CHAPTER 1

An Overview of Econometrics

1.1 The importance of econometrics

The Duke of Wellington, a British commander of the Napoleonic Wars, once said: ‘All the business of war, indeed all the business of life, is to endeavour to find out what you don’t know by what you do; that’s what I call “guessing what is on the other side of the hill”’. This is an apt description of what econometrics is all about.

Economics is full of unanswered questions such as: ‘Will a change in interest rates affect the exchange rate?’ ‘Do the long-term unemployed have a more difficult time getting jobs than the short-term unemployed?’ ‘What is the impact of gas prices on the choice of whether to drive or take the bus to work?’. These are examples of the ‘what we don't know’ of economics. The ‘what we do know’ of economics are data. All sorts of agencies (e.g. governments, newspapers, companies even individuals) collect facts that shed light on the ‘what we don't know’. Look, for instance, in most newspapers and you will find lots of information about the prices of various assets (e.g. interest rates, exchange rates, stock prices, etc.). Most governments carry out surveys or censuses of many activities of their citizens, and these can, for example, be used to compare the experience of the long-term unemployed with that of the short-term unemployed. Economic researchers have carried out surveys of commuters, and some of the information provided can be used to investigate factors that influence the choice between the private car and public transport.

Wellington knew that one had to appeal to the facts to make a good military decision. The same applies in economics. Without an appeal to the facts (i.e. the data), economic debates can degenerate into a sterile repetition of fixed opinions. Or they can become informal storytelling sessions where economists support their views with their favorite anecdotes. When making military preparations, anyone can ‘guess what is on the other
side of this hill, but it takes a great commander to combine all the available information and draw the most sensible conclusions. To continue the analogy, the purpose of econometrics is to show the economist how to be a great commander, to use ‘what we know’ in the most effective manner in order to try and resolve ‘what we don't know’. In other words, econometrics shows us how to use data in a sensible and systematic manner to shed light on economic questions.

The purpose of this chapter is to provide you with an understanding of the basic concepts and tools that are used by econometricians. Given the primary role of data in econometrics, it won't surprise you to learn that much of this chapter is about data. We discuss the types of data commonly used by economists and offer a brief discussion about where data are obtained. Following this, we discuss some simple ways of analyzing data (e.g. graphical methods and descriptive statistics) and offer an introduction to some of the basic theoretical tools used by the econometrician (e.g. expected values and variances). These basic concepts and tools are then used in all the remaining chapters of this book.

### 1.2 Types of economic data

This section introduces the types of data used by economists and defines the notation and terminology associated with them.

#### 1.2.1 Time series data

Macroeconomists and financial economists are often interested in concepts such as gross domestic product (GDP), stock prices, interest rates, exchange rates, etc. Such data are collected at specific points in time. In all of these examples, the data are ordered by time and are referred to as time series data. The underlying phenomenon that we are measuring (e.g. GDP, stock prices, interest rates, etc.) is referred to as a variable. Time series data can be observed at many frequencies. Commonly used frequencies are: annual (i.e. a variable is observed every year), quarterly (i.e. 4 times a year), monthly, weekly or daily.

In this book, we will use the notation $Y_t$ to indicate an observation on variable $Y$ (e.g. an exchange rate) at time $t$. A series of data runs from period $t = 1$ to $t = T$. Here, $T$ is used to indicate the total number of time periods covered in a dataset. To give an example, if we were to use monthly time series data from January 1947 to October 1996 on the UK pound/US dollar exchange rate – a period of 598 months – then $t = 1$ would indicate January 1947, $t = 598$ would indicate October 1996 and $T = 598$ would be the total number of months. Hence, $Y_1$ would be the pound/dollar exchange rate in January 1947, $Y_2$ would be this exchange rate in February 1947, etc. Time series data are presented in chronological order.

Working with time series data often requires some special tools, which are discussed in Chapters 6 and 7.
1.2.2 Cross-sectional data

In contrast to the above, researchers often work with data that are characterized by individual units. These units might refer to companies, people, or countries. For instance, a financial economist investigating theories relating to portfolio allocation might collect data on the return earned on the stocks of many different companies. With such cross-sectional data, the ordering of the data typically does not matter (unlike time series data).

In this book, we use the notation \( Y_i \) to indicate an observation on variable \( Y \) for individual \( i \). Observations in a cross-sectional dataset run from unit \( i = 1 \) to \( N \). By convention, \( N \) indicates the number of cross-sectional units (e.g. the number of companies surveyed). For instance, a researcher might collect data on the share price of \( N = 100 \) companies at a certain point in time. In this case, \( Y_1 \) will be equal to the share price of the first company, \( Y_2 \) will be equal to the share price of the second company, and so on.

It is worthwhile stressing another important distinction between types of data. In the preceding example, the researcher collecting data on share prices will have a number corresponding to each company (e.g. the price of a share of company 1 is $25). This is referred to as quantitative data.

However, there are many cases where data do not come in the form of single numbers. For instance, the labour economist, when asking whether or not each surveyed employee belongs to a union, receives either a Yes or a No answer. These answers are referred to as qualitative data. Such data arise often in economics when choices are involved (e.g. the choice to buy or not to buy a product, to take public transport or a private car). Econometricians usually convert these qualitative answers into numeric data. For instance, the labor economist might set \( \text{Yes} = 1 \) and \( \text{No} = 0 \). Hence, \( Y_1 = 1 \) means that the first individual surveyed does belong to a union, and \( Y_2 = 0 \) means that the second individual does not. When variables can take on only the values 0 or 1, they are referred to as dummy (or binary) variables.

1.2.3 Panel data

Some datasets will have both a time series and a cross-sectional component. Such data are referred to as panel data. Economists working on issues related to economic growth often make use of panel data. They might work, for instance, with data for 90 countries for the years 1950–2000 for the variable \( Y = \text{GDP} \). Such a dataset would contain the value of GDP for each country in 1950 (\( N = 90 \) observations), followed by GDP for each country in 1951 (another \( N = 90 \) observations), and so on. Over a period of \( T \) years, there would be \( TN \) observations on \( Y \). We will use the notation \( Y_{it} \) to indicate an observation on variable \( Y \) for country \( i \) at time \( t \). Panel datasets are often used by labour economists. For instance, the government often carries out surveys of many people asking them questions about their employment, income, education, etc. From such a survey the labour economist might work with the variable \( Y = \text{the wage of} \ N = 1000 \text{ individuals for} \ T = 5 \text{ years.} \)
1.2.4 Obtaining data

All of the data you need in order to understand the basic concepts and to carry out the analyses covered in this book can be downloaded from the website associated with this book. However, in the future you may need to gather your own data for an essay, dissertation, or report. Economic data come from many different sources, and it is hard to offer general comments on the collection of data. Below are a few key points that you should note about common datasets and where to find them.

It is becoming increasingly common for economists to obtain their data over the internet, and many relevant websites now exist from which data can be downloaded. You should be forewarned that the web is a rapidly growing and changing place, so that the information and addresses provided here might soon be outdated. Accordingly, this section is provided only to give an indication of what can be obtained over the internet, and as such is far from complete.

