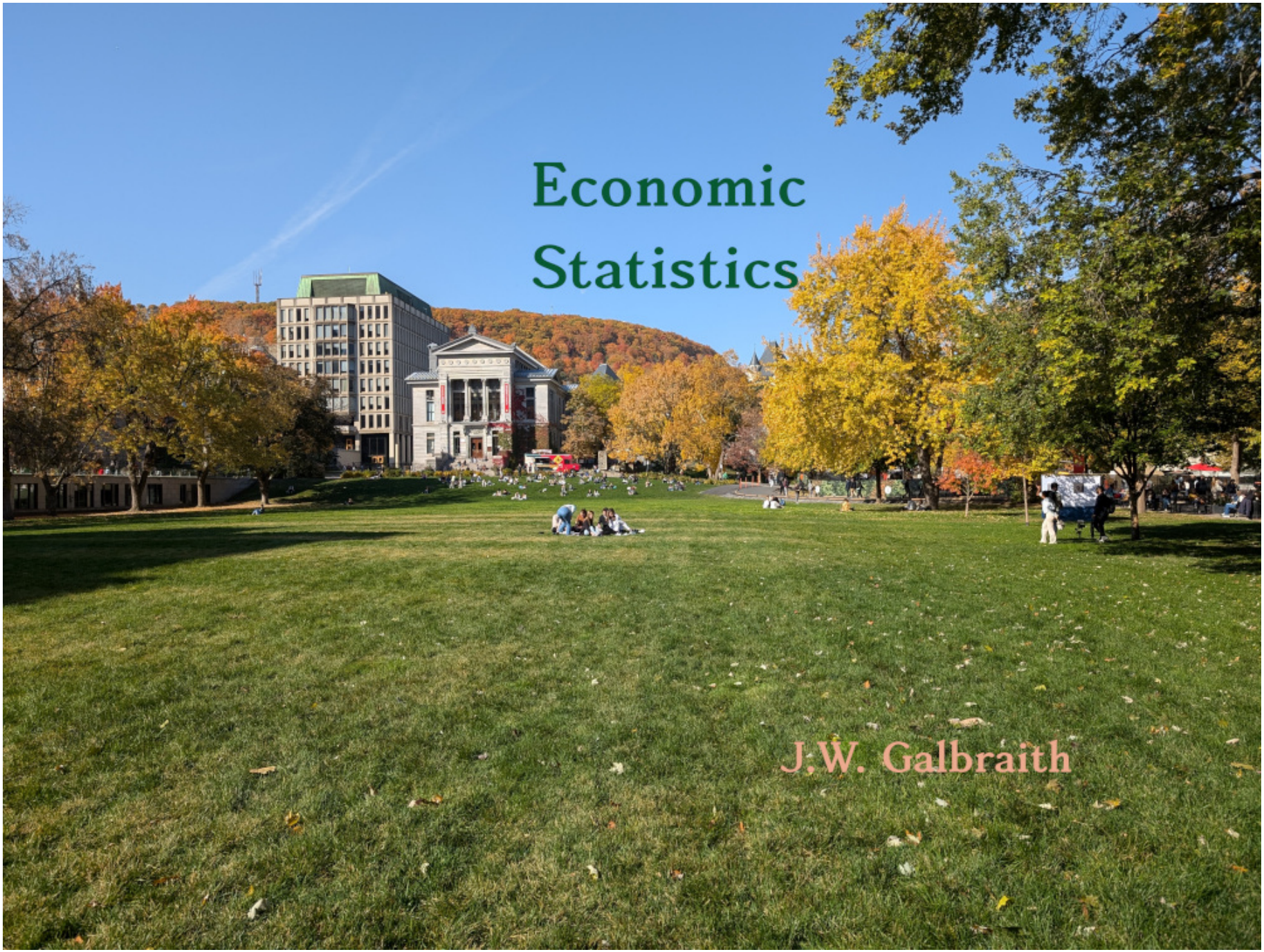

Economic
Statistics
J.W. Galbraith

Economic Statistics

J. W. Galbraith
McGill University

Version of: November 2025

Preface

I hope that these draft chapters are useful. Comments on this draft, particularly any errors that you might spot, will be gratefully received.

These chapters are for the use of students of McGill University. Please do not circulate more widely.

JWG

Acknowledgements

Sonia Laszlo, PLSS data
Simon van Norden, created Box-Whisker plots (Galbraith van Norden paper)
Douglas Hodgson, Canadian art price data

Contents

Part I: Introduction

Part II: Some theoretical foundations

Part III: Statistical methods

Part I: Introduction

Chapter 1

Statistical Reasoning

Statistical reasoning consists of more than the formal analysis of sets of data. It allows us to improve our thinking about uncertain situations, and therefore to make better decisions, or simply to understand better or predict better what is happening, or will happen. It allows us to analyze common errors in reasoning that might otherwise lead us to misunderstand uncertain situations. It also allows us to describe and communicate better what we do know about these environments.

Sharpening our reasoning is one of the purposes of studying statistics, and we will be trying to do so throughout this book. For now, it may be best to illustrate some ways in which thinking about uncertain events is difficult, but where ideas described here will help us to understand more. We will try to do this in the following examples.

First, we need to begin being clear about some terms that we will use, and in particular the word 'uncertainty'. We will mean by this any situation in which we cannot predict some event perfectly, even if someone with better information could do so (we will use the word 'random' similarly; a formal definition of a random variable will be given in Chapter 6. For example, imagine that you lend your car to friends to drive to Florida. You receive an e-mail from South Carolina saying that no one is hurt but ... there has been an accident. If your car is not fully insured for collision damage, then the extent of your loss is uncertain; there may be no uncertainty for your friends, but there is for you. Notice that the uncertainty reflects your lack of information about the situation, not some inherent property of the world; what has happened to your car is already determined and may be known to others. If your car is insured for collision with a deductible of \$1000,[1] then your uncertainty is reduced; any loss over \$1000 will be paid for, and your maximum loss is the \$1000, but may be less. We describe both of these as situations of uncertainty; the existence of some predictable component does not change the fact that the outcome is uncertain – it is not completely predictable. This use of the word 'uncertainty' is compatible with the distinction often made by economists,

[1] deductible is an amount which is deducted by the insurer from a claim settlement.

following Frank Knight (1921), between 'risk' and 'uncertainty', which we will be able to understand later when we have introduced some formal methods.

The formal study of statistics began with gambling problems, so let us do likewise.

Example 1: Returns from casino gambling or lotteries.

Anyone who gambles in casinos (unless engaging in what the casino would consider cheating, including card-memorization methods) would, if they persisted long enough, eventually be a net loser.[2] Of course, in a few rare cases where a person gets a very lucky start, he or she might need to live far longer than a human lifespan for this to happen, but eventually losses would swamp any gains. The same is true of lottery ticket buyers, although again in the rare case of someone who wins millions, it's true that a lifespan of many thousands of years of weekly lottery purchases might be necessary to end up a net loser. Nonetheless, weak laws of large numbers (*v.i.* Chapter 12) tell us that this will be the eventual result, and for most people it shows up sooner rather than later. The proliferation of casinos suggests that this point is not widely appreciated.

If a large number of people are playing a particular game of chance, we can also use statistics to give a good estimate of the proportion of them who will be net losers after playing any given number of times (again, see Chapter 12); the previous paragraph implies that this proportion will get ever closer to one as the number of rounds increases.

Example 2: Using information in uncertain situations.

The famous Monty Hall puzzle illustrates the fact that, even though we are in an uncertain situation, we may be able to use imperfect information to improve our chances of getting a desired outcome. This puzzle is named for the host, Monty Hall, of a game show ('Lets Make a Deal'), played roughly as follows.[3] There are three closed doors, behind one of which is a valuable prize (*e.g.* a car), and behind two of which are approximately-valueless prizes. The contestant picks a door, which remains closed. The host, who knows where the prize is, opens one of the doors containing the low-value items. He then offers the contestant the chance to switch her choice to the remaining unopened door, or to stay with her initial choice.

Since the prize could be behind either remaining door, it often seems that the chances are equal of winning with either door. But this is not so. Contestants who switch are in effect exploiting the information revealed when the host opened a door; they win $\frac{2}{3}$ of the time. Those who do not switch do not exploit this information, and win only $\frac{1}{3}$ of the time. (A weak law of large

[2] In a casino or lottery, a certain proportion of the bet goes to the casino or lottery organizer each time, so the bet is not a *fair gamble* : see Chapter 5

[3] You can watch old segments at **www.letsmakeadeal.com**, but it would be more instructive to go to lectures.

numbers, as mentioned above, implies that in many repeated plays of this game, the actual observed proportion of the games in which the switchers win will get arbitrarily close to $\frac{2}{3}$, whereas the observed proportion of the games that non-switchers win will get arbitrarily close to $\frac{1}{3}$.) To see why, we need to study probability; this example will be treated in Chapter 5

Example 3: Description

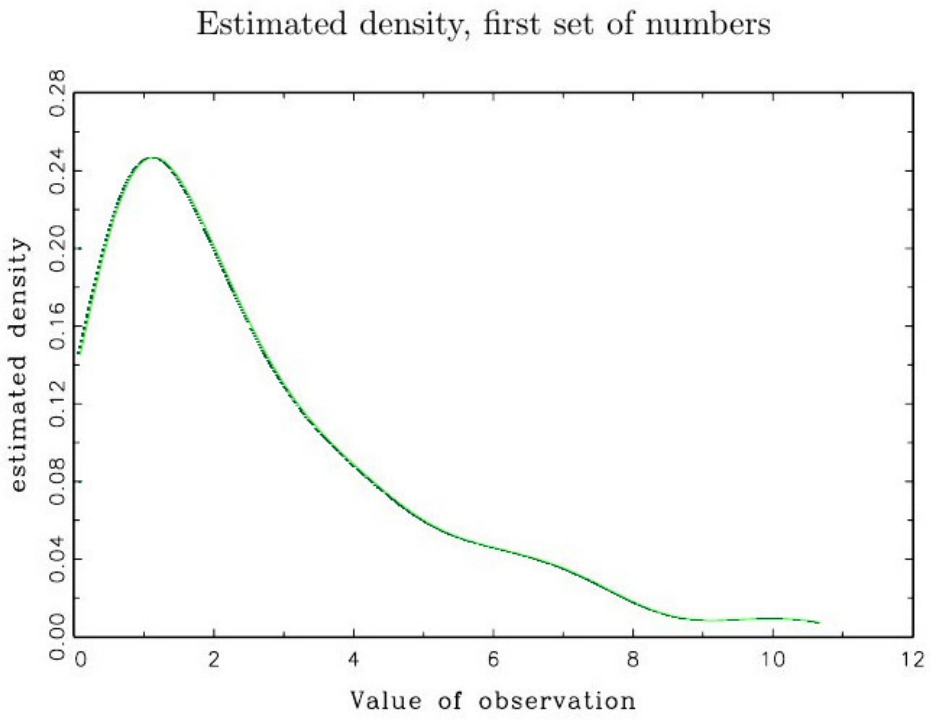
Here is a set of numbers (which we will write down only to two decimal places of precision):

7.94	1.88	7.35	0.74	0.79	6.63	3.15	9.50	3.91	2.26
0.91	1.31	0.85	3.37	4.20	1.77	0.22	0.91	1.00	2.10
1.05	2.23	0.99	0.53	10.26	5.42	3.22	2.91	1.83	0.56
2.70	0.65	0.43	2.78	1.55	2.29	3.08	2.03	0.53	4.63
7.15	2.45	4.28	2.75	0.20	2.32	0.72	2.30	2.29	1.85
0.88	3.09	0.94	1.61	0.99	2.21	2.22	1.37	1.89	1.03
0.75	1.95	3.92	1.09	3.14	1.80	0.80	6.46	2.56	2.27
5.52	4.39	7.26	9.88	0.24	0.15	1.41	2.46	1.44	0.96
3.28	0.63	1.32	0.75	1.63	1.27	7.45	0.52	5.81	5.09
0.51	3.26	6.05	0.85	4.19	2.66	0.27	1.25	5.79	4.10
0.39	0.70	1.29	0.87	2.26	6.76	1.63	1.19	3.33	3.34
0.92	7.00	0.40	1.61	0.58	1.39	7.33	3.45	0.27	3.08
3.41	1.06	2.41	1.72	6.16	4.82	0.20	1.23	0.62	3.64
0.31	1.98	1.46	5.30	1.04	0.37	3.30	6.57	8.36	0.76
1.73	0.58	2.43	0.54	4.35	1.66	0.79	5.49	1.17	1.10
4.36	0.91	0.68	0.75	0.48	3.79	0.29	0.57	6.38	2.10
0.96	2.84	4.22	1.29	2.26	10.68	0.46	1.79	1.95	0.49
5.10	1.70	1.15	2.40	1.22	5.10	4.34	3.21	0.62	1.30
0.93	6.08	4.10	0.09	5.74	2.43	3.69	1.95	0.95	1.98
4.52	1.62	2.20	3.53	1.17	0.44	0.77	4.65	0.65	4.02

You might be in a position where you expect to see more numbers representing the same thing, and you would like to be able to describe the set and any patterns that you can find. What might you have said about these numbers before studying statistics? Perhaps that they are all positive, almost all (not literally all) between zero and ten, or that values less than three seem to be particularly frequent. This could be useful, but it would be much more revealing to be able to present the density (Chapter 6) of these data, which can be estimated as follows.

Ranges of observation where the density is high are relatively likely to occur; we can see from this density that, for example, the most likely range for an observation is somewhere between one and two. Chapter 13 describes how to obtain an estimated density function like this. These data were actually generated from a χ^2_3 distribution (Chapter 9). The wavy part of this

FIGURE 1.1



estimated density for observation values above about five does not occur in the true density; Chapter 13 explains why this occurs here and how more observations would solve the problem. We could also (or instead) measure properties of these numbers that would tell us in numerical (rather than graphical) form about the location, dispersion, asymmetry, relative frequency of extreme events, and so on: Chapter 3 begins this subject.

Example 4: Learning from samples

Learning to distinguish genuine differences between measured quantities from random variation that is of no significance is a core element of statistical reasoning.

We are often presented with survey results which purport to show differences between groups or types of people, or differences over time, or samples which suggest other distinctions between two or more classes of item. A survey of 200 people (100 men and 100 women), for example, finds that 21% of men are smokers, and 24% of women. A survey of consumer confidence with 200 subjects suggests that 60% of consumers were optimistic in May, while a similar survey in June suggests only 58% were optimistic.

Is it really true that more women smoke than men? Did consumer confidence really drop in June? Maybe, but we can't tell from these surveys. Think for a minute of flipping a coin, where each flip is like one point in a sample of a coin's tendency toward the head or tail. Even if we know that the coin will fall on each side half the time, we know that in 100 repetitions of the experiment, we might well get, for example, 46 heads and 54 tails; if we do it again we might get 52 heads and 48 tails. Our sample will not reproduce exactly the true probabilities. Our samples of male and female smokers, or of consumers in May and consumers in June, might differ simply for this reason: random variation. Alternatively, there could be genuine differences. How many people would we really need to survey to be quite confident that differences like these are genuine? Part III of this book will help us to work this out; moreover, the methods described there will allow us to put a precise numerical measure on our degree of confidence: rather than being quite confident, we may be able to say that we are 95 or 98% confident.

Example 5: Memory in random processes

Consider the following statements:

i This coin has come up tails four times in a row. We're due for a change; the next one will probably be heads.

ii It's been unusually hot for two days. It will probably be hotter than usual again tomorrow.

iii It's been unusually hot for two days. It will probably be cooler than usual tomorrow.

iv The Canadiens are on a hot streak – they've won the last four games. So they'll probably beat Boston tonight.

v The stock market has been down for the last three days, so it will probably be up today.

vi The stock market has been very volatile for the last three days, so prices will probably continue to move around a lot today.

Each of these statements makes a prediction of the future of some process based on its past. Sometimes this can be done, and sometimes not.

Statement i is invalid, and in fact is an example of what is sometimes called the Gambler's fallacy, sometimes the Monte Carlo fallacy. The coin has no memory; what happened in the last few throws has nothing to do with what it will do next. Each new flip is independent (in a sense that will be made precise later; Chapter 5) of past ones. The fact that in a large number of trials the proportion of heads must come to 50% in a fair coin often leads people to think along the lines of statement 1, on the grounds that heads will have to be offset by later tails to get to 50%. But this is not quite what happens; see Chapters 11 and 12.

Either (not both of course) of statements ii and iii could be true; the weather is related to the weather in the recent past. Which, if either, of them is true is an empirical question, not something that we can work out by reasoning alone. In fact it's ii that's true; unusual temperatures do tend to have some positive persistence. Of course, statement ii simply says that relatively high temperatures are probable, not certain, following relatively high temperatures the previous day; there are plenty of occasions when a new front passes through and the weather changes.

Statement iv is a kind of statement that we hear often, and which could in principle be true or false. It is a relatively tricky one to think about. What is meant by a 'hot streak'? If, as here, it is used to mean something that has predictive power for the future, then it must mean that the team is in some condition that makes winning more likely than usual, not simply that they actually did win several games in a row. A coin could not be said to be on a hot streak in this sense: we know that three heads in a row does not make a fourth head more likelyits still $\frac{50}{50}$. Is a hockey team's hot streak just like that, or does it actually enter into phases (no injuries, players feeling happy, unpopular coach just fired, salary increases all around ...) in which winning is more likely than usual? The author does not know the answer. The question has been studied in various sports, however. Note that statement iv goes further and says that, because of this supposed hot streak, beating Boston is likely; to evaluate this statement, we would have to know how likely it normally is that Montreal would beat Boston, how much that probability has changed (if at all) in this 'streak', and what the new probability of beating Boston is. These things can all be estimated, but involve us in estimation of unconditional and conditional probabilities, which will use methods presented in a number of later chapters.

Statements v and vi are again things that could be true or false (because, unlike the coin-flipping case, we don't know exactly what the mechanism is for changes in stock prices, so we can't be sure a priori what the right answer is). In fact, daily changes in stock prices are at least close to being unpredictable from past changes. This is probably not literally true, but a great deal of statistical research in financial data suggests that it is at least a very good approximation for most purposes. So leaving aside the approximation, statement v is false. (If we had been talking about changes in a stock's price in the last few seconds, however, a statement related to v might be true; statistical tests of relations between changes over very short periods do often find evidence of some relation.) Statement vi, however, is true. While stock market price changes themselves are approximately unpredictable, periods of relatively large changes tend to be persistent: we will see an example in some data in Chapter 20.

Working out whether statements such as these are true typically requires a mixture of a priori reasoning and the careful statistical analysis of data.

Although these cases all remain uncertain, we find that in some instances we can make useful statements about what is likely to happen.

Example 6: Association

We are very often interested in whether or not two (or more) different variables are related in some way: whether they tend to move together, or opposite; whether one causes another; whether they share the same trend although they may move apart in the short term. As in Example 3 (sampling), we may see samples in which two variables appear to be associated, but is this genuine, or random variation? Chapters 15 and 16 describe how we can test this.

A well-known financial example is the apparent association between a Super Bowl victory by a team from the original National Football League (as opposed to a team from the pre-merger American Football League) and an increase in the Dow Jones Industrial Average for the rest of the year. (Both of these are binary variables – they take on only two values (up or down for the Dow Jones, yes or no for a win by an original NFL team; we could record these numerically as zero or one).[4] Establishing whether or not this association really exists brings us to problems related to data mining, including the paradoxical point that the probability of finding an association in a statistical test on a given data set will depend on whether the data were known when the hypothesis was formulated: see Chapter 15 In this case it's hard to imagine that the association could be genuine, let alone causal, although some people have been tempted to take a victory by an NFC team as a 'buy' signal (remember, of course, that the DJIA goes up most years anyway).

Apart from this problem of hypotheses suggested by data and 'tested' on the same data, we can measure the association between two variables easily (Chapter 8). But learning about causation (in general, and in particular from non-experimental statistical data) is, as David Hume (1739), (1748) famously argued, something that we can only do imperfectly and tentatively. This is an area where, once attuned to the problem, you will spot many examples of fallacious statements in second-hand accounts of research (and sometimes in the research itself) as well as in conversation; we will give specific examples below. The next section elaborates further on this.

Example 7: Conditional association

It is very often the case the two variables are associated, but not because there is any genuine affective one on the other. Instead, both of the variables may be related to some other underlying factor.

Consider the following example. A study is made of the health of male and female professors in economics departments in universities. Measures of

[4] The Dow Jones Industrial Average and related indices are described at the site https://www.spglobal.com/spdji/en.

health which are not gender specific, such as perhaps deviation of blood pressure from some optimum value or oxygen uptake per kilogram of body mass, are recorded for each person in the sample. When the results are collated and analyzed, it is found that there is a strong negative association between being male and having good measures of health: the men are much less healthy than the women. Policies are proposed to address health information specifically to male academics, studies are recommended to investigate the causes of academic male ill health, and so on. Whats wrong with this?

The existence of some association would probably be genuine in this case. But among other things, there is a problem and assuming that this association arises because of some direct relationship. Instead, the following mechanism may be at work: women have been taking PhD's and entering academic work in much larger numbers in recent years; in most academic fields the proportion of men among the oldest faculty members (those hired, say, 40 years ago) is much higher than the proportion of men among the youngest (typically those hired most recently). This is especially true in economics, at the time of writing. Therefore, the male faculty members will typically be on average older. An association between being male and having poor health would be produced simply by the fact that older people (at least within the relevant age ranges) tend to have poorer health. By looking at the health of men and women in groups where the average ages are different, we may be misled in attributing to gender and effect is in fact a result of age. There is an association between being male and having poor health, in this example, but the association conditional on taking account of age may well be zero. It is typically this conditional association that interests us, and not the unconditional Association: we want to remove the effects of other variables, and concentrate on the effect of one particular thing.

A simple way to deal with this problem would be to look at our sample by different ages; if we have a sufficiently large sample size we could consider all of the 65-year-old professors and ask whether men and women do equally well on health measures, all of the 64-year-old professors, all of the 63-year-old professors, and so on. We might find that within each age category, there is no difference in typical health measures of men and women: there is no association, conditional on age. The overall (unconditional) association between being male and having poor health is nonetheless a fact in samples of this type; it's simply not the fact that's relevant to us if we're trying to figure out whether academic men are on average less healthy than academic women.

Even if we do classify our data points by age as just suggested, this may not be sufficient: there may be other variables that we would want to 'condition on'. But what we really need is statistical technique that allows us to condition on a large number of variables, that is, remove their effects and ask whether there is a remaining component to the association once those effects have been accounted for. This is one of the things that regression

methods, and related methods, attempt to do, and is one of the main objects of interest in the related field of econometrics.

Notice that this entire problem is arising because we have to deal with non-experimental data. In an ideal controlled experiment, one changes a single quantity keeping everything else constant, and observes the effect on an outcome. It is in general impossible for us to do this in the social sciences, including economics, although there is a field of experimental economics in which experiments are conducted on consenting subjects. A great deal of statistical and econometric technique is therefore devoted to trying to 'net out', or control for, the effects of numerous variables that we cannot hold constant. For example, in the present problem, in order to run a controlled experiment holding all other factors constant, we would have to take samples of male and female professors and ensure that their lives are identical in all respects that could possibly be relevant to their health, leaving no difference between the two groups except gender. Good luck.

Example 8: Prediction

Here is another set of numbers reported to two decimal places of precision:

-0.79	-0.81	-0.48	-0.87	-0.31	0.01	0.79	-0.46	-0.35	-0.49
0.95	-0.39	-0.75	-0.02	-0.08	-0.16	-0.67	-0.78	-0.82	-0.51
0.89	0.83	-0.69	-0.43	0.93	-0.65	0.27	0.13	0.30	0.21
0.28	0.27	0.71	0.23	0.96	-0.98	0.88	-0.58	-0.41	-0.18
0.85	-0.80	0.26	-0.48	0.19	0.49	0.93	0.56	-0.78	0.38
-0.71	0.24	0.01	0.99	0.94	-0.24	0.67	-0.57	-0.07	0.12
-0.85	0.35	0.19	-0.08	0.79	-0.01	0.77	-0.71	0.61	0.07
-0.28	-0.52	0.41	0.45	-0.94	0.16	0.03	0.19	0.47	0.69
0.20	-0.96	0.71	0.71	-0.41	0.44	-0.57	-0.53	0.36	-0.12
0.29	0.25	0.78	0.23	-0.87	0.91	-0.85	-0.99	0.81	0.37

You need to make the best possible prediction of the next number in this sequence (the 100 numbers are ordered like words on a page, so that for example the last two are 0.81 and 0.37). By 'best possible', we will not mean getting the right answer each time; in practical problems, that will be impossible. Instead we will mean that you should follow a prediction rule that will lead to the lowest value of a loss function (a measure of the 'badness' of your predictions) in repeated forecasts. For this set of numbers and for standard, symmetric loss functions, the best prediction that you can make here is zero each time. This may sound strange, but it reflects the fact that the past values do not contain any useful information; see Chapter 20

Here is another set of numbers.

-0.04	-1.05	-0.04	-0.82	-0.97	-0.87	-0.06	-0.16	-0.38	-0.84
-1.01	-1.54	-1.24	-1.62	-1.13	-0.31	-0.46	0.16	-0.62	-0.40
-0.59	0.34	-0.65	0.42	-0.65	-0.25	-1.24	-1.40	-0.84	-1.29
-0.44	0.02	-0.31	0.68	0.43	0.89	-0.24	-0.32	-0.46	-0.50
0.63	0.86	1.17	0.78	0.04	0.39	-0.68	-0.69	0.25	-0.27
-0.05	0.03	0.63	0.87	1.56	0.91	0.78	0.90	0.99	1.26
1.24	0.51	-0.10	0.76	0.28	-0.36	-0.49	-0.99	0.21	-0.73
-0.39	0.08	0.38	0.02	0.61	0.93	-0.27	0.06	-0.26	-0.91
-1.08	0.08	-0.32	0.50	-0.01	-0.85	-0.35	-1.23	-0.54	0.17
-0.53	-0.59	-0.66	-0.24	-1.19	-1.58	-0.28	0.25	0.58	-0.43

This process has some 'memory' which can be exploited, so we can use past values in making a prediction. (How would we find this out? (Chapter 20). Here, the prediction rule that will do best by standard loss functions (for example, the sum of squared prediction errors or of absolute prediction errors) is to predict that the next value will be 0.5 $\times$ the last, plus 0.2 $\times$ the second last. How would you work out a rule like this? Again, see Chapter 20

These examples are intended to illustrate the point that there is much more to be gained from an understanding of statistics than just an ability to generate numbers by applying statistical programs to data. But we of course also want to generate useful numbers. First, we will look at the kinds of data that will occur in economic and financial contexts, and at a few things that it will be useful to understand before we begin to compute any statistics.

Chapter 2

Economic and financial data

Before beginning to learn about computing statistics, it will be useful to know something about the kinds of data that we will meet, and how different algebraic transformations that we can apply to them can produce new data series with different properties and can reveal different features of the original data series.

2.1 Graphical representations of data

One way to classify economic and financial data sets is as containing time series, cross sectional, or panel data. In a time series, the data are ordered (usually with a historical date attached); the order cannot be changed without changing the meaning of the time series. A cross section is a sample of units (such as people, firms, countries, and so on), and in general does not have any ordering that needs to be retained for statistical analysis. A panel combines both time series and cross sectional dimensions by following particular units through time, so that we have a set of units, each of which is observed or sampled at various points in time.

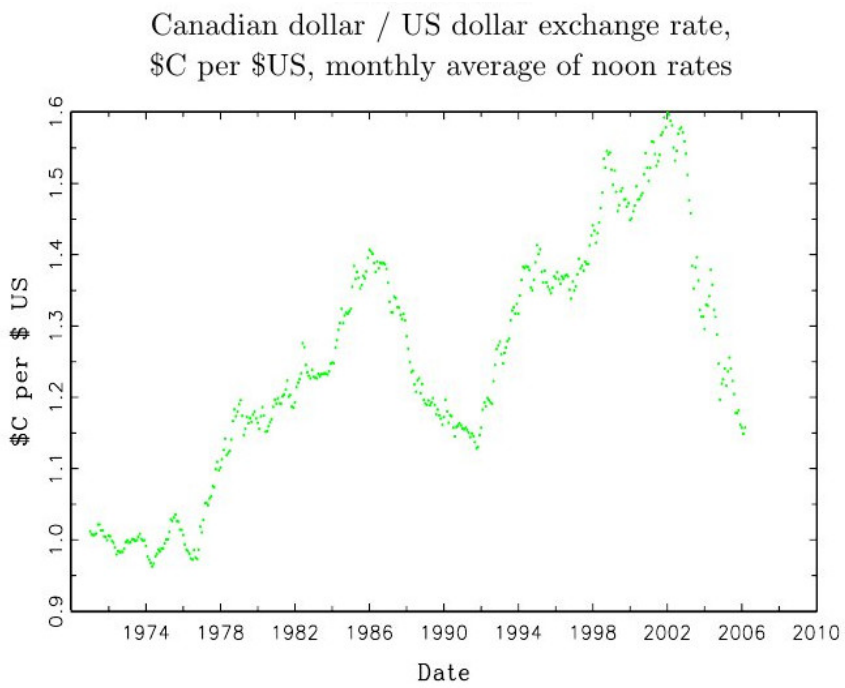
2.1.1 *Time series data*

We begin by considering data which come in an ordered sequence of observations made at different (usually equally-spaced) points in time, which typically arise in macroeconomics, finance, international economics, monetary economics, and so on. Figure 2.1.1 illustrates such a time series, the exchange rate between Canadian and U.S. dollars, recorded each month from the beginning of 1971. Monthly data may represent a single observation from a month (usually the beginning or end), or some combination of values at points within the month; in this case each monthly observation is constructed as an average of daily (12 noon) quotations.[1]

[1] These data come from the FRED (Federal Reserve Economic Data) data set maintained by the Federal Reserve Bank of St. Louis, and available at http://research.stlouisfed.org. We will use a number of data series from FRED in this chapter.

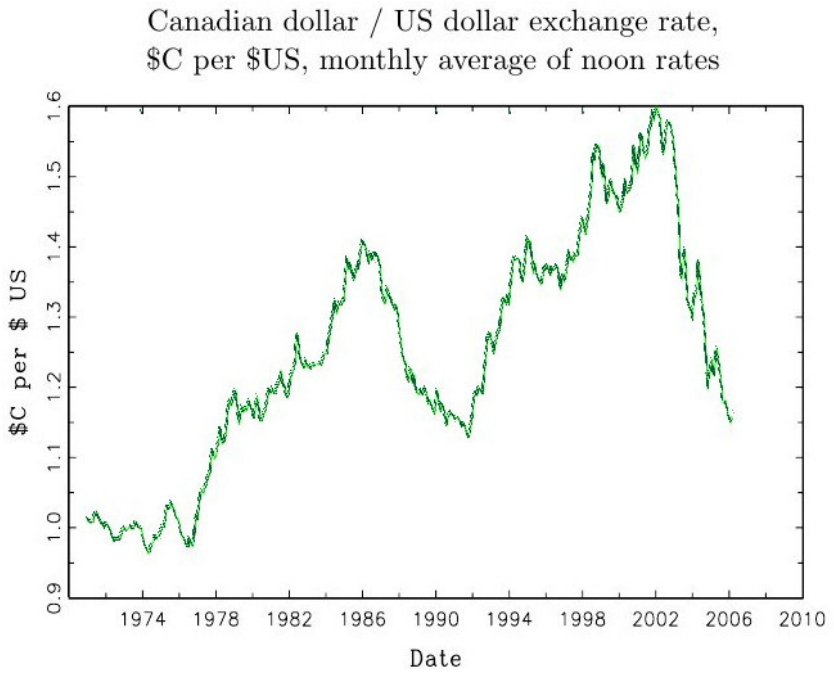
Although this graphic is straightforward to interpret, it is much more common to represent the data with an unbroken line, as in Figure 2.1.2. Although the data are in fact measured at monthly intervals rather than continuously, the continuous line sometimes makes a graph easier to read, particularly relative to one with small symbols. The line also suggests (correctly, here) that there is a progression from one point to the next along the line; in a data set with no meaningful order, this could be misleading. We will, as is common, generally use this form for plotting time series. We need to bear in mind however that the data behind the graph are not continuous, and that the graph smooths out the variation within the month.

