope of this textbook.

In the remainder of this textbook, we will focus on the most common kind of Big Data: very large numbers of observations. In addition, we'll note some issues with data that has a massive number of variables. We'll ignore complexity and continuous data collection. And we won't discuss technical issues such as the need for additional computational power or specific data management tools to store and manipulate data with billions of observations. Our focus will be on what the massive number of observations, or sometimes variables, implies for the substantive conclusions of data analysis. Our list of References and further reading will offer some help in the fast-evolving field of technical solutions.

A final comment: most of the traditional, "small data," issues and solutions we will discuss in this textbook will remain relevant for Big Data as well. We shall always note when that is not the case. Similarly, when relevant, we shall always discuss the additional issues Big Data may imply for the methods and tools we cover.

Review Box 1.6 ***Big Data***

- Big Data is characterized by a massive number of observations or variables.
- Sometimes Big Data is also complex in its structure and/or continuously updated.
- Typically, Big Data is collected for purposes other than analysis. Thus it shares all the advantages and disadvantages of other kinds of admin data.
- Its size, complexity, and continuous updating present additional challenges for its collection, structuring, storage, and analysis.

18 Good practices in data collection

Several good practices in data collection are recognized to increase or help assess data quality. Some are general across many methods; others are specific.

Carrying out one or more pilot studies before data collection is general advice. To pilot a data collection method means to try it out in microcosm before doing the whole thing. Piloting is more powerful the more features of the final data collection are included. In web scraping this may mean small-scale collection of data from all websites across all kinds of items that will be relevant. In web surveys it may include recruiting a few respondents as well as asking them to fill out the entire questionnaire. With complex data collection, piloting may come in several steps, such as identifying the sampling frame, drawing a sample, identifying observations or recruiting respondents, and collecting the data itself by scraping, interviewing, etc. Sometimes these steps are given different names as they get to include more and more parts of the entire data collection process (pilot, pretest, field rehearsal, etc.).

When people are involved in data collection, it is good practice to give them precise instructions to follow. An important objective of these is to get the actual content of measured variables as close as possible to their intended contents, thus increasing their validity. These practices also help with comparability and reliability by inducing different people to measure similar things in similar ways. For example, in interview-based surveys, precise instructions usually include questions to be read out (as opposed to letting interviewers ask questions using their own words), when and exactly how to clarify things, and how to translate answers into what is to be recorded. Instructions need to be easy to follow, so a balance needs to be found in how detailed and how accessible instructions are.

Another good practice is training people that participate in data collection in how to follow those instructions and how to make other kinds of decisions. Good training involves many hands-on exercises with examples that are likely to come up during data collection. Training of interviewers for complex surveys may take several days and is often very costly. Nevertheless, it is important to give thorough training to people involved in the data collection in order to ensure high data quality.

Less frequent but very useful practices aim at assessing data quality as part the data collection. For example, the validity of measures in surveys may be assessed with the help of cognitive interviews. These ask respondents to explain why they answered a survey question the way they did. Another technique is asking a survey question in slightly different ways to different respondents (or the same respondents) to see if differences in wording that should not matter make a difference in the answers.

A useful practice to evaluate reliability is test-retest measurement: measuring the same thing more than once within the same data collection process. For example, the price of the same product in the same store may be recorded by two people, independent of each other. Or the same question may be asked of the same respondent twice within the same questionnaire, preferably with many questions in-between. Such a test-retest measurement took place within the World Management Survey: it re-interviewed several hundred firms to assess the reliability of the management quality scores.

There are good practices that help assess coverage issues, too. Whether non-response in a survey leads to severe biases may be assessed by giving some of the would-be respondents higher incentives to participate. If that results in a higher response rate, we may compare the distributions of variables across respondents with and without the extra incentives to see if different response rates lead to different distributions.

There are many other techniques and practices that data collection may include to assess various dimensions of data quality. Making use of all is practically impossible. Nevertheless, it can be very useful to include one or two of them if data collectors are concerned with one or two issues in particular. The results of these techniques can not only shed light on the extent of particular issues but they may be used to mitigate their consequences in the course of the analysis.

Very often, data collection is a complex task. Teamwork here is especially useful as designing and implementing data collection may require a wide range of expertise. The more complex the process, the larger the benefits of advice and collaboration. However, even seemingly simple data collection tasks may have issues that inexperienced researchers are not aware of and can result in inferior data quality. Thus, we think it always makes sense to seek advice and, if needed, mentoring during all stages of data collection. Garbage in, garbage out: if the data we collect ends up having crucial flaws, our analysis will not be able to answer our question. It's better to minimize that possibility if we can.

Review Box 1.7 ***Some good practices for data collection***

Piloting data collection

- Assessing the validity and reliability of most important variables by, for example, cognitive interviews or test-retest measurement – when feasible and economical.
- Examining sources of imperfect coverage to assess potential selection bias.
- Working in teams with experts to design data collection.

19 Ethical and legal issues of data collection

Observations in data analyzed in business, economics or policy most often concern people or firms. Collecting and using such data is subject to strong legal and ethical constraints and rules. These constraints and rules are meant to protect the subjects of data collection: the business interests of firms, the physical and mental health, safety, and integrity of people. Observing these constraints and rules is extremely important: breaching them can have severe consequences for the ongoing data collection and beyond. When the rules are not observed, firms or people may decline participation or take legal action during or after data collection. These, in turn, may affect not only the ongoing data collection but also the general attitude of potential respondents toward data collection in society.

One general principle is confidentiality. In general, data users should not be able to learn sensitive information about identifiable subjects of the data collection. Sensitive information includes, but is not restricted to, information that may harm firms or individuals. When the data contains sensitive information, the principle of confidentiality implies that respondents should be impossible to identify. At a minimum that means that data needs to be de-identified: names and addresses should not be part of the dataset. But it is more than that. Some variables, or combinations of variables, may help identify firms or persons even without names and addresses. Ensuring confidentiality also means ensuring that no such combination allows respondents to be identified from observations.

The collection of data on people, or, as sometimes referred to, human subjects, is subject to a large body of regulation at both international and national levels. The regulation originates from medical research, but it has been adapted for data collection for non-medical purposes, too. The most important guiding principles include respect for persons (people should be treated as autonomous human beings), their protection from harm, and their equal treatment during data collection. It is good practice to obtain informed consent from people for collecting data on them. This means not only data collected here and now but also potential linkages to other data sources. In fact, when data collection is supported by research grants from government or non-governmental foundations, these, and many more, principles are required to be observed.

Another general principle is ownership of information. A lot of information and a lot of data is available on the web or offline. However, availability does not imply the right to analyze that data for any purpose. Who owns that information and what the owner permits to be done with that information is not always easy to find out. Nevertheless, one should always aim to understand the details of ownership and usage rights to make sure it is ethical and legal to collect and use the data.

The rules of data collection are complex. One seemingly good way to think about these issues is to consider oneself to be a subject of the data collection and think about what protection one would need to feel safe and be willing to participate. Another seemingly good starting point is to consider whether similar data was collected recently. But these are not enough. Practices may not be OK just because we, as data analyst, would feel comfortable with them, or a recent data collection project got away with them. Instead, it is strongly advised to consult experts in the legal and ethical aspects of data collection.

Review Box 1.8 ***General principles***

- Ethical and legal rules need to be fully observed; consulting experts is good practice before collecting data.
- Important rules include ensuring confidentiality and observing ownership rights.

20 Summary and practice

20.1 Main takeaways

- Know your data: how it was born and what its major advantages and disadvantages are to answer your question.
  - Data quality is determined by how the data was born.
  - Data is stored in data tables, with observations in rows and variables in columns.

20.2 Practice questions

1. What are in the rows and columns of a data table? What are ID variables?
2. What are xsec, tseries, and xt panel data? Give an example for each.
3. What's the validity and what's the reliability of a variable? Give an example of a variable with high validity and one with low validity.
4. What's selection bias? Give an example of data with selection bias and without.
5. List two common advantages of admin data and two potential disadvantages.
6. How can we tell if a sample is representative of a population?
7. List two sampling rules that likely lead to a representative sample and two sampling rules that don't.
8. List three common features of Big Data. Why does each feature make data analysis difficult?
9. An important principle for research is maintaining confidentiality. How can we achieve that when we collect survey data?
10. You want to collect data on the learning habits of students in your data analysis class. List two survey methods that you may use and highlight their advantages and disadvantages.
11. You want to collect data on the friendship network of students in a class. You consider two options: (1) collect their networks of Facebook users using data there (80% of them are on Facebook), or (2) conduct an online survey where they are asked to mark their friends from a list of all students. List arguments for each option, paying attention to representation, costs, and ethical issues.
12. You consider surveying a sample of employees at a large firm. List four selection methods and assess whether each would result in a representative sample.
13. You want to examine the growth of manufacturing firms in a country. You have data on all firms that are listed on the stock exchange. Discuss the potential issues of coverage and its consequences. Does it matter which country it is?
14. 1000 firms are randomly selected from all the SMEs (small and medium enterprises) in a country. What is the population in this example? What's the sample?
15. You are doing a survey about the smoking habits of the students of your university and want to reach a 20% sample. Here are some potential sampling rules; would each lead to a representative sample? Why or why not? (1) Stand at the main entrance and select every fifth entering student. (2) Get the students' email list from the administration and select every fifth person in alphabetic order. (3) The same, but select the first fifth of the students in alphabetic order. (4) The same, but now sort the students according to a random number generated by a computer and select the first fifth of them.

20.3 Data exercises

Easier and/or shorter exercises are denoted with [*] Harder and/or longer exercises are denoted with [**]

1. Take the `hotels-vienna` data used in this chapter and use your computer to pick samples of size 25, 50, and 200. Calculate the simple average of hotel price in each sample and compare them to those in the entire dataset. Repeat this exercise three times and record the results. Comment on how the average varies across samples of different sizes. [*]
2. Choose a course that you take, and design a short survey to measure how much time your classmates spend on this course, broken down by activities (lectures, practice sessions, study groups, individual study time, etc.). Carry out a web survey (using Survey Monkey, Google Forms etc.) among your classmates. Report the response rate and mean of the main variables. Comment on your results. Write a short report on the challenges of designing and executing the survey. [**]
3. Choose two products that are sold in most supermarkets and gas stations. Visit ten retailers and record the price of each product. What difficulties, if any, did you encounter during this data collection? Are there differences in the prices you collected? Do you see an interesting pattern there? [**]
4. Collect data on used cars of a specific make and model (e.g., Ford Focus) in a city, from the web, using a classified ads website or a used cars website as a source. Use web scraping to collect all available data on these cars. (Alternatively, collect the most important variables by hand from the 100 most recently advertised cars.) Write a short report on what you did, how many cars you ended up with in the data, and what difficulties you encountered, if any. [**]
5. Download country–year panel data on three variables ("indicators") of your choice from the World Bank website. Write a short report on what you did, how many countries and years you ended up with in the data, and what difficulties you encountered, if any. [*]

21 References and further reading

On surveys, we recommend Roger Tourangeau & Rasinski (2000). On sampling, a classic book is Kish (1965). A more recent overview is Bethlehem (2009).

On Big Data and data collection in marketing, the review piece Faro & Ohana (2018) is an informative discussion.

Regarding technical issues, there is more on the World Bank API at `datahelpdesk.worldbank.org`

Bloom & Van Reenen (2007), Cavallo (2017), Athey & Imbens (2016), Bandiera et al. (2018), Banerjee & Duflo (2012) Agrawal et al. (2018), Siegel (2013), Rohde & Breuer (2017)

Chapter 2

Preparing data for analysis

How to organize, manage, and clean data

Motivation

What are the benefits of immunization of infants against measles? In particular, does immunization save lives? To answer that question you can use data on immunization rates and mortality in various countries in various years. International organizations collect such data, and a lot more, on many countries for many years. The data is free to download, but it's complex. How should you import, store, organize and use the data to have all relevant information in an accessible format that lends itself to meaningful analysis? And what problems should you look for in the data, how can you identify those problems, and how should you address them?

You want to know who the most successful managers are (as coaches are also called in football, or soccer) in the top English football league. To investigate this question, you have downloaded data on all games played in the league, as well as data on managers, including which team they worked at and when. How should you combine this data to investigate your question? Moreover, how would you uncover whether there are issues with the data that prevent linking the data and investigating it, and how would you address those issues?

Before analyzing their data, data analysts spend a lot of time on organizing, managing, and cleaning it to prepare it for analysis. This is called data wrangling or data munging. It is often said that 80% of data analysis time is spent on these tasks. Data wrangling is an iterative process: we usually start by organizing and cleaning our data, then start doing the analysis, and then go back to the cleaning process as problems emerge during analysis.

This chapter is about preparing data for analysis: how to start working with data. First, we clarify some concepts: types of variables, types of observations, data tables, and datasets. We then turn to the concept of tidy data: data tables with data on the same kinds of observations. We discuss potential issues with observations and variables, and how to deal with those issues. We describe good practices for the process of data cleaning and discuss the additional challenges of working with big data.

This chapter includes three case studies. The first one, **Finding a good deal among hotels: data**

preparation, continues to work towards finding hotels that are underpriced relative to their location and quality. In this chapter, the case study illustrates how to find problems with observations and variables and how to solve those problems. It uses the `hotels-vienna` data. The second case study, called **Identifying successful football managers**, combines information on English football (soccer) games and managers, using the `football` data. We'll use this data in a case study in Chapter 24 to uncover whether replacing football managers improves team performance. This case study illustrates linking data tables with different kinds of observations, problems that may arise with such linkages, and their solutions. The third case study, **Displaying immunization rates across countries**, illustrates how to store multi-dimensional data. It uses the `world-bank-vaccination` data. We'll use this data in a case study to investigate whether immunization saves lives, in Chapter 23.

Learning outcomes. After working through this chapter, you should be able to

- understand types of variables and observations;
- organize data in a tidy way;
- clean the data: identify and address problems with observations and variables;
- create a reproducible workflow to clean and organize data;
- document data cleaning and understand such documentation.

1 Types of variables

Data is made up of observations and variables. Observations are the units for which the information is collected (customers, countries, days when an asset is traded). Variables contain the information (income, size, price). Variables take on specific values. The name variable comes from the fact that they have more than one value: the values vary across observations.

We first discuss the various kinds of variables, and the following sections will discuss the various kinds of observations. We describe types of variables by what kind of information they capture and how the values of the variable are stored in the data. It is useful to understand the variable types as they help understand what we can do with the variables. Sometimes statistical software also asks data analysts to specify the type of each variable.

Quantitative variables are born as numbers, and they are stored as numbers, in numeric format. Typically, they can take many values. Examples include prices, numbers of countries, costs, revenues, age, distance. Date and time are a special case of quantitative variables. They are often measured in specific scales and stored in a specific date/time format.

Qualitative variables, also called **categorical variables** or **factor variables**, are not born as numbers. Instead, their values have a specific interpretation, typically denoting that the observation belongs to a category. Types or brands of products, name of countries, highest levels of education of individuals are examples. Most often, qualitative variables have few values, but sometimes they have many values. A special type of qualitative variables have the sole purpose of identifying observations – they are identifiers, or ID variables. With many observations, an ID variable has many values. (We introduced ID variables earlier in Chapter 1 Section 1.)

Finally, quantitative variables are also called **continuous** and qualitative are also called **discrete** variables. These names come from mathematics, where a continuous variable have values without gaps,

while a discrete variable can have specific values only, with gaps in-between. These labels are a little misleading for real data where few variables are measured in a truly continuous way because of integer units of measurement (dollars, thousand dollars, kilometers, etc.).

The values of qualitative variables are sometimes stored as text, describing the categories. Text in data is also called a **string**. Most data analysts prefer storing qualitative variables as numbers. In that case each number should correspond to a category, and value labels show this correspondence, giving the meaning of each number value. For example, the brand of chocolate chosen by a chocolate customer, a qualitative variable, may be stored as string with the name or abbreviation of the brand, or as a number with appropriate labels (1 = Lindt, 2 = Godiva, etc.).

Binary variables are a special case of qualitative variables: they can take on two values. Most often the information represented by binary variables is a yes/no answer to whether the observation belongs to some group. Examples include whether the respondent to a survey is female or not, whether a firm is in the manufacturing sector or not, etc. For the purpose of data analysis it is best to have them take values 0 or 1: 0 for no, 1 for yes. Binary variables with 0/1 values are also called **dummy variables** or **indicator variables**.

In terms of measurement scale, data analysts often distinguish among four types of data: nominal, ordinal, interval, and ratio. These may be thought as refinements of the qualitative/quantitative classification.

- **Nominal variables** are qualitative variables with values that cannot be unambiguously ordered. Examples include whether the customer purchased a certain brand of a product, or a firm chose a certain location for its headquarters. Individual decision makers may have some ordering of these options, but there is no universally agreed ranking of all options for these types of variables.
- **Ordinal**, or **ordered** variables take on values that are unambiguously ordered. All quantitative variables can be ordered; some qualitative variables can be ordered, too. Examples include subjective health measures (whether someone rates their health as poor, fair, good, very good, or excellent) or the strength of an opinion (e.g., whether one strongly agrees, agrees, disagrees, or strongly disagrees with a statement).
- **Interval variables** have the property that a difference between values means the same thing regardless of the magnitudes. All quantitative variables have this property, but qualitative variables don't have this property. A one degree Celsius difference in temperature is the same when 20 is compared with 21 or 30 is compared with 31. A one dollar price difference of $3 versus $4 is the same a $10 versus $11.
- **Ratio variables**, also known as **scale variables**, are interval variables with the additional property that their ratios mean the same regardless of the magnitudes. This additional property also implies a meaningful zero in the scale. Many but not all quantitative variables have this property. Measures of length, elapsed time, age, value, or size are typically ratio variables. Zero distance is unambiguous, and a 10km run is twice as long as a 5km run. A used car sold for zero dollars costs nothing unambiguously, and a used car sold for $8000 is twice as expensive as one sold for $4000. An example of an interval variable that is not a ratio variable is temperature: 20 degrees is not twice as warm as 10 degrees, be it Celsius of Fahrenheit.

Often, the raw data already has variables that fall into the groups described above, and they are ready for cleaning and then analysis. Sometimes, though, variables need to be created from the information we access. That may be challenging. Such difficult cases include texts, pictures, voice, and videos. For example, variables that we may want to create from text are the frequency of specific words or

the proportion of positive or negative adjectives. Working with such information is a fast developing branch of data science, but it's beyond the scope of our textbook.

Review Box 2.1 ***Types of variables***

- Qualitative (factor, categorical) variables have few values, often denoting a category to which the observation belongs. They may be nominal or ordered.
- Binary (dummy) variables are special case of qualitative variables, with only two values; it's best to have those as 0 and 1.
- Quantitative (numeric) variables are born as numbers and can take many values that are meaningful in themselves. They are always interval variables with meaningful differences, and sometimes they are ratio variables.

2 Stock variables, flow variables

Before turning to observations, let's consider one more way we can distinguish quantitative variables. In business, economics, and policy analysis we often work with quantities that could measure a flow or capture a stock.

Flow variables are results of processes during some time. Typically, they are the result of activities over time. The textbook example of a flow variable is the amount of water in a river that flowed through a reservoir gate yesterday; economic examples include sales of chocolate last month and government deficit last year.

Stock variables refer to quantities at a given point in time. Often, they are a snapshot of a business, a market, an economy, or a society. The textbook example is the amount of water in a reservoir at 8am this morning. Economic examples include the inventory of chocolate in a shop at the end of last month, or the amount of government debt at the end of last year.

The importance of distinguishing flow and stock variables comes from how we typically work with them. For example, their meaningful aggregation differs: often, flow variables are summed (monthly sales to annual sales); stock variables are averaged (average inventory at the end of each month last year). But flow and stock variables may be related: the difference between stocks at different points in time are often related to flow variables. For example, the difference between inventories of a chocolate factory at the end of this month and the end of last month is the difference between chocolate production and chocolate sales during last month, two flow variables.

Not every quantitative variable is either flow nor stock: counterexamples include price and distance. But when a variable is either a flow or a stock, it may be important to keep that in mind.

3 Types of observations

Just as it makes life easier to distinguish various kinds of variables, it makes sense to introduce concepts for types of observations. Recall that observations are the rows of a data table (Chapter 1, Section 1). But in different data tables observations may mean very different things. Those "things" are related to

the actual physical or legal entities, or dimensions, our data is about, such as companies, individuals, days, or transactions. There are two main entity types in economic data analysis: cross-sectional entities and time series entities. These may result in data with cross-sectional observations, time series observations, or observations that have both a cross-sectional and a time series dimension.

In **cross-sectional data** (**xsec data**), observations are cross-sectional entities, such as people, firms, or countries. Such entities are also called cross-sectional units. In xsec data, each variable refers to the same time across units (same month, same year). Cross-sectional units are denoted by index i, so that x_i refers to variable x for cross-sectional unit i. The ID variable in cross-sectional data tables should identify each cross-sectional unit.

In **time series data** (**ts data**), observations are different points or intervals in time, such as closing time of the stock market, or the duration of a calendar year. These time series entities are also called time periods. In ts data, all observations refer to the same cross-sectional unit (person, firm, country), and they are different in terms of the time period of observation. Time series observations are denoted by index t. The ID variables in time series data denote the time period of observation.

Time periods are characterized by their **time series frequency**, also called periodicity. Time series frequency is the time difference between observations. Time series frequency may be yearly, monthly, weekly, daily, hourly, etc. Time series frequency is lower if observations are less frequent; time series frequency is higher if the observations are more frequent. For example, yearly data is of lower frequency than monthly data; weekly data is of higher frequency than monthly data.

Time series frequency may also be irregular. For example, adjacent observations may refer to times of transactions, separated by as much time as happens to pass between the transactions. In fact most frequencies have some irregularities: years and months may have different numbers of days, or the effective time relevant for the variable may be different (such as the number of working days in a year to produce GDP in a country or the number of working hours in a month to produce output at a company). Whether and how one should address such irregularities is to be decided on a case by case basis.

Multi-dimensional data have observations across multiple dimensions. The most common multi-dimensional data is **cross-section time series data** – also called **xt data**, **longitudinal data** or **panel data**. xt data include multiple cross-sectional units observed multiple times. Examples include yearly financial data on all companies in a country, weekly sales at various retail stores, or quarterly macroeconomic data on various countries.

Observations in xt data are one unit observed in one time period. Typically, observations in xt data are referred to using two indices – i for the cross-section and t for the time series – so that x_{it} denotes variable x for xsec unit i at time t.

Table 2.1: Types of observation

Observation type	Observations	ID variable	Example
Cross-sectional (xsec)	x_i: different cross sectional units observed at same time	identifies each cross-sectional unit	People, companies, countries observed in the same year/month
Time series (tseries)	x_t: same cross-sectional unit observed at different time periods	identifies each time period	A single person/company/country observed at different times
Cross-section-time series (xt, longitudinal, or panel)	x_{it} multiple cross-sectional units observed across multiple time periods	one ID identifies cross-sectional units; one ID identifies time periods	retailer stores observed weekly, countries observed yearly

4 Tidy data

After learning about types of variables and observations, let's introduce some concepts on how data is, or should be, organized. A useful guide to organize and storing data is the **tidy data** approach. In this section we'll introduce the principles of tidy data organization.

Data is stored in one or more **data tables**, each consisting of rows of observations and columns of variables. A **dataset** may consist of a single data table or multiple data tables that are related. In short, the tidy data approach prescribes that

1. In a data table, each observation forms a row.
2. In a data table, each variable forms a column.
3. Each kind of observation forms a data table.

The last point means that when the dataset has information on different kinds of observations made up of different kinds of entities, the information on them should be stored in different data tables. For example, a dataset on customer purchases of various products may include a data table on customers (their age, income, etc.), a data table on products (type, brand, price, quality indicators) and a data table on purchase events (which customer purchased which product when).

We have introduced ID variables earlier, in Section 1. They are especially important for datasets with multiple data tables. To keep track of entities within and across data tables, each kind of entity should have its own ID variable, and that should be stored in each data table with information on that kind of entity. The purpose of an ID variable is to uniquely and unambiguously identify each entity in the data table as well as across data tables in a dataset.

Consider our example of a dataset with three data tables, the first one on customers and their characteristics, the second one on products and their characteristics, and the third one on purchase events. Customers need to have their ID that identifies them in the first and the third data table; products need to have their ID that identifies them in the second and the third data table; the third data table should have yet one more ID variable that identifies the time of purchase.

The tidy data approach prescribes that data tables should have one observation in one and only one row. It's a simple principle, but various issues may arise in practice. Sometimes there are non-observation rows in the data table. These should be erased. If such rows contain important information they should be stored in some other way, e.g., in the data documentation. One potential exception is the header row that contains the names of the variables. In spreadsheet and text format, the header occupies the first row in the data table. When read into statistical software, this header row becomes a separate object so that when one looks at the data table, the first row is the first observation.

The tidy data approach has many advantages. It offers a simple guiding principle to organize all kinds of data, including complicated datasets with many data tables. Working with data, such as finding and resolving issues with observations and variables, is easier with tidy data tables. Tidy data tables are also transparent, which helps other users to understand them and work with them. Moreover, tidy data can be extended easily. New observations are added as new rows; new variables as new columns.

Data tables are a fundamental unit of data analysis. All processes of cleaning and analyzing are done within a data table. For the actual analysis, data analysts create a new data table from those tidy data tables. This new data table is called a **workfile**. With a tidy dataset including multiple data tables, this requires combining variables from those multiple data tables, with the help of appropriate ID variables. We will discuss the details of this process in the next few sections.

5 A1 Case Study – Finding a good deal among hotels: data preparation

Types of variables, tidy table

Let us continue our case study from the previous chapter. Recall that the question of the case study is to find a good deal among hotels in a particular European city on a particular night. We collected the data from the web, which we call the `hotels-vienna` data. The dataset consists of information on hotels that have available room for a weekday night in November 2017. We have 428 observations in the dataset. In this chapter we'll learn how to start working with this data. Before doing so, let's consider its most important variables.

The first variable is a number that denotes each hotel. We created these numbers to replace the names of the hotels for confidentiality reasons. It is a qualitative, and nominal variable stored in numeric format. A particular value has no meaning besides identifying a particular hotel. This variable is our identifier (ID).

The type of accommodation is another qualitative and nominal variable. It denotes whether the hotel is in fact a hotel, or a hostel, an apartment-hotel, etc.

Important quantitative variables are price and distance. Both are stored as numeric and are measured in a ratio scale.

Stars is a qualitative variable, and an ordinal one. To earn stars, hotels need to satisfy certain criteria. The more criteria checked, the more stars they have. More stars mean more criteria checked. Stars have values of 1, 2, ... 5, and there are some in-between values, too (2.5, 3.5, 4.5).

Customer rating was originally collected as an ordered qualitative variable, 1 through 5. This allowed the price comparison site to compute the average customer rating. The variable in our data is the average of those individual ratings. It is a number with many values, so we treat it as a quantitative

variable. However, we need to keep in mind that its source is an ordered qualitative variable, thus the same difference may mean different things at different levels (4.2 versus 4.0 is not the same as 3.2 versus 3.0).

The number of ratings that were used for this average is a quantitative variable, and a ratio variable with a meaningful zero.

Table 2.2 summarizes the types of these variables and gives an example of each from the actual data.

Table 2.2: List of the variables in the `hotels-vienna` data

Variable name	**Type**	**Example value**
ID (hotel identifier)	qualitative, nominal	21897
accommodation type	qualitative, nominal	hotel
price	quantitative	81
distance	quantitative	1.7
star	qualitative, ordered	4
average customer rating	assumed quantitative, based on a qualitative, ordered variable	3.9
number of customer ratings	quantitative	189

The data is stored as tidy data. Its rows are observations, its columns are the variables, and this table contains one kind of observations, hotels. The data table is Table 2.3 shows a few rows and columns of a data table, using the `hotels-vienna` data.

Table 2.3: A simple tidy table

	variables/columns		
	hotel_id	price	distance
observations/rows	21897	81	1.7
	21901	85	1.4
	21902	83	1.7

Source: `hotels-vienna` data. Vienna, 2017 November weekday.

6 Tidy approach for multi-dimensional data

Multi-dimensional data may be stored in more than one way in data tables. Let us focus on xt data (cross-section time series data).

The tidy approach recommends storing xt data in a data table with each row referring to one cross-sectional unit observed in one time period. Thus, one row is one observation *it*. This is called the **long format for xt data**. The first row is the first time period of the first cross-sectional unit. The next row is the same cross-sectional unit observed in the next time period. After the last time period observed for this cross-sectional unit, the next row in the data table is the first time period for the next cross-sectional unit. Correspondingly, observations in xt data are identified by two ID variables, i for cross-sectional units, and t for time periods.

An alternative, but not tidy, way of storing multi-dimensional data is called the **wide format for xt data**. Here one row would refer to one cross-sectional unit, and different time periods are represented in different columns. Thus, there are as many columns for each variable as the number of time periods in the data. Here each observation is a different cross-sectional unit to be identified by the cross-sectional identifier i. Time should be "identified" in the names of the variables.

Sometimes doing analysis is easier with wide format, especially if there are only a few time periods. However, the good practice is not to store data in wide format. Instead, the tidy approach prescribes storing multi-dimensional data in long format and transforming it for analysis when necessary. The advantages to the long format for xt data are transparency and ease of management. It is straightforward to add new observations to long format tables, be those new cross-sectional units or new time periods, and it is easier to transform and clean variables

7 B1 Case Study –Displaying immunization rates across countries

Tidy tables

As an example, consider an xt panel of countries with yearly observations, downloaded from the World Development Indicators data website maintained by the World Bank. We'll use this `world-bank-vaccination` data in Chapter 23, Section 9 where we'll try to uncover the extent to which measles immunization saves the lives of children in poor countries. Here we illustrate the data structure focusing on the two ID variables (country and year) and two other variables, immunization rate and GDP per capita.

Table 2.4 show parts of this xt panel data table in long format.Table 2.5 shows the same part of this xt panel data table, but now in wide format. We can see the main difference: in the long format we have multiple rows for a country, as each country is shown three times for the three years we cover. We can also see the advantage of the long and tidy format: if we were to add a new variable such as population, it would be simply adding a new column.

Table 2.4: Country-year panel on immunization and GDP per capita – tidy, long data table

Country	Year	imm	gdppc
India	2015	87	5743
India	2016	88	6145
India	2017	88	6516
Pakistan	2015	75	4459
Pakistan	2016	75	4608
Pakistan	2017	76	4771

Note: *Tidy (long) format of country-year panel data, each row is one country in one year. imm: rate of immunization against measles among 12–13-month-old infants. gdppc: GDP per capital, PPP, constant 2011 USD.N=xxx*
Source: `world-bank-vaccination` data.

Table 2.5: Country–year panel on immunization and GDP per capita – wide data table

Country	imm2015	imm2016	imm2017	gdppc2015	gdppc2016	gdppc2017
India	87	88	88	5743	6145	6516
Pakistan	75	75	76	4459	4608	4771

N
ote: *Wide format of country-year panel data, each row is one country, different years are different variables. imm: rate of immunization against measles among 12–13-month-old infants. gdppc: GDP per capital, PPP, constant 2011 USD.N=xxx*
S
ource: `world-bank-vaccination` data.

8 Relational data and linking data tables across observations

After discussing how to organize and store data in tidy data tables with appropriate ID variables, we turn to how to combine such table into a workfile. Sometimes, with complex data, that can be challenging. We may have macroeconomic data on many countries in many years, geographic information on the same countries for continents with many countries that is the same across time, and variables describing the global economy in each year that are the same for all countries. As another example, we may have data on companies and their managers, with some variables describing companies that may change through time, other variables describing the characteristics of managers that don't change, and data that tell us which manger worked for which company and when.

Relational data is the term often used for such datasets: they have various kinds of observations that are linked to each other through various relations.

Data analysts clean and describe the data tables one by one in a tidy relational dataset. Then, to actually work with such data, they combine the variables they need from the various data tables to form a single data table, their workfile. Sometimes they use all variables from all data tables to create the workfile. More often, they pick certain variables from each data table and work with those only. In either case data analysts need appropriate ID variables to find the same entities across data tables.

The process of pulling different variables from different data tables for well-identified entities to create a new data table is called **linking**, **joining**, **merging**, or **matching** data tables.

To do so, data analysts start with one data table, and they merge to it observations from another data table. The most straightforward linkage is one-to-one (1:1) matching: merging tables with the same type of observations. For example, a data table with customers as observations with their age and income as variables may be merged to another data table on customers with a variable describing how they rated their last visit to the retail store. As another example, a country–year data table with population as a variable may be merged to another country–year data table with GDP per capita as a variable.

Other times we may want to do many-to-one (m:1) matching. For example, we may want to link average family income in zip code locations, a data table with observations on zip codes, to a data table with customers as observations whose zip code of residence is a variable. In the customer-level data table multiple customers may live in the same zip code area, thus the same zip code information may be merged with many customers. As another example, a data table with country–year observations with macro data may be merged with a data table with countries as observations and area as a variable. Here

many country–year observations on the same country are merged with the same country observation.

An alternative is one-to-many (1:m) matching. One-to-many matching can be thought of as the same as many-to-one (m:1) matching but changing the order of the data tables: starting with the data table with the more aggregated observations (zip codes, countries) and merging to it the observations from the other data set (customers, country–year observations).

The most complex case is many-to-many (m:m) matching: one observation in the first table may be matched with many observations in the second table, and an observation in the second data table may be matched with many observations in the first data table. In such an instance, we may need to have a separate table that connects IDs. For example we may want to link information on companies and information on managers. Each company may have had more than one manager during the time span of our data; each manager may have worked for more than one company. In such complicated cases, we often need to think a bit and rephrase the problem as a set of 1:m and m:1 matching problems.

When merging two data tables, be that 1:1, 1:m, m:1, or m:m matching, data analysts may be able to link all observations or only some of them. This latter may happen in two ways: observations in the first data table have no matching observations in the second data table, or observations in the second data table have no matching observations in the first data table. Depending on their goals, data analysts may keep all observations whether matched and unmatched, or they may keep matched observations only.

Review Box 2.2 ***Tidy data tables***

1. Tidy data tables – each observation is a row; each column is a variable.
2. Multi-dimensional data may need to be reformatted to a tidy data version, with an ID for each dimension (such as country–year)
3. Relational datasets have a set of tidy data tables, the observations in which can be linked via IDs.

9 C1 Case Study –Identifying successful football managers

Introducing data and concepts

In this case study, we are interested to identify the most successful football managers in England. We'll extend upon this case study focusing on the impact of replacing managers later, in Chapter 24, Section 6. We combine data from two sources for this analysis, one on teams and games, and one on managers. Our focus in this chapter is how to combine the two data sources and what problems may arise while doing so.

Before we start, let us introduce some concepts. We will talk about football, called soccer in the U.S.A. and a few other countries. In England and some other countries, coaches are called managers – as they take on managerial duties beyond coaching. We will focus on the English Premier League (EPL, for short) – the top football division in England. A season runs for about 9 months, from mid-August to mid-May, and the league consists of 20 teams each season.

Our data covers 11 seasons of EPL games – 2008/2009 to 2018/2019; the data comes from the publicly available Football Data website. At the end of every season, some teams are relegated to the second

division, while some are promoted to join the Premier League from the second division. Some teams feature in all seasons (such as Arsenal, Chelsea, or Manchester United), while others play in the first division just once (such as Cardiff).

Each observation in the data table is a single game. Key variables are the date of the game, the name of the home team, the name of the away team, the goals scored by the home team, and the goals scored by the away team. Each team features 19 times as "home_team" and 19 times as "away_team". We have 380 rows for each season. With 11 seasons, the total number of observations is 4,180. A small snippet of the data is presented in Table 2.6.

Table 2.6: Football game results – game level data

Date	HomeTeam	AwayTeam	Home goals	Away goals
2018-08-19	Brighton	Man United	3	2
2018-08-19	Burnley	Watford	1	3
2018-08-19	Man City	Huddersfield	6	1
2018-08-20	Crystal Palace	Liverpool	0	2
2018-08-25	Arsenal	West Ham	3	1
2018-08-25	Bournemouth	Everton	2	2
2018-08-25	Huddersfield	Cardiff	0	0

Note: *English Premier League, all games, 11 seasons from 2008/9 through 2018/9. One observation is one game. N=4,180*
Source: `football` data.

Is this a tidy data table? It is; each observation is a game, and each game is a separate row in the data table. Three ID variables identify each observation: date, home team, away team. The other variables describe the result of the game. From the two scores we know who won, by what margin, how many goals they scored, and how many goals they conceded.

But there is an alternative way to structure the same data table, which will serve our analysis better – in this data table, each row is a game played by a team. It includes variables from the perspective of that team: when played, who the opponent was, and what the score was. That is also a tidy data table, albeit a different one. It has twice as many rows as the original data table: 760 observations per season, 8,360 observations in total.

Table 2.7 shows a small part of this tidy data table. Each game appears twice in this data table, once for each of the playing team's perspectives. For each row here we had to introduce a new variable to denote whether the team at that game was the home team or the away team. Now we have two ID variables, one denoting the team, and one denoting the date of the game. The identity of the opponent team is a qualitative variable.

Table 2.7: Football game results – team-game level data

Date	Team	Opponent team	Goals	Opponent goals	Home/away	Points
2018-08-19	Brighton	Man United	3	2	home	3
2018-08-19	Burnley	Watford	1	3	home	0
2018-08-19	Man City	Huddersfield	6	1	home	3
2018-08-19	Man United	Brighton	2	3	away	0
2018-08-19	Watford	Burnley	3	1	away	3
2018-08-19	Huddersfield	Man City	1	6	away	0

Note: *English Premier League, all games, 11 seasons from 2008/9 through 2018/9. One observation is one game for one team. N=8,360*
Source: `football` data.

Our second data table is on managers. One row is one manager-team relationship. So each manager may feature more than once in this data table if they worked for multiple teams. For each observation, we have the name of the manager, their nationality, the name of the team (club), the start time of the manager's work at the team, and the end time. Table 2.8 lists a few rows.

Table 2.8: Football managers - spells at teams

Name	Nat.	Club	From	Until
Arsene Wenger	France	Arsenal	1 Oct 1996	13 May 2018
Unai Emery	Spain	Arsenal	23 May 2018	Present*
Ron Atkinson	England	Aston Villa	7 June 1991	10 Nov 1994
Brian Little	England	Aston Villa	25 Nov 1995	24 Feb 1998
John Gregory	England	Aston Villa	25 Feb 1998	24 Jan 2002
Dean Smith	England	Aston Villa	10 Oct 2018	Present*
Alan Pardew	England	Crystal Palace	2 Jan 2015	22 Dec 2016
Alan Pardew	England	Newcastle	9 Dec 2010	2 Jan 2015

Note: *Managers in EPL. One observation is a job spell by a manager at a team. N=395*
Source: `football` data.

As we can see, some managers had a long spell, others a shorter spell at teams. Some managers coached more than one team: Alan Pardew, for instance, worked for both Crystal Palace and Newcastle. We have 395 observations for 241 managers.

So we have a relational dataset. It has one data table with team-game observations, and one data table with manager-team observations. The first data table contains dates for the games; the second data table contains the start dates and end dates for the time each manager worked for each team. To work with the information in the two data tables together, we need to create a workfile, which is a single data table that is at the team-game level with the additional variable of who the manager was at the time of that game.

We have all the information to link managers to team-games: which manager was in charge at the time a team played a game. But creating that linkage is not straightforward. We discuss some of the problems and its solutions in the next section, and we'll return to our case study afterwards.

10 Entity resolution: Duplicates, ambiguous identification, and non-entity rows

In many data tables that we start working with, we may observe strange things: the ID variable is not unique when it should be, we appear to have multiple observations with different ID variables for entities that should be the same, or we may have rows that are not for the kinds of entities we want. These are issues with the entities in the data table. Before doing any meaningful analysis, such issues need to be resolved, to the extent it is possible. That process is called **entity resolution**.

One potential issue is having **duplicate observations**, or, simply, duplicates, in the data table. Duplicates appear when some entities that should be a single observation appear more than once in the data table. While the name suggests two appearances, it may refer to three or more appearances as well. Duplicates may be the result of human error (when data is entered by hand), or the features of the data source (e.g., data scraped from classified ads with some items posted more than once). When possible, duplicates need to be reduced to a single observation.

In the simplest case, duplicates are perfect: the value of all variables is the same. In such cases one has to delete the duplicates and leave one data row for each observation only. In more difficult cases, the value of one or more variables is different across data rows that appear to correspond to the same observation. Then, one has to make a decision about whether to keep all observations or select one, and if the latter, which value to keep – or maybe create a new value, such as the average.

A related, but conceptually different, issue is to have **ambiguous identification**: the same entity having different IDs across different data tables. The task here is to make sure that each entity has the same ID across data tables. That is necessary to link them properly. Entities are frequently identified by names. Unfortunately, though, names may cause issues for two main reasons: they are not unique (more than one person may be called John Smith), and different data tables may have different versions of the same name (e.g., middle names sometimes used, sometimes not: Ronald Fisher, Ronald A. Fisher, and Sir Ronald Aylmer Fisher is the same person, a famous statistician). This task is called **disambiguation**: making identification of entities not ambiguous.

Yet another issue is having non-entity observations: rows that do not belong to an entity we want in the data table. Examples include a summary row in a table that adds up, or averages, variables across all, or some, entities. For example, a data table downloaded from the World Bank on countries often includes observations on larger regions, such as Sub-Saharan Africa, or the entire World, and they are included just like other rows in the data table together with the rows for individual countries. Before any meaningful analysis can be done, we need to erase such rows from the data table.

The important message is this: assigning unique IDs is important. As a ground rule, we should avoid names as ID, people or firm names are not unique (remember, even the two authors have the same first name) and can be frequently misspelled. Using numerical ID variables is the good practice.

Finally, note that, very often, there is no single best solution to entity resolution. We may not be 100% certain if different rows in the data table belong to the same real-life observation or not. Moreover, often it is not evident what to do when different values of a variable show up for the same real-life observations. As a result, our clean data table may end up not really being 100% clean. But that is all right for the purpose of data analysis. Data analysts have to learn to live with such imperfections. When the magnitude of the problem is small, it is unlikely to have a substantial effect on the results of our subsequent analysis. When the magnitude of the problem is large, we may have to try different versions of solving it and see if they have an effect on the results of our analysis.

11 C2 Case Study – Identifying successful football managers

Entity resolution

Let us look at data issues for the football manager case study and see if we need to resolve entities. Indeed, the data does have ambiguous entities that we need to deal with. For instance, consider Manchester City and Manchester United, the two major football teams from the city of Manchester. When we looked through official sites, news reports and datasets, we could find many different versions of how teams were named, as summarized in Table 2.9. Entity resolution here is defining unique IDs and deciding which names do actually belong to the same team.

Table 2.9: Different names of football teams in Manchester, UK

Team ID	Unified name	Original name
19	Man City	Manchester City
19	Man City	Man City
19	Man City	Man. City
19	Man City	Manchester City F.C.
20	Man United	Manchester United
20	Man United	Manchester United F.C.
20	Man United	Manchester United Football Club
20	Man United	Man United

Source: Various sources including source for the football data

For the manager names, we sometimes see one or more space characters in the name, and sometimes we don't see them. Another issue is whether and how accents are included in names, such as "Arsène Wenger", the manager of Arsenal. On a very few occasions, the manager's name may be missing. One example is that there are no records for the team Reading for mid-March 2013. In our case this comes from a caretaker manager who somehow was not recorded. In such cases we can create a separate ID, with the name missing.