Some of the datasets available on the web are free, but many are not. Most university libraries or computer centres subscribe to various databases that the student can use. You are advised to check with your own university library or computer centre to see what datasets you have access to. Most universities will at a minimum have access to the major datasets collected by the government. For instance, in the UK the Office of National Statistics (ONS) collects all sorts of data, and these are usually available through UK university libraries. The UK Data Archive (http://www.data-archive.ac.uk/) is another useful source. An extremely useful American site is ‘Resources for Economists on the Internet’ (http://rfe.org). This site contains all sorts of interesting material on a wide range of economic topics and provides links to many different data sources. On this site you can also find links to journal data archives. Many journals encourage their authors to make their data publicly available, and hence, in many cases, you can get data from published academic papers through journal data archives. A good example is the Journal of Applied Econometrics Data Archive (http://www.econ.queensu.ca/iae/).

Another site with useful links is the National Bureau of Economic Research (http://www.nber.org/). One good data source available through this site is the Penn World Table (PWT), which gives macroeconomic data for over 100 countries for many years. We will refer to the PWT below. Most countries also have large panel datasets where large groups of individuals are surveyed every year. In America, the Panel Study of Income Dynamics (http://psidonline.isr.umich.edu/) is a valuable resource for researchers in many fields. In the UK, the comparable panel dataset is the British Household Panel Survey (http://www.iser.essex.ac.uk/ulsc/bhps/).

With regard to financial data, there are many excellent databases of stock prices and accounting information for all sorts of companies for many years. Unfortunately, these tend to be very expensive and, hence, you should see whether your university has a subscription to a financial database. Two of the more popular ones are DataStream by Thompson Financial (http://www.datastream.com/) and Wharton Research Data Services (http://wrds.wharton.upenn.edu/). For free data, a more limited choice of financial data is available through popular internet ports such as Yahoo (http://yahoo.finance.com). The Federal Reserve Bank of St Louis also maintains a free database with a wide variety of
data, including some financial time series (http://research.stlouisfed.org/fred2/). The Financial Data Finder (http://www.cob.ohio-state.edu/fin/osudata.htm), provided by the Fisher College of Business at the Ohio State University, is also a useful resource. Many academics also make the datasets they have used available on their websites. For instance, Robert Shiller at Yale University has a website that provides links to many different interesting financial datasets (http://aida.econ.yale.edu/%7Eshiller/index.html). A general point worth stressing is that spending some time searching the web can often be very fruitful.

1.2.5 Data transformations: levels and growth rates

In this book, we will mainly assume that the data of interest, \( Y \), are directly available. However, in practice, it is common to take raw data from one source and then transform it into a different form for empirical analysis. For instance, the financial economist may take raw time series data on the variables \( X = \) company earnings and \( W = \) number of shares and create a new variable: \( Y = \) earnings per share. Here, the transformation would be

\[
Y = \frac{X}{W}.
\]

The exact nature of the transformation required depends on the problem at hand, so it is hard to offer any general recommendations on data transformation. Some special cases are considered in later chapters. Here, it is useful to introduce some transformations that often arise with time series data.

To motivate this transformation, note that in many cases macroeconomists and financial economists are not interested directly in a variable (e.g. GDP), but rather how it is changing over time. To make things concrete, consider financial economists. In many cases they would not be interested in the price of an asset, but rather in the return that an investor would make from purchase of the asset. This depends on how much the price of the asset will change over time. Suppose, for instance, that the financial economist has annual data on the price of a share in a particular company for 1950–1998 (i.e. 49 years of data), denoted by \( Y_t \) for \( t = 1 \rightarrow 49 \). In some cases this might be the variable of primary interest. Such a variable is referred to as a level (i.e. we refer to the ‘level of the share price’). However, people are often more interested in the growth of the share price. A simple way to measure growth is to take the share price series and calculate a percentage change for each year. The percentage change in the share price between period \( t - 1 \) and \( t \) is calculated according to the formula

\[
\% \text{ change} = \frac{Y_t - Y_{t-1}}{Y_{t-1}} \times 100.
\]

It is worth stressing that a percentage change always has a timescale associated with it (e.g. the percentage change between period \( t - 1 \) and \( t \)). For instance, with annual data this formula would produce an annual percentage change, with monthly data the formula would produce a monthly percentage change, etc. As will be discussed in later chapters, it is sometimes convenient to take the natural logarithm of variables. The definition and
properties of logarithms can be found in Appendix A: Mathematical Basics at the end of this book. Using the properties of logarithms, it can be shown that the percentage change in a variable is approximately

\[ \% \text{ change} \approx \frac{\ln(Y_t) - \ln(Y_{t-1})}{\ln(Y_t/\bar{Y}_0)} \times 100. \]

In practice, the ‘×100’ is often dropped, so that, say, 5% would be 0.05. The percentage change in an asset’s price is often referred to as the growth of the price or the change in the price. Changes in variables are often used with macroeconomic variables. For instance, macroeconomists sometimes study GDP growth (instead of the level of GDP) or inflation (which is the change in the price level). Chapters 6 and 7 cover time series econometrics, and in these chapters it is often important to distinguish between the level of a variable and its growth rate.

1.3 Working with data: graphical methods

Once you have your data, it is important for you to summarize it. After all, anybody who reads your work will not be interested in the dozens or, more likely, hundreds or more observations contained in the original raw dataset. Indeed, you can think of the whole field of econometrics as one devoted to the development and dissemination of methods whereby information in datasets is summarized in informative ways. Charts and tables are very useful ways of presenting your data. There are many different types (e.g. bar chart, pie chart, etc.). In this section, we will illustrate a few of the commonly used types of chart. Since most economic data are either in time series or cross-sectional form, we will briefly introduce simple techniques for graphing both types of data.

1.3.1 Time series graphs

Monthly time series data from January 1947 to October 1996 on the UK pound/US dollar exchange rate are plotted in Figure 1.1. Such charts are commonly referred to as time series graphs. The dataset contains 598 observations – far too many to be presented as raw numbers for a reader to comprehend. However, a reader can easily capture the main features of the data by looking at the graph. One can see, for instance, the attempt by the UK government to hold the exchange rate fixed until the end of 1971 (apart from large devaluations in September 1949 and November 1967) and the gradual depreciation of the pound as it floated downwards through the middle of the 1970s.

1.3.2 Histograms

With time series data, a graph that shows how a variable evolves over time is often very informative. However, in the case of cross-sectional data, such methods are not appropriate and we must summarize the data in other ways.
The Penn Word Table allows us to obtain cross-sectional data on real GDP per capita in 1992 for 90 countries. Real GDP per capita in every country has been converted into US dollars using purchasing power parity exchange rates. This allows us to make direct comparisons across countries. One convenient way of summarizing these data is through a histogram. To construct a histogram, begin by choosing class intervals or bins that divide the countries into groups based on their GDP per capita. In our dataset, GDP per person varies from $408 in Chad to $17,945 in the USA. One possible set of class intervals is 0–2,000, 2,001–4,000, 4,001–6,000, 6,001–8,000, 8,001–10,000, 10,001–12,000, 12,001–14,000, 14,001–16,000, and 16,001–18,000 (where all figures are in US dollars). Note that each class interval is $2,000 wide. In other words, the class width for each of our bins is 2,000. For each class interval we can count up the number of countries that have GDP per capita in that interval. For instance, there are seven countries in our dataset with real GDP per capita between $4,001 and $6,000. The number of countries lying in one class interval is referred to as the frequency of that interval. A histogram is a bar chart that plots frequencies against class intervals.