FIGURE 2.1.1



In fact, as in Figure 2.1.3, we can plot the daily data instead of monthly average; we see that the broad pattern is of course the same, but in Figure 2.1.3 the variation within the month is more visible. However, with 8851 points plotted, the resolution of the figure on the page is now not adequate to capture all of the information in the time series, and the result is a line which appears thicker in places as the daily variation is squeezed into a small space (if we plotted the daily data of Figure 2.1.3 with individual symbols instead of a line, it would be very crowded, to the point where our sequence of points

FIGURE 2.1.2



would begin to resemble a line in any event). Nonetheless, we can see some of the extra intra-month variability which is smoothed out in the monthly data.

Note that while the data behind Figures 2.1.1 and 2.1.2 are measured at a lower frequency than those yielding Figure 2.1.3, each series represents measurements on the same underlying economic process (currency exchange between Canada and the U.S.).

2.1.2 *Cross-sectional data*

Although cross-sectional data may have no single correct ordering, it can be useful nonetheless to plot points from the data set; the relative magnitudes of different observations on a variable may be perceived much more quickly than by inspecting a table of numbers. Figure 2.1.4A plots Gross Domestic Product per capita in the year 2000 in a cross-sectional sample of 96 countries taken from the Penn World Tables, and ordered alphabetically.[2]

[2] These data are available at https://pwt.econ.upenn.edu; see Heston et al. (2002) for a description.

FIGURE 2.1.3

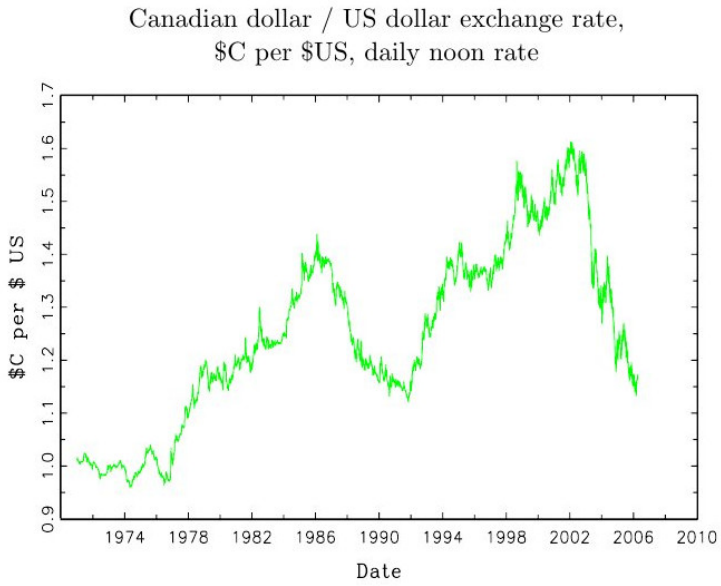
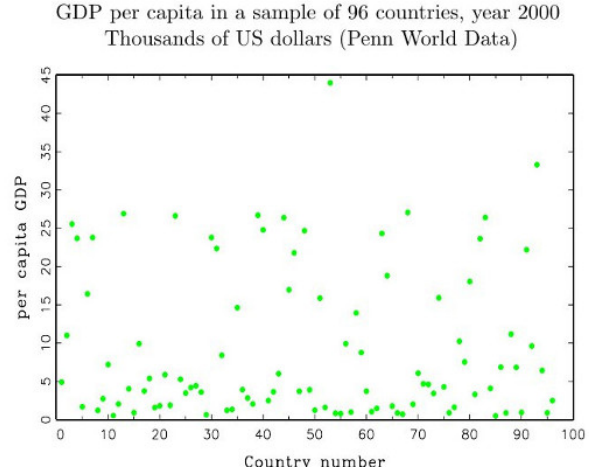


FIGURE 2.1.4A



We see immediately many countries with low income, a smaller stratum of relatively high income countries with GDP around $25,000 US per capita, and one substantially higher value near the middle of the set (Luxembourg, with annual per capita GDP of nearly $45,000 per capita in 2000). This set of

data is not a random sampling of all countries, however, and fails in various ways to provide an accurate picture of the dispersion of national incomes on this planet. In particular, the countries selected are those for which a complete set of data on real GDP per capita was available over the years 1960-2000 inclusive. Richer, developed countries tend to have much better data collection, and so this set includes virtually all developed countries, but excludes many poor, less-developed countries. (If we insisted on data available from 1950, we would lose further countries from this samplealmost all of which would be poorer countries along the bottom of the figure.) As well, this figure gives equal visual weight to large countries such as the US and small countries such as Luxembourg: in some presentations of data such as these, the figure is instead constructed to use a size of circle which is proportional to population size in each country, to give a better idea of typical outcomes for individuals on this planet.

A cross-sectional example from using data related to economic development is plotted in Figure 2.1.4B, where we have data on individuals incomes from the 1991 Peruvian Living Standards Survey.[3]

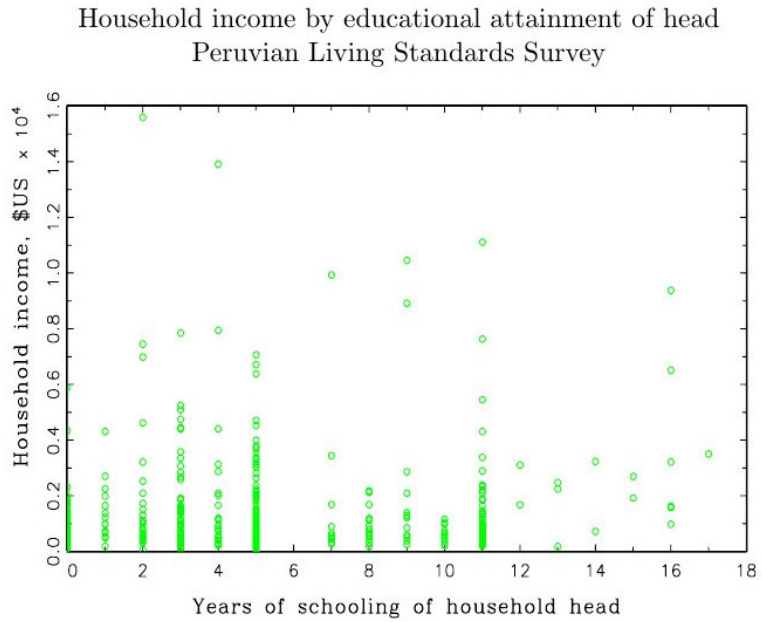
The part of this data set represented here comprises information on 423 households in rural Peru, for which a number of quantities in addition to income are measured, including number of individuals in the household, educational attainment of the head of household, number of male and female children, and other variables. Since these data points have no unique ordering, and since a number of variables are present in the data set, it is sensible to plot pairs of variables rather than the set of 423 incomes in a random ordering. In Figure 2.1.4B we plot income against years of education of the head of household. Note that Figure 2.1.4b tells us about both income and years of education of the household head; the reason that the points are in a set of vertical lines is that schooling is only measured to integer numbers of years, so no points can occur between integers on the lower axis.

2.1.3 *Panel data*

The distinguishing feature of a panel of data is that it combines both time series and cross sectional data on a particular variable (although the word panel is sometimes used loosely to describe a data set with a number of time series on different variables, this is not the technical sense that we want to distinguish, in which we have two dimensions recorded on the same variable). A panel of survey data might, for example, consist of observations over twelve years on 800 individuals, each of whom reports his or her income, employment status, years of completed education, marital status, number of children, and so on. Panels of this type (such as the US Panel Survey of Income Dynamics,

[3] The Peruvian Living Standards Survey is one of the World Banks Living Standards Measurement surveys.

FIGURE 2.1.4B



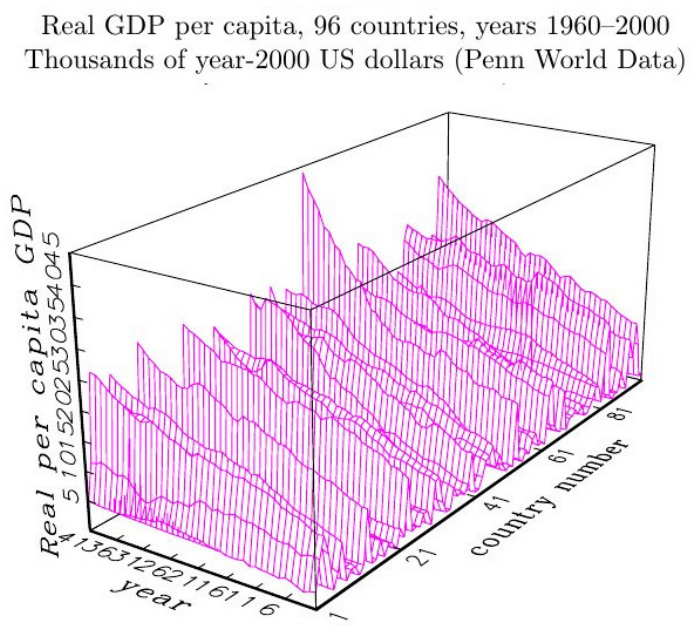
PSID, or Canadian Survey of Labour and Income Dynamics, SLID)[4] are called 'balanced' if the same individuals are in the sample at each observation date. In surveys of human beings, inevitably some individuals drop out of the survey over time, leaving a smaller part of the panel that remains balanced. Another example is given in Figure 2.1.5, which plots real GDP per capita, in thousands of year-2000 US dollars, for the sample of 96 countries mentioned above but over the full set of years 1960-2000 (note that the origin of the figure is at the bottom right corner, so that time advances from right to left in this figure).

This figure is difficult to read. It is clear that there is a great deal of variability, but the lower-income countries time series are obscured behind those of the higher-income countries. If we change the ordering of the cross-sectional dimension of the data from the alphabetical to an ordering by output in the last year of the sample, the range of time paths is easier to see, as in Figure 2.1.6.

We now see more clearly both the wide range of different incomes across countries, and the fact that there is a large set of low-income countries which remained low-income over this period. If we wish to concentrate on growth per se, however, rather than the absolute levels of income, it may be useful

[4] See **https://psidonline.isr.umich.edu** and **https://www.statcan.ca/en/start** respectively.

FIGURE 2.1.5



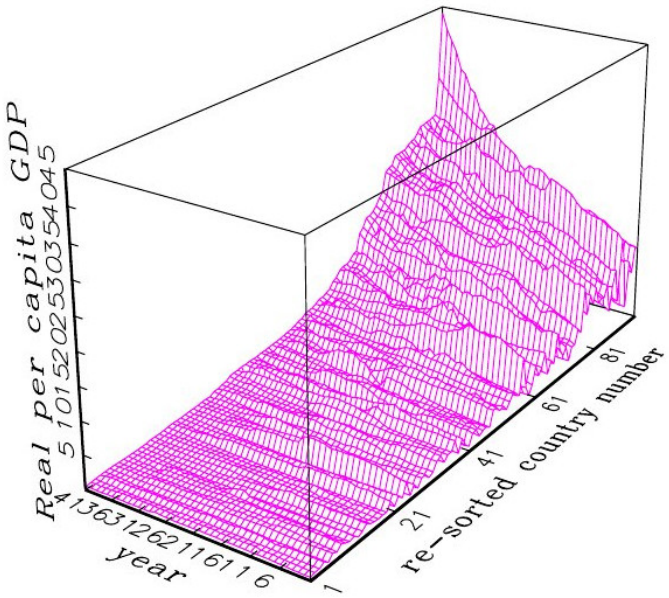
to re-scale the data to a common starting point, so that the graphic reflects proportionate growth rather than absolute levels. Figure 2.1.7 does so, by dividing each countrys individual time series of output by the value in the first year, 1960. Each countrys time series therefore begins at 1, and the figure comprises a set of lines which diverge from this starting point to values generally, but not always, above 1 by the end of the sample. A final value of 3.0, for instance, then indicates that the countrys real per capita GDP grew by a factor of 3, or by 200%, over the time period. The lines which end at the highest points, near 8 and 10 respectively, are those for the fastest-growing countries (Hong Kong and South Korea), not those with the highest absolute output levels.

Figure 2.1.7 is again difficult to read because the more slowly-growing countries are obscured. We can again choose an ordering to make the figure easier to inspect, and order by the ratio of real per capita GDP in 2000 to that of 1960, to get Figure 2.1.8.

We can see from the left side of Figure 2.1.8 that a number of less-developed countries experienced growth, followed by decline, in output; some ended the period with lower real per-capita GDP than in 1960.

Figure 2.1.6

Real GDP per capita, 96 countries, years 1960–2000, sorted
Thousands of year-2000 US dollars (Penn World Data)



2.2 Transformations of data series

At times it is easier to understand some properties of a data series by doing some simple calculations before graphing the data. We begin by looking at some of the transformations common in time series data, for which transformations are often particularly useful. Consider first of all one of the best-known economic time series, Gross Domestic Product (GDP) of the United States, a measure of total output of the US economy. Figure 2.2.1 plots this quantity, measured every three months (one quarter) from 1947 through 2005.[5] This series looks very smooth and regular. Recessions (periods of decline in output) are hardly visible in the context of the regular overall trend. In fact, this time series almost resembles a deterministic function such as

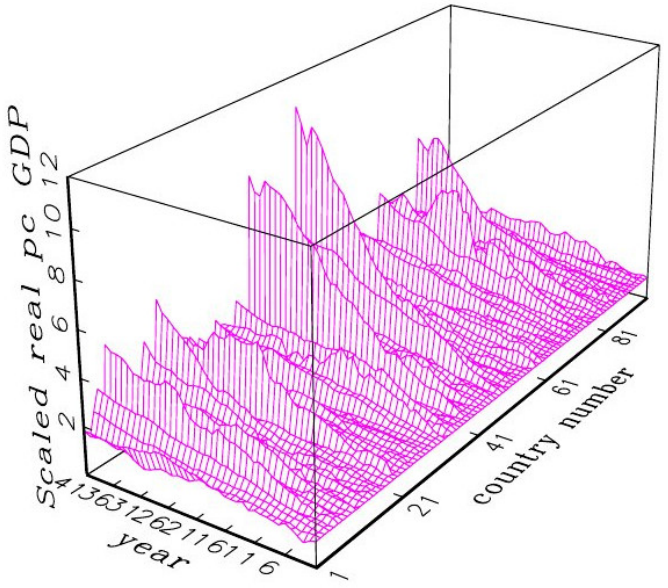
$$y_t = \alpha(1+\beta)^t, \tag{2.2.1}$$

where t is a discrete index representing time, for example $t = 1, 2, 3, \ldots$. This function grows by the factor $(1 + \beta)$ each time the time index advances

[5] These data also come from the FRED data base through the Federal Reserve Bank of St. Louis; the original source is the Bureau of Economic Analysis of the US Bureau of Commerce.

Figure 2.1.7

Real GDP per capita, 96 countries, years 1960–2000, scaled (1960=1)
Thousands of year-2000 US dollars (Penn World Data)

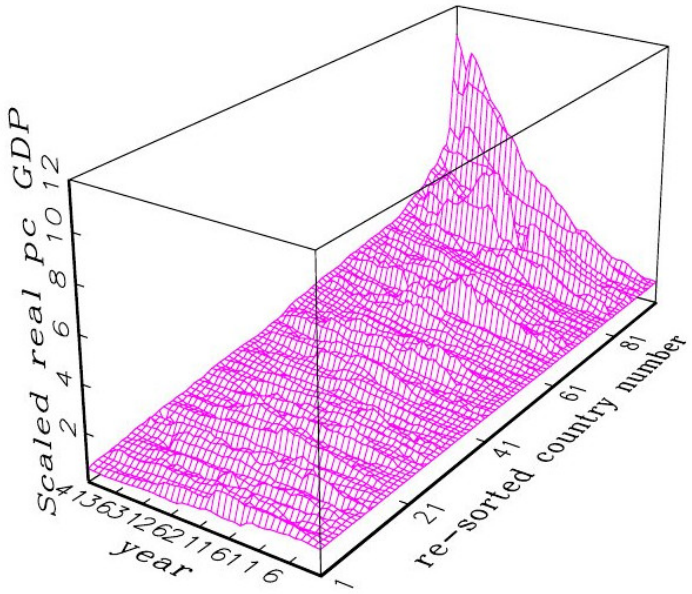


by 1, so that for, *e.g.* $\beta = 0.02$, y grows by 2% each period. This constant percentage growth implies that the absolute growth in the series is constantly rising, so that changes near the beginning of the time series come to seem trivial when plotted together with later values. This effect is obscured in these data since percentage changes near the end of the sample have tended to be smaller; in stock price data (*v.i.* Figures 2.2.7 and 2.2.8, it will be easy to see.

There are several algebraic transformations that are commonly applied to series such as GDP to reveal features that may be obscured in the raw (unprocessed, or un-transformed) data. The first transformation recognizes that GDP measured in constant dollars may be misleading for some purposes, since the value of a dollar has fallen substantially over this historical period (i.e., there has been substantial inflation). To adjust for this, we divide the nominal (measured in currency) GDP series by a price index, in this case a GDP deflator, which is designed to account for the changing value of the currency. The result is a real measure of the output of the economy, intended to remove the effect of inflation and to reveal more clearly the capacity of the economy to produce goods and services. Figure 2.2.2 plots this time series. While growth in the real measure is less dramatic, the same feature is visible

FIGURE 2.1.8

Real GDP per capita, 96 countries, years 1960–2000, scaled & sorted
Thousands of year-2000 US dollars (Penn World Data)



as in Figure 2.2.1: real GDP has tended to follow an upward curve, reflecting growth which is closer to being constant in percentage terms than in absolute terms (which would produce a straight line). Changes near the beginning of the series continue to be obscured by their small absolute size relative to later values of the series.

To understand how this is routinely handled, take the logarithm of the function (2.2.1) (the logarithmic function is reviewed in the appendix to this chapter). If $y_t = \alpha(1+\beta)^t$ then

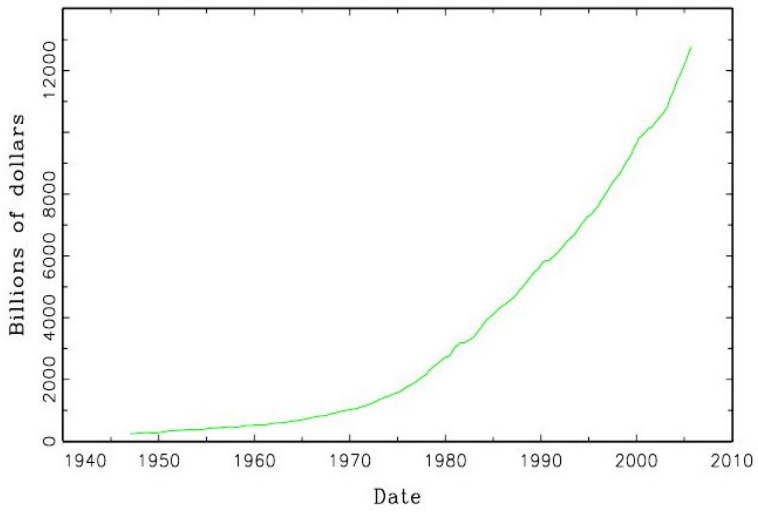
$$\log(y_t) = \log\big(\alpha(1+\beta)^t\big) = \log(\alpha) + t \cdot \log(1+\beta), \tag{2.2.2}$$

using the properties of the logarithmic function given in the appendix. Note that when the index t increases by 1, $\log(y)$ increases by *addition* of $\log(1+\beta)$ rather than by a multiplicative term of $1+\ \beta$ as was the case with untransformed y_t

That is, taking the logarithm of (2.2.1) has turned a case of constant proportional or percentage growth into one of constant absolute growth. Moreover, $\log(1+\beta) \approx \beta$ for small values of β, so that $\log(y_t)$ increases by roughly

FIGURE 2.2.1

US gross domestic product
billions of dollars, seasonally adjusted



β each time the index advances by 1. (Again, see the appendix for an explanation of this property.)

Figures 2.2.3 and 2.2.4 plot the logarithms of US GDP and real GDP respectively (equivalent to plotting the original series on a logarithmic scale). The two series are now closer to being straight lines than exponential curves (although the nominal series, Figure 2.2.3, does grow more quickly in the 1970's and early 1980's, reflecting the high inflation rates at that time), and percentage changes at the beginning of the sample are now on a visual 'level playing field' relative to those at the end: a fall of 5% at the beginning of the sample will show up as the same distance on this graph whether it occurs near the beginning, where absolute numbers are lower, or at the end.

A final transformation that is commonly applied has a more dramatic effect. Consider the change between two neighbouring points,

$$\log(y_t) - \log(y_{t-1}) = \log(1+\beta), \tag{2.2.3}$$

by (2.2.2). As we noted earlier, this is approximately equal to β for small values of β. So the difference between neighbouring values of the logarithm of this function is approximately equal to the one-period proportional change. We usually use the symbol 'Δ' for the change, so $\Delta\log(y_t\ ) \approx \beta$. Since this

FIGURE 2.2.2

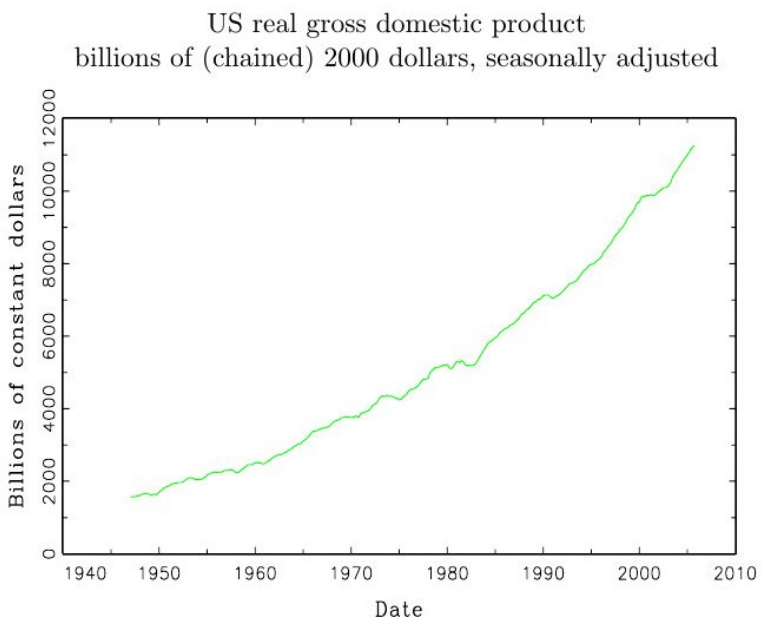
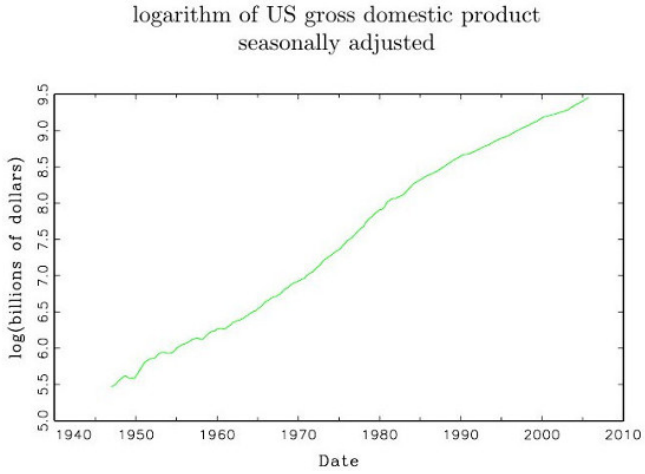


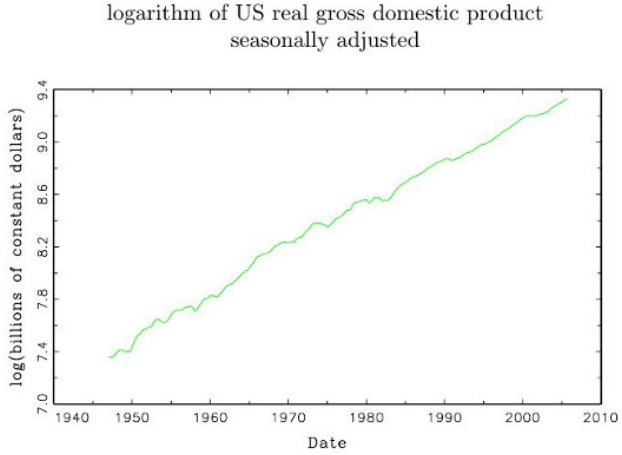
FIGURE 2.2.3



relationship is not exact, it is usually preferable to compute the proportionate change directly, however, as

$$\frac{y_t - y_{t-1}}{y_{t-1}} \tag{2.2.4}$$

FIGURE 2.2.4



To compare these two computations, note that $\Delta\log(y_t) = \log(y_t) - \log(y_{t-1}) = \log(y_t/y_{t-1})$; for growth by the proportion β (100 β%), $y_t/y_{t-1} = 1+\beta$, and so $(y_t - y_{t-1})/y_{t-1} = y_t/y_{t-1} - 1 = \beta$. By contrast, $\Delta\log(y_t) = \log(1+\beta) \approx \beta$, but $\log(1+\beta)$ is not exactly equal to β unless $\beta = 0$. If the aim is to compute the true proportionate growth rate each period, equation (2.2.4) gives the exact growth rate, and (2.2.3) the approximation.

Equation (2.2.4) gives the transformation represented in Figures 2.2.5 and 2.2.6, again for nominal and real US GDP respectively; if we had instead used the difference in logarithms, by (2.2.3), the resulting figures would be hard to distinguish visually except at a few points. We now see much more detail in the fluctuations of GDP than was apparent in any of the previous figures, although the same information was present in another form in the earlier figures.

Notice that measured output growth has tended to fluctuate less over time.

The various sequences of observations plotted in Figures 2.2.1–2.2.6 all represent different time series, although they are all based on the series of measurements of US GDP. These series have very different properties, which may be important in analyzing data of this (time series) type. Chapters 18 and 20 emphasize a number of these differences, and ways in which data of these types may be much trickier to analyze than non-time-dependent samples.

Before we leave this topic, we noted above that changes near the beginning of a time series tend to be obscured in a sample for which proportionate growth, rather than absolute growth, is stable over time. Consider the Dow Jones Industrial Index over the years 1915 through 2005, in Figure 2.2.7, and the logarithmic transformation of this series, Figure 2.2.8. In the raw data of 2.2.7, the speculative boom through mid-1929 (3 September 1929, 381.2) and

FIGURE 2.2.5

Proportionate change in US gross domestic product
seasonally adjusted

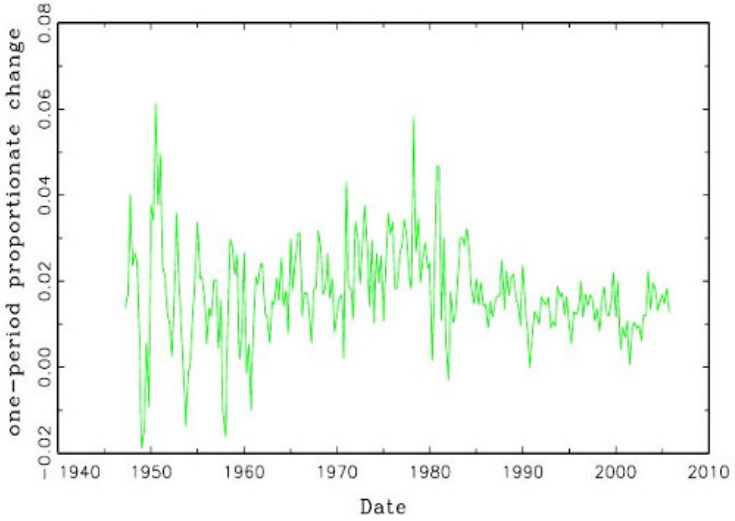
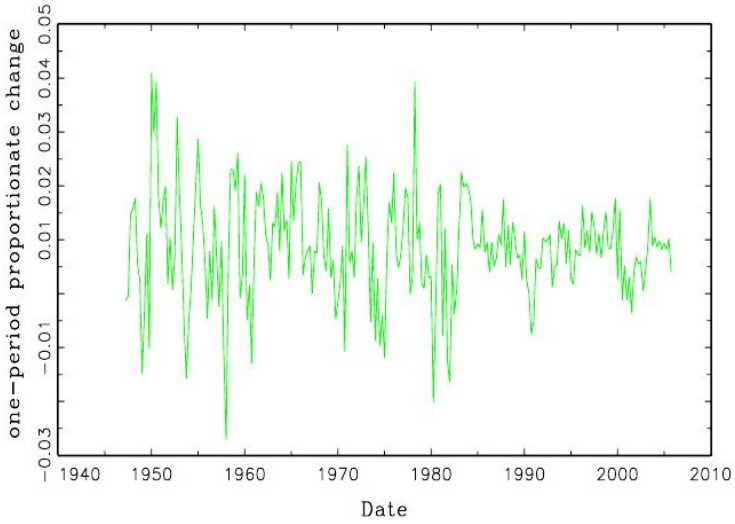


FIGURE 2.2.6

Proportionate change in US real gross domestic product
seasonally adjusted



the long decline to the bottom during the depths of the Depression in 1932 (8 July 1932, 41.2: a decline of 89%) appear as a small blip; this is because the absolute decline of 340 points from peak to trough is small in the context of recent index values. In the logarithm of the data plotted in 2.2.8 this period shows up, as it should, very clearly.

FIGURE 2.2.7

Dow Jones Industrial Average

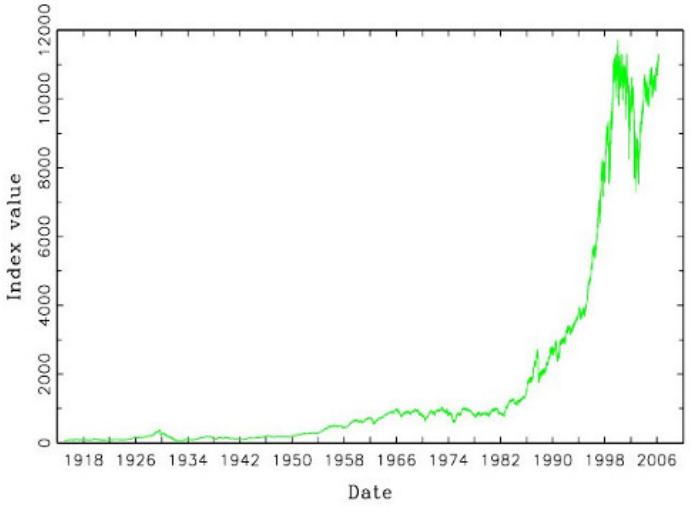
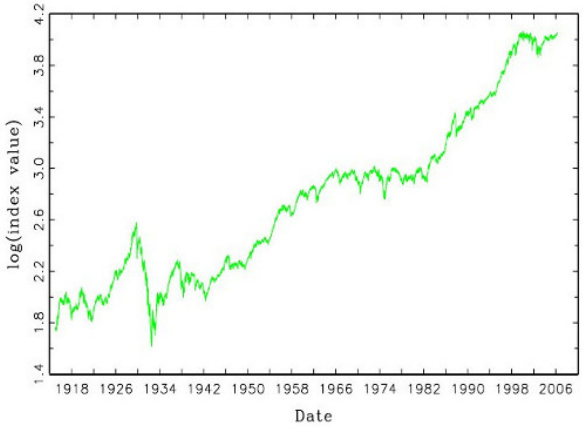


FIGURE 2.2.8

logarithm of Dow Jones Industrial Average



Data transformations usually play a smaller role in cross-sectional data (or the cross-sectional dimension of a panel); the individual observations are usually in some sense directly comparable, unlike time series observations which are made at different points with different price levels, and so on. Nonetheless, particularly for visual presentation, some transformation may be convenient. For example, a graph depicting the incomes of different individuals may be difficult to read if there is one extremely high income person in the sample; other individuals data points will be concentrated near zero on a scale wide enough to accommodate an income of, say, $5 million per year. In this case, a logarithmic transformation of the income data may be visually useful.

2.3 Some data sources

The quality and variety of some types of economic data have improved dramatically as electronic recording and storage of large numbers of observations have become routine. Many types of data, however, must still be collected manually, by survey. Some data are also recorded experimentally, using voluntary subjects in laboratories.