Having made these corrections and combined the datasets, we can create unambiguous ID variables for teams in the first data table and managers and teams in the second data table. With these ID variables and the dates of the games and the date for the managers' job spells, we can create a workfile by joining the managers data table to the team-games data table of the appropriate date. This procedure is not straightforward, and it can be done in multiple ways. You are invited to check our code on the course website to see one way to do it.

With the workfile at hand, we can describe it. The workfile has 8360 team-game observations: in each of the 11 seasons, 20 teams playing 38 games (19 opponent teams twice; $11 \times 20 \times 38 = 8360$). For this 11 seasons, we have 137 managers in the data.

From this case study, we learned that entity resolution, having the data structured right, and having proper ID variables are essential when working with relational data. Only with clearly identified teams and managers can we hope to match managers to teams and the games they play. The actual joining of the two data tables is a little involved, and you are invited to look at our solution to it in the appropriate code on the course website.

12 Discovering missing values

With tidy data and no issues with entities to resolve, we can turn our attention to the actual content of variables. And there may be issues there, too. A frequent and important issue with variables is **missing values**. Missing values mean that the value of a variable is not available for some observations. They present an important problem, for three main reasons.

First, missing values are not always straightforward to identify, and they may be mistakenly interpreted as some valid value. That is not a problem when missing values are represented by a specific character in the data, such as "NA" (for "not available"), a dot ".", an empty space "". Statistical software recognize missing values if stored appropriately. Sometimes, however, missing values are recorded with number values, outside the range (e.g., 0 for no, 1 for yes, 9 for missing). Such values need to be

replaced with a value that the software recognizes as missing. Identifying missing values and storing them in formats recognizable by the statistical software is always a must. It should be the first step for dealing with missing values, and it needs to be done even if it affects one single observation.

The second issue with missing values is that they mean fewer observations in the data with valid information. As we shall see in subsequent chapters, the number of observations is an important determinant of how confidently we can generalize our results from the particular dataset to the situation we truly care about. When a large fraction of the observations is missing for a key variable in the analysis, we have a lot fewer observations to work with than the size of the original dataset.

The magnitude of the problem matters in two ways: what fraction of the observations is affected, and how many variables are affected. The problem is small if values are missing for one or two variables and only a few percent of the observations for each variable. The problem is bigger the larger fraction missing and/or the more variables affected. For an example with small missing rates of many variables, consider a dataset on 1000 people with 50 variables that aim to capture personality characteristics. For each of these variables, 2% of the observations have missing values – i.e., we do not have information on a given characteristic for 20 of the 1000 people. Suppose, moreover, that for each variable on personal characteristics, the occurrence of missing values is independent across variables. In this case we end up with as few as 360 people with valid values for all 50 variables, which is 36% of the original number of observations (0.98 raised on the power of 50=0.36). This is a large drop even though each missing rate is tiny in itself.

The third issue is potential **selection bias**. One way to think about missing values is that they lead to a dataset used for the analysis that covers fewer observations than the entire dataset. We discussed coverage earlier in Chapter 1, Section 4. As always, when coverage is incomplete, an important question is whether this smaller sample represents the larger dataset. In the case of missing data, the terminology used is whether data is **missing at random**.

How can we tell whether values are missing at random or there is selection bias? The two approaches that work for assessing whether a sample represents a population work here as well: benchmarking and understanding the selection process.

Benchmarking means comparing the distribution of variables that are available for all observations. Think of missing values of a variable y. Then benchmarking involves comparing some statistics, such as the mean or median (see more in the next chapter 9) of variables x, z, etc., each of which is thought to be related to variable y, in two groups: observations with missing y and observations with non-missing y. If these key characteristics are different, we know there is a problem.

Understanding the selection process requires knowing how the data was born and a good understanding of the content of the variables. In some other cases, missing is really just that: no information. Then we should understand why it is so – e.g., why some respondents refused to answer the survey question, or why some companies failed to report a value. However, in some cases missing doesn't really mean missing but zero, only the value zero was not filled in for the relevant observations. For example, a variable for export revenues in company-level data may be left missing for companies that do not have such revenue. But it means zero. When missing values can be replaced with meaningful ones, we should do so. In other cases, missing values decrease the quality of data, just like incomplete coverage.

13 Managing missing values

Having missing values for some variables is a frequent problem. Whatever the magnitude, we need to do something about them. But what can one do about them?

There are two main options. First, we can work with observations that have non-missing values for all variables used in the analysis. This is the most natural and most common approach. It is usually a reasonable choice if the fraction of missing values is not too high, there are few variables affected, and selection bias is not too severe. One version of this approach is to work with observations that make up a well-defined subsample, in which the missing data problem is a lot less severe. For example, if missing values are a lot more prevalent among small firms in administrative data on firms, we may exclude all small firms from the data. The advantage of such a choice is transparency: the results of the analysis will refer to medium and large firms.

The second option is filling in some value for the missing values, such as the average value, or a value randomly selected value from the other observations. This is called **imputation**. Imputation may make sense in some cases but not in others. In any case, imputation does not add information to the data. For that reason, it is usually not advised to do imputation for the most important variables in the analysis. When data analysts use many variables, imputation may make sense for some of them. There are sophisticated methods of imputation, but those are beyond the scope of this textbook. Partly for that reason, we advise against imputing missing values in general.

Let us offer three practical pieces of advice regarding managing missing observations.

First, when possible, focus on more fully filled variables. Sometimes the variable that best captures the information we need has many missing values, but a less perfect variable is available with few missing variables. For instance, working with data on customers, we may have a variable describing family income for each customer, but that information may be missing for most customers. Instead, we may have information on the zip code of their residence available for virtually all customers. In this case it makes sense not to use the family income variable at all, and instead use a variable on the average family income in zip code locations available from another source.

Second, for a qualitative variable, we may want to add missing as a new value. For instance, we may consider missing value as another category for hotel stars in addition to the categories allowed, resulting in one more category. Note that this addition may change the type of the variable; in the case of stars, it changes it from ordinal to nominal. That may be a price worth paying as, in any case, we may transform this single variable to many binary variables (yes-no to each category) for our analysis (more details in Chapter 10).

Third, and most importantly, whatever you choose to do with missing values, make a conscious choice and document that choice. Some choices are more reasonable than others, depending on the situation. But all choices have consequences for the analysis.

As with other steps of data cleaning, there may not be an obviously best solution to deal with missing values. Such imperfections are a fact of life for data analysts. Magnitudes matter: small issues are likely to have small effects on the results of our analysis; issues that affect many observations may have substantial effects. In the latter case, data analysts often try alternative decisions during the cleaning process and see whether and how they affect their results in the end.

Review Box 2.3 ***Missing values***

- Missing values of variables may be an issue for several reasons:
  - they may be interpreted as values, often as extreme values;
  - they yield fewer observations to analyze;
  - they may the in selection bias during analysis if the observations with missing values are very different.
- There are several options to deal with missing values:
  - drop variables with too many missing values;
  - drop observations with missing values if there aren't many;
  - replace missing values with some other value (imputation) and add a binary variable to indicate missing values (a flag).

14 A2 Case Study – Finding a good deal among hotels: data preparation

Duplicates and missing values

We illustrate the issues with the raw `hotels-vienna` dataset with observations on hotels in Vienna for a weekday in November 2017.

Recall that we replaced hotel names with a numerical ID variable for confidentiality reasons. The way we did that ensured that each hotel name corresponds to a single number. Thus, duplicates with hotel names would show up as duplicates in the ID variable, too. And there are duplicates. There are 430 observations in the raw data, yet there are only 428 different ID values. It turns out that the reason for this difference is that there are two hotels that are featured twice in the raw data table. They are listed in Table 2.10 below, together with the most important variables.

Table 2.10: Duplicates in the `hotels-vienna` data

ID	Accommodation type	Price	Distance to center	Stars	Avg. rating	Number of ratings
22050	hotel	242	0.0	4	4.8	404
22050	hotel	242	0.0	4	4.8	404
22185	hotel	84	0.8	3	2.2	5
22185	hotel	84	0.8	3	2.2	5

The table shows that these duplicates are of the simple kind: all variables have the same value. To resolve the issue we need to drop one of each of the duplicates. The result is 428 hotels in the data table (this is the number of observations we described in Chapter 1 Section 3).

Next let's turn to missing values. The most important variables have no missing values except for

average customer rating, which is missing for 35 of the 428 observations. This is an 8% missing rate. When we dig deeper into the data, we can see that the type of accommodation is strongly related to whether average customer rating is missing. In particular, 34 of the 35 observations with missing values are for apartments, guest houses, or vacation homes. Of the 264 regular hotels in the data, only 1 has a missing value for average customer rating. It's a hotel with 2.5 stars, 0.7 miles from the city center, charging 106 dollars. Later, when we analyze the data to find a good deal, we'll restrict the data to regular hotels with 3 to 4 stars; this hotel would not be in that data (see Chapter 7, Section 4).

15 The process of cleaning data

The data cleaning process starts with the raw data and results in clean and tidy data. It involves making sure the observations are structured in an appropriate way and variables are in appropriate formats.

Data cleaning may involve many steps. It is good practice to document every step of data cleaning.

An important step of data cleaning is to make sure all important variables are stored in an appropriate format. Binary variables are best stored as 0 or 1. Qualitative variables with several values may be stored as text or number. It is usually good practice to store them as numbers, in which case there should be a correspondence to say what number means what: this is called value labeling. Cleaning variables may include slicing up text, extracting numerical information, or transforming text into numbers.

Another important step is identifying missing values and making appropriate decisions on them. As we have described above, we have several options. We may opt to leave them as missing; then missing values need to be stored in a way that the statistical software recognizes.

Data cleaning also involves making sure that values of each variable are within their admissible range. Values outside the range are best replaced as missing unless it is obvious what the value was meant to be (e.g., an obvious misplacement of digits). Sometimes we need to change the units of measurement, such as prices in another currency or replacing very large numbers with measures in thousands or millions. Units of measurement should always be indicated with the variables. Sometimes it makes sense to keep both the original and the new, generated variable to be able to cross-check them later in the analysis.

As part of data cleaning, variable description (also called variable labels) should be prepared showing the content of variables with all the important details. Similarly, when qualitative variables are stored as numbers, value labels that show the correspondence between values and the content of the categories need to be stored as well. In a spreadsheet, those labels should be stored in a separate sheet or document. Some software can include such labels together with the data table.

In any case, it is important to be economical with one's time. Often, there are variables in the data that we think we won't use for our analysis. Most of the time it does not make sense to invest time and energy to clean such variables. It is good practice to make the tidy data file leaner, containing the variables we need and work with that data file. As the raw data is stored anyway, we can always go back and add and clean additional variables if necessary.

Data wrangling is part of a larger process. Data analysts start with structuring and cleaning the data, then turn to data description and analysis. They almost always find errors and problems in those latter stages, and they go back to structuring and cleaning. For instance, in some cases, missing values

tell us that there is a problem in the data management and cleaning process, and in some cases we can improve the process and reduce the number of missing values. Maybe we actually made some coding mistakes causing errors, such as mishandled joining of two data tables.

Data analysts make many choices during the process of data wrangling. Those choices need to be the results of conscious decisions and have to be properly documented. Often, there are multiple ways to address the same issue, and there may not be a single best solution available. We have seen examples of this for handling apparent duplicates in the data with different values of the same variable, or handling missing values. When the issue in question affects few observations, that ambiguity is unlikely to be a problem. When it affects many observations, it may have effects on the results of our analysis. In such a case it is good practice to try different solutions to the problem and see whether and how that affects the results of the analysis. This process is called **robustness checking**.

16 Reproducible workflow: Write code and document your steps

Data wrangling should be executed in a way that is easy to repeat and reproduce. This means documenting what we do and writing code so that steps can be repeated.

Documentation is very important. It is good practice to produce a short document – called README.txt that describes the most important information about the data used in any analysis. This document may be helpful during the analysis to recall important features of the data.

Another, often longer, document can describe all data management and cleaning steps we have done. It is also essential to enable other analysts to check your work, replicate your analysis, or build on your analysis in the future (including the future version of yourself). Such a document should allow recalling the steps as well as communicating them. It is important to cite the data source, too.

Table 2.11 offers a checklist to guide describing datasets.

Table 2.11: Describing data

Topic	**Content to include**
Birth of data	When, how, for what purpose it was collected
	By whom, and how is the data available
	Whether it's the result of an experiment
Observations	Type: cross-sectional, time series, etc.
	What an observation is, how it is identified, number of observations
Variables	List of variables to be used, their type, their content, range of values
	Number or percentage of missing values
	If a generated variable, how it was created
Data cleaning	Steps of the process

It is also useful to write code for for all data wrangling steps. Yes, writing code takes more time than executing commands one by one or clicking through such commands. However, investing some time can pay out fast. In most cases, we have to redo data cleaning after new issues emerge or the raw data changes. There the benefits of code tend to massively outweigh its costs.

1. It makes it easy to modify part of the procedure and re-do the entire procedure from the beginning to the end.

2. An automated process can be repeated when needed – maybe a slight change in the underlying raw data.

3. It makes it easy for anyone else to reproduce the procedure thus increasing the credibility of the subsequent analysis.

4. Code becomes the skeleton for documentation: it shows the steps in the procedure itself.

Of course, there are trade-offs between all those benefits and the work needed to write code, especially for short tasks, small datasets, and for novice data analysts. Yes, it's OK to use a spreadsheet to make some quick changes to save time. However, we think that it makes sense to write code all the time, or very frequently, even if it seems too big an effort at first sight. We typically don't know in advance what data cleaning decisions we'll have to make and what their consequences will be. Thus, it is quite possible that we would have to re-do the whole process when things change, or for a robustness check, even for a straightforward-looking data cleaning process. Moreover, writing code all the time helps in mastering the coding required for data cleaning and thus helps in future projects.

As we have emphasized many times, data cleaning is an iterative process. One starts with cleaning and creating a tidy data version and then a workfile. Then, while describing the data and working on the analysis, it is very common to discover further issues that require going back to the data cleaning step. This is yet one more reason to work with code: repeating everything and adding new elements to data cleaning is a lot easier and a lot less time consuming if written in code.

Review Box 2.4 ***Data wrangling: common steps***

0. Write code – it can be repeated and improved later
1. Understand types of observations and what actual entities make an observation.
2. Understand types of variables and select the ones you'll need.
3. Store data in tidy data tables.
4. Resolve entities: find and deal with duplicates, ambiguous entities, non-entity rows.
5. Get each variable in an appropriate format; give variable and value labels when necessary.
6. Make sure values are in meaningful ranges; correct non-admissible values or set them as missing.
7. Identify missing values and store them in an appropriate format; make edits if needed.
8. Make a habit of looking into the actual the data tables to spot issues you didn't think of.
9. Have a description of variables.
10. Document every step of data cleaning.

17 Organizing data tables for a project

After having discussed how to detect issues with observations and variables in a data table and what to do with such issues, let's turn to how the various data tables should be organized. It is good practice to organize and store the data at three levels. These are

- Raw data tables
- Clean and tidy data tables
- Workfile(s) for analysis

Raw data, the data as it was obtained in the first place, should always be stored. Raw data may come in a single file or in multiple files. It should be stored in the original format before anything is done to it. This way we can always go back and modify steps of data cleaning if necessary. And, most often it is necessary: something may go wrong, or we may uncover new issues during the process of data cleaning or data analysis. Having the raw data is also key for replicating one's analysis. That becomes especially important if similar analyses are to be done in the future.

The next step is producing clean and tidy data from raw data. That's the process of data cleaning. It involves making sure each observation is in one row, each variable is in one column, and variables are ready for analysis. This process takes much of that 80% of total time that we wrote about earlier. In the next chapter we'll discuss a few of the most important steps in data cleaning.

Often tidy data means multiple data tables. It often makes sense to create different data tables for data coming from different sources. That makes the process of creating tidy data from raw data transparent. One should always create different data tables for data with different kinds of observations.

In cross-sectional data this means potentially different kinds of entities, such as individuals, families, neighborhoods, schools of children, employers of workers, etc. In time series data this means potentially different time periods, such as daily observation of prices but weekly observations of quantities. For multi-dimensional data this may mean all of the above, plus different data tables for information at different levels of observations.

In cross-section time series data one may have data tables with one row for each cross-sectional unit i without any change in time, data tables with aggregate time series with observations t (that are the same for all cross-sectional units), and some data tables with observations i, t – in the long format of course.

The last of the three levels of the file structure is the workfile. This is the one file on which the analysis is to be done. The rows of workfiles are observations that form the basis of the analysis. Typically, they contain only a subset of all available variables, and, with more than one tidy data file, they may use data from all or only a subset of those files. Work files may or may not be in tidy data format.

Let us emphasize the advantage of having tidy data before turning them into a workfile for the analysis. Tidy data files tend to be more transparent. Thus they are better suited to identifying problems in the data, addressing those problems and producing clean data, and adding or taking away observations and variables. With tidy data we can add new variables to be included in the analysis, and we can produce various kinds of workfiles for various kinds of analyses.

Consider the organization of files for this textbook. For all case studies, we store raw files as they were collected. The clean folder contains cleaned tidy data tables as well as the code that produces these

clean files from the raw ones. The work folder for each chapter includes the analytic files and work data that we use for the analysis. We also added an output folder for storing graphs and tables.

18 B3 Case Study – Identifying successful football managers

Organizing files

As always, we produced three kinds of files during the process of cleaning the football managers data: data files, code, and documentation files. We have three kinds of data files: raw data files (the original ones without any modification), tidy data files, and a workfile for our analysis.

We have two raw data files here, one on every game from 11 seasons of the English Premier League, and one on football managers that worked for any of the teams that appeared in the English Premier league during those 11 years. In fact, the data on games from those 11 years are 11 separate data files that we downloaded. We then merged (appended) these 11 files to form a single data file. We collected data on the managers of these teams during these 11 years from Wikipedia.

Of these raw data files, we created three tidy data files, one with games as an observation (epl_games), one with one team playing one game as an observation (epl_teams_games), and one with one spell for a manager working for a team as an observation (epl_managers). Then, for our analysis, we combined the second and third tidy data table to form a workfile, with one team playing one game as an observation, together with the information of who the manager was then. That's our workfile for this chapter (ch02_football_manager_change).

Table 2.12 shows these data files together with the file name of the Stata and R code that produce them (.R and .do). Besides data and code, the file list contains a README file that contains the most important information on the data and our code.

Table 2.12: Structure of storing data and code files

path	folder	files
da_data_repo	football	README
	football/raw	fdata_pl_2008.csv - fdata_pl_2018.csv
		managers_epl.xls
	football/clean	epl_games.csv
		epl_teams_games.csv
		epl_managers.csv
		football_managers_workfile.dta
		football_managers_workfile.csv
		football-epl-maker.do
		football-epl-maker.R
		VARIABLES.xls
da_case_studies	ch02-football-manager-success	ch02-football-manager-success.do
		ch02-football-manager-success.R

19 B4 Case Study – Identifying successful football managers

Finding the most successful managers

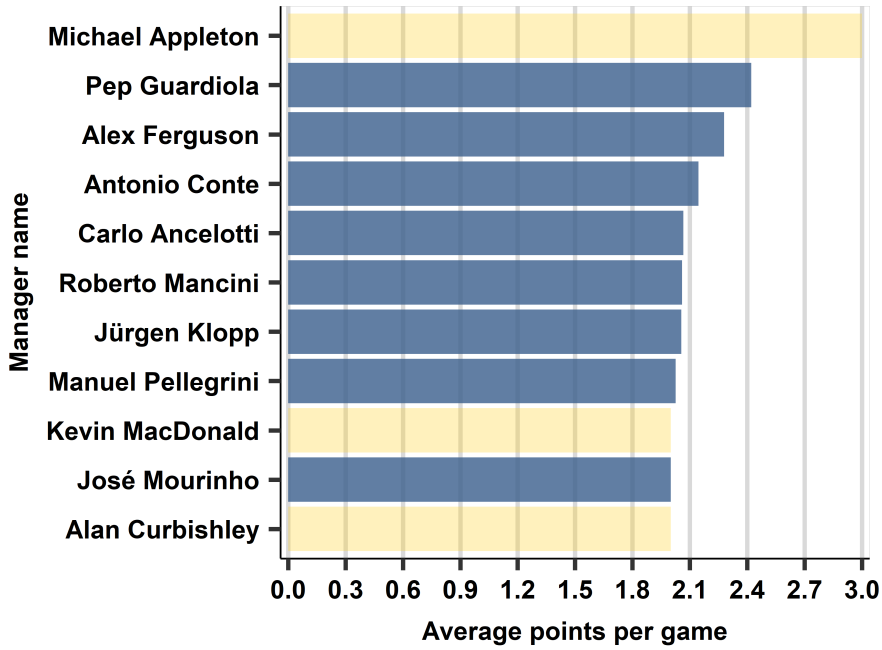
Having compiled the data we need, we can answer our question: which managers have been the most successful ones over the 11 seasons of EPL. We continue considering managerial spells at teams: if a manager worked for two teams, we consider it two cases. First, we need to define success. Let us consider average points per game as a measure of success.

To calculate it, we simply add up both the points earned over a career at a team and the number of games he managed, and divide total points with number of games. We rank them, and take a look at those with at least 2 points per game – we have eleven such manager-team combinations.

Graph 2.1 shows the top managers. We can see some well-known names in European football, like Alex Ferguson (Manchester United) Pep Guardiola (Manchester City) or Carlo Ancelotti (Chelsea) among top managers.

One interesting aspect is the presence of a few lesser known names, starting with Michael Appleton topping the chart with a perfect 3/3 ratio. The reason behind that is some managers were only sitting in on a few games, and if lucky, they could get high win ratio. To make this clear, we used a different color for manager-team spells less than 18 games, such as caretakers.

Figure 2.1: The most successful managers



Note: *English Premier League, 11 seasons, manager-team spells. Yellow denotes spells less than 18 games (caretakers).*
Source: `football` data.

Once we discount caretakers, we have top winner managers. In terms of average points per game,

Pep Guardiola tops our league of managers, with 2.42 points per game for his entire spell of 3 years (114 games) at Manchester City followed by Alex Ferguson (2.28 points) having managed Manchester United for 190 games.

20 Summary and practice

20.1 Main takeaways

- You'll have to spend a lot of time on data wrangling, and you'll have to make many decisions.
  - Your data wrangling decisions may affect the conclusions of your analysis.
  - Write code to repeat, reproduce, or change your decisions.
  - Document everything you do.
  - Tidy structure helps all subsequent work with data.

20.2 Practice questions

1. What are in the rows and columns of a typical data table?
2. What's the difference between cross-sectional and time series data?
3. What's a binary variable? Give two examples.
4. What are nominal and ordinal qualitative variables? Give two examples for each.
5. What are interval and ratio quantitative variables? Give two examples for each.
6. What are stock and flow variables? Give two examples for each.
7. What's the difference between long and wide format xt panel data? Which one would you prefer and why?
8. In a dataset on individuals we have information on their income, how many hours they work in a typical week, and whether they live in a metropolitan area, a small town or a rural area. What kinds of variables are these?
9. In a dataset on countries we have information on the name of the country, the continent, their population, and their GDP. What kinds of variables are these?
10. Decide what types of variables the following are. Are they qualitative or quantitative? Are they binary variables? Also think about whether they are measured on a nominal, ordinal, interval, or ratio scale.
11. List four topics that data cleaning documentation should address.
12. What are the benefits to writing code for cleaning data?
13. What are the benefits to documenting data cleaning?

    (a) IQ

    (b) Country of origin

(c) Number of years spent in higher education

(d) The answer to the question in a survey that says: “Indicate on the scale below how much you agree with the following statement: Everyone has to learn data analysis for at least two semesters." with options “5 – Fully agree" “4 – Somewhat agree" “3 – Indifferent" “2 – Somewhat disagree" “1 – Fully disagree"

(e) A variable that is 1 if an individual bought a car in a given month

(f) Eye color

14. Consider the following data tables. Data table 1 includes countries in its rows; its columns are the name of the country, its area, and whether it has access to sea. Data table 2 includes countries in its rows; its columns are GDP and population for various years, each year in different column. Is this a tidy dataset? If yes, why? If not, why not, and how can you transform it into a tidy dataset?

15. Consider the following data tables. Data table 1 includes single-plant manufacturing companies in its rows; its columns are the name of the company, the zip code of its plant, its industry, and the average wage the company pays in 2019. Data table 2 includes zip codes in its rows; its columns are total population, population density, and average house prices in 2019 in the zip code area. You want to analyze how the average wage paid by companies is related to the average house prices. How would you create a workfile from these data tables for that analysis?

20.3 Data exercises

Easier and/or shorter exercises are denoted with [*] Harder and/or longer exercises are denotes with [**]

1. Use hotel price data for another city (not Vienna). Load and clean the data, document the cleaning, and describe the clean dataset. In particular, look at the raw data and make it into tidy data. [*]

2. Use a classified ads site, such as Craigslist, and get data on the price and some of the most important characteristics of used cars. Pick a city and a make and type of car (e.g., Ford Fusion). Load and clean the data, document the cleaning, and describe the clean dataset. Discuss any decisions you may have made [**]

3. Consider a department that uses this textbook for a sequence of 4 courses: DA1, DA2, DA3, and DA4. Below you will find textual descriptions of the organization of teaching. Create a tidy table with this info in any spreadsheet (Excel, Google Sheets, etc). [*]

   (courses and programs) The department has five programs running some courses: EC1, EC2, EC3, BA, FIN. A course may be core or elective. Courses DA1 and DA2 are core for all programs. DA3 is core for BA and elective for Fin. DA4 is core for EC1 and elective for all other programs.

   (lectures and seminars) Each course is composed of a lecture and a seminar. For Fin, there is one lecture for DA1 and DA2, and another lecture for all other programs. For all relevant programs there is one shared lecture in DA3 and DA4. In courses DA1 and DA2, there is one seminar for BA, one for EC1, EC2 and EC3 together, and one for FIN. In DA3 there is a single seminar for relevant programs. In DA4, there is a joint seminar for BA and FIN, and another one for EC1, EC2, EC3.

   (professors) For FIN, lectures and seminars in DA1 and DA2 are taught by Prof X. All other lectures are taught by Prof Y. Seminars with students from BA are taught by Prof W. Remaining seminars are taught by Prof Z.

4. Consider the `world-bank-vaccination` dataset based on World Bank data. Generate a new variable for the growth rate of GDP per capita, in percentage terms, $(gdppc_{it}/gdppc_{i,t-1}-1)*100$. Add this variable to the long and wide format of the data and to Tables 2.4 and 2.5. [*]

5. Consider the `football` dataset we used. Create a table counting the number of seasons each teams spent in the Premier League. List the teams that featured in all 11 seasons. List the teams that were in the League only once. [*]

21 Under the hood: Naming files

Many seemingly mundane issues can make a big difference when working with data and code. Naming files is one of them.

Good names are easy to understand and informative for other people who may want to work with our files. That includes our future selves: one shouldn't count on remembering all the details that are not documented.

There are three useful criteria for naming files: names should be machine readable, human readable, and work well with default ordering.

1. Machine readable Machine readable means it will be easy to extract information by splitting bits of text – for searching files, or narrowing the list of files by terms such as "graph", "table", "chapter05", "cleaner".

It also means to keep it plain and simple so that variety of software could read it. In particular, it means having **no**

- spaces
- punctuation(.,;)
- accented characters
- capitalized letters

2. Human readable The name should also contain information on content – so that it is easy to figure out what a particular program or document is about. We may add a word or two on the purpose or the topic of the file (such as data-cleaner, financials-merger) as well as adding the type of the file (draft, presentation, conf-talk).

We suggest a deliberate use of "_" and "-", which allows us to recover meta-data from the file names. In particular,

- "_" underscore used to delimit units to use for search, differentiate between parts of information. These parts will be easy to search on later. E.g., "firms_financials_2012"
- "-" hyphen used to delimit words for easy readability such as "compare-means-age"

3. Works well with default ordering as the computer default sorting already makes some sense.

For cases where we have several similar files, put something numeric first (or early on). This will help natural ordering (01, 02 or ch01, ch02). The date can be at the beginning, if it is crucial, or it can go at the end.

A useful little trick is to **left pad** other numbers with zeros, i.e. have 01 and not 1. Otherwise 10 will be before 2.

Finally, use the ISO 8601 standard for dates: "YYYY-MM-DD" – e.g., "2018-06-01". While this order may look strange to people in the U.S.A., this is indeed the global standard.

These are not essential to data science, but they are easy to implement, and the payoffs will accumulate as your skills evolve and projects get more complex.

Below are some examples of file names that we like:

- "bekes-kezdi_textbook-presentation_2018-06-01.pdf" - it uses "-" and "_" deliberately, has a date in ISO format, all is in lower case, no special characters, and accents on names are gone.
- "ch02_organizing-data_world-bank-download_2017-06-01.tex" - it starts with a number early on, can be ordered, has left padding (02 and not 2), combines title (organizing data) and type (draft-text), all lower case

Examples of file names we do not like include: thesis.pdf, mytext.doc, calculations1112018.xls, Gábor's-01.nov.19_work.pdf.

22 References and further reading

On tidy data, the original paper is Wickham (2014).

A similar set of suggestions to ours for spreadsheet users is Broman & Woo (2018).

The discussion on naming convention and quite a few points come from Brian (2015).

Chapter 3

Exploratory data analysis

How to describe the information contained in variables

Motivation

You want to identify hotels in a city that are underpriced for their location and quality. You have scraped the web for data on all hotels in the city, including prices for a particular date, and many features of the hotels. How can you check whether the data you have is clean enough for further analysis? Can you assess the quality of the data in a simple but powerful way? And how should you start the analysis itself?

You want to learn the extent of home advantage in football (soccer): how much more likely it is that a team wins if it plays in its home stadium, and how many more goals it scores. You have data from all games from a league, including who won, the score difference, and which team played in their home stadium. How should you summarize the data in a graph or a number that best describes the extent of home advantage? What additional graphs or numbers could help you understand its potential causes?

After collecting the data, assessing its quality, cleaning it, and structuring it, the next step is exploratory data analysis (EDA). Exploratory data analysis aims to describe variables in a dataset. EDA is important for understanding potential problems with the data and making analysts and their audiences familiar with the most important variables. The results of EDA help additional data cleaning, decisions for further steps of the analysis, and giving context to the results of subsequent analysis.

The chapter starts with why we use exploratory data analysis. It then discusses some basic concepts such as frequencies, probabilities, distributions, and extreme values. It includes guidelines for producing informative graphs and tables for presentation and describes the most important summary statistics. The chapter, and its appendix, also cover some of the most important theoretical distributions and their uses.

There are four case studies in this chapter. The first one, **Finding a good deal among hotels: data exploration** continues using the `hotels-vienna` data to illustrate the description of distributions

and their use in identifying problems in the data. The second case study, **Comparing hotel prices in Europe: Vienna vs. London** uses the `hotels-europe` data, which contains hotels from several European cities, to illustrate the comparison of distributions. The third case study, **Measuring home team advantage in football**, examines whether and to what extent football (soccer) teams tend to perform better if they play in their home stadium, using the `football-games` data. It illustrates the use of exploratory data analysis to answer a substantive question. The last one, **Distributions of body height and income**, uses data from a large survey to illustrate theoretical distributions, using the `height-income-distributions` data.

Learning outcomes. After working through this chapter, you should be able to

- understand the importance and uses of exploratory data analysis;
- understand distributions, visualize them, and describe their most important features;
- identify situations when extreme values matter and make conscious decisions about extreme values;
- produce graphs and tables of presentation that are focused, informative, and easy to read;
- know the main features of the most important theoretical distributions and assess whether they are good approximations of distributions of variables in actual data.

1 Why do exploratory data analysis?

Informative description of variables is an essential first step of data analysis. It is called **exploratory data analysis**, abbreviated as **EDA**, also called **descriptive analysis**. There are five important reasons to do EDA.

First, to know if our data is clean and ready for analysis. EDA is part of the iterative process of data cleaning that we discussed previously, in Chapter 2. Informative description of variables tells us about problems that we need to address in the data cleaning process. Results of EDA may show too few observations because we discarded too many by mistake; they may suggest that the true units of measurement are different from what we thought; or they may tell us that there are values that a variable should not have because of mistakes or because such values are used to signify missing data. These are mundane but surprisingly common issues that could result in the nonsensical result of an analysis if not addressed. Therefore, the results of EDA may make data analysts go back to the cleaning process and start over.

Second, to guide subsequent analysis. Some features of variables, such as how spread the values are or whether there are extreme values, have important consequences for further analysis. The may suggest that data analysts transform a variable, or what method may be best suited to analyze patterns related the a variable. This is a little cryptic for now, but we'll see many instances of this in the subsequent chapters. The results of EDA may also help in getting a sense of what we can expect from future analysis. For example, we can't expect to understand what may cause differences in a variable if that variable has little spread in the data – i.e., if most observations have the same value.

Third, to give context. Describing the most important variables in our data is an important part of presenting the results of data analysis. Decision makers who will use those results need to know what is contained in key variables and how their characteristics affect the interpretation of the main results. For example, when our results show the effect of one variable on another variable, we want to know if

that effect is large or small. Exploratory data analysis helps in answering that question by uncovering typical values, or typical differences, in variables, which we can use as benchmarks to compare our results to.

Fourth, sometimes, we can answer our question with very simple tools that are used as parts of exploratory data analysis. That's quite rare, but when that's the case, EDA is the end of our analysis.

Fifth, to ask more questions. Quite often, exploratory data analysis uncovers an interesting phenomenon for a variable of interest. This should lead to asking questions that further analysis with more data and more sophisticated methods may answer.

In the remainder of this chapter we discuss what features of variables data analysts need to uncover, what methods they have to do that, and how they can produce graphs and tables that summarize the most important information in an accessible way.

Review Box 3.1 ***The use of exploratory data analysis (EDA)***

- Exploratory data analysis (EDA) describes the features of the most important variables in the data.
- We may use EDA for five purposes
  - To check data cleaning
  - To guide subsequent analysis
  - To give context of the results of subsequent analysis
  - To answer simple questions
  - To ask additional questions

2 Frequencies and probabilities

Let's start with the most basic property of variables: what values they can take and how often they take each of those values. We first introduce the concept of frequency that makes sense from the viewpoint of the data we have. In the next section we generalize frequencies to the concept of probability that makes sense in more abstract settings.

The **absolute frequency**, or count, of a value of a variable is simply the number of observations with that particular value in the data. The **relative frequency** is the frequency expressed in relative, or percentage, terms: the proportion of observations with that particular value among all observations in the data. If a variable has missing values, this proportion can be relative to all observations including the missing values or only to observations with non-missing values. Most often we express proportions excluding observations with missing values. When our goal is to check the data cleaning process, absolute frequencies including missing values is the better option.

Probability is a general concept that is closely related to relative frequency. Probability is a measure of the likelihood of an event. An event is something that may or may not happen.

In the context of data, an event is the occurrence of a particular value of a variable. For example, an event in a data table is that the manager of the firm is female. This event may occur various times in

the data. The probability of the event is its relative frequency: the proportion of firms with a female manager among all firms in the data. Considering a more abstract example, whether there is pasta for lunch at the canteen today is an event, and the probability of pasta for lunch today is a measure of its likelihood.

Probabilities are always between zero and one. We denote the probability of an event as $P(event)$, so that $0 \leq P(event) \leq 1$. Sometimes they are expressed in percentage terms so they are between 0% and 100%: $0\% \leq P(event) \leq 100\%$.

Considering a single event, it either happens or does not. These two are mutually exclusive: the probability of an event happening and it also not happening is zero. We denote an event not happening as $\sim event$ so $P(event \,\&\, \sim event) = 0$.

Probabilities are more general than relative frequencies as they can describe events without data. However, with some creative thinking, we can often think of potential data that contains the event so that its probability is a proportion. Data on what's for lunch at our canteen for many days is an example. If we had such data, the frequency of the event (pasta served for lunch) would give its probability.

But not always. Sometimes probabilities may be defined for events for which no data can be imagined. These include **subjective probabilities** that describe the degree of uncertainty an individual feels about events that may happen in the future but have no precedents. Examples include the probability that you, the reader, will like this textbook enough to keep it for future reference, or the probability that rising sea levels will flood the underground system of Singapore within 20 years.

Abstract probabilities are interesting and important from various points of view. But data analysts work with actual data. For most of them, probabilities and frequencies are closely related. They tend to think of probabilities as relative frequencies in actual data or data they can imagine.

Review Box 3.2 ***Frequencies and probabilities***

- Frequencies describe the occurrence of specific values (or groups of values) of a variable.
  - Absolute frequency is the number of observations (count).
  - Relative frequency is the percentage, or proportion.
- Probability is a measure of the likelihood of an event; it's value is always between 0 and 1 ($0 \leq P(event) \leq 1$.
- In data, probabilities are relative frequencies; the "events" are the specific values of the variable.
-

3 Visualizing distributions

Frequencies are summarized by distributions. The **distribution** of a variable gives the frequency of each value of the variable in the data, either in terms of absolute frequency (number of observations), or relative frequency (proportion or percent). The distribution of a variable completely describes the variable as it occurs in the data. It does so focusing on the variable itself, without considering the

values of other variables in the data.

It is good practice to examine distributions of all important variables as the first step of exploratory data analysis.

The simplest and most popular way to visualize a distribution is the **histogram**. The histogram is a bar chart that shows the frequency (absolute or relative) of each value. The bars can be presented horizontally; that's more common in business presentations and when there are few bars. The more traditional presentation is vertical bars.

For binary variables, the distribution is the frequency of the two possible values and thus the histogram consists of two bars. For variables that take on a few values, the histogram shows as many bars as the number of possible values.

For variables with many potential values, showing bars for each value is usually uninformative. Instead, it's better to group the many values in fewer groupings or bins. The histogram with binned variables shows bars with the number or percentage of observations within each bin. It is good practice to create bins of equal width so each bin covers the same range of values. As the case study will illustrate, the size of bins can have important consequences for how histograms look and how much they reveal about the properties of distributions.

For any histogram, we need to decide on the bin size, either by setting the number bins or the bin width. (Letting our statistical software set the bin size is also one such decision.) Very wide bins may lump together multiple modes. But very narrow bins may show a lot of ups and downs for random reasons and thus can blur important properties of a distribution. It is good practice to experiment with a few alternative bin sizes to make sure that important features of the distribution don't remain hidden.

Visual inspection of a histogram reveals many important properties of a distribution. It can inform us about the number and location of **modes**: these are the peaks in the distribution that stand out from their immediate neighborhood. Most distributions with many values have a center and tails, and the histogram shows the approximate regions for the center and the tails. Some distributions are more symmetric than others. Asymmetric distributions, also called **skewed distributions**, have a long left tail or a long right tail; histograms visualize those. Histograms also show if there are **extreme values** in a distribution: values that are very different from the rest. We'll discuss extreme values in detail in Section 5.

Density plots – also called **kernel density estimates** – are an alternative to histograms for variables with many values. Instead of bars, density plots show continuous curves. We may think of them as curves that wrap around the corresponding histograms. Similarly to histograms, there are details to set for density plots, and those details may make them look very different. The most important detail to specify is the bandwidth, the closest thing to bin size for histograms. But there are more things to set for density plots, which are beyond the scope of this textbook. For this reason we advise you not to draw density plots on their own, but only to complement histograms – unless you have more detailed knowledge of how to produce them.

Review Box 3.3 ***Distributions and their visualization***

- A distribution of a variable shows the frequency, absolute or relative, of each potential value of the variable in the data.
- The histogram is a bar graph showing the frequency of each value of a variable if the variable has few potential values.
- Density plots (also known as kernel density estimates) are continuous curves; they can be viewed as wrapped around corresponding histograms.

4 A1: Case Study – Finding a good deal among hotels: data exploration

Describing distributions

The broader question on hotels we will be discussing in chapters to come is how to find a good deal among hotels. Here, we explore the most important variables we'll use in that analysis, with a focus on location – asking where hotels are located and what is the share of hotels close to the city center. Describing the distribution of hotels in quality or distance can help business decisions on hotel development.

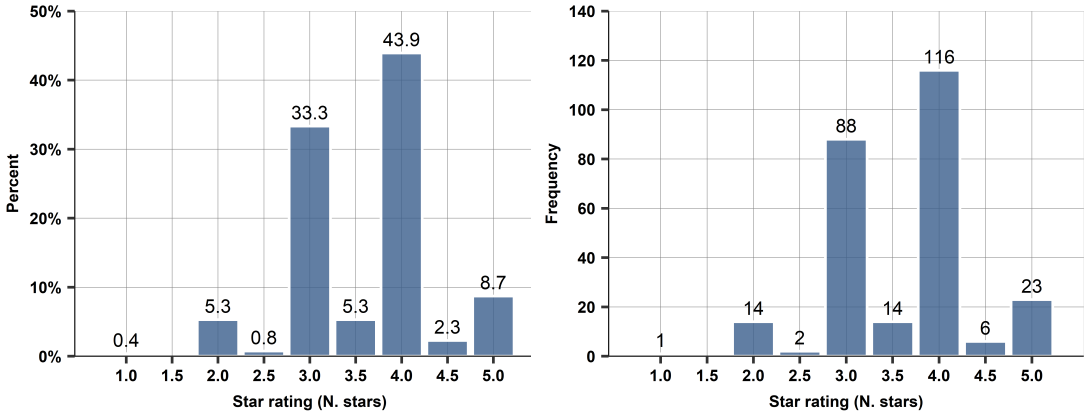
This case study uses the hotels data. We introduced this data in Chapter 1 Section 3. We will keep focusing on the data from Vienna for a particular night. The data we work with contains proper hotels only, without apartment houses etc. We have $N = 264$ hotels.

Let us start with stars, a measure of hotel quality. Stars are determined by the services hotels offer, according to detailed criteria. Hotel stars can take on a few distinct values; thus a histogram becomes a set of bars – one for each value.

Figures 3.1a and 3.1b show the histogram of stars in our data. The horizontal axis shows the potential values of this variable such as 3 stars, 4 stars, and also 3.5 stars (3 stars plus). In Figure 3.1b, the vertical axis shows absolute frequencies: the number of observations corresponding to each value. In Figure 3.1a, we show relative frequency (percentage).

According to the histogram, there are 88 hotels with 3 stars, 14 with 3.5 stars, and 116 hotels with 4 stars. In relative terms this means 37% (88/264=0.33) with 3 stars, 5% with 3.5 stars, and 44% with 4 stars. Indeed, most properties are either 3 or 4 star hotels.

Figure 3.1: Histogram of hotel stars



(a) Relative frequency (percent)

(b) Absolute frequency (count)

Note: *Histogram for a qualitative variable.*
Source: `hotels-vienna` data. Vienna, hotels only, for a 2017 November weekday. N=264.

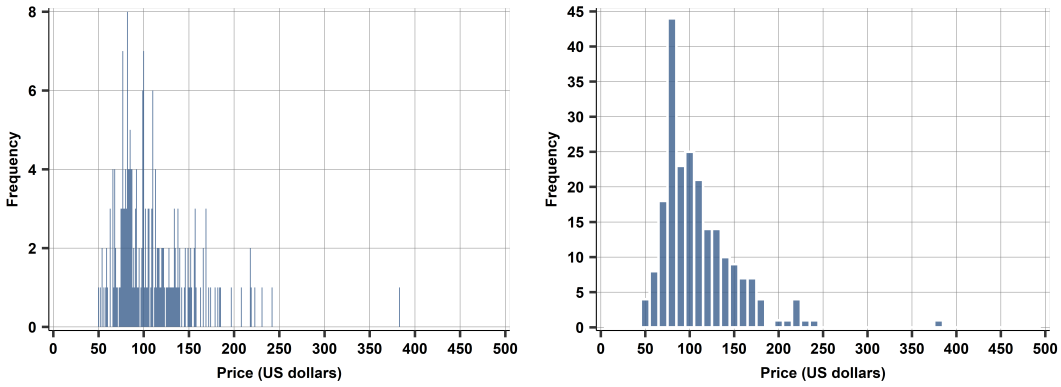
From now on we will focus on proper hotels and the mid-quality segment: our data will consist of hotels in Vienna with three to four stars. This is a subset of the `hotels-vienna` data described in Chapter 1, excluding non-hotel accommodation, hotels with fewer than three stars or more than four stars. We have $N = 218$ such hotels in Vienna.