Figure 1.2 is a histogram of our cross-country GDP per capita dataset that uses the class intervals specified in the previous paragraph. Note that, if you do not wish to specify class intervals, any relevant computer software package will do it automatically for you. Most computer packages will also create a frequency table, which we have put into Table 1.1. The frequency table indicates the number of countries belonging to each class interval.

Figure 1.1 Time series plot of UK pound/US dollar exchange rate.
For instance, we can read that there are 33 countries with GDP per capita less than $2,000, 22 countries with GDP per capita above $2,000 but less than $4,000, and so on. The last row indicates that there are four countries with GDP per capita between $16,000 and $18,000.

This same information is graphed in a simple fashion in the histogram in Figure 1.2. Graphing allows for a quick visual summary of the cross-country distribution of GDP per capita.

### Table 1.1 Frequency table for GDP per capita data.

<table>
<thead>
<tr>
<th>Class interval ($)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–2,000</td>
<td>33</td>
</tr>
<tr>
<td>2,001–4,000</td>
<td>22</td>
</tr>
<tr>
<td>4,001–6,000</td>
<td>7</td>
</tr>
<tr>
<td>6,001–8,000</td>
<td>3</td>
</tr>
<tr>
<td>8,001–10,000</td>
<td>4</td>
</tr>
<tr>
<td>10,001–12,000</td>
<td>2</td>
</tr>
<tr>
<td>12,001–14,000</td>
<td>9</td>
</tr>
<tr>
<td>14,001–16,000</td>
<td>6</td>
</tr>
<tr>
<td>16,001–18,000</td>
<td>4</td>
</tr>
</tbody>
</table>

**Figure 1.2** Histogram of GDP per capita for 90 countries.
per capita. We can see from the histogram that many countries are very poor, but that there is also a ‘clump’ of countries that are quite rich (e.g. 19 countries have GDP per capita greater than $12,000). There are relatively few countries in between these poor and rich groups (i.e. few countries fall in the bins between $6,001 and $12,000). Researchers often refer to this clumping of countries into poor and rich groups as the ‘twin peaks’ phenomenon. In other words, if we imagine that the histogram is a mountain range, we can see a peak at the class intervals $0–$2,000 and a smaller peak at the class interval $12,000–$14,000. These features of the data can be seen easily from the histogram, but would be difficult to comprehend simply by looking at the raw data.

1.3.3 XY plots

Economists are often interested in the nature of the relationships between two or more variables. For instance: ‘What is the relationship between capital structure (i.e. the division between debt and equity financing) and firm performance (e.g. profit)?’, ‘Are higher education levels and work experience associated with higher wages among workers in a given industry?’, ‘Are changes in the money supply a reliable indicator of inflation changes?’, ‘Do differences in financial regulation explain why some countries are growing faster than others?’, etc. All these questions involve two or more different variables. The techniques described previously are suitable for describing the behaviour of only one variable; for instance, the properties of the single variable real GDP per capita are illustrated in Figure 1.2 above. They are not, however, suitable for examining relationships between pairs of variables.

Once we are interested in understanding the nature of the relationships between two or more variables, it becomes harder to use graphs. Beginning in the next chapter, we will discuss regression analysis, which is the most important tool used by economists working with many variables. However, graphical methods can be used to draw out some simple aspects of the relationship between two variables. XY plots (also called scatter diagrams) are particularly useful in this regard.

Below you will find a graph of data on deforestation (i.e. the average annual forest loss over the period 1981–1990 expressed as a percentage of total forested area) for 70 tropical countries, along with data on population density (i.e. the number of people per thousand hectares). It is commonly thought that countries with a high population density will likely deforest more quickly than those with low population densities, since high population density may increase the pressure to cut down forests for fuel wood or for agricultural land required to grow more food.

Figure 1.3 is an XY plot of these two variables. Each point on the chart represents a particular country. Reading up the Y axis (i.e. the vertical axis) gives us the rate of deforestation in that country. Reading across the X axis (i.e. the horizontal axis) gives us population density. It is possible to label each point with its corresponding country name. We have not done so here, as labels for 70 countries would clutter the chart and make it difficult to read. However, one country, Nicaragua, has been labeled. Note that
this country has a deforestation rate of 2.6% per year ($Y = 2.6$) and a population density of 640 people per thousand hectares ($X = 640$).

The $XY$ plot can be used to give a quick visual impression of the relationship between deforestation and population density. An examination of this graph indicates some support for the idea that a relationship between deforestation and population density does exist. For instance, if we look at countries with a low population density (less than 500 people per hectare, say), almost all of them have very low deforestation rates (less than 1% per year). If we look at countries with high population densities (e.g. over 1500 people per thousand hectares), almost all of them have high deforestation rates (more than 2% per year). This indicates that there may be a positive relationship between population density and deforestation (i.e. high values of one variable tend to be associated with high values of the other, and low values tend to be associated with low values). It is also possible to have a negative relationship between two variables. This might occur, for instance, if we substituted urbanization for population density in an $XY$ plot. In this case, high levels of urbanization might be associated with low levels of deforestation since expansion of cities would possibly reduce population pressures in rural areas where forests are located.

It is worth noting that the positive or negative relationships found in the data are only tendencies and, as such, do not hold necessarily for every country. That is, there may be exceptions to the general pattern of an association between high population density and
high rates of deforestation. For example, on the XY plot we can observe one country with a high population density of roughly 1300 and a low deforestation rate of 0.7%. Similarly, low population density can also be associated with high rates of deforestation, as evidenced by one country with a low population density of roughly 150 but a high deforestation rate of almost 2.5% per year. As economists, we are usually interested in drawing out general patterns or tendencies in the data. However, we should always keep in mind that exceptions (in statistical jargon outliers) to these patterns typically exist. In some cases, finding out which countries don’t fit the general pattern can be as interesting as the pattern itself.

1.4 Working with data: descriptive statistics and correlation

Graphs have an immediate visual impact that is useful for livening up an essay or report. However, in many cases it is important to be numerically precise. The following chapters of this book will describe common numerical methods for summarizing the relationship between several variables in detail. Here, we discuss briefly a few descriptive statistics for summarizing the properties of a single variable or at most two variables. By way of motivation, we will return to the concept of a distribution illustrated briefly in our discussion of histograms. The concept of a probability distribution is a fundamental one in statistics and we will use it throughout this book. Appendix B: Probability Basics, at the end of this book, formally defines and motivates probability distributions. Here we provide some informal intuition about distributions and ways of summarizing their properties.