2.3.1 *Financial data*

Financial data, particularly concerning asset prices (and related transformations such as rates of return) are among the highest-quality data available. For many assets traded on major exchanges, each transaction is recorded with time, date, and other associated information. Regularly-spaced samples from the data, for example daily closing prices of companies shares, or hourly mid-market exchange rate quotations, are often available in reasonably long historical time series. While these data are generally very precise (that is, recorded to several significant digits) and accurate (that is, free of errors), errors and omissions do arise because of occasional system failures, or other anomalies.[6] Data on interest rates (rates of return on bonds of different types), exchange rates, commodity prices, options and other derivative contracts are of similar quality. Note however that some assets (e.g. gold, US dollars) trade on more than one exchange, and that arbitrage keeps prices similar but not identical on different exchanges. The price of gold as traded on the London Bullion Market, the Chicago Board of Trade and the New York Mercantile Exchange is not exactly the same thing even if transacted in a common currency. There is no unique time series giving the price of gold. But the U.S. dollar price

[6] For example, the daily closing Dow Jones Industrial Average data recorded in Figure 2.2.7 were put together from two sources; one of the sources was missing an observation in 2004, while the other was a shorter time series. Data from the two were checked against each other and spliced to obtain the long time series.

time series on London, Chicago and New York exchanges will only differ by small amounts. Financial data may also describe cross-sectional samples of properties of firms or households. Data such as these will generally be obtained by survey, are subject to substantial reporting or recording error, and will generally have neither the precision nor the accuracy of high-quality asset price data. Nonetheless such data may contain a great deal of information.

2.3.2 *Macroeconomic data*

Macroeconomic data come from a variety of sources, but apart from cases where they overlap with financial data (exchange rates and interest rates, for example) and so are recorded in asset markets, macroeconomic data are typically obtained from surveys or administrative sources. A number of the most interesting data series, such as those pertaining to national income or output (gross domestic product, industrial production) are obtained by responses of private enterprises to survey questionnaires, which are used to estimate an overall output measure for the economy. These data are usually revised at least twice following the initial estimate, and the revisions are substantial relative to the fluctuations in the series themselves. Figure xxxx shows a sequence of initial, first revision, and final revision estimates of US GDP; while the final estimates are probably the most accurate, it is clear that substantial measurement error in this series is inevitable. Estimates of the unemployment rate, consumer (and other) price levels, money supply, and numerous other series are subject to similar considerations.

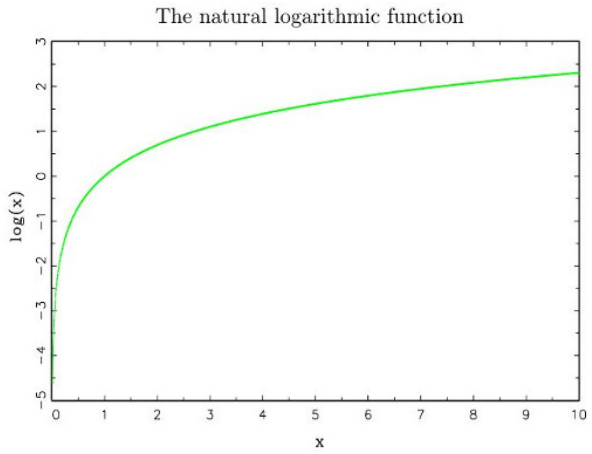
Some data, such as new unemployment insurance claims, do not come from surveys but from recording of individuals direct dealings with governments. While still subject to error, such data series at least attempt to record every relevant observation rather than using surveys of a subset of relevant cases.

2.3.3 *Data on individuals*

In many areas of economics such as labour, development and health economics where the behaviour of individuals is being studied, the relevant data typically pertain to individual human beings. In some cases such data are recorded on a large scale by governments: income tax data provide an example, but also illustrate the fact that the use of such data is typically subject to restrictions to protect confidentiality. Most data sets concerning individuals are however collected from individuals who participate voluntarily in surveys. Responses will typically show some *sample-selection effect* ; see Chapter 21.

The logarithmic function

Figure A2.1



A logarithmic function is the inverse of an exponential function. An exponential function is one that has the form

$$y = a^z,$$

where $a > 0$ and $a \neq 1$; z may be any real number ($z \in \mathcal{R}$), and therefore y may take on any strictly positive real value ($y \in \mathcal{R}^{++}$). The logarithmic function with base a applied to y gives back z: $\log_a(y) = z$ (*i.e.* $\log_a(a^z) = z$), and conversely $a^{\log_a(y)} = y$.

For $a > 1$, the exponential function increases without bound and has an increasing slope as z increases; the logarithmic function with base $a > 1$ also increases without bound, but with decreasing slope, as z increases.

A very commonly used value of the base is the constant e, defined as

$$\mathrm{e} = \lim_{h \to 0} (1 + h)^{1/h},$$

or, as a convergent power series:

$$\mathrm{e} = \sum_{n=0}^{\infty} \frac{1}{n!} = 1 + 1 + \frac{1}{2} + \frac{1}{6} \ldots.$$

This constant has the property that the exponential function based on e is its own derivative, *i.e.* if $y = \mathrm{e}^x$, then $\mathrm{d}y/\mathrm{d}x = \mathrm{e}^x$. In general for a base a,

if $y = a^x$ then $\mathrm{d}y/\mathrm{d}x = ka^x$, for some constant k which is not equal to unity unless $a = \mathrm{e}$.

The logarithmic function with base e is called the *natural* logarithmic function and is plotted in Figure A2.1. It is often written simply as $\log(y)$ with no base indicated; the symbol $\ln(y)$ is also used.

In addition to the properties given in section 2, note that, with all logs to an arbitrary base a,

$$\log(y^h) = h \log(y) \text{ and } \log(yz) = \log(y) + \log(z)$$

With h = -1, we can see that $\log(x^{-1}) = \log(1/x)$ = -log x, and so $\log(y/z)$ = $\log(y.1/z) = \log(y)$ - $\log(z)$. Logarithms taken to different bases differ by a multiplicative constant:

$$\log_b y = \log_a y \log_b a = \log_a y / \log_a b,$$

and this implies that $\log_b a = 1/\log_a b$.

Chapter 3

Elementary data description

It is often useful to have simple numerical measures of the properties of data series, even when plots of the data or other descriptions of the entire set of data may be available. Simple summary measures are useful in communicating properties, in making general comparisons, and in reasoning about data. For example, we may argue that a certain change in financial markets will tend to make the fluctuations in returns more pronounced; this will be hard to see in 'before' and 'after' graphs of sequences of returns unless the change is dramatic, but a number measuring the typical size of fluctuations may increase. Similarly, we may wonder if undergoing a training program is associated with higher wages for workers. Plotting the data on a thousand workers randomly selected for the program, and another thousand who were not selected, is unlikely to reveal a clear distinction given the many other differences among workers. But if we calculate an average wage for both groups, a difference may well be discernible.

Many such measures, or statistics, have been defined. (Note that a statistic is a quantity calculated from data, as opposed to an observation from a data set.) In this chapter we will review only a few of the most commonly used statistics, to introduce the subject and to help us in discussing other topics. Our discussion in this chapter will refer only to the observed data, and not to any theoretical quantities that these statistics may measure; in later chapters, when we have defined the relevant theoretical concepts, we will begin to link these measures to them, and we will see that these quantities can be understood as estimates of underlying theoretical properties of distributions.

3.1 Measures of location or central tendency

The most common statistics used to describe data are those that indicate where a typical observation lies: where the 'centre' of a set of observations is located. Some of these statistics are sufficiently well known to be used (albeit sometimes imprecisely) in common speech.

In order to define terms precisely, we now need clear notation to describe a data set. Let n be the number of observations in our data set, and let these observations be labelled $x_1, x_2, \ldots, x_n$. Standard notation for summation uses the symbol $\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \ldots + x_n$; i is called the index, and we say that i indexes the set of n observations.

Here are a few simple measures of location.

D3.0 Sample mean:[1]

$$\bar{X} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i.$$

To define the next measures, we need to sort the data and have notation for the ordered data set. Let the observations sorted from smallest to largest be labelled $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, so that $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$: these are called the *order statistics* of the data. We can then define measures that refer to points in the ordered set.

D3.1 Trimmed mean:

$$\bar{X}_{[k,k]} \equiv \frac{1}{(n-2k)}\sum_{i=k+1}^{n-k} x_{(i)}.$$

As we have defined it here, the trimmed mean drops the k smallest and the k largest observations from the sample, leaving n-2k. The trimmed mean is also often defined as the estimator that drops a certain percentage of the sample at the small and large ends rather than a fixed number of points. In the latter case, the percentage must tend to zero as the sample size increases in order for the estimator to converge to the true mean. Trimmed estimators are sometimes useful where data may contain occasional extreme observations which can have a large effect, in small or moderate samples, on the estimated mean; trimming makes the estimator more robust.In some cases asymmetrical trimming is used (differing numbers or percentages are trimmed from each end of the sample).

D3.2 Sample median:

$$\hat{q}_{0.5} = \begin{cases} x_{(j)}, & j = (n+1)/2, \text{ if } n \text{ is odd} \\ \dfrac{x_{j)} + x_{j+1}}{2}, & j = n/2, \text{ if } n \text{ is even.} \end{cases}$$

The sample median describes a point at or below which half of the observed data lie (with an odd number of data points, of course, we cannot find

[1] This is the sample arithmetic mean. The sample geometric mean is

$$\left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

that is, $(x_1 \cdot x_2 \cdot \ldots \cdot x_n)^{1/n}$, which is appropriate in cases where the relevant total is given by a product rather than a sum, as in compound growth.

such a point, so we take the middle point as an approximation). We can divide the ordered data into other equal parts such as fourths (the quartiles), fifths (quintiles), tenths (deciles) and hundredths (percentiles); in general, these are called **quantiles**.[2]

Different definitions of sample quantiles or sample percentiles are used. A simple one takes the α quantile as the smallest point in the sample such that at least a proportion α of the data lie at or below that point: that is,

D3.3 $\hat{q}_\alpha = \min_j(x_{(j)})$ such that $j/n \geq \alpha$.

For example if $\alpha = 0.682$, then $\hat{q}_\alpha$ is the smallest order statistic (the smallest value in the sample) which is such that at least 68.2% of the sample values are less than or equal to that value. If we have the nine (ordered) data points {0.12, 0.38, 1.57, 2.02, 2.93, 3.45, 5.12, 6.74, 8.55} then $\hat{q}_{0.682} = 5.12$.[3]

Notice that the median (which is usually thought of as synonymous with the 50th percentile) as defined in D3.2 is not a special case of this: if we applied D3.3 to compute the fiftieth percentile ($\alpha = 0.5$), we would get

$$\hat{q}_{0.5} = x_{(j)}, \qquad \begin{array}{ll} j = (n+1)/2, & \text{if } n \text{ is odd} \\ j = n/2 & \text{if } n \text{ is even.} \end{array}$$

Although D3.3 corresponds well with a theoretical quantity which we will define in Chapter 6, we may, particularly in small samples, sometimes wish to use an alternative definition of a quantile such as is implicit in D3.2 for the median. We will consider examples below after we have studied the concept of a cumulative distribution function.

To illustrate some features of these statistics, let us return to the set of numbers used in Example 5 of Chapter 1, and compute these measures. The sample mean and median are respectively 2.58 and 1.88. That the mean exceeds the median reflects the fact that the data have the long upper tail that we saw in Figure 1.1; observations above the median tend to be farther from it than observations below, so they increase the sample mean more than observations below the median lower it. (Another way of thinking about this is to note that replacing a moderately large observation with a very large one increases the mean, but doesn't affect the median; the large observations in the upper tail raise the mean, but the median is just the same as it would be if the top observations were cut down to just above the median.)

[2] Note the distinction between *quartiles* and *quantiles*; quartiles, like deciles and percentiles, are a particular case of the general concept of quantile. We can discuss, for example, the quantile 0.473. The *quantile* 0.25 is the first *quartile*.

[3] We obtain this value because 68.2% of our sample size = $0.682 \times 9 = 6.138$, so we have to have at least 6.138 data points below our 0.682 quantile. We cannot have a fractional number of data points, so the smallest value that gives us at least 6.138 points is the seventh; the seventh order statistic is 5.12. (Alternatively, if j/n is $7/9 = 0.777$, this is big enough, but if j/n were $6/9 = 0.667$, this would be too small: we need at least 0.682 of the data below the chosen point, so j must be seven, and the seventh point is 5.12).

3.2 Measures of dispersion

After learning about the centre of a set of data, the next thing that we typically want to learn about is the degree of dispersion about the centre; are they heavily concentrated there, so that most observations lie quite close to the sample mean, or are they widely dispersed, so that knowing the sample mean says little about where a typical point lies?

Two of the most commonly used measures directly describe dispersion around the sample mean, so the sample mean enters their definitions.

D3.4 Sample variance:

$$s^2 \equiv (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{X})^2.$$

D3.5 Standard error: [4]

$$s = +\left[(n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{X})^2\right]^{1/2}.$$

Note that these definitions divide by n 1 rather than n, although we are averaging n items (the squared deviations of each observation from the sample mean). The reasons for doing this are slightly subtle and depend on definitions not only of these sample statistics, but also of underlying theoretical quantities that will be introduced later; Chapter 11 explains.

The sample variance is the typical squared deviation of a point from the centre. If (*e.g.*) the sample mean is zero, and observations are at ± 0.1, then the sample variance will be 0.01. The standard error, however, would be 0.1, and so is a more directly interpretable measure of the typical deviation of a point from the centre of the data.[5]

Another set of dispersion measures makes no reference to the sample mean, but is related instead to the percentiles of the data. To define these, we again need to refer to the order statistics $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ rather than to the original, unsorted, data.

D3.6 Range: $x_{(n)} - x_{(1)}$.

D3.7 Interquartile range: $q_{0.75} - q_{0.25}$

We could of course define other measures, taking differences between any two percentiles, but the interquartile range is commonly used.

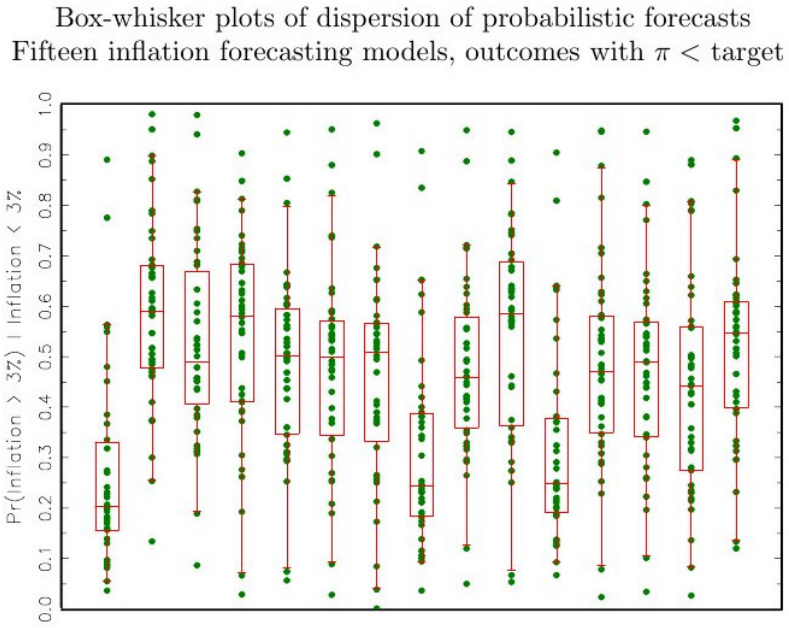
Sometimes it is useful to combine some of these percentile-related measures in a graphical form. A **box plot** or **box-whisker plot**, for example, plots

[4] The '+' denotes the positive square root.

[5] The French term for this measure, *écart-type* (literally 'typical gap'), is more descriptive.

the median, first and third quartiles, and more extreme parts of the data. A box covers the interquartile range from $q_{0.25}$ to $q_{0.75}$, the median is indicated by a line in this box, and lines ('whiskers') move outward; not all users define the whiskers in the same way, however. The original definition (Tukey 19xxx) defines the whiskers as moving out to the minimum $x_{(1)}$ and maximum $x_{(n)}$ of the data, unless these points lie more than 1.5 times the interquartile range from the median; in the latter case, the whiskers move out only to 1.5 times the interquartile range from the median, and larger or smaller observations are indicated as isolated points. The whiskers are sometimes defined instead as moving outward from the box to fixed percentiles of the data such as the 10th and 90th, or 5th and 95th. An example is given in Figure 3.2.1, which shows sets of forecasts of the probability that inflation will fall below a 3% target, for sets of cases in which inflation did in fact turn out to fall below the target. The fifteen plots show the results from fifteen forecasting models, with box-whisker plots superimposed on the observed forecasts. (These data are from Galbraith and van Norden 2011.)

FIGURE 3.2.1



3.3 Measures of skewness and kurtosis

A set of data that is symmetrically distributed has the feature that the patterns describing the data above and below the mean are mirror images of each other (*i.e.* the data are distributed symmetrically around the mean). Data such as those depicted in Figure 1.1 are said to be skewed rather than symmetric; on one side of the mean, the data are dispersed more widely. A sample measure of skewness can be defined using the standard error (D3.5) as a scaling factor.

D3.8 Coefficient of **skewness**:

$$\frac{[(n-1)^{-1}\sum_{i=1}^{n}(x_i-\bar{X})^3]}{s^3}.$$

The purpose of scaling by the cube of the standard error is to concentrate on skewness, removing the effect of dispersion: the numerator of the expression will tend to be larger for higher-variance data sets, but we want to abstract from this feature and concentrate on asymmetry alone. Dividing by the cube of the standard error is one adjustment that we can perform to do this.

Note that as the variance used the sum of the second power of terms in $(x_i-\bar{X})$, the coefficient of skewness uses the third; the measure of kurtosis that we will define below uses the fourth, and each of these can be seen as estimates of *moments* of the underlying distribution of data, which will be defined in Chapter 7: that is, each of these measures can be seen as an estimate of an underlying theoretical quantity.

The theoretical value of the coefficient of skewness is zero for a symmetrically distributed set of data; however the converse is false (a coefficient of zero does not imply symmetry – it is possible that different patterns on either side of the mean can nonetheless lead to positive and negative terms exactly cancelling in the numerator of D3.8).

Recall also that the mean is equal to the median for a symmetric distribution. Another measure of skewness is based on this relation:

D3.9 Alternative skewness measure:

$$(\bar{X}-\hat{q}_{0.5})/s.$$

In words: mean minus median, divided by standard error. Again, symmetry implies that the true theoretical quantity that this measures will be zero, but the converse is not true. This quantity will always lie in the interval [1, 1]. Sums of higher powers of $(x_i-\bar{X})$ can also be given meaningful interpretations (note that the variance and coefficient of skewness were based on sums of second and third powers of this quantity). In particular, the fourth power is often used as follows:

D3.10 Coefficient of **kurtosis**:

$$\frac{(n-1)^{-1}\sum_{i=1}^{n}(x_i-\bar{X})^4}{s^4}-3.$$

This is useful as a measure of the frequency of extreme events relative to the *Normal distribution*, (Chapter 7), for which the theoretical quantity measured by

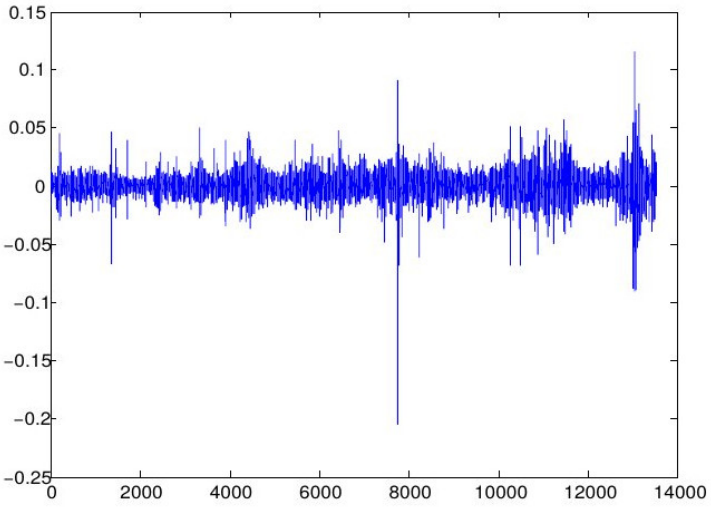
$$\frac{(n-1)^{-1}\sum_{i=1}^{n}(x_i-\bar{X})^4}{s^4}$$

is equal to three; therefore D3.10 measures the excess of this over the theoretical value for the Normal. We will study this property in Chapter 6

To illustrate these measures on a readily-available data set, consider the S&P500 index, which we will label St at a time index t. With daily data from 2 January 1957 through 30 September 2010, we can compute the daily returns $(S_t - S_{t-1})/S_{t-1}$ beginning with the second day. Here is a plot of the 13531 data points on returns:

FIGURE 3.3.1

Daily returns on the S&P 500 index, 2 January 1957–30 September 2010



Multiplying by 100 gives daily percentage returns. The sample mean X of these daily percentage returns is 0.0287 (that is, on an 'average' day, the S&P500 rises by 0.000287 or 0.0287 percent). The trimmed mean dropping the five smallest and five largest returns is 0.0294; the fact that it is slightly larger reflects the fact that the smallest returns (*i.e.* the largest daily losses) have been greater in absolute value than the largest returns.

The standard error of the percentage returns is 0.994 and the mean absolute return (the sample mean of the absolute values of these data) is 0.669: although the mean return is just slightly above zero, this is the mean of many positive and many negative returns which approximately offset each other; a typical days change is well away from zero.

The coefficient of skewness is -0.642, a fairly small degree of leftward skewing. The coefficient of kurtosis is however 22.39, far above the value 3 that applies to a Normal distribution, reflecting the fact that financial return data such as these typically show a much higher frequency of extreme events than would a Normal distribution with the same mean and variance. If one were to fit a Normal distribution to these return data to predict risk, it would provide a very poor characterization and would lead to severe underestimates of the probabilities of large losses or gains. In Chapter xxxx, after parametric distributions such as the Normal have been discussed, we will return to this example to illustrate this point.

3.4 Measures of association

The descriptive measures given so far apply to single series of data. Often we want to describe the association between two data series: do they tend to rise and fall together, for example? Commonly used measures of association are the covariance, an extension of the concept of variance given above, and correlation, which scales the covariance into a fixed interval.

D3.10 Sample covariance between variables X and Y:

$$(n-1)^{-1}\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y}).$$

Note the analogy to the definition of the sample variance above: in that definition the term $(x_i - \bar{X})$ appears multiplied by itself (that is, squared) whereas in this definition the second of these terms is replaced by the corresponding quantity and the other variable. The sample variance of the variable is therefore analogous to the sample covariance of the variable with itself.

This raw sample variance can be difficult to interpret: a small value may be consistent with the strong association, or a large value with a weak association. The covariance depends upon the variances of the underlying variables, and so those variances need to be known or estimated in order to judge whether a particular sample code variance is a 'large' or 'small' value. The correlation deals with this by scaling the sample covariance, using the square roots of the sample variances of the two underlying variables.

D3.11 Sample correlation between variables X and Y:

$$(n-1)^{-1}\frac{\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})}{s_X s_Y},$$

where s_X and s_Y are the standard errors (square roots of the sample variances) of X and Y respectively. The correlation lies in the interval from -1 to +1.

A positive correlation means that when one of the variables is above its mean, the other variable will tend to be above its mean as well; the closer is the correlation to one, the less likely it is that one of the variables will be below its means and the other is above. By contrast, a negative correlation means that when one of the variables is above its mean, the other variable will tend to be below its mean. It's important to bear in mind that covariance, correlation, or other measures of association do not imply causation. For example, in a large sample of people who are surveyed about their health, we might find that there is a strong negative correlation between the amount of coffee that an individual consumes and the typical number of hours sleep that he or she gets. Does this imply that drinking coffee (even in the morning) causes people to sleep less overall? Not necessarily: this might be true, or it might be that people who sleep less well drink more coffee in an attempt to compensate, or it might be that some third factor tends to produce the association, for example that people with high-stress jobs tend to sleep less but also drink more coffee in order to be able to work harder during the day. In other words, the causation might run from more coffee to less sleep, from less sleep to more coffee, from something else to both less sleep and more coffee, or perhaps causation runs in two of these directions, or in all three directions. From the statistic alone, we have no idea. We just know from this measure that coffee consumption and sleep tend to be negatively associated.

It is often tempting for people to impute causation. To take another example, a study may find a positive correlation between amounts that elderly individuals take of a certain vitamin, and measures of cognitive ability. Readers of journalistic accounts of the study might think that they should take more of this vitamin to prevent decline in their mental abilities. But while it's possible that the vitamin does indeed have this effect, it's also possible that causation runs the other way: elderly people who are suffering cognitive impairment may be less inclined to take their vitamins. It may also be that some third factor produces an association: those who have developed habits of taking good care of themselves may be more likely to take vitamins, and also may be more likely to retain their cognitive ability into advanced age. The existence of the correlation does not imply that the vitamins are having any effect.

Chapter 4

Some philosophy of (empirical) science

One of the main concerns of the philosophy of science is to understand the methods we can use to understand the world better through empirical observation and experiment. Much of this literature is relevant to research in the social as well as natural sciences, and in fact is relevant to any attempt to learn from observations. Understanding the limits of what we can learn from data is one of the things that will help us to think clearly about the results of statistical analyses, and a brief look at some philosophy of science will be useful. This chapter will give only a very short overview of some of the ideas that seem to me to be useful and to underlie much modern thinking about scientific method and empirical knowledge. This discussion is largely based on influential work of Popper (1935, 1959).

We often want to find explanations of features of the world that we observe: for example, why can countries with apparently similar endowments of natural resources have such different levels of income? In many cases such as this, we cannot perform experiments to investigate different explanations, and we rely on empirical observation. Numerous different explanations can be suggested for an observation such as this, and we want to know what we can do to distinguish these different explanations and preferably to narrow down the set of possibilities to a smaller number. An important point here is that empirical data can in principle allow us to show that some theories are false, but cannot prove theories to be true. Essentially, and this is the idea of falsification that is associated with Popper (and which also seems to describe much scientific method as practiced for centuries) we use empirical evidence to try to eliminate some theories from contention. Those that remain are tenable, at least until further observations come along (which may later show them to be unable to explain some further piece of evidence).

Three key concepts to understand are falsification, corroboration and induction.

4.1 Falsification

In order to be falsifiable, an explanation or theory must make some statements that are in principle capable of being contradicted by an observation or experience. If this is the case we can compare the predictions of the theory to empirical experience, and can ask whether the theory appears to be compatible with what we observe. If not, then the theory in its current state is not sustainable. If so, then the theory may be said to have been corroborated, but certainly not to have been proven true; there may be many other theories that are also compatible with the observations we have made.

For example, consider a very simple form of the Fisher hypothesis stating that the nominal interest rate is equal to the inflation rate plus a constant real interest rate ($i = \pi + \bar{r}$, in a standard notation). We might look at data from a given country over time and find that the theory seems to have some value, since certainly inflation and nominal interest rates tend to rise together by similar though not identical amounts; however, because they don't rise by identical amounts, we would have to reject the basic theory. Nonetheless, we might move on to try to formulate a better theory which retains this property but allows for some variation in the real interest rate; we might for example consider a theory that says that the nominal interest rate in a given country is equal to the inflation rate in that country plus a real interest rate which is equal to the US interest rate plus a small margin. We could then test that theory on new empirical observations. This refined theory would perform better, but will still not be perfectly compatible with the data: in a small sample of data we might not have enough evidence to reject it, but in a larger sample of data we would do so, and we would then have to move on to consider further refinements. It's entirely possible that a theory may be compatible with one set of data and not with an augmented version of the data set which contains additional information and observations. In this case as we get more data we refine the theory, because we are able to discriminate more and more subtle effects as more data accumulate and more information accrues.

When predictions and observations are clear and unambiguous, it may be entirely clear that an explanation has been falsified. When random samples are involved however, things will typically not be so clear. For example, one may believe that a person's income is related only to factors which affect productivity, such as ability, education and work effort. One may therefore believe that income is independent of characteristics such as height, at least when we control for these other factors. In a random sample of data, we may nonetheless observe an association between height and income, again after controlling for these other factors, using methods that we will see later. It's possible, however, that this association is just the result of an unusual sample of data. One purpose of statistics is to try to determine the probability that this is true; that is the probability that a result has arisen because of sampling error. Nonetheless, it will never be possible with random samples of data entirely to eliminate the possibility that sampling error is responsible

for a result. What we can do, however, is to estimate the probability that this is happening. If we use an appropriate technique, and accumulate more and more data, then we will typically be able to drive this probability to a small number so that our confidence that falsification has genuinely occurred becomes quite high.

4.2 Corroboration and induction

It is often the case that theory may be tested against data many times and repeatedly make successful predictions. While theories or explanations are often said to be corroborated by this process, it's important to remember that this does not entail proving the theory true, nor does it even entail making it probable that the theory is true; we have no basis for any formal probability statement about this process.

It's always possible that another theory makes identical predictions to the theory that we tested, but differs in some other circumstances that we have not yet observed. A good example of this is the classical (Newtonian) mechanics developed by Sir Isaac Newton in the 17th century. Newtonian mechanics was exceptionally successful at explaining the movements of bodies both large and small, from planetary motions down to the movements of objects small enough to hold in our hands. Within the limits of measurement, virtually all observations about such bodies' movements could be explained through Newton's laws of motion. This corroboration through repeated successful experiments persisted for many many years; experiment after experiment after experiment was consistent with the predictions of Newtonian mechanics. It would have been a mistake nonetheless to state that the theory had been proven to be true or even that had been shown to be probably true.

It turned out of course that the reason that experimenters of the 17th, 18th and 19th centuries did not have observations that were incompatible with the theory was that the circumstances or technologies required for such incompatible observations were not available at that time. Early in the 20th century, Einstein produced his theories of special relativity and general relativity that implied that there would be small deviations from the movements predicted by Newton's laws of motion. These would be essentially undetectable given the measurement precision available before the 20th century, but in the presence of very massive objects or objects moving very close to the speed of light, there would be observable deviations from Newtonian mechanics once the technology existed to make measurements of this type. It soon became clear that there were indeed deviations from the predictions of Newtons theory which could be accounted for by the relativistic theories.

In this case it bacame clear that, as in the previous example, with more information we are able to uncover more subtle effects, and that these effects may be incompatible with an existing theory; a new theory is advanced which in this case also explains the patterns explained by previous theory, and explains further observations as well. Hundreds of years of successful

predictions by the Newtonian theory therefore did not imply that it was the truth: it was simply a theory that was a good approximation through a wide range of circumstances. That is, it was an adequate theory for explanation of all observations made until the early 1900's, but was eventually superseded by a more subtle theory that explained more. It would have been a mistake in reasoning therefore to presume that one was proving the Newtonian theory correct with repeated observations in which it made successful predictions; we simply had not yet met circumstances in which it could be distinguished from another theory.

What we learn from corroboration is that we can continue to use a theory to make deductions or predictions for the time being. We recognize that more evidence may emerge that will show that the theory has inadequacies, and we recognize that the theory may therefore be replaced someday with another, but corroboration indicates that the theory can reasonably be used for practical purposes until this time.

This is related to the traditional problem of induction. Popper (1959: p.27) writes

> It is usual to call an inference inductive if it passes from singular statements (sometimes also called particular statements), such as accounts of the results of observations or experiments, to universal statements, such as hypotheses or theories.

For example, we might repeatedly see cows that have some white markings or patches on them. When can we conclude from these observations that all cows have white patches? The answer of course is that we can't.

This is not to say, of course, that inductive reasoning is never useful. For example, over many years of driving to my cottage, Ive noticed that Im much more likely to see the police on the Eastern Townships autoroute giving out speeding tickets on warm sunny days, than on cold dark rainy nights. (It's possible of course that I simply don't see the cars as well at night, but I don't think that's it.) One might be tempted to make an inductive leap and say that traffic laws are more heavily enforced when driving conditions are safe than when driving conditions are dangerous perhaps because the police like to work office hours, and don't like to get out of their cars if it's rainy. However, we can't be sure of this. Nonetheless, it's useful to have evidence that there are often a lot of police out on the highway on sunny days when I might otherwise be tempted to drive fast and rack up demerit points, and it's reasonable to act on this information. I could in fact go further and use methods that we will learn later, together with some systematic data collection, to test the hypothesis that the probability of seeing the police giving someone a ticket is the same in the day and evening.

4.3 Asymmetry

One of the things that we always need to bear in mind is there is an asymmetry between proving things true and proving things false. This asymmetry occurs in simple forms of logical reasoning as well as in statistical reasoning.