The next variable we explore is hotel room price. The distribution of prices may be visualized in multiple ways. In this data, prices range from 50 dollars to 383 dollars plus a single hotel offering a room for 1012 dollars. Let us disregard that extreme value for the rest of the section, we'll come back soon. We will work with $N = 217$ observations in this section.

A histogram with a bar for each value is not a particularly informative visualization, but we show it for educational purposes in Figure 3.2a. Most values occur once, with some values (like 110 dollars) occurring several times. This histogram shows more bars, thus more hotels, with room price below 200 dollars, with most values between 80 and 150 dollars.

The right panel shows a histogram with 20-dollar bins. This graph is far more informative. Most of the distribution is between 60 and 180 dollars, with few observations above 180 dollars. It suggests a distribution with a mode at 100 dollars. The distribution is skewed, with a long right tail. A long right tail means that for many values on the right end of the histogram – in our case above 180 dollars – there are scattered values with only a few observations each. There is one extreme value around 400 dollars.

Figure 3.2: Histogram of hotel price



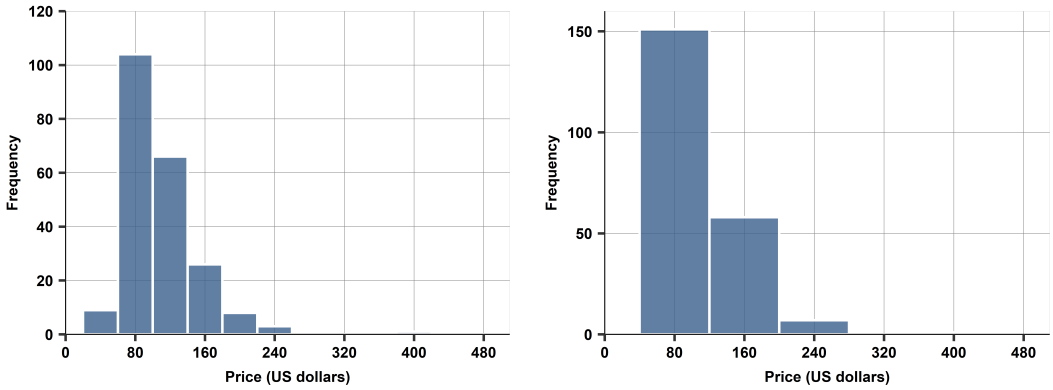
(a) Each value separately

(b) 20-USD-wide bins

Note: *Panel (a): bars at each value; panel (b): bars for 20-dollarbins; excluding extreme value.*
Source: `hotels-vienna` data. Vienna, 3–4 stars hotels only, for a 2017 November weekday. N = 217.

We can increase bin width further, lumping together more values in a bin and thus reducing the number of bars. The next figures show histograms of the same price distribution with wider bins, first of 40 dollars, second of 80 dollars. The wider the bin, the fewer details of the distribution are revealed by the histogram. All show the same long right tail, but the 80-dollar bin puts the mode in the leftmost bin, whereas it was in the second bin earlier. The histograms with the wider bins suggest a more gradual decline in frequency.

Figure 3.3: The distribution of hotel price: Histograms with wide bins



(a) 40-eurobins

(b) 80-eurobins

Note: *Panel (a): 40-dollar bins; panle (b): 80-dollar bins; excluding extreme value.*
Source: `hotels-vienna` data. Vienna, 3-4 stars hotels only, for a 2017 November weekday. N = 217.

These graphs (Figures 3.2a, 3.2b, 3.3a, and 3.3b) taken together suggest a trade-off. Narrower bins give you a more detailed view but make it harder to focus on the really important properties of the distribution. Finding a good compromise may depend on the purpose. To design further analytic work, a 20-dollar bin is very useful. For presenting to a more generalist audience, a 50-dollar bin may

be better. Designing bin size is a necessary task, and requires practice.

5 Extreme values

Some quantitative variables have **extreme values**: substantially larger or smaller values for one or a handful of observations than rest.

Sometimes extreme values don't belong in the distribution either because they are errors. Most frequently, such errors are due to mistakes in digits or units of measurement, such as company revenues in dollars instead of millions of dollars or number of visits to a supermarket per year instead of per week. In addition, extreme values may not belong in the distribution because they represent patterns that we aren't interested in.

But other times extreme values are an integral part of the distribution. Indeed, they may be the most important observations. For example, in investment decisions extreme future returns and their frequencies are among the most important things to know as they may lead to huge losses or huge gains. The overall gain or loss of such an investment may be determined by a few extreme events as opposed to the cumulative result of many small gains or losses. Similarly, when the question is about what damages to expect from floods or earthquakes, it's the largest of damages that we should care about.

Some data analysts call extreme values **outliers**. We prefer not using that term as it implies that they do not belong in the distribution, which may or may not be the case.

The most worrying problem with extreme values is that they may not show up in a dataset even if they are an inherent characteristic of a variable. Extreme values, by nature, are indeed rare. For example, data spanning a few years may not have instances of the most extreme losses on an asset that may come in the future.

Visualizing the distribution via a histogram is a good way to catch extreme values. What to do, if anything, to extreme values is a more difficult question. It depends both on why extreme values occur and what the ultimate question of the analysis is.

Extreme values that occur due to error should be replaced by a missing value marker or, in the rare case when one can infer the correct value, by that correct value. Error-caused extreme values are rare but are often straightforward to catch. For example, earnings a thousand times higher than the average in low-skilled occupations are surely an error and so are more than 168 working days a week.

More often, extreme values are an inherent part of the distribution. The size of the largest countries or firms, the salaries of chief executives in a firm, the best-paid actors among all actors, or a large drop in asset prices on the day of a stock-market crash are not errors in the data. What we do in such cases depends on the question of our analysis.

If the question is more relevant for the rest of the observations, we may discard the observations with extreme values and restrict the analysis to the rest. When we look into how salaries and other incentives may keep employees at a firm, it makes sense not to focus on chief executives but restrict the analysis to the other employees. When we want to know how distance to main highways affects the price of commercial real estate, we may discard the largest cities and focus on smaller towns. It is good practice to be explicit about such decisions when presenting the results of the analysis, saying that the analysis is relevant for some kinds of observations but not for others.

In fact, what, if anything, we should do with observations with extreme values of a variable depends also on the role of the variable in our analysis. Starting with Chapter 4, Section 1, we will distinguish a y variable and one or more x variables in the analysis. Typically, our analysis will aim to uncover patterns in how values of y tend to differ for observations that have different values of x: hotel price differences by how far they are from the city center, differences in the quality of management by firm size, etc.

Data analysts tend to be conservative when discarding observations with extreme y values: they usually keep them unless they are errors. However, data analysts tend to be less conservative when discarding observations with extreme x values: they often discard them even if they are parts of the distribution. The reason is in the different roles y and x play in the analysis.

Discarding observations with extreme x values narrows the scope of the comparisons. That's a transparent decision, defining, or modifying, the focus of the analysis. In contrast, discarding observations with extreme y values changes the result of the comparisons. The consequences of this decision are not straightforward, and it's often safer to avoid those consequences. We'll discuss this issue in more detail in Chapter 8, Section 12.

Review Box 3.4 ***Extreme values***

- Some variables have extreme values: substantially larger or smaller values for a few observations than the values for the rest of the observations.
- Extreme values may be genuine or they may be an error.
  - When errors, extreme values should be corrected or treated as missing values.
  - When genuine, extreme values should be kept, unless we have a substantive reason to restrict the analysis to the subset of observations without extreme values.

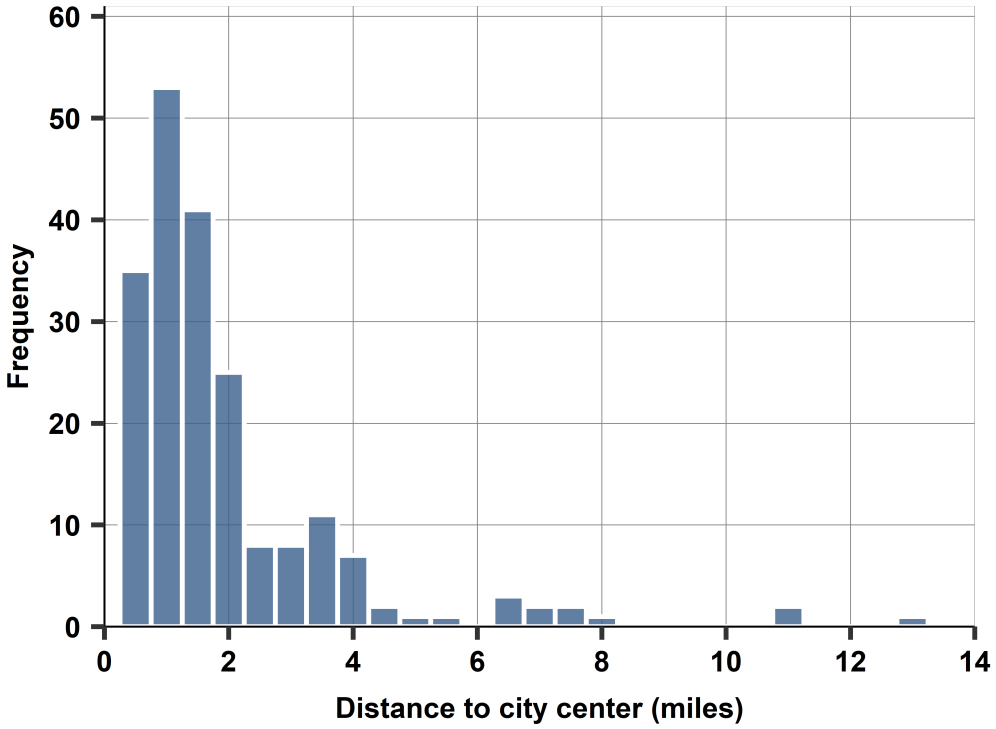
6 A2: Case Study – Finding a good deal among hotels: data exploration

Identifying and managing extreme values

Both price and distance to the city center have extreme values in the Vienna `hotels-vienna` data. We will use this data to find hotels that are underpriced for their location and quality. Hotel price will be our y variable, and so we drop extreme values only if they are errors. In contrast, distance will be an x variable, so we can drop extreme values even if they are parts of the distribution if we want to narrow our analysis.

We first look at distance. The data here includes all hotels in Vienna with 3 to 4 stars, regardless of their distance to the city center. Figure 3.4 shows the histogram of distance.

Figure 3.4: Histogram of distance to the city center.



Source: `hotels-vienna` data. Vienna, all hotels with 3 to 4 stars N = 217

For this histogram we used 0.5-mile-wide bins. This way we can see the extreme values in more detail even though the rest of the histogram looks a bit less nice (too uneven). The y axis shows the frequency. The histogram shows three hotels above 8 miles: two at around 11 miles and one at close to 13 miles. We see another group of hotels between 6 and 8 miles that are a little separated from the rest of the distribution (no value between 8 and 11 miles).

We decided to drop the three hotels that are more than 8 miles away from the city center and keep the ones at 6 to 8 miles. The extreme values we dropped are not errors. But they are values that would not be relevant for the main question of the analysis: finding hotels that are underpriced relative to their distance to the city center (and their quality). Eleven and 13 miles are far enough from the city center that we think we wouldn't choose these hotels if our aim was to stay in Vienna to explore the city. At the same time we didn't discard the hotels at 6 to 8 miles, thinking that maybe some of them are so inexpensive that they could be good deals even factoring in their distance. Note that this decision is arbitrary, but one such decision is necessary. In a thorough analysis, we would see if including the 8+ miles hotels, or excluding the 6–8 miles hotels, changes the answer to our question.

To better understand the features of hotels far from the center, we investigated our "city_actual" variable. It turns out that even within the 8-mile radius, a few hotels are in villages (such as Voesendorf) that are related to Vienna but are not Vienna proper. Hence, we decided to drop these 7 hotels, too. The result is a sample of 208 hotels.

Next we looked at prices, using the data of the 208 hotels that are within 8 miles from the city center and are in Vienna proper. Earlier, we pointed out a single observation with a price of 1012 dollars. This

is an extreme value indeed. We dropped it because we decided that it is almost surely an error. It's a room in a 3 star hotel is unlikely to cost over a thousand dollars. Moreover, in the `hotels-europe` dataset that contains prices for several dates for these same, and many more, hotels, the price of this hotel is around 100 dollars on all other dates not 1000 dollars.

We have identified, on Figure 3.2b, an additional observation with an extreme value, close to 400 dollars. We decided to keep this observation. This is a high price for this kind of a hotel (3 stars). At the same time, inspecting the the `hotels-europe` dataset reveals another date with a similarly high price, and the prices on the rest of the dates, while considerable lower, are not an lower by an order of magnitude that would indicate a digit error. Thus, we can't conclude that this extreme value is an error.

To summarize, our exploratory analysis led us focus our sample. Our key steps were:

1. We started with full data N=428.
2. We inspected the histograms of the qualitative variables
    - Accommodation type - could be apartment, etc.; kept hotels only N=264.
    - Stars - kept on 3, 3.5 4 stars N=218.
3. We looked at quantitative variables, focusing on extreme values.
    - Price: the extreme value of 1012 dollars is a likely error, dropped it; kept all others N= 217.
    - Distance: some hotels are far away; defined cutoff; dropped beyond 8 miles N=214.
    - Distance, one more step: looked at variable city_actual, realised that some hotels are not in Vienna proper; dropped them N=207.
4. The final sample is **Hotels, 3 to 4 stars, below 400 dollars, less than 8 miles from center, in Vienna proper N=207.**

7 Good graphs: Guidelines for data visualization

Now that we have introduced visualization of distributions, let's pause and spend some time on how to produce good graphs in general. These thoughts are meant to guide all decisions that go into producing graphs. They correspond to the practice of data visualization professionals; see some important references at the end of the chapter.

Before we begin, let us point out that our principles and suggestions are aimed at data analysts not visual designers. Typically, data analysts want to spend less time designing graphs than visual designers. As a result, they are more likely to use ready-made graph types and templates, and they benefit more from following a few simple rules instead of engaging in a creative process each and every time. Just like with any of our advice, this is not a must do list. Instead it shows how to think about decisions data analysts must take. You may take other decisions, of course. But those decisions should be conscious instead of letting default settings determine the look of your graphs.

The starting principle is that all of our decisions should be guided by the usage of the graph. The **usage of a graph** is a summary concept to capture what we want to show and to whom. Its main elements are purpose, focus, and audience. Table 3.1 explains these concepts and gives some examples. Note

Table 3.1: Usage of visualization

Concept	Explanation	Typical cases	Examples
Purpose	The message that the graph should convey.	Main conclusion of the analysis.	y and x are positively
		An important feature of the data.	There are extreme values of y at the right tail of the distribution.
		Documenting many features of a variable	All potentially important properties of the distribution of y.
Focus	On graph, one message.	Multiple related graphs for multiple messages.	A histogram of y that identifies extreme values, plus a box plot of y that summarizes many other features of its distribution
Audience	To whom the graph wants to convey its message.	Wide audience	Journalism
		Non-data-analysts with domain knowledge.	Decision makers
		Analysts	Fellow data analysts, or our future selves when we want reproduce the analysis

that some of the examples use graphs that we haven't introduced yet; this is because we want the advice to serve as reference later on.

Once usage is clear, the first set of decisions to make are about how we convey information: how to show what we want to show. For those decisions it is helpful to understand the entire graph as the overlay of three graphical objects:

1. Geometric object; the geometric visualization of the information we want to convey, such as a set of bars, a set of points, a line; multiple geometric objects may be combined.
2. Scaffolding: elements that support understanding the geometric object, such as axes, labels, and legends.
3. Annotation: adding anything else to emphasize specific values or explain more detail.

When we design a graph, there are many decisions to make. In particular, we need to decide how the information is conveyed: we need to choose a **geometric object**, which is the main object of our graph that visualizes the information we want to show. The geometric object is often abbreviated as a geom. The same information may be conveyed with the help of different geometric objects, such as bar charts for a histogram or a curved line for a density plot. In practice, a graph may contain more than one geometric object, such as a set of points together with a line, or multiple lines.

Choosing the details of the geometric object, or objects, is called **encoding**. Encoding is about how we convey the information we want using the data we have, and it means choosing elements such as height, position, color shade. For a graph with a set of bars as the geometric object, the information

may be encoded in the height of these bars. But we need to make additional choices, too. These include general ones such as color and shade as well as choices specific to the geometric object, such as width of the bars or lines, or size of the dots.

In principle, a graph can be built up freely from all kinds of graphical objects. Data visualization experts and artists tend to follow this bottom-up approach. In contrast, most data analysts start with choosing a predefined **type of graph**: a single or a set of some geometric objects and a scaffolding, possibly some annotation. One graph type is the histogram that we introduced earlier in this chapter. Histograms are made of a set of bars as a geometric object, with the information in the data (frequency) encoded in the height of the bar. The scaffolding includes an x-axis with information on the bins, and a y axis denoting either the absolute (count) or relative frequency (percent). We'll introduce many more graph types in this chapter and subsequent chapters of the textbook. Table 3.2 offers some details and advice.

Table 3.2: The geometric object, its encoding and graph types

Concept	General advice	Examples
Geometric object	Pick an object suitable for the information to be conveyed	Set of bars comparing quantity A line showing value overtime
One or more geoms	May combine more than one geoms to support message or add context	Dots for the values of a time series variable over time, together with a trend line
Encoding	Pick one encoding only	Histogram: height of bars encodes information (frequency); don't apply different colors or shades.
Graph type	Can pick a standard object to convey information	Histogram: bars to show frequency Scatterplot: values of two variables shown as a set of dots.

The next step is **scaffolding**: deciding on the supporting features of the graph such as axes, labels, and titles. This decision includes content as well as format, such as font type and size. Table 3.3 summarizes and explains the most important elements of scaffolding.

Lastly, we may add **annotation** – if there is something else we want to add or emphasize. Such additional information can help put the graph into context, emphasize some part of it. The two main elements of annotation are notes and visual emphasis, see Table 3.4 for some advice

Table 3.3: Scaffolding

Element	General advice	Examples
Graph title	Title should be part of the text; it should be short emphasizing main message	Swimming pool ticket sales and temperature (or: Swimming pool ticket sales higher when temperature is high)
Axis title	Each axis should have a title, with the name of the actual variable and unit of measurement	Distance to city center (miles) Household income (thousand US dollars)
Axis labels	Value numbers on each axis, next to tics, should be easy to read	0, 2, 4, 6,8,10,12 14 kms for distance to city center
Gridlines	Add horizontal and, if applicable, vertical gridlines to help reading off numbers	Vertical gridlines for histogram Both horizontal and vertical gridlines for scatterplots
Legends	Add legend to explain different elements of the geometric object Legends are best if next to the element they explain	Two groups, such as "male", "female" Time series graphs for two variables; it's best to put legends next to each line
Fonts	Large enough size so the audience can read them	Font size "10 "

Table 3.4: Annotation

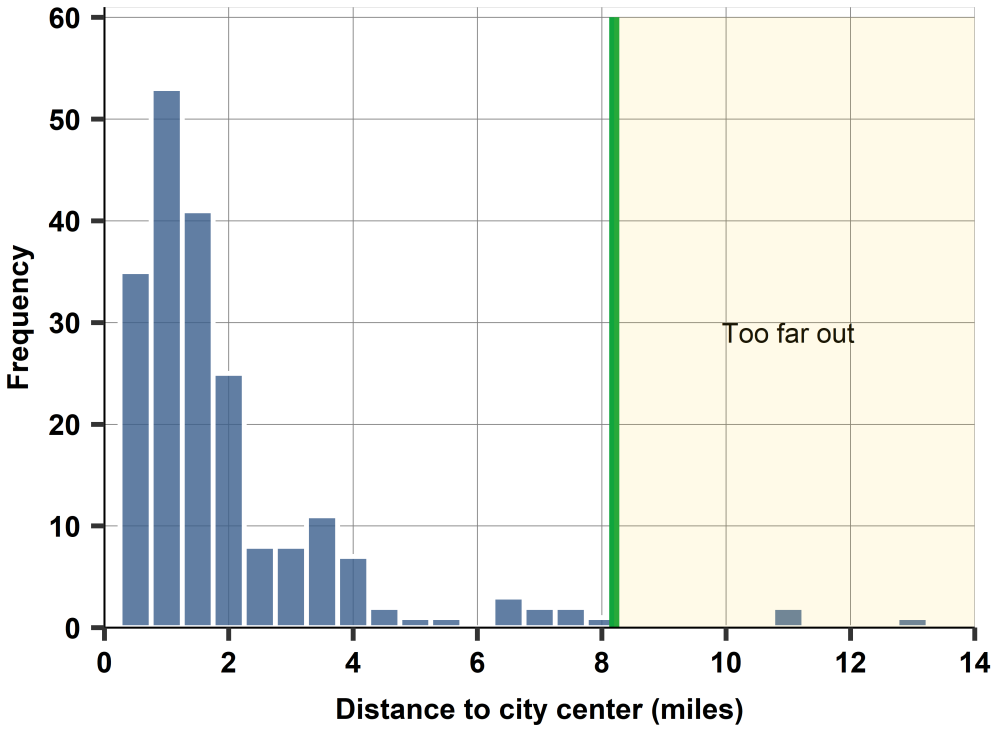
Concept	General advice	Examples
Graph notes	Add notes to describe all important details about how the graph was produced and using what data.	Lowess nonparametric regression with scatterplot. Hotels-vienna dataset. Vienna, all hotels with 3 to 4 stars. N=217
Added emphasis	May add extra annotation to graph to emphasize main message	A vertical line or circle showing extreme values on a histogram An arrow pointing to a specific observation on a scatterplot

8 A3 Case study – Finding a good deal among hotels: data exploration

The anatomy of a graph

Let us use a previous graph here, to illustrate the most important parts of a good graph. Recall that we should should keep in mind usage, encoding, scaffolding, annotation. We use are previous graph with the histogram of hotel distance to the city center (Figure 3.5), but here we added some annotation.

Figure 3.5: Histogram of distance to the city center



Source: `hotels-vienna` data. Vienna, all hotels with 3 to 4 stars. N=217.

Usage. We use this graph to search for extreme values. The main message is that the three hotels that are located more than 8 miles away from the city center are separate from the rest. The target audience is a specialized one: fellow data analysts. The figure may serve to document the reasons of our decisions, again a special usage.

Encoding. The graph shows the distribution in enough detail to spot extreme values, but it also shows the main part of the distribution to put those extreme values in context. Our encoding choice was a histogram with a 0.5-mile bin. Alternative choices would have had wider or narrower bins for the histogram or a density plot. Choosing the histogram with a 0.5-mile led to a balance for the usage of the graph: showing the main part distribution and the extreme values. One message, one encoding: we use a single color as bar height is enough to help compare through distance bins.

Scaffolding. The x axis denotes distance to city center. Although bins are every 0.5 mile, the labels are at 2-mile intervals to help readability. The y axis denotes the number of hotels per bin. It is absolute

frequency here not percentage, because our focus is on counting observations with extreme values. Notice the labels on the y axis: they are in increments of 10. The scaffolding includes horizontal and vertical gridlines to help reading off numbers.

Annotation. We point out the main message of the graph: the three hotels beyond 8 miles from the center appear to form their own cluster. We used a colored rectangle, but we could have circled them, or had an arrow pointed at them. For a scientific audience we could have skipped that annotation because that audience would understand the issue anyway and may appreciate a clean histogram.

9 Summary statistics for quantitative variables

A histogram of a quantitative variable can inform us about the shape of the distribution, whether it has extreme values, or where its center is approximately. But visual displays don't produce the numbers that are often important to answer our questions. How far are hotels from the city center in general? What's the spread of prices? How skewed is the distribution of customer ratings? To answer such questions we need numerical summaries of variables. They are called statistics, and this section covers the most important ones.

A **statistic** of a variable is a meaningful number that we can compute from the data. Examples include mean income or the range of prices. Basic **summary statistics** are numbers that describe the most important features of the distribution of a variable. Summary statistics can answer questions about variables in our data, and they often lead to further questions to examine.

Most readers are probably well acquainted with many summary statistics, including the **mean**, the **median** (the middle value), various **quantiles** (terciles, quartiles, **percentiles**, etc.), and the **mode** (the value with the highest frequency in the data). The mean, median, and mode are also called measures of central tendency, because they give an answer to where the center of the distribution is. Those answers may be the same (the mean, median, and mode may be equal, or very close to each other), or they may be different.

Importantly, we use the terms mean, average, and expected value as synonyms, and we use the notation $E[x]$ as well as $\bar{x}$.

The mean of a quantitative variable is the value that one can expect for a randomly chosen observation. The mean of a 0/1 binary variable is the proportion of observations with value 1, and no observation would have the expected value as it is between 0 and 1 whereas the value can be only 0 or 1.

Similarly, most readers know the most important measures of spread, such as the **range** (the difference between the largest and smallest value), **inter-quantile ranges** (e.g., the 90–10 percentile range, or inter-quartile range), the **standard deviation** and the **variance**. The standard deviation captures the typical difference between a randomly chosen observation and the mean. The variance is the square of the standard deviation. The variance is a less intuitive measure, but it is easier to work with because it is a mean value itself. The formulae are respectively:

$$Var[x] = \frac{\sum(x_i - \overline{x})^2}{n} \tag{3.1}$$

$$Std[x] = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n}} \tag{3.2}$$

Note that alternative formulae for the variance and the standard deviation divide by $n-1$ not n. Most data are large enough that this makes no practical difference. It turns out that dividing by $n-1$ is the

correct formula if we use the statistics in the data to infer the standard deviation in the population that our data represents (see more details in Chapter 5, Section 12). Since it makes little difference in practice, and dividing by n is easier to remember, we will continue to divide by n in this textbook.

The standard deviation is often used to re-calculate differences between values of a quantitative variable, in order to express those values relative to what a typical difference would be. In a formula, this amounts to dividing the difference by the standard deviation. Such measures are called **standardized differences**. A widely used standardized difference is from the mean value; it is called the **standardized value of a variable** or the **z-score** of the variable.

$$x_{standardized} = \frac{(x - \bar{x})}{Std[x]} \tag{3.3}$$

While measures of central value (mean, median, etc.) and spread (range, standard deviation, etc.) are usually well known, summary statistics that measure skewness are less frequently used. At the same time skewness can be an important feature of a distribution, showing whether a few observations are responsible for much of the spread. Moreover, there is a very intuitive measure for skewness (which exists in a few variants).

Recall that a distribution is **skewed** if it isn't symmetric. A distribution may be skewed in two ways, having a long left tail or having a long right tail. A long left tail means having a few observations with small values with most observations having larger values. A long right tail means having a few observations with large values with most observations having smaller values. Earlier we showed that the hotel price distributions has a long right tail – such as in Figure 3.3b. That is quite typical: skewness with a long right tail is frequent among variables in business, economics, and policy, such as with prices, incomes, population, etc.

The statistic of skewness compares the mean and the median and is called the **mean–median measure of skewness**. When the distribution is symmetric, its mean and median are the same. When it is skewed with a long right tail, the mean is larger than the median: the few very large values in the right tail tilt the mean further to the right. Conversely, when a distribution is skewed with a long left tail, the mean is smaller than the median: the few very small values in the left tail tilt the mean further to the left. The mean–median measure of skewness captures this intuition. In order to make this measure comparable across various distributions, we use a standardized measure, dividing the difference by the standard deviation. (Sometimes this measure is multiplied by 3, and then it's called Pearson's second measure of skewness. Yet other times the difference is divided by the mean, median or some other statistic.)

$$Skewness = \frac{\overline{x} - median[x]}{Std[x]} \tag{3.4}$$

The next table summarizes the most important descriptive statistics we discussed.

Table 3.5: Summary table of descriptive statistics

Type of statistic	Name of statistic	Formula	Intuitive content
Central value	Mean	$\bar{x} = \frac{1}{n}\sum x_i$	The value we expect from a randomly chosen observation
	Median	-	The value of the observation in the middle
	Mode	-	The value (bin) with the highest frequency
Spread	Range	$max[x] - min[x]$	Width of the interval of possible values
	Inter-quantile range	$q_{upper}[x] - q_{lower}[x]$	Distance between the upper quantile and the lower quantile
	Variance	$Var[x] = \frac{\sum(x_i-\bar{x})^2}{n}$	-
	Standard deviation	$Std[x] = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}$	Typical distance between observations and the mean
Skewness	Mean–median skewness	$Skewness = \frac{\bar{x}-median(x)}{Std[x]}$	The extent to which values in the tail pull the mean

10 B1 Case Study – Comparing hotel prices in Europe: Vienna vs London

Comparing distributions over two groups.

We are interested in comparing the hotel markets over Europe, and would like to learn about characteristics of hotel prices. To do that, let us focus on comparing the distribution of prices in Vienna to another city, London.

The data we use is an extended version of the dataset we used so far. The dataset `hotels-europe` includes the same information we saw, for 46 cities and 10 different dates. We will explore it more in Chapter 9. For this case study, we consider the same date as earlier (weekday in November 2017) for Vienna and London.

We focus on hotels with 3 to 4 stars that are in the actual city of Vienna or London. We have no extreme value of price in London, so we need to drop the single, above 1000 dollars priced hotel in Vienna. In our sample, there are $N = 435$ hotels in the London dataset compared to the $N = 207$ hotels in Vienna.

The next figure shows two histograms side by side. To make them comparable, they have the same bin size (20 dollars), the same range of axes, and each histogram shows relative frequencies. The same range of axes means that the x axis goes until 500 dollars for Vienna, too, even though the highest price there is below 500 dollars.

The histograms reveal many important features. Here is a selected set of observations we can make:

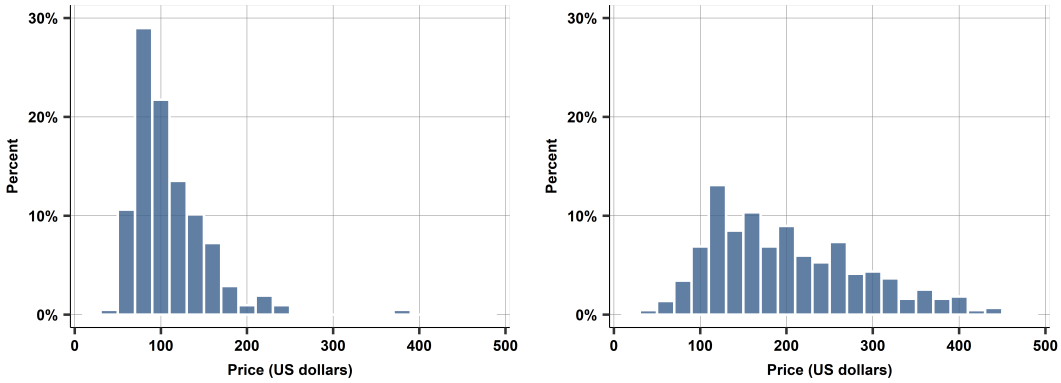
- The range starts slightly below 100 dollars in both cities but it ends below 500 dollars in Vienna while it goes above 800 dollars in London.

- The London distribution of prices covers more of the higher values, and it is more spread out (i.e. has a larger difference between the minimum and maximum price).

- Hotel prices tend to be higher in London.

- Both distributions have a single mode, but their location differs. The mode in Vienna is at the lower end of the price distribution, around 100 dollars. The mode in London is around 200 dollars even though both distributions start around 50 dollars.

Figure 3.6: The distribution of hotel price in Vienna and London

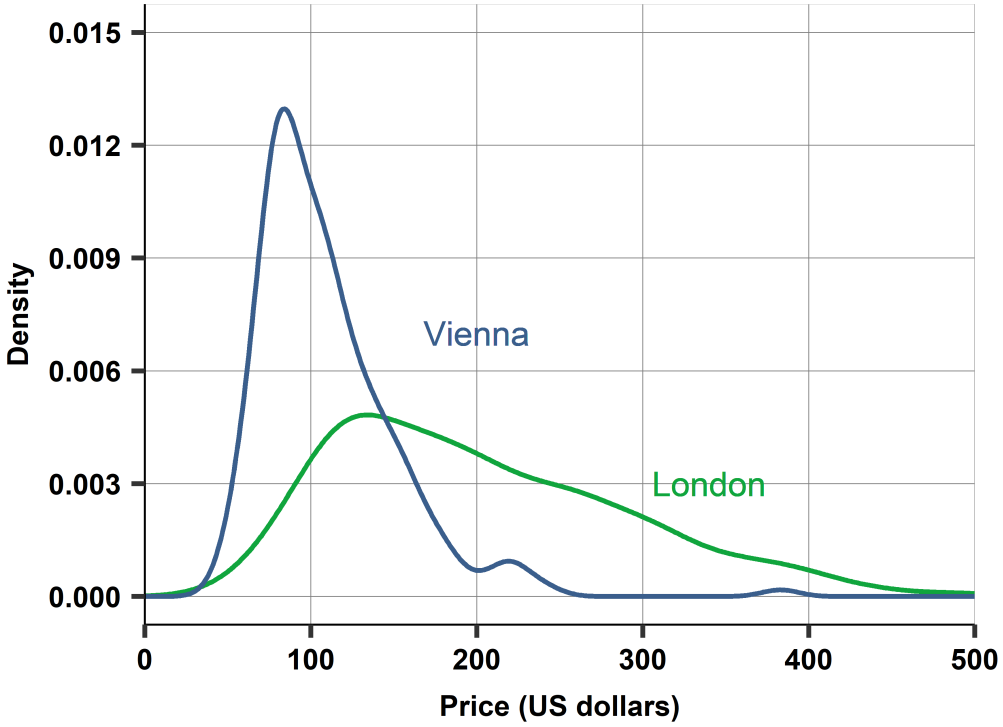


(a) Vienna (b) London

Source: `hotels-europe` dataset. Vienna and London, 3-4 stars hotels only, for a 2017 November weekday. Vienna: N=207, London: N=435.

The same price distributions can be visualized with the help of density plots. Figure 3.7 shows the Vienna and London distributions laid on top of each other. The density plots do not convey more information than the histograms. In fact, they are less specific in showing the exact range or the prevalence of extreme values. But comparing density plots on a single graph is just easier – we can see immediately where the mass of hotels are in Vienna and in London.

Figure 3.7: Density plots of hotel prices: Vienna and London



Note: *Kernel density estimates with Epanechnikov smoothing method.*
Source: `hotels-europe` dataset. Vienna and London, 3-4 stars hotels only, for a 2017 November weekday. Vienna: N=207, London: N=435.

We can quantify some aspects of the distributions, too. Table 3.6 contains some important summary statistics in the two datasets.

Table 3.6: Descriptive statistics for hotel prices in two cities.

city	n	mean	median	min	max	sd	skew
London	435	202.36	186	49	491	88.13	0.186
Vienna	207	109.98	100	50	383	42.22	0.236

Source: `hotels-europe` dataset. Vienna and London, weekday, November 2017

Average price is 110 dollars in Vienna and 202 dollars in London. The difference is 92 dollars, which is almost an extra 90% relative to the Vienna average. The mean is higher than the median in both cities, indicating a somewhat skewed distributions with a long right tail. Indeed, we can calculate the standardized mean-median measures of skewness, and it is more positive in Vienna $((110-100)/42 = 0.236)$ than in London $((202-186)/88 = 0.186)$.

The range of prices is substantially wider in London $(491-49 = 442)$ than in Vienna $(383-50 = 343)$, and the standard deviation shows a substantially larger spread in London (88 versus 42). The first column shows that the London dataset has about two times as many observations (435 versus 207).

These summary statistics are in line with the conclusions we drew by inspecting the visualized distributions. Hotel prices in London tend to be substantially higher on average. They are also more spread, with a minimum close to the Vienna minimum, but many hotels above 200 dollars. These together imply that there are many hotels in London with a price comparable to hotel prices in Vienna, but there are also many hotels with substantially higher prices.

11 Visualizing summary statistics

The summary statistics we discussed are often presented in table format. But there are creative ways to combine some of them in graphs. We consider two such graphs.

The more traditional visualization is the **box plot**, also called the box and whiskers plot shown in Figure 3.8a. The box plot is really a one-dimensional vertical graph, only it is shown with some width so it looks better. The center of a box plot is a horizontal line at the median value of the variable, placed within a box. The upper side of the box is the third quartile (the 75^{th} percentile) and the lower side is the first quartile (the 25^{th} percentile). Vertical line segments on both the upper and lower side of the box capture most of the rest of the distribution. The ends of these line segments are usually drawn at 1.5 times the inter-quartile range added to the third quartile and subtracted from the first quartile. These endpoints are called **adjacent values** in the box plot. The lines between the lower (upper) adjacent value and 25th (75th) percentile range are called **whiskers**. Observations with values not contained within those values are usually added to the box plot as dots with their respective values. The box plot conveys many important features of distributions such as their skewness and shows some of the quantiles in an explicit way.

A smarter-looking alternative is the **violin plot** shown in Figure 3.8b. In essence, the violin plot adds a twist to the box plot by overlaying a density plot on it. Violin plots show the density plot on both sides of the vertical line, but there is no difference between the two sides. In a sense, similarly to a box plot, we have two sides here purely to achieve a better look. Compared to the traditional box plot, the basic violin plot shows fewer statistics and does not show the extreme values. In exchange, it gives a better feel for the shape of the distribution. As there are many complementing features of box plots and violin plots, we advise you to consider both.

Figure 3.8: The structure of the box plot and the violin plot

(a) Box plot

(b) Violin plot

Review Box 3.5 ***Summary statistics and their visualization***

- Measures of central value: Mean (average), median, other quantiles (percentiles), mode.
- Measures of spread: Range, inter-quantiles ranges, variance, standard deviation
- Measure of skewness: The mean – median difference.
- The box plot is a visual representation of many quantiles and extreme values.
- The violin plot mixes elements of a box plot and a density plot.

12 C1: Case Study –Measuring home team advantage in football

Distribution and summary statistics

The idea of home team advantage is that teams that play on their home turf are more likely to play better and win compared to the same game played at the other team's stadium (the other team is also called the "away" team). In particular, this case study asks whether professional football (soccer) teams playing in their home stadium have an advantage, what is the extent of that advantage, and what are the causes of such an advantage.

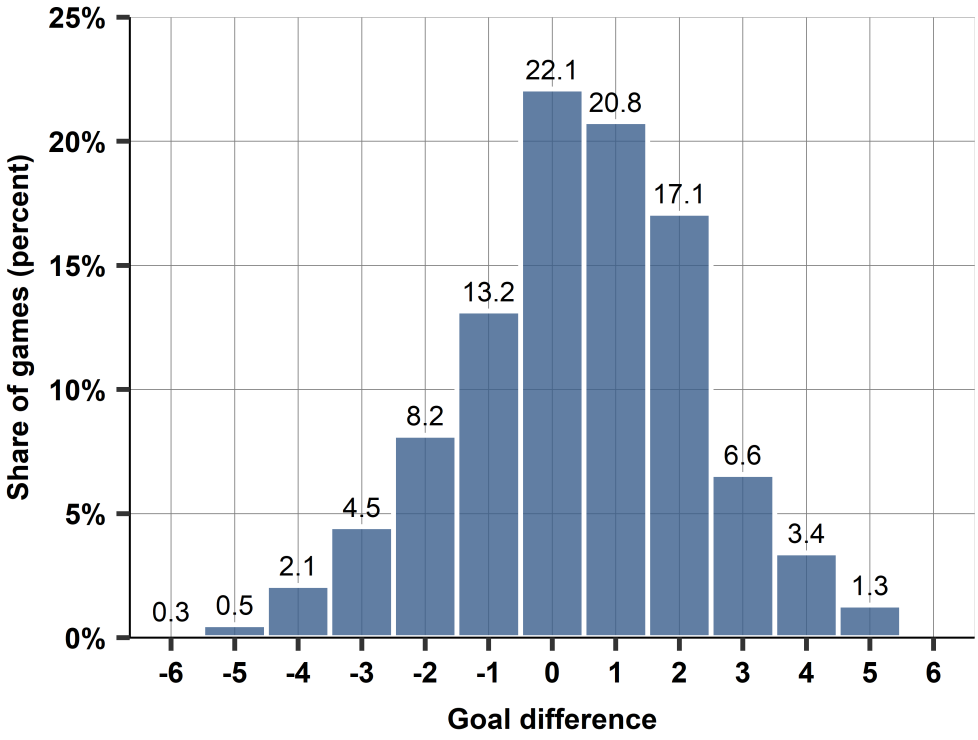
These questions are interesting in order for fans to know what to expect from a game, but also for professional managers of football teams who want to maximize the number of wins and the support of fans. If home team advantage is important, managers need to know its magnitude to benchmark the performance of their teams. More ambitiously, they would want to understand its sources in order to choose strategies that strengthen its effect when playing at home and dampen its effect when playing away.

Here we use data from the English Premier League, with the same data we started with in Chapter 2. Here we focus on games in the 2016/7 season. In each season, each pair of teams plays twice, once in the home stadium of one team, and once in the home stadium of the other team. This gives 380 games total (20×19: each of the 20 teams plays once at home against each of the other 19 teams).

The most important variables are the names of the home team and the away team and the goals scored by each team. From these two quantitative variables we created a goal difference variable: home team goals – away team goals. Examining the distribution of this goal difference can directly answer our question of whether there is a home team advantage and how large it is.

Let's look at the distribution of the home team – away team goal difference first. Figure 3.9 shows the histogram. While the goal difference is a quantitative variable, it doesn't have too many values so we show a histogram that shows the percentage of each value instead of bins.

Figure 3.9: The distribution of home team – away team goal difference



Source: `football_games` data. English Premier League, season 2016–2017, all games. N=380.

The mode is zero: 22.1% of the games end with a zero goal difference (a draw). All other goal differences are of smaller percentage – the larger the difference, the smaller their frequency. This gives an approximately bell-shaped curve, except it is skewed with more observations (taller bars) to the right of zero. That suggests home team advantage already. But let's look more closely into this histogram.

The most striking feature of the histogram is that for each absolute goal difference, the positive value is more frequent than the negative value. The home team – away team goal difference is +1 in 20.8% of the games while it is -1 in 13.2% of the games; it's +2 in 17.1% of the games and -2 in only 8.2% of the games, etc. This clearly shows that home teams score more goals so there is a home team advantage. It also seams pretty large. But how large is it? To answer that we need to provide a number and put it into context.

To that end, we calculated the mean and the standard deviation of the goal difference as shown in Table 3.7. Moreover, we calculated the relative frequency of games in which the home team wins (positive goal difference), games in which the away team wins (negative goal difference), and games that end with a draw (zero goal difference: we know their proportion is 22%). Table 3.7 shows the results.

Table 3.7: Describing the home team – away team goal difference.

Statistic	value
Mean	0.4
Standard deviation	1.9
Percent positive	49
Percent zero	22
Percent negative	29
Number of observations	380

Source: `football` dataset. English Premier League, season 2016–2017, all games. N=380.

The average goal difference is 0.4: on a randomly chosen game in the 2016/7 season of the English Premier League, the home team is expected to score 0.4 more goals.