In our cross-country dataset, real GDP per capita varies across the 90 countries. This variability can be seen by looking at the histogram in Figure 1.2, which plots the distribution of GDP per capita across countries. Suppose you wanted to summarize the information contained in the histogram numerically. One thing you could do is to present the numbers in the frequency table (see Table 1.1). However, even this table may provide too many numbers to be easily interpretable. Instead it is common to present two simple numbers called the mean and standard deviation. The mean is the statistical term for the average. Remember that $Y_1, \ldots, Y_N$ can be used to denote the $N$ different observations on our variable. This is referred to as a sample. The mathematical formula for the mean is given by

$$\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N},$$

where $N$ is referred to as the sample size (here this is the number of countries) and $\sum_{i=1}^{N}$ is the summation operator (i.e. it adds up real GDP per capita for all countries). The summation operator is defined and discussed in Appendix A: Mathematical Basics. In our case, mean GDP per capita is $5,443.80$. Throughout this book, we will place a bar over a variable to indicate its mean (i.e. $\bar{Y}$ is the mean of the variable $Y$, $\bar{X}$ is the mean of the variable $X$, etc.).
The concept of the mean is associated with the middle of a distribution. For example, if we look at the previous histogram in Figure 1.2, $5443.80 lies somewhere in the middle of the distribution. The cross-country distribution of real GDP per capita is quite unusual, having the twin peaks property described earlier. It is more common for distributions of economic variables to have a single peak and to be bell-shaped. Figure 1.4 is a histogram that plots just such a bell-shaped distribution. For such distributions, the mean is located precisely in the middle of the distribution, under the single peak. The leading example of a distribution of this sort is the normal distribution (see Appendix B, Definition B.10).

Of course, the mean or average hides a great deal of variability across countries. Other useful summary statistics, which shed light on the cross-country variation in GDP per capita, are the minimum and maximum. For our dataset, minimum GDP per capita is $408 (Chad) and maximum GDP is $17945 (USA). By looking at the distance between the maximum and minimum, we can see how dispersed the distribution is.

The concept of dispersion is quite important in economics and is closely related to the concepts of variability and inequality. For instance, real GDP per capita in 1992 in our dataset varies from $408 to $17945. If poorer countries were, in the near future, to grow quickly, and richer countries to stagnate, then the dispersion of real GDP per capita in, say, 2012, might be significantly less. It may be the case that the poorest country at this time will have real GDP per capita of $10000 while the richest country will remain
at $17,945. If this were to happen, then the cross-country distribution of real GDP per capita would be more equal (less dispersed, less variable). Intuitively, the notions of dispersion, variability, and inequality are closely related.

The minimum and maximum, however, can be unreliable guidelines to dispersion. For instance, what if, with the exception of Chad, all the poor countries experienced rapid economic growth between 1992 and 2012 while the richer countries did not grow at all? In this case, cross-country dispersion or inequality would decrease over time. However, since Chad and the USA did not grow, the minimum and maximum would remain at $408 and $17,945 respectively. A more common measure of dispersion, which does not suffer from this drawback, is the standard deviation. Its formula is given by

\[ s = \sqrt{\frac{\sum(Y_i - \bar{Y})^2}{N-1}}. \]

Statisticians refer to the square of the standard deviation as the variance \((s^2)\), and it is common to see either terminology used. It is important to remember this close relationship between the standard deviation and the variance.

The standard deviation has little direct intuition. In our cross-country GDP dataset, the standard deviation is $5,369.496 and it is difficult to get a direct feel for what this number means in an absolute sense. However, the standard deviation can be interpreted in a comparative sense. That is, if you compare the standard deviations of two different distributions, the one with the smaller standard deviation will always exhibit less dispersion. In our example, if the poorer countries were to suddenly experience economic growth and the richer countries to stagnate, the standard deviation would decrease over time.

### 1.4.1 Expected values and variances

In the previous section we talked about means and variances. We should actually have called them sample means and sample variances. The word ‘sample’ is added to emphasize that they are calculated using an actual ‘sample’ of data. For instance, in our cross-country GDP dataset we took the data we had and calculated exact numbers for \(\bar{Y}\) and \(s\). We found these to be $5,443.80 and $5,369.496 respectively. These are the sample mean and standard deviation calculated using the dataset at hand.

As another example, suppose we have collected data on the return to holding stock in a company for the past 100 months. We can use these data to calculate the sample mean and sample variance. However, these numbers are calculated on the basis of the historical performance of the company. In finance, we are often interested in predicting future stock returns. By definition, we do not know exactly what these will be, so we need to extend the concepts of means and variances to cases where we do not have a sample of data. Potential investors would be interested in something like a mean and a variance. That is, investors would be interested in the typical return that they might expect. They might also be interested in the risk involved in purchasing the stock. The concept of a typical
or *expected value* sounds similar to the ideas we discussed relating to the mean. The concept of riskiness sounds similar to the idea of a variance we discussed above. In short, we need concepts like the sample mean and variance, but for cases where we do not actually have data to calculate them. The relevant concepts are the *population mean* and the *population variance*.

In Appendix B: Probability Basics, we formally define the population mean and variance and discuss their properties. Here, we provide some intuition and definitions. A common way to motivate the distinction between motivating population and sample concepts is through an example. Consider, for instance, the height of every individual in the USA. In the population as a whole there is some average height (the population mean height) and some variance of heights (the population variance). This population mean and variance will be unknown, unless someone actually went out and measured the height of every person in the USA. However, a researcher might have data on the actual heights of 100 people (e.g. a medical researcher might measure the height of each of 100 patients). Using the data for 100 people, the researcher could calculate $\bar{Y}$ and $s^2$. These are the sample mean and variance and will be actual numbers. The medical researcher could then use these numbers as estimates (or approximations) for what is going on in the country as a whole (i.e. sample means and variances can be used as estimates for population means and variances). This is an important distinction in statistics, and it is important to stress that sample and population concepts are different, with the former being actual numbers calculated using the data at hand and the latter being unobserved.

Perhaps the previous two paragraphs are enough intuitively to motivate the distinction between sample and population means and variances. To see why financial analysts need to know this distinction (and to introduce some notation), let us use our example of potential investors interested in the potential return they might make from buying a stock. Let $Y$ denote next month’s return on this stock. From the investors’ point of view, $Y$ is unknown. The typical return they might expect is measured by the population mean and is referred to as the expected value. We use the notation $E(Y)$ to denote the expected return (also known as the mean return and often labelled $\mu$). Its name accurately reflects the intuition for this statistical concept. The ‘expected value’ sheds light on what we expect will occur.

However, the return on a stock is rarely exactly what is expected (i.e. rarely will you find $Y$ to turn out to be exactly $E(Y)$). Stock markets are highly unpredictable – sometimes the return on the stock could be higher than expected, sometimes it could be lower than expected. In other words, there is always risk associated with purchasing a stock. A potential investor will be interested in a measure of this risk. Variance is a common way of measuring this risk. We use the notation $\text{var}(Y)$ for this.