For example, if we say that all cows have white patches on them, we can prove the statement false by observing a single cow that has no white patches; the statement is a universal one and purports to describe all instances, and therefore a single instance which is incompatible with it is sufficient to disprove it. By contrast, no number of observations of cows that do have white patches is sufficient to prove that all of them have white patches; it is always possible in principle, even if we haven't observed one, that cows without white patches exist. No number of empirical observations will allow us to prove a universal statement to be correct, but a single observation can allow us to prove the statement false.

There is also an asymmetry between proving things true and proving things false in the context of statistics. When we study hypothesis testing, this asymmetry shows up in the way we interpret rejections or non-rejections of tests. If we reject a hypothesis in a statistical test, then this tells us that (apart from some probability of a misleading result, which we can quantify) the hypothesis is false; but if we do not reject it it does not prove that the hypothesis is true. It is always possible that there are other hypotheses that would have generated the same predictions, but would be better in other circumstances. When we failed to reject the hypothesis, we learned that our observations have not provided strong evidence against it, but this is not the same thing as having shown the hypothesis to be true, because we can't be sure that other observations will not do so. As Popper (1959: p.41[1]) writes,[2]

> My proposal is based upon an asymmetry between verifiability and falsifiability; an asymmetry which results from the logical form of universal statements. For these are never derivable from singular statements, but can be contradicted by singular statements. Consequently it is possible

[1] A footnote is omitted from this quotation.

[2] In classical logic, *modus tollens* (Latin: roughly 'way of removing') refers to an argument of the following form: If X is true then Y is true; Y is false; therefore X is false. *Modus ponens* (Latin: roughly 'way of placing') refers to an argument in this form: If X is true then Y is true; X is true; therefore Y is true. These are two valid forms of deductive argument. There are two similar-looking invalid forms of reasoning, or fallacies: If X is true then Y is true; Y is true; therefore X is true. For example, if cows can fly, then airplanes can fly; airplanes can fly, therefore cows can fly. Whoops! This is called the fallacy of affirming the consequent. Finally we have the fallacy of denying the antecedent: if X is true then Y is true; X is false; therefore Y is false. For example, if cows can fly then there are living beings in the air; cows can't fly; therefore there are no living beings in the air.

by means of purely deductive inferences (with the help of the *modus tollens* of classical logic) to argue from the truth of singular statements to the falsity of universal statements.

4.4 Summary

1. Genuinely scientific theories must be falsifiable. A statement which is not falsifiable is some other kind of statement rather than a scientific one. (By 'scientific' we refer of course to using the scientific method, rather than to natural science in particular.)

2. Theories or explanations should be tested, that is, attempts should be made to falsify them in order to narrow down the set of possible explanations that may potentially be valid.

3. Theories that have not been falsified can be retained as tentatively adequate explanations, but can't be taken to have been proven true, however much corroboration they may receive. It's always possible that new observations will come along to falsify a theory.

4. We cannot be literally certain that falsification has occurred when we work with statistical observations that have uncertainties associated with them, but as we obtain more data we can narrow our uncertainty bands and be more confident in a falsification. We can then better distinguish similar theories that may differ in relatively subtle ways.

Part II: Some theoretical foundations

Chapter 5

Probability Theory

Statistical reasoning depends on a theory of probability. Understanding the meaning and interpretation of test statistics, estimation methods, and even of the simple descriptive statistics that we reviewed in Chapter 3, requires an understanding of probability. Probability statements are also made in common, non-technical conversation, and a better understanding of probability will also help us in making more precise statements and in evaluating statements that we hear. For example, we often say things such as 'She'll probably be here before noon.' By 'probably' we presumably mean that the probability of her arrival by noon is greater than 50%; a more precise version of this statement might be 'I'm 90% sure that she'll be here before noon.' But what do these statements mean? In the end she'll either be here by noon or she won't, so how (if at all) could we ever check on the accuracy of the statements?

People have considered probability *a priori*, or through purely theoretical reasoning, and empirically or *a posteriori*, through actual observations of random events. These two ways of learning about probability produce compatible results, if we interpret them sensibly. We will begin by discussing simple, traditional ways of defining probability along each of these lines, and we will see that although we can understand a good deal using these simple definitions, they will not allow us to solve all of the problems that we want to solve. Next we will set out a formal system that allows us to write clear and precise rules for describing and calculating probabilities. This system will operate using sets of events, so we will need to study some set theory in order to work with the system. We will then be able to understand enough about probability, unconditional and conditional, to follow the statistical arguments in the rest of this book.

5.1 'Classical' probability

As we have noted, the earliest formal study of probability was motivated by gambling problems. Many such problems have the feature that the outcomes (rolls of dice, card selections, *etc.*) can be divided into equally likely outcomes – that is, we may not know the probability of something, but we may feel confident that two or more things have the same probability, whatever that is. Thinking about probability in this way will not allow us to solve

all problems that interest us, but will take us quite a way , and will help to introduce a more general approach.

We first define a random experiment.

Definition 5.1.1 A **random experiment** is an experiment for which the outcome is not known with certainty.

Note that, as in our discussion in Chapter 2, randomness requires only that there be some uncertainty about the outcome.

Definition 5.1.2: Classical probability Suppose that a random experiment can result in any one of n outcomes, which are mutually exclusive[1] and equally probable. If n_A of these outcomes have a characteristic A, the classical probability of an outcome with characteristic A is n_A/n .

Note that the probability of A arising is defined with reference to other probabilities, of each of the outcomes, which we assumed to be equal. We can compute some probabilities from others, but we have to rely on some assumed *a priori* knowledge to do so.

The traditional examples illustrating this definition use dice, cards and coins, where it is sensible to assume that certain outcomes have identical or virtually identical probabilities. If a head and a tail are equally likely when we flip a coin, then the probability of either outcome, by **D5.1.2**, is 1/2. If we roll a six-sided die and each of the integers 1, 2, . . . , 6 is equally likely to arise, then the probability of any one of these outcomes is 1/6 ; the probability of an even number arising is 3/6 or 1/2; of a number greater than 4 arising is $2/6 = 1/3$, and so on. In picking a card from a standard deck at random, the probability of a club is $13/52 = 1/4$; the probability of picking out a Queen is $4/52 = 1/13$.[2]

In calculating probabilities in this way, we must be careful not to assume that a division of possible outcomes into k categories means that each category is equally likely to arise. For example, if we roll two dice, the possible outcomes for the sum of the two values are the integers 2, 3, . . . , 12. However, these values are not all equally likely, because some of them can arise in more ways than others, and so have a higher probability of occurring. There are six possible values for the first die, and six for the second, or 36 possible permutations. Of these, one (1 and 1) yields a sum of 2, so the probability of a 2 is 1/36. The most likely number to arise is 7, which can be produced by (1,6), (2,5), (3,4), (4,3), (5,2) or (6,1), or six of the 36 equally likely possibilities, so its probability is 1/6. Similarly, imagine that a couple decides to have two children. What is the probability that they will both be girls? The possible

[1] Mutually exclusive: any one outcome excludes the possibility of any other outcome.

[2] For those readers brought up in strict Methodist households, a standard deck has 52 cards divided into four suits (clubs, diamonds, hearts, spades), each of which contains cards numbered 1 (ace) through 10, plus Jack, Queen, King.

outcomes might be categorized as two girls, two boys, and one of each. This does not imply that the probability of each of these outcomes is identical (if they were, then the probability of two girls would be 1/3). If it is equally probable that each child be a boy or a girl, then a categorization into equally probable outcomes is GG, GB, BG, BB. Having two girls is one of four equally probable outcomes, so the probability is 1/4 .

Now what if the probabilities of a boy or a girl being born are not exactly 1/2 but, for example, 0.505 and 0.495? We no longer have the equally probable events necessary to compute the probability that we want using **D5.1.2** (although in this case using 0.5 would give a good approximation for most purposes, we cannot get the exact answer). Even this simple problem is beyond the scope of **D5.1.2**, and this is true of other kinds of problems that we will want to solve, involving for example a set of outcomes which is of unknown size, or potentially infinite. We need to have more elaborate methods for deductive reasoning about probability.

5.2 A posteriori probability

Another traditional approach to probabilistic reasoning begins from the other end of the problem. Rather than reasoning purely deductively, we can observe outcomes, and attempt to infer probabilities from these outcomes.

Definition 5.2.1: A posteriori probability. Suppose that a random experiment is repeated n times, and produces an observable outcome each time. If n_A of these outcomes have a characteristic A, the **A posteriori probability** of an outcome with characteristic A is n/n_A.

Notice that while the notation used (A, n, n_A) is the same as in **D5.1.2**, we now speak of an experiment that is actually carried out, from which we will infer probability. We do not need to assume that any two events are equally likely, but we do need a set of observations to work with.

Reasoning in this way will allow us to handle problems that we could not handle with **D5.1.2**, because equal probabilities are not required; however, different people carrying out the experiments will usually get different answers (if they do a very large number of experiments, however, the differences will typically be small). For example, I might flip a coin 100 times and get 53 tails; you might get 48 tails (in fact, the whole range of answers that will typically arise when many people do this experiment can be described precisely, as we will see in later chapters.)

Now consider again the example of the couple with two children, and the probability that a baby will be a girl or boy.[3]The proportion of newborns who

[3] Note again that we can speak of probability to describe our (imperfect) knowledge of a situation even if someone with greater knowledge could describe the situation with certainty. Before ultrasound or other tests, one might have said just before a birth that the probability of having a girl was about 12 , although

are boys has frequently been measured and, while the actual proportions of course differ slightly from place to place, it is typically found to be around 0.512. We would infer from this that there is some probability that a newborn will be male (we might qualify this by saying that the probability is specific to a particular region or time), and that we have measured this probability at 0.512, recognizing that this is a measurement subject to some degree of error, rather than an exact answer. We have no means of putting approximate bounds on this error or otherwise giving any indication of its importance, although this is a key element in statistical reasoning which we will want to develop methods to address. In this case (that is, reasoning *a posteriori*), we could also directly measure the proportion of couples who had two girls among those who had exactly two children.

Health economists as well as medics are often interested in survival probabilities after life-threatening events. Consider two possible procedures for treating a patient brought to hospital after having had a heart attack. What is the probability of survival if procedure A is followed, versus the probability if procedure B is followed? How do these compare with the probability if no treatment is given? We can imagine estimating these probabilities by assigning incoming patients randomly to procedure A, procedure B, or to no treatment, and computing the survival proportions to estimate probabilities. If it is clear that the procedures are better than nothing, however, assigning some to have no treatment would be unacceptable; moreover, if admissions staff tend to assign certain types of patients to treatment A and some to treatment B, then our sampling is biased and the proportions will not reflect the desired probabilities: for example, admissions staff may believe that procedure A is more effective in the worst cases, and so may assign the worst cases to procedure A: procedure A then has a group of cases which are typically less likely to lead to survival, and so may show a lower survival proportion for that reason, rather than because it is less effective.

These examples illustrate the strengths and limitations of the *a posteriori* approach. As long as we have a large number of observations made under comparable conditions, in which a certain feature can be observed to occur or not, we can estimate a probability. This is very flexible and does not require that we assume things that might not be true, such as equality of birth probabilities, in order to get an answer. On the other hand, it gives us no means of using reasoning to supplement our knowledge beyond what can be directly measured in repeated experiments, and gives us no means of answering questions such as 'what is the probability that we will be in a recession in three months?' or even 'what is the probability that I will die of a heart attack?'.

In order to be able to solve a wide variety of problems, we will need to set out a precise and more general description of what we mean by 'probability'.

the child's sex was already determined with certainty and would have been observable to someone with the right technology.

We will do so by starting with some definitions, using these to define probability via several axioms, and then using further mathematical methods with the definitions and axioms to compute probabilities. In order to make the necessary definitions, we need to refer to some concepts from set theory, so we will begin with some fundamental set-theoretic concepts. We will complete our review of set theory in a later section.

5.3 Set theory: basic concepts

This section contains only a few very elementary definitions and examples.

Definition 5.3.1: Set A **set** is a collection of items.

The items in a set may be tangible or intangible. For example, we may have a set of ideas: $A = \{$ capitalism, socialism, communism, Maoism$\}$ is a set of political philosophies which were influential in the 20th century. We usually write the elements of a set within parentheses, separated by commas.

Definition 5.3.2: Element An **element** of a set is one of the items in that set.

Definition 5.3.3: Space A **space** is the collection of all possible elements from which sets may be defined in a particular context.

The set A just given may be thought of as containing elements from the space of all possible political philosophies.

Definition 5.3.4: Subset A **subset** is a part of another set, such that if B is a subset of A, every element of B is also an element of A.

The symbols '$\subset$' and '$\supset$' are typically used to describe subsets; $B \subset A$ means that B is contained in A, or B is a subset of A; similarly $B \subset A$ means that B contains A, so that A is a subset of B.

Definition 5.3.5: Complement The **complement** of a set A,denoted here by $\bar{A}$, consists of all elements of the space which which are not in the set.

Definition 5.3.6: Union The **union** of two sets A and B, denoted $A \cup B$, is the set of all elements which belong to at least one of the sets. Similarly, the union of ℓ sets is the set of all elements belonging to at least one of the ℓ sets.

Definition 5.3.7: Intersection The **intersection** of two sets A and B, denoted $A \cap B$ is the set of all elements which belong to both of the sets, and the intersection of ℓ sets is the set of all elements which belong to every one of the ℓ sets.

Our aim below will be to use set theory to help develop some statements about probability that can be used for computing probabilities. We will therefore need to take advantage of any known relations among these set-theoretic concepts, in the next section, to develop rules applicable to probabilities. Some of these rules are summarized in the following theorem.

Theorem 5.3.1 Set operations: For any sets A, B, C, *etc* .,

i (Commutativity) $A \cap B = B \cap A$; $A \cup B = B \cup A$.

ii (Associativity) $A \cap B \cap C = (A \cap B) \cap C = A \cap (B \cap C)$ and $A \cup B \cup C = (A \cup B) \cup C = A \cup (B \cup C)$.

iii (Distributivity) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

iv (De Morgan's laws) $\overline{(A \cup B \cup C)} = \overline{A} \cap \overline{B} \cap \overline{C}$ and $\overline{(A \cap B \cap C)} = \overline{A} \cup \overline{B} \cup \overline{C}$.

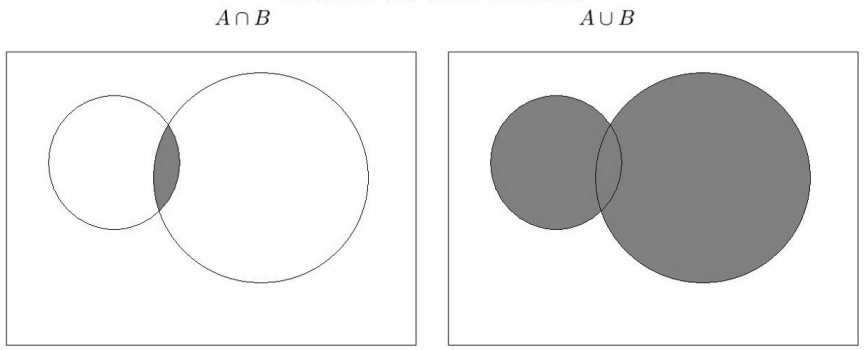
While we have used only the sets A, B, C in **T5.3.1**, we could extend the theorem to cover any number of sets. For example, the first part of **T5.3.1** (iv) can be written for ℓ sets, $A_1, A_2, \ldots, A_\ell$ as

$$\overline{(A_1 \cup A_2 \cup \ldots \cup A_\ell)} = \overline{A_1} \cap \overline{A_2} \cap \ldots \cap \overline{A_\ell}.$$

These theorems and others are often illustrated and visualized using *Venn diagrams*.[4] Figures 5.3.1A/B illustrate the intersection and union of two sets A and B, Figures 5.3.2A/B the intersection and union of three sets A, B,C, and Figures 5.3.3A/B the complements of these quantities, addressed in De Morgan's laws for the case of three sets.

Figure 5.3.1 A/B

Intersection and union of two sets



[4] Named for John Venn, 1834-1923.

FIGURE 5.3.2 A/B

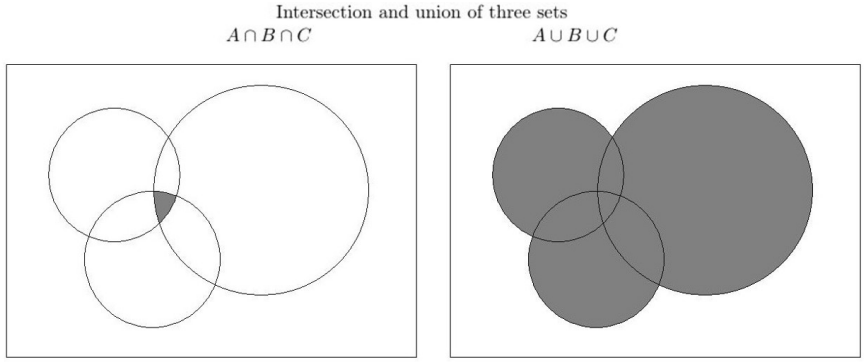
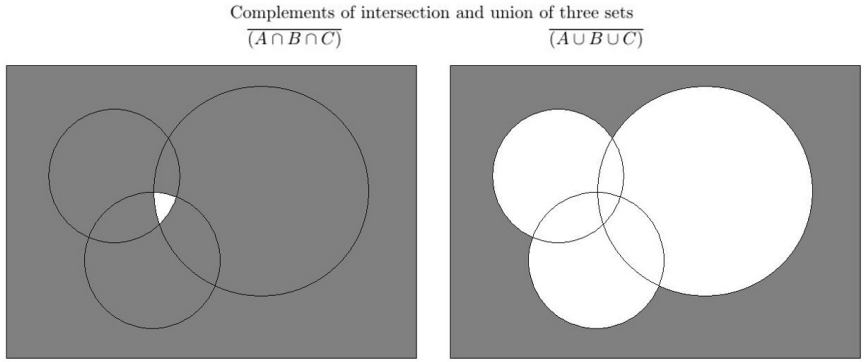


FIGURE 5.3.3 A/B



5.4 Axiomatic probability

We can now use the concepts and definitions established above, together with some essential principles or axioms which we will presume must hold for probability, in order to derive some rules that probabilities must follow. As we build up a set of such rules, or theorems of the system, we are building a catalogue of results that we can use in assigning probabilities to certain types of events, or to solving problems that involve determining probabilities.[5] First we will need to make further definitions of concepts directly related to probability, and will need to write down axioms which correspond with our concept of probability.

[5] A formal axiomatic system begins with a few unproven statements or axioms which are presumed to be true, and derives further statements or theorems from them by deductive logic alone. If therefore we are confident in the truth of the axioms, we can be equally confident in the truth of the derived theorems.

Using the set-theoretic definitions above, we can define events, sample spaces, probability mass functions and probability spaces. With these definitions, we will have assembled the key elements necessary in order to begin to determine probabilities in practical problems.

Definition 5.4.1: Sample space A **sample space**, or **outcome space**, for which we will use the symbol Ω, is the collection of all possible outcomes of a random experiment.

Definition 5.4.2: Event An **event** is a subset of the outcome space. Since an event is a subset, we will usually use a notation for an event similar to that used for a set above, typically an upper case Latin letter.

Definition 5.4.3: Mutually exclusive events Two events A and B, subsets of Ω, are **mutually exclusive** if the intersection of these subsets is the empty set, *i.e.* $A \cap B = \emptyset$.

Definition 5.4.3: Exhaustive set of events A set of events $\{A_1, A_2, \ldots, A_k\}$ is an **exhaustive set** if $\bigcup_{i=1}^{k} A_i = \Omega$, that is, the events $A_1, A_2, \ldots, A_k$ are the only possible events.

Two mutually exclusive events cannot both happen; each excludes the other. Note that we have introduced the commonly-used symbol $\emptyset$ for the empty set, that is, the set which contains no elements. Again, we can also extend this definition to an arbitrary number ℓ of events; the events $A_1, A_2, \ldots, A_\ell$ are mutually exclusive if no two of the sets have any elements in common: that is, $A_i \cap A_j = \emptyset \quad \forall\, i \neq j$ (Recall that the symbols $\forall\, i \neq j$ are read as 'for all i not equal to j,' indicating that the statement holds for any two subsets A_i and A_j as long as they are not the same; by contrast, if $i = j$, then we have $A_i \cap A_i$, the intersection of a set with itself, and of course $A_i \cap A_i = A_i$.)

We now define a **probability function** or **probability measure**, from which we will later derive rules for manipulating probabilities that will allow us to solve specific problems.

Definition 5.4.5: A **probability function** or **probability measure** $P(\cdot)$ is a function defined on a $\boldsymbol{\sigma}$**-field** $\mathcal{F}$, of events, (sigma-field) with range the closed interval [0, 1], which has the following properties:

i $P(A) \geq 0 \; \forall\, A \in \mathcal{F}$

ii $P(\Omega) = 1$

iii Let $A_1, A_2, \ldots, A_\ell$ be a set of mutually exclusive events in $\mathcal{F}$ and define $B = A_1 \cup A_2 \cup \ldots \cup A_\ell$. Then $P(B) = \sum_{i=1}^{\ell} P(A_i)$.

We have used a finite set of events with ℓ elements, but the definition continues to hold if we replace ℓ with ∞ and define $B = A_1 \cup A_2 \cup \ldots$ Part iii

of the definition may instead by derived as a consequence of this more general statement.

It follows from the fact that $\mathcal{F}$ is a σ-field that $A_1 \cup A_2 \cup \ldots \cup A_\ell \in \mathcal{F}$, that is, the union of all of these events is also an element of $\mathcal{F}$. We will need to make a formal definition of a field, also called an **algebra**, of events in order to define a probability space.

Definition 5.4.6: A **σ-field** or **σ-algebra** of events $\mathcal{F}$ is a collection of events with the following properties:

i $\Omega \in \mathcal{F}$

ii if $A_1 \in \mathcal{F}$ and $A_2 \in \mathcal{F}$ then $A_1 \cup A_2 \in \mathcal{F}$

iii if $A \in \mathcal{F}$ then $\overline{A} \in \mathcal{F}$.

It follows from property (ii) that the union of any finite set of events must also be in $\mathcal{F}$. We may want to allow a potentially infinite set of events in which case we have properties (i) and (iii) above and also

$$\text{if } A_1, A_2, \ldots \in \mathcal{F} \text{ then } \mathcal{B} = A_1 \cup A_2 \cup \ldots \in \mathcal{F}.$$

In the rest of this chapter, we will assume that any events discussed come from a probability space, defined on a *sigma*-field as necessary. While we will not state this condition explicitly each time, it is important in that it allows us to conclude that all of the sets constructed from other sets (complements, unions, intersections, etc) are also contained in the sample space and the σ-field.

From **D5.4.5** we can derive a number of theorems which follow from the parts of the definition of a probability function, and can be used directly in solving problems involving probabilities.

Theorem 5.4.1 Let A and B be two events in $\mathcal{F}$.

i Let the sample space be divided into n equally-likely outcomes. Let n_A of these outcomes imply the event A. Then $P(A) = n_A/n$.

ii $P(B) = P(A \cap B) + P(\overline{A} \cap B)$.

iii $P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(\overline{A} \cap B)$.

iv Let A and B be mutually exclusive events. Then $P(A \cup B) = P(A) + P(B)$.

v Let $\{A, B\}$ be an exhaustive set of events. Then $P(A) + P(B) \geq 1$. If $A \cap B = \emptyset$, $P(A) + P(B) = 1$.

vi $P(\overline{A}) = 1 - P(A)$.

vii $P(\emptyset) = 0$.

In cases in which we can identify a finite number of equally-likely outcomes, it is possible to use **T5.4.1** (i) directly to obtain probability statements; the remaining parts of the theorem do not require partition into equally-likely outcomes.

Sometimes when we try to make such a partition, we find that there is a large number of equally-likely outcomes, and counting them can be extremely cumbersome. It may be simpler to do the counting using the definitions of the factorial operator, and of permutations and combinations. These devices will allow us to exploit the definition and theorem in cases that are too bulky to allow counting cases unsystematically.

Definition 5.4.7 The **factorial operator** applied to an integer m, written as $m!$, is defined as $m \cdot (m-1) \cdot \ldots \cdot 1 = \prod_{j=1}^{m-1}(m-j) = \prod_{i=1}^{m} i$, and $0! \equiv 1$.

Definition 5.4.8 The number of possible **combinations** of k items chosen from a set of n items is the number of distinct sets of k that can be assembled when order is not considered (different orderings are considered to be equivalent), and is written as ${}^nC_k$ or $\binom{n}{k}$.

That is, in choosing combinations of three objects from a set of five, the sets $\{A, B, C\}$ and $\{C, B, A\}$ are considered the same combination. They would be different *permutations*, however.

The number of combinations can be computed as

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}.$$

Definition 5.4.9 The number of possible **permutations** of k items chosen from a set of n items is the number of distinct sets of k that can be assembled when different orderings are considered to be different, and is written as ${}_nP_k$; it can be computed as ${}^nP_k = n!/(n-k)!$

We will use these definitions below when we meet problems involving large numbers of equally-likely outcomes, which it would be difficult to count without these expressions.

5.5 Conditional probability

In what we have seen so far, we have considered probabilities of events without reference to the occurrence of other events (unless, of course, we are speaking about a union or intersection of events, in which case we can think of this union or intersection as defining a new event). The probabilities, treated above, are called *unconditional* probabilities. It is also interesting to be able to describe the way in which probability statements can be made more precise if relevant information from another event is observed. For example, the probability that a randomly-selected individual will have a heart attack before age 65 may be a. The probability that a randomly-selected individual who eats mainly fried foods and saturated or trans fats will have a heart

attack before age 65 may be $b > a$. The probability that a randomly-selected individual with this diet, who also smokes heavily, will have a heart attack before age 65 may be $c > b$. The information about diet and smoking, on the assumption that the current state of medical knowledge is correct on these points, is relevant to determining heart attack probabilities; this is useful conditioning information; that is, when we compute probabilities with this additional information, we may be able to get more precise answers than if we had simply treated the individual as part of a general population.

The definition of conditional probability is given in **D5.5.1** and is easily illustrated with a Venn diagram, as in Figure 5.5.1, if we think of areas as representing probabilities.

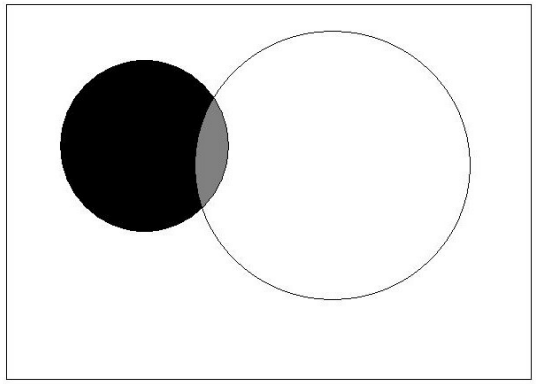
Definition 5.5.1 The conditional probability of A given that B holds, written $P(A|B)$, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

In Figure 5.5.1, the conditional probability of B given A is equal to the ratio of the light-shaded area ($P(A \cap B) = P(B \cap A)$ to the dark shaded + light shaded areas, $P(A)$; *i.e.* $P(B|A) = P(B \cap A)/P(A)$.

FIGURE 5.5.1

Conditional probability in sets A and B



D5.5.1 immediately implies that we can re-write the probability of the intersection as $P(A \cap B) = P(A|B)P(B)$.

An important result called Bayes' Theorem or Bayes' rule[6] follows from this definition and from the commutativity property, Theorem 5.3.1 i. We have $P(A \cap B) = P(A|B)P(B)$ and by symmetry (simply re-labeling), $P(B \cap A) = P(B|A)P(A)$. Since $P(B \cap A) = P(A \cap B)$ by the commutativity property, it follows that $P(A|B)P(B) = P(B|A)P(A)$, and dividing by $P(B)$ we obtain:

[6] First obtained by the Reverend Thomas Bayes, 1702-1761.

Theorem 5.5.1 Bayes' Theorem. Let A and B be two events in $\mathcal{F}$. Then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

A number of other important results can also be stated using conditional probabilities.

Theorem 5.5.2 Let A, B and C be events in $\mathcal{F}$.

i Let $P(B) > 0$. Then $P(\overline{A}|B) = 1 - P(A|B)$.

ii Let $P(B) > 0$. Then $P(A|B) = P(A \cap C|B) + P(A \cap \overline{C}|B)$

iii Let $0 < P(B) < 1$. Then $P(A) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B})$.

iv Let $\{A, B\}$ be an exhaustive set of events, let A and B be mutually exclusive and $P(A) > 0, P(B) > 0$. Then $P(C) = P(C|A)P(A) + P(C|B)P(B)$.

Bayes' Theorem, in particular, allows us simple solutions to some potentially tricky problems. For example, consider the well-known problem of interpreting a medical test result. A test is given which has a false positive rate of 2%; that is, in people who do not have the condition, the test falsely indicates that they do 2% of the time (and 98% of the time correctly indicates its absence of the condition). If an individual does have the condition, the test identifies this 99% of the time. The condition is present in 1 person out of 1000. If someone receives a positive test result, assuming that people are randomly selected for testing, what is the probability of having the condition? Let A indicate that an individual has the condition, and let B indicate that an individual receives a positive test result. Then $P(B|A) = 0.99, P(A) = 0.001$, and using **T5.5.2** (iii), $P(B) = 0.001(.99) + .999(.02) = 0.02097$. Therefore $P(A|B) = (0.99 \cdot 0.001)/0.02097$, or approximately 4.7%. That is, given a positive test, the probability that one nonetheless does not have the condition is 95.3%, and this in a test that might be described loosely as being 98% or 99% accurate. (If this result appears strange, note that in 1000 individuals randomly selected for testing, roughly one might be expected to have the condition, whereas around 20 would be expected to get false positives.)

Conditional probability statements are often much more precise and useful than unconditional statements, because they embody more information. For example, the unconditional probability that a car will be stolen in Montreal in any given year may be 1/200: that is, of all cars legally registered to owners living in Montreal, one in two hundred of them is stolen in any given year. If we know that someone lives in Montreal and has a car, but know nothing else, then the best we can do to estimate his or her probability of

having the car stolen is to use the unconditional probability, 1/200. However, if we have information on relevant *conditioning variables*, (that is, variables which help us to predict car theft), then we may be able to make a better estimate of the probability applicable to this individual. If car is a new BMW, the owner has had two previous cars stolen within the last five years, and he or she parks on the street rather than in a garage, then the probability of having this car stolen in the next year will be much higher than 1/200.

Of course, not all events or types of information are useful as conditioning information. For example, learning that an individual grew up on a farm may convey no information about whether he or she will have a heart attack before age 65.[7] In this example, if A is having a heart attack before age 60 and B is having grown up on a farm, then $P(A|B) = P(A)$: event B is irrelevant.

Independence in this statistical sense, often called statistical independence, can be defined by different equivalent conditions.

Definition 5.5.2a Two events A and B are **independent** if and only if $P(A|B) = P(A)$ (for $P(B) > 0$); equivalently $P(B|A) = P(B)$ (for $P(A) > 0$).

Definition 5.5.2b Two events A and B are independent if and only if $P(A \cap B) = P(A)P(B)$.

The statement of **D5.5.2a** implies **D5.5.2b** and *vice versa* .

Extensions of the many of the results that we have stated above are available for an arbitrary number of events, rather than just the two events A and B. The next theorem collects some of these more general forms of result, of which results above are special cases.

Theorem 5.5.3 Let $A_1, A_2, \ldots A_\ell$ be events in $\mathcal{F}$. Then

i (Bayes' Theorem) $P(A_i|B) = \big(P(B|A_i)P(A_i)\big)/P(B)$.

ii If the events $A_1, A_2, \ldots A_\ell$ are independent of each other, then $P(A_1 \cap A_2 \cap \ldots \cap A_\ell) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_\ell)$.

iii If the events $A_1, A_2, \ldots A_\ell$ are mutually exclusive, then $P(A_1 \cup A_2 \cup \ldots \cup A_\ell) = \sum_{i=1}^{\ell} P(A_i)$.

iv If the events are mutually exclusive and exhaustive, then

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots + P(B|A_\ell)(P(A_\ell).$$

[7] Of course, it is possible that growing up on a farm may confer a lifelong serenity which lowers pre-65 heart attack risk. To check on this would require investigation in a multivariate model to control for multiple effects measurable in a cross-sectional sample of individuals; Chapter 18.

Examples of proofs of some simple theorems of the system

Theorem A1. Let i index a set of mutually exclusive and exhaustive events A_i. Then $\sum_{i\in\Omega} P(A_i) = 1$.