Is this a large difference? We can answer that in two ways. First, 0.4 goals appears to be a sizeable difference: it's almost one goal every two games. Second, we can put it into the context of the spread of the goal difference across games, with the help of the standard deviation. The standard deviation is 1.9, showing the typical deviation from the average. The average goal difference, 0.4, is thus about one fifth of the typical deviation. Viewed this way, the average goal difference is not negligible but not huge, either.

While football is played for goals, what truly matters is who wins a game. From this point of view, the home team advantage looks large again. Forty-nine percent of the games are won by the home team, and only 29% are won by the away team. That's two thirds more.

To conclude, we have uncovered a sizeable home team advantage in the English Premier League. Home teams score 0.4 goals more on average, and they win almost 50% of the time but lose only about 30% of the time. This should serve as a useful benchmark for what to expect from each game. It also shows the importance of the phenomenon. So anyone who cares about the home team advantage should do more analysis to try to learn about its cause, or causes.

What did we learn from this exercise? The simple tools of exploratory analysis revealed something interesting and important. The histogram showed clear evidence for the home team advantage. The mean goal difference summarized the extent of the home team advantage in a single number that we could interpret on its own and in comparison with the standard deviation. Finally, we use the goal difference variable to compute the relative frequency of home wins versus away wins that showed the magnitude of home team advantage from a different, and arguably more relevant, point of view.

13 Good tables

In Section 7 we gave advice on how to produce good graphs. Here we do the same for tables. While graphs show visual representation of some results from the data, tables present statistics: meaningful numbers computed from the data. Despite these differences, the structured thinking that helps produce good graphs works just the same here. To produce a good table, data analysts need to think about its **usage**, choose **encoding** and **scaffolding** accordingly, and may add **annotation**.

First, usage: what's the purpose and who's the target audience. One important type wants to commu-

nicate the main result, or results, of the analysis to people who would use those results. A table of this type is called a **results table**, or a communication table. Just like good graphs, good communication tables are focused on one message. The other main table type is the **documentation table**. Its aim is to document exploratory data analysis. Documentation tables describe the structure of the data, one or more variables, or some other features such as missing values or extreme values. Such tables are also used to summarize the results of data cleaning and restructuring processes – e.g., by showing numbers of observations and statistics of key variables for the original data and the data we chose to work with in the end. Documentation tables are produced for other analysts who may want to have a deep understanding of the data, possibly to carry out their own analysis by reproducing or modifying what we did. As usual, those other analysts may include our future selves: documentation tables help us remember certain details of the data that may prove to be important during our analysis.

Usage of tables should guide all subsequent decisions. Encoding here means what numbers to present in the table, and in what detail. The numbers may be averages, counts, standard deviations, or other statistics – for the entire data table or for subsets of observations. Documentation tables tend to be large and include everything that is, or may become, important. In contrast, communication tables should be simple and focused. For example, when analyzing company data to understand churn (why some employees quit while others stay), a communication table on churn rates may show the percent who quit within, say, one year of entry, by a handful of major categories of job rank. A documentation table, on the other hand, may include more job categories, present numbers by year entry, and show numbers of employees that quit and stayed together with their ratio. The additional detail in the documentation table helps other analysts to reproduce our analysis in the future, and it helps us catch potential anomalies in the data.

A good practice of encoding is to show totals together with components. Even though showing totals is often redundant, it can be very informative as it helps grasp what the components mean, by highlighting what they should add up to. For example, when a table shows the percent of sales by day of the week in the rows of a table, it is good practice to show the total of 100 percent in the bottom row. Similarly, in a documentation table with numbers of observations in subgroups in the data, the total number of observations is usually included, too.

Another good practice is to include the number of observations in all tables. This helps the audience have a rough idea about the extent to which the results may generalize beyond the data (see Chapter 5). It may also help remind them whether the statistics refer to all observations or a group of them.

The second question of encoding is how much detail the numbers should have in the table. Here, too, usage should be our guide. Communication tables should have no more detail than necessary. For example, a table on churn percentages should include few, if any, decimal digits. If 123 employees leave the firm within six months in a dataset of 1234 employees, this proportion is 9.967585% to six decimal digits. We find it ridiculous to present this number to such a precision. 9.97% is better; 10% is even better. Usually, documentation tables have numbers in more detail. When reproducing steps of analysis, we want to see, step by step, whether we get the same numbers as the first analysis did. But, there too, many decimals of percentages can be confusing. Instead, it's more helpful to include the numbers of observations that make up a percentage.

Third, scaffolding. Tables should have titles that say what's in the table. They need to have row headings and column headings that say what's in each row and column. And they need to have notes to give all the important detail. The main message is often a good choice for a title (e.g., "Churn is twice as high in sales than in other job roles."). A more conservative choice is to describe what kind of comparison the table does (e.g., "Churn by job rank). It is good practice to keep the title short and focused. Similarly, row and column headings should be short. The notes then should contain all detailed information, including the longer description of row and column headings, source of data, if the statistics are computed from a subsample, etc.

Finally, we may want to add annotation to our table to highlight specific numbers. We may circle a number, use color or bold typeface, etc. One frequent annotation is to put asterisks next to certain statistics, something that we'll see from Chapter 9 on for regression tables.

As with all of our advice, these principles should be viewed as starting points. You should keep to them or deviate from them if you have reasons to. The important thing is to make decisions in conscious ways.

Table 3.8: Good tables

Layer	What that means here	Specifics	Examples
Usage	Purpose and audience	Presentation table	Average in groups; percent of observations in groups
		Documentation table	Summary statistics of key variables
Encoding	What numbers	Include totals; include number of observations	Have row/column totals; include 100% in table with percentages
	In what detail	Few details for presentation; more for documentation	Few or no decimals for percentages; exact numbers of observations for documentation
Scaffolding	Title	Focused titles	What statistics, in what groups
	Row and column heading	All rows and columns should have short and informative headings	
	Notes	All necessary information in notes	How variables and/or groups are constructed; source of data
Annotation	Optional: may highlight specific numbers	Circle, use different color, boldface, etc.	

14 C2 Case Study – Measuring home team advantage in football

The anatomy of a table

Let us return to our football case study that investigates home team advantage, and let's review the table that shows some statistics of the home versus away goal difference. We first repeat Table 3.7 above:

Table 3.9: Describing the home team – away team goal difference.

Statistic	value
Mean	0.4
Standard deviation	1.9
Percent positive	49
Percent zero	22
Percent negative	29
Number of observations	380

Note: *Repeating Table 3.7*
Source: `football` dataset. English Premier League, season 2016–2017, all games. N=380.

Let's start with the usage of this table. Its purpose is showing that there is home team advantage in this data, and it's not small. It's a presentation table: its target audience is people who want to learn about home team advantage in football but may not be data analysts. Thus, we kept the table simple. It has a single column and six rows, presenting six numbers altogether.

Usage dictates encoding: what numbers to present in the table. The first number we present is the mean of the home versus away goal difference. It's positive, showing that, on average, there is a home team advantage in this data. Its magnitude is 0.4, which can be appreciated in comparison with the standard deviation (1.9) shown in the next row. The next three rows show the percent of games with a positive home-away difference (when the home team won), the percent with zero difference (draws), and the percent with negative difference (the home team lost). Note that here we didn't put a total row, going against our own advice. That's because we decided that it would make the table look odd. Thus, we made a conscious decision. Finally, the last row shows the number of observations, adhering to our other general advice.

The title of the table is relatively short and meaningful. Note that it is about the variable (goal difference), not the concept it is meant to measure (home advantage). The table has column and row headings that are relatively short and to the point. The notes say what the data is about (games from a league in one season) and the name of the data source (so you can look for it in our data repository).

We could have made other choices with encoding and scaffolding to serve the same usage. Among the data exercises, you'll be invited to create your own version of this table with possibly other numbers and text, and you'll be invited to explain your choices.

15 Theoretical distributions

Before closing our chapter by reviewing the process of exploratory data analysis, let's take a detour on theoretical distributions. Theoretical distributions are distributions of variables with idealized properties. Instead of showing frequencies in data, they show more abstract probabilities: the likelihood of each value (or each interval of values) in a more abstract setting. That more abstract setting is a hypothetical "data" or "population," or the abstract space of the possible realizations of events.

But why should we care about theoretical distributions? The main reason is that they can be of great help when we want to understand important characteristics of variables in our data. Theoretical distributions are usually simple to describe and have a few well-established properties. If a variable in

our data is well approximated by a theoretical distribution, we can simply attribute those properties to the variable without having to check those properties over and over. It tuns out that there are surprisingly many real-world variables whose distributions are quite well approximated by one of the theoretical distributions.

Another reason to know about theoretical distributions is that some of them are useful in understanding what happens when we want to generalize from the data we have. This second reason may sound very cryptic for now, but it should be much clearer after having read Chapters 5 and 6.

Many theoretical distributions are known in statistics. We focus on two in this section and discuss a few more in Section 19.

Before we review these two, let's note something about language. When describing variables in the abstract, statisticians also call them random variables. The idea behind that terminology is that there is randomness in their value: we don't know what the value is before looking at it. While this terminology certainly makes sense, we do not use that expression in this textbook. Instead we simply call variables variables, without the qualifier "random," whether in a very abstract setting or in a particular dataset. Besides simplicity, our choice of terminology helps reserve the term "random" for events and things without pattern, as in random sampling or values missing at random.

The **normal distribution** is the best known and most widely used theoretical distribution of quantitative variables. It is a pure theoretical construct in the sense that it was derived mathematically from another distribution, the binomial (see Section 19). Variables with a normal distribution can in principle take on any value from negative infinity to positive infinity. The histogram of the normal distribution is bell shaped. For that reason the popular name for the normal distribution is the bell curve. It is also called the Gaussian distribution after the German mathematician who played a role in popularizing it (it was the French mathematician Laplace who first derived it).

The normal distribution is characterized by two parameters, usually denoted as μ and σ. They refer to the mean (μ) and the standard deviation (σ). The variance is the square of the standard deviation, σ^2.

A special case of the normal is the **standard normal distribution**. It is a normal distribution with parameters $\mu = 0$ and $\sigma = 1$: its mean is zero and its standard deviation is one (and thus its variance is also one). If a variable x is normally distributed with mean μ and standard deviation σ, its transformed version is distributed standard normal if we take out the mean and divide this difference by the standard deviation: $\frac{(x-\mu)}{\sigma}$.

It turns out that when we transform a normally distributed variable by adding or multiplying by a number, the result is another normally distributed variable, with appropriately transformed mean and standard deviation. It also turns out that when we add two normally distributed variables, the resulting new variable is also normally distributed, and its mean is the sum of the means of the two original variables. (The standard deviation is a function of the original standard deviations and the correlation of the two variables.)

Some variables in real data are well approximated by the normal distribution. The height of people in a population is usually approximately normal, and so is their IQ, a measure of intelligence (although that is in part because the tests behind the IQ measure are constructed that way). Variables in real life are well approximated by the normal distribution if they are a result of adding up many small things. That's because the normal is a generalization of the binomial, which is the sum of Bernoulli variables.

The normal is a bad approximation to real-life variables with extreme values. Extreme values are very

unlikely in the normal distribution. For example, the normal distribution formula (see in the Under the hood section in Section 19) suggests that only 0.1% of the values of a normally distributed variable should be more than three standard deviations above the mean; only 0.0000001% (one in a billion) should be higher than six standard deviations above the mean.

Besides the absence of extreme values, symmetry is another feature of the normal distribution that makes it a bad approximation to many economic variables. Earnings, income, wealth, and firm productivity are usually asymmetrically distributed with long right tails. One theoretical distribution that may be a better approximation to such variables is the lognormal distribution.

The **lognormal distribution** is very asymmetric, with a long right tail, potentially including many extreme values at the positive end (but not at the other end). The lognormal distribution is derived from the normal distribution. If we take a variable that is distributed normally (x) and raise e to its power (e^x), the resulting variable is distributed lognormally. The old variable is the natural logarithm of the new variable, hence the name of the new distribution (the log of which is normal). Because we raised e to the power of the original variable, the values of the resulting lognormal variable are always positive. They range between zero and positive infinity (never reaching either). By convention, the parameters of the lognormal are the mean μ and standard deviation σ of the original, normally distributed variable, which is the logarithm of the new variable. Thus the mean and the standard deviation of the lognormal are complicated functions of these parameters. They are: $e^{\mu+\frac{\sigma^2}{2}}$ and $\sqrt{e^{\mu+\frac{\sigma^2}{2}}e^{\sigma^2-1}}$.

There are real life variables that are approximately lognormally distributed. These include distributions of price, income, and firm size. Variables are well approximated by the lognormal if they are the result of many things multiplied together (the natural log of them is thus a sum). Another way to think about lognormal variables is that their percentage differences are normally distributed. Thus, they do not have many extreme values in terms of percentage differences. (Note that normally distributed percentage differences translate into quite extreme differences in terms of absolute values.)

16 D1 Case Study – Distributions of body height and income

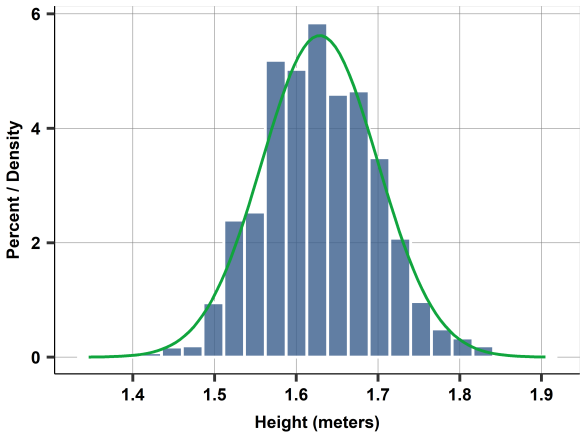
Data and describing distributions

Let us consider two examples using population data from the U.S.A. For this purpose, we use the `height-income-distributions` data from 2014.

Our first example is adult height, which is well approximated by the normal distribution for the vast part of the distribution, but not for extreme values. Average height among adult women in the U.S.A. population is around 164 cm (5'4"), with standard deviation 6.5 cm (2.5"). Thus women taller than 203 cm (6'10") should be less than one in a billion in the population, meaning that no such women are expected to live in the U.S.A. Yet there are quite a few American women that tall. Similarly to height, many real-life variables with some extreme values are still well approximated by the normal for the larger middle part of the distribution, often as much as over 99% of the distribution. Whether that makes the normal a good approximation or not for the purpose of data analysis depends on whether we are interested in extreme values or not.

Figure 3.10 below shows an example: the histogram of the height of American women aged 55–60. We have overlaid the density plot of the bell-shaped theoretical normal distribution that has the same mean and standard deviation as height in the data (1.63, 0.07).

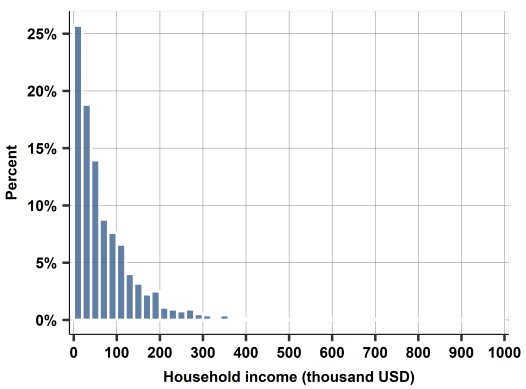
Figure 3.10: Example for approximately normal distribution: women's height



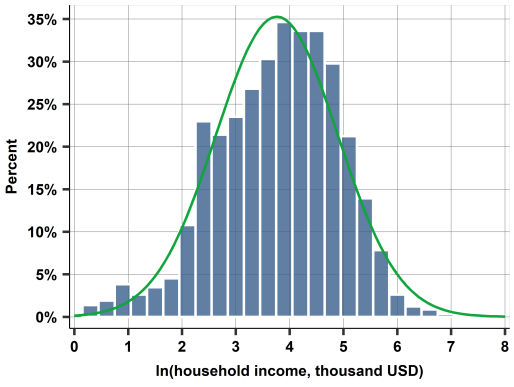
Note: *Histogram of the height of women aged 55–60, U.S.A. Overlayed with bell curve of the normal distribution.*
Source:Health and Retirement Study, `height-income-distributions` data, 2014. N=15662 (1988 females of age 55–60).

Second, let us consider a variable that has a few large values: household income. Figure 3.11 below shows the distribution of household income among households of women age 55 to 60 in the U.S.A. The left panel shows the histogram of household income (in 1000 dollars). The right panel shows the histogram of the log of household income, overlaid with the density plot of the normal distribution with the same mean and standard deviation.

Figure 3.11: An example for an approximately lognormal variable: Household income



(a) Distribution of household income



(b) Distribution of the natural log of household income

Note: *Household income of women age 55–60, U.S.A., 2014.*
Source: `Health and Retirement Study` data, 2014. N=15662 (1988 females of age 55–60).

17 Steps of exploratory data analysis

At the end of this chapter let us review the steps of exploratory data analysis (EDA). With clean and tidy data we can start our analysis. EDA is the first step of analysis: we should describe variables on their own before doing anything else. EDA helps data analysts "know their data," which is one of our most important pieces of advice. Getting familiar with the details of the data is important. Results of EDA may feed back into data wrangling (cleaning and restructuring data) if they uncover further issues. Moreover, these results should influence the next steps of analysis and may lead to transforming or changing the data. They are also essential for putting results in context. Finally, the results of EDA may lead to asking new questions that we previously haven't thought of.

Exploratory data analysis and data wrangling form an iterative process. EDA may uncover issues that call for additional data cleaning, stepping back again. Then, further steps of data analysis may raise additional questions or issues that call for additional EDA and, possibly, additional data wrangling. For example, when exploring employee churn at our company (whether and why some employees quit within a short time), we may uncover a 100% churn rate in one department (everybody seems to have quit). That may be true churn that we would explore in detail. Or, it may be something else – e.g., an error in the data or closing of a department. Those cases would call for fixing the error or modifying the data for the analysis by excluding employees of that department from the analysis.

Exploratory data analysis should start by focusing on the variables that are the most important for the analysis. If other variables turn out to be important, we can always come back to explore those variables later.

For each variable, it is good practice to start by describing its entire distribution. For qualitative variables with a few values, that is best done by listing the frequencies and producing simple bar charts. For quantitative variables we usually produce histograms to visualize the distributions.

An important first question is whether the distribution has extreme values. If yes, we need to decide what to do with them: if they are obvious errors, we replace them with missing values (or, less frequently, correct them). Similarly, if extreme values are not relevant for the question of the analysis, it makes sense to drop those observations. If they are not obvious errors and may be relevant for the question of the analysis, it is good practice to keep them. We may also decide to transform variables with extreme values.

Examining the distribution of a variable can answer other important questions such as whether the distribution is symmetric or skewed, how many modes it has, etc. We can also go further and see if it is well approximated by one of the theoretical distributions we considered.

The next step is looking at descriptive statistics. These summarize some important features of distributions in more precise ways. Having numeric values is necessary to appreciate the magnitude of subsequent results, understand the amount of variation in each variable, see if some values are very rare, etc. Summary statistics may call for data transformations. Examples include changing units of measurement (e.g., expressing values in thousand dollars) and creating relative measures (e.g., GDP per capita). It usually makes sense to look at minimum, maximum, mean, and median values, standard deviation and the number of valid observations. Other measures of spread may also be informative, and we may compute the mean–median measure of skewness, standardized by the standard deviation or the mean.

Exploratory data analysis may go substantially further, comparing variables or comparing distributions across groups of observations. As in all cases, our advice should be considered as a starting point. Data analysts should make conscious decisions at many points of the analysis, including the focus and

details of exploratory data analysis.

Review Box 3.6 ***Recommended steps of exploratory data analysis***

1. First focus on the most important variables. Go back to look at others if subsequent analysis suggests to.
2. For qualitative variables, list relative frequencies.
3. For quantitative variables, look at histograms.
4. Check for extreme values. Decide what to do with them.
5. Look at summary statistics.
6. Do further exploration if necessary (time series data, comparisons across groups of observations, etc.)

18 Summary and practice

18.1 Main takeaways

- Start all data analysis with exploratory data analysis.
  - Its results will be important for data cleaning, reality checks, understanding context, or asking further questions.
  - Explore all aspects of the distribution of your most important variables, including potential extreme values.
  - Produce visualization and tables guided by their usage.

18.2 Practice questions

1. The distribution of quantitative variables may be visualized by a histogram or a density plot (kernel density estimate). What's the difference between the two and which one would you use? List at least one advantage for each. How about qualitative variables with a few values?
2. The mean, median, and mode are statistics of central tendency. Explain what they are precisely.
3. The standard deviation, variance, and inter-quartile range are statistics of spread. Explain what they are and give the formula for each.
4. What are percentiles, quartiles and quintiles? Is the median equal to a percentile?
5. Why do we take the sum of squared deviations from the mean as a measure of spread, not the sum of the deviations themselves?
6. A distribution with a mean higher than the median is skewed. In what direction? Why? Give an intuitive explanation.

7. Extreme values are a challenge to data analysis if they are relevant for the question of the analysis. List two reasons why.
8. What kind of real-life variables are likely to be well approximated by the normal distribution? What are well approximated by the lognormal distribution? Give an example for each.
9. Are extreme values more likely in a normal or a lognormal distribution? Why?
10. Based on what you have learnt about measurement scales and descriptive statistics, decide if it is possible to calculate the mean, mode, and median of the following variables that tell us information about the employees at a company:
    (a) number of years spent in higher education
    (b) the level of education (high school, undergraduate, graduate, doctoral school)
    (c) field of education (e.g., IT, engineering, business administration)
    (d) binary variable that shows whether someone has a university degree
11. Take Figure 3.9 in Section 10. Describe its usage, encoding, and scaffolding. Would you want to add annotation to it? What and why?
12. Take Table **??** in Section 9 Describe its usage, encoding, and scaffolding. Would you do some things differently? What and why?
13. What kind of real-life variables are likely to be well approximated by the Bernoulli distribution? Give two examples.
14. What kind of real-life variables are likely to be well approximated by the binomial distribution? Give two examples.
15. What kind of real-life variables are likely to be well approximated by the power law distribution? Give two examples.

18.3 Data exercises

Easier and/or shorter exercises are denoted by [*] Harder and/or longer exercises are denoted by [**].

1. Pick another city beyond Vienna from the `hotels` data, and create a data table comparable to the one used in our case study. Visualize the distribution of distance and the distribution of price and compute their summary statistics. Are there extreme values? What would you do to them? Describe the two distributions in a few sentences. [*]
2. Use the data on used cars collected from a classified ads site (according to the Chapter 1 data exercise). Visualize the distribution of price and the distribution of age, and compute their summary statistics. Are there extreme values? What would you do to them? Describe the two distributions in a few sentences. [*]
3. Pick another season from the `football` data and examine the extent of home team advantage in ways similar to our case study. Compare the results and discuss what you find. [*]
4. Choose the same 2016/7 season from the `football` as in our data exercise and produce a different table with possibly different statistics to show the extent of home team advantage. Compare the results and discuss what you find. [*]

5. Choose a large country (e.g., China, Japan, Great Britain) and find data on the population of its largest cities. Plot the histogram of the distribution and create a table with the most important summary statistics. Plot the histogram of log population as well. Finally, create a log rank – log population plot. Is the normal, the lognormal, or the power-law distribution a good approximation of the distribution? Why? [*]

19 Under the hood: More on theoretical distributions

Let's introduce a few more useful concepts that theoretical statisticians use to describe theoretical distributions. We will rarely use these concepts in this textbook. However, they are frequently used in more traditional and more advanced statistical textbooks as well as in statistically more sophisticated analyses.

The first concept is the probability distribution function, or *pdf*. In essence, the pdf is the theoretical generalization of the histogram and its smoother cousin, the density plot. The pdf of non-continuous variables, such as Bernoulli, binomial, and other qualitative variables, shows the probability of each value in the distribution. The pdf of continuous variables, such as normal, lognormal, power-law, or uniform variables, shows the probability of each value and the values in its close neighborhood in the distribution. The pdf is expressed in a formula as a function of the parameters of the distribution, and it is often represented graphically as a bar chart like the histogram (for qualitative variables) or as a continuous curve like a density plot (for continuous variables).

A variation on the pdf is the *cdf*, the cumulative distribution function. For each value in the distribution, the cdf shows the probability that the variable is equal to that value or a lower value. Thus, at each value of the distribution, the cdf equals the pdf plus the sum of all pdf below that value. Hence the name "cumulative": the cdf accumulates the pdf up to that point. Similarly to the pdf, the cdf is expressed as a function of the parameters of the theoretical distribution and is often represented graphically.

The third advanced concept good to know about is the *moments* of distributions. Moments are a more general name for statistics like the mean and the variance for theoretical distributions. These are the expected value of the variable or the expected value of the square, cube, etc., of the variable, or, like the variance, the expected value of the square, cube, etc. of the variable minus its mean. Moments are numbered according to the power they take, so that the first moment is $E[x]$, the second moment is $E[x^2]$, etc. The variance is a second moment of the variable minus its mean so it is also called the second centered moment: $E[(x - E[x])^2]$.

Theoretical distributions are fully captured by a few *parameters*: these are statistics that determine the distributions. For each distribution we introduce their parameters, establish the range of possible values, show the shape of the histogram, and describe how the mean and standard deviation are related to the parameters of the distribution.

19.1 Bernoulli distribution

The distribution of a zero-one binary variable is called *Bernoulli*. The name comes from Jacob Bernoulli, the mathematician from the 1600s who first examined it. The Bernoulli distribution is one of those rare theoretical distributions that we observe over an over: all zero-one variables are distributed Bernoulli. (Note the use of words: if the distribution of a variable is Bernoulli, we say "it

is distributed Bernoulli"; we'll use this expression for other theoretical distributions, too.) Examples include whether a customer makes a purchase, whether the CEO of a firm is young, whether a portfolio produces a large negative loss, or whether the online price of a product is lower than its offline price. The Bernoulli distribution has one parameter: p, the probability of observing value one (instead of value zero).

With only two possible values, zero and one, the range of the Bernoulli distribution is zero to one, and its histogram consists of two bars: the frequency of observations with value zero, and the frequency of observations with value one. If, instead of frequency, the histogram shows the proportion of each value, the height of the bar at value one is equal to p, and the height of the bar at zero equals $1 - p$. The mean of a Bernoulli variable is simply p, the proportion of ones. (To verify this try $p = 0$ or $p = 1$ or $p = 0.5$.) Its variance is $p(1 - p)$ so its standard deviation is $\sqrt{p(1 - p)}$.

19.2 Binomial distribution

The *binomial distribution* is based on the Bernoulli distribution. A variable has a binomial distribution if it is the *sum of independent Bernoulli variables* with the same p parameter. Some actual variables that may have a binomial distribution include the number of car accidents, the number of times our portfolio experiences a large loss, or the number of times an expert correctly predicts if a new movie will be profitable. Binomial variables have two parameters: p, the probability of one for each Bernoulli variable and n, the number of Bernoulli variables that are added up.

The possible values of a binomial variable are zero, one, and all other integer numbers up to n. Its range is therefore 0 through n. The histogram of a binomial variable has $n+1$ bars (zero, one, through n) if not grouped in bins. The binomial distribution has one mode in the middle, and it is symmetric so its median, mean, and mode are the same. With large n the histogram of a binomial variable is bell shaped. The mean of a binomial variable is np, and its variance is $np(1 - p)$, so its standard deviation is $\sqrt{np(1 - p)}$.

The other distributions we cover in this section may approximate *quantitative variables* (as defined in Chapter 1). These theoretical distributions are for continuous variables that include fractions as well as irrational numbers such as π or the square root of two. In real data few variables can take on such values. Even variables that may in principle be continuous such as distance or time are almost always recorded with countable values such as integers or fractions rounded to a few decimal places. The continuous distributions are best seen as potential *approximations* of the distribution of real-life quantitative variables.

19.3 Uniform distribution

The *uniform distribution* characterizes continuous variables with values that are *equally likely* to occur within a range spanned by a minimum value and a maximum value. Examples of real life variables that may be approximately uniformly distributed are rare; the day of birth of people is an example. The uniform distribution is more often used as a benchmark to which other distributions may be compared. The uniform distribution has two parameters, the minimum value a and the maximum value b. The histogram of the uniform distribution is completely flat between a and b with zero frequency below a and above b. It is therefore symmetric. It has no mode: any value is just as frequent as any other value. The mean of a uniformly distributed variable is $\frac{a+b}{2}$, the variance is $\frac{(b-a)^2}{12}$, the standard deviation is $\sqrt{\frac{(b-a)^2}{12}}$.

19.4 Power-law distribution

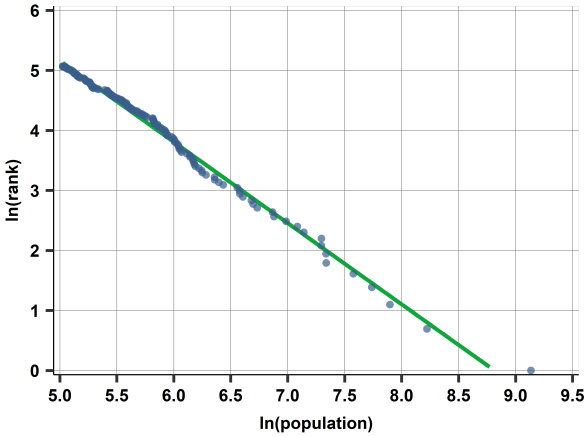
While the lognormal distribution may well approximate the distribution of variables with skewness and some extreme values, it is usually a bad approximation when those extreme values are very extreme. Distributions with very large extreme values may be better approximated by the the power-law distribution.

The *power-law distribution* is also known as the *scale-free distribution* or the *Pareto distribution*, and it is closely related to *Zipf's law*.

The power-law distribution is a very specific distribution with large extreme values. Its specificity is perhaps best captured by its scale invariance property (hence the name scale-free distribution). Let's take two values in the distribution and their ratio (say, 2:1). Let's compute the number of observations with one value (or within its neighborhood) relative to the number of observations with the other value (its neighborhood). For example, there may be 0.6 times as many cities with a population around 200 thousand than around 100 thousand in a country. If the variable has a power-law distribution, this proportion is the same for all value-pairs with the same ratio through the entire distribution. In the city population example, this means that there should be 0.6 times as many cities of with 600 thousand inhabitants than 300 thousand, 2 million than 1 million, etc. This is the scale invariance property.

A related, though less intuitive, property, of the power law distribution is that a scatterplot of the log of the rank of each observation against the log of the value of the variable yields a straight line. The log of the rank of the observation with the largest value is $ln(1) = 0$, the log of the rank of the second largest observation is $ln(2)$, etc. The figure below shows an example using the population of the 159 largest Japanese cities in 2015 (those larger than 150 thousand inhabitants). You will be asked to produce similar plots using other data as data exercises.

Figure 3.12: Log rank – log value plot. Size of Japanese cities



Note: *The natural log of the rank in the distribution (1st largest, 2nd largest, etc.) and the natural log of population*

Source: www.citypopulation.de.N=159

This latter property of the power-law distribution is the consequence of the fact that the probability of values in the close neighborhood of any value x is proportional to $x^{-\alpha}$: the value to a negative power, therefore the name power-law. Larger values are less likely than smaller values, and how much less

likely depends only on the parameter α. The larger the α, the smaller the likelihood of large values. This α is the only parameter of the power-law distribution.

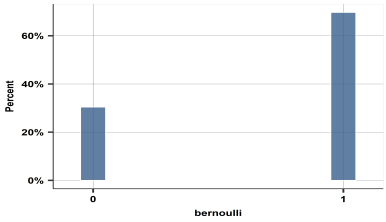
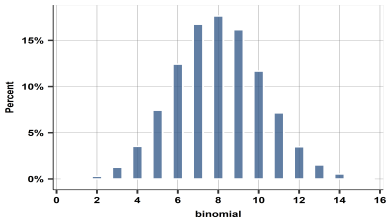
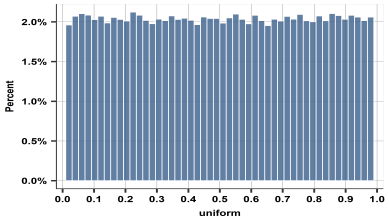
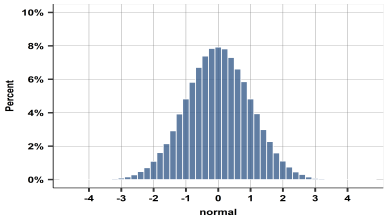
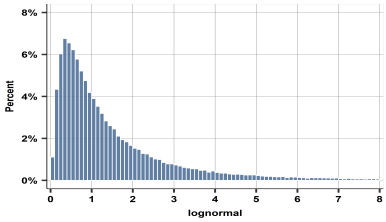
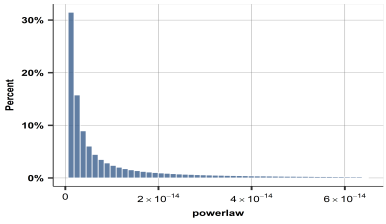
The range of the power-law distribution is zero to infinity (neither included). The power-law distribution is similar to the lognormal in that it is very asymmetric with a long right tail and thus allows for very extreme values. It is different from the lognormal by having a substantially larger likelihood of very extreme large values and thus a substantially fatter right tail.

The left part of the power-law distribution is also very different from the lognormal: its histogram continuously decreases in contrast with the up-then-down histogram of the lognormal. The mean and standard deviation of power-law distributions are functions of α, but they are not particularly informative statistics. In fact, with small α they may be infinitely large. Instead, more revealing statistics of power-law distributions compare the prevalence of certain values – we shall see examples of this later.

There are many variables in real life that are well approximated by a power-law distribution. To be more precise, this is usually true for the upper part of the distribution of such real-life variables. In other words, they are well approximated by power-law above certain values but not below. Examples include the population of cities, the size of firms, individuals' wealth, the magnitude of earthquakes, or the frequency of words in texts (ranking word frequency as "the" would be most frequent, followed by "be," etc.). These variables have extremely large values with low proportion but still a much higher proportion than what a lognormal distribution would imply.

An informative statistic of power-law distributions is the share of the values that is attributed to the top x percent of the distribution. One example is that in the U.S.A., the richest 1% own 40% of the total wealth. Another example is the so-called 80–20 "rule" that posits that 80% of total sales for a product will be concentrated among 20% of customers. Of course, 80–20 is not so much a rule as an empirical observation that holds in some cases but not necessarily in others. The point is that the fraction of wealth (or population in cities, workers in firms, energy released by earthquakes, etc.) in the top 20% (or top 5%, top 1% etc.) characterizes the power-law distribution.

Review Box 3.7 ***Important theoretical distributions***

Distribution	histogram	parameters	range	mean	std. deviation
Bernoulli		p	$[0,1]$	p	$\sqrt{p(1-p)}$
Binomial		p, n	$[0,n]$	np	$\sqrt{np(1-p)}$
Uniform		a, b	$[a,b]$	$\frac{a+b}{2}$	$\sqrt{\frac{(b-a)^2}{12}}$
Normal		μ, σ	$(-\infty,+\infty)$	μ	σ
Lognormal		μ, σ	$(0,+\infty)$	$e^{\mu+\frac{\sigma^2}{2}}$	$\sqrt{e^{\mu+\frac{\sigma^2}{2}}e^{\sigma^2-1}}$
Power-law		x_{min}, α	$(0,+\infty)$	$x_{min}\frac{\alpha-1}{\alpha-2}$	$\sqrt{x_{min}^2\frac{\alpha-1}{\alpha-3}}$

19.5 References and further reading

There are quite a few papers on the phenomenon of home advantage (Pollard 2006), and two excellent books on many aspects of soccer and data – Sally & Anderson (2013) and Kuper & Szymanski (2012).

On the idea of extreme values and their potentially large role, an interesting book is Nassim Nicholas Taleb: The Black Swan (Taleb 2007).

A great book on the emergence of statistics is (Salsburg 2001). It offers a non-technically demanding yet precise discussion of how key concepts such the distribution, random sampling, or correlation emerged.

An early apostle of good graphs is Ronald A Fisher. In his 1925 book (Fisher 1925) the first chapter after the introduction is called Diagrams.

Data visualization has now a robust literature as well as a large variety of online resources. In particular, our section on graphs has been shaped by the approach of Alberto Cairo, see for instance "The Functional Art: An introduction to information graphics and visualization" (Cairo 2012) "How Charts Lie" (Cairo 2019).

There are many great books on data visualization. Edward Tufte "Visual Explanations: Images and Quantities, Evidence and Narrative" (Tufte 1997) or any other book Tufte are classics. Two recent wonderful books, both with R code are Kieran Healy: "Data Visualization – A practical introduction" (Healy 2019) and Claus O. Wilke "Fundamentals of Data Visualization" (Wilke 2019). Another great resource is Chapter 03 in the book R for Data Science by Garrett Grolemund and Hadley Wickham (Grolemund & Wickham 2017).

Chapter 4

Comparison and correlation

Simple tools to uncover patterns of association by comparing values of y by values of x

Motivation

Are larger companies better managed? To answer this question, you downloaded data from the World Management Survey. How should you describe the relationship between firm size and the quality of management? In particular, can you describe that with the help of a single number, or an informative graph?

To answer the previous question, you have to use the variables in the data to measure the quality of management. In particular, you can use the quality of the many kinds of management practices separately, or you can use a summary measure. What are the advantages and disadvantages of each approach? If you want to use a summary measure, what's the best way to create it, and how should you interpret its magnitude?

Many questions that data analysis can answer are based on comparing values of one variable, y, against values of another variable, x, and often other variables. Such comparisons are the basis of uncovering the effect of x: if x affects y, the value of y should be different if we were to change the value of x. Uncovering differences in y for different values of one or more x variables is also essential for prediction: to arrive at a good guess of the value of y when we don't know it but we know the value of x.

We start by emphasizing that we need to measure both y and x well for meaningful comparisons. We introduce conditioning, which is the statistical term for uncovering information related to one variable as values of another variable change. We discuss conditional comparisons, or further conditioning, which takes values of other variables into account as well. We discuss conditional probabilities, conditional distributions, and conditional means. We introduce the related concepts of dependence, mean dependence, and correlation. Throughout the chapter, we discuss informative visualizations of the various kinds of comparisons.

In this chapter, we use the **Management quality and firm size: describing patterns of association** case study that uses the `wms-management-survey` data. The question is to what extent the quality

of management tends to be different when comparing larger firms to smaller firms. We illustrate conditional probabilities, conditional means, and various aspects of conditional distributions with this case study.

Learning outcomes. After working through this chapter, you should be able to

- identify the y and x variable (or x variables) in the data;
- understand the concepts of conditional probability, conditional distribution, conditional mean;
- create informative figures to visualize conditional means (bin scatter), other conditional statistics (box plots), and joint distributions of quantitative variables (scatterplots);
- understand the concepts of dependence, mean dependence, and correlation, and produce and interpret correlation coefficients.

1 The y and the x

Much of data analysis is built on comparing values of a y variable by values of an x variable, or more x variables. Such a comparison can uncover the **patterns of association** between the two variables: whether and how observations with particular values of one variable (x) tend have particular values of the other variable (y). The y and x notation for these variables is as common in data analysis as it is for the axes of the Cartesian coordinate system in calculus.

The role of y is different from the role of x. It's the values of y we are interested in, and we compare observations that are different in their x values. This asymmetry comes from the goal of our analysis.

On of the two most frequent goals of data analysis is predicting the value of a y variable with the help of other variables. The prediction itself takes place when we know the values of those other variables but not the y variable. To predict y based on the other variables we need rules that tell us what the predicted y value is as a function of the values of the other variables. Such a rule can be devised by analyzing data where we know the y values, too.

Those other variables are best thought of as many x variables, such as x_1, x_2, ... We use the same letter, x, for these variables and distinguish them by subscripts only because their role in the prediction is similar. For instance, to predict the price of AirBnB rentals, we need x variables that matter for that price, such as number of rooms and beds, location, etc. (we will investigate this in Chapter 1). To predict whether applicants for unsecured loans will repay their loans, we need variables that matter for that repayment probability, such as applicants' income, occupation, age, family status, etc.

The other most frequent goal of data analysis is to learn about the effect of a causal variable x on an outcome variable y. Here, we typically are interested in what the value of y would be if we could change x: how sales would change if we raised prices; whether the proportion of people getting sick in a group would decrease if they received vaccination; how the employment chances of unemployed people would increase if they participated in a training program.

Data analysis can help uncover such effects by examining data with both y and x and comparing values of y between observations with different values of x. Examples of observations with different x values indicate weeks with different prices, groups with different vaccination rates, or people who participated in a training program versus people who didn't. Often, when trying to uncover the effect of x on y, data analysts consider other variables, too. E.g., they want to compare weeks in which

the price of the product is different but the price of competing products is the same; groups with different vaccination rates but of the same size and living conditions; people who participate and don't participate in the training program but have the same level of skills and motivation. We often denote these other variables by another letter, z, to emphasize that their role is different from the role of the x variable.

In sum, deciding on what's y and what's x in the data is the first step before doing any meaningful analysis. In this chapter we discuss some general concepts of comparisons and introduce some simple and intuitive methods.

Review Box 4.1 *y and x in data analysis*

Most of data analysis is based on comparing values of a y variable by values of an x variable (or more variables).

- For prediction we want to know what value of y to expect for different values of various x variables, such as x_1, x_2, ...
- For causal analysis we want to know what value of y to expect if we changed the value of x, often comparing observations that are similar in other variables (z_1, z_2, ...).

2 A1 Case Study – Management quality and firm size: describing patterns of association

Question and data

In this case study we explore whether, and to what extent, larger firms are better managed.

Answering this question can help benchmarking management practices in a specific company. It can also help understand why some firms are better managed than others. Size itself may be a cause in itself as achieving better management may require fixed costs that are independent of firm size with benefits that are larger for larger firms. Whether firm size is an important determinant of better management can help in answering questions such as the potential benefits of company mergers (merged companies are larger than their component companies), or what kinds of firms are more likely to implement changes that require good management.

We use data from the World Management Survey to investigate our question. We introduced the survey in Chapter 1 Section 6. In this case study we analyze a cross-section of Mexican firms from the 2013 wave of the survey.

Mexico is a medium-income and medium-sized open economy with a substantial heterogeneity of its firms along many dimensions. Thus, Mexican firms provide a good example to study the quality of management. We excluded 33 firms with fewer than 100 employees and 2 firms with more than 5000 employees. There are 300 firms in the data, all surveyed in 2013. (Among the Data Exercises, you'll be invited to carry out an analogous analysis of firms from a different country.)