Appendix B: Probability Basics describes how to obtain $E(Y)$ and $\text{var}(Y)$ given a probability distribution. To give some intuition, here we provide an example from the field of finance. Suppose you are an investor trying to decide whether to buy a stock on the basis of its return next month. You do not know what this return will be. You are quite confident (say, 70 % sure) that the markets will be stable, in which case you will earn a 1 % return. However, you also know there is a 10 % chance the stock market will crash,
in which case the stock return will be $-10\%$. There is also a $20\%$ probability that good news will boost the stock markets and you will get a $5\%$ return on your stock.

In this example, there are three possible outcomes (good, normal, bad) which are 0.05, 0.01, and $-0.10$ (i.e. the possible returns are 5, 1, or $-10\%$). We will use the symbol ‘Pr’ for probability. Thus, \( \Pr(Y = 0.05) = 0.20 \) says that there is a $20\%$ chance of obtaining the $5\%$ return. We can now define the expected return as a weighted average of all the three possible outcomes, where the weights are given as the probability that each occurs:

\[
E(Y) = \Pr(Y = 0.05)0.05 + \Pr(Y = 0.01)0.01 + \Pr(Y = -0.10)(-0.10)
= 0.20 \times 0.05 + 0.70 \times 0.01 + 0.10(-0.10)
= 0.007
\]

In words, the expected return on the stock next month is $0.7\%$ (i.e. a bit less than $1\%$).

In our example, we have assumed that there are only three possible outcomes next month. In general, if there are \( K \) possible outcomes (label them \( y_1, y_2, \ldots, y_K \)), the formula for the expected value is

\[
E(Y) = \sum_{i=1}^{K} \Pr(Y = y_i)y_i.
\]

The formula for the expected value when the variable is continuous (and, thus, there are an infinite number of possible outcomes) is given in Appendix B, Definition B.8. It has similar intuition but is slightly more complicated.

The formula for \( \text{var}(Y) \) is also given in Appendix B. Suffice it to note here that it can be calculated by means of the expected value operator using the formula

\[
\text{var}(Y) = E(Y^2) - [E(Y)]^2.
\]

In our financial example, we have already calculated \( E(Y) = 0.007 \). However, to calculate the variance we still need to calculate \( E(Y^2) \). This can be done in the same manner as before, except using \( Y^2 \) instead of \( Y \). That is, the general formula, if there are \( K \) possible outcomes, is

\[
E(Y^2) = \sum_{i=1}^{K} \Pr(Y^2 = y_i^2)y_i^2.
\]

In the particular example, the three possible outcomes for \( Y^2 \) are \((0.05)^2 = 0.0025\), \((0.01)^2 = 0.0001\), and \((-0.10)^2 = 0.01\). We can plug these into the formula for calculating \( E(Y^2) \):

\[
E(Y^2) = \Pr(Y^2 = 0.0025) \times 0.0025 + \Pr(Y^2 = 0.0001) \times 0.0001
+ \Pr(Y^2 = 0.01) \times 0.01
= 0.20 \times 0.0025 + 0.70 \times 0.0001 + 0.10 \times 0.01
= 0.00157.
\]
We can use this result to obtain
\[
\text{var}(Y) = E(Y^2) - [E(Y)]^2 \\
= 0.00157 - (0.007)^2 \\
= 0.001521.
\]

The standard deviation, being the square root of the variance, can be calculated to be 0.039.

To summarize, in the previous section on descriptive statistics we motivated the use of the sample mean and variance, \(\bar{Y}\) and \(s^2\), to give the researcher an idea of the average value and dispersion, respectively, in a dataset. In this section, we have motivated their population counterparts, \(E(Y)\) and \(\text{var}(Y)\), as having similar intuition but being relevant for summarizing information about an uncertain outcome (e.g. the return on a stock next month). In the following chapters, we will use the expected value and variance operators extensively, and their formal properties are given in Appendix B. However, it is always important to have an intuitive understanding about what the mean and variance operator are, and the intention of this section is to provide this intuition.

### 1.4.2 Correlation

The mean and variance are properties of one variable. However, economists are often interested in investigating the nature of the relationship between two (or more) variables. Most of this textbook is about investigating such relationships, and we will develop many different approaches for doing so. Here, we take a first step in this direction by introducing the idea of correlation. Correlation is an important way of numerically quantifying the relationship between two variables. In this chapter, we will first describe the theory behind correlation, and then work through a few examples designed to think about the concept in different ways.

Let \(X\) and \(Y\) be two variables (e.g. population density and deforestation, respectively), and let us also suppose that we have data on \(i = 1, \ldots, N\) different units (e.g. countries). The correlation between \(X\) and \(Y\) is denoted by \(r\) and its mathematical formula is

\[
\rho = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}}}
\]

The variables to which \(r\) refers are usually clear from the context. However, in some cases we will use subscripts to indicate that \(r_{XY}\) is the correlation between variables \(X\) and \(Y\), \(r_{XZ}\) is the correlation between variables \(X\) and \(Z\), etc.

Once you have calculated the correlation between two variables, you will obtain a number (e.g. \(r = 0.55\)). It is important that you know how to interpret this number. In this section, we will try to develop some intuition about correlation. Firstly, however, let us briefly list some of the numerical properties of correlation.
Properties of correlation

1. \( r \) always lies between \(-1\) and \(1\).
2. Positive values of \( r \) indicate a positive correlation between \(X\) and \(Y\). Negative values indicate a negative correlation. \( r = 0 \) indicates that \(X\) and \(Y\) are uncorrelated.
3. Larger positive values of \( r \) indicate stronger positive correlation. \( r = 1 \) indicates perfect positive correlation. Larger negative values of \( r \) indicate stronger negative correlation. \( r = -1 \) indicates perfect negative correlation.
4. The correlation between \(Y\) and \(X\) is the same as the correlation between \(X\) and \(Y\).
5. The correlation between any variable and itself is \(1\).

Econometricians use the word ‘correlation’, in much the same way as the layperson does, as measuring the degree of association or the strength of the relationship between two variables. The following continuation of the deforestation/population density example will serve to illustrate verbal ways of thinking about the concept of correlation.

Example: The correlation between deforestation and population density

Let us suppose that we are interested in investigating the relationship between deforestation and population density. Remember that the data set used to create Figure 1.3 contains these variables for a cross-section of 70 tropical countries. The correlation between deforestation (\(Y\)) and population density (\(X\)) turns out to be 0.66. Being greater than zero, this number allows us to make statements of the following form:

1. There is a positive relationship (or positive association) between deforestation and population density.
2. Countries with high population densities tend to have high deforestation rates. Countries with low population densities tend to have low deforestation rates. Note that we use the word ‘tend’ here. A positive correlation does not mean that every country with a high population density necessarily has a high deforestation rate, but, rather, that this is the general tendency. It is possible that a few individual countries do not follow this pattern.
3. Deforestation rates vary across countries, as do population densities. Some countries have high deforestation rates, others have low rates. This high/low cross-country variance in deforestation rates tends to ‘match up’ with the high/low variance observed in population densities.