Proof: We could re-write (iii) in Definition 5.4.5 as $P(B) = \sum_{i\in B} P(A_i)$, that is, B is the union of all events A_i and its probability is the sum of the probabilities of all events such that A_i in B. Now let $B = \Omega$. Then $P(\Omega) = \sum_{i\in\Omega} P(A_i)$. But $P(\Omega) = 1$ and so $\sum_{i\in\Omega} P(A_i) = 1$. ■

Theorem A2. Let the sample space Ω be the union of of n mutually exclusive and equally likely events C_i. Then $P(C_i) = 1/n \; \forall i$, where i indexes the events.

Proof: We have $\sum_{i\in\Omega} P(A_i) = 1$ from Theorem A1 and so $\sum_{i\in\Omega} P(C_i) = 1$ since the events C_i are also mutually exclusive. Since they are equally likely, $P(C_i) = c$, a constant, for all i. Then $\sum_{i\in\Omega} P(C_i) = \sum_{i\in\Omega}(c) = nc$, since we add up the same item, c, n times. Then $nc = 1$, so $c = 1/n = P(C_i)$. ■

Theorem 5.4.1 (i) Let the sample space Ω consist of n mutually exclusive and equally likely outcomes C_i. Let n_A of these outcomes imply the event A. Then $P(A) = n_A/n$.

Proof: From Theorem A2 we have that $P(C_i) = 1/n \; \forall i$. Let $j = 1, 2, \ldots n_A$ index outcomes that imply event A, which we can label $A_1, A_2, \ldots A_{n_A}$. Since any one is sufficient to imply A, the probability $P(A)$ is the probability of the union of these events $A_1, A_2, \ldots A_{n_A}$. Now from D5.4.5 (iii) we know that if $A_1, A_2, \ldots A_\ell$ is a set of mutually exclusive events, and define $B = A_1 \cup A_2 \cup \ldots \cup A_\ell$, then $P(B) = \sum_{i=1}^{\ell} P(A_i)$. In the notation of this theorem, we have $A = A_1 \cup A_2 \cup \ldots \cup A_{n_A}$, and so $P(A) = \sum_{i=j}^{n_A} P(C_i) = \sum_{i=j}^{n_A}(1/n) = n_A(1/n) = n_A/n$. ■

Theorem 5.4.1 (ii first part)

Proof: Begin with a few things we already know: $P(\overline{A}) - 1 - P(A)$, and

$$P(B) = P(A \cap B) + P(\overline{A} \cap B). \qquad (A5.1)$$

since A and $\overline{A}$ are mutually exclusive and exhaustive (they cover the entire sample space). Therefore also A and $(\overline{A} \cap B)$ must be mutually exclusive, so that by axiom (iii), Definition 5.4.5, the sum of their probabilities is the probability of their union:

$$P(A \cup B) = P(A) + P(\overline{A} \cap B). \qquad (A5.2)$$

Now we can just re-arrange (A5.1) to write:

$$P(\overline{A} \cap B) = P(B) - P(A \cap B). \qquad (A5.1')$$

Substituting (A5.1') into (A5.2) to eliminate the term $P(\overline{A} \cap B)$, we obtain $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. This corresponds with the intuitive idea that to compute the probability of a union, we need to avoid double-counting the intersection, which is what we would do if we just added up $P(A)$ and $P(B)$. ■

Theorem 5.5.2 (iv): Theorem of Total Probabilities, two-set version.

Let $\{A, B\}$ be an exhaustive set of events, let A and B be mutually exclusive and $P(A) > 0, P(B) > 0$. Then $P(C) = P(C|A)P(A) + P(C|B)P(B)$.

Proof: $P(C|A)P(A) + P(C|B)P(B) = [P(C \cap A)/P(A)]P(A) + [P(C \cap B)/P(B)]P(B)$, substituting from the definition of conditional probability, and so $P(C|A)P(A) + P(C|B)P(B) = P(C \cap A) + P(C \cap B)$. Now A and B are mutually exclusive, and therefore $P(C \cap A) + P(C \cap B) = P[(C \cap A) \cup (C \cap B)] = P[C \cap (A \cup B)]$, with the first of these equalities following from Theorem 5.4.1 part (iv) and the latter following from the distributive property, Theorem 5.3.1 part (iii). Finally A and B are also exhaustive, i.e. $A \cup B = \Omega$, the sample space. Therefore $P[C \cap (A \cup B)] = P(C \cap \Omega) = P(C)$. ■

One of De Morgan's laws (other direction is similar)

Theorem: Let A and$B$ be two sets in S. Then $\overline{(A \cup B)} = \overline{A} \cap \overline{B}$.

Proof: Two sets C and D are equal if each is contained in the other, i.e. $C \subset D$ and $D \subset C$. So first we show that $\overline{(A \cup B)} \subset (\overline{A} \cap \overline{B})$. Consider an individual element γ of the first set, *i.e.* $\gamma \in \overline{(A \cup B)}$. hen $\gamma \in \overline{(A \cup B)} \Rightarrow \gamma \notin (A \cup B)$. Therefore $\gamma \notin A$ and $\gamma \notin B$; therefore $\gamma \in \overline{A}$ and $\gamma \in \overline{B}$. So $\gamma \in \overline{A} \cap \overline{B}$ and any element of the first set $\overline{(A \cup B)}$ must be contained in the second set, $overline A \cap \overline{B}$.

Now we show the converse, $(\overline{A} \cap \overline{B}) \subset \overline{(A \cup B)}$. Consider an individual element ν of the set $(\overline{A} \cap \overline{B})$ Then $\nu \notin A$ and $\nu \notin B$. Therefore $\nu \notin (A \cup B)$ (to be in the union of the two sets A and B, ν would have to be in one or the other). Finally, $\nu \notin (A \cup B) \Rightarrow \nu \in \overline{(A \cup B)}$. So any element of the set $\overline{A} \cap \overline{B}$ must be contained in the set $\overline{(A \cup B)}$. Since each set is contained in the other, they are equivalent. ■

Chapter 6

Random variables and distribution theory

This chapter uses the theory of probability described in Chapter 5 to introduce fundamental statistical concepts underlying virtually all statistical analysis. It provides a set of definitions and quantities which will form a basis for statistical reasoning and discussion quite apart from any formal analysis, as well as theoretical counterparts to the descriptive statistics introduced in Chapter 3. In particular, the concepts of cumulative distribution function, probability mass function, and probability density function, expectation and moments (introduced in the next chapter) will reappear constantly in this book and in other statistical material that the reader will see.

6.1 Random variables

The random variable is the fundamental unit of statistical analysis. Since it is not deterministic – its values cannot be predicted with certainty – we require statistical concepts to describe and predict the outcomes of the random variable and relate it to other random variables.

D6.1.1 A *random variable* is a real-valued quantity which depends on a random event.

Because it depends on a random event, the random variable cannot be perfectly predictable. Note also that this definition makes the random variable a real number, which rules out some things that we might speak of informally as random variables. For example, the next car to pass on the street might be made in North America or elsewhere, and we might be interested in estimating the proportion made in North America. For each car that passes, we could record the variable $y_i =$ 'NA' or $y_i =$ 'not NA', where i indexes the observations, that is, $i = 1, 2, 3, \ldots$ for a sequence of observations. The variable Y which can take either of these values is not a random variable by the definition above, because these place-of-manufacture labels are not real numbers. However, if we assign $z_i = 1$ for North American cars and $z_i = 0$ for others, then the variable Z is a random variable by definition 6.1.1.

A more formal definition than **D6.1.1** would refer explicitly to the probability theory that we developed in Chapter 5. The following definition makes **D6.1.1** somewhat more precise, and uses the definition of a probability space. Recall that ω is an outcome from this space Ω.

D6.1.2 A *random variable* X is a real-valued function on a probability space Ω. (It is often specified also that this random variable is such that $A_\ell = \{\omega : X(\omega) \leq \ell\}$ belongs to $\mathcal{A}\ \forall\, \ell \in \mathcal{R}$.

We are now able to define the fundamental concept of a *cumulative distribution function*, or 'cdf'. Every well-defined random variable possesses a cumulative distribution function, but does not necessarily possess a probability density function, another core concept which we will define soon. As with a random variable, we may define a cdf in different ways and with different degrees of formality. Here is one definition.

Definition 6.1.3 The cumulative distribution function of a random variable X is a function $F_X(x)$ defined on the real line, with range [0, 1], such that $F_X(x) = P(X \leq x)$ for every real value x.

(Notice that we use an upper-case subscript on the function, indicating the name of the variable ('X') and a lower-case argument ('x') indicating a particular value of the random variable.) We could have written last part of **D6.1.3** somewhat more formally as $F_X(x) = P(\{\omega : X(\omega) \leq x\})\ \forall\, x \in \mathcal{R}$ making reference to the elements ω of the sample space Ω on which the random variable is defined.[1]

The cdf is sometimes referred to as a 'distribution function', omitting 'cumulative.' Mathematicians sometime say simply 'distribution'. It is however a cumulative function: for any real value x, it gives the total probability corresponding with all outcomes of the random variable up to and including x. Because it is a probability, it must be a real number in [0, 1], as the definition requires.

A number of properties of the cdf follow from the facts that it is a cumulative function and that its values are probabilities. In particular:

- $0 \leq F_X(x) \leq 1; \quad \lim_{x\to-\infty} F_X(x) = 0; \quad \lim_{x\to+\infty} F_X(x) = 1;$
- $F_X(x_1) \leq F_X(x_2) \quad \forall\, x_1 < x_2;$
- $\lim_{h\downarrow 0} F_X(x+h) = F_X(x).$

Note: The notation $\lim_{h\downarrow 0} F_X(x+h)$ means that the limit is taken from above. That is, $h \to 0$ only for positive values of h. An alternative notation is $\lim_{h\to 0+}$.

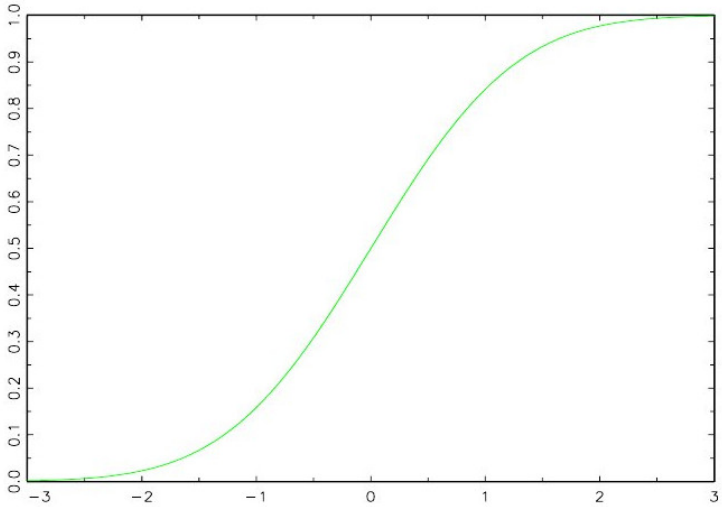
The first statement corresponds with probabilities bounded into the interval [0, 1]; no matter how small x is the probability of being at or below it cannot be less than zero, and no matter how large it is, the probability of being at or below it cannot exceed one. The second statement reflects the fact that the cdf is a cumulative function of non-negative values, so it can never decline as x rises. In the third statement, we use the symbols $h \downarrow 0$ to indicate that h is approaching zero from above, so that h is positive. This statement is described verbally as 'right continuity' and follows from the fact that we used an inequality in **D6.1.2**; if we had used a strict inequality, we would have left continuity instead.

A typical cdf looks like the one illustrated in Figure 6.1.1.

[1] Some sources define the inequality as strict, *i.e.* $F_X(x) = P(X < x)$. Whether this difference will have practical importance will depend on the nature of the random variable, as we will see below.

FIGURE 6.1.1

Example of a cumulative distribution function



This figure plots the cdf of a random variable X over the interval [3, 3], but this random variable can in fact take any real value, so it is not bounded into this or any other interval. Because this random variable can (albeit rarely in this case) take values below -3 and above 3, the value of the cdf at -3 is not zero, but slightly above zero, and the value at +3 is slightly below 1: in other words, a bit of the probability is left on each side of the [3, 3] interval.[2] (If X had a lower bound at a and an upper bound at b, then we would have $F_X(a) = 0$ and $F_X(b) = 1$, but no such bounds exist here). The fact that the slope is higher near the midpoint of the figure indicates that there are more realizations of the random variable near the midpoint 0, so that small increases in the value that we consider raise the cdf more in that region. Notice that the vertical axis ranges from zero to one, the values that bound the value of the cdf.

[2] This is in fact the cdf of a standard Normal random variable; this distribution will be described in Chapter .

The cdf is a theoretical or population quantity, referring as it does to the true probability that a random variable lies at or below a particular point. The sample counterpart, an *estimate* of the population cdf, is a very straightforward quantity to obtain; it replaces true probabilities with the sample frequencies of observing values at or below a point. Although in principle we can define a cdf at any point, the sample or empirical cdf is usually computed using the sample points as the points of evaluation of the function. In order to define this empirical cdf it is useful to introduce first the concept of the *order statistics* of a sample.

Definition 6.1.4 Let $\{x_1, x_2, \ldots, x_n\}$ be n sample realizations of a random variable X. Let $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ be the same n values, sorted in ascending order, such that $x_{(j+i)} \geq x_{(j)} \quad \forall\, i > 0$. The sorted values $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ are the **order statistics** of the sample.

Note that $x_{(1)}$ and $x_{(n)}$ are therefore the smallest and largest values in the sample.

We can now define the empirical cdf, usually referred to as simply the empirical distribution function, or EDF.

Definition 6.1.5 The empirical cumulative distribution function or **empirical distribution function (EDF)** is defined at each order statistic as $\hat{F}_X(x_{(i)}) \equiv n_i/n$, where n_i is the number of values in the sample less than or equal to $x_{(i)}$ and n is the sample size.

If there are no tied values in the sample, $n_i = i$, because the first order statistic has one sample value at or below it, the second order statistic has two sample values at or below it (itself and the first order statistic) and so on. If there are ties, however, then (for example) the third and fourth order statistics might be equal to the second. In that case the number of values less than or equal to the second order statistic would exceed two, because all four of the first order statistics would qualify. The distinction between samples where there may be ties and samples in which there cannot be ties suggests the distinction between continuous and discrete random variables, which we will treat in the next section. First, we will look at a few examples.

Figure 6.1.2 shows two empirical cdf's based on the country-income data that were plotted in Chapter 2.

The relatively steep parts of the empirical cdf at lower incomes indicate that there are relatively many observations in this region; by contrast, the very high-income regions are relatively flat. The horizontal scales in the two cdf's differ, reflecting the fact that incomes were generally higher in 2000, but the shapes are similar.

Figure 6.1.3 shows the empirical cdf of daily percentage returns in the Dow Jones Industrial Average.

In this example there are many data points, over 22,000. Plotting the empirical cdf nonetheless gives a clear picture of where most of the data lie:

FIGURE 6.1.2

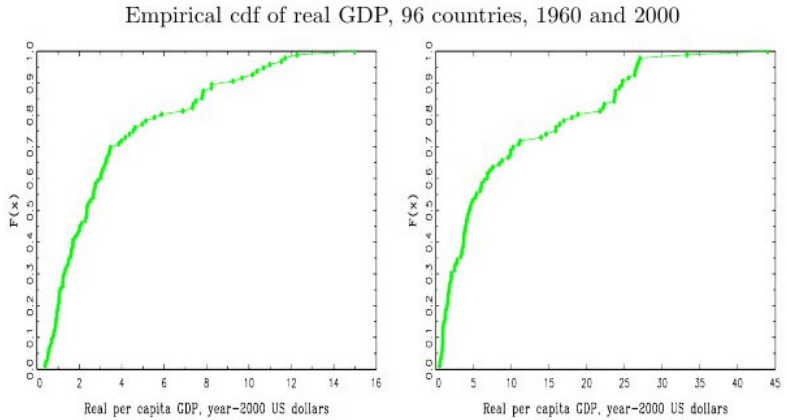
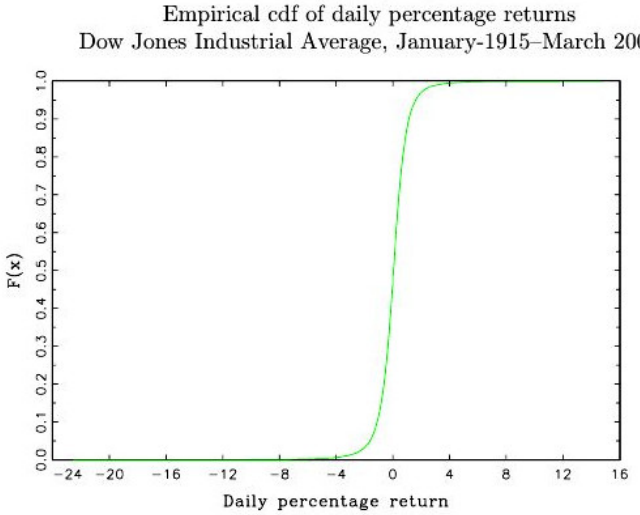


FIGURE 6.1.3



most are within a few percentage points of zero, with rare very large negative or positive returns.

6.2 Continuous and discrete random variables

Some random variables can take on any value in an interval, whereas others can take on only one of a set of values. Some have elements of both of these qualities. For example, respondents to a survey may be asked their weight, how many children they have, and how much time last week they

spent watching television. The first of these can take on any value in some interval (at least if we measure finely enough), while the second can take on only the discrete values 0, 1, 2, ... In the third case, a number of people will have the exact value zero (those who watched no television whatsoever), while the rest will have watched a number of hours which may be anywhere in the interval from zero to 168, the number of hours in a week.[3]

The first type of random variable will be called continuous; the word 'continuous' here is used to indicate that there is some (possibly unbounded) interval, with no gaps, where the variable can take values. The second type will be called discrete, referring to the fact that there are separated points at which values of the random variable can arise. In the third case of time spent watching television, it is again true that any value in an interval is admissible. The random variable nonetheless has some characteristics of a discrete random variable, in that the exact value zero may apply to a number of individuals surveyed, whereas other individuals' values will lies at different points up to 168 hours so that in sufficiently finely measured data, no two will have the same value unless it is zero.

In order to make these ideas precise we now need some more definitions. We begin with the discrete case.

Definition 6.2.1 A random variable is called **discrete** if the values that it can take on form a countable set.

In this case the cdf of the random variable may also be referred to as a discrete cdf.

Since the set of values is countable, the probability of a realization being at or below some point is the sum of the probabilities associated with every value up to and including that point, and the sum of the probabilities of all possible values must be 1. We can define a function which gives these probabilities.

Definition 6.2.2 The **probability mass function (PMF)** $f_X(x)$ of a discrete random variable X which can take on values $x_1, x_2, x_3, \ldots$ is the function $f_X(x_i) = P(X = x_i),\ i = 1, 2, 3, \ldots$

Consider an example from health economics. Figure 6.2.1A-D shows the hypothetical probability that a randomly selected member of a population will be taking a given number of medications at the time of sampling.[4] (Although

[3] Again this assumes that our measurement is arbitrarily fine. In practice, most people will report a number of hours which will be a whole number or a simple fraction, such as 3 1/2. But in principle any time in the interval (0, 168] hours could be reported for those who watched some television during the week. We might also imagine recording this information electronically by a device attached to a television, which measures the exact amount of time that the television is turned on, to a very fine unit of time.

[4] We may consider the figure for 6 medications to represent values of 6 or greater.

these are hypothetical probabilities, they are based on sample values taken from the *Survey of the Health of Canadians*, and a number of the features of these probability mass functions, in particular male/female differences and age-specific differences, are genuine features of the population.) The first two figures (A/B) show females and males respectively for the entire population; the latter two (C/D) show only the probability functions applying to individuals 75 years of age and older.

FIGURE 6.2.1A/B

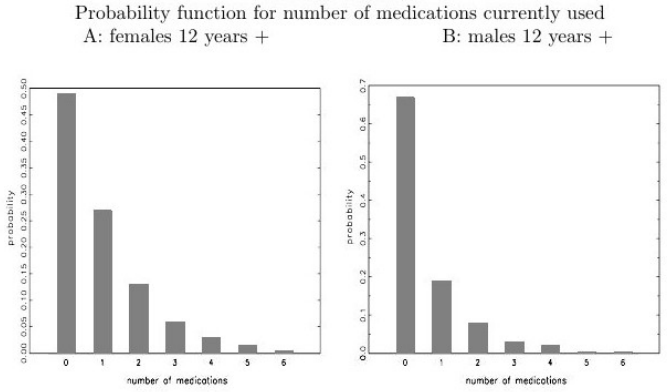
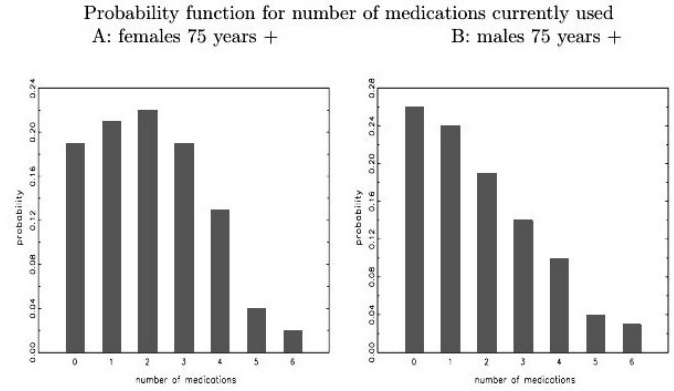


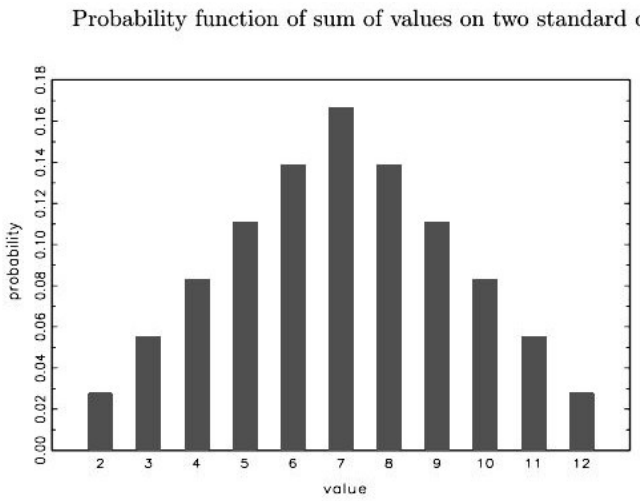
FIGURE 6.2.1C/D



The outcomes form a countable set: if one is taking any quantity at all of a medication, then that medication is included in the set; there can be no non-integer results.

A traditional example where we can determine the probabilities theoretically is the probability mass function for the sum of the values on two dice,

FIGURE 6.2.2



which we worked out in Chapter 5. This probability function is depicted in Figure 6.2.2.

Notice that the probability mass function for the sum of values on two dice is symmetrical around the mean value of 7; as we move to 6 or 8, 5 or 9, and so on the probabilities fall by the same amount whether we go up or down from the mean. For number of unemployment spells the function is asymmetrical (skewed right, with a long upper tail).

For a continuous random variable, we cannot define a set of probabilities as we have just done for the discrete case; the random variable can take on any value in some continuum of points, rather than at a finite set of points, and so we cannot speak of the probability of the random variable taking on a particular value (the probability at any point is zero, a point being an infinitesimal quantity). Instead we can compute the probability that the random variable lies in some interval, and we use the term *probability density* to reflect this difference.

For example, let X be the amount of time that passes between two trades of a particular security on a particular stock exchange, measured to an arbitrarily high precision. Then the probability that the time between two trades is exactly 8 seconds that is, 8.000000... – is zero. However the probability of that time being 8 seconds to the nearest integer number of seconds, that is $7.5 \leq X < 8.5$, is not zero.

In general we can compute the probability that a random variable X lies in an interval from point a to point b as

$$P(a \leq X < b) = \int_a^b f_X(x)\,\mathrm{d}x,$$

where $f_X(x)$ is the *probability density function*, which we now define.

Definition 6.2.3 The **probability density function** of a random variable X is the function $f_X(x)$ such that $\int_{-\infty}^{x} f_X(y)\,\mathrm{d}y = F_X(x)$, where $F_X(x)$ is the cumulative distribution function of X.

The probability density function is often called simply the density.

If the derivative of $F_X(x)$ exists at the point x, then that derivative is the density at point x, $f_X(x)$.

For discrete distributions, we can compute the probability of a random variable lying in an interval by the simple sum of probabilities of possible values lying the interval:

$$P(a \leq X < b) = \sum_{i=1}^{k} P(x_i),$$

where k is the number of possible values of the random variable lying between a and b.

Probability functions and probability density functions must possess a few key properties, which arise because they describe probabilities, and all possible probabilities, for random variables. In particular, they must always give probabilities which are bounded into the interval [0, 1], and must assign total probability of 1 to the universe of possible outcomes.

Definition 6.2.4 Properties of a discrete probability function. Let $\{x_i\}_{i=1}^{\ell}$ be the set of all possible values taken by a random variable X. Then if $P_X(x)$ is the probability mass function of X,

- $P_X(x) > 0$ for $x_i,\ i = 1, 2, \ldots, \ell$
- $P_X(x) = 0$ for $x \notin \{x_i\}_{i=1}^{\ell}$
- $\sum_{i=1}^{\ell} P_X(x_i) = 1$
- $F_X(x_i) = \sum_{x \leq x_i} P_X(x)$.

Analogous properties apply to the probability density function for a continuous random variable.

Definition 6.2.4 Properties of a probability density function. Let X be a continuous random variable which takes values in the interval $[a, b]$. Then if $f_X(x)$ is the probability density function of X,

- $f_X(x) \geq 0\ \forall\ x$

- $f_X(x) = 0$ for $x \notin [a, b]$
- $\int_a^b f_X(x)\,\mathrm{d}x = 1$
- $F_X(x) = \int_a^x f_X(y)\,\mathrm{d}y$.

If the random variable is unbounded on one or both sides, then we replace a or b or both with $-\infty$ or ∞; for a random variable unbounded on both sides, $\int_{-\infty}^{\infty} f_X(x) = 1$.

Note that these definitions do not state explicitly, but imply, that $P_X(x_i) \leq 1$ and $\int_c^d f_X(x)\,\mathrm{d}x \leq 1$ for any x_j or interval $[c, d]$.

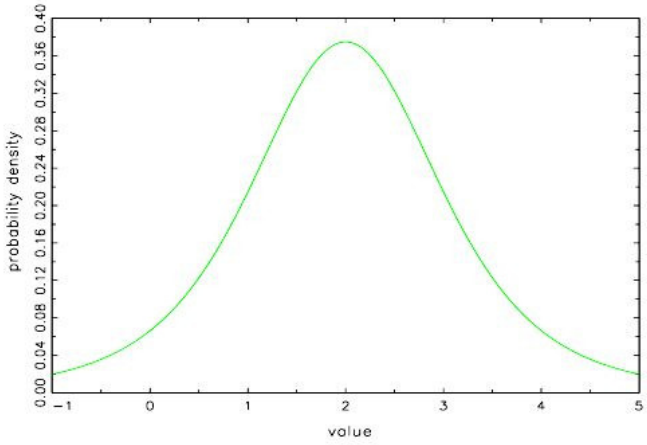
One feature of distribution functions, probability functions or probability density functions which is often interesting is **symmetry** or asymmetry.

Definition 6.2.5 Symmetry. A random variable X is symmetrically distributed around a point μ if $(X-\mu)$ and $-(X-\mu)$ have the same distribution.

That is, the random variable has the same distribution if we reflect it in a line drawn at the point μ. Figure 6.2.3A shows a continuous random variable which is symmetric around the value 2; again, Figure 6.2.2 above showed a discrete random variable symmetrically distributed around the value 7.

FIGURE 6.2.3A

Probability density function of a symmetrically-distributed random variable



This variable is plotted from -1 to 5, that is 2 ± 3; its mean is 2. Although the density is only plotted in this range, it in fact is unbounded in each direction. Because this is the density of a symmetric random variable, the parts of the density on either side of the mean are mirror images of each other.

Although we will be introducing a number of standard distributions in Chapter 9, it will be useful to have a few to work with as examples before that point. We will therefore introduce here the 'Normal' or 'Gaussian' distribution.

A random variable X having a Normal distribution has the density

$$f_X(x) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],$$

where μ and σ^2 are the mean and variance of the random variable. Any finite mean and any finite positive variance are compatible with a Normal distribution, and when these two values are specified, the particular Normal density is fully specified: that is, different densities can be Normal, because they can have different means and variances, but there can only be one Normal distribution with a given mean and variance. The **standard Normal** is the Normal distribution with mean zero and variance of 1. A commonly used notation for the Normal with mean μ and variance σ^2 is $\mathrm{N}(\mu, \sigma^2)$, so that the standard Normal is N(0,1).

Note that the word 'normal' should not be taken to indicate that it is normal, or typical, that data will have this distribution.[5] As we will see in Chapter 12, it will be true that certain functions of data (such as the sum or mean of observations) will tend to be approximately Normal; however, nothing tells us that observations on a random variable itself will be Normal, and in fact simple reasoning will often tell us that data cannot be Normally distributed (because for example the random variable is bounded or asymmetrically distributed, neither of which is compatible with having a Normal distribution).

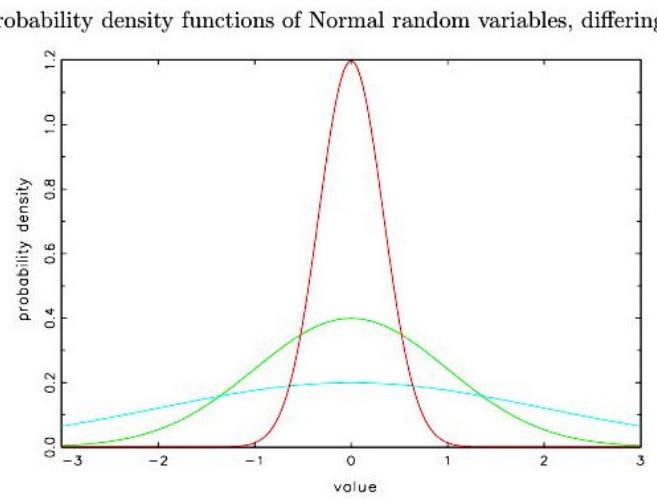
The Normal is a symmetrical and unbounded distribution. The densities of several Normal random variables having different variances are plotted in Figure 6.2.3B.

Note that these Normal distributions look different when plotted on the same scale, and indeed are different, although they are all Normal. By contrast, the distribution of Figure 6.2.3A looks like the Normal with variance of 1; however, it is a different distribution (it is in fact a t-distribution with four 'degrees of freedom', shifted by 2) with some important differences in its properties, as we will see in Chapter 9.

A density function, unlike a distribution function, may not exist; that is, any well defined random variable must have a distribution function, but

[5] Correspondingly, we will capitalize the word 'Normal' in referring to the distribution because it is a proper name, not an adjective. Some prefer the name 'Gaussian' in order to avoid this confusion, referring to the fact that the distribution was used in justifying the least squares method (Chapter 18) by Carl Friedrich Gauß (1777-1855); the distribution was however first described by Abraham de Moivre (1667-1754).

FIGURE 6.2.3B



density may fail to exist because the cdf may not have a finite derivative at one or more points. Nonetheless, when the density does exist and can be estimated, it can be a revealing form in which to represent data.

Consider the data represented in Figures 6.2.4A and 6.2.4B, prices paid at selected auctions, 1968-2001, for paintings by any of a set of 152 Canadian artists; there are approximately 8000 observed sale prices in this sample.[6] Figure 6.2.4A gives the full empirical cdf of these sale prices. The distribution of sale prices is clearly asymmetric; the upper tail is quite long (that is, a small proportion of paintings sell for prices far above the mean, while the lowest prices are much closer to the mean). The largest price paid exceeded 4.5 million dollars; however, very few works sold at prices exceeding one million dollars, and most sold for $10,000 or less. When we plot the entire empirical cdf, therefore, keeping the very high-price observations on the graph compresses the region where most paintings lie to a very small part of the horizontal scale. In Figure 6.2.4B we plot only a part of the empirical cdf, comprising paintings which sold for $10,000 or less; leaving aside the few very large observations allows us to see detail in the region where most observations lie. Figure 6.2.4B reveals more readily that approximately half of sales were at $5,000. or below ($F(5000) \approx 0.5$), and about two thirds at $10,000 or below.