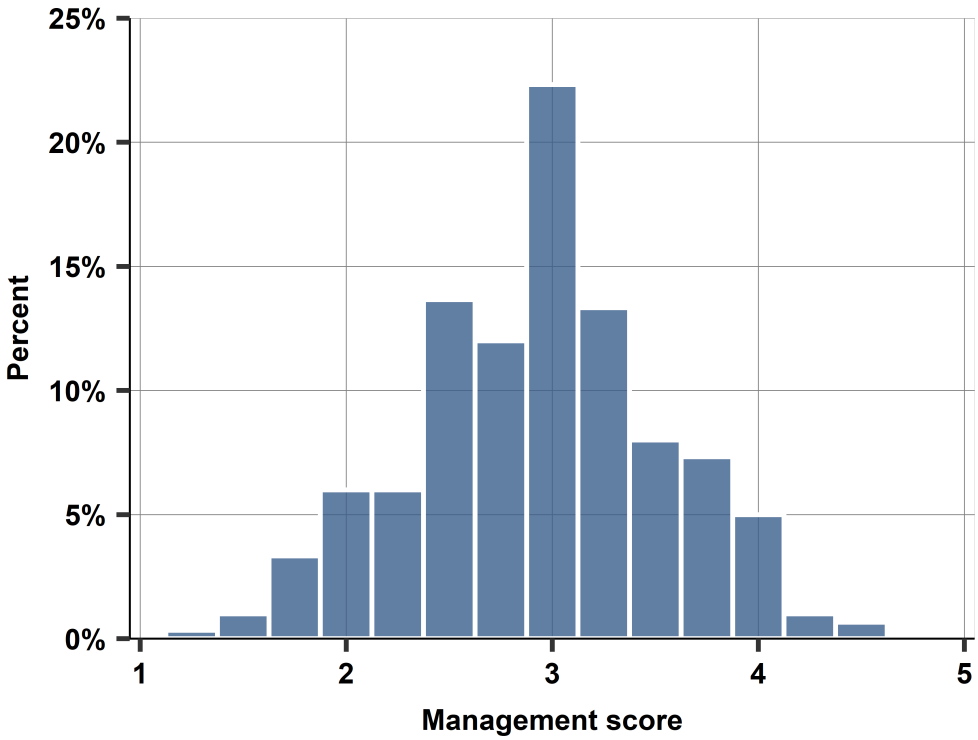
The y variable in this case study should by a measure of the quality of management. The x variable

should be a measure of firm size.

Recall that the main purpose of this data collection was to measure the quality of management in each firm. The survey included eighteen “score" variables. Each score is an assessment by the survey interviewers of management practices in a particular domain (tracking and reviewing performance or time horizon and of targets) measured on a scale of 1 (worst practice) to 5 (best practice).

Our measure of the quality of management is the simple average of these 18 scores. We – following the researchers who collected the data – call it “the" management score. By construction, the range of the management score is between 1 and 5 because that’s the range of all 18 items within the average. The mean is 2.9, the median is also 2.9, and the standard deviation is 0.6. The histogram (Figure 4.1 shows a more-or-less symmetric distribution with the vast majority of the firms having an average score between 2 and 4.

Figure 4.1: Distribution of the management score variable

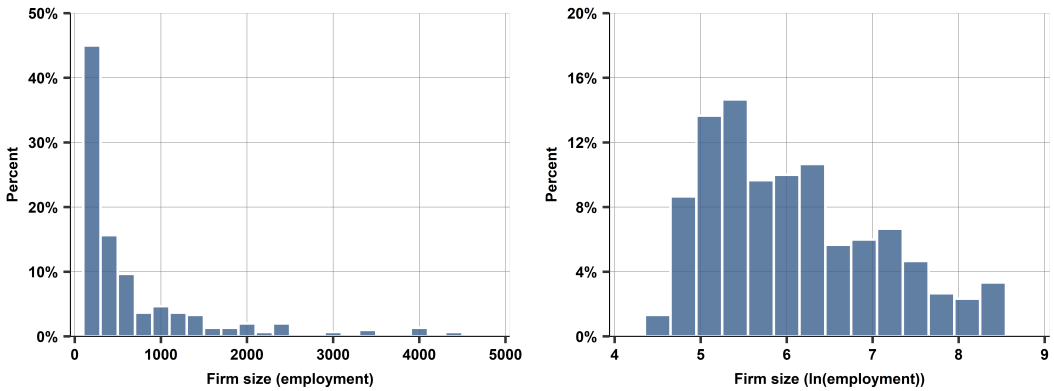


Note: *Histogram, bin width set to 0.25.*
Source: `management-survey` data. Mexico, all firms, N=300.

We measure firm size by employment: the number of workers employed by the firm. The range of employment in the Mexican data is 100 to 5000. The mean is 760 and the median is 350, signaling substantial skewness with a long right tail. The histogram shows this skewness. It also shows two extreme values: the largest four firms have 5000 employees, followed by two firms with about 4500 employees and three with 4000 employees. Recall from Chapter 3 Section 15 that distributions with long right tails may be well approximated by the lognormal distribution. To check this, we also show the histogram of the natural log of employment. That histogram is still skewed with a longer right tail, but it is substantially more symmetric. Also note that the extreme values are not so extreme

anymore with log employment. Thus, we can conclude that the distribution of employment in this data is skewed, it is closer to lognormal than normal, but even the lognormal is not the best approximation.

Figure 4.2: The distribution of employment



(a) Number of employees (b) Natural log of number of employees

Note: *Histograms*
Source: `wms-management-survey` data. Mexican sample, N=300.

3 Conditioning

The word statisticians use for comparison is **conditioning**. When we compare the values of y by the values of x, we condition y on x. y is also called the **outcome variable**; x is also called the **conditioning variable**. Most of the time, we will simply call them y and x.

When data analysts want to uncover values of y for observations that are different in x but similar in z, they do one one more step of conditioning: they compare y by x conditional on z. That is called **further conditioning** or a **conditional comparison**. Thus the word conditioning can be confusing if used without more context. It may mean a simple comparison – uncovering values of y for different values of x – or it may mean a conditional comparison – uncovering values of y for observations that are different in x but the same in z. Therefore we will try to be more specific and always add the necessary context.

Conditioning is an abstract concept. In practice we explore conditional probabilities, conditional means, and conditional distributions. The next sections discuss these in detail.

Review Box 4.2 ***Conditioning***

- Conditioning is the statistical term used for comparing values, or statistics, of y by values of x.
- Going one more step, we can further condition on z: compare y by values of x among observations with similar value for z.

4 Conditional probabilities

In Section 2 of the previous chapter, we introduced probability as the generalization of relative frequency. The probability of a value of a variable in a dataset is its relative frequency (percentage). In more abstract settings, the probability of an event is the likelihood that it occurs. In this section we discuss comparing the probability of events, or the probability of values of variables, by another event, or by values of another variable.

Conditional probability is the probability of one event if another event happens. The event of which the conditional probability is about is called the **conditional event**; the other event is called the **conditioning event**. Conditional probabilities are denoted as $P(event_1 \,|\, event_2)$: the probability of $event_1$ conditional on (or "given") $event_2$.

Note that the conditional probability is not symmetrical: $P(event_1 \,|\, event_2) \neq P(event_2 \,|\, event_1)$ in general. Pairs of probabilities of this sort are called **inverse probabilities**. Thus, inverse probabilities are not equal in general. In fact, inverse probabilities are related to each other in a somewhat complicated way. Understanding their relation can be important to understand a lot of real life problems, such as understanding the probability of having a condition after receiving a test result. It is also useful to understand the logic of generalizing results from the data that we'll discuss in Chapter 5. We discuss inverse probabilities and their relationship in more detail in the Under the hood section 14.

Joint probabilities are related to conditional probabilities. The **joint probability** of two events is the probability that both occur: $P(event_1 \,\&\, event_2)$. When two events are mutually exclusive, their joint probability is zero (the two never happen together).

Another probability related to two events denotes the likelihood that one event or the other happens. This is the sum of the two probabilities minus their joint probability: $P(event_1 \, OR \, event_2) = P(event_1) + P(event_2) - P(event_1 \,\&\, event_2)$. If the two events are mutually exclusive, we subtract zero from the sum of the two probabilities.

The conditional probability can be expressed as the corresponding joint probability divided by the probability of the conditioning event:

$$P(event_1 \,|\, event_2) = \frac{P(event_1 \,\&\, event_2)}{P(event_2)} \tag{4.1}$$

Two events are **independent** if the probability of one of the events is the same regardless of whether or not the other event occurs. In the language of conditional probabilities this means that the conditional probabilities are the same as the unconditional probabilities: $P(event_1 \,|\, event_2) = P(event_1)$ and $P(event_2 \,|\, event_1) = P(event_2)$.

Less intuitive, but also true is that the joint probability of independent events equals the product of their individual probabilities: $P(event_1 \,\&\, event_2) = P(event_1) \times P(event_2)$. (You can see this after plugging it into the formula that relates conditional and joint probabilities.)

In data, the events refer to values of variables. Most often, the conditional variable, y, is binary: $y = 0$ or $y = 1$. Then the conditional probability is the probability that $y = 1$ if x has some value: $P(y = 1|x = value)$. Since y is binary, we know the probability of $y = 0$ if we know the probability of $y = 1$, be it a conditional or unconditional probability; e.g., $P(y = 0|x = value) = 1 - P(y = 1|x = value)$. When x is binary too, there are two conditional probabilities:

$$P(y = 1|x = 1) \tag{4.2}$$
$$P(y = 1|x = 0) \tag{4.3}$$

With more values for any of the two variables (y, x), we have more numbers to compare: $P(y = value|x = value)$. With relatively few values, visualization often helps. There are many options for using standard graphs as well as creating individualized graphs. A good solution is the stacked bar chart. It presents the relative frequencies within bars that are on top of each other and thus always add up to a height of 100%. To visualize conditional probabilities if both y and x have few values, a good visualization is to show stacked bar charts of y for the values of x.

Review Box 4.3 ***Conditional probability***

- Conditional probability of an event: the probability of an event if another event (the conditioning event) happens.
  $P(event_1 \,|\, event_2)$
- Two events are independent if the probability of one of the events is the same regardless of whether the other event occurs or not:
  $P(event_1 \,|\, event_2) = P(event_1)$ and $P(event_2 \,|\, event_1) = P(event_2)$.
- The joint probability of two events is the probability that both happen at the same time.
  $P(event_1 \,\&\, event_2)$

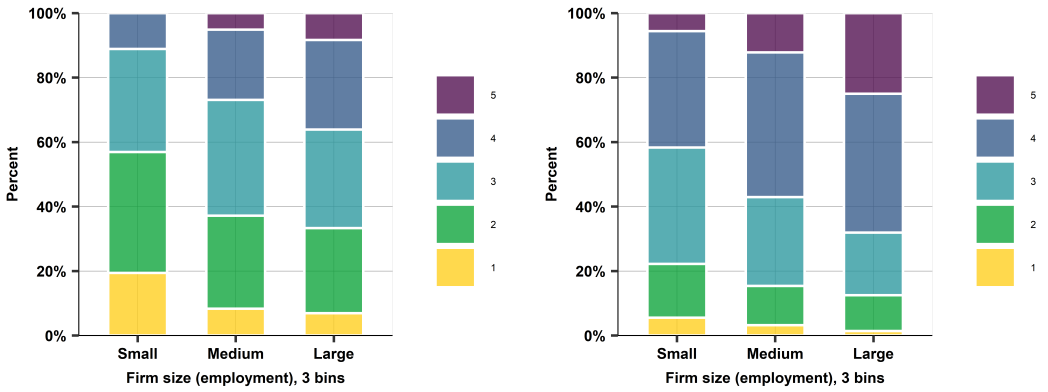
5 A2 Case Study – Management quality and firm size: describing patterns of association

Conditional probabilities

Both the management score and employment are quantitative variables with many values. They do not lend themselves to investigating conditional probabilities, at least not without transforming them. Thus, to illustrate conditional probabilities, we consider the individual score variables as y – recall that there are 18 of them, each with five potential values, 1 to 5. For x, we created a qualitative variables by creating three bins of employment: small, medium, large. These three bins are obviously arbitrary. We have chosen them to be bounded by round numbers: 100–199, 200–999, and 1000+ (with 72, 156, and 72 firms, respectively). Thus, for each score variable we have 15 conditional probabilities: the probability of each of the 5 values of y by each of the three values of x – e.g., $P(y = 1|x = small)$.

Listing 15 conditional probabilities in a table is not a great way to present them. But stacked bar charts are a great way to visualize them. The next figure shows two examples, one for lean management and one for performance tracking (each with values 1,2,..,5), each separately by small, medium, and large as firm size bins.

Figure 4.3: Quality of specific management practices by three bins of firm size: Conditional probabilities



(a) Lean management

(b) Performance tracking

Note: *Firm size as defined by number of employees. Stacked bar charts by firm size bins (bins are 100–199, 200–999, 1000+, with 72, 156, and 72 firms, respectively)*
Source: `wms-management-survey` data. Mexican sample, N=300.

For both lean management and performance tracking, the figures show the same pattern of association between the quality of management and firm size. Small firms are more likely to have low scores and less likely to have high scores than medium-sized firms, and, in turn, medium-sized firms are more likely to have low scores and less likely to have high scores than large firms. For lean management, scores 4 and 5 take up 11% of small firms, 27% of medium sized ones and 36% of large ones. For performance tracking, the corresponding percentages are 40%, 57% and 68%. These results suggest that larger firms are more likely to be better managed. You will be asked to produce similar stacked bar charts to confirm that the patterns with other management variables are the same.

6 Conditional distribution, conditional expectation

Just as all variables have a distribution (Chapter 3, Section 2), all y variables have a **conditional distribution** if conditioned on an x variable. This is a straightforward concept if the x variable has few values. The simplest case is a binary x when the conditional distribution of y is two distributions, one for each of the two x values.

Conditional distributions with few x values are best presented by visualization: for each x value we show stacked bar charts if y has few values or histograms if y is a quantitative variable with many values. The previous section covered conditional stacked bar charts. Here we discuss conditional histograms. Whatever we said about histograms in Chapter 3 Section 3 applies here: the look of the histograms are affected by our choice of bin size; density plots may be an alternative but are more sensitive to parameter settings. Our additional advice here is to make sure that the histograms are fully comparable across the x values: most importantly, they have the same bins of y, and they have the same scale on the vertical axis.

Comparing histograms can reveal qualitative differences in the distributions. We can tell which distribution tends to be to the left or right of the other, how their modes compare, which one looks more spread out, which one looks more skewed. To make quantitative comparisons, however, we need to

compare statistics that summarize the important features of distributions.

The most important conditional statistic is the **conditional mean**, also known as the **conditional expectation**, which shows the mean (average, expected value) of y for each value of x. The abstract formula for the conditional mean is

$$E[y|x] \tag{4.4}$$

From a mathematical point of view this is a function: if we feed in a number value for variable x, the conditional expectation gives us the number that is the mean of y for observations that have that particular x value. Note that while the overall mean of y, $E[y]$, is a single number in a data table, the conditional mean $E[y|x]$ varies in the data, because it can be different for different values of x.

Analogously, we can look at conditional medians, other conditional quantiles, conditional standard deviations, etc. Comparing box plots and violin plots are great ways to visualize conditional statistics of y when x has few values.

Review Box 4.4 ***Conditional distribution and conditional mean – quantitative y, few values of x***

- Conditional distribution of y by x is the distribution of y among observations with specific values of x.
- $E[y|x]$ denotes the conditional mean (conditional expectation) of y for various values of x.
- If x has few values, it is straightforward to visualize conditional distributions and calculate conditional means and other conditional statistics.

7 Conditional distribution, conditional expectation with quantitative x

When the x is quantitative with many values, things are more complicated. It is usually impossible or impractical to plot histograms or compute the conditional mean for each and every value of x. First, there are too many values of x and, typically, too few observations for each value. Second, even if we had enough observations, the resulting statistics or pictures would typically be too complex to make sense of.

We have two approaches to deal with these problems. The first approach circumvents the problem of too many x values by reducing them through creating bins. With few bins and thus many observations within each bin, we can calculate conditional means, plot histograms, box plots, etc., in ways that are easy to interpret. Creating bins from x is not only the simplest approach, but it often produces powerful results. Usually with just three bins of x – small, medium, large – we can capture the most important patterns of association between a quantitative y and a quantitative x.

Visualization of conditional means of y for bins of x is called a **bin scatter**. A bin scatter is a figure with the values of the binned x on the horizontal axis and the corresponding conditional mean values of y on the vertical axis. It is good practice to visualize bin scatters with meaningful x values for the x bins, such as their midpoint or the median value of observations within the bin.

The second approach keeps x as it is and uses other techniques for comparison. One technique is drawing a **scatterplot**, which is a visualization of the **joint distribution** of two variables. The joint distribution of two variables is the frequency of each value combination of the two variables. A

scatterplot is a two-dimensional graph with the x and y values on its two axes, and dots entered for each observation in the data with the corresponding x and y values. With small or moderately large datasets, scatterplots can reveal interesting things. With a very large number of observations, scatterplots can look like a big cloud and allow one to infer less information.

Starting with Chapter 7, we'll consider the most widely used method to uncover $E[y|x]$: regression analysis. This is a natural continuation of the methods considered in this chapter. But we don't discuss regression here as it requires more time and space. And, before moving on to regression analysis, we consider a few more topics within this chapter and the next two chapters.

Review Box 4.5 ***Joint and conditional distributions of two quantitative variables***

- The joint distribution of two variables shows the frequency of each value combination of the two variables.
- The scatterplot is a good way to visualize joint distributions.
- Conditional expectations may be computed and visualized in bins created from the conditioning variable.

8 A3 Case Study – Management quality and firm size: describing patterns of association

Conditional mean and joint distribution

Let's return to our case study on management quality and firm size, using our data of Mexican firms. y is the management score, and x is employment. Recall that this data contains firms with 100 to 5000 employees, and the distribution of employment is skewed with a long right tail (Figure 4.5a).

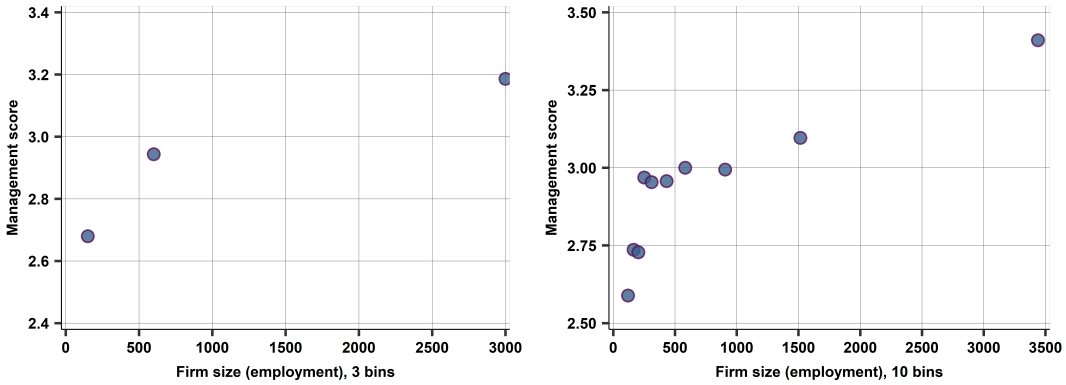
The next two charts (Figures 4.4a and 4.4b) show two bin scatters with mean y conditional on three x bins and ten x bins. We created the three bins in the same way as earlier, 100–199, 200–999 and 1000+ employees, with 72, 156, and 72 firms in the bins, respectively. Our approach to create the ten bins was different: instead of looking for round numbers to separate them, we simply split the data into ten equal-sized subsamples by the number of employees. Most statistical software can create such bins in a straightforward way. On both bin scatter graphs, we show the average management score as a point corresponding to the midpoint in the employment bin (e.g., 150 for the 100–199 bin).

The bin scatter with thee bins implies a clear positive association: larger firms are better managed, on average. The bin scatter with ten bins shows a somewhat more blurred picture, with a very similar mean in bins 4 through 8. But, overall, that picture too shows higher means in bins of larger firm size. The magnitude of the difference in mean management quality between large and small firms is moderate. Returning to the clearer picture with three bins, we see that the mean management score is 2.68 for small firms, 2.94 for medium sized ones, and it is 3.19 for large. When we compare these differences to the size difference between bins of employment, we see that the difference in mean management quality tends to be smaller when comparing bins of larger size, suggesting a positive but nonlinear, concave pattern of association (a positive concave function increases at a decreasing rate).

Finally, note that the bin scatters reflect the very skewed distribution of employment by having larger distances between bin midpoints at larger sizes. We could have presented the bin scatters with the

same bins but showing log employment on the x axis instead of employment; that would have shown a more even distribution of the bins and a more linear pattern. Such figures would show the exact same association: management tends to be of higher quality for larger firms, and that difference is smaller, in terms of absolute employment, at higher levels of employment. You'll be invited to do this as a data exercise.

Figure 4.4: Mean management quality score and firm size



(a) Three bins of employment

(b) Ten bins of employment

Note: *Bin scatters*
Source: `wms-management-survey` data, Mexican firms. N=300.

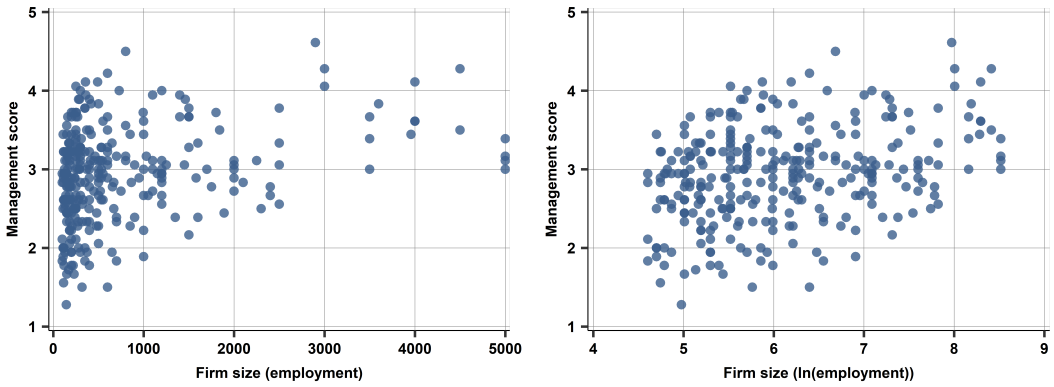
The bin scatters show a positive pattern of association on average. But does that mean that all larger firms are better managed in this data? Not at all. To appreciate the distribution of the management score around its conditional mean values, we look at the scatterplot.

On the left panel (Figure 4.5a) we see the consequences of the very skewed employment distribution: most observations are clustered in the leftmost part of the figure. We can see a positive slope on this graph among larger firms, but it's hard to see any pattern among smaller firms.

To make the patterns more visible we apply a transformation that we'll use quite often later on (see Chapter 8, Section 2). The idea here is that the distribution of employment is very skewed, closer to lognormal than normal, so we take the natural logarithm of employment. The right panel (4.5b) shows the same scatterplot with the natural log of employment on the x axis instead of employment itself. This amounts to stretching the employment differences between firms at lower levels of employment and compressing those differences at higher levels. (We'll spend a lot more time on what such a transformation does to variables and comparisons of variables later, in Chapter 8.) This scatterplot leads to a more spread out picture, reflecting the more symmetric distribution of the x variable. Here the positive association between mean management score and (log) employment is more visible.

In any case, we also see a lot of variation of the management score at every level of employment. Thus, there is a lot of spread of the management score among firms with the same size. As for other features of the distribution, the scatterplot doesn't show a clear pattern between employment (or log employment) and either spread or skewness of the distributions. But that may be because such features are not always easy to read off from a scatterplot.

Figure 4.5: The joint distribution of the management quality score and firm size



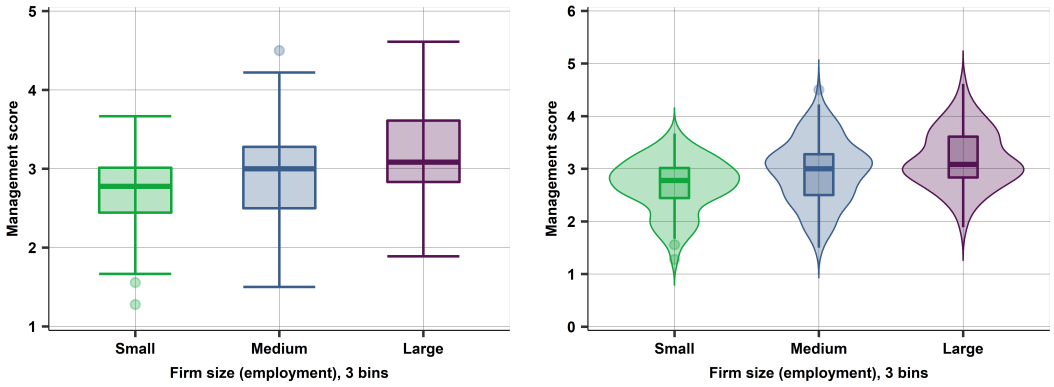
(a) By employment (b) By log employment

Note: *Scatterplots*
Source: `wms-management-survey` data, Mexican manufacturing firms, N=300.

To gain yet more insight into whether, and to what extent, the spread or skewness of the management score distribution differ at different levels of employment, we produced box plots and violin plots of the management score for three employment bins.

Both the box plots and the violin plots reveal that the median management score is higher in larger firms, reflecting the same positive association as the bin scatters and the scatterplot. That positive pattern is true when we compare almost any statistic of the management score: median, upper and lower quartiles, minimum and maximum. These figures also show that the spread of management score is somewhat smaller in smaller firms. That means that small firms are more similar to each other in their management scores, besides having lower scores on average. In contrast, larger firms differ more from each other in terms of their management score.

Figure 4.6: Conditional summary statistics of the management score by bins of firm size



(a) Box plots (b) Violin plots

Note: *Visuals of the conditional summary statistics: Box plot and violin plot.*
Source: `wms-management-survey` data. Mexican manufacturing firms, N=300.

9 Dependence, covariance, correlation

After discussing the conditional mean and the conditional distribution and their visualizations, let's introduce a few related concepts that are often used in data analysis.

Dependence of two variables, also called **statistical dependence** means that the conditional distributions of one variable (y) are not the same when conditional on different values of the other variable (x). There is something different in the distribution of y when compared across values of x. In contrast, **independence of variables** means that the distribution of one conditional on the other is the same, regardless of the value of the conditioning variable. These concepts may be viewed as generalizations of independent events (see Section 4 above).

Dependence of y and x may take many forms. For example y may be more spread out or more skewed for some x values. But the most important form of dependence is **mean-dependence**: the mean of y is different when the value of x is different. In other words, the conditional expectation $E[y|x]$ is not always the same but varies with the value of x.

The **covariance** and the **correlation coefficient** are measures of this mean dependence. To be a bit more precise, they measure mean dependence in an average sense. $E[y|x]$ may have ups and downs by the value of x, and the covariance and correlation coefficient are average measures of those ups and downs. When y and x are positively correlated, $E[y|x]$ tends to be higher when the value of x is higher. When y and x are negatively correlated, $E[y|x]$ tends to be lower when the value of x is higher. The two measures are very closely related: the correlation coefficient is the standardized version of the covariance.

The formula for the covariance between two variables x and y in a dataset with n observations is:

$$Cov[x, y] = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n} \tag{4.5}$$

The correlation coefficient divides this by the product of the two standard deviations:

$$Corr[x, y] = \frac{Cov[x, y]}{Std[x]Std[y]} \tag{4.6}$$

$$-1 \leq Corr[x, y] \leq 1 \tag{4.7}$$

The covariance may be any positive or negative number, while the correlation coefficient is bound to be between negative one and positive one. But their sign is always the same: the covariance is zero when the correlation coefficient is zero; the covariance is positive when the correlation coefficient is positive; the covariance is negative when the correlation coefficient is negative.

When the correlation coefficient is zero we say that y and x are **uncorrelated**. With positive correlation, y and x are **positively correlated**. With negative covariance and correlation we say that y and x are **negatively correlated**. The magnitude of the correlation coefficient shows the strength of the association: the larger the magnitude, the stronger the mean-dependence between the two variables. Data analysts tend to consider a correlation of 0.2 or less (in absolute value) weak, and a correlation above 0.7 in absolute value is usually considered strong.

If two variables are independent, they are also mean-independent and thus the conditional expectations are all the same: $E[y|x] = E[y]$ of any value of x. The covariance and the correlation coefficient are both zero in this case. In short, we say that independence implies mean independence and zero correlation. But the reverse is not true. We can have zero correlation but mean dependence (e.g., a symmetrical U-shaped conditional expectation has an average of zero), and we can have zero correlation and zero

mean dependence without complete independence (e.g., the spread of y may be different for different values of x).

Spending a little time with its formula can help understand how the covariance is an average measure of mean-dependence. The product within the sum in the numerator multiplies the deviation of x from its mean $(x_i - \bar{x})$ with the deviation of y from its mean $(y_i - \bar{y})$, for each observation i. The entire formula is the average of these products across all observations. If a positive deviation of x from its mean goes along with a positive deviation of y from its mean, the product is positive. Thus, the average of this product across all observations is positive. The more often a positive $x_i - \bar{x}$ goes together with a positive $y_i - \bar{y}$, the more positive is the covariance. Or, the larger are the positive deviations that go together the larger the covariance.

If, on the other hand, a positive $x_i - \bar{x}$ goes along with a negative $y_i - \bar{y}$, the product tends to be negative. Thus the average of this product is negative. The more often a positive $x_i - \bar{x}$ goes together with a negative $y_i - \bar{y}$, the more negative the covariance. Or, the more frequently we observe [positive deviation – negative deviation] pairs, the more negative the covariance.

Finally, if a positive $x_i - \bar{x}$ goes along with a positive $y_i - \bar{y}$ some of the time and a negative $y_i - \bar{y}$ at other times, and these two balance each other out, the positive values of the product and the negative values of the product cancel out. In this case the average is zero. Exact zero covariance rarely happens in real data because that would require an exact balancing out. The more balanced the [positive deviation – positive deviation] instances are with the [positive deviation – negative deviation] instances, the closer the covariance is to zero.

Also note that the formulas of the covariance or the correlation coefficient allow for all kinds of variables, including binary variables and ordered qualitative variables as well as quantitative variables. The covariance and the correlation coefficient will always be zero if the two variables are mean-independent, positive if positively mean-dependent, and negative if negatively mean-dependent. Thus, they give a quick and not completely meaningless picture about mean-dependence among binary and ordered qualitative variables. However, they are more appropriate measures for quantitative variables. That's because the differences $y_i - \bar{y}$ and $x_i - \bar{x}$ make more sense when y and x are quantitative variables. For that reason, data analysts use other correlation-like measures for qualitative variables, but those measures are beyond the scope of our textbook.

Review Box 4.6 ***Covariance and correlation***

- The covariance measures the mean-dependence of two variables.
  $Cov[x, y] = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n}$
- The correlation coefficient is a standardized version of the covariance and ranges between -1 and 1.
  $Corr[x, y] = \frac{Cov[x,y]}{Std[x]Std[y]}$
- The covariance and the correlation are both zero if the variables are independent, positive if the variables are positively dependent, and negative if they are negatively dependent.

10 From latent variables to observed variables

Before closing the chapter, let's discuss two more topics briefly. The first one is the concept of latent variables.

Often, the y and/or x variables we have in of our question are abstract concepts: quality of management of a firm, skills of an employee, risk tolerance of an investor, health of a person, wealth of a country. Typically, such variables are not parts of an actual dataset, and they can't be because they are too abstract. Such variables are called **latent variables**.

Data analysis can help answer questions involving latent variables only by substituting observed variables for them. Those observed variables are called **proxy variables**, where "proxy" means substitute, and has common roots with "approximate" (the word proxy is used as noun, adjective, and verb). The quality of management may be proxied by answers to survey questions on management practices, as in our case study; employee skills may be proxied by qualifications or measures of past performance; etc.

The most important thing to keep in mind here is that data analysis compares values of measured variables. Even if those variables are supposed to measure abstract concepts, it's never the abstract concepts themselves that we have in our data. Thus we can never examine things such as skills or attitudes or health; instead, we examine proxies such as measures of performance, answers to survey questions, or results of doctors' diagnoses. This is simply re-iterating the point we made earlier in Chapter 1 Section 4: the content of a variable is determined by how it is measured – not by what name somebody attached to it.

A specific issue arises when our data contains not one but more variables that could serve as proxies to the latent variable we want to examine. The question here is how to *combine multiple observed variables*. Data analysts use one of three main approaches

- Use one of the observed variables
- Take the average (or sum) of the observed variables
- Use principal component analysis (PCA) to combine the observed variables

Using one measured variable and excluding the rest has the advantage of easy interpretation. It has the disadvantage of discarding potentially useful information contained in the other measured variables.

Taking the average of all measured variables makes use of all information in a simple way. When all of those variables are measured using the same scale, this approach yields a combined measure with a natural interpretation. When the variables are measured at different scales, we need to to bring the observed variables to the same scale. Usually, we do that by standardizing them: subtracting the mean and dividing by the standard deviation (see Chapter 3 Section 9). This standardized measure is also called a **z-score**.

By taking a simple average of all these variables we give them equal weight. This may be a disadvantage if some of the variables are better measures of the latent variable than others. The third approach remedies that problem. **Principal component analysis** (**PCA**) is a method to give higher weights to the observable variables that are better measures. PCA finds those weights by examining how strongly they would be related with the weighted average. The logic is an iterative process: create an average, examine how each variable is related to it by looking at their correlations, give higher weights to those

with stronger correlation, start over. The actual technique takes an ingenious approach to do the whole thing in one step.

Of the three approaches, we recommend the second one: taking a simple average of the observed variables after making sure that they are measured at the same scale. This is the simplest way to combine all variables in a meaningful way. In principle, PCA produces a better combined measure, but it is more complicated to produce and harder to present to non-expert audiences. Moreover, it often gives similar results to a simple average. Thus we recommend that PCA is used as a robustness check, if at all. If the results of our analysis are very different with a simple average and with a PCA measure, some of our observed variables are very differently related to the average measure than others. It is good practice then to go back and understand the content of those variables and, perhaps, discard some of them from the analysis.

Review Box 4.7 ***Latent and proxy variables***

- Latent variables are abstract concepts that are not actual variables in the data.
- Proxy variables (proxies) are variables in the data that measure latent variables.
- When more proxy variables are available for a single latent variables it is good practice to take their average, after making sure that they are measured on the same scale (i.e., by standardizing them).

11 A4 Case Study – Management quality and firm size: describing patterns of association

Correlation and latent variable

The covariance between firm size and the management score in the Mexican sample we use is 177. The standard deviation of firm size is 977, the standard deviation of management score is 0.6. the correlation coefficient is 0.30 $(177/(977 * 0.6) = 0.30)$.

This result a positive association: firms with more employees tend to have a higher management score. The magnitude of the correlation is moderate, presumably because many other things matter for the quality of management besides the size of a firm.

Table 4.1 shows the correlation coefficient in six broad categories of industrial classification (plus one “other" category with the industries with very few forms, combined).

Table 4.1: Management score and employment: correlation and average management score by industry.

Industry	Management–firm size correlation	Observations
Auto	0.50	26
Chemicals	0.05	69
Electronics	0.33	24
Food, drinks, tobacco	0.05	34
Materials, metals	0.32	50
Textile, apparel	0.29	43
Wood, furniture, paper	0.28	29
Other	0.44	25
All	0.30	300

Source: `wms-management-survey` data. Mexican manufacturing firms, N=300.

The table reveals that the management quality–firm size correlation varies considerably across industries. The correlation is strongest in the auto industry. At the same time, we see hardly any correlation among firms in the chemicals and food industries.

Before concluding our case study, note that it illustrates the measurement issues related to latent variables, too. From a conceptual point of view, the y variable in our case study is management quality, a latent variable. We have 18 measures for this latent variable in the data; those are the 18 score variables on the quality of various aspects of management. Each of these 18 variables is measured by the survey (as we discussed in Chapter 1, Section 6), and each is measured on the same, 1-to-5 scale.

For the measure of the overall quality of management, we used two of the three strategies we recommended in Section 10. To illustrate conditional probabilities, visualized by the stacked bar charts in Figures 4.3a and 4.3b, we used two of the 18 score variables. Each one is an imperfect measure of the overall quality of management, but each has a clear interpretation: the rating of the particular aspect of management quality by the interviewer. For most of the case study, we used the average score: the simple average of the 18 scores. We could use this simple average because each of the 18 variables aimed to measure the same thing, management quality, and each was measured on the same scale (1 to 5).

As a data exercise you are invited to try the third option we recommended in Section 10, and create a principal component from the 18 scores instead of their simple average. When analyzing the relationship between firm size and management quality, using this principal component measure turns out to give very similar results to what we have uncovered using the average score measure.

This concludes our case study. What did we learn from it about the association between firm size and management quality? We found that, among Mexican manufacturing firms, larger firms tend to be better managed. Large firms (with 1000-5000 employees) have an average score of 3.19, compared to 2.94 for medium sized firms (with 200-999 employees) and 2.68 for small ones (with 100-199 employees).

We also found that the correlation, while positive, is not very strong, perhaps because other things matter for the quality of management besides firm size. When disaggregating the results into smaller industry groups, we found that the strength of the management–size correlation differs in some industries from the rest, but we haven't seen any clear pattern that would tell us why. Finally, we have seen that management quality is not only better, on average, among larger firms, but it is also somewhat more spread among larger firms.

These results inform the business or policy questions we may be interested in. When considering the management practices of a specific firm, we should have firms of similar size as a benchmark. And, better management of a larger firm may be a potential benefit of increased firm size – e.g., through a merger between companies.

As for the methods discussed in this chapter, this case study illustrated what we can do to uncover patterns of associations and conditional distributions when both y and x are quantitative variables. We have seen that creating bins from x can lead to informative visualizations, such as a bin scatter or box plots of y by bins of x. Three bins (small, medium, large) appeared a good choice in our case. For example, the bin scatter with ten bins did not give much more information than the bin scatter with three bins. We have also seen that the correlation coefficient is a useful statistic to summarize mean dependence between y and x, and it allows us to dig a little deeper by showing whether and how the correlation differs across groups by a third variable (here industry). Finally, we have seen that, with rich enough data, we can use an average score variable calculated from many (here 18) variables to measure a latent variable, management quality in our case study.

12 Sources of variation in x

Our final section in this chapter is a note on variation in x, the variable (or variables) we condition on to make comparisons in y. The first thing to note is that we need variation in x, and the more variation we have the better in general. In data with no variation in x, all observations have the same values and it's impossible to make comparisons. This may sound trivial, but it's essential to keep in mind. Similarly, the more variation in x, the better the chances for comparison.

For example, when data analysts want to uncover the effect of price changes on sales, they need many observations with different price values. If prices don't change at all, there is no way to learn how they may affect sales. If prices change very rarely or the changes are negligible in magnitude, there isn't much room for comparison and thus there isn't much to learn.

The second question is where that variation in x comes from. As we shall see in subsequent chapters (e.g., Chapters 19 through 24), data analysts need to understand the **sources of variation** in x. This is a somewhat fancy way of saying that data analysts should have a good understanding of why values of x may differ across observations. From this perspective, there are two main types of data: experimental data and observational data.

In **experimental data**, the value of x differs across observations because somebody made them different. In a medical experiment assessing the effects of a drug, some patients receive the drug while others receive a placebo, and who receives what is determined by a rule designed by the experimenter, such as a coin flip or a computer generated sequence of numbers. Here x is the binary variable indicating whether the person received the drug, instead of the placebo. Such variation is called **controlled variation**. Uncovering the effect an experiment amounts to comparing y (such as whether a subject recovers from the illness or not) across the various values of x (whether a subject received the drug or not).

In contrast, in **observational data**, no variable is fully controlled by an experimenter or any other person. Most data used in business, economics, and policy analysis are observational. Typical variables in such data are the results of the decisions of many people with diverging goals and circumstances, such as customers, managers of firms, administrators in a government, or members of the board of the monetary authority. Thus, typically, variation in these variables has multiple sources.

Whether the variation in conditioning variables is controlled (experimental) or not (observational) is extremely important for causal analysis. Learning the effects of a variable x is a lot easier when we have data from an experiment, in which variation in x is controlled. With observational variation, of the many other things that affect an intervention variable, some may affect the outcome variable in a direct way, too. Disentangling those effects requires data analysts to further condition on many variables at the same time, using methods that we'll cover later in this textbook. Even with the best methods, conditioning on variables is possible only if those variables are measured in the data, which typically is an issue. We'll return to these questions in Chapter 10 and, in more detail, in Chapters 19 through 24.

For example, the price of a product (x) sold by a retailer may vary across time. In a sense that variation has one source, the decisions of people in charge at the retail company. But that decision, in turn, is likely affected by many things, including costs, predicted customer demand, and the pricing decisions of competitors, all of which may change through time and thus lead to variation in prices. Here a data analyst may want to uncover what would happen if the retailer increased its price (x) on sales (y), using observational data. That requires conditioning on price: looking at differences in sales across observations with different prices. But the results of this comparison won't tell us what would happen if the retailer increased the price. That's because the question is about changing x as an autonomous decision, whereas, in the data, x tends to change together with some other things. The data analyst then may go on to try to further condition on other those other things that are sources of variation, such as the price charged by the competitors, airtime advertising, or seasonal variation in demand. If lucky, the data analyst may be ale to measure all those variables. But that's a tall order in most cases. The power of experimental data is that there is no need to measure anything else.

The other frequent goal of data analysis, making predictions about y, poses somewhat different requirements for variation in x. Understanding the sources of variation in x or, more realistically, the many x_1, x_2,..., variables, is still useful, although not in the way it is in causal analysis. Here the main question is stability: whether the patterns of association between y and all those x variables are the same in our data as in the situation for which we make the prediction. Controlled variation in x helps only in the rare case when x would also be controlled in the situation we care about. But uncovering cause and effect relationships can be helpful in prediction in general, as such relationships tend to be stable. We shall discuss these issues in more detail in Part 3, from Chapter 13 through Chapter 18.

This chapter introduced some fundamental concepts and methods of conditioning y on x, the statistical concept of comparing values of y by values of x (or more x variables). We'll return to conditioning y on x in Chapter 7 where we introduce regression analysis. Before doing so, we discuss some general principles and methods in the next chapter that help draw conclusions from our data about the situation we are really interested in.

13 Summary and practice

13.1 Main takeaways

- Data analysis answers most questions by comparing values of y by values of x.
  - Be explicit about what y and x are in your data and how they are related to the question of your analysis.
  - $E[y|x]$ is mean y conditional on x.
  - Often more variables are used: x_1, x_2, ... for prediction, x and further conditioning on z_1,

$z_2, \ldots$ for causal analysis.

13.2 Practice questions

1. Give an example with two independent events. Can independent events happen at the same time?
2. Give an example of two mutually exclusive events. Can mutually exclusive events happen at the same time?
3. What's the conditional probability of an event? Give an example.
4. What's the conditional mean of a variable? Give an example.
5. How is the correlation coefficient related to the covariance? What is the sign of each when two variables are negatively dependent, positively dependent, or independent?
6. Describe in words what it means that hotel prices and distance to the city center are negatively correlated.
7. Why do we need variation in the conditioning variable?
8. What's the difference between the sources of variation in x in experimental data and observational data?
9. What's the joint distribution of two variables, and how can we visualize it?
10. What's a scatterplot? How does it look like for two quantitative variables, each of which can be positive only, if the two variables are positively correlated?
11. What's a bin scatter, and what is it used for?
12. What's a latent variable, and how can we use latent variables in data analysis?
13. List two ways to combine multiple measures of the same latent variable in your data for further analysis, and list an advantage and a disadvantage of each way.
14. You want to know if working on case studies in groups or working on them independently is a better way to learn coding in R. What would be your y and x variable here and how would you measure them?
15. Can you tell from the shape of a bin scatter if y and x are positively correlated? Can you tell from it how strong their correlation is?