All that the preceding statements require is for \( r \) to be positive. If \( r \) were negative, the opposite of these statements would hold. For instance, high values of \( X \) would be
associated with low values of $Y$, etc. It is somewhat more difficult to get an intuitive feel for the exact number of the correlation (e.g. how the correlation 0.66 differs from 0.26). The $XY$ plots discussed below offer some help, but here we will briefly note an important point to which we shall return when we discuss regression. The degree to which deforestation rates vary across countries can be measured numerically using the formula for the standard deviation discussed previously. As mentioned in point 3 in the above example, the fact that deforestation and population density are positively correlated means that their patterns of cross-country variability tend to match up. It turns out that the correlation squared ($r^2$) measures the proportion of the cross-country variability in deforestation that matches up with, or is explained by, the variance in population density. In other words, correlation is a numerical measure of the degree to which patterns in $X$ and $Y$ correspond. In our population/deforestation example, since $0.66^2 = 0.44$, we can say that 44% of the cross-country variance in deforestation can be explained by the cross-country variance in population density.

---

**Example: House prices in Windsor, Canada**

The file HPRICE.XLS, available on the website associated with this book, contains data relating to $N = 546$ houses sold in Windsor, Canada, in the summer of 1987. It contains the selling price (in Canadian dollars) along with many characteristics for each house. We will use this dataset in future chapters, but for now let us focus on just a few variables. In particular, let us focus on the relationship between $Y =$ the sales price of the house and $X =$ the size of its lot in square feet. The correlation between these two variables is $r_{XY} = 0.54$.

The following statements can be made about house prices in Windsor:

1. Houses with large lots tend to be worth more than those with small lots.
2. There is a positive relationship between lot size and sales price.
3. The variation in lot size accounts for 29% (i.e. $0.54^2 = 0.29$) of the variability in house prices.

Now let us add a third variable, $Z =$ number of bedrooms. Calculating the correlation between house prices and number of bedrooms, we obtain $r_{YZ} = 0.37$. This result says, as we would expect, that houses with more bedrooms tend to be worth more than houses with fewer bedrooms. Similarly, we can calculate the correlation between number of bedrooms and lot size. This correlation turns out to be $r_{XZ} = 0.15$ and indicates that houses with larger lots also tend to have more bedrooms. However, this correlation is small and, unexpectedly, suggests that the link between lot size and number of bedrooms is quite weak. In other words, you may have expected that houses on larger lots, being bigger, would have more bedrooms than houses on smaller lots. However, the correlation indicates that there is only a weak tendency for this to occur.
The above example allows us to motivate briefly an issue of importance, namely that of causality. Researchers are often interested in finding out whether one variable ‘causes’ another. We will not provide a formal definition of causality here but instead will use the word in its everyday meaning. In this example, it is sensible to use the positive correlation between house price and lot size to reflect a causal relationship. That is, lot size is a variable that directly influences (or causes) house prices. However, house prices do not influence (or cause) lot size. In other words, the direction of causality flows from lot size to house prices, not the other way around.

Another way of thinking about these issues is to ask yourself what would happen if a homeowner were to purchase some adjacent land and thereby increase the lot size of the house. This action would tend to increase the value of the house (i.e. an increase in lot size would cause the price of the house to increase). However, if you reflect on the opposite question, ‘Will increasing the price of the house cause lot size to increase?’ you will see that the opposite causality does not hold (i.e. house price increases do not cause lot size increases). For instance, if house prices in Windsor were suddenly to rise for some reason (e.g. owing to a boom in the economy), this would not mean that houses in Windsor suddenly acquired bigger lots.

The discussion in the previous paragraph could be repeated with ‘lot size’ replaced by ‘number of bedrooms’. That is, it is reasonable to assume that the positive correlation between \( Y = \) house prices and \( Z = \) number of bedrooms is due to \( Z \) influencing (or causing) \( Y \), rather than the opposite. Note, however, that it is difficult to interpret the positive (but weak) correlation between \( X = \) lot size and \( Z = \) number of bedrooms as reflecting causality. That is, there is a tendency for houses with many bedrooms to occupy large lots, but this tendency does not imply that the former causes the latter.

One of the most important things in empirical work is knowing how to interpret your results. The house example illustrates this difficulty well. It is not enough just to report a number for a correlation (e.g. \( r_{XY} = 0.54 \)). Interpretation is important too. Interpretation requires a good intuitive knowledge of what a correlation is in addition to a great deal of common sense about the economic phenomenon under study. Given the importance of interpretation in empirical work, the following section will present several examples to show why variables are correlated and how common sense can guide us in interpreting them.

**Understanding why variables are correlated**

In our deforestation/population density example, we discovered that deforestation and population density are indeed correlated positively, indicating a positive relationship between the two. But what exact form does this relationship take? As discussed above, we often like to think in terms of causality or influence, and it may indeed be the case that correlation and causality are closely related. For instance, the finding that population density and deforestation are correlated could mean that the former directly causes the latter. Similarly, the finding of a positive correlation between education levels and wages could be interpreted as meaning that more education does directly influence the wage
one earns. However, as the following examples demonstrate, the interpretation that correlation implies causality is not always necessarily an accurate one.

**Example: Correlation does not necessarily imply causality**

It is widely accepted that cigarette smoking causes lung cancer. Let us assume that we have collected data from many people on (a) the number of cigarettes each person smokes per week ($X$) and (b) whether they have ever had or now have lung cancer ($Y$). Since smoking causes cancer, we would undoubtedly find that $r_{XY} > 0$; that is, that people who smoked tend to have higher rates of lung cancer than non-smokers. Here, the positive correlation between $X$ and $Y$ indicates direct causality.

Now suppose that we also have data on the same people, measuring the amount of alcohol they drink in a typical week. Let us call this variable $Z$. In practice it is the case that heavy drinkers also tend to smoke, and hence $r_{XZ} > 0$. This correlation does not mean that cigarette smoking also causes people to drink. Rather it probably reflects some underlying psychological or social attitude: people who smoke often tend to drink as well. Thus, a correlation between two variables does not necessarily mean that one causes the other. It may be the case that an underlying third variable is responsible.

Now consider the correlation between lung cancer and drinking. Since people who smoke tend to get lung cancer more, and people who smoke also tend to drink more, it is not unreasonable to expect that lung cancer rates will be higher among heavy drinkers (i.e. $r_{YZ} > 0$). Note that this positive correlation does not imply that alcohol consumption causes lung cancer. Rather, it is cigarette smoking that causes cancer, but smoking and drinking are related to some underlying psychological or social attitude. This example serves to indicate the kind of complicated patterns of causality that occur in practice and how care must be taken when trying to relate the concepts of correlation and causality.

Another important distinction is that between direct (or immediate) and indirect (or proximate) causality. Recall that, in our deforestation/population density example, population density ($X$) and deforestation ($Y$) were found to be positively correlated (i.e. $r_{XY} > 0$). One reason for this positive correlation is that high population pressures in rural areas cause farmers to cut down forests to clear new land in order to grow food. It is this latter ongoing process of agricultural expansion that directly causes deforestation. If we calculated the correlation between deforestation and agricultural expansion ($Z$), we would find $r_{YZ} > 0$. In this case, population density would be an indirect cause and agricultural expansion a direct cause of deforestation. In other words, we can say that $X$ (population pressures) causes $Z$ (agricultural expansion), which in turn causes $Y$ (deforestation). Such a pattern of causality is consistent with $r_{XY} > 0$ and $r_{YZ} > 0$. In our house price example, however, it is likely that the positive correlations we observed reflect direct causality. For instance, having a larger lot is considered by most people to be a good thing in and
of itself, so that increasing the lot size should directly increase the value of a house. There is no other intervening variable here, and hence we say that the causality is direct.