Notice also that there are many 'jumps' in the empirical cdf clearly observable in the finer scale of Figure 6.2.4B: these reflect the fact that sale

[6] See Hodgson and Vorkink (2004).

FIGURE 6.2.4A/B

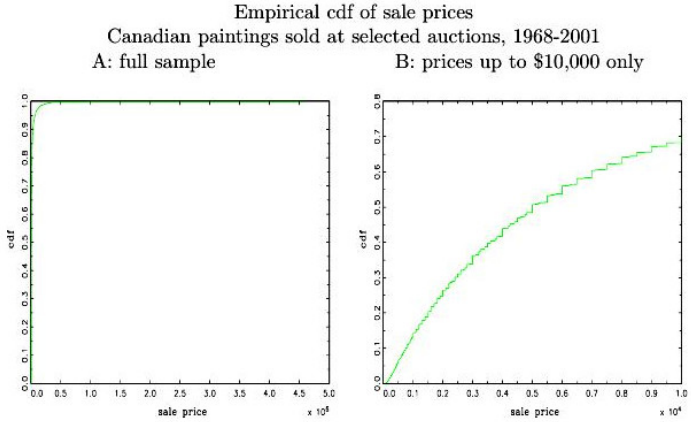
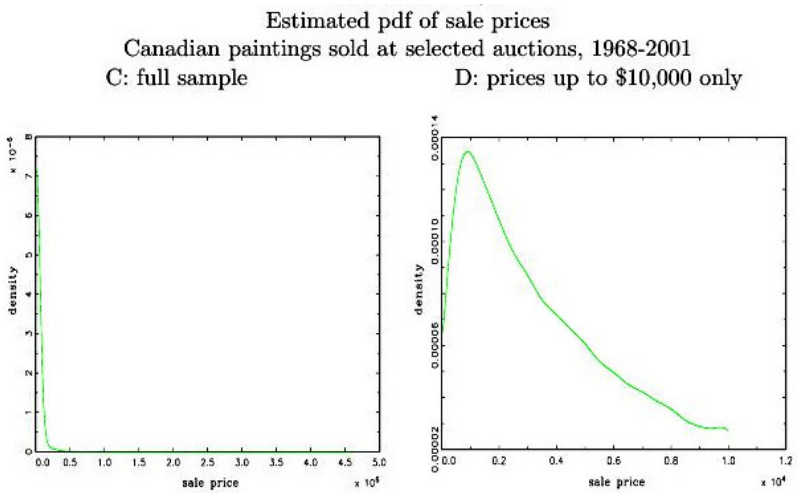


FIGURE 6.2.4C/D



prices tend to be at particular numbers. For example, there are many sales at exactly $5,000, but none at prices between $5,001 and $5,099.

Figures 6.2.4C and 6.2.4D plot the corresponding probability density functions. Again, the first of these plots the pdf for the entire sample; the small number of observations at very high values leads to the large upper tail clearly visible in 6.2.4C, but makes interpreting the values in the region below $10,000 difficult, because of the compressed scale. Figure 6.2.4D plots only the part of the pdf pertaining to observations below $10,000; we now see

clearly that the most common sale prices at auction are in the region of only $1,000, and that the density of prices drops steadily from that point.

Although the cdf and pdf contain the same information in different forms, some conclusions are much easier to draw from the cdf, and others from the pdf.

Many quantities can be computed from the cdf or pdf. In the next chapter we look at some of the most important of these, the moments of the distribution, and in particular the expectation.

Chapter 7

Expectation and moments

One of the most widely applied statistical concepts is that of the 'average', or mean. In this chapter we will define the theoretical quantity corresponding to the sample mean defined in Chapter 3 and will learn about some of its properties. We will also define the general concept of population moments, and show that a number of other moments correspond with sample quantities such as the sample variance.

7.1 Expectation

In the case of a discrete distribution, the expectation or expected value is defined with a simple sum.[1]

Definition 7.1.1 Let X be a discrete random variable taking on values in the finite set $\{x_1, x_2, \ldots, x_k\}$, with corresponding probabilities $\{p_1, p_2, \ldots, p_k\}$. Then the expected value of X is

$$\mathrm{E}(X) = \sum_{i=1}^{k} p_i x_i. \tag{7.1.1}$$

For example, consider the average value obtained from the sum of two dice. The expected value of the sum, using the probabilities of the outcomes 2, 3, ... , 12 that we obtained earlier, is $2(1/36) + 3(2/36) + 4(3/36) + 5(4/36) + 6(5/36) + 7(6/36) + 8(5/36) + 9(4/36) + 10(3/36) + 11(2/36) + 12(1/36) = 252/36 = 7$.

We can also define a discrete random variable with an infinite set of outcomes, as long as the (infinite) set of probabilities continues to sum to one.

[1] Of course, 'expected value' does not mean 'the value that we expect to get' each time we draw from a distribution: in fact the expected value may be one which could not possibly arise. For example, the expected value of the number of children born to a couple may be 2.1. Expected value does have an simple interpretation as a monetary value; if we play n times a game with an expected value of $\$v$, then if n is a large number we will typically gain approximately $\$nv$. This statement can be made much more precise using the laws of large numbers and central limit theorems discussed below.

Recall that when we defined the sample counterpart of this quantity, the sample mean, we took a simple sum divided by the number of observations: $\overline{X} = (1/N)\sum_{i=1}^{N} x_i$. This sample value will converge to the theoretical or population value in (7.1.1) because the x_i's will tend to occur in proportion to their population probabilities. If for example $p_1 \equiv p(x_1) = 0.3$ and $p_2 \equiv p(x_2) = 0.2$, then in the theoretical expression (7.1.1) we put weights of 0.3 and 0.2 on the values of x_1 and x_2. In the sample mean expression, we do not use (or know) these weights: however x_1 will tend to occur 3/2 times as often as x_2, so the same weighting will tend to emerge in large samples without our knowing it a priori. Laws of large numbers (Chapter 10) formalize this idea.

For a continuous distribution, the expectation is an analogue of (7.1.1).

Definition 7.1.2 Let X be a continuous random variable defined on the interval $[a, b]$ and having probability density function $f_X(x)$ at the value x. Then the expectation or expected value of X is

$$\mathrm{E}(X) = \int_a^b x f_X(x)\,\mathrm{d}x. \tag{7.1.2}$$

If the distribution is unbounded on one side or the other, then we have $a = -\infty, b = \infty$, or both.

The point made above concerning the sample mean continues to apply: in the sample mean we use a simple sum rather than the integral of X weighted by the pdf, but values of X will tend to occur relatively often in regions where $f_X(x)$ is relatively high, and the sample mean will again converge to $\mathrm{E}(X)$, assuming that the latter exists.

The expectation has a number of properties that allow us to make statements about the expectations of quantities related to a particular random variable. The following theorem summarizes two of these.

Theorem 7.1.1 Let X be a random variable with expected value $\mathrm{E}(X) = \mu_X$, and let a, b be constants. Then:

i $\mathrm{E}(a + bX) = a + b\mathrm{E}(X) = a + b\mu_X$

ii (Jensen's inequality) Let $g(\cdot)$ be a convex function. [2] Then $\mathrm{E}(g(X)) \geq g(\mathrm{E}(X))$.

Theorem 7.1.1 (ii) is known as Jensen's inequality. Part (i) of the theorem states that if we consider the linear function $a + bX$, it doesn't matter whether we take the expectation first, then plug it into the function, or compute a function of X first, then take the expectation: the result is the same.

[2] A (real-valued) convex function is such that for any real number x_i and point $(x_i, g(x_i))$, there exists a line $h(x)$ such that $h(x_i) = g(x_i)$ and $h(x_j) \leq g(x_j)$ for any other point x_j. That is, at any point, a line tangent to the function lies everywhere at or below the function.

Jensen's inequality states that for a nonlinear function, this does not in general hold: we get a different result if we take the expectation of a function of X, versus taking the same function of the expectation of X. The order of operation matters with non-linearity. If for example $g(X) = X^2$, then Jensen's inequality states that $\mathrm{E}(X^2) \geq (\mathrm{E}(X))^2$: squaring the random variable first and then taking its mean will produce a result at least as great as taking the mean first, then squaring. In the case of a random variable X having the standard Normal distribution (N(0,1)) introduced in the last chapter, we have $\mathrm{E}(X) = 0$, so $(\mathrm{E}(X))^2 = 0$, while $\mathrm{E}(X^2) = 1$ since the mean of X is 0, so its variance is $\mathrm{E}(X^2)$, which by definition of the standard Normal is 1. The inequality is strict in this case.

For a general function $h(x)$, we can define the expectation of the function of the original random variable X as $\mathrm{E}(h(X)) = \sum_{i=1}^{k} p_i h(x_i)$ in the discrete case, and as $\mathrm{E}(h(X)) = \int_a^b h(x) f_X(x)\,\mathrm{d}x$ for a continuous random variable defined on $[a, b]$. To return to the example of two dice, imagine that we play a game in which we receive a payoff of \$5 for every point on the two dice. Then the random variable X is again the sum of the points on the two dice, $h(X) = 5X$, and the expected value of the payoff is given by $5{\cdot}2(1/36) + 5{\cdot}3(2/36) + 5{\cdot}4(3/36) + 5{\cdot}5(4/36) + 5{\cdot}6(5/36) + 5{\cdot}7(6/36) + 5{\cdot}8(5/36) + 5{\cdot}9(4/36) + 5{\cdot}10(3/36) + 5{\cdot}11(2/36) + 5{\cdot}12(1/36) = 5{\cdot}252/36 = 5{\cdot}7 = 35$.

The expected value or mean is the first *moment* of the distribution. Higher moments can also be defined by similar expressions, and lead to population quantities which are estimated by sample quantities in the way that expectation, or population mean, is estimated by the sample mean.

7.2 Higher moments

Expressions analogous to (7.1.1) and (7.1.2) can be used to define a full set of raw or central moments.[3] These moments – as with the first – may not exist for a given random variable, but for the moment (that is, for the time being) we will leave this aside and explore the implications of these definitions.

Definition 7.2.1 The ℓ^{th} raw moment of a random variable is $\mathrm{E}(X^\ell), \ell = 1, 2, \ldots$.

It follows that for a discrete random variable, the ℓ^{th} raw moment is

$$\sum_{i=1}^{k} p_i x_i^\ell \tag{7.2.1}$$

whereas for a continuous random variable it is

$$\int_a^b x^\ell f_X(x)\,\mathrm{d}x. \tag{7.2.2}$$

[3] The word 'moment' is used because of an analogy to concepts from elementary physics, such as moment of inertia.

Raw moments are affected by the mean of the distribution of the random variable. By contrast, the *central moments* remove the effect of the mean, so that two distributions which are identical except for a shift caused by differing means will have the same central moments of order two and higher.

Definition 7.2.2 The ℓ^{th} central moment of a random variable X is $\mathrm{E}((X - \mu_X)^\ell)$, $\quad \ell = 2, \ldots$ The second central moment is called the variance, $\sigma_X^2 \equiv \mathrm{E}((X - \mu_X)^2)$.

Note that the first central moment would simply be zero for any distribution. For a discrete random variable, the ℓ^{th} central moment is

$$\sum_{i=1}^{k} p_i (x_i - \mu_X)^\ell, \tag{7.2.3}$$

whereas for a continuous random variable it is

$$\int_a^b (x - \mu_X)^\ell f_X(x)\,\mathrm{d}x. \tag{7.2.4}$$

A number of the descriptive statistics that we defined earlier can now be seen as estimates, or sample counterparts, of these population central moments. The sample variance is an estimate of the second central moment (the population variance), whereas the skewness and kurtosis measures that we defined are functions of the third and fourth central moments.[4] To define these functions, let $m_2, m_3, m_4 \ldots$ be the central moments of a distribution. Then the population variance is $m_2 = \mathrm{E}((X - \mu_X)^2)$, the population coefficient of skewness corresponding to the sample measure in **D3.8** is

$$\frac{m_3}{(m_2)^{3/2}}, \tag{7.2.5}$$

and the population coefficient of kurtosis corresponding with the sample measure in **D3.10** is

$$\frac{m_4}{(m_2)^2}. \tag{7.2.6}$$

The moments of a distribution, or functions based on them such as these, are useful as descriptive statistics because they reveal different properties of the distribution. The second central moment, or variance, is an indicator of the dispersion of values around the mean. The third and other odd-numbered moments are zero for symmetric distributions since positive and negative values cancel; therefore deviations from zero, scaled to account for differing degrees of dispersion, can be taken to measure asymmetry.[5]

[4] We often refer to the square root of the population variance, σ, as the standard deviation; its estimate or sample counterpart is called the standard error.

[5] Positive and negative values can also cancel, yielding a third moment of zero, for some non-symmetric distributions; therefore a zero third moment is a necessary, but not sufficient, condition for symmetry.

The variance or second central moment is of great importance in characterizing our uncertain knowledge of where an estimated quantity lies, or more precisely in allowing us to define intervals within which random variables will fall with particular probability. Chebychev's inequality (section 7.3 below) provides one way of using the variance (or its square root, the standard deviation) in this way, although more precise statements will be possible when we can characterize estimated values as having particular distributions (Chapters 14 and 16 in particular). Note that if we are dealing with a discrete distribution, using the definition of expectation for discrete random variables **D7.1.1**, we can translate the expression $\mathrm{E}(X - \mu)^2$ into $\sum_{i=1}^{k} p_i (x_i - \mu)^2$ for this case; for a continuous distribution, using **D7.1.2** we obtain that $\mathrm{E}(X - \mu)^2 = \int_a^b (x - \mu)^2 f_X(x)\,\mathrm{d}x$. Similar expressions for higher-order moments follow also from the definitions of expectation for discrete and continuous random variables.

Recall that we have $\mathrm{E}(a + bX) = a + b\mathrm{E}(X) = a + b\mu_X$ for the first moment of a linear transformation of a random variable. The variance is a central moment, however, and so the shift by a in $a + bX$ is subtracted out in computing the variance and has no effect. In fact it is easy to see, using the notation $\mathrm{Var}(X)$ for the variance of X, that $\mathrm{Var}(a + bX) = \mathrm{E}[a + bX - \mathrm{E}(a + bX)]^2 = \mathrm{E}[a + bX - (a + b\mu)]^2 = \mathrm{E}[bX - b\mu)]^2 = \mathrm{E}[b(X - \mu)]^2 = b^2\mathrm{E}[X - \mu]^2 = b^2\,\mathrm{Var}(X)$.

As we move to higher order even-numbered moments, we give progressively more emphasis to the most extreme (farthest from the mean) values in a distribution. To illustrate this, consider the following mean-zero discrete distributions represented through their probability functions:

$$X = \begin{cases} 1 & (p = 0.49) \\ -1 & (p = 0.49) \\ 10 & (p = 0.01) \\ -10 & (p = 0.01) \end{cases}$$

$$Y = \begin{cases} 2 & (p = 0.49) \\ -2 & (p = 0.49) \\ 10 & (p = 0.01) \\ -10 & (p = 0.01) \end{cases}$$

$$Z = \begin{cases} a & (p = 0.49) \\ -a & (p = 0.49) \\ 4 & (p = 0.01) \\ -4 & (p = 0.01) \end{cases}$$

where we choose the value a in the definition of Z to give Z the same variance as Y, which requires $a = 2.3904572$. Table 7.2.1 gives the second, fourth and sixth moments of these distributions, where the ℓ^{th} central moment is computed using equation (7.2.2) as $\mathrm{E}(V - 0)^\ell = \sum_{i=1}^{4} p_i v_i^\ell$ for $V = \{X, Y, Z\}$ and the four possible values v_i and their probabilities, just given; note that each random variable is constructed to have a mean of zero.

Table 7.2.1
Moments 2, 4 and 6 of the random variables X, Y and Z

Central moment	X	Y	Z
2	2.98	5.92	5.92
4	200.98	215/68	37.12
6	20000.98	20062.72	264.78

The second moment of X is less than that of Y and Z; the second moment computation is dominated by the values that occur 98% of the time. When we look at the fourth moment, however, X and Y are now close, and far above Z, since Z lacks the very large values (± 10) occuring in the other random variables; by the sixth moment, the 98% of the distribution composed of relatively small values makes little difference, and the computation is dominated by the 2% of extreme values. Now X and Y are very close, while Z's fourth moment is far smaller despite the fact that on 98% of values, Z has the larger deviations from the mean. That is, as we go to higher and higher moments, results are increasingly dominated by the most extreme values in the distribution.

With these definitions of moments, we are now able to state the Chebychev inequality.

7.3 The Chebychev inequality

One of the key questions that we will often want to answer about any distribution is: what is the probability of an observation above or below a certain value, or in a certain interval? If we know the form of the distribution, as in examples that we will see in Chapter 9, then we can usually answer this question in one of various ways. However, with an empirical distribution of unknown form, we cannot compute probabilities using methods that depend on knowledge of this unknown mathematical form. Chebychev's inequality nonetheless allows us to state a bound (not an exact value) for one important question of this type, that is, the probability that an observation will lie in a symmetric interval around the mean. The theorem requires only weak conditions; nonetheless it is important to note that these conditions do not always hold, so that the Chebychev inequality does not hold for every possible distribution.

Theorem 7.3.1 Let X be a random variable having finite mean and variance, μ and σ^2. Then, for any $\ell > 0$,

$$P(|X - \mu| \geq \ell\sigma) \leq \frac{1}{\ell^2}. \tag{7.3.1}$$

That is, the probability of X being greater than ℓ standard deviations from its mean is no greater than ℓ^{-2}. The statement has useful content only

for $\ell > 1$, since for $\ell \leq 1$ it does not bound the probability to any value below 1. Alternatively, we can write $P(|X - \mu| < \ell\sigma) \geq (1 - 1/\ell^2)$ or: the probability that X is less than ℓ standard deviations from its mean is at least $1/\ell^2$. Again, note that this is only a lower bound on the probability contained in the interval of ℓ standard deviations from the mean: if we knew the form of the distribution function, then we could calculate the exact value, which will typically be substantially higher. For example, the Normal distribution contains about 95% of the probability within two standard errors of the mean, well above the bound of at least 75% specified by the Chebychev inequality. Both statements are correct, but the statement based on knowledge of the exact distribution, where this is available, is more precise.

This inequality can be proved as a consequence of a more general relationship involving probability and expectation, called the *Markov inequality*, which is stated and proved in the Appendix:

Theorem 7.3.2 Let X be a random variable and $g(\cdot)$ a non-negative function on $\mathcal{R}$ such that $\mathrm{E}(g(X))$ exists. Then

$$P(g(X) \geq \ell) \leq \frac{\mathrm{E}(g(X))}{\ell} \quad \forall\, \ell > 0. \tag{7.3.2}$$

Proof of Theorem 7.3.2.[6]

For simplicity, we will assume that the density $f_X(x)$ exists. Define the *indicator functions*:

$$\mathrm{I}(g(x) < \ell) = \begin{cases} 1 & \text{if } g(x) < \ell \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \mathrm{I}(g(x) \geq \ell) = \begin{cases} 1 & \text{if } g(x) \geq \ell \\ 0 & \text{otherwise.} \end{cases}$$

for any $\ell > 0$. Note that $\mathrm{I}(g(x) < \ell) + \mathrm{I}(g(x) \geq \ell) = 1$. Then

$$\begin{aligned} \mathrm{E}(g(X)) &= \int_{-\infty}^{\infty} g(x) f_X(x)\, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} g(x)\mathrm{I}(g(x) < \ell) f_X(x)\, \mathrm{d}x + \int_{-\infty}^{\infty} g(x)\mathrm{I}(g(x) \geq \ell)) f_X(x)\, \mathrm{d}x. \end{aligned}$$

Since $g(x) \geq 0$, the first integral above is non-negative, and so $\mathrm{E}(g(X))$ is no smaller than the second integral, for which we have

$$\int_{-\infty}^{\infty} g(x)\mathrm{I}(g(x) \geq \ell) f_X(x)\, \mathrm{d}x \geq \ell \int_{-\infty}^{\infty} \mathrm{I}(g(x) \geq \ell) f_X(x)\, \mathrm{d}x = \ell\, P(g(X) \geq \ell),$$

and so

$$\mathrm{E}(g(X)) \geq \ell\, P(g(X) \geq \ell) \quad \text{or} \quad P(g(X) \geq \ell) < \frac{\mathrm{E}(g(X))}{\ell}.$$

This completes the proof. ∎

Proof of Theorem 7.3.1

Notice first that, if a random variable Y is centred, with expectation zero, then its variance is $\mathrm{E}Y^2 = \mathrm{E}|Y|^2$. Notice next that that the variance of $(X - \mu)/\sigma$ is 1, as therefore is the variance of $|(X - \mu)/\sigma|$, which is

$$\int_{-\infty}^{\infty} \left(\frac{x - \mu}{\sigma}\right)^2 f_X(x)\, \mathrm{d}x.$$

The inequality $|(x - \mu)/\sigma| \geq \ell$ holds if and only if one of the following two inequalities holds: $(x - \mu)/\sigma \geq \ell$ or $(x - \mu)/\sigma \leq -\ell$, that is, $x \geq \mu + \sigma\ell$ or $x \leq \mu - \sigma\ell$. The integrand in the variance above is non-negative, and so the integral is no smaller than the integral of the same integrand over a subset of the range of integration. The variance is therefore no smaller than

$$\left[\int_{-\infty}^{\mu - \sigma\ell} + \int_{\mu + \sigma\ell}^{\infty}\right] \left(\frac{x - \mu}{\sigma}\right)^2 f_X(x)\, \mathrm{d}x.$$

which in turn is no smaller than $\ell^2 P(|X - \mu|/\sigma \geq \ell)$. But, since the variance is 1, this leads to the inequality

$$1 \geq \ell^2 P(|X - \mu|/\sigma \geq \ell) \quad \text{or} \quad P(|X - \mu| \geq \ell\sigma) \leq \frac{1}{\ell^2}.$$

This completes the proof. ∎

[6] These proofs are based on those given by Hogg and Craig (1959) and Mood *et al.* (1974).

Chapter 8

Joint and conditional distributions

So far we have discussed only one random variable at a time. Often this is not sufficient: the relationships among random variables (and therefore among data series that we may measure) are some of the most interesting things that we can learn about through statistics.

When we discuss a joint distribution of random variables, we are referring to the way in which two or more random variables are distributed in relation to each other, and we can ask for example not just how likely it is that $X = 1$, (to take a discrete case), but how likely it is that $X = 1$ when $Y = 2$, or when $Y = 3$. These variables have individual ('univariate' or 'marginal') distributions, as we would use if we were analyzing them one at a time, but the joint distribution can reveal additional information. When we treat joint distributions, we are able to define corresponding quantities of interest such as the covariance, correlation, and conditional probability which help us to describe some of the relationships among variables just as the sample mean or variance help us to describe a single variable.

8.1 Joint discrete and continuous distributions

A joint cumulative distribution function is defined analogously to the CDF for a single random variable, and is again a quantity that exists for any well-defined random variable and so is a fundamental element in our treatment of random phenomena.

Definition 8.1.1 The *joint cumulative distribution function* of a set of k random variables $X_1, X_2, \ldots, X_k$ is a function $F_{X_1,X_2,\ldots,X_k}(x_1, x_2, \ldots, x_k)$ defined on the real line, with range [0,1], such that

$$F_{X_1,X_2,\ldots,X_k}(x_1, x_2, \ldots, x_k) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_k \leq x_k)$$

for every set of real values $x_1, x_2, \ldots, x_k$.

The joint CDF therefore describes the probability that a condition holds on each one of k random variables. For example, consider the joint distribution of two approximately continuous random variables that we could measure in a survey of individuals, income and the individual's net asset value.[1] Then we

[1] As usual in economic data we have to be careful to define precisely what we are measuring. By income we would normally mean income as defined in an income

might be interested in $F(20,000,0)$ and $F(100,000,0)$, the cumulative probabilities that an individual has income up to \$20,000 and has negative or zero net assets, and that an individual has income up to \$100,000 and has negative or zero net assets. Note that the latter number, perhaps counter-intuitively, is higher: we are not describing only people with incomes of \$100,000 (who are less likely to have net debt than people with incomes of \$20,000), but rather we describe the probability that two conditions both hold: income is less than some value, and net assets ≤ 0. The higher the income that we specify, the more people will fulfill the first condition and the higher will be the cumulative probability (unless there is absolutely no one with income between \$20,000 and \$100,000 who has net assets ≤ 0.)

A joint distribution may also include a mix of continuous and discrete random variables. Consider the joint distribution of individual income and years of formal schooling, where the latter is measured as an integer. In this case the joint distribution of the continuous random variable income and the discrete random variable 'years of schooling' is a set of functions describing the accumulation of probability as income increases for any value of the number of years of schooling, increasing to higher levels as we move discretely to higher values of the number of years.

8.2 Conditional distribution functions

Conditional distributions describe random variables when information about one or more other random variables is also available. Because the links among different variables are among the most interesting features of economic data, our data analysis typically involves conditional distributions, implicitly or explicitly. For example, when we forecast a random variable we will usually consider an expected value given ('conditional on') values of the same or other random variables observed in the past. As another example, when we try to understand an economic process such as an individual's income, we consider what income is likely to be conditional on information about education, age, place of residence, and so on. It may be interesting to know that mean income of full-time workers in a given country and time are (*e.g.*) \$54,000 (the unconditional mean); it may also be useful that, for a worker aged 55 with a professional degree and living in Vancouver, mean earnings are (*e.g.*) \$78,500 (the conditional mean, given the three pieces of information just mentioned).

It may be easiest to understand conditional distributions by beginning with a simple discrete example. In Table 8.2.1 we illustrate a case of two

tax code, which would therefore include not only employment income but also rental income, interest and dividend income, and taxable capital gains income. We might prefer a different measure, including not only earned income, but the true change in value of capital assets, but this may be unavailable to us. The value of the individual's net assets may be similarly difficult to obtain, and we might use a simple measure comprising an approximate value of real estate assets, pensions, and net financial assets at a particular point in time.

jointly distributed discrete random variables, each of which can take on only three values ($-1, 0, 1$ for X; $2, 3, 4$ for Y). The table gives the probabilities of each of the pairings (x_i, y_i), $i, j = 1, 2, 3$.

Table 8.2.1
Example of a joint probability distribution
for two discrete random variables

	$X=-1$	$X=0$	$X=1$
$Y=2$	0.16	0.39	0.20
$Y=3$	0.08	0.08	0.04
$Y=4$	0.01	0.03	0.01

The table indicates, for example, that the probability that $Y = 2$ and $X = -1$ is 0.16; similarly, $P(Y = 4, X = 0) = 0.03$. This table fully describes the discrete joint distribution, and we can use it to make both unconditional and conditional statements.

By adding the columns, we can see that $P(X = -1) = 0.25(0.16 + 0.08 + 0.01)$, $P(X = 0) = 0.50$, and $P(X = 1) = 0.25$. Adding the rows, we have $P(Y = 2) = 0.75, P(Y = 3) = 0.2, P(Y = 4) = 0.05$. These unconditional (or 'marginal', a traditional term reflecting the fact that the unconditional probabilities were written in the margins of tables such as this) probabilities do not refer to the value of the other random variable.

Now consider the case in which $Y = 4$. This condition holds with $p = 0.05$; the joint probability that $Y = 4$ and $X = 0$ is 0.03. That is, of the 5% of probability accounted for by cases with $Y = 4$, 60% (0.03/0.05) involves $X = 0$ as well; the probability that $X = 0$ given that $Y = 4$ is 0.6. When $Y = 3$, which accounts for 20% of total probability, the (conditional) probability that $X = 0$ is only 0.4 (0.08/0.20): of the 20% of total probability accounted for by the $Y = 3$ case, 40% entails $X = 0$ as well. These two conditional probability statements can be expressed as $P(X = 0|Y = 4) = 0.6$; $P(X = 0|Y = 3) = 0.4$. Note that each of these conditional probabilities differs from the unconditional $P(X = 0) = 0.5$; knowing the value of Y helps us to refine out knowledge of the distribution of X, and the conditioning information concerning Y has value. Note also that the probability that the condition will arise is irrelevant to the calculation: we are computing the probability of X being 0 if Y is 4, and the probability that Y actually does have the value 4 is irrelevant, as long as it is not zero. The probability of Y taking on the given value is relevant to the joint probability but not to the conditional.

The information in the table also allows us to make statements conditional on X. For example, $P(Y = 4|X = 0) = 0.03/0.50 = 0.06$, slightly higher than the unconditional probability that $Y = 4$; $P(Y = 2|X = -1) = 0.16/0.25 = 0.64$, lower than the unconditional $P(Y = 2) = 0.75$.

In each of these cases, we compute the conditional by dividing the joint probability (the probability that both conditions hold) by the probability of

the condition, because we want to compute a proportion of the total probability accounted for by the condition that we use. We can formalize this in the following definitions (given here for two random variables, although these can be extended to an arbitrary number).

Definition 8.2.1 The conditional probability function of Y given X for two jointly distributed discrete random variables with joint probability function $P_{X,Y}(x,y)$ is defined as

$$P_{Y|X}(y|x) = \frac{P_{X,Y}(x,y)}{P_X(x)}, \qquad \text{for } P_X(x) > 0.$$

An analogous definition applies to the continuous case, where we use the density function.

Definition 8.2.2 The conditional density function of Y given X for two jointly distributed continuous random variables having the joint probability density function $f_{X,Y}(x,y)$ is defined as

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \qquad \text{for } f_X(x) > 0$$

Note that the discrete conditional probability function and the continuous conditional density function have the properties that each takes on non-negative values, the sum of the conditional probabilities across all outcomes y_i is 1, and the integral of the conditional p.d.f. with respect to Y is 1.

8.3 Independence

We indicated above examples in which conditioning information was useful, because that information allowed us to make a probability statement specific to given circumstances, as opposed to relying on probability statements applicable on average across various circumstances. However, information that we might condition on is not always useful; sometimes two random variables tell us nothing about each other. We can formalize this through the concept of statistical independence of random variables, analogous to the statistical independence of two events defined in Chapter 5.

Definition 8.3.1 Two continuous random variables X and Y with joint probability density function $f_{X,Y}(x,y)$ are statistically independent if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

The analogous definition holds for discrete random variables where we use the probability function instead of the p.d.f. As in the previous section, we have used two random variables in the definition, but the same holds for any number of random variables: if a set of random variables are statistically independent, then the joint p.d.f. is equal to the product of the densities of the individual random variables.

Recall the definitions **8.2.1** and **8.2.2** of conditional density or probability functions. In the continuous case, for example, we can re-arrange the definition to state that $f_{X,Y}(x,y) = f_{Y|X}(y|x)f_X(x)$ or $f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y)$. Consistency with definition **8.3.1** then implies that, if the two random variables are independent, $f_{X|Y}(x|y) = f_X(x)$ and $f_{Y|X}(y|x) = f_Y(y)$: that is, the conditional distributions of X given Y and Y given X are the same as the unconditional distributions of X and Y ; neither random variable conveys any information about the distribution of the other.

It can be difficult to think of economic variables that are literally independent of each other, although independence may be a reasonable approximation in many cases. For example, we might imagine that (within the class of people who own cars) the colour of an individual's car has nothing to do with his or her income, so that these two variables are independent. However, numerous mechanisms could create a weak relationship between the two, and break independence. Perhaps green cars are unfashionable, therefore cheaper to buy used, so lower-income people may be more likely to buy used green cars. Perhaps the mix of colours produced in new cars has tended to change, with more silver cars and fewer blue cars produced than six or seven years ago. In this case higher income people, who are relatively likely to have new cars, will be relatively likely to own silver cars and relatively unlikely to own blue ones. While it is hard to imagine such effects being very strong, even a tiny link is sufficient to break independence.

8.4 Covariance and correlation

The concepts of covariance and correlation are extremely widely used, and often used in misleading arguments involving non-experimental data, so it is particularly important for students of such data to understand clearly what is implied by these concepts.

Recall that in Chapter 7 we defined the (population) variance of a random variable X as $\mathrm{E}(X-\mu)^2$, the mean of the distribution of the random variable $(X-\mu)^2$. We also define covariance as the mean of the distribution of a new random variable formed from the underlying variables X and Y.[2]

Definition 8.4.1 The covariance between two random variables X and Y is defined as

$$\mathrm{cov}(X,Y) = \mathrm{E}\big((X-\mu_X)(Y-\mu_Y)\big).$$

[2] It is assumed in **D8.4.1** that the expectation exists; otherwise the covariance is not defined. Recall from the earlier definition of an expectation that we can write

$$\mathrm{cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(X-\mu_X)(Y-\mu_Y)f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y,$$

and so as usual the existence of the expectation depends on existence of an integral.