13.3 Data exercises

Easier and/or shorter exercises are denoted with [*] Harder and/or longer exercises are denoted with [**]

1. Are central hotels better? To answer this, using the `hotels-vienna` data (as discussed in Chapter 3, section 4), create two categories by the distance from center: close and far (by picking a cutoff of your choice). Show summary statistics, compare star ratings and prices for close and far hotels. Create a stacked bar chart as well as a box plot and a violin plot. Summarize your findings. [*]

2. Using the `wms-management-survey` data, pick a country different from Mexico, reproduce all figures and tables of our case study, and compare your results to what we found for Mexico. [*]

3. Use the `wms-management-survey` data from a country of your choice, and pick 2 of the 18 management scores. Produce bin scatters and scatterplots, stacked bar charts, and calculate conditional statistics to uncover the patterns of their association with employment. Summarize what you find, and comment on which visualization you find the most useful. [*]

4. Use the `wms-management-survey` data from a country of your choice, and produce a principal component using all 18 items to form an alternative management score variable. Use this principal component and the simple average management score to produce bin scatters, scatterplots, and and calculate conditional statistics to uncover the patterns of their association with employment. Compare your results and comment on which y measure you would use in presenting them. [*]

5. Use the `football` data and pick a season. Create three groups of teams, based on their performance in the previous season (new teams come from the lower division, and you may put them in the lowest bin). Examine the extent of home team advantage (as in Chapter 3, section 12) by comparing it across these three groups of teams. Produce bin scatters and scatterplots, and calculate conditional statistics. Discuss what you find, and comment on which visualization you find the most useful. [**]

14 Under the hood: Inverse conditional probabilities, Bayes rule

As we introduced in Section 4, inverse conditional probabilities are two conditional probabilities, in which the role of the conditioning event and the conditional event are switched: $P(event_1 \,|\, event_2)$ and $P(event_2 \,|\, event_1)$. In this section we discuss their relationship to each other.

Suppose that we want to know if an athlete used illegal substance (doping). For this case we collect lab tests. Does the positive result of the test indicate that there is illegal substance in the body of the athlete? We we are interested in is whether the athlete has doped given the positive test result. But tests are imperfect so a test result will not tell doping for sure. Instead, what we may hope for is a probability: the likelihood that someone doped, or not doped, given the result of the test. These are conditional probabilities: $P(doped \,|\, positive)$ and $P(not\, doped \,|\, positive)$. (Knowing one of these two gives the other one as the two sum up to one.) Imperfection of tests mean that they may give positive results even if athletes don't dope: $P(positive \,|\, not\, doped) > 0$.

Tests that are used in real life are usually validated so the level of their imperfection is known. Thus we typically know $P(positive \,|\, not\, doped)$. What we are interested is the inverse probability: $P(not\, doped \,|\, positive)$ The relation of inverse conditional probabilities tells us how the imperfect nature of a doping test determines how confident we can be concluding that an athlete doped if the result of the test is positive.

The two inverse conditional probabilities are related although their relation might seem complicated. We can derive one from the other using the formula that links conditional probabilities and joint probabilities as both are related to the same joint probability, the probability of both $event_1$ and $event_2$ occurring. The relation is called Bayes' rule after the reverend Bayes who was the first to express this formula, in the 17^{th} century.

$$P(event_2 \,|\, event_1) = \frac{P(event_1 \,|\, event_2)P(event_2)}{P(event_1)} \tag{4.8}$$

Which, in turn, can be rewritten as

$$P(event_2 \,|\, event_1) = \frac{P(event_1 \,|\, event_2)P(event_2)}{P(event_1 \,|\, event_2)P(event_2) + P(event_1 \,|\, event_2) * P(event_2)} \tag{4.9}$$

The most important message of this formula is that inverse conditional probabilities are not the same in general. The formula is also complicated. Instead of memorizing the formula we suggest using a different approach. This approach amounts to thinking in terms frequencies and proportions in place of abstract probabilities.

Consider our doping example: what's the likelihood that an athlete is a doper (or a non-doper) if they receive a positive test result? Start with assuming that a fifth of the athletes dope. Out of, say, 1000 athletes that means 200 doping and 800 not doping. Consider a test that is imperfect but not bad: it always shows a positive result when the athlete dopes, but it also shows positive results 10 percent of the times if an athlete does not dope. The former means that the test will be positive for all 200 dopers.

The latter means that the test will also be positive for 10% of the non-dopers, which would be 80 out of the 800. In total we have 280 positive tests out of 1000. Of these 280 positives 200 are dopers and 80 non-dopers. We don't know which 200 is a doper and which 80 is a non-doper, but we can use these figures to calculate probabilities. The probability that an athlete is a doper if their test is positive is $200/280 = 71\%$ approximately. The probability that an athlete is not a doper if their test is positive is $40/240 = 29\%$ approximately. This may look surprising: a relatively small imperfection (10% of positive results for non-dopers) results in a much larger drop in our confidence: the chance that a positive tester did not dope in fact is 29%. (Working through the formulae gives the same result.) The inverse conditional probability is larger because we started with the assumption that only 20% of athletes dope.

This example highlights several important things. First, working through the frequencies is not super-easy, but it is doable. Second, however we carry out the calculation we need the probability that the test comes out positive for each of the groups, dopers and non-dopers (these are $P(event_1 \,|\, event_2)$ and $P(event_1 \,|\, event)$).

Third, we need the overall fraction of athletes that dope. That is $P(event_2)$ in the formulae above. This proportion is sometimes called the *base rate*. Without the base rate we can't compute the inverse probability. Unfortunately, we may not know the base rate. A good practice in such cases is to use several plausible values and give a range of inverse conditional probabilities.

In our example, we assumed a base rate of 20%: 200 out of 1000 athletes use doping. If, instead we assumed a 5% base rate (50 out of 1000 doped) a test with a 10% positive rate for non-dopers (and 100% positive rate for dopers) would result in 50 dopers among the positively tested and 95 non-dopers (10% of 950). Thus the likelihood of doping conditional on a positive test is only 0.34 $(= 50/(50+95)$. On the other hand, if we assumed a 50% base rate (500 out of 1000 doped) the same test would result in 500 dopers among the positively tested and 50 non-dopers. In this case the likelihood of doping conditional on a positive test is a high 0.91 $(= 500/(500+50)$. The base rate has a substantial effect on the result. With few dopers an imperfect test would give more misleading results than the same imperfect test with many dopers.

15 References and further reading

Regarding the world management survey, you may find plenty of reading at the survey website for academic material at https://worldmanagementsurvey.org/academic-research/manufacturing-2/ - with links to papers. For a business perspective, you may have a look at this Harvard Business Review article Bloom et al. (2017). For a more detailed review of the project, consider reading Bloom et al. (2014).

Chapter 5

Generalizing from data

How to generalize the results of analysis of your data and how to assess the validity and the limits of such generalizations

Motivation

How likely is it that you will experience a large loss on your investment portfolio of company stocks? To answer this, you have collected data on past returns of your portfolio and calculated the frequency of large losses. Based on this frequency, how can you tell what likelihood to expect in the coming calendar year? And can you quantify the uncertainty about that expectation in a meaningful way?

How is the quality of management among manufacturing firms related to how much they export to foreign markets? You have data on a sample of manufacturing firms, and you can calculate various measures of association in your data. But you are interested in that association among all firms, not just this sample, and you are interested in how that may look like in the near future. Can you generalize the findings in the sample to all firms and to the future? And can you assess the quality, and uncertainty, of such generalization?

Most often, we analyze the data we have to help a decision in a situation that's not included in the data. To do so, we need to generalize the results of our analysis from the data we have to the situation we care about. The most important questions are whether we can generalize the results, and whether we can quantify the additional uncertainty brought about by such a generalization.

We start this chapter by discussing the two steps of the process of generalization: generalizing from the data to a general pattern it represents, such as a population, and assessing how the situation we care about relates to the general pattern our data represents. The first task is statistical inference. We introduce the conceptual framework of repeated samples and estimation. We introduce the standard error and the confidence interval that quantify the uncertainty of this step of generalization. We introduce two methods to estimate the standard error, the bootstrap and the standard error formula. We then discuss external validity of the results of an analysis, which is the second step of generalization:

from the general pattern our data represents to the general pattern behind the situation we care about. While there are no readily available methods to quantify the uncertainty this step brings to the results, we discuss how we can think about it and how we can use the results of additional data analysis to assess it.

The case study **What likelihood of loss to expect on a stock portfolio?** examines the probability of large negative returns on financial assets and how to infer that probability in future investment situations using the data at hand. It uses the `stock-market` data. This case study illustrates not only how statistical inference works but also its limitations.

Learning outcomes. After working through this chapter, you should be able to

- think about generalization beyond the actual data;
- understand the concept and logic of statistical inference and external validity;
- understand the concepts of repeated samples and bootstrap;
- compute and use standard errors to create confidence intervals and use those for statistical inference;
- understand whether and how additional data analysis can help assess external validity.

1 When to generalize and to what?

Sometimes we analyze our data with the goal of learning about patterns in that data itself, among the observations it contains. When we search for a good deal among offers of hotels or used cars, all we care about are the hotels, or cars, in our data. In such cases there is no need to generalize our findings to any other situation. More often, though, we analyze a dataset in order to learn about patterns that are likely to be true in other situations.

Examples for the need to generalize are plentiful. We may analyze data on used cars to determine what price we may get for a similar car that we own. Or, we may use data on a sample of working people from a country in certain occupations to assess what wage current students studying for such occupations may expect. Or, we may analyze data on the history of returns on an investment portfolio to assess the probability of losses of a certain magnitude in the future.

In all of these latter cases we want to infer something from the data at hand for a situation we care about, but which is not contained in the data. The data on used cars does not include our car. The sample of working people includes only a subset of all working people in the country, and it is about the past not about the future we care about. The data on the history of portfolio returns covers the past, whereas we are interested in the likelihood of losses in the future.

This chapter focuses on how to generalize the results from a limited dataset to other situations. The act of generalization is called **inference**: we infer something from our data about a more general phenomenon because we want to use that knowledge in some other situation. Inference is best broken down into two steps: generalizing from the data to the general pattern it represents, and generalizing from that general pattern to the general pattern that is behind the situation we truly care about. These may sound a little cryptic at first, so let's discuss them in detail.

The first step is **statistical inference**, which aims at generalizing to the situation that our data represents, using statistical methods. Our data may represent a population if it's a representative sample

(see Chapter 1, Section 14). More generally, it may represent a general pattern. To be able to generalize to it, that general pattern needs to exist, and it needs to be stable over time and space.

The simplest case here is a random sample that represents a well-defined **population**. As we defined it earlier (see Chapter 1, Section 14), as sample is representative of a population if the distribution of all variables is very similar in both. Having learned more about distributions of many variables in Chapter 4, we can be more specific now: the joint distribution of all variables needs to be similar. that includes conditional distributions and correlations, too. For example, a dataset on market analysts in your country from last year with variables age, gender, and salary represents the population of all market analysts in your country last year if the distribution of gender, age, and salary is very similar in the data and the population, including conditional statistics such as average salary by age and gender.

The **general pattern** is a more abstract concept, but, often it can also be something rather specific. Examples include the probability of a large loss on our portfolio, or how much more likely better managed firms in medium-sized medium-income countries are to export to foreign markets. For our data to represent this general pattern, it is not necessary that all variables have the same distribution. Instead, we require that the specific pattern (a likelihood, a conditional probability, a conditional mean) is the same. But we want that similarity not purely for a well-defined population but reflecting something more general that's behind our data. Examples include what determined the history of asset returns, or all the things that influenced the relationship between management quality and exporting in medium-sized medium-income economies during the time our data was collected.

Generalizing to a general pattern that our data represents can make sense even when our data covers an entire population. For example, our data may include all companies in India that were in business in 2016. Thus, the management quality – exporting association in the data is what it is in the population of firms in 2016. However, we rarely carry out data analysis to learn about a specific population in a specific time in the past. Instead, we want to learn about something more general: in India at other times, or in other countries like India.

When making statistical inference, we can use readily available tools to learn about the general pattern our data represents. Those tools can quantify the uncertainty about this step of generalization. We'll introduce the toolbox of statistical inference in the subsequent sections of this chapter.

The second step is assessing the **external validity** of the results. External validity is about whether the general pattern our data represents is the same as the general pattern that would be relevant for the situation we truly care about. Whatever lies behind gender differences in earnings of market analysts in the population that our data represents may or may not be similar to whatever will determine gender differences five years from now in our country, or in a different country. Whatever made large losses likely during the time that our data is from may or may not be similar to what would make large losses possible in the future we care about.

If the general pattern behind the situation we care about and the data we have is the same, what we uncover from our data has high external validity for the situation we care about. If, on the other hand, the general pattern behind the situation we care about is different from the general pattern that our data represents, whatever we uncover from our data would have low external validity. External validity is best thought of as a continuum, from very high (our data represents the same general pattern we are interested in) through medium (the two general patterns are similar but not the same) to very low (the two general patterns are very different). Assessing external validity requires thinking and knowledge about the situations; statistical methods can help in indirect ways only. We shall return to external validity and challenges to it later in this chapter.

In contrast with statistical inference, we don't have statistical methods that give direct answers to the degree of external validity. And there are no methods that can quantify the extra degree of uncertainty

(although one branch of advanced statistics, called Bayesian statistics, makes an attempt at this, but that's beyond the scope of our textbook). Instead, assessing external validity is a thinking exercise that requires knowledge about the patterns themselves, and it usually results in qualitative statements about how high or low external validity is in our case.

In practice, the goal of most analyses is a quantitative summary of patterns: in the simplest case, it is a **statistic**. A statistic, as we discussed, is a number that one can calculate from the data at hand. Examples include uncovering a probability, a correlation coefficient, or a difference between two groups. Using a somewhat simplistic but useful terminology, we want to infer the **true value** of the statistic after having computed its **estimated value** from actual data. The true value is a shorthand for the value in a population or a general pattern. Using this terminology, external validity is about the extent to which the true value of the statistic is similar in the general pattern our data represents and the general pattern that's behind the situation we care about. The goal of statistical inference is to uncover the true value of the statistic in the general pattern, or population, that our data represents. The starting point of statistical inference is the estimated value of the statistic that we calculate from our data. It is called an estimated value because the reason for calculating it is to uncover (estimate) the true value.

Review Box 5.1 ***Generalization from a dataset***

- Inference is the act of generalizing from the data to some more general pattern.
- In practice it amounts to conjecturing the true value of a statistic given its estimated value.
- The true value of the statistic is its value in a population, or general pattern.
- It is good practice to split the inference process into two parts:
  - Use statistical inference to uncover what the true value may be in the population, or general pattern, the data represents.
  - Assess external validity: Define the population, or general pattern, you are interested in; define the population, or general pattern, the data represents; compare the two.

2 A1 Case Study – What likelihood of loss to expect on a stock portfolio?

Question and data

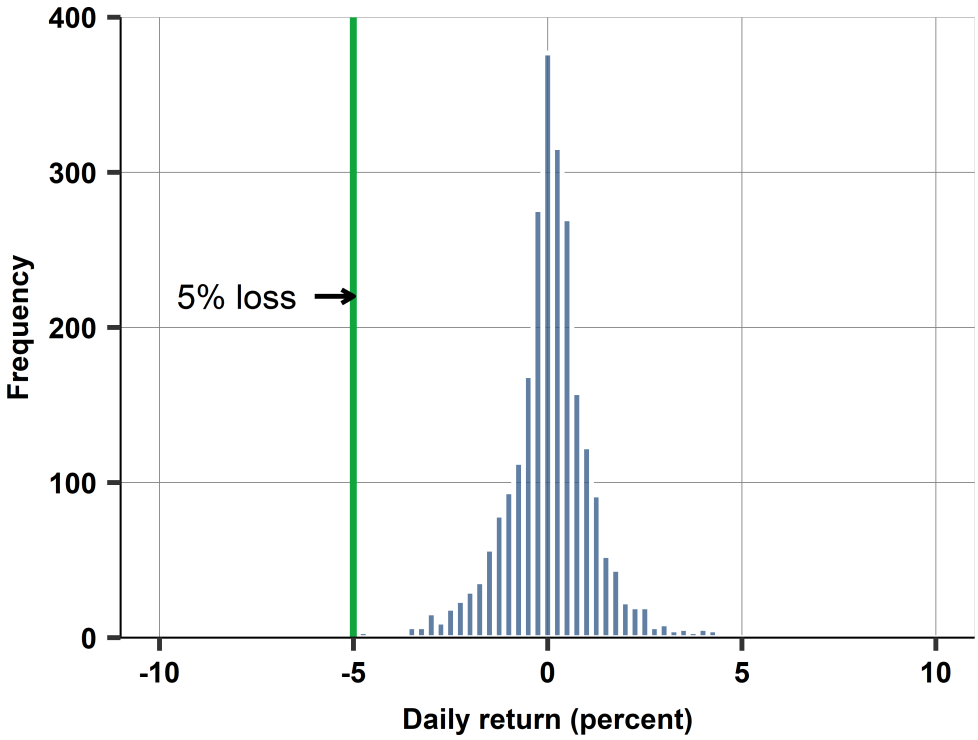
The case study in this chapter aims to assess the likelihood of experiencing a loss of a certain magnitude on an investment portfolio from one day to the next day. The goal is to guess the frequency of such a loss for the coming calendar year. This is an inference problem: we use data from the past to learn something about a different situation, the future. The first step is to assess external validity. The general pattern we care about is the pattern that determines the probability of large losses in the coming year. That may or may not be the same general pattern that determined the frequency of large losses in our data, covering past years. For now, let's proceed assuming that the general pattern remains the same, and let's return to this question at the end of our case study. The second step is assessing the likelihood of a large loss in the general pattern represented by our data.

The investment portfolio in this case study is the S&P 500, a stock market index based on company

shares listed on the New York Stock Exchange and Nasdaq. It is the weighted average of the price of 500 company shares and thus it is an investment portfolio with 500 company shares, each with its appropriate proportion.

The data includes day-to-day returns on the S&P 500, defined as percentage changes in the closing price of the index between two consecutive days. It covers 11 years starting with August 25, 2006 and ending with August 26, 2016. It includes 2519 days. This is time series data at daily frequency. The original data has gaps as there are no observations for the days the markets are closed, such as weekends and holidays. The data used in this case study ignores those gaps and simply takes returns between two consecutive days the markets were open. Figure 5.1 shows the histogram of daily returns.

Figure 5.1: Histogram of daily returns in the entire data.



Note: *Day to day (gaps ignored) changes, as percentages. From August 25, 2006 to August 26, 2016. N=2519*
Source: `sandp-stocks` data. S&P 500 market index. N=2519

To define large loss, we take a day-to-day loss exceeding 5 percent. That is a rare event, but it's not extremely rare. The histogram shows that -5 percent cutoff with a vertical line. We define large loss as a binary variable, taking the value one when the day-to-day loss exceeds 5 percent and zero otherwise. The statistic in the data is the proportion of days with such losses. It is 0.5 percent in this data: the S&P500 portfolio lost more than 5 percent of its value on 13 out of the 2519 market days between August 25, 2006 and August 26, 2016. What can we infer from this 0.5 percent chance for the next calendar year?

Note a limitation to our question: it is about the likelihood of a 5% or larger loss. It is not about the probability of a very large loss of, say, 20%. Nor is it about the expected value of losses. That's not

because these other questions are less interesting. It's because our data does not allow for answering them. The largest loss on the S & P 500 is 9% in our data, so the likelihood of larger losses is impossible to infer from our data, even though they are not impossible of course.

3 Repeated samples, sampling distribution, standard error

The conceptual background to statistical inference is **repeated samples**. The basic idea is that the data we observe is one example of many datasets that could have been observed. Each of those potentially observed datasets can be viewed as samples drawn from the population, or the more abstract general pattern. When our data is in fact a sample from a well-defined population, it is easy to see that many other samples could have turned out instead of what we have.

That would be the case with data on the earnings of market analysts. We have data from a sample taken from the population of all market analysts from your country last year; another random draw would have resulted in a different sample of individuals.

When the data is representative of a more abstract general pattern, the concept of repeated samples is more abstract, too. The data of returns on an investment portfolio is an example: our data is best thought of as a particular realization of the history of returns that could have turned out differently. Another example is when our data consists of a population, such as all firms in India that were in business in 2016. But those firms, with their management practices and performance, are just one realization of what could have happened in a country like India in a year like 2016.

The goal of statistical inference is learning the value of a statistic in the population, or general pattern, represented by our data. With repeated samples, the statistic has a distribution: its value may differ from sample to sample. The distribution of the statistic of interest is called its **sampling distribution**. The difference between female and male market analysts could have turned out differently in a different sample. The distribution of the quality of management may have been different if we chose a different sample of firms. The proportion of days with a 5 percent or larger loss on an investment portfolio may have turned out different if we could re-run history.

The sampling distribution of a statistic shows the values that the statistic takes across repeated samples. The most important aspect of this sampling distribution is its standard deviation: how much spread there is across repeated samples. The standard deviation of the sampling distribution has a specific name: it is the **standard error**, abbreviated as **SE**, of the statistic. The name originates from the fact that any particular estimate is likely to be an erroneous estimate of the true value. The magnitude of that typical error is one SE.

Review Box 5.2 ***Sampling distribution***

- The sampling distribution is the distribution of the estimates of the statistic across many repeated samples of the same size.
- The standard deviation of the sampling distribution is called the Standard Error (SE) of the estimated statistic.

4 A2 Case Study – What likelihood of loss to expect on a stock portfolio?

Repeated samples, Standard Error

The situation we care about in our case study is returns on the S&P 500 portfolio next year. The data we have is returns on the S$P 500 portfolio for a period of 11 years. The statistic of interest to us is the proportion of days with 5 percent or larger losses. This is a single number in our data: 0.5%. Statistical inference answers the question of what the proportion of such days could be in the general pattern that governed the observed history of returns.

The fraction of 5%+ losses could have turned out different in those 11 years if history played out differently. That history could have turned out differently is an assumption, but one that sounds reasonable, and we'll maintain that assumption throughout this case study. This is the main idea behind the concept of repeated samples.

Ideally, we would like to analyze many repeated samples of that 11-year history: alternative series of daily returns that could have happened instead of the one that did happen. We cannot re-run history, of course. In the following sections we'll learn methods that help make an educated guess of what could have happened had we been able to re-run history. For now, let's consider an artificial but educational example.

In this artificial example we carry out a **simulation exercise** to illustrate the concept of the sampling distribution. Simulation exercises are used often in statistics. The main idea behind simulation exercises is to create an artificial world where we know the true value of the statistic and see whether and how a particular method can uncover it. In our case that means knowing the true fraction of 5%+ losses in the general pattern that is behind what we observe in the data we have, and see what its estimated value looks like across repeated samples from that general pattern.

Of course we don't know the true value of the general pattern behind the 11-year history we observe. Instead, in this exercise we replace the question with one to which we know the answer. Suppose, for the sake of this artificial example, that the 11-year data is all there is. It is the population, with the general pattern: the fraction of days with 5%+ losses is 0.5% in the entire 11 years' data. That's the true value. The question is how we could uncover this true value if we had data not on all 11 years but only a part of it.

So, in the next step of the simulation exercise let's forget that we know this true value. Assume, instead, that we have only three years' worth of daily returns in our data. Our task is then to estimate the true value (in the entire 11-year period) from the data we have (three years). In the course of this exercise we can directly look at repeated samples because we can draw many random samples of 3-year time series from the 11-year time series we have. Those 3-year time series may be different from each other, thus the fraction of days with 5%+ losses may vary across them.

There are many ways to draw a sample of three years. For simplicity let's define three years of stock market data as 900 days of data. We could start with the first 900 days, then the next 900 days and so on – but that would yield only three samples. Alternatively, we could start with the first 900 days, then the 900 days starting with the second day, and so on. Instead, we opt for a third approach: simple random sampling. We start with the 11-year data, consider each day in it, one after the other, and we select or don't select each day in an independent random fashion.

(Note that this kind of sampling destroys the time series nature of our original data as the observations in the samples are not necessarily days next to each other. That is fine if the variable of interest, daily

returns, is independent across days in the original data: whatever the return is on one day has nothing to do with whatever it is on any other day, including the previous day or the subsequent day. It turns out that there is some dependence across daily returns, but it's weak and we'll ignore it in this exercise.)

So that's what we did. We took many random samples of 900 days and computed the percentage share of days with 5%+ loss in each sample. The power of this exercise lies in the fact that we can do this many times, creating many repeated samples from the same population.

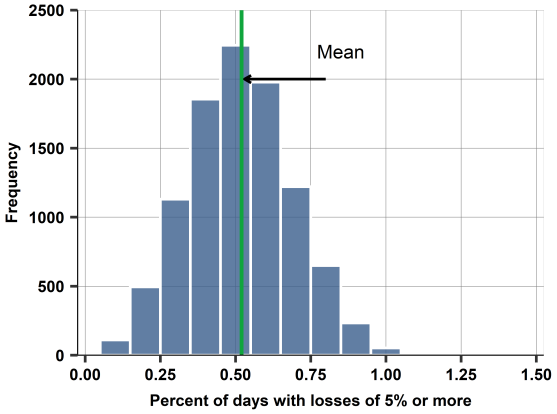
Figure 5.2 below shows the histogram created from the 10,000 random samples, each with 900 observations, drawn from the entire data. For each of these 10,000 samples, we have calculated the percentage of days with losses of 5% of more. The histogram shows the absolute frequency (out of 10,000) of the value across the repeated samples.

The histogram shows a distribution that has quite some spread: in some of the samples none of the days experienced such losses (0%); in other samples as many as 2% of the days did. Most values are closer to 0.5%. In fact, it turns out that the average of these values across the repeated samples is 0.5%, very close to the true value in the entire 11-year dataset that we aim to estimate (the difference is in the second decimal point).

The standard deviation of this sampling distribution is 0.2. That means that in a particular sample, we can expect to be 0.2% off from the true mean value, which, as we have seen, is also equal to the true value. This is the standard error of the statistic of interest (the fraction of 5%+ losses).

The third thing to note about this sampling distribution is its shape. The histogram shows a bell-shaped, or approximately normal, distribution (see Chapter 3, Section 15.)

Figure 5.2: Histogram of simulated number of days with big losses



Note: *Histogram of the proportion of days with losses of 5 percent or more, across repeated samples of size N=900. 10,000 random samples*
Source: `sandp-stocks` data. S&P 500 market index.

5 Properties of the sampling distribution

As we introduced in the previous section, the sampling distribution of a statistic is the distribution of this statistic across repeated samples. The sampling distribution has three important properties that are true in general. (To be more precise, they are true whenever the statistic is an average.) These properties are:

1. The average of the values in repeated samples is equal to its true value (the value in the entire population, or general pattern, represented by our data).

2. The sampling distribution is approximately normal.

3. The standard error (the standard deviation of the sampling distribution) is smaller the larger the samples, with a proportionality factor of the square root of the sample size.

The first property is called **unbiasedness**: the average computed from a representative sample is an unbiased estimate of the average in the entire population.

The second property is called **asymptotic normality**, or **approximate normality**. The “approximate" adjective means that the distribution is close to normal. The “asymptotic" adjective means that the larger the sample, the closer the distribution is to normal. Looking at larger and larger samples, the closer and closer the sampling distribution is to normal.

The third property is called **root-n convergence**: the standard error (the standard deviation of the sampling distribution) is inversely proportional to the square root of the sample size. As we look at larger and larger samples, the standard deviation is smaller and smaller by the square root of the sample size.

One consequence of root-n convergence is that the standard deviation is very small in very large samples. Taking this to its mathematical extreme, as we look at samples that are infinitely large, the standard deviation shrinks to zero. In that hypothetical extreme case, the sampling distribution is not really a distribution but yields the same value in virtually all of the repeated samples.

Because the sampling distribution is approximately normal, we know that the measured values fall within plus or minus two standard errors of the truth approximately 95 percent of the time. Similarly, we know that they fall within plus or minus 1.6 SE of the truth with a 90 percent chance, and within 2.6 SE with a 99 percent chance.

In reality, we don’t get to observe the sampling distribution. If we did, we would know what the true value was: it would be just its center, the mean of the sampling distribution. Instead, we observe data that is one of the many potential samples that could have been drawn from the population, or general pattern. However, it turns out that we can get a very good idea of what the sampling distribution would look like even from a single sample. We won’t be able to locate its center. But, with appropriate tools, we’ll get a pretty good estimate of the standard error. And with that, we’ll be able to come up with a range of where the true value may be, using our knowledge that the sampling distribution is approximately normal.

Review Box 5.3 ***Properties of the sampling distribution***

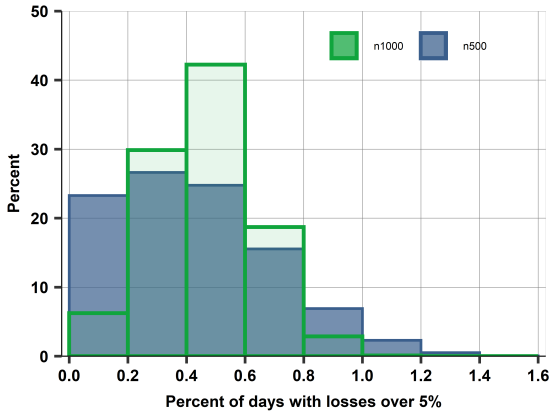
- The sampling distribution has three properties:
  1. Its mean approximately equals the true value.
  2. It is approximately normal.
  3. Its standard deviation is inversely proportional to the square root of the size of the repeated samples.
- The latter two properties are approximate; the approximation is better, the larger the samples.

6 A3 Case Study – What likelihood of loss to expect on a stock portfolio?

Sampling distribution with different sample sizes

We can carry out similar simulation exercises with different sample sizes. For example, let's look at samples of 450 days, half as many as we had earlier. The distribution of the fraction of days with 5 percent or larger losses are summarized by the two histograms shown together in Figure 5.3. The histograms have the same bins; within each bin, they show the number of draws (out of 10,000) for which the estimate (the proportion of days with large losses) falls within the bin. For example, in the bin (0.88, 1), we have 512 cases for the n=900 and 1250 cases for the n=450 exercise with values of 0.88 or 1.

Figure 5.3: Histograms of big loss simulations with N=450 and N=900



Note: *Histogram of the proportion of days with losses of 5 percent or more, across repeated samples in two simulation exercises, with N=450 and N=900. 10,000 random samples*
Source: `sandp-stocks` data. S&P 500 market index.

The figures show that, with 450 observations, more of the estimates are in the smallest and the largest bins, and fewer of the estimates are in the middle bins, than with 900 observations. The mean of the

two sampling distributions is very similar, around 0.5 percent, but the distribution that corresponds to the smaller sample is more spread out. The standard deviation of the fraction estimates across the repeated samples is 0.3 for 450 observations, and it is 0.2 for 900 observations. Both distributions are approximately bell-shaped, with some skewness with a longer right tail. The distribution of the smaller sample (450 observations) is more skewed.

Recall that the main properties of the sampling distribution are unbiasedness, approximate (asymptotic) normality, and root-n convergence. The results in this simulation exercise are in line with those properties. Recall also that the properties hold if the statistic is an average. Our statistic turns out to fit that criterion. The proportion of days with a 5%+ loss is the average of a binary indicator variable with a value of one on all days with a 5%+ loss and a value of zero on all other days (the average is 0.5% or 0.005 in our data).

A side note: how can then the fraction of 5+% losses be normally distributed across repeated samples when it can never be negative? The answer is that the normal distribution is never literally true. Instead, it is always an approximation. Taken literally, the normal distribution would allow not only for negative values but also irrational values such as pi or the square root of two. Returns on portfolios don't take these values either. The question is how close the distribution is to this ideal normal distribution. Going back to the histogram of the sampling distribution with N=900, we can see that it's fairly close. So saying that the distribution is approximately normal may make sense even if it can't be literally normal.

7 The confidence interval

The *Confidence Interval (CI)* is the most important measure of statistical inference. Recall that we use statistical inference when we analyze a dataset to infer the true value of a statistic: its value in the population, or general pattern, represented by our data. The CI defines a range where we can expect the true value in the population, or the general pattern, to lie. More precisely, the CI gives a range for the true value with a probability. That probability is something data analysts need to pick; it says how likely it is that the true value is in that range. Typical probability picks are 95 percent or 90 percent. The "95 percent CI" gives the range of values where we think that true value falls with a 95 percent likelihood. It also says that we think that with 5 percent likelihood, the true value will fall outside the confidence interval.

The confidence interval is constructed as a symmetric range around the estimated value of the statistic in our data. We first calculate the estimated value of the statistic in our data. Then we specify the length of the CI on the two sides of the estimated value. But how do know the length of the appropriate CI? It turns out that the key is the standard error (SE): the SE is the general unit of measurement when constructing confidence intervals. The appropriate 95 percent CI is the $\pm 2SE$ interval around the estimate from the data. The appropriate 90 percent CI is the $\pm 1.6SE$ interval around the estimate from the data; and the appropriate 99 percent CI is the $\pm 2.6SE$ interval around it. (We will discuss details in Section 12.)

In practice we don't know the sampling distribution so we don't know the appropriate standard error. The next two sections describe two approaches that give good estimates of the SE. Before discussing them, let's spend some time understanding why the CI constructed this way may tell us where to expect the true value.

Here is an intuitive argument. Assume that you know the sampling distribution. The histogram of the sampling distribution shows the frequency of particular values of the estimates across repeated

samples from the same population (or general pattern). Ninety-five percent of all estimates are within two standard errors of its center ($\pm 2SE$). Now take any value within that range. Measure the 95 percent interval around this value: it is $\pm 2SE$. The center of the sampling distribution will be within this interval. Recall that the center of the sampling distribution is the true value. Thus, the true value will be within the interval as long as we measure it around a point within $\pm 2SE$ of the center. If we were to measure the same interval around points outside that range, the true value would not be contained in it. The fraction of the samples for which this interval contains the true value is 95 percent; the fraction of the samples for which this interval does not contain the true value is 5 percent. That's why this is the 95 percent CI we are after: this interval contains the truth in 95 percent of the times among repeated samples. Viewed from the perspective of a single sample, the chance that the truth is within the CI measured around the value estimated from that single sample is 95 percent.

8 A4 Case Study – What likelihood of loss to expect on a stock portfolio?

Repeated samples and Confidence Interval

Let's return to our simulation example of repeated samples of 900 days where the statistic is the proportion of days with losses of 5 percent or more. If we know the sampling distribution of this proportion, we know its standard deviation. That's the standard error (SE), and its value is 0.2 here. The mean of the sampling distribution is 0.5. Thus, the fraction of days with losses of 5 percent or more is within 0.1 and 0.9 percent in 95 percent of the samples ($0.5 \pm 2 \times 0.2$).

Recall that this sampling distribution is from 900-day samples. If, instead of 900 days, we measured the fraction of 5+ percent loss days in samples of 450 days, the SE would be 0.3; the fraction of days with losses of 5 percent or more would be within 0 and 1.1 percent in 95 percent of the samples ($0.5 \pm 2 \times 0.3$ and recognizing that the fraction can't be negative so the lower bound is zero).

9 Discussion of the CI: Confidence or probability?

Having introduced the confidence interval and how we can compute it if we know the SE, let us discuss a bit more what we exactly mean by a CI. It turns out that there are two major leaps in the intuitive reasoning we presented in the previous section.

The first one is conceptual. The 95% probability that a statistic is within a range in a particular sample is the 95% frequency that it would occur in repeated samples. This probability is thus a relative frequency. In contrast, our "95% confidence" that the true value is within a range has no corresponding frequency to consider. The population, or general pattern, is not repeated. Rather, this 95% confidence characterizes our personal feeling of uncertainty. That's why we call it "confidence" and not probability. But for the argument to work, that confidence needs to be a probability, even if it cannot be defined through frequencies. It is also called "subjective probability," a concept that we mentioned when we first discussed probabilities in Chapter 3 Section 2.

The second leap builds on the first one and takes the 95% confidence as a probability. If so, it is a conditional probability: the probability that the true value is within an interval conditional on the estimated value. This conditional probability is the inverse of the probability that the estimated value falls in an interval conditional on the true value. It is this latter probability that we can derive from

the approximately normal sampling distribution. And, as we discussed in the Under the hood Section of Chapter 4, inverse conditional probabilities are not equal in general. Here, for the two to be equal, we need to assume that, without the data, we would have no idea what the true value could be. In particular, we have no prior knowledge, or even a guess, of the true value. Note that, more often than not, we do have some idea about the true value: the average price of cars like ours cannot be anything, the fraction of days with large losses is unlikely to be very high, etc. Nevertheless, we usually go ahead and interpret the confidence interval as if we had no prior idea.

Bayesian statistics, a branch of statistics that we mentioned earlier in Section 1, takes these issues seriously. It works with the assumption that analysts can quantify their prior ideas and takes those into account for inference. We don't cover Bayesian statistics in this book. Instead, we construct and interpret confidence intervals the way most data analysts do, maintaining the assumption of classical statistical inference: as analysts we approach every inference problem with a completely open mind about what the true value may be.

So far all this is just theory. We can't come up with a CI in practice unless we know the standard error, but we can't know it by looking at the distribution of repeated samples as we only observe a single one of those repeated samples. However, statistics offers not one but two solutions to make a good guess of the SE from that single sample. Both solutions are quite ingenious, but neither is simple to understand. Some of us would argue that these solutions are among the things that make statistics exciting – you could say, beautiful.

10 Estimating the standard error with the bootstrap method

Recall our simulation study: we took many samples of the same size from the original data to construct the sampling distribution of the statistic we were after. We introduced three important properties of the sampling distribution: unbiasedness, root-n convergence, and approximate (asymptotic) normality. We illustrated how to get a sampling distribution with the help of an artificial example: the true value of a statistic to figure out was actually its value in the entire dataset. We wanted to see the sampling distribution of this true value's estimates from data with fewer observations.

But that was an artificial situation. In practice we examine all the data we have, and we would like to uncover the sampling distribution of a statistic in samples similar to that data. In particular, we want the sampling distribution of samples of the same size as the original data, all of which represent the same population or, more generally, the same general pattern.

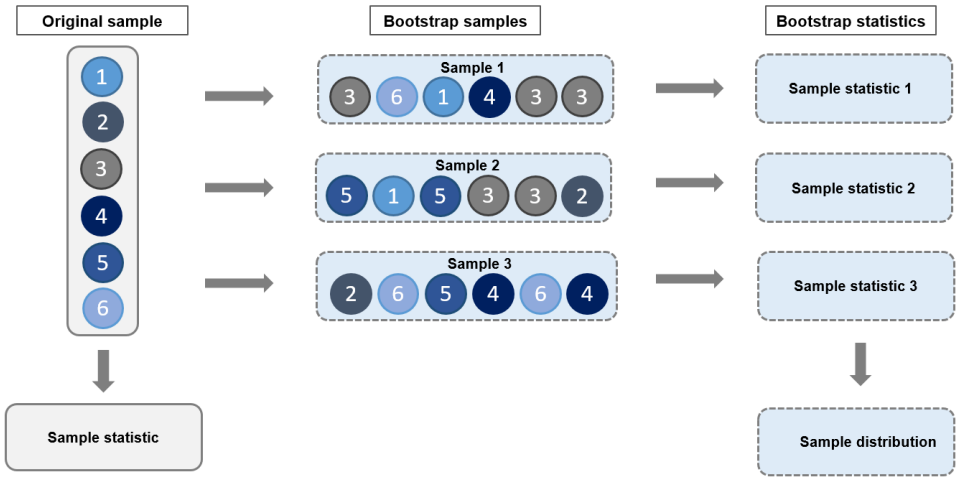
The **bootstrap** is a method to achieve just that: it creates the sampling distribution across repeated samples that are representative of the same population, or general pattern, as the original data and also have the same size. The bootstrap method stretches the logic of the simulation exercise to draw samples using only the data we have, and those samples have the same size as the original data yet are different from it. We might think that this should be impossible just like it is impossible to pull yourself out of the swamp by your bootstraps. Hence the name of the method. It turns out, however, that the trick is possible here.

The bootstrap method takes the original data and draws many repeated samples of the size of that data. The trick is that the samples are drawn with replacement. The observations are drawn randomly one by one from the original data; once an observation is drawn it is "replaced" into the pool so that it can be drawn again, with the same probability as any other observation. The drawing stops when the sample reaches the size of the original data. The result is a sample of the same size as the original data, yielding a single **bootstrap sample**. This bootstrap sample includes some of the original observations

multiple times, and it does not include some of other original observations. Perhaps surprisingly, in data that are moderately large, more than one third of the original observations in the data don't end up in a typical bootstrap sample.

Repeating this sampling many times, the bootstrap method creates many repeated samples that are different from each other, but each has the same size as the original data. The miracle of the bootstrap is that the distribution of a statistic across these repeated bootstrap samples is a good approximation to the sampling distribution we are after: what the distribution would look like across samples similar to the original data.

Figure 5.4: Bootstrap samples



Note: *Our data and repeated samples. Some observations are not included, others may be included multiple times.*

The bootstrap method works if the variable we consider is independent across observations: its value is not related to its previous values or subsequent values. By sampling observations one by one and then replacing them into the pool to be possibly drawn again, we destroy the before-after relationships in the original data. That is not a problem if there is no relationship across observations, which is true in most cross-sectional data. If there is an important relationship, as in many kinds of time series data, this bootstrap method does not work. (Some time series variables may be independent across observations, as we had in our case study.) Note that there are more sophisticated bootstrap methods that work with time series data by preserving relationships across observations, but we don't consider those in this textbook.

Since the distribution of a statistic across bootstrap samples is a good approximation of the sampling distribution, it gives a good approximation of the standard error, too. The **bootstrap estimate of the standard error** is simply the standard deviation of the statistic across the bootstrap samples.

Once we have the standard error, we have everything we need for constructing the confidence interval of our choice. First, we take the value of the statistic estimated from the original data. Second, we take the standard error computed with the bootstrap. Using these two numbers, we calculate the confidence interval, for instance taking the $\pm 2SE$ for the 95 percent CI.

The bootstrap is based on a clever yet conceptually simple idea. It is computationally intensive: the computer has to take many random draws, calculate the estimated value of the statistic across many

samples, store those results, and then calculate the standard deviation. All this is done by the machine. It may take some time, but it is not the data analyst's time.

If the bootstrap sounds like a miracle, something too good to be true, it is because there is something to that. The bootstrap does indeed have its limitations, on top of the independence we discussed above. For example, it tends to provide very poor standard error estimates in the presence of extreme values. In fact, if the data has very extreme values, such as in the power-law distribution, the standard error estimates are way too small and very misleading. Moreover, the bootstrap may produce poor standard error estimates in small samples if the statistic is not the mean but the median or some other quantile. At the same time, there are many extensions of the bootstrap that can deal with some of its problems. In this textbook we focus on the average and related statistics of variables that are independent across observations: situations that are well suited for bootstrap standard error estimation.

It turns out that the largest advantage to SE estimation by the bootstrap method is that, in large datasets, it allows us to estimate confidence intervals for statistics other than averages. This includes quantiles or nonlinear functions of averages. This is a big advantage as the other method, that we'll discuss in the next section, is not available for many of those statistics.

Review Box 5.4 ***The bootstrap method***

- The bootstrap is a method to generate samples that represent the same population, or general pattern, as our original data and also have the same size.
- Each bootstrap sample is a random sample from the original data with replacement.

  The bootstrap procedure generates many such samples, approximating the sampling distribution of a statistic.
- We use the bootstrap method to get the standard error (SE) of a statistic: it is the standard deviation of the estimated value of the statistic across the bootstrap samples.

11 A5 Case Study – What likelihood of loss to expect on a stock portfolio?

Bootstrap standard error estimation

Going back to our example, the task is to estimate the confidence interval for the proportion of days with large negative returns. First, we estimate that proportion from the data: we have already done that, it is 0.5 percent. The second step is estimating its standard error. The third step is measuring plus and minus two standard errors around that 0.5 percent value. The result will be the 95 percent CI that tells us where to expect the majority of days with large negative returns in the general pattern represented by our data.

We estimate the standard error by bootstrap. We apply the bootstrap procedure we studied that is applicable for variables that are independent across observations. As we noted earlier, there appears to be some dependence in this data but it's weak and we ignore it.