The general message that should be taken from these examples is that correlations can be very suggestive but cannot on their own establish causality. In the smoking/cancer example above, the finding of a positive correlation between smoking and lung cancer, in conjunction with medical evidence on the manner in which substances in cigarettes trigger changes in the human body, has convinced most people that smoking causes cancer. In the house price example, common sense tells us that the variable 'number of bedrooms' directly influences house prices. In economics, the concept of correlation can be used in conjunction with common sense or a convincing economic theory to establish causality.

**Understanding correlation through XY plots**

Intuition about the meaning of correlations can also be obtained from the XY plots. Recall that, when interpreting Figure 1.3, we discussed positive and negative relationships based on whether the XY plots exhibited a general upward or downward slope. If two variables are correlated, then an XY plot of one against the other will exhibit such patterns. For instance, the XY plot of population density against deforestation exhibits an upward sloping pattern (see Figure 1.3). This plot implies that these two variables should be positively correlated, and we find that this is indeed the case from the correlation $r = 0.66$. The important point here is that positive correlation is associated with upward sloping patterns in the XY plot, and negative correlation is associated with downward sloping patterns. All the intuition we developed about XY plots in the previous section can now be used to develop intuition about correlation.

Figure 1.5 uses the Windsor house price dataset to produce an XY plot of $X = $ lot size against $Y = $ house price. Recall that the correlation between these two variables was calculated as $r_{XY} = 0.54$, which is a positive number. This positive (upward sloping) relationship between lot size and house price can clearly be seen in Figure 1.5. That is, houses with small lots (i.e. small $X$-axis values) also tend to have small prices (i.e. small $Y$-axis values). Conversely, houses with large lots tend to have high prices.

The previous discussion relates mainly to the sign of the correlation. However, XY plots can also be used to develop intuition about how to interpret the magnitude of a correlation, as the following examples illustrate. Figure 1.6 is an XY plot of two perfectly correlated variables (i.e. $r = 1$). They do not correspond to any actual economic data, but were simulated on the computer. All the points lie exactly on a straight line.

At the other extreme, Figure 1.7 is an XY plot of two completely uncorrelated variables ($r = 0$). Note that the points are randomly scattered over the entire graph. Real-world datasets (e.g. in Figure 1.3 or 1.5) tend to be between the two extremes.

We have illustrated these points about correlation using positively correlated variables. Plots for negative correlation exhibit downward (instead of upward) sloping patterns, but otherwise the same sorts of pattern noted above hold for them. These figures illustrate one way of thinking about correlation: correlation indicates how well a straight line can
Figure 1.5 *XY* plot of lot size against house price.

Figure 1.6 *XY* plot of two perfectly correlated variables.
be fitted through an XY plot. Variables that are strongly correlated have observations that fit on or close to a straight line. Variables that are weakly correlated have observations that are more scattered in an XY plot.

**Correlation between several variables**

Correlation is a property that relates two variables. Frequently, however, researchers must work with several variables. For instance, house prices depend on the lot size, number of bedrooms, number of bathrooms and many other characteristics of the house. As we will see in subsequent chapters, regression is the most appropriate tool for use if the analysis contains more than two variables. However, it is also not unusual for empirical researchers, when working with several variables, to calculate the correlation between each pair. Note that there are many correlations when the number of variables is large. For instance, if we have three variables, $X$, $Y$, and $Z$, then there are three possible correlations (i.e. $r_{XY}$, $r_{XZ}$, and $r_{YZ}$). However, if we add a fourth variable, $W$, the number increases to six (i.e. $r_{XY}$, $r_{XZ}$, $r_{XW}$, $r_{YZ}$, $r_{YW}$, and $r_{ZW}$). In general, for $M$ different variables there will be $\frac{M(M-1)}{2}$ possible correlations. A convenient way of ordering all these correlations is to construct a matrix or table, as illustrated by the following example.

Using data on three variables labelled $X$, $Y$, and $Z$, we calculate the correlation matrix in Table 1.2.

To explain the interpretation of a correlation matrix, note that the number 0.318 is $r_{XY}$ since it appears in the column labelled $X$ and the row labelled $Y$. Similarly,
\[ r_{XZ} = -0.131 \] and \[ r_{YZ} = 0.097. \] Note that the 1.000 values in the correlation matrix indicate that any variable is perfectly correlated with itself, and the upper right-hand corner of the matrix is left blank since it would be identical to the lower left-hand corner (e.g. since \( r_{XY} = r_{YX} \)).

### 1.4.3 Population Correlations and Covariances

Previously we discussed means and variances and distinguished between sample and population variants. For instance, the sample mean was denoted by \( \overline{Y} \) and was the average calculated using the data at hand. The population mean was denoted by \( E(Y) \) (or \( \mu \)) and called the expected value. It was a more theoretical concept. We motivated it with an example where \( Y \) was next month’s return on a stock. This is not known exactly, but financial analysts are often able to predict what they would expect the return to be. This is \( E(Y) \). However, there is uncertainty associated with the analyst’s prediction, and this is measured through the (population) variance, denoted \( \text{var}(Y) \).

The same sample/population distinction holds with correlations. We will use the notation \( \text{corr}(X, Y) \) to denote the population correlation (remember \( r \) is our notation for the sample correlation). A formal definition of the population correlation is given in Appendix B: Probability Basics, Definition B.8. Here, we try informally to motivate why this concept might be useful using a financial example. Consider a portfolio consisting of the shares of two companies with returns \( X \) and \( Y \). The expected return of the portfolio depends on the expected returns of the two individual stocks (i.e. \( E(X) \) and \( E(Y) \)). What is the risk of this portfolio? The risk of an individual stock can be related to its variance. However, with a portfolio of stocks the correlation between their returns is also important. The financial analyst is, thus, interested in \( \text{corr}(X, Y) \) when evaluating the riskiness of a portfolio.

To illustrate the previous point, suppose an investor is interested in investing over the summer months in the shares of two companies: an umbrella manufacturer and an ice cream maker. Sales of these two companies are susceptible to the weather. If it is a hot, sunny summer, then ice cream makers do well (and owners of their stock make large returns). But if the summer is rainy, sales are very poor for the ice cream makers (and owners of their stock make small or negative returns). Hence, it seems like shares in the ice cream company are very risky. Shares in the umbrella manufacturer are also very risky – but

<table>
<thead>
<tr>
<th></th>
<th>( X )</th>
<th>( Y )</th>
<th>( Z )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X )</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Y )</td>
<td>0.318</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>( Z )</td>
<td>-0.131</td>
<td>0.097</td>
<td>1.000</td>
</tr>
</tbody>
</table>

*Table 1.2 A correlation matrix.*
for exactly the opposite reasons. Sunny summers are bad for umbrella sales, whereas rainy summers ensure good sales.