The covariance has a number of limitations for practical purposes, which lead us to work with a transformation of it (the correlation) in many circumstances. For example, we might be interested in the way in which changes in new vehicle purchases (X) move with changes in the oil price (Y). If we measure new vehicle purchases in thousands instead of individual units, the variable X and its mean μ_X will be smaller by a factor of 1000, so by **D8.4.1** the covariance will be smaller by 1000 as well; the covariance is not independent of scale. Relatedly, the covariance can take any value, so we cannot point to any value of the covariance as being objectively large or small, indicating a strong or weak relation. By transforming to correlation, we remove these difficulties.[3]

Definition 8.4.2 The correlation between two random variables X and Y is defined as

$$\text{corr}(X,Y) = \frac{\text{E}\big((X-\mu_X)(Y-\mu_Y)\big)}{\sigma_X \sigma_Y},\ \sigma_X, \sigma_Y > 0.$$

Recall that we have used the standard notation σ^2 for the variance of a random variable, so that σ_X, σ_Y are the standard deviations of the two random variables.

The correlation is free of scale effects. If we measure new vehicle purchases in thousands instead of units, we take a factor of 1000 out from both the numerator ($(X-\mu_X)$) and the denominator (σ_X), and the correlation is unaffected; any scaling of either random variable enters numerator and denominator and cancels. Moreover, the correlation always lies in the interval $[-1,1]$, a fact which follows from the Cauchy-Schwarz inequality.

Theorem 8.4.1 Let W and Z be two random variables with $\text{E}(W^2)$ and $\text{E}(Z^2) < \infty$. Then $(\text{E}(WZ))^2 \leq \text{E}(W^2)\text{E}(Z^2)$; $(\text{E}(WZ))^2 = \text{E}(W^2)\text{E}(Z^2)$ if and only if $P(W = kZ) = 1$ for some k.

The latter condition, that $P(W = kZ) = 1$, essentially states that W is simply a re-scaling of Z except possibly on a finite set of points.

The proof that $|\text{corr}(X,Y)| \leq 1$ is listed as an exercise at the end of this chapter. This result means that we can always think of values near 1 as 'high' correlations, near zero as 'low', and so on.

A positive correlation means that relatively high (*i.e.*, above the mean) values of one variable have some tendency to be accompanied by relatively high values of the other; conversely a negative correlation implies means that relatively high values of one tend to be accompanied by relatively low values of the other. In either case, if the absolute value of the correlation is near zero, this tendency is weak and may be difficult to spot in empirical data.

[3] We again assume the existence of the expectations (plural, because σ_X and σ_Y are also defined by expectations) in making this definition.

Correlation never implies causation, but it seems to be a constant temptation to empirical researchers to think in this way. In some cases, of course, a causal link can be responsible for a correlation, but in others a correlation may arise because each of two variables is linked to some underlying third factor or set of other factors, while having no direct impact on the other. For example, income and education are positively correlated in cross-sectional samples of individuals. In this case there is almost certainly a causal link: education tends to expand opportunities, allowing individuals to choose higher-income employment (although there are many well-known examples of very wealthy people, usually self-employed, who dropped out of school early).[4] We infer causality from reasoning about the way in which markets for skilled labour function, however, not from the existence of a positive correlation. As another example, in samples of people of approximately the same age, consumption of champagne and strawberries may be negatively correlated with indicators of health problems such as number of doctor visits, days spent in hospital, number of medications prescribed, *etc* .[5] While there may be some protective impact on health of champagne or strawberry consumption (or some negative effect, particularly in extreme champagne-consumption cases), the more important effect seems likely to be that individuals' champagne and strawberry consumption will tend to be higher in higher-income individuals, who also tend to have fewer health problems (a fact often explained by their being better able to afford health-related services, better able to avoid unhealthy conditions, and being on average better informed, being on average better educated as well). Other hidden factors, not direct causation, are behind the correlation.

8.5 Conditional expectation

Conditional expectation is another concept that is very widely used in analyzing the relationships among economic variables. It is defined using the conditional density function defined earlier.

Definition 8.5.1 The conditional expectation of a continuous random variable Y, conditional on a continuous random variable X taking the value x, is

$$\text{E}(Y|X=x) = \int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\, \text{d}y.$$

An analogous definition applies to discrete random variables.

Although we are conditioning on a particular value of the random variable X in making this definition, and so use the notation '$\text{E}(Y|X=x)$', we also frequently write simply $\text{E}(Y|X)$.

[4] Don't try this at home.

[5] The author has never seen a study to this effect, but is inclined to guess that it's true.

Conditioning on a particular value of a random variable allows us to make more informative statements than we can make by using the unconditional expectation alone. For example, consider annual earnings in a sample of people whose incomes are available from tax records. The mean earnings in the group of people that we sample may be, say, $51,245. If we pick out a person at random from the sample, then, this might be the best estimate of what that person's earnings will turn out to be. But if we have other information, for example the individual's age and number of years of schooling, we may be able to make a more precise statement; earnings tend to increase with age, up to a certain point, and tend to increase also with number of years of schooling. So if we observe an individual aged 52 with 16 years of formal education, we would expect income above the overall mean (since 52 and 16 are relatively high values of variables associated with higher income).

The function relating the conditional expectation of a random variable to the conditioning variables is called the *conditional expectation function*, and is also called the *regression function*. Linear functions are often used to approximate the conditional expectation function in empirical work; correspondingly, *linear regression* is a technique which is extremely commonly applied in empirical economics and finance. In the example just given, a linear conditional expectation function could be written as

$$\mathrm{E}(Y|X_1, X_2) = a + bX_1 + cX_2, \tag{8.5.1}$$

where Y is income, X_1 is age and X_2 is years of formal education. The parameters of this linear model, a, b, c, can be estimated by a variety of techniques; Chapter 18 gives an introduction. Of course, conditional expectation functions can also be non-linear, and non-linear functions or functions of unknown form can also be estimated using standard statistical techniques.

Notice that the relationship given in equation 8.5.1 simply refers to the mean of the distribution of Y given values of X_1 and X_2; it does not imply that other variables are irrelevant (that is, a more elaborate conditional expectation function could be written if we observed additional conditioning variables), and it does not imply a causal link between the values X_1, X_2 and Y.

8.6 The bivariate Normal distribution

In Chapter 6 we introduced the univariate Normal distribution. The bivariate Normal is a straightforward extension, and a simple example of a joint continuous distribution. Inspecting the graphs of densities of some examples of bivariate Normals can help improve our intuition about joint and marginal distributions, correlations, and so on. Since this distribution involves two Normally distributed random variables, the form of the joint distribution will depend on the correlation between them; the definition will involve therefore not only the means and variances (or standard deviations) of each random variable, but an additional parameter representing the correlation between the two.

For two jointly Normally distributed random variables X_1 and X_2, having means μ_1 and μ_2 , and standard deviations σ_1 and σ_2 respectively, and with correlation ρ, the joint probability density function is

$$f_{1,2}(X_1, X_2) = [2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}]^{-1} \cdot \exp\left[-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right].$$

Since the term in the exponential is scaled by a negative quantity, the density is higher where the inner term in square brackets is smaller. Therefore the density is higher as X_1 is close to its mean, X_2 is close to its mean, and (to make the final term small) the deviations of each from their means have the same sign if ρ is positive, opposite signs if ρ is negative. Note that the third term in the inner square brackets can be no larger than the sum of the other two, so that this whole term is always positive.

Figure 8.1.1 A/B graphs the densities in two simple cases of equal variance for each random variable, and correlation of zero between them. Probability density drops off symmetrically in any direction around the peak, which lies above the point on the plane where each random variable is equal to its mean. Contour lines (lines joining points of equal height) on these graphs would be circles. In Figure 8.1.1B the means are shifted to (2,2) rather than (0,0); the graph is correspondingly moved so that its peak lies over this point, but has the identical shape.

In Figure 8.1.1C the correlation between the two random variables remains zero, but they now have different variances; one of the random variables has a marginal distribution like the flatter density in Figure 6.2.3B (*v.i.* Chapter 6), the other has marginal distribution more like the most peaked distribution in Figure 6.2.3B. Similarly, in the multivariate case probability density drops more quickly from the peak along the axis representing movements in the lower variance random variable, more slowly for the high variance random variable.

Figure 8.1.1D embodies a non-zero (positive) correlation between the two random variables. We now see that density is relatively high along and near the axis where $X_1 = X_2$, reflecting the term $-2\rho\big((x_1-\mu_1)/\sigma_1\big)\big((x_2-\mu_2)/\sigma_2\big)$ in the joint density. In Figures 8.1.1.E and 8.1.1.F we plot joint densities with higher values of the correlation ρ, and as ρ approaches 1, we see that scaled (by standard deviation) discrepancies between each variable and its mean are close to being identical, so that the joint density approaches a two-dimensional curve.

FIGURE 8.1.1.A/B

Bivariate Normal densities

$\sigma_1^2 = \sigma_2^2 = 1; \rho = 0$

A: $\mu_1 = \mu_2 = 0$ B: $\mu_1 = \mu_2 = 2$

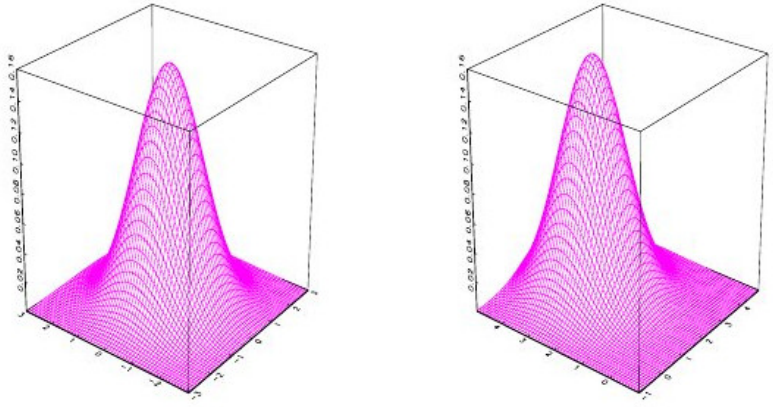


FIGURE 8.1.1.C/D

Bivariate Normal densities

$\mu_1 = \mu_2 = 0;$

C: $\sigma_1^2 = 4, \sigma_2^2 = \frac{1}{2}, \rho = 0.0$ D: $\sigma_1^2 = \sigma_2^2 = 1, \rho = 0.5$

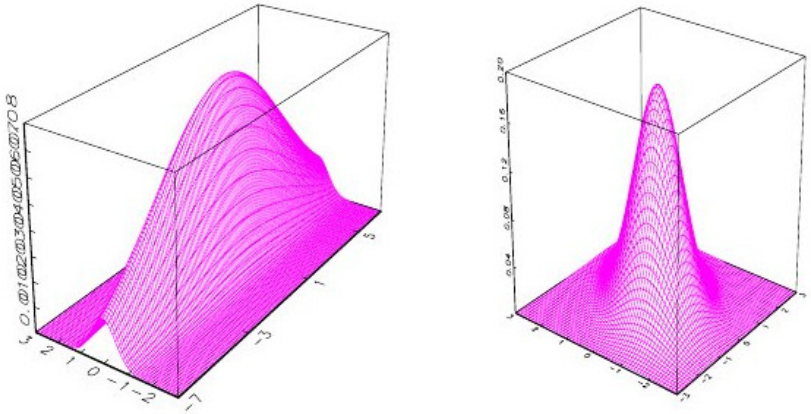
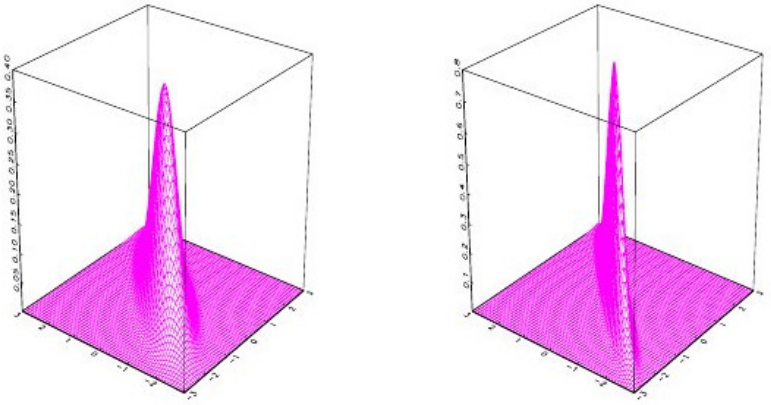


FIGURE 8.1.1.E/F

Bivariate Normal densities

$\mu_1 = \mu_2 = 0;\ \sigma_1^2 = \sigma_2^2 = 1;$

E: $\rho = 0.90$ F: $\rho = 0.98$



Chapter 9

A few standard distributions

In previous chapters we discussed properties of distribution, probability and (where they exist) density functions in general. In this chapter we will look at examples of specific distributions with known properties. Although we can estimate CDFs, PDFs or PMFs from data without imposing any functional form, there are a number of reasons for being interested in specific forms.

First, in some circumstances, theorems such as central limit theorems will tell us that a specific form should provide a good approximation to some quantity that interests us. A *central limit theorem* (CLT) tells us that, under some conditions, a sample mean will have a distribution that converges toward the Normal as sample size increases; see Chapter 10. As well, relations among these known distributions may allow us to deduce that if one distribution applies to a given random variable, it follows that another distribution will apply to a function of that random variable. For example, the square of a Normal random variable, standardized to zero expectation and unit variance, has a χ^2_1 distribution (that is, the Chi-squared with one degree of freedom).

In other circumstances, empirical experience may suggest that some distribution provides an adequate approximation. For example, the Pareto distribution has long been used to describe distributions of income beyond some threshold point. As well, we can sometimes use a particular distribution as a base and generalise it to embody information gained from further empirical experience. For example, the t distribution has thicker tails than the Normal (that is, a greater relative frequency of extreme events), but although the relatively thick tails of low-degree-of-freedom t distributions have been found useful in some problems involving asset return data, the symmetry of the t has turned out not to be strictly valid in important cases. Generalizations of the t to allow different upper and lower tail thickness, and other forms of asymmetry, were developed to improve modelling in such cases.

When we do have a specific functional form that provides an adequate approximation, there are a number of advantages. We can compute important quantities such as tail area probabilities or probabilities of values occurring between two points, from the known function. If the known form truly is an adequate approximation, we can take advantage of the ability to estimate the entire form of the distribution from only a few parameters (for example,

the expectation and variance fully characterize a Normal density, so if we feel confident in using a Normal to fit some data, we need only estimate the expectation and variance to estimate the full density, via the known functional form).

When unfamiliar problems arise, a specific distribution may give us a quick first-round approximation in a problem, if we can answer a few simple questions about the distribution (Continuous or discrete? Skewed or approximately symmetric? Are the data bounded to a particular region, such as the positive integers or positive reals?). That being the case, its useful to have a catalogue of potential distributions in mind. This chapter will indicate a few, but there are large references available that list many more.

9.1 Random numbers

Computers have *random-number generators*, or RNGs, and what they generate are sequences of *random numbers*. Such a sequence has most of the mathematical properties of a sequence of mutually independent realizations from the uniform distribution on the interval $[0, 1]$. See Knuth (1998), Chapter 3 for a very thorough discussion of this point. One property not shared by a sequence generated by a computer and a sequence that satisfies the mathematical requirements of realizations from the uniform distribution is that the elements of a computer-generated sequence are rational numbers that can be expressed with a finite number of bits, whereas a realisation of the uniform distribution may be any real number in the $[0, 1]$ interval. However, this and all other differences between what the computer generates and the mathematical ideal have no bad consequences for the simulations needed in econometrics.

Random numbers can be transformed into realizations from distributions other than the uniform. A valuable reference for many of these transformations is Devroye (1986). Thus we can incorporate any desired form of randomness that we can specify into a model.

The "random" numbers generated by computers are not random according to some meanings of the word. For instance, a computer can be made to spit out exactly the same sequence of supposedly random numbers more than once. In addition, a digital computer is a perfectly deterministic device. Therefore, if random means the opposite of deterministic, only computers that are not functioning properly would be capable of generating truly random numbers. Because of this, some people prefer to speak of computer-generated random numbers as **pseudo-random**. However, for the purposes of simulations, the numbers computers provide have all the properties of random numbers that we need, and so we will call them simply random rather than pseudo-random.

Computer-generated random numbers are mutually independent *drawings*, or realisations, from specific probability distributions, usually the uniform $\mathrm{U}(0, 1)$ distribution or the standard normal distribution $\mathrm{N}(0, 1)$.

For most of the distributions defined in this chapter, we will give two seemingly different definitions, or characterisations. The first is just an analytic expression for the density or the CDF. But the second is a *recipe for simulation*. Such a recipe starts with specifying some random numbers, either $\mathrm{U}(0,1)$ or $\mathrm{N}(0,1)$. These random numbers are then transformed, using an algebraic formula, and the resulting deterministic function of the random numbers is a random variable that has the distribution in question.

9.2 Discrete distributions

We will represent these distributions in the form of their probability mass functions.

9.2.1 *Uniform*

The uniform distribution may describe either discrete or continuous random variables. In the discrete case, it indicates that the probability is the same on each of the finite number of possible outcomes. If we have a uniform distribution on the integers from a to b inclusive, then there are $b-a+1$ values in total. If X is a random variable having this distribution, its probability mass function is then:

$$p_X(x) = (b-a+1)^{-1}, \quad x \in [a,b].$$

It is easy to check the condition that

$$\sum_{x=a}^{x=b} p_X(x) = 1.$$

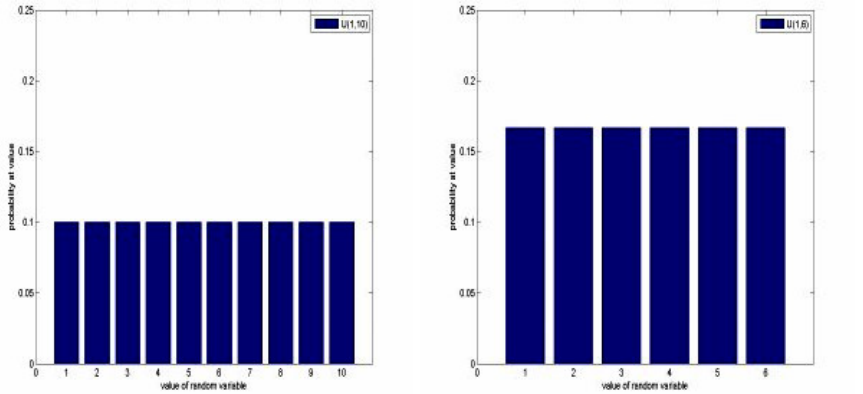
Figure 9.1.1 plots two uniform probability mass functions, over the integers from 1 to 10 and 1 to 6 respectively. Note that the sum of the probabilities is 1 in each case.

Let U denote a random number with the $\mathrm{U}(0,1)$ distribution. In order to simulate a variable with a discrete uniform distribution from a to b, define the random variable X by the formula

$$X = a + \lfloor (b-a+1)U \rfloor.$$

Here $\lfloor \cdot \rfloor$ is the *floor function*, which rounds down non-integer numbers to integer immediately less, while leaving integers unchanged. The *ceiling function*, denoted $\lceil \cdot \rceil$, is defined similarly: it rounds up non-integer numbers to the immediately greater integer. The $[0,1]$ interval can be partitioned into subintervals of the form $[i/(b-a+1), (i+1)/(b-a+1)]$, $i = 0, 1, \ldots, b-a$, and U assigns the probability $1/(b-a+1)$ to each interval. A value of U in the subinterval indexed by i leads to X taking on the value $a+i$. If $i = 0$, we get $X = a$, and if $i = b-a$, then $X = b$.

Figure 9.1.1 a/b
Discrete Uniform probability functions



The uniform distribution has an interesting and useful feature (which we will revisit below when we discuss the continuous case): for any random variable z, if z has cumulative distribution function $F(z)$, then a sample of values $F(z_i)$ will have the uniform distribution. That is, if we take sample of values $z_1, z_2, \ldots, z_n$, then this sample has the distribution $F(\cdot)$. But if we apply $F(\cdot)$ to each sample point, to find the value of the cdf at that point, then the distribution of this new set of values is uniform on [0,1]. So if we think that we know the distribution function applying to a sample of data, we can compute the sequence of values $F(z_i)$ and this sequence should be Uniform on [0,1] if we are correct in our belief about the distribution function.

9.2.2 *Binomial*

Consider a situation in which there are two possible outcomes, so that the random variable of interest X can take only the values x_1 or x_2, with corresponding probabilities p and $1-p$. For example, we might be flipping a coin, asking a survey participant which of two political parties he or she will vote for, or recording the response to a yes-or-no question. In n such instances (n outcomes of the random variable X) we can have any integer number between 0 and n of x_1's or of x_2's, where of course the numbers of each sum to n. The *binomial distribution* describes the probabilities of each of the possible overall outcomes of n samples: zero x_1's and n x_2's, one x_1 and $n-1$ x_2's, and so on.

Let ω be the number of x_1's in the sample of n realizations of X. Then the probability mass function of the $\mathrm{Bin}(n,p)$ distribution is

$$P(\omega) = \frac{n!}{\omega!(n-\omega)!} p^{\omega}(1-p)^{n-\omega}, \text{ for } \omega = 0, 1, 2, \ldots, n.$$

A special case of this, for $n = 1$, is the *Bernoulli distribution*, *i.e.*

$$P(x) = p^x(1-p)^{1-x}, \text{ for } x = 0 \text{ or } x = 1,$$

with $P(x) = 0$ otherwise. The random variable ω, which has the binomial distribution, can be thought of as the number of x_1's in n independent Bernoulli trials.

Simulating the Bernoulli distribution is spectacularly easy. We have

$$X = \begin{cases} 1 & \text{if } U < p \\ 0 & \text{if } U > p. \end{cases}$$

Note that we do not consider the case with $U = p$, because this is an event with probability zero. This leads on to a way to simulate the Binomial distribution itself. For a $\text{Bin}(n, p)$ variable, X say, generate n IID variables B_i that have the Bernoulli distribution with parameter p. Then we have

$$X = \sum_{i=1}^{n} B_i.$$

This means that X is the number of Bernoulli variables equal to 1, as required.

The probability p is sometimes described as the probability of 'success', but of course it can be the probability of either of two specific outcomes. The following figures show binomial distributions for two different numbers of trials and two different probabilities p of 'success'. For example, the first of these figures considers 12 trials with the probability of success of 50%. We can read the first of the graphs as giving us the probability that, in 12 trials, we would get exactly one success, or exactly two, or exactly three, and so on up to the maximum of twelve, where the probability is 0.5. the second graph, on the right, gives the same set of probabilities for each possible outcome, but this time where the probability of success is 0.2 each time rather than 0.5.

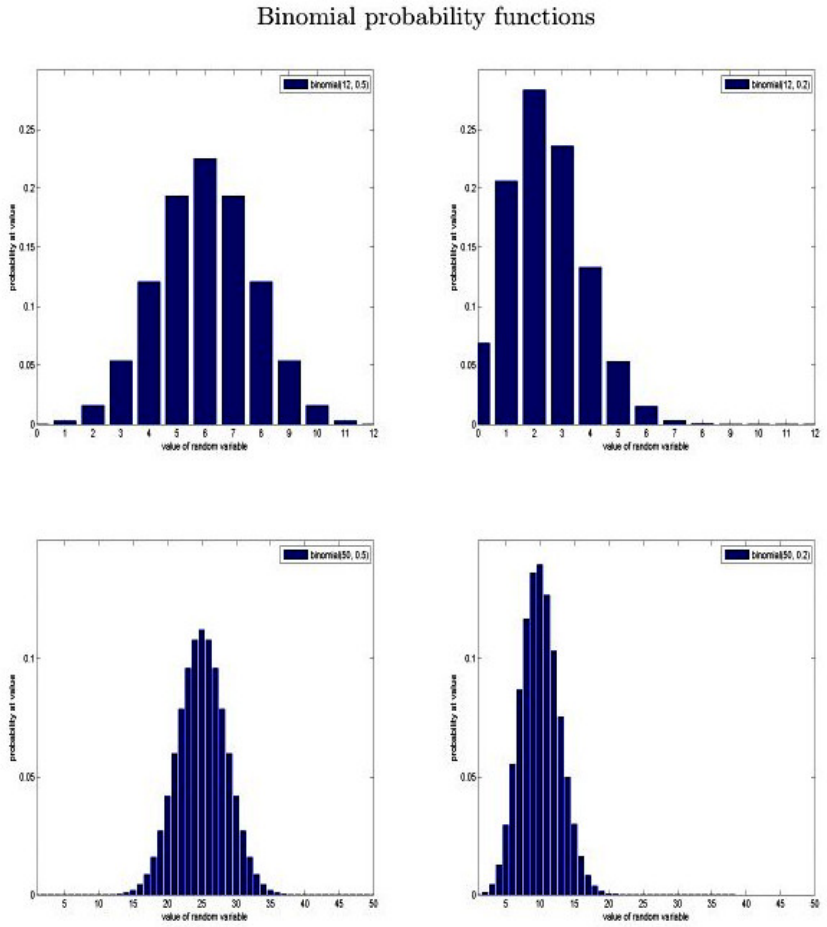
The block of graphs in Figure 9.1.2 with 50 trials, may look a bit like the Normal distribution. This is not a coincidence; the chapter on the central limit theorem below explains why the binomial distribution in fact converges to the standard Normal.

9.2.3 *Poisson*

The *Poisson distribution* is sometimes used to model data that are counts of the number of times that something occurs. It is therefore defined over the set of whole numbers $\{0, 1, 2, \ldots\}$: something that is being counted can happen either zero times, or one time, or twice, or three times, and so on. The probability mass function for the outcomes $n = 0, 1, 2, \ldots$ is defined as a function of a strictly positive parameter λ:

$$P(n) = \frac{\lambda^n}{n!}\,\mathrm{e}^{-\lambda} \quad \text{for } n = 0, 1, 2, \ldots.$$

FIGURE 9.1.2 A-D



One way in which the exponential function can be defined is by its power series expansion, as follows,

$$\mathrm{e}^x = \sum_{n=1}^{\infty} \frac{x^n}{n!}.$$

Then we can check that the probabilities sum to one:

$$\sum_{n=0}^{\infty} P(n) = \mathrm{e}^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = \mathrm{e}^{-\lambda}\mathrm{e}^{\lambda} = 1.$$

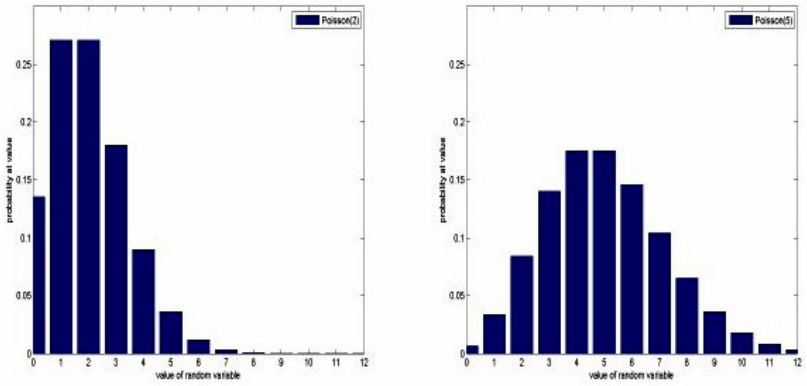
The parameter λ is called the *intensity* of the distribution and it is also the expectation, as we can check:

$$\sum_{n=0}^{\infty} n\frac{\lambda^n}{n!}\,\mathrm{e}^{-\lambda} = \mathrm{e}^{-\lambda}\sum_{n=1}^{\infty}\frac{\lambda^n}{(n-1)!} = \lambda\mathrm{e}^{-\lambda}\sum_{m=0}^{\infty}\frac{\lambda^m}{m!} = \lambda.$$

Figure 9.1.3 shows probability functions for $\lambda = 2$ and 5.

FIGURE 9.1.3 A/B

Poisson probability functions



9.3 CONTINUOUS DISTRIBUTIONS

Continuous distributions will be depicted using their probability density functions. It is relatively easy to show several of these densities at the same time, so we will now combine several parameter values or sets of parameter values into a single graph.

9.3.1 *Uniform distributions*

As we said earlier, the Uniform also has a continuous form. Let a random variable X be uniform on the continuous interval $[a, b]$, which we write $\mathrm{U}(a, b)$; then its density is

$$f_X(x) = \begin{cases} (b-a)^{-1}, & x \in [a,b] \\ 0 & \text{otherwise.} \end{cases}$$

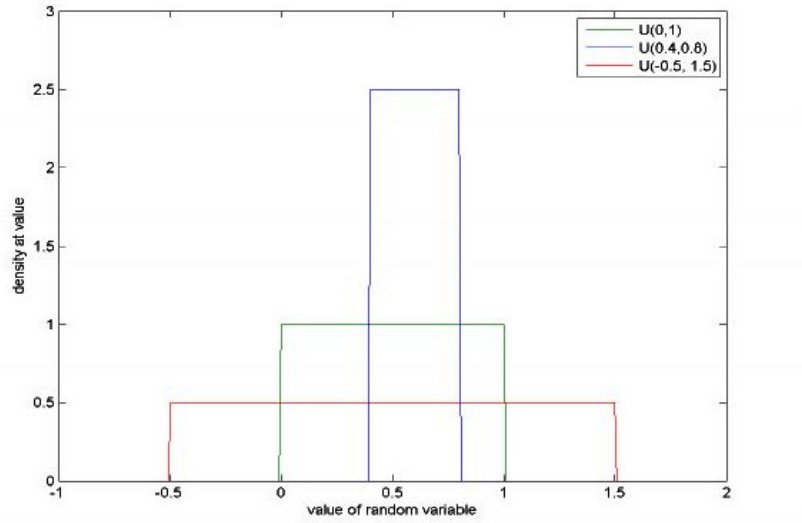
Simulation of $X \sim \mathrm{U}(a, b)$ is easy: with $U \sim \mathrm{U}(0, 1)$, we have

$$X = a + (b - a)U.$$

If Z is any continuous random variable with cumulative distribution function $F(\cdot)$, we again have that the cdf of $F(Z)$ is U(0,1).

FIGURE 9.2.1

Continuous uniform densities



Notice that in Figure 9.2.1 it is easy to calculate that the area under any one of these densities is exactly one. This is of course true of the other continuous densities as well, although this may be harder to see in the other examples that follow.

9.3.2 *Normal distributions*

The Normal distribution is extremely widely used, partly for good reason, partly for not-so-good reasons.

The good reason is that central limit theory, to be discussed in a later chapter, tells us that under quite a wide range of circumstances, a sum or average of random variables will have a distribution that converges on the Normal. So when one is working with sums or averages, it is often sensible to use the Normal to approximate their distribution.

However, many data are generated by processes that are not summed or averaged, and the Normal distribution is sometimes lazily applied to data whose distribution is of unknown form. A moment's reflection will tell us that many data series will be skewed or bounded on one side, for example, and so cannot strictly be Normal. As well, the Normal distribution has very thin tails: that is, the relative frequency of extreme events is quite low relative to what arises in many other distributions. The Normal is inadequate to characterize the proportion of extreme events that arise in many random processes. This is something to be particularly wary of in financial data,

where Normal distributions will often provide very poor fits to details of data, and if they are used inappropriately the risk of extreme events can be severely underestimated.

We have already defined the Normal density; again, for $X \sim \mathrm{N}(\mu, \sigma^2)$, it is

$$f_X(x) = (2\pi\sigma^2)^{-1/2} \exp\Big[-\frac{(x-\mu)^2}{2\sigma^2}\Big],$$

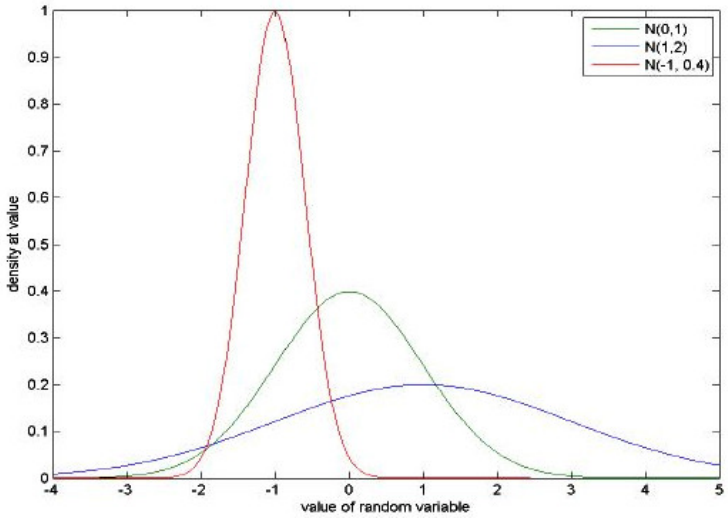
a function of the expectation and standard deviation parameters, μ and $\sigma > 0$. For simulation, let $Z \sim \mathrm{N}(0,1)$. Then

$$X = \mu + \sigma Z.$$

Notice that different standard deviation (scale) parameters σ change the appearance of the distribution substantially; some of these Normal distributions look less like the stereotypical ('bell'-shaped) Normal than do the t-distributions in the following section. Appearances can be deceiving.