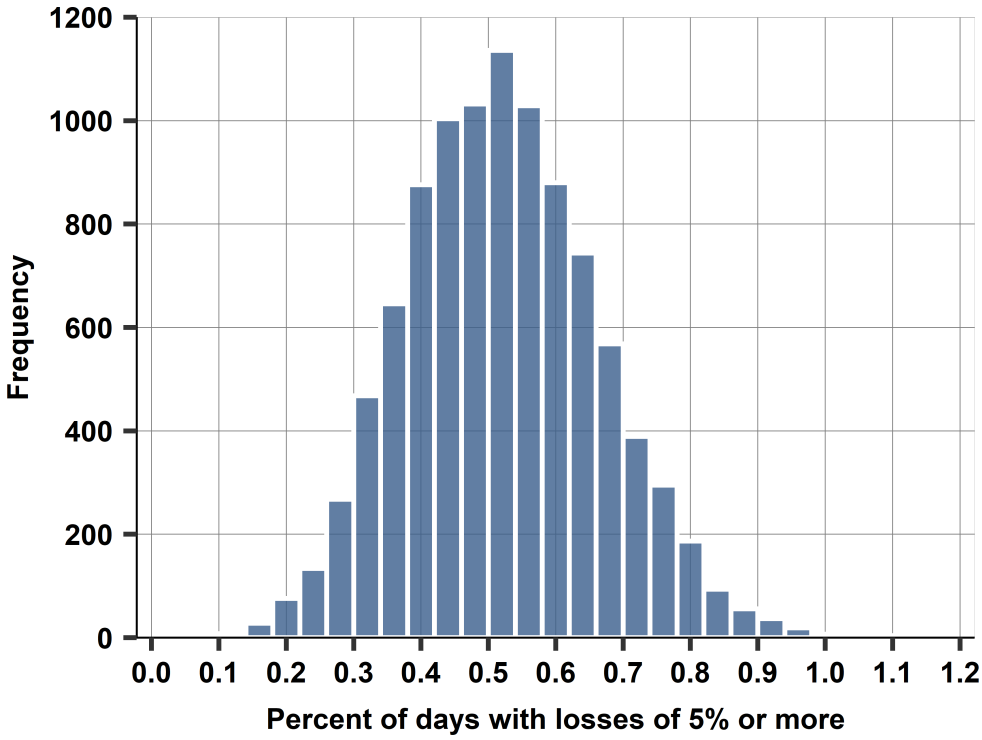
Then the procedure is as follows. Take the original data and draw a bootstrap sample. Calculate the proportions of days with 5%+ loss in that sample. Save that value. Then go back to the original data and take another bootstrap sample. Calculate the proportion of days with 5%+ loss and save

that value, too. And so on, repeated many times. In the end, we end up with a new a data table, in which one observations stands for one bootstrap sample, and the only variable contained in it is the estimated proportion for each bootstrap sample. The standard error we are after is simply the standard deviation of those estimated values in this new data.

We created 10 000 bootstrap samples from the original data. Each bootstrap sample had 2519 observations just like the original data. We calculated the estimate of the statistic in each: the proportion of days with 5+ percent loss. Its value varied across the bootstrap samples, from as low as 0.1 percent to as high as 1.2 percent. On average, the value of the statistic was 0.5, equal to its value in the original data. The median is equal to the mean, indicating a symmetric distribution.

Figure 5.5 shows the bootstrap distribution: the sampling distribution of the fraction of days with 5+% losses across the 10 000 bootstrap samples. The histogram shows a bell-shaped distribution, as it should.

Figure 5.5: Bootstrap distribution of the proportion of days with losses of 5 percent or more



Note: *10 000 bootstrap samples from the original data; each had 2519 observations just like the original data.*
Source: `sandp-stocks` data. S&P 500 market index.

The goal of this exercise was to get the standard error of the statistic, which is simply the standard deviation across the bootstrap samples. Here that value is 0.14: this is the bootstrap SE estimate. With the standard error estimate at hand, we are ready to construct the confidence interval. The 95 percent CI is 0.5 percent $\pm 2 \times 0.14\%$, which is $[0.22\%, 0.78\%]$ using values rounded to the second decimal. Its interpretation: in the general pattern represented by the 11-year history of returns in our data, we can be 95 percent confident that the chance of daily losses of more than 5 percent is

somewhere between 0.22 and 0.78 percent (rounded to the first decimal).

12 The standard error formula

The bootstrap method reconstructs the entire sampling distribution and calculates the standard error directly from that distribution. In this section we show an alternative approach: using a formula for the standard error, without the need to simulate the entire sampling distribution by bootstrap or any other method.

The case we consider is a variable whose value is independent across observations in the data, just like for the bootstrap. The statistic we consider is the average. In this case we know that the sampling distribution is approximately normal, with the true value as its mean. We have seen that the standard error, which is the standard deviation of the sampling distribution, depends on the size of the sample. We called that dependence root-n convergence, as the standard error decreases by the square root of the sample size.

There is a relatively *simple formula for the standard error (SE) of the average* ($\bar{x}$). The formula includes the root-n factor and just one other element. $\bar{x}$ is the estimate of the true mean value of x in the general pattern (or the population).

The standard error formula for the estimated $\bar{x}$ is

$$SE(\bar{x}) = \frac{1}{\sqrt{n}} Std[x] \tag{5.1}$$

where $Std[x]$ is the standard deviation of the variable in the data (the square root of its variance), and n is the number of observations in the data. We emphasize again that this formula is valid if the values of x are independent across observations in the data. There are corresponding SE formulae for variables with dependence across observations. We'll consider some of them in Chapters 9, 12, and 23 in this textbook.

The standard error is larger, the smaller the sample and the larger the standard deviation of the variable. The first of these we know already, it's root-n convergence. The second one is new. The relation of $SE(\bar{x})$ to $Std[x]$ is not an obvious one, but it makes intuitive sense. Recall that the standard error of $\bar{x}$ is the standard deviation of the various $\bar{x}$ estimates across repeated samples. The larger the standard deviation of x itself, the more variation we can expect in $\bar{x}$ across repeated samples. At the extreme, if x is always the same so that $Std[x] = 0$, its average value estimated in various repeated samples should be the same, too, so that $SE(\bar{x}) = 0$.

The standard error formula is built into all statistical packages, including R and Stata. Therefore, we rarely compute it using the formula ourselves. Nevertheless, we think it is instructive to memorize this simple formula for the intuition it contains.

Review Box 5.5 ***The standard error formula***

- The standard error formula of the estimate of an average is

$$SE(\bar{x}) = \frac{1}{\sqrt{n}} \times Std[x] \tag{5.2}$$

- This formula gives a good estimate of the standard error if the values of x are independent across observations in the data.
- $SE(\bar{x})$ is larger if
  - the number of observations, n, is smaller;
  - the standard deviation of the variable, $Std[x]$, is larger.

13 A6 Case Study – What likelihood of loss to expect on a stock portfolio?

The standard error formula

Let's consider our example of 11 years' data on daily returns of the S&P 500 portfolio. The size of the sample is n = 2519 so that $\sqrt{(1/n)} = 0.02$. The standard deviation of the fraction of 5+% losses is 0.07. Multiplying 0.07 by 0.02 results in 0.0014, or 0.14 percent. This is the same number that we got from the bootstrap exercise.

With the same SE as in the bootstrap, the CI and its interpretation are the same, too: the 95 percent CI is [0.22, 0.78]. This means that in the general pattern represented by the 11-year history of returns in our data, we can be 95 percent confident that daily losses of more than 5 percent occur with a 0.2 to 0.8 percent chance.

14 External validity

The confidence interval summarizes uncertainty about the true value of the statistic in the population, or the general pattern, that our data represents. An important part in inference is defining the population, or general pattern, we care about, and assessing how close it is to the population, or general pattern, our data represents. As we introduced earlier in Section 1. **External validity** is the concept that captures the similarity of those two general patterns. We say of our data analysis that the external validity is high when the two are the same. When the two general patterns are very different, we say that the external validity is low.

With high external validity, the confidence interval captures all uncertainty about our estimate. With low external validity, it does not. With low external validity, our estimate of the statistic is likely to be off with a larger likelihood than is suggested by the confidence interval. How much larger is very difficult to quantify. Therefore, we usually describe that extra uncertainty in qualitative terms only. For example, we may say that the CI captures all or most of the uncertainty when we believe that external validity is high. In contrast, when we believe that external validity is low, we say that the

estimated value may be off with a higher likelihood than the CI suggests.

External validity is as important as statistical inference. However, it is not a statistical question. By that we mean that the methods of data analysis and statistics cannot be used to assess it in a direct fashion. Instead, we need substantive knowledge of the situations. Some data analysis may help in that assessment, such as benchmarking variables for which we have some information in the population we care about (benchmarking was introduced in Chapter 1, Section 14). But those methods can never fully assess external validity.

The most important challenges to external validity may be collected in three groups: time, space, and sub-groups. Viewed from a more abstract perspective, all three concern the stability of the general pattern, between the one our data represents and the one we care about.

First, stability in time. The data we analyze is always from the past, whereas the situation we care about is almost always in the future. Thus, in order to generalize from the data, we need to assume that the general pattern that was relevant in the past, will remain relevant for the future.

In many applications, there is a great deal of stability. The impact of rain and temperature on crops is pretty stable from one year to another (even if the patterns of weather may change). At the same time, whatever predicts the success of businesses in an industry in the past may be different from what will predict their success in the future (think of the rising and then diminishing role of access to navigable waterways).

A specific issue with time is the occurrence of extreme values. Our data may be free of some extreme values because some very rare extreme events may not have happened in the historic period that our data represents. However, this does not imply that such a rare extreme event could not happen in the future we care about.

Second, stability through space. The data we analyze may come from a different country, region, or city than the situation we care about. Taste and behaviour of people may vary and so does historic background or the transport infrastructure. Whether that means that the general pattern of interest varies through space is an important question. For example, gender differences in earnings among market analysts may differ from country to country; the extent to which better management is associated with exporting to foreign markets may be more stale across countries.

Third, stability across subgroups. A pattern may be strong among one group of people, but it may not be there for people in different professional, age, gender, cultural, or income groups. For example, an experiment may show that working from home makes employees more productive and less likely to quit a company in the service industry in China (this will be our case study in Chapter 20). Would working form home have similar effect on accountants or programmers at a Chinese IT company?

Review Box 5.6 ***External validity of the results of data analysis***

- External validity means the extent to which the results of an analysis generalize from the population, or general pattern, represented by the data to the population, or general pattern, we care about.
- When the two are close, we say the external validity of the analysis is high. When the two are far, we say its external validity is low.

15 A7 Case Study – What likelihood of loss to expect on a stock portfolio?

External validity

The 95 percent confidence interval of the probability of a daily loss greater than 5 percent was [0.2, 0.8] percent in our example. This is where the probability of such losses is likely to be found in the general pattern that our data represents.

The general pattern represented by our data is how things worked in the 11 years contained in our data; the general pattern we care about represents how things will work in the coming year. External validity of the 95 percent confidence interval is high if the future one year will be like the past 11 years in terms of the general pattern that determines returns on our investment portfolio. With high external validity, this confidence interval would tell us where to expect the probability of 5%+ loss in the coming year.

However, external validity may not be that high in our case. Whether the next year will be like the past eleven years is difficult to tell. Even if we have no particular reason to expect the future to be systematically different, we can't rule that out. For instance, our 2006-2016 data includes the financial crisis and great recession of 2008–2009. It does not include the dotcom boom and bust of 2000–2001. We have no way to know which crisis is representative of future crises to come.

Therefore, it makes sense to consider a wider interval for 5%+ loss in the general pattern in the future. How much wider is impossible to determine by data analysis. This is quite common, as we can't, in general, put numbers on uncertainty coming from low external validity. At the very least it is good practice to consider the 95 percent CI of [0.2, 0.8] to be too narrow. Sound investment decisions based on our analysis should acknowledge that daily losses of 5+ percent may have a higher probability than 0.8 percent, the upper end of this interval.

This concludes our case study. It illustrated how to make statistical inference in two ways, using the bootstrap and using a formula for the standard error. The two methods gave the same result, which is how it should be, and is almost always the case when the statistic is an average and the sample is large enough. Then we used this SE to construct a 95% confidence interval to quantify the degree of uncertainty about the true value of the statistic: where we can expect the probability of a 5% or larger loss may be in the general pattern that determined the 11-year history of our data. Before carrying out that statistical inference, we used the data of this case study for a simulation exercise to illustrate the concept of repeated samples. Finally, we illustrated the way to think about inference in general: start thinking external validity, go ahead with statistical inference if there is a chance that external validity is high, and then return to the question when the analysis is done, to qualify its results in case external validity is not that high.

16 Big data, statistical inference, external validity

Big data and the advance of computing has changed the way we conduct statistical inference in several important ways. Recall that big data means a great many observations, or, less often, a great many variables. The case of many variables can pose a unique challenge for inference, and we'll discuss that in Chapter 6, Sections 11 and 14. Here we focus on big data with very many observations.

With very many observations, confidence intervals are typically very narrow. That should be obvious

by now, from the more abstract concept of root-n convergence and the more tangible role of the number of observations in the SE formula. Whatever statistic we estimate from our big data, it will be very close to its true value in the population, or general pattern, our data represents. Thus, statistical inference is just not very important with such big data.

In case we still wanted to carry out statistical inference, the advances in computing makes it a lot easier to use computationally intensive methods, such as the bootstrap, instead of formulae. This is actually more important with not-so-big data. With very big data, the bootstrap process may take a lot of time.

Importantly, however, big data in itself does not change the need to assess external validity or the way to do that. In particular, our analysis does not have higher external validity just because we have many observations. External validity is not affected by the size of the dataset or the ability to carry out better statistical inference by computationally intensive methods. From the viewpoint of external validity, having very many observations that poorly represent the general pattern that we are after is just as problematic as having few such observations. Indeed, a smaller representative dataset may be highly superior to a huge non-representative data.

This concludes our chapter that introduces the conceptual questions of generalizing from data, and the methods to carry out statistical inference. The next chapter introduces testing hypotheses, a way to look at the same question from a more specific angle.

17 Summary and practice

17.1 Main takeaways

- Generalizing results from your data means estimating how they would be reflected in the general pattern that describes the situation you care about.
    - First step is statistical inference: to the general pattern your data represents
    - Second step is about external validity: from the general pattern your data represents to the general pattern that describes the situation you care about
    - With big data (large N), the first step is not interesting but the second step is just as important.

17.2 Practice questions

1. When do we call a sample representative and how it is connected to random sampling? Why do people favor random sampling?
2. What are the two parts of the inference process? List them, explain them, and give an example with the two parts.
3. When do we say that results from analyzing our data have high external validity? Low external validity?
4. What's the population, or general pattern, represented by the monthly time series of unemployment rates in Chile between 2000 and 2018?

5. What's the population, or general pattern, represented by data on exports per total sales in a cross-sectional random sample of companies in Vietnam in 2018?
6. Does it make sense to create a confidence interval for a statistic that you calculated using cross-country data with all countries in Earth? If not, why not? If yes, what's the interpretation of that CI?
7. What does the confidence interval show? What are the typical likelihoods used for them?
8. The proportion of daily losses of more than 2% on an investment portfolio is 5 percent in the data. Its confidence interval is [4,6] percent. Interpret these numbers.
9. In the data that is a random sample of the population of your country from last year, 30-year-old market analysts earn 30 percent more than 25-year-old market analysts, on average. The 95% CI of this difference is [25,35] percent. Interpret these numbers.
10. In the example above, what can you conclude about the expected wage difference between 30 and 25 year-old market analysts in your country five years from now? In a different country five years from now?
11. How would you estimate the bootstrap standard error of a statistic? Under what assumption does the bootstrap with random sampling work?
12. What's the standard error formula for an average, and what does it imply about what makes the SE larger or small? Under what assumption does this formula work?
13. How do you create the 95% CI of a statistic if you know its SE? How do you create the 90% CI?
14. Name two kinds of stability that are important challenges to external validity. Give an example for each.
15. You downloaded data from the World Development Indicators database on GDP per capita and CO_2 emission per capita, and find that their correlation coefficient is 0.7, with SE=0.05. Create a 95% CI and interpret it.

17.3 Data exercises

Easier and/or shorter exercises are denoted with [*] Harder and/or longer exercises are denoted with [**]

1. Download ten years of daily data on the price of another financial asset, such as an individual stock, or another stock market index. Document the main features of the data, create daily percentage returns, and create a binary variable indicating large losses by choosing your own cut-off. Estimate the standard error of the estimated likelihood of large daily losses by bootstrap and using the SE formula; compare the two, and create 95% confidence intervals. Conclude by giving advice on how to use these results in future investment decisions. [*]
2. Download ten years of daily data on the price of another financial asset, such as an individual stock, or another stock market index. Create daily percentage returns, and create a binary variable indicating large losses by choosing your own cut-off. Carry out a simulation exercise pretending that the truth is contained in your entire data and you want to infer that from a sample of 300 days. Take repeated samples, visualize the distribution of large daily losses, and describe that distribution. Repeat the simulation with samples of 900 days instead of 300 days, and compare the results [**]

3. Use the `hotels-europe` data and pick two cities and the same date. In each city, take hotels with three stars and calculate the average price. Estimate the standard error of the estimated average price by bootstrap and using the SE formula, and create 95% confidence intervals. Compare the average price and the confidence intervals across the two cities, and explain why you have a narrower CI for one city than the other. Conclude by giving advice on how to use these results for a hotel manager who will decide on the advertised price if their three-star hotel one of the cities. [*]

4. Use the `wms-management-survey` data and pick two countries. Estimate the average management quality score in each. Estimate their standard error by bootstrap and using the SE formula, and create 95% confidence intervals. Compare the average price and the confidence intervals across the two countries, and explain why you may have a narrower CI for one country than the other. Discuss the external validity of you results for the quality of management in your country of origin in the current year. [*]

5. Download the most recent data from the World Development Indicators website on GDP per capita and CO_2 emission per capita. Divide countries into two groups by their GDP per capita and calculate the average difference in CO_2 emission per capita the between the two groups. Use bootstrap to estimate its standard error, create the appropriate 95% CI, and interpret it. [**]

18 Under the hood: The Law of Large Numbers

The **Law of Large Numbers** (LLN) and the **Central Limit Theorem** (CLT) are the two most important tools to assess how an estimate in a dataset compares to its true value: its value in the population, or general pattern, represented by the data. The LLN and the CLT are true when our estimates are averages.

In a nutshell,

- the LLN tells us that the estimated average ($\bar{x}$) is likely to be very close to the true mean if our sample is large enough, while
- the CLT tells us in what range $\bar{x}$ is around the true mean and helps estimating that range from the sample we have.

The results of the LLN and the CLT are mathematically correct when the samples are infinitely large. Of course, in practice they never are. We use these results to describe how things look like in samples that are “large enough." What “large enough" means is not defined by the theorems.

There are, however, many simulation studies that give us some guidance. If the variable is close to normally distributed, it seems that these theorems give surprisingly good approximation for samples as small as 40 or even smaller. The further away the variable is from normality, the larger the sample needs to be. A few hundred observations almost always justify the use of the LLN and CLT results. There are special cases, though, with large extreme values, when even hundreds of observations may not be enough, so some alertness is helpful.

The LLN and CLT we present below are derived for **i.i.d. variables**: identical and independently distributed variables. For such variables, the value of this variable in one observation conveys no information about its value in another observation, and knowing nothing more than the position of an

observation in the data, we should expect that it is just like any other observation. The LLN is true for many other kinds of variables, too. There are versions of the CLT that are true for other kinds of variables; they differ from the version we present by the standard deviation of the normal distribution. We do not consider those other LLN and CLT results here.

The LLN tells us that in large enough samples, our estimates of the average will be very close to their population value. The larger the sample, the closer to that value our estimate will be.

The theorem itself tells us that the sample average of a variable gets very close to the population mean of that variable if the sample is large.

The most straightforward version is derived for i.i.d. variables. As a formula, LLN says that

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \rightarrow E[x] \; in \; probability \tag{5.3}$$

where $E[x]$ means the true expected value of x, and $\rightarrow$ in probability means that the left hand-side of the formula approaches the right hand-side, with probability 1 if n goes to infinity. Somewhat more precisely this means the following: tell me how close you want to be to the population value and I can give you a large enough sample size that will get you that close with as high a probability as you want: 90%, 99%, 99.99%, etc. Another notation of the same thing is that the "probability limit," abbreviation plim, of the sample average is the expected value:

$$plim \; \bar{x} = E[x] \tag{5.4}$$

19 Under the hood: The Central Limit Theorem

The LLN assures us that the estimated value is likely close to the true value, but it does not tell us how close it is. This is where the CLT can help.

The CLT states that, in large enough data, the estimated average will be distributed normally around the true expected value, and the variance of this normal distribution is related to the variance of the original variable that makes up the average.

The CLT we discuss here is derived for i.i.d. variables. The CLT states the following:

$$\sqrt{n}(\bar{x} - E[x]) \overset{A}{\sim} N(0, Var[x]) \tag{5.5}$$

In English, the sample average minus the expected value, multiplied by the square root of the sample size, is distributed approximately normally (the wiggle with letter A on top means approximate distribution) with mean zero and variance that equals the variance of the variable itself.

Of course we rarely are interested in the distribution of things like $\sqrt{n}(\bar{x} - E[x])$. Instead, we are usually interested in the distribution of $\bar{x}$ itself. The question is how we can get that from the CLT. The formula is relatively straightforward:

$$\bar{x} \overset{A}{\sim} N\left(E[x], \frac{1}{n}Var[x]\right) \tag{5.6}$$

In large samples, the sample average is distributed approximately normally around the population mean, and the variance of that distribution equals the variance of the variable itself divided by the sample size.

One may wonder why we state the CLT in the first form at all instead of proceeding right away with this second form. The reason is that the theorem is literally true as n goes to infinity; but then the variance in the second formula would go to zero (as it should, otherwise it would not be consistent with the LLN, which says that sample mean should go to the population mean with probability one). In practice, there is of course no such thing as an infinitely large sample. Instead, we want to interpret these theoretical results as approximations in large samples. So we can use the second formula without problems.

In order to understand the main messages of the CLT, let's go back to the theoretically derived first formulation. It is instructive to break down the message of the CLT into three pieces.

1. In large enough samples, the estimated average minus the true expected value, multiplied by the square root of the sample size, falls within a range around zero. Technically, it is a random variable, which means that we can't tell its value for sure but we can tell its distribution. This means that for any probability, we can find a range within which the variable falls with that particular probability. If you think about it, this is already a remarkable result. As the sample size goes to infinity, $\sqrt{n}$ goes to infinity. At the same time, $\bar{x}$, the sample average, goes to the expected value $E[x]$ with probability one (by the LLN), so that their difference $\bar{x} - E[x]$ goes to zero with probability one. And here is the interesting thing: the product of the two goes neither to infinity nor to zero. Actually, we know more: it will fall within certain ranges around zero with certain probabilities.

2. The distribution of this thing, the sample average minus the expected value multiplied by the square root of the sample size, is normal. This is true regardless of the distribution of the variable itself. This is a remarkable result, too. Whatever the distribution of x, the distribution of $\sqrt{n}(\bar{x}-E[x])$ will be normal. That's right: whatever the distribution of x is! In fact, you can think of the normal distribution as defined by the CLT, so that we can expect something to be distributed normally if it is the average of many independent things.

3. The mean and the variance of that distribution equal the mean and the variance of the original variable itself. Thus, if we know the sample size n, the mean of x, and the variance of x, we know everything about the distribution. That is, of course, if the sample is large enough. The great thing is that we have pretty good estimates for these: the LLN guarantees that we can substitute the sample average and the within-sample variance in for the population mean and variance, respectively. Why? The LLN tells us that the sample mean x is close to the population mean $E[x]$. But the LLN also tells us that the within-sample variance is close to the population variance because the latter is a kind of a mean, too: $Var[x] = E[(x_i - E[x])^2]$. So the variance within the sample $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is close to the population variance in large samples.

20 References and further reading

A classic but still relevant introduction to the logic of statistical inference is the introductory chapter of Ronald A Fisher's 1925 book (Fisher 1925).

On the central role that the confidence interval should play in statistical inference in economic applications see (Romer 2020).

The way statistics represents uncertainty and potential issues with this approach is discussed in a historical context by Salsburg (2001).

A related, but broader question is examined in great detail by (Manski 2020): why should results of data analysis try to quantify the degree of uncertainty when data analysts want to influence decisions?

On Bayesian analysis, a comprehensive text is Gelman & Rubin (2018). Andrew Gelman also has a great blog on statistics and causal analysis in social sciences at `statmodeling.stat.columbia.edu/`.

Chapter 6

Testing hypotheses

How to formulate a hypothesis, and how to use evidence in the data to help decide if we can maintain or reject it

Motivation

You want to know whether online and offline prices differ in your country for products that are sold in both ways. You have access to data on a sample of products with their online and offline prices. How would you use this data to establish whether prices tend to be different or the same for all products?

You have conducted an experiment to see whether a new design for the online ad of your product would yield more customers purchasing your product. In your experiment, customers were randomly chosen to see one of the two versions of the ad. You have data on whether each customer followed up and made a purchase decision. You see more follow-ups for those who saw the new version in your data, but the follow-up rates are very small to begin with, and their difference is not very large. How can you tell from this evidence whether to expect the new version to result in more follow-ups in the future?

Generalizing the results of our analysis from the data we have to the situation we care about can be carried out in a more focused way than we discussed in the previous chapter. One such focused approach uses a statistic (e.g., a difference in two means) computed from our data to see whether its true value is equal to something we assume (e.g., the difference is zero). This is called hypothesis testing: using results in the data to see if we have enough evidence to tell whether a hypothesis (the two means are equal) is wrong (they are not equal) or whether we don't have enough evidence (they may be equal).

This chapter introduces the logic and practice of testing hypotheses. We describe the steps of hypothesis testing and discuss two alternative ways to carry it out: one with the help of a test statistic and a critical value, and another one with the help of a p-value. We discuss how decision rules are derived from our desire to control the likelihood of making erroneous decisions, and how significance levels, power, and p-values are related to the likelihood of those errors. We focus on testing hypotheses about

averages, but, as we show in one of our case studies, this focus is less restrictive than it may appear. The chapter covers one-sided versus two-sided alternatives, issues with testing multiple hypotheses, the perils of p-hacking, and some issues with testing on big data.

The main case study **Comparing online and offline prices: testing the difference** in this chapter is based on the `billion-prices` data we described in Chapter 1, Section 5. This data includes online and offline prices of selected products sold by selected retailers. The data we analyze in this case study is for retailers in the United States. The question we test is whether online and offline prices are the same, on average. In addition, we continue the the case study called **Testing the likelihood of loss on a stock portfolio** we started in Chapter 5. In this case study we ask about the probability of a large loss on an investment portfolio, using the `stock-market` data. Here our question is whether such losses are likely to occur more frequently than a threshold frequency that we can tolerate.

Learning outcomes. After working through this chapter, you should be able to

- formulate a null and an alternative hypothesis that correspond to the question you want to answer;
- think in terms of false positives and false negatives;
- understand the logic of hypothesis testing with the help of a test statistic and a critical value;
- carry out hypothesis testing with the help of a p-value;
- interpret the result of a hypothesis test;
- Understand the additional issues that arise with testing multiple hypotheses from the same data.

1 The logic of testing hypotheses

A hypothesis is a statement about a population, or general pattern. Testing a hypothesis amounts to gathering information from a dataset and, based on that information, deciding whether that hypothesis is false or true in the population, or general pattern.

Thus, **hypothesis testing** means analyzing the data at hand to make a decision about the hypothesis. Two decisions are possible: rejecting the hypothesis or not rejecting it. We reject the hypothesis if there is enough evidence against it. We don't reject it if there isn't enough evidence against it. We may have insufficient evidence against a hypothesis either if the hypothesis is true or if it is not true but the evidence is weak. Rejecting a hypothesis is a more conclusive decision than not rejecting it; we'll discuss this asymmetry more in the next section.

Testing a hypothesis is a way of making an inference, with a focus on a specific statement. As with any kind of inference, we have to assess external validity, too: the extent to which the population, or general pattern, represented by our data is the same as the population, or general pattern, we are truly interested in.

Hypothesis testing is a formal procedure. It is educational to break that procedure into specific steps. In what follows, we'll describe the steps we suggest to follow.

The first step is defining the statistic that would answer our question. Recall that a statistic is anything whose value can be computed from data. Often the statistic that would answer our question is an average, or the difference between two average values. In fact, we'll focus on averages in this chapter.

For convenience, let's call this statistic s. We are interested in the true value of s, which is its value in the population, or general pattern, our data represents. Let's denote this true value by s_{true}. The value of the statistic in our data is its estimated value, denoted by a hat on top, $\hat{s}$.

For example, you may look at data from an experiment that presented some of your potential customers the old version of an online ad and other customers a new version. You want to know if the new version is different in terms of how many people follow up and visit your website after being presented one of the two versions of the ad. Here the statistic is the difference in the follow-up frequency in the two groups of people. If we denote the follow-up rate among customers who saw the old version as r_{old} and among customers who saw the new version as r_{new}, the statistic we are interested in is $s = r_{new} - r_{old}$

Or, you may use company data on firm growth rates (e.g., in terms of sales), and the age of their managers, to learn if firms with a young manager tend to grow at a different rate than firms with an old manager. Here the statistic is the difference in average growth between firms with older managers and firms with younger managers: $s = \overline{growth}_{youngmanager} - \overline{growth}_{oldmanager}$.

It is good practice to pause at this point and, as a second step, think about external validity. Having defined the statistic that translates our question into something we can compute from our data, we should get a sense of whether, and to what extent, that statistic in the data is generalizable to the situation we care about. This amounts to defining the population, or general pattern, our data represents, and making sure we understand how it compares to the population, or general pattern we are interested in. Fortunately, we don't need the two to be similar in every respect. The question is whether the value of the statistic is the same in the two. However, if there is very little chance that the two are similar it makes little sense to continue with the analysis.

If our data on firms with growth rates and the age of their managers was collected in a country that has very different cultural norms about age than our country, whatever age-related differences we may find in our data would have low external validity for our country. In that case we may decide not to use that data to learn about whether firms with younger versus older managers grow at the same rate in our country. Or, we may uncover that the firms in our data are all very small, whereas we are interested in the growth of medium-sized firms. Here, too, external validity of our findings may be low for the kinds of firms we care about, and we may decide not to go ahead with the analysis using this particular data. On the other hand, we may conclude that external validity may be high enough from the viewpoint of the patterns of association between manager age and firm growth rate, thus the general pattern represented by the firms in our data and what we care about may be similar. In this case we should go ahead with our analysis. As we'll see, it makes sense to return to external validity at the end of the process, to qualify the results if necessary.

In our example of testing the new versus old version of an online ad, we are interested in whether the new version would result in more follow-ups among customers in the future. Thus, we need to think about whether the customers in our experiment are a representative sample of our future customers, and whether the circumstances during the experiment are similar to what future customers would face. To be more precise, the sample and the circumstances need to be similar in the experiment and the future roll-out of the ad in terms of the the new versus old follow-up rate difference. An example when it is satisfied is

We shall discuss the other steps in detail in the following sections of this chapter.

2 A1 Case Study – Comparing online and offline prices: testing the difference

Question, data, statistics, external validity

The question of this case study is whether online and offline prices of the same products tend to be different or the same. We use the `billion-prices` data from the Billion Prices project. We introduced this data earlier, in Chapter 1, Section 5. It contains the online and offline price of the same product measured, ideally, at the same time, in various countries over various years. In this case study we use data from the United States, and we include products with their regular prices recorded (as opposed to sales or discounted prices). This data was collected in 2015 and 2016.

The general question of the case study is whether online and offline prices of the same products tend to be different. The "tend to be" part of the question is vague. We can turn this into various statistics. We choose the average, both because it's an intuitive statistic of central value (Chapter 3, Section 9), and because the tools of statistical inference are best used for the average. Alternatives to the mean of the price differences may be the median or the mode of the online and offline price distributions, but we don't consider those statistics in this chapter.

Thus, our statistic is the average of the difference of the online versus offline price. Each product i has both an online and an offline price in the data, $p_{i,online}$ and $p_{i,offline}$. Let the variable $pdiff$ denote their difference:

$$pdiff_i = p_{i,online} - p_{i,offline} \tag{6.1}$$

Then, the statistic with n observations (products) in the data, is:

$$s = \overline{pdiff} = \frac{1}{n}\sum_{i=1}^{n}(p_{i,online} - p_{i,offline}) \tag{6.2}$$

Note that the average of the price differences is equal to the difference of the average prices; thus this s statistic also measures the difference between the average of online prices and the average of offline prices among products with both kinds of price: $\frac{1}{n}\sum_{i=1}^{n}(p_{i,online} - p_{i,offline}) = \frac{1}{n}\sum_{i=1}^{n} p_{i,online} - \frac{1}{n}\sum_{i=1}^{n} p_{i,offline}$.

The mean difference is USD -0.05: online prices are, on average, 5 cents lower in this data. That's the value of our statistic in the data. Denoting that value with $\hat{s}$,

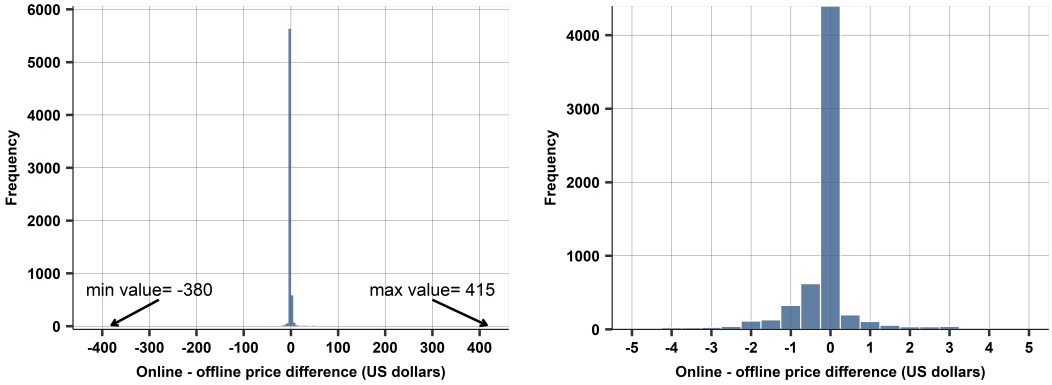
$$\hat{s} = -0.05 \tag{6.3}$$

There is substantial spread around this average: the standard deviation of the price differences is USD 10. But that larger spread is due to a few very large price differences, in both directions. The minimum price difference is USD -380, the maximum is USD +415. At the same time, 5,593 (87%) of the 6,439 products in this data have the difference within ±1 dollars. Furthermore, 4,112 (64%) have the exact same online and offline price.

To get a sense of how the distribution looks, let's look at the histogram. Figures 6.1a and 6.1b show the histograms of the same distribution, except the first one has all observations and shows the distribution over its entire range, while the second one includes observations within a ±5 dollar price difference. Except for how large the range is, the first histogram shows little of the features of the distribution, The second one is more informative. It shows the large spike at zero, which corresponds to the fact that 64% of the products in this data have the exact same online and offline price (see above). In addition, this histogram suggests that the distribution is not perfectly symmetrical. Indeed, 25% of

the products have a lower online price and 11% have a higher online price. This asymmetry shows up in another way: the distribution is slightly skewed with a longer left tail (the median is zero, so the mean–median difference is negative).

Figure 6.1: The distribution of online-offline price differences.



(a) Entire distribution

(b) Between USD -5 and +5

Note: *NB: Mean = -0.05, Std.Dev. = 10.0, Median = 0. For Figure 6.1a N=6439 and for Figure 6.1b N=6200*
Source: `billion-prices` data.

After having defined the statistic of interest and calculated its value from the data, let's review what we know about the source of this data so we can get some sense of whether we can generalize the average price difference in this data to a situation we care about. As we discussed in Chapter 1, Section 5, this data includes 10 to 50 products in each retail store included in the survey (which are the largest retailers in the U.S.A. that sell their products both online and offline). The products were selected by the data collectors in offline stores, and they were matched to the same products the same stores sold online. The products in the data may not represent all products sold at these stores. While there is no particular reason to suspect that online-offline price differences are different for other kinds of products, we can't rule out that possibility. Thus, strictly speaking, the general pattern of the statistic represented by this data is average online-offline price differences in large retail store chains for the population of the kinds of products that data collectors would select with a high likelihood.

That's a rather narrow population. Instead, we would want to further generalize the findings to another general pattern of interest, such as the online-offline price differences among all products in the U.S.A. sold both online and offline by the same retailers. This extra step of generalization is best thought of as a question of external validity. External validity is high if the distribution of price differences among the products represented in our data is similar to that more general pattern among all products that are sold both online and offline in the U.S.A.

Of course, we may care about the online-offline price difference in another country or a different year (the data was collected in the U.S.A. in 2015–6). For that additional step of generalization, the question is whether the general patterns that determine the average price difference in the country and the year of interest are similar to the U.S.A. in 2015–6. For now, we assume that that's the case and continue with our analysis. We shall return to the external validity of our findings at the end of the case study.

3 Null hypothesis, alternative hypothesis

After specifying the statistic that helps answer our question, and after convincing ourselves that the statistic computed from our data may help answer the question in the situation we truly care about, the next step is formally stating the question in hypotheses. More specifically, we need to state two competing hypotheses, of which only one can be true. The first one is the **null hypothesis**, denoted as H_0, and also simply called as the null. The second one is the **alternative hypothesis**, denoted as H_A and called the alternative for short. These hypotheses are formulated in terms of the unknown **true value** of the statistic. This is statistical inference (Chapter 5), so the true value means the value in the population, or general pattern represented by our data. The null specifies a specific value or a range of values, while the alternative specifies other possible values. Together, the null and the alternative should cover all possibilities we are interested in.

With our definition of statistic s, the most widely examined null and alternative hypotheses are the following:

$$H_0 : s_{true} = 0 \tag{6.4}$$
$$H_A : s_{true} \neq 0 \tag{6.5}$$

The null says that the true value of the statistic is zero; the alternative says that it's not zero. Together, these cover all logical possibilities for the true value of the statistic.

It may seem odd to have $H_0 : s_{true} = 0$ when, presumably, we analyze the data because we suspect that the true value of s is not zero (online and offline prices may differ on average; companies with younger managers may grow at a different rate, on average). This seemingly twisted logic comes from the fact that testing a hypothesis amounts to seeing if there is enough evidence in our data to reject the null. It is sometimes said that the null is protected: it should not be too easy to reject it otherwise the conclusions of hypothesis testing would not be strong.

When learning about hypothesis testing, some of us found it helpful to relate its logic to the logic of a criminal court procedure. At court the task is to decide whether an accused person is guilty or innocent of a certain crime. In most modern societies the starting point is the assumption of innocence: the accused person should be judged guilty only if there is enough evidence against their innocence. This is so even though the accused person was brought before court presumably because there was a suspicion of their guilt. To translate this procedure to the language of hypothesis testing, H_0 is that the person is innocent, and H_A is that the person is guilty.

Medical tests are another instructive example. When testing whether a person has a certain medical condition, the null is that the person does not have the condition (healthy), and the alternative is that they have it (sick). The testing procedure amounts to gathering information to see if there is evidence to decide that the person has the condition.

The case when we test if $H_A : s_{true} \neq 0$ is called a **two-sided alternative** as it allows for s_{true} to be either greater than zero or less than zero. For instance, we focus on the difference in online and offline prices, with H_0 being the equality. In such a case we are not really interested if the difference is positive or not, or whether it is negative or not.

The other case is working with a **one-sided alternative**, when we are indeed interested if a statistic is positive or not. The null and the alternative should be set up so that the hypothesis we are truly interested in is in the alternative set. So when we want to know if s_{true} is positive or not, we want to

put $s_{true} > 0$ in the alternative thus, making the null $s_{true} \leq 0$:

$$H_0 : s_{true} \leq 0 \tag{6.6}$$
$$H_A : s_{true} > 0 \tag{6.7}$$

For reasons that we'll see later, the null needs to contain the equality sign, and the alternative should contain the strict inequality.

An example where a one-sided alternative would make sense is the likelihood of large losses on a portfolio in the "Should we worry about a large loss on a stock portfolio?" case study. There we want to know if this likelihood is greater than a threshold value. In other words, we want to know if the difference of the likelihood and the threshold value is greater than zero.

Review Box 6.1 ***Null hypothesis, alternative hypothesis***

- To carry out a test we need to have a clearly stated null hypothesis ("the null," H_0) and a clearly stated alternative hypothesis ("the alternative," H_A).
- H_0 and H_A are statements about the true value of the statistic (the value in the population, or general pattern, represented by our data)
- A typical null-alternative pair with a two-sided alternative is $H_0 : s_{true} = 0$, $H_A : s_{true} \neq 0$.
- A typical null-alternative pair with a one-sided alternative is $H_0 : s_{true} \leq 0$, $H_A : s_{true} > 0$.
- Together, the null and the alternative cover all possibilities (or all interesting possibilities).
- Statistical testing amounts to producing evidence from the data and seeing if it's strong enough so we can reject the null.

4 t-test

After introducing the logic of hypothesis testing, let's dive into the actual process. In this chapter we focus on one specific test, called the **t-test**. This testing procedure is based on the **t-statistic** that we'll introduce in this section.

Following the logic of hypothesis testing, we start from the assumption that the null is true and thus $s_{true} = 0$. Then we look at the evidence to see if we want to reject this null or maintain our assumption that it's true. Intuitively, the evidence we look for is how far the estimated value of the statistic, $\hat{s}$, is from zero, its hypothesized value. We reject the null if the distance is large – i.e., if the estimate is far from its hypothesized value. Conversely, we do not reject the null if the estimate is not very far – i.e., when there is not enough evidence against it. How far is far enough for rejecting requires a measure of the distance. That measure is the t-statistic that we'll introduce here.

More generally, the t-statistic is a **test statistic**. A test statistic is a measure of the distance of the estimated value of the statistic from what its true value would be if H_0 were true.

Consider $H_0 : s_{true} = 0, H_A : s_{true} \neq 0$, where s is an average, or the difference of that average from a number, or a difference of two averages. The null and alternative are about the true value of the statistic: s_{true}. Our data contains information to compute the estimated value of the statistic: $\hat{s}$. The

t-statistic for this hypotheses uses the estimated value $\hat{s}$ plus its standard error:

$$t = \frac{\hat{s}}{SE(\hat{s})} \tag{6.8}$$

When $\hat{s}$ is the average of a variable x, the t-statistic is simply

$$t = \frac{\bar{x}}{SE(\bar{x})} \tag{6.9}$$

When $\hat{s}$ is the average of a variable x minus a number, the t-statistic is

$$t = \frac{\bar{x} - number}{SE(\bar{x})} \tag{6.10}$$

When $\hat{s}$ is the difference between two averages, say, $\bar{x}_A$ and $\bar{x}_B$, the t-statistic is

$$t = \frac{\bar{x}_A - \bar{x}_B}{SE(\bar{x}_A - \bar{x}_B)} \tag{6.11}$$

In the first two formulae, the denominator is the SE of average x, something that we know how to estimate in two ways, by bootstrap or by formula (Chapter 5, sections 10 and 12.) In the last formula we have something different: the SE of the difference of two averages. We haven't discussed ways to estimate it, although the intuition behind the bootstrap suggests that that may be an appropriate method. Moreover, there is a formula for it as well that makes use of the potentially different standard deviations of x in the two groups. Fortunately, we don't need to learn that formula as all statistical software we use has it built in and uses it when carrying out such a test.