However, the overall portfolio is much less risky than the individual stocks. Whenever one of the stocks does poorly, the other does well. In a rainy summer, investors will earn a good return on the part of their portfolio in umbrella stocks but a bad return on the part in ice cream stocks. In a sunny summer, the opposite will occur. Hence, the investors’ portfolios will be quite safe — earning an adequate return regardless of the weather.

In statistical language, the previous example shows how the correlation between the returns on the shares in the two companies is a crucial factor in assessing the riskiness of a portfolio. In our example, this correlation was negative (i.e. whenever one stock made a good return, the other made a bad return). In practice, of course, the correlations between the returns in shares of two different companies may be positive or negative.

The previous discussion is meant to motivate why correlation is an important concept for the financial economist. To develop a formula for exactly what the population correlation is requires us to take a slight detour and introduce the concept of a covariance. Covariance is defined as

\[ \text{cov}(X, Y) = E(XY) - E(X)E(Y). \]

Previously, we have illustrated how expected values are calculated (and Appendix B provides a formal definition) and \( E(X) \) and \( E(Y) \) can be calculated as described above. To calculate \( E(XY) \), the same formula is used, except that the variable is \( XY \). The population correlation is the covariance normalized so as to have the same properties as the sample correlation (see the Properties of Correlation list near the beginning of this section and replace \( r \) with \( \text{corr}(X, Y) \)). It has the following formula:

\[ \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}. \]

Knowledge of this exact formula is only occasionally required in this textbook. However, it is crucial to have some intuition about correlation and how it depends on the variances and covariances of two variables.

As with means and variances, it is common for sample concepts to be used as estimates of population concepts. Thus, to return to our ice cream/umbrella example, the portfolio manager would be interested in knowing \( \text{corr}(X, Y) \): the population correlation between the stock returns in the two companies. The portfolio manager might collect data from the last 20 summers on stock returns for the two companies and use these data to calculate \( r \): the sample correlation. The sample correlation could then be used as an estimate of \( \text{corr}(X, Y) \).
1.5 Chapter summary

1. Economic data come in many forms. Common types are time series, cross-sectional, and panel data.
2. Economic data can be obtained from many sources. The internet is becoming an increasingly valuable repository for many datasets.
3. Simple graphical techniques, including histograms and XY plots, are useful ways of summarizing the information in a dataset.
4. Many numerical summaries can be used. The most important are the mean, a measure of the location of a distribution, and the standard deviation, a measure of how spread out or dispersed a distribution is.
5. If $Y$ is a variable that could have many outcomes, then the expected value, $E(Y)$, is a measure of the typical or expected outcome and the variance, $\text{var}(Y)$, is a measure of the dispersion of possible outcomes.
6. Correlation is a numerical measure of the relationship or association between two variables.
7. Correlation can also be interpreted graphically by means of XY plots. That is, the sign of the correlation relates to the slope of a best-fitting line through an XY plot. The magnitude of the correlation relates to how scattered the data points are around the best fitting line.
8. There are many reasons why two variables might be correlated with each other. However, correlation does not necessarily imply causality between two variables.
9. $\text{corr}(X, Y)$ is the population correlation and is a useful concept when talking about many issues in economics and finance (e.g. portfolio management).

Exercises

All datasets mentioned below are available on the website associated with this book.

1. (a) File INCOME.XLS contains data on the natural logarithm of personal income and consumption in the USA from 1954Q1 to 1994Q2. Make one time series graph that contains both of these variables.
   (b) Transform the logged personal income data to growth rates. Remember that the percentage change in personal income between period $t - 1$ and $t$ is approximately $\left[\ln(Y_t) - \ln(Y_{t-1})\right] \times 100$ and the data provided in INCOME.XLS are already logged. Make a time series graph of the series you have created.
2. (a) Recreate the histogram in Figure 1.2 using the dataset GDPPC.XLS.
(b) Create histograms using different class intervals. For instance, begin by letting your software package choose default values and see what you get, then try values of your own.

3. The file FOREST.XLS contains data on both the percentage increase in cropland from 1980 to 1990 and on the percentage increase in pasture land over the same period. Construct and interpret XY plots of these two variables (one at a time) against deforestation. Does there seem to be a positive relationship between deforestation and expansion of pasture land? How about between deforestation and the expansion of cropland?

4. Construct and interpret descriptive statistics for the pasture change and cropland change variables in FOREST.XLS.

5. (a) Using the data in HPRICE.XLS, calculate and interpret the mean, standard deviation, minimum, and maximum of $Y$ = house price, $X$ = lot size, and $Z$ = number of bedrooms.

(b) Verify that the correlation between $X$ and $Y$ is the same as given in the example in this chapter. Repeat for $X$ and $Z$, then for $Y$ and $Z$.

(c) Now add a new variable, $W$ = number of bathrooms. Calculate the mean of $W$.

(d) Calculate and interpret the correlation between $W$ and $Y$. Discuss to what extent it can be said that $W$ causes $Y$.

(e) Repeat part (d) for $W$ and $X$ and then for $W$ and $Z$.

6. People with a university education tend to hold higher-paying jobs than those with fewer educational qualifications. This could be due to the fact that a university education provides important skills that employers value highly. Alternatively, it could be the case that smart people tend to go to university and that employers want to hire smart people (i.e. a university degree is of no interest in and of itself to employers). Suppose you have data on $Y$ = income, $X$ = number of years of schooling, and $Z$ = the results of an intelligence test of many people, and that you have calculated $r_{XY}$, $r_{XZ}$, and $r_{YZ}$. In practice, what signs would you expect these correlations to have? Assuming the correlations do have the signs you expect, can you tell which of the two stories in the paragraph above is correct?

7. The file EXCHAP1.XLS contains four variables: $Y$, $X_1$, $X_2$, and $X_3$.

(a) Calculate the correlation between $Y$ and $X_1$. Repeat for $Y$ and $X_2$ and for $Y$ and $X_3$.

(b) Create an XY plot involving $Y$ and $X_1$. Repeat for $Y$ and $X_2$ and for $Y$ and $X_3$.

(c) Interpret your results for (a) and (b).

8. (a) Using the data in FOREST.XLS, calculate and interpret a correlation matrix involving deforestation, population density, change in pasture, and change in cropland.

(b) Repeat part (a) using the following variables in the dataset HPRICE.XLS: house price, lot size, number of bedrooms, number of bathrooms, and number of storeys. How many individual correlations have you calculated?
Endnotes

1. This book (like most comparable textbooks) is about analyzing data (not obtaining data). There are so many possible places from which data can be obtained, with each economic problem having its own data requirements, that it is not possible to offer more than a superficial discussion of obtaining data in this book.

2. The return that an investor earns on owning a share is its change in price plus dividends paid. Thus, the percentage change calculated here can be interpreted as the return (exclusive of dividends). Alternatively, adding dividends in to calculate the return involves a simple change to this formula.

3. The $Q$ notation denotes the quarter of the year in which the observation occurs. For instance, $1954Q1$ means the first quarter (January to March) of 1954.