The sign of the t-statistic is the same as the sign of $\hat{s}$. If $\hat{s}$ is positive, the t-statistic is positive; if $\hat{s}$ is negative, the t-statistic is negative. The magnitude of the t-statistic measures the distance of $\hat{s}$ from what s_{true} would be if the null were true. The unit of distance is the standard error. For example, the t-statistic is one (or negative one) if $\hat{s}$ is exactly one standard error away from zero.

Review Box 6.2 ***t-test***

- The t-test is a procedure to decide whether we can reject the null.
- The t-test is designed to test hypotheses about the mean of a variable.
- The t-statistic transforms the original statistic of interest into a standardized version.
- When the question is whether the mean of a variable is equal to a number, the t-statistic is $t = \frac{\bar{x} - number}{SE(\bar{x})}$
- When the question is whether the mean of a variable is the same in group A and group B, the t-statistic is $t = \frac{\bar{x}_A - \bar{x}_B}{SE(\bar{x}_A - \bar{x}_B)}$

5 Making a decision; false negatives, false positives

The following step is making a decision: either rejecting the null or not rejecting it. In hypothesis testing this decision is based on a clear rule specified in advance. Having such a decision rule makes the decision straightforward.

A clear rule also makes the decision transparent, which helps avoid biases in the decision. That's important because biases may occur, sometimes unconsciously. Unfortunately, we humans are often tempted to use evidence to support our pre-existing views or prejudices. If, for example, we think that firms with younger managers tend to grow faster, we may pay more attention to the evidence that supports that belief than to the evidence against it. In particular, we may be tempted to say that the estimated growth rate difference $\hat{s}$ is large enough to reject the null, because we believe that the null isn't true. Clear decision rules are designed to minimize the room for such temptations.

To be specific, the decision rule in statistical testing is comparing the test statistic to a **critical value**. The critical value thus tells us whether the test statistic is large enough to reject the null. The null is to be rejected if the test statistic is larger than the critical value; the null is not to be rejected if the test statistic isn't larger than the critical value. To be transparent, we need to set the critical value before looking at the test statistic. As that is impossible to document, the practice is to use a commonly accepted critical value.

But how to select the critical value? Or, more to the point, why is a particular critical value commonly used for a test?

To answer that question, we first need to understand that the critical value reflects a preference for how conservative we want to be with the evidence. If we set the critical value high, we require that far means very, very far. That makes rejecting the null hard. We often say that a choice for such a critical value is a conservative choice. If we set the value low, we accept not-so-far as far, too. That makes rejecting the null easier. That would be a lenient choice. The process of setting the critical value involves a trade-off between being too conservative or too lenient. To explore that trade-off we need to dig a little deeper in the consequences of the decision.

We can make one of two decisions with hypothesis testing: we reject the null or we don't reject the null. That decision may be right or wrong. There is no way to know if we are right or wrong when we make the decision. Nevertheless, we can think about these possibilities. We can be right in our decision in two ways: we reject the null when it is not true, or we do not reject the null when it is true. We can be wrong in our decision in two ways, too: we reject the null even though it is true, or we do not reject the null even though is not true. Let's focus on the two ways of being wrong.

We say that our decision is a **false positive** if we reject the null when it is true. The decision is "false" because it is erroneous. It is "positive" because we take the active decision to reject the protected null. The language may seem familiar from medical testing: a result is "positive" if it suggests that the person has the condition that they were tested against (rather negative news in most cases...).

We say that our decision is a **false negative** if we do not reject the null even though it's not true. This decision is "negative" because we do not take the active decision and, instead, leave the null not rejected. In medical testing a result is "negative" if it suggests that the person does not have the condition that they were tested against (which often is rather good news...).

In formal statistical language, a false positive decision is called a **Type-I error** and a false negative decision is called a **Type-II error**.

Table 6.1 summarizes these possibilities.

Table 6.1: Two ways to be right, and two ways to be wrong when testing a null hypothesis

	Null Hypothesis is true	Null Hypothesis is false
Don't reject the null	True Negative (TN) Correct	False Negative (FN) Type II error
Reject the null	False Positive (FP) Type I error	True Positive(TP) Correct

False positives and false negatives are both wrong, but they are not equally wrong. The way we set up the null and the alternative, wrongly rejecting the null (a false positive) is a bigger mistake than wrongly accepting it (a false negative). Thus, the testing procedure needs to protect the null: we want to reject it only if the evidence is strong. Therefore, the the critical value is to be chosen in a way that makes false positives rare. But that is a balancing act. It would be easy to completely avoid false positives just by never making a positive decision: if we never reject the null we could not possibly reject it wrongly. That is, of course, not a useful procedure. Otherwise, false positives are always a possibility. The usual solution is to go for a very small chance for false positives.

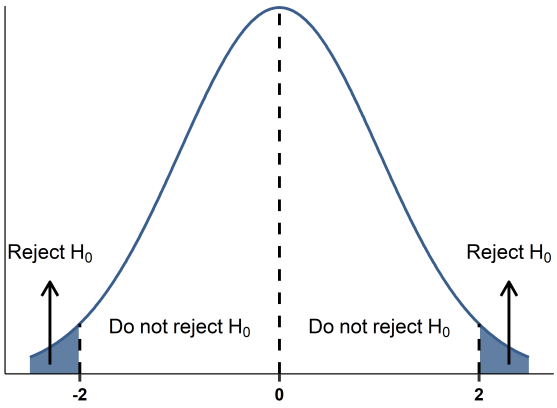
Let's focus on a t-test for a null with a two-sided alternative, $H_0 : s_{true} = 0$, $H_A : s_{true} \neq 0$. (We shall return to cases with a one-sided alternative later, in Section 9.) A commonly applied critical value for a t-statistic is ± 2: reject the null if the t-statistic is smaller than -2 or larger than $+2$; don't reject the null if the t-statistic is between -2 and $+2$. That way the probability of a false positive is 5%.

Why? Answering this question is not very easy, but we know everything already to do so. We can calculate the likelihood of a false positive because we know what the sampling distribution of the test statistic would be if the null were true. Recall, from Chapter 5, Section 3, that the sampling distribution of a statistic is its distribution across repeated samples of the same size from the same population. The sampling distribution of an average is approximately normal, its mean is equal to the true mean, and its standard deviation is called the standard error. The t-statistic has the average in its numerator, so that its distribution is also approximately normal, but its standard deviation is one because the denominator is the SE of $\hat{s}$. (It turns out that with fewer than 30 observations, the normal approximation to the distribution of the t-statistic is not very good. Instead, under some specific circumstances, the distribution is closer to something called a t-distribution – hence the name of the t-statistic.)

Here we ask how the sampling distribution would look if the null hypothesis were true. In that case the distribution of the t-statistic would be standard normal: a normal distribution with mean zero and standard deviation of one. The mean would be zero because $s_{true} = 0$ if the null were true. The standard deviation would be one one because the t-statistic is standardized (Chapter 3, Section 9): it has the SE in the denominator.

By the properties of the standard normal distribution, the probability that the t-statistic is less than -2 is approximately 2.5%, and the probability that it is greater than $+2$ is also about 2.5%. Therefore the probability that the t-statistic is either below -2 or above $+2$ is 5% if the null is true. And that 5% is the probability of false positives if we apply the critical values of ± 2 because that's the probability that we would reject the null if it was true. Figure 6.2 illustrates these probabilities.

Figure 6.2: The probability of a false positive



The sampling distribution of the test statistic when the null is true, and the probability of a false positive when the decision is to reject the null if the test statistic is larger than 2 in absolute value.

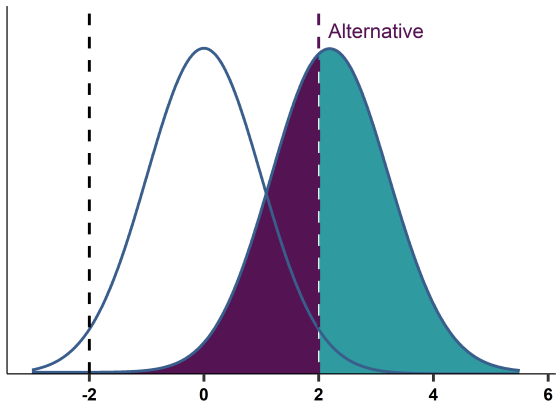
Based on the same argument, we can set other critical values that correspond to different probabilities of a false positive. If we make the critical values -2.6 and $+2.6$, the chance of a false positive is 1%. With critical values -1.6 and $+1.6$ it is 10%. While these other critical values are also used sometimes, the ± 2 critical value and the corresponding 5% chance is the conventional choice. That choice means that we tolerate a 5% chance for being wrong when rejecting the null.

So this is how data analysts avoid biases in their decision when testing hypotheses. They use the same ± 2 critical value of the t-test regardless of the data and hypothesis they are testing. That gives a 5% chance of a false positive decision.

After the probability of a false positive, let's turn to false negatives. Fixing the chance of false positives affects the chance of false negatives at the same time. Recall that a false negative arises when we don't reject the null even though it's not true. We don't reject the null when the t-statistic is within the critical values. Can we tell how likely it is that we make a mistake by not rejecting it?

Figure 6.3 illustrates how we can answer this question. It shows the two kinds of mistakes we can make when applying a threshold for rejecting the null. The upper part shows what happens if our test statistic is large, beyond the critical value, even though the null is true. This is the probability of a false positive, which we discussed above. The bottom part is new. It shows the opposite mistake: what happens if our test statistic is not large enough, within the critical value, even though the null is not true but, instead, the true value of the statistic is something different. That's the probability of a false negative.

Figure 6.3: The probability of a false positive and a false negative



The sampling distribution of the test statistic when the null is true (upper panel) versus the alternative being true (lower panel), and the corresponding probabilities of making a mistake. Upper panel: a false positive (rejecting the null when it's true); Lower panel: a false negative (not rejecting the null when the alternative is true).

It turns out that the chance of a false negative depends on two things: how far the true value is from the value defined in the null hypothesis, and how large the sample is. We can see that if we think how Figure 6.3 would change under various scenarios. First, the further away the true value is from zero, the smaller that bright red region. Thus the smaller the probability of a false negative. Second, the larger the sample, the smaller the bright red region, again. So the probability of a false negative is smaller.

In the next section we'll build on all this and introduce an alternative way to make the decision – one that is easier in practice and involves fewer steps. Before that, let's introduce a few more concepts that are often used when testing hypotheses.

The probability of a false positive is called the **size of the test**. The maximum probability that we tolerate is the **level of significance**. When we fix the level of significance at 5% and end up rejecting the null, we say that the statistic we tested is significant at 5%. The probability of avoiding a false negative is the **power of the test**. Thus, we usually fix the level of significance at 5% and hope for a high power (i.e., a high probability of avoiding a false negative), which is more likely the larger the sample and the further away the true value is from what's in a null.

As we have seen, the power of the test is larger the further away the truth is from what we stated in the null and the more observations we have in our data. The power of the test means our ability to tell if the null is in fact not true. With high power (e.g., because of many observations) we have a good chance that we can reject the null if it's not true. In contrast, with low power (e.g., because of too few observations), we don't have a good chance to reject the null if it's not true.

Let's consider the example of the experiment that shows some potential customers the old version of an online ad and other potential customers the new version of the ad. Here the question is whether the new ad has a different effect. The statistic, as we discussed earlier, is the difference between the fraction of people who follow up to the product website after being presented the ad in the two groups.

Translating the question into a null and a two-sided alternative, the null here is no difference, and the alternative is a difference. A false negative would be to decide that the follow-up rates are the same in the two versions when in fact they are different. If we have many subjects, a well designed experiment has high power and would reject the null of no difference if there is indeed a difference. However, with too few subjects, even an otherwise well designed experiment has low power. Thus, it would not reject the null even if the two versions are different in the follow-up rates. What's too few and what's many depends on a lot of things. We'll return to this question when we discuss the design of experiments in Chapter 20.

Review Box 6.3 ***False positives, false negatives, size, significance, and power***

- A false positive is a decision to reject the null hypothesis when it is in fact true.
- A false negative is a decision not to reject the null hypothesis when it is in fact not true.
- The level of significance is the maximum probability of a false positive that we tolerate.
- The power of the test is the probability of avoiding a false negative.
- In statistical testing we fix the level of significance of the test to be small (5%, 1%).
- Tests in larger data with more observations have more power in general.

6 p-value

The **p-value** makes testing substantially easier. In effect it saves us from the need to calculate test statistics and specify critical values. Instead, we can make an informed decision based on the p-value only.

The p-value informs us about the probability of a false positive. To be more precise, the p-value is the probability that the test statistic will be as large as, or larger than, the critical value, if the null hypothesis is true. Thus, the p-value is the smallest significance level at which we can reject H_0 given the value of the test statistic in the data. It is sometimes denoted by lowercase p or $P >$ test statistic, as in "$P > |t|$" for a two-sided t-test. Because the p-value tells us the smallest level of significance at we can reject the null, it summarizes all the information we need to make the accept/reject decision. The p-value depends on the test statistic and the sampling distribution of the test statistic.

Let's say that we get a p-value of 0.04. That means that the smallest level of significance at which we can reject the null is 4%. In this case we should reject if our initially set target was 5% or higher. If, on the other hand, our initially set target was 1%, we shouldn't reject the null if we get a p-value of 0.04. If, instead, we get a tiny p-value of, say, 0.0001, we should reject the null if our initial target level is 10%, 5%, 1%, or even 0.1%.

To be transparent, we should set the level of significance before carrying out the test, as we explained earlier in Section 5. In practice, that means applying the convention of 5% or 1%.

The great thing about the p-value is that statistical software calculates it for us. Of course, to do so, we have to define the statistic and the null and alternative hypotheses. Then we just look at the p-value and reject the null if the p-value is below the pre-set level of significance (5% or 1%), and we don't reject it otherwise.

Review Box 6.4 ***The p-value***

- The p-value is the smallest significance level at which we can reject H_0 given the value of the test statistic in the sample.
- More intuitively, it is the probability that the test statistic will be as large as, or larger than, the critical value, if the null hypothesis is true.
- The p-value summarizes all the information in the data we need to make a decision when testing a hypothesis.
- Reject the null if the p-value is less than the level of significance we set for ourselves (say, 5%). Don't reject the null otherwise.

7 A2 Case Study – Comparing online and offline prices: testing the difference

In our case study, the hypotheses are about whether the online versus offline price difference is zero on average in the population of products represented by the data.

$$H_0 : s_{true} = \overline{pdiff}_{true} = 0 \tag{6.12}$$

$$H_A : s_{true} = \overline{pdiff}_{true} \neq 0 \tag{6.13}$$

Following usual practice, let's fix the level of significance at 5%. By doing so, we tolerate a 5% chance for a false positive. That is, we allow ourselves a 5% chance to be wrong if we reject the null hypothesis of zero average price difference. A 5% level of significance translates to ± 2 bounds for the t-statistic.

The value of the statistic in the data is -0.054. Its standard error is 0.124. Thus the t-statistic is

$$t = \frac{-0.054}{0.124} = -0.44 \tag{6.14}$$

This is well within ± 2. Thus we don't reject the null hypothesis of zero difference. Doing so would be inconsistent with our goal to keep the probability of false positive below 5%. Thus we can't reject the hypothesis that the average price difference is zero in the population of products represented by the data.

So our decision is not to reject the null. That can be the result of two things. One: the null is true indeed, and online and offline prices are equal, on average, in the population of products represented by the data. Two: the null is not true, only we didn't have enough evidence against it in this data. While we can't tell which of these two is the case, we have evidence to help. Notably, note that our dataset is large, with over six thousand observations. Thus, the power of the test is likely high. A caveat is that the power may be low against very small alternatives: if the true average price difference were not zero but very close to zero. Taking these together, this data would quite likely produce a large enough test statistic that would make us reject the null if the true average price difference were quite different from zero. Instead, we got a small test statistic, and we couldn't reject. Thus either the null is true or, if it isn't true, the true price difference is tiny. That means that we can be quite certain that the average online-offline price difference is zero or very small in the population of products represented by our data.

We arrive at the same conclusion with the help of the p-value. The software calculated that the p-value of the test is 0.66. That means that the smallest level of significance at which we can reject the null is 66%. In other words, the chance that we would make a mistake if we rejected the null is at most 66%. So we don't reject the null as this 66% is well over the 5% level of significance we have chosen. This is of course the same decision that we made with the t-statistic and the critical value.

Finally, note that we would arrive to a similar conclusion if, instead of carrying out a test, we just looked at the 95% confidence interval of the average price difference. We have all the ingredients to calculate that CI: average price difference in the data is -0.054, with SE=0.124. Thus the 95% CI is

$$CI_{95\%} = [-0.054 - 2SE, -0.053 + 2SE] = [-0.30, +19] \tag{6.15}$$

This CI contains zero. So the value of the true average price difference (the average price difference in the population of products represented by our data) may very well be zero.

In fact, the CI has more information than a p-value or a t-statistic. It also shows that it's very unlikely that the true average price difference is more than -30 cents or +20 cents.

This concludes our case study. From it, we learned that online and offline prices are, on average, very similar among the products represented by our data. Indeed, we can't reject our hypothesis that they are the same on average.

What does that imply for products we are interested in? For example, we can conclude that differences between online and offline prices of the same products are likely zero on average among all products in the U.S.A. that were sold by large retailers in the U.S.A. in 2915–6. For this step of generalization, we need to assume that the products sampled in this dataset represent all products in those stores, at least in terms of the average online-offline price difference. Also, keeping in mind that we restricted our data to regular prices (without sales or promotions), we would need to assume that the average zero difference holds for sales and promotion if we wanted to generalize to them. Moreover, we need to assume that the retailers sampled in this data represent all larger retailers in the U.S.A., at least in terms of the difference of their online versus offline prices. Finally, if we are interested in another year, say this year, then we would need to assume that the pattern is stable enough from 2015–6 to this year. If we want to make further generalizations, we would need to assume more. For example, these results may have a lower external validity if we wanted to compare online and offline prices of the same products sold by different retailers. Similarly, they may have low external validity for other countries.

This case study illustrated how to carry out a hypothesis test. First, we translated a question (do online and offline prices tend to be the same?) into a statistic (the average of the difference of online and offline prices of the same products). Then we formed a null hypothesis and an alternative hypothesis; here that was a two-sided alternative. We calculated the statistic from the data, together with its standard error, and divided the two to calculate the t-statistic. We compared the value of this t-statistic to the critical values and decided that we couldn't reject the null. We repeated this decision with the help of the p-value that we calculated directly using the software. We also looked at the 95% CI and arrived at the same decision. We then thought about what that decision really meant, and then, the extent to which we could generalize that decision from the population of products represented by the data to the population of products we were interested in.

Perhaps the most important take-away from this case study is that carrying out hypothesis testing is actually not that hard. We had to follow some steps, but most of the work was done by the computer. A harder task is to define the statistic of interest, the null, and the alternative. And the even harder task is to understand what rejecting or not rejecting the null really means, and whether we could generalize that decision to the situation we really care about.

8 Steps of hypothesis testing

The previous sections described all the ingredients to carry out a hypothesis test. We have illustrated how to do that in our case study. But there are specific cases that call for doing things a bit differently. Before moving to those, let us summarize the steps of hypothesis testing.

Review Box 6.5 ***Steps of hypothesis testing***

1. Define the statistic that corresponds to your question.
2. Initial assessment of external validity: general pattern of interest, general pattern represented by the data, does it make sense to continue?
3. State the null and alternative hypotheses.
4. Choose a critical value. In practice, that should correspond to 5% – present an argument if you want to choose some other value.
5. Calculate the test statistic from the data.
6. Make a decision based on the calculated test statistic and critical value.
7. Alternatively to the previous three steps, calculate the p-value and make a decision by comparing it to a pre-set level of significance, that is 5% in practice – again, make an argument if you want to choose some other value.
8. Interpret the results.
9. Final assessment of external validity. Would the same decision be fine for the population, or general pattern, of interest?

9 One-sided alternatives

When we discussed false positives and false negatives we focused on two-sided hypotheses, with $H_0 : s_{true} = 0$ against $H_A : s_{true} \neq 0$. And that's what we had in our case study, too. However, as we mentioned earlier in Section 3, sometimes we want to do a one-sided test instead.

Testing a null hypothesis against a one-sided alternative means having an inequality in H_A. Most often, it goes with having an inequality in the null hypothesis instead of an equality. The two most frequent examples are $H_0 : s_{true} \leq 0$ against $H_A : s_{true} > 0$ and $H_0 : s_{true} \geq 0$ against $H_A : s_{true} < 0$.

In order to reject the null $s_{true} \leq 0$, we need to reject each and every possible value in the hypothesized interval in favor of $s > 0$. The hardest of those possible values to reject would be $s = 0$. This is because if we can reject $s = 0$, we can reject all $s < 0$ values, too. For that reason we need the equality to be always part of the null. Otherwise we would not know the value under the null that is closest to the boundary. Therefore, in practice, testing $H_0 : s_{true} \leq 0$ against $H_A : s_{true} > 0$ is the same as testing $H_0 : s_{true} = 0$ against $H_A : s_{true} > 0$. Similarly, testing $H_0 : s_{true} \geq 0$ against $H_A : s_{true} < 0$ is the same as testing $H_0 : s_{true} = 0$ against $H_A : s_{true} < 0$.

Having only one of the inequalities in the alternative leads to focusing on one side of the test statistic

only. Here we don't care how far the estimate is from the hypothesized value in one direction; we only care about the deviation in the other direction. Focusing on deviations in one direction means that we care about one half of the sampling distribution of the test statistic. The question therefore is the probability that the test statistic would fall outside the critical value in the "wrong direction," given that it deviates from the hypothesized value in that "wrong direction." With $H_0 : s_{true} \leq 0$ against $H_A : s_{true} > 0$, we care about whether $\hat{s}$ is large positive enough in order to reject the null; if it is negative, we don't reject the null whatever its size.

The probability of a false positive is smaller in this case. We don't reject the null if the test statistic falls in the region that is specified in the null hypothesis – we may reject it only if it falls in the other region. Thus, we make a false positive decision only half of the time. With the standard t-test of a two-sided hypothesis, the p-value can be thought of as the sum of two probabilities: that the t-statistic is below the negative critical value and that the t-statistic is above the positive critical value. When we are testing a one sided hypothesis instead, we never reject the null if the t-statistic is on the side that's contained in the null. We may reject it only if it is on the other side. The probability of a false positive decision here is therefore half of the probability of a false positive in the corresponding two-sided test.

Therefore, the practical way to test a one-sided hypothesis is a two-step procedure.

1. If the test statistic is in the region of the null, don't reject the null. This happens if $\hat{s}$ is in the region of the null (e.g., $\hat{s} < 0$ for $H_0 : s_{true} \leq 0$ against $H_A : s_{true} > 0$);
2. If the test statistic is in the region of the alternative, proceed with testing the usual way with some modification. Ask the software to calculate the p-value of the null hypothesis of the equality (for example, $H_0 : s_{true} = 0$ if the true null is $H_0 : s_{true} \leq 0$) and divide the p-value by two.

10 B1 Case Study – Testing the likelihood of loss on a stock portfolio

One-sided test of a probability

To illustrate how to carry out a one-sided hypothesis test, let's go back to our case study in Chapter 5, on stock market returns. The question there was the probability of a loss on our portfolio of at least 5% from one day to another. Our portfolio was the company stocks contained in the S&P500 index in the U.S.A. We used the `stock-market` data on the daily value of the S&P500 stock market index in 11 years, from 2006 through 2016, with 2519 observations. Recall from Chapter 5, Section 13, that the proportion of such losses was 0.5 percent of the days in the data.

For this illustrative example, let's say that we can tolerate such losses as long as they are not more frequent than 1% of the days. That 1% chance is larger than the 0.5% chance we see in our data. Thus, we would want to know if we can safely reject that larger chance. And we are interested in what would happen next year. Assume for the sake of this illustrative example that external validity is high, and let's focus on the general pattern represented by our 11 years of data.

The statistic is the proportion of days with a loss of 5% or more. This is a relative frequency or, in a more abstract sense, a probability. Up to this point we looked at averages: we introduced the t-test for averages, and we used our knowledge of the sampling distribution of averages. Can we use that knowledge to test a probability?

The answer is yes. In fact, we already answered this question earlier in Chapter 5, Section 6. The

proportion of days with a 5%+ loss is nothing else than the average of a binary indicator variable. This binary indicator variable, which we may call $lossof5$, is one on days with a loss of 5% or more, and it is zero on other days. The average of this variable is the proportion of the times it has the value one. In our data, that average is 0.005, which corresponds to a frequency of 0.5% of the days.

We want to see its difference from the 1% frequency that we can tolerate. So our s here is the proportion of days with such a loss in our data minus 0.01: $s = \overline{lossof5} - 0.01$ where $lossof5$ is our binary variable.

Here the tricky part is translating our question into an appropriate null hypothesis and an appropriate alternative hypothesis. We want to make sure that the proportion of large losses is less than 1%. So this should go into the alternative: $H_A : s_{true} < 0$. The null should cover all other possibilities, including the equality sign: $H_0 : s_{true} \geq 0$. This is one-sided testing.

In our data, we found a proportion of days of such losses to be 0.005, so $\overline{lossof5} = 0.005$. Then the statistic calculated from our data is $\hat{s} = 0.005 - 0.01 = -0.005$. The proportion of days with large losses is half of a percent in our data. It's smaller than the threshold value of one percent, so the test statistic is less than zero.

Following the decision rules we outlined, first note that the test statistic is outside the region of the null. The region of the null is positive numbers and zero; the test statistic is negative. Then, the next step is to make the software calculate the p-value for the corresponding two-sided test ($H_0 : s_{true} = 0\, H_A : s_{true} \neq 0$). That p-value is shown to be 0.0000, which means that it's less than 0.00005. According to step two above, we can further divide this by two, and that would lead to an even smaller p-value.

In any case, the p-value is very small so we reject the null. The conclusion: we can safely say that large losses occur less frequently than one percent of the days. This is true for the general pattern behind the 11-year-old history of the S&P500 in our data. It is true for the general pattern behind next year, too, if external validity is high. But the p-value is very small, so we can draw the same conclusion even if we allow for some extra uncertainty due to potentially not-super-high external validity.

Note, finally, that the 95% CI would have given us a similar conclusion. The SE we calculated to be 0.0014 so the 95% CI of large losses is $[0.22\%, 0.78\%]$. This CI doesn't contain the threshold value of 1% we are concerned about. So, with high confidence we can say that we can expect daily losses of 5% or more to occur on less than 1% of the days. Of course, whether the same general pattern will prevail in the future is an open question, so we may qualify this decision.

11 Testing multiple hypotheses

Our last technical topic in this already pretty technical chapter deals with how to interpret the results of tests when we carry out many of them using the same data. That situation is called **testing multiple hypotheses**.

The simplest decision rule to testing a hypothesis consists of calculating a p-value from the data and comparing it to a pre-set level of significance. So far we have considered cases with one pair of hypotheses: a null and an alternative. Often, though, data analysis involves testing multiple null hypotheses, each with its corresponding alternative. For example, we may be interested in the correlation of firm size and various dimensions of the quality of management, using the 18 different scores in the World Management Survey, or we may want to see if that correlation is positive in each industry separately (see our case study in Chapter 4, Section 11). Or we may be interested in losses on a portfolio of

multiple magnitudes: 2%, 3%, 4%, 5%, etc. Or we may be interested in whether online and offline prices are the same, on average, in various groups of products. So, by asking slightly more detailed questions that correspond to our original question, we can easily end up with dozens of hypotheses to test.

Testing multiple hypotheses is an issue because the statistical theory of testing, and the corresponding tools, have been developed for testing a single null hypothesis. The way the p-value is calculated makes it capture the likelihood of a false positive for a single null hypothesis. When testing multiple hypotheses, we may be tempted to use these tools, such as the p-value, for each hypothesis. But that would lead to misleading results.

To understand why using a decision rule, which is designed for a single test, would lead to misleading conclusions when applied to multiple tests, consider a situation in which we are testing as many as 20 hypotheses. An example is whether the online-offline price differential is zero in each of 20 groups of products. Assume that all of those 20 null hypotheses are true. If we set the level of significance for a single hypothesis, we allow for a 5% chance to be wrong when rejecting the null. That means that we are allowed to be wrong 5% of the time by rejecting the null. Thus, we can expect that, following that rule, we would reject the null about once when we test our 20 null hypotheses. In practice, that would mean that we can expect to see a p-value less than 0.05 in 1 out of the 20 tests. Naively we may pick that one null hypothesis and say that there is enough evidence to reject it, and say that we can't reject the other 19. But that would be a mistake: we started out assuming that all 20 nulls are true. Yes, rejecting the null when it's true is a risk we always have. But at the heart of the procedure is to make sure that happens rarely – here 5% of the time. And, yet, here we can expect that to happen with a much higher chance.

That is because the p-value for a single null hypothesis is designed to calculate the likelihood that we were wrong to reject that particular null hypothesis if tested by itself. In multiple hypothesis testing, the question is the likelihood that any one of the true nulls will be wrongfully rejected. These are just two different probabilities.

There are various ways to deal with probabilities of false positives when testing multiple hypotheses. Multiple testing is a rather complicated part of statistics, and its tools are beyond the scope of this textbook. In any case, it is good practice to be especially cautious and conservative when testing multiple hypotheses and to use conservative criteria (such as a 1%, or even lower, level of significance instead of the customary 5%) for rejecting null hypotheses.

12 A3 Case Study – Comparing online and offline prices: testing the difference

Testing multiple hypotheses

To illustrate the issues with testing multiple hypotheses, let's go back to our case study of online and offline prices of the same products. Besides pooling all products as we did so far, we may be interested in whether online and offline prices are the same, on average, for each retail store in the data. The U.S.A. sample has 16 retailers. Thus testing whether the average price differential is zero in each is testing multiple hypotheses – 16 of them. One way to express these 16 hypotheses in a condensed format is denoting each store by index j

$$H_0 : s_{j,true} = \overline{pdiff}_{j,true} = 0 \tag{6.16}$$

$$H_A : s_{j,true} = \overline{pdiff}_{j,true} \neq 0 \tag{6.17}$$

The next table shows the 16 p-values that we get if we carry out each test one by one.

Table 6.2: Multiple test of price differentials

Retailer ID:	44	45	46	47	48	49	50	51
Diff	3.74	-1.2	-0.43	0.05	0.42	2.41	0.61	0.28
p-value	0.04	0.22	0.00	0.10	0.04	0.20	0.10	0.06
Retailer ID:	53	54	56	57	58	59	60	62
Diff	-0.97	-0.03	-0.49	0.93	-0.17	-0.53	-0.14	1.36
p-value	0.01	0.80	0.04	0.00	0.00	0.00	0.70	0.12

Note: *The table reports 16 tests – one for each retailer.*
Source: `billion-prices` data, U.S. retailers in 2015–6, sales and promotions excluded.

If the null hypothesis were true in each store, we would expect about one of the p-values below 0.05 (5% of 16 is slightly less than one). However, the p-value is less than 0.05 in 8 of the 16 hypotheses. This is more than what we would expect if there was zero price difference in each store.

As an alternative way to assess multiple hypotheses, let's apply a more conservative criterion for rejection, say, a 1% level of significance. That would mean that we would reject four of the nulls.

Thus, we may conclude that the average online-offline price difference is not zero in some of the retail chains in the data, even if we factor in the fact that we are testing multiple hypotheses.

How come, then, that we could not reject the assertion that it's zero when averaged over all retailer stores? Mechanically, this is because the average price difference is negative in some stores and positive in others. Of the eight stores with p-value that is less than 0.05, five have a negative price difference and three a positive difference.

To make sense of what we see, let's ask why would some retail stores charge higher prices online while other stores charge higher prices offline. A likely answer is that they don't in fact do that. More likely, what we see is that there are errors in the data, or differences for some products due to specific reasons. Those errors or deviations are sometimes positive, sometimes negative, and their magnitude is large for some retailers and small for others. And, crucially, they appear to average out when considering all 16 stores, but not so much for some of the stores when they are considered separately.

Recall from Chapter 1 that data collectors in retail stores registered offline prices first and then they sent images to data collectors in the survey center so they can collect the corresponding online prices. As we dig deeper in the data collection process, we learn that in our data, 25% of the products had their online price collected on the same day; the remaining 75% were collected within two to eight days. As prices may change over time, the larger the time lag between the registration of the two prices, the more likely their difference is due to an error.

We can check this by going back to the raw data. Indeed, the standard deviation of price differences is a lot smaller for products whose online price is registered at the same time than for products with a time lag in the registration. That is in line with more error for the latter. Moreover, it turns out that the proportion of products with online prices registered later is larger among the retailers with p-values below 0.05 in the previous table. Thus their measured price differences are more likely to be due to error. As a data exercise, you will be invited to re-do the analysis including products whose offline and online price was evaluated on the same day.

Going back to the main story, what do these findings imply for our original question of whether online and offline prices are the same, on average? Our question is about the general pattern represented

by the data that includes all 16 retailers. Averaging over the retailers may reduce potential errors in data collection. In principle, the hypothesis would also allow for some retailers charging higher prices online while others do the reverse. Thus the conclusions of the two kinds of tests (single test on the full set vs. multiple tests on individual retailers) are not in conflict. The multiple tests either reveal artificial features of the data (measurement error), or they may add new information on differences in pricing by retailer. In this case study the former is more likely.

13 p-hacking

Earlier, we discussed why it is good practice to use a conventional level of significance (5% or 1%) when making a decision. This rule eliminates a potentially arbitrary decision of how large a likelihood we would tolerate for the false positive decision. And, as we argued, that's good because allowing for an arbitrary decision may make data analysts make testing decisions that are biased towards their preconceptions. For example, if we think that firms with younger managers tend to grow at equal rates, and we test the null of no difference and arrive at a p-value of 0.04, we may decide that we can't tolerate a 4% chance of for the false positive and decide not to reject the null. That decision would reinforce our belief of no difference. In contrast, if we believed that the firms with younger managers tend to go at different rates than firms with older managers, and we see the same p-value of 0.04, we may decide that this 4% chance is something we can live with and reject the null, reinforcing our belief of differences in growth rates. All that controversy can be avoided if we adopt a conventional 5% threshold. That would make us reject the null and say the growth rates are different, regardless of what our prior belief was.

Unfortunately, however, data analysts can make other decisions during their analysis to arrive at test results that would reinforce their prior beliefs. For example, we may face some issues during data cleaning – e.g., what to do with extreme values (see Chapters 2 and 3). For example, when analyzing growth rates of firms by the age of their managers, some firms may have unusually large positive growth in the data. Suppose, moreover, that it's not obvious whether those observations with extreme values should be kept or discarded (e.g., those extremely high growth rates may not be the results of mergers or acquisitions, but may be errors). Then, suppose that we end up rejecting the null of equal average growth rates if we decide to keep the observations with extreme values, but we end up not rejecting the same null if we discard those observations. Then, we, as humans, may prefer the version that reinforces our prior beliefs, and go ahead with the data cleaning decision that produces the test results that are in line with those beliefs. A specific, but apparently frequent, case is when a data analyst prefers to show "significant" results – i.e., results when the null is rejected. Recall that rejecting the null is the stronger, more informative decision. Thus, a data analyst may prefer to show the results of tests that lead to rejecting the null, because those are "more conclusive."

These are examples of **p-hacking**. In a narrower sense, p-hacking refers to showing only results of hypothesis tests that suggest one decision, when different choices in the process of data wrangling or data analysis would lead to a different decision for those hypotheses. Often, this means presenting positive decisions, where the null is rejected, and not presenting negative results, where the null isn't rejected. Hence the name p-hacking: carrying out tests with different versions of the data and the analysis, and presenting only the positive results. In a broader sense, p-hacking refers to presenting results of any analysis that reinforce prior beliefs even though other results that may be similarly meaningful are not presented.

P-hacking is bad because it deceives our audience, and it can deceive ourselves, too. It's a particularly dangerous way to lie with statistics because the methods are fine, only not everything is presented. The seemingly obvious way to avoid it would be to present all results, including those that would lead

to a conflicting, or less strong, conclusion. Unfortunately, that goes against our other general advice to stay focused and show few results that the audience can digest. Thus, a good practice is to produce may robustness checks, and report their results next to the main results, sometimes only as a qualifier. For example, we may say that including extreme values would lead to a less strong conclusion.

Another consequence of p-hacking is that we can't fully trust the results of previous research. As an extreme example, suppose that there is a medical intervention that turns out not to have any effect. (Think of your favorite hoax here.) But it was tested in 100 well-designed experiments. In 5 of the 100 experiments, the researchers found that they can reject the null that it has no effect. So these 5 are false positives. That's actually what we would expect if all experiments used a 5% levels of significance for their test. This is all very well, but now suppose that only those 5 positive results are published, the other 95 are not. Looking through the published results, we would see only five experiments, all of which with a positive result, concluding that the intervention had an effect.

While p-hacking is in reality probably not as extreme as we have described it here, it surely exists. Because much published research shows something that may not always be true, the phenomenon is also called **publication bias**: answers to a question that end up being published may be a biased sample of all answers to the question that were produced during analyzing available data. There are some ingenious methods to detect p-hacking and publication bias and correct results for those biases. Discussing those is beyond the scope of our textbook. Instead, we advise treating published evidence with some healthy skepticism. Yes, there is usually a lot to learn about a question from reading previous research. But no, the answer that emerges from all that reading is not always the final and only true answer.

14 Testing hypotheses with Big Data

As we discussed in Chapter 5, Section 16, generalizing to the population, or general pattern, represented by the data is not really an issue with big data that has a lot of observations. The CI produced from such big data would be so narrow that it would be practically the same as our estimated value. However, generalizing from the population, or general pattern, represented by the data to the population, or general pattern, we care about, is a different question. Here it does not help if the data is big, with very many observations. By the same logic, the process of testing hypotheses with big data is not usually a big deal: if the statistic computed from the data is not literally equal to the value in the null, we would always reject the null. But whether that decision is readily generalizable to the situation we care about is a matter of external validity, where big data doesn't help.

One caveat here is that big data with very many observations may not be so big if we examine very rare phenomena. Consider our example of presenting two versions of the same online ad to two groups of people, and examining whether the follow-up rate in the two groups is the same. In many experiments of this kind, the actual follow-up rates are tiny. Suppose, for example, that of 1 million people that were shown the old version of the ad, 100 followed up to visit the company's website. Of the other 1 million people that were shown the new version, 130 followed up. There is a difference in the follow-up rates in this data, but that doesn't necessarily mean that the difference is there in the general pattern represented by this data. That's even though we have 2 million observations, which is a lot. However, the follow-up rates are tiny. Thus, the standard error that tells us whether that difference of 30 people is large enough is not that tiny in the end. Thus, tiny differences need to be properly tested even with big data. You will be invited to carry out this test among the practice questions.

However, big data has a lot of variables instead of, or on top of, a lot of observations. Examples include data from DNA sequencing, or detailed data on online transactions. In some cases the number

of variables exceeds the number of observations – sometimes by orders of magnitude. The issues with testing multiple hypotheses are especially relevant in such data. When, for example, we want to see which parts of the DNA are associated with a certain medical condition, the number of hypotheses to test may be in the millions. Applying conventional 5% or even 1% levels of significance would yield many false positives. Instead, it becomes critically important to apply appropriate methods to handle this issue in an explicit manner – e.g., by applying very conservative levels of significance such as one millionth of a percent. Testing multiple hypotheses with that many variables is at the frontier of statistical research on using big data. Thus we just describe the issue; solutions to it are beyond the scope of our textbook.

15 Summary and practice

15.1 Main takeaway

- Testing in statistics means making a decision about the value of a quantity (statistic, parameter) in the general pattern represented by the data.
  - It starts with explicitly stating H_0 and H_A.
  - A statistical test rejects H_0 if there is enough evidence against it; otherwise it doesn't reject it.
  - Testing multiple hypotheses at the same time is a tricky business; it pays to be very conservative with rejecting the null.

15.2 Practice questions

1. Write down an example for a null and a two-sided alternative hypothesis.
2. Write down an example for a null and a one-sided alternative hypothesis.
3. What is false positive? What is false negative? Give an example of a test, and explain what a false positive and a false negative would look like.
4. What is the level of significance or size of a test, and what is the power of a test?
5. Tests on larger datasets have more power in general. Why?
6. What is the p-value? The p-value of a test is 0.01. What does that mean and what can you do with that information?
7. Why is testing multiple hypotheses problematic, and what can you do about it? Give an example.
8. What is p-hacking, and how can you minimize its perils when presenting the results of your analysis? Give an example.
9. What is the effect of p-hacking on published results that examine the same question? What does that imply for the usefulness of reading through the literature of published results? Give an example.

10. You examine the wages of recent college graduates, and you want to test whether the starting wage of women is the same, on average, as the starting wage of men. Define the statistic you want to test. Define the population for which you can carry out the test if your data is a random sample of college graduates from your country surveyed in 2015. Write down the appropriate null and alternative hypotheses, and describe how you would carry out the test. What would be a false negative in this case? What would be a false positive?
11. You are testing whether an online advertising campaign had an effect on sales by comparing the average spending by customers who were exposed to the campaign with the average spending of customers who were not exposed. Define the statistic you want to test. What is the null and the alternative if your question is whether the campaign had a positive effect? What would be a false negative in this case? What would be a false positive? Which of the two do you think would have more severe business consequences?
12. Consider the null hypothesis that there is no differences between the likelihood of bankruptcy of firms that were established more than three years ago and firms that were established less than three years ago. You carry out a test using data on all firms from a country in 2015, and the test produces a p-value of 0.001. What is your conclusion? What if the p-value was 0.20?
13. A randomly selected half of the employees of a customer service firm participated in a training program, while the other half didn't. How would you test whether the training had an effect on the satisfaction of the customers they serve? Describe all steps of the test procedure.
14. Consider our example of online ads in Section 14, in which two versions of the same ad were shown to 1 million people each, and 100 followed up from the group that was shown the old version, while 130 followed through from the other group. Write down the test statistic, the null and the alternative (go for two-sided alternative). Carry out the test using the t-statistic with the help of the following statistics: the follow-up rate for the new version is 0.00013; for the old version 0.00010; their difference is 0.00003; the SE of that difference is 0.000015. Interpret your decision.
15. Consider the same online ad example as in the previous question, but now go for a one-sided alternative. Write down the test statistic, the null and the alternative, and argue why you did it that way. Carry out the test using the information that the p-value of the two-sided test would be 0.05.

15.3 Data exercises

Easier and/or shorter exercises are denoted by [*] Harder and/or longer exercises are denoted by [**]

1. Use the same `billion-prices` data as in the case study. Pick another country, exclude sales and promotions, and test whether online and offline prices are the same, on average. Compare your results to the case study. [*]
2. Use the same `billion-prices` data as in the case study. Pick the same country, but now include sales and promotions, and test whether online and offline prices are the same, on average. Compare your results to the case study. [*]
3. Use the same `billion-prices` data as in the case study. Pick the same country, exclude sales and promotions, and keep only products whose offline and online price was assessed on the same day. Test whether online and offline prices are the same, on average. Also test the multiple hypotheses of whether they are the same in each store. Compare your results to the case study. [*]

4. Use the same `sandp500` data as in the case study. Pick a different definition for a large loss, and/or a different threshold for its frequency. Test whether the frequency of large losses is larger or smaller than that threshold frequency. Carry out the test and interpret its results, including what to expect for a similar investment portfolio next year. [*]

5. Consider the `hotels-europe` data. Pick a city and two different dates. Test if prices are the same, on average, on those two days. Watch out for potential differences in the included hotels for those two dates. One way to see if the two groups are different hotels is comparing their average customer ratings – so test if their average ratings are the same. Discuss your findings. [*]

16 References and further reading

On p-hacking and publication bias, an informative article is Annie Franco & Simonovits (2014).