FIGURE 9.2.2

Normal densities



9.3.3 *Chi-squared (χ^2) distributions*

The χ^2 distribution arises in testing problems. Sums of independent squared standard Normal random variables have χ^2 distributions, with degrees of freedom equal to the number of independent squared standard Normal random variables that are added. Because standard Normal random

variables can often be shown to arise asymptotically through the application of a central limit theorem to a sum or average, it follows that taking an inner product of a vector of such variables with itself (*i.e.* taking the sum of the squared underlying Normal random variables) will lead to a χ^2 distribution if the variables being added are independent. The very wide applicability of central limit theorems that imply asymptotic normality therefore implies a correspondingly wide applicability of joint tests combining several of these variables, yielding χ^2 distributions.

The χ^2 distribution is characterised by a *degrees-of-freedom* parameter, ν : its density is

$$f_X(x) = \frac{x^{(\nu-2)/2} \mathrm{e}^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}$$

where $\Gamma(k)$ is the Gamma function,

$$\Gamma(k) = \int_0^\infty \mathrm{e}^{-s} s^{k-1}\, \mathrm{d}s.$$

The Gamma function is continuous and is defined on any real value except 0 and the negative integers, but for positive integer values of k it yields the factorial function: that is $\Gamma(k) = (k-1)\,!$.

A χ^2 variable Y with ν degrees of freedom can be simulated using ν IID standard Normal variables Z_i, $i = 1, \ldots, \nu$:

$$Y = \sum_{i=1}^{\nu} Z_i^2.$$

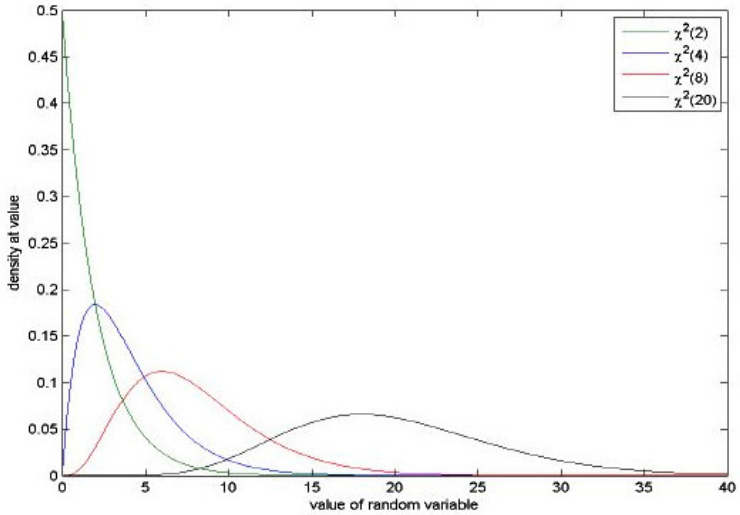
The expectation of a χ^2 distribution is its degrees-of-freedom parameter ν, and the variance is 2ν. Visually, then, χ^2 distributions move to the right as we consider higher degrees of freedom. The χ^2 densities with one or two degrees of freedom are monotonically declining; with three or more d.f., the densities have an interior maximum.

9.3.4 *(Student's) t distributions*

The t-distribution also arises in testing problems, where a standard Normal random variable is divided by an estimated standard error, which under some conditions will have a χ^2 distribution. While the conditions required for an exact t-distribution to arise in these testing problems are often not met, the t is nonetheless widely used, in part because it provides a slightly more conservative test then does the use of the asymptotic Normal which would be justified under a wider set of circumstances for the same testing problems. The t-distribution has a degrees-of-freedom parameter, and as that parameter increases without bound, the t-distribution approaches the standard Normal. This is sometimes written as $t_\nu \xrightarrow{D} \mathrm{N}(0,1)$ as $\nu \to \infty$.

FIGURE 9.2.4

χ^2 densities



The t may be difficult to distinguish visually from the Normal. However, as we have seen elsewhere, the proportion of its probability mass beyond a certain distance from the origin differs from the same proportion in the standard Normal to an ever greater degree as one moves farther from the origin: that is, the distribution appears to be a very good approximation in the central region, but (for any given degrees of freedom parameter) the farther out in the tails we go, the poorer is the approximation that the standard Normal provides to the t.

The t is also sometimes used to model data where extreme events arise more frequently than could be accounted for in the standard Normal; for example, low-degrees-of-freedom t-distributions are sometimes used in financial econometrics to describe data series that have a higher relative frequency of extreme events than would arise in a Normal distribution.

Its density is:

$$f_X(x) = (\nu\pi)^{-1/2}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2},$$

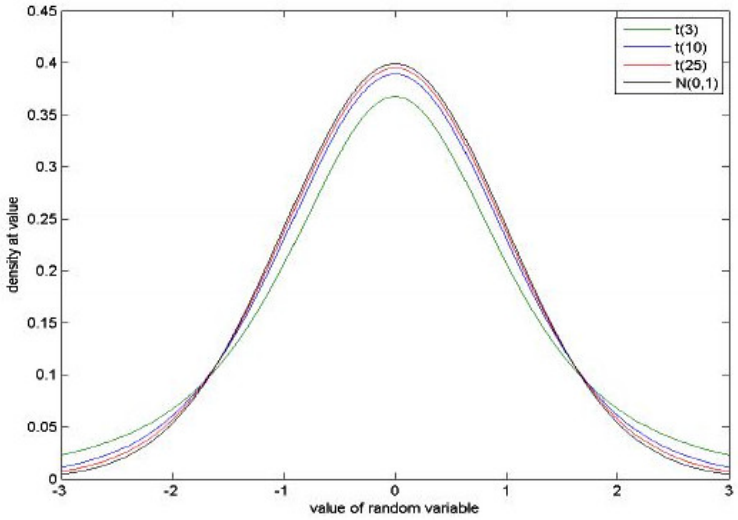
where again $\Gamma(k)$ is the Gamma function defined above.

For simulation, start from a standard Normal variable Z and a χ^2 variable Y with ν degrees of freedom independent of Z. Then if T has the t distribution with ν degrees of freedom, we have

$$T = Z/\sqrt{Y/\nu}.$$

FIGURE 9.2.3

$t-$ densities, varying degrees of freedom, compared with N(0,1)



9.3.5 *F distributions*

F distributions characterise ratios of independent χ^2 distributions divided by their degrees of freedom. Again, therefore, this distribution is one that arises in testing problems. Because the χ^2 distributions are divided by their degrees of freedom, however, the expectations of numerator and denominator are set to one. Of course, this does not imply that the expectation of an F distribution is equal to one, because the ratio is a nonlinear transformation, and the expectation of the ratio is not in general equal to the ratio of the expectations. However, the expectation of an F distribution is close to one.

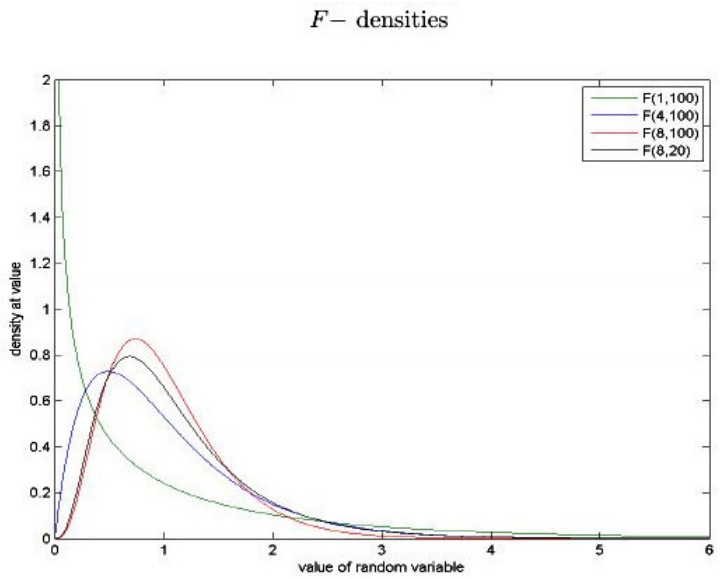
The F distribution has two degrees-of-freedom parameters, one corresponding with the numerator and one with the denominator, in a ratio of χ^2 distributions. The density has the form:

$$f_X(x) = \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}\frac{\Gamma\big((\nu_1+\nu_2)/2\big)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\left[1+\left(\frac{\nu_1}{\nu_2}\right)x\right]^{-(\nu_1+\nu_2)/2}x^{(\nu_1-2)/2}.$$

For simulation of a variable $X \sim F_{\nu_1,\nu_2}$, start from two independent χ^2 variables Y_1 and Y_2 with ν_1 and ν_2 degrees of freedom respectively. Then

$$X = (Y_1/\nu_1)/(Y_2/\nu_2).$$

FIGURE 9.2.5



9.3.6 *Exponential distributions*

The exponential distribution is often used as a model for continuous, positive quantities such as times: for example, completion times or waiting times. Its CDF has the simple form

$$F_X(x) = 1 - \mathrm{e}^{-x/\mu}, \quad x \geq 0.$$

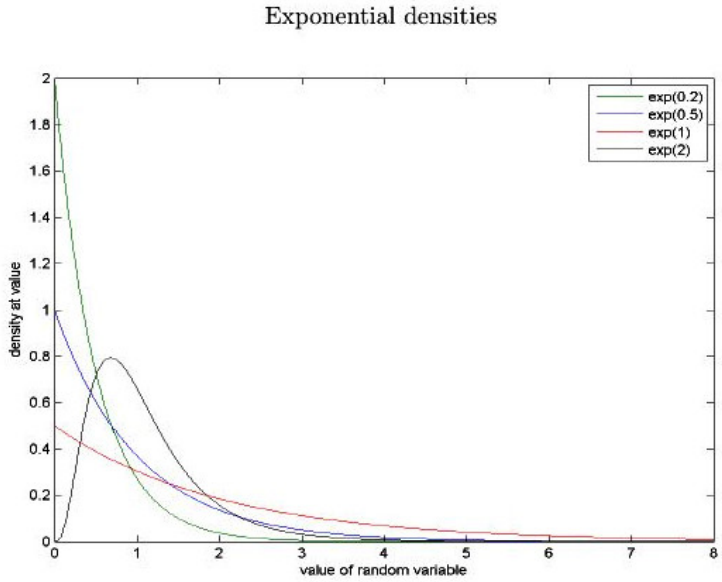
a function of the single positive parameter μ. The density $f_X(x)$ is the derivative of the CDF:

$$f_X(x) = \mu^{-1}\mathrm{e}^{-x/\mu}.$$

If we set $\mu = 1$ and note that a density must integrate to 1, it follows that

$$\int_0^\infty \mathrm{e}^{-x}\,\mathrm{d}x = 1.$$

FIGURE 9.2.6



The expectation can be calculated as follows:

$$\mu^{-1}\int_0^\infty x\,\mathrm{e}^{-x/\mu}\,\mathrm{d}x = \mu\int_0^\infty y\,\mathrm{e}^{-y}\,\mathrm{d}y = \mu,$$

where the changes of variables is $y = x/\mu$, and integration by parts shows that

$$\int_0^\infty y\,\mathrm{e}^{-y}\,\mathrm{d}y = -\big[y\,\mathrm{e}^{-y}\big]_0^\infty + \int_0^\infty \mathrm{e}^{-y}\,\mathrm{d}y = 1.$$

An unusual link between the $\mathrm{U}(0,1)$ distribution and the exponential distribution with expectation 1 leads to a simple way to simulate the latter. Consider the variable $X = -\log(U)$ where $U \sim \mathrm{U}(0,1)$. Its CDF is

$$F_X(x) = P\big(-\log(U) \leq x\big) = P\big(\log(U) \geq -x\big) = P\big(U \geq \mathrm{e}^{-x}\big) = 1 - \mathrm{e}^{-x},$$

which shows that the distribution of $-\log(U)$ is the exponential with $\mu = 1$. Thus to simulate the exponential distribution, the recipe is $X = -\log(U)$, and, for $\mu \neq 1$,

$$X = -\mu\log(U).$$

9.3.7 *Relation between exponential and Poisson distributions*

9.3.8 *Lognormal distributions*

If a random variable has a lognormal distribution with parameters μ and σ, then the logarithm of that random variable has a Normal distribution with the same parameters. It is useful to know the form of this distribution because we often are in the position of taking the logarithm of random variables, some of which may be approximately Normal, so that the lognormal can provide a good approximation to the distribution of the result. The lognormal is also often used as a general model for the distribution of skewed data, where the exact form of the distribution is unknown.

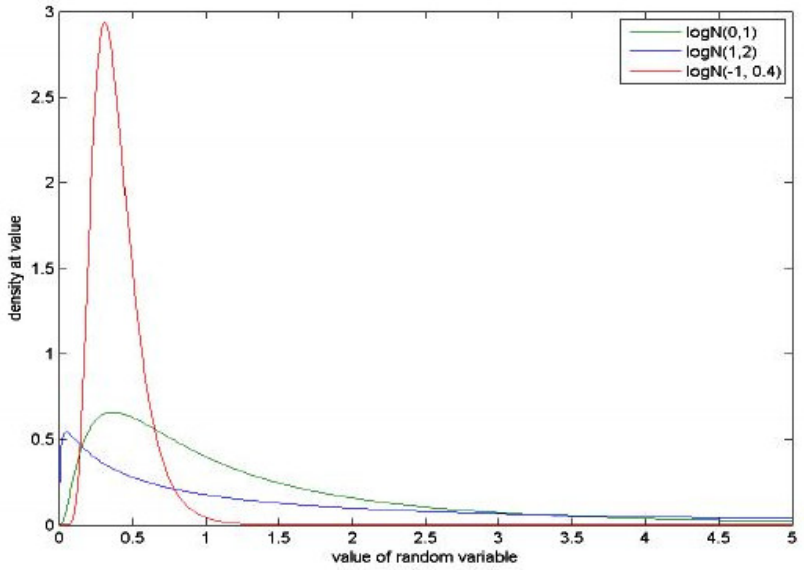
Its density has a form reminiscent of the Normal:

$$f_X(x) = (2\pi\sigma^2)^{-1/2}(x\sigma)^{-1}\exp\Big[-\frac{(\log x - \mu)^2}{2\sigma^2}\Big].$$

If X has the lognormal distribution with $\log X \sim \mathrm{N}(\mu, \sigma^2)$ and $Z \sim \mathrm{N}(0,1)$, X can be simulated as

$$X = \exp(\mu + \sigma Z).$$

FIGURE 9.2.7
lognormal densities



PART III: STATISTICAL METHODS

Chapter 10

Introduction to sampling and sampling distributions

We look at samples of data in order to learn about something, usually about something more than the sample itself. Typically we are hoping to find out about a set with many members, such that it is impossible to look at every member of the set. For example, conservation officials interested in the question of whether a license should be required to fish on a certain lake might investigate whether fishing is reducing the average size of fish by taking away relatively more of the mature fish (perhaps of a specific species). They might catch, weigh and return to the water a certain number of fish, and repeat the exercise every year for some years. (In order to judge whether any differences they see over time are genuine, as opposed to being simply the result of random differences in which fish they catch, they would want to apply methods for statistical inference described in chapters below.) In this case, it is clear that they are intending to learn about the entire population of fish in the lake from their sample of fish. A pollster who asks a sample of people how they intend to vote in the coming election is hoping to be able to predict the outcome of the election, which requires learning about the voting intentions of those who will actually vote. Note that the intentions of people who are not going to vote are irrelevant to the outcome, and so are not of concern to the pollster: the population of interest is those who are actually going to cast a vote, so that even in this case there is some subtlety to the question of what the relevant population is.

Sometimes it is less clear what population we can learn about. For example, we might survey a group of workers in a particular company in Toronto to determine whether those who undertook a training program benefited through relatively higher wages, promotions, and so on. But who is it are we learning about? If the results apply only to these workers and to this company, then they might not be of much interest to anyone else, and we might even be able to speak with every employee at the company if it is small enough, so that sampling would not be necessary. Would the results apply to any worker, anywhere, who undertakes training? This seems unlikely, given the diversity of the workforce and the conditions of work around the world. We might conclude, however, that the results could provide a good indicator of the likely benefits of a particular type of training program for North American

workers in a certain kind of industry (so that this is the population being studied), for as long as certain general conditions remain in place. In any event, determining what population we are learning about requires some thought.

Once we are clear about what the population being studied is, we want to know how the quantities to be obtained from the sample are related to the true characteristics of the population that are of interest to us. They will not of course be identical in general, but we hope that they will be close and will tend to get closer as we take larger and larger samples. The purpose of this and the following chapters is to characterize what is known about the relation between sample quantities and population quantities: that is, what is the distribution of a sample quantity relative to a population ('true') quantity?

10.1 Sample and population

D10.1 Sample: A sample is a subset of a population that can be observed by an investigator.

The aim in sampling is to obtain a sample which is representative of the population. For example, we might be interested in how the vote will go in the upcoming (at the time of writing) referendum on Scottish independence. If we sample the population in Scotland by setting up a booth on campus at the University of Edinburgh, then almost everyone we asked will either be a university student or have a degree, will have above average (expected lifetime) income, and so on. We will not be learning the views of the poor, chronically unemployed, rural, or elderly voters. Unless university students and staff happen by coincidence to have the same distribution of views as the general population, we will get a misleading view of overall voting intentions.

Normally we would prefer a 'random' sample.

D10.2 Simple random sample: A simple random sample is a sample from a population such that every member of the population is equally likely to be chosen for the sample, and successive observations in the sample are independent.

Note that this definition of 'random' is somewhat different from what might be used in other contexts in statistics; for example a random stochastic process is one which is not fully predictable, but may have some predictable part.

In some cases, it is difficult to achieve the goal of a random sample, because some members of the population are less likely to be observed than others. Researchers may therefore sometimes use a 'stratified sample'. A stratified sample is one in which the population is divided into mutually exclusive and exhaustive classes, and the final sample is designed to have the same proportion of each class as does the population. Simple random sampling may be used within each class, with the goal of obtaining an overall sample which is representative of the population. For example, we may have 8% of a particular population which is elderly (let's say, 70 or over). If we try

to sample randomly from the population, however, we may find that we are getting in touch with elderly people less often, either because they are less likely to answer the phone, or come to the door, or be contacted by whatever other means we are using; perhaps we would only end up with 3% of our sample being elderly people; if their behavior patterns are different in a way that is relevant to what we are investigating, our results could therefore be misleading. We might therefore continue sampling only elderly people until we have enough to make up 8% of the overall sample. The goal remains to obtain a sample which is representative of the entire population.

When we have a sample, we will want to use it to learn about some characteristic of the population. Often, we will start by estimating the mean of some characteristic in the population, for example, the mean weight of a fish in the lake. But we know that the mean weight in our sample will not, except by outrageous coincidence, be the same to the nearest gram as in the population. What then does the sample mean tell us about the population mean? We can hope that they are close, but that is not very useful. In order to answer the question well, we would like to be able to characterize the entire distribution of the sample mean, given some population mean and size of sample. If the mean weight of a trout in the lake is 746 grams, and if we catch, measure, and release a sample of 100 trout, what is the distribution of possible sample means that we could find? We can answer this question, at least approximately (and with an approximation error that declines to zero as sample size increases) in a very wide variety of cases.

10.2 Sampling and distributions of samples

We can begin with a simple example that we have seen earlier, in which we can compute the exact distribution of the sample mean, to help us understand what we are trying to obtain and how to interpret it.

Consider a simple game played by two people. **A** flips a fair coin (the probability of a head = the probability a tail = 0.5) and pays \$1 to **B** if the coin comes up heads, and receives \$1 from **B** if the coin comes up tails. Clearly, each person is a symmetric position, and has the same probability of being a winner, loser, or breaking even after playing n times. The population mean payoff to each player is -1(0.5) + 1(0.5) = 0, regardless of the number of times a game is played.

The sample mean – the average of what is won or lost – may of course differ. If they play three times, **A**'s possible outcomes are $\{-3, -1, 1, 3\}$ and the sample mean outcomes are $\{-1, -1/3, 1/3, 1\}$, and the same is of course true for **B**. These outcomes are not equally likely, of course, and we have seen that the probabilities can be computed in various ways. The probabilities of the four outcomes are 1/8, 3/8, 3/8, 1/8. Notice that because the number of rounds is odd, it's actually impossible to break even exactly, so it's impossible for the sample mean to equal the population mean in this case. Nonetheless,

although zero is not a possible outcome, the mean of the sampling distribution is zero, just as the mean of the population distribution is zero.

This is the sampling distribution of the mean payoff after playing the game three times, and it fully describes what outcomes could emerge for the mean and what their probabilities are, given the conditions of the game.

If we were to repeat this exercise for a game of, say, 10 rounds, the distribution would be different although still centered on zero. With 10 rounds, the probabilities would be more heavily clustered near zero, and we could compute them exactly using the binomial distribution.[1] As we have seen in earlier chapters, with the distribution we can answer questions such as this: if they play the game 10 times, what is the probability that **A** has a mean loss of greater than 0.25, *i.e.* a total loss of more than \$2.50? (It's the sum of the first four probabilities, or 176/1024.) The sampling distribution allows us to make statements about the probability of the sample mean lying in different regions, given the population mean. Conversely, given an observed sample mean, it allows us to make statements about where the true population mean is likely to lie, in the more usual case where the population mean is not known.

Now let's do a much larger exercise, using a computer simulation. We use a computer random-number generator to generate pseudo-random variables from either the Uniform[0,2] distribution or from the Chi-squared distribution with 1 degree of freedom. Both of these distributions have a mean of 1, so the population mean in both of these experiments is 1.

In each case we take samples of size n from the distribution, and take the mean of each sample. We do this 100,000 times for each sample size, so that we have many examples of sample means, and then we can actually estimate the distribution that applies to the sample mean. We could use a kernel density estimator (which is at present not described in this book, but can be thought of for now as a development of the idea of the histogram, producing a smooth curve instead of a set of bars). See Silverman (1986) for an exposition. There are three sample sizes, so that we can observe something about the way in which the sampling distribution changes as the sample size changes.

Figure 10.2.1 shows results from Uniform random variables. The Uniform distribution is symmetric, and the sampling distributions are apparently symmetric as well. In the case of the input data which are χ^2_1 , shown in Figure 10.2.2, the sampling distribution for $n = 10$ is noticeably skewed; in fact

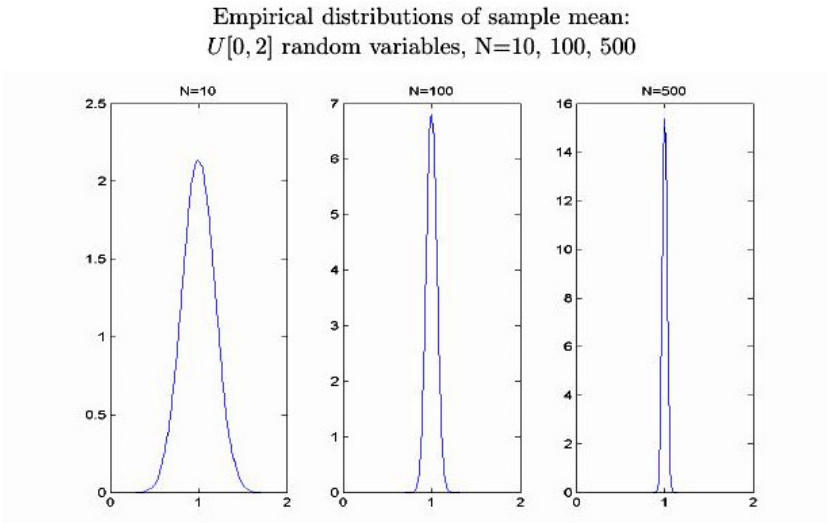
[1] The possible outcomes are $\{-10, -8, -6, -4, -2, 0, 2, 4, 6, 8, 10\}$ with corresponding means $\{-1, -0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8, 1\}$. The probabilities are

$$\left(\frac{1}{1024}\right)\{1, 10, 45, 120, 210, 252, 210, 120, 45, 10, 1\}.$$

if one looks very closely (compare the heights of the density function around 0.5 and 1.5), a tiny degree of skewness is visible at $n = 100$ as well. In the largest sample size, no skewness is visible to the naked eye.

Notice that the vertical scales are different: all of these density functions integrate to 1, so that as they become thinner they must become taller as well: that is, they become more tightly concentrated around the population mean.

FIGURE 10.2.1

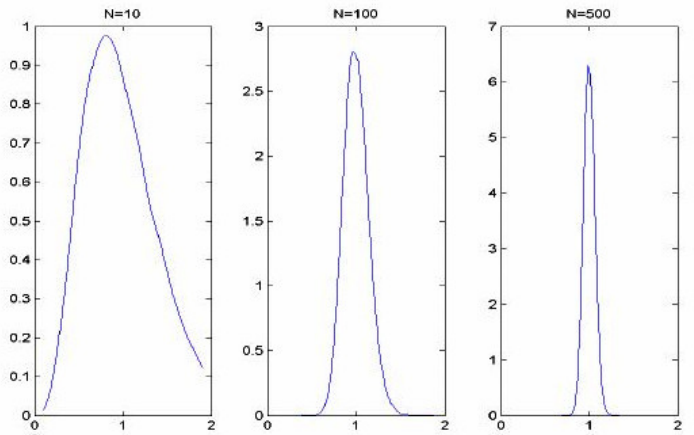


These results suggest that the distribution of the sample mean tends to become ever more highly concentrated around the true value as the number of sample points increases, and also that the distribution of the sample values around the true value tends toward a single peaked (unimodal) and symmetric distribution as the sample size increases. Both of these results are borne out by theory, as we shall see later in Chapter 11

10.3 A simple, if unrealistic, case

A simple case in which we can work out the exact distribution of the sample mean is that in which the data actually come from a Normal distribution. Typically, however, we will observe some feature of the data that makes it impossible that the data could truly be Normal. For example, the data may be bounded on one or both sides (the unemployment rate, or the proportion of survey respondents who say they'll vote for a particular party, cannot go below zero or above 100%). Alternatively, a simple plot of the histogram of the data set may show substantial skewness. Nonetheless its useful to start by learning about this case, for several reasons:

FIGURE 10.2.2



- The sampling distribution that emerges in more realistic cases, where the data distribution is unknown, will turn out to be approximately the same as the distribution that results in this case.
- The Normal-data case will help us to understand the reasons for the use of the t distribution in some problems.
- We will gain some understanding, through this and results in the chapter covering Central Limit Theorems, of the distinction between exact finite-sample results and asymptotic results.
- The Normal-data case has a direct application in some circumstances, particularly in computer simulations where the input data are created to have a particular distribution.

In order to obtain the sampling distribution of the mean from a Normal population of data, we need the following result.

Theorem 10.1: Linear combinations of multivariate Normal random variables are Normal. Let $Z_1, Z_2, \ldots, Z_n$ be independent Normal variables each of which has expectation 0, and variance of Z_i is σ_i^2, $i = 1, 2 \ldots, n$. Then the linear combination $a_1Z_1+a_2Z_2+\ldots+a_nZ_n$ has the distribution $\mathrm{N}(0, \sum_{i=1}^n a_i^2\sigma_i^2)$.

Proof: See Kendall *et al.* (1991), example 11.2. ■

(If the expectations of the random variables are non-zero, then the expectation of the Normal distribution applying to the linear combination is simply the weighted sum of the expectations, $\sum_{i=1}^n a_i\mu_i$.)

To apply this result to obtain the distribution of the sample mean, note that the sample mean is a linear combination of the sample data, $\overline{X}_n = \sum_{i=1}^n X_i/n$, with the weight on each data point being the constant $1/n$.

Here we are treating the case in which the data are independent samples from an $\mathrm{N}(\mu, \sigma^2)$ distribution. The expectation of the sample mean is

$$\mathrm{E}\Big(\sum_{i=1}^n \frac{1}{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^n \mathrm{E}(X_i) = \frac{1}{n}(n\mu) = \mu.$$

So the expectation of $\overline{X}_n$ is the same as the expectation of the sample data that are being averaged, the X_i's. This is not true for the variance; the variance of the sample mean in this independent sampling case is smaller than the variance of the data: using our earlier results on the variance of a linear combination,

$$\mathrm{Var}\Big(\sum_{i=1}^n \frac{1}{n} X_i\Big) = \mathrm{Var}\Big(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \ldots + \frac{1}{n}X_n\Big)$$

$$= \Big(\sum_{i=1}^n \frac{1}{n^2}\,\mathrm{Var}(X_i)\Big) = \frac{1}{n^2}\Big(\sum_{i=1}^n \sigma^2\Big) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

There are several important points to note and remember about this.

- The variance declines with sample size. That is, as we get more sample points our estimator has less dispersion, and we have a better and better idea of where the true value lies. This is reflected in the graphs above, where we see the densities becoming more tightly concentrated around the true value as the sample size increases.
- The computation of the variance is very straightforward in this case because there are no covariance terms: we have assumed that we have an independent sample. If the data were correlated, additional terms appear in the computation of the variance, and it would be larger than σ^2/n; however, as long as the correlation between subsequent observations is not perfect, the variance of the sample mean will still decline as sample information accumulates.
- Putting together the expectation and variance of the distribution of the sample mean with the fact from the theorem that it must have a Normal form, we obtain the result in this case that $\overline{X}_n \sim \mathrm{N}(\mu, \sigma^2/n)$.
- We can standardize the sample mean to obtain a distribution which does not change with the sample size: subtracting the mean and dividing by the square root of the variance, we find $(\overline{X}_n - \mu)/(\sigma/\sqrt{n}) \sim \mathrm{N}(0,1)$, the standard Normal distribution.

- By multiplying numerator and denominator by the square root of the sample size n in the distribution just given, the result may be rewritten as $\sqrt{n}\big((\overline{X}_n - \mu)/\sigma\big) \sim \mathrm{N}(0,1)$. This implies that scaling up the discrepancy in the estimate of the expectation by the square root of the sample size leads to a fixed, non-degenerate distribution. It follows therefore that the discrepancy itself is declining at the rate of the square root of sample size. This is an example of 'root-n' convergence, which appears in many standard parametric problems.

Thus the discrepancy between the sample (estimated) and population ('true') means, divided by the standard deviation of the sample mean (we might say: the discrepancy 'measured in standard deviations') has a standard Normal distribution. Note that we write standard deviation rather than standard error, because we are referring to the population value, σ.

This is what is sometimes called an 'infeasible' or 'non-operational' statistic. Not everything on the left-hand side of the expression is observable: we dont know σ. Because we usually don't have this value, we can't actually compute this statistic.

In practice, we have to replace σ with s, that is, we replace the standard deviation with the standard error (or sample standard deviation) of the data. Does this change the sampling distribution?

Given the sampling conditions assumed, s converges in probability to σ in a sense that we will define precisely in the next chapter. So in large samples,

$$\frac{\overline{X}_n - \mu}{SE(\overline{X}_n)} \text{ or } \frac{\overline{X}_n - \mu}{s/\sqrt{n}} \text{ should be very close to } \frac{\overline{X}_n - \mu}{SD(\overline{X}_n)} \text{ or } \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}},$$

and so the former should have a distribution close to N(0,1). This turns out to be true.

But in fact, for this case, the exact distribution that applies for any given sample size n (not just the asymptotic result) has been worked out, and we do not need to use an approximation. This was done in 1908, by William S. Gosset, 1876-1937. Because Gosset used the pseudonym Student, the distribution is often called Student's t distribution.

Theorem 10.2: t distribution. Let Z be a random variable with the standard Normal distribution $\big((Z \sim \mathrm{N}(0,1)\big)$ and let W the random variable with the Chi-squared distribution with r degrees of freedom ($W \sim \chi^2_r$). Then if Z and W are independent, the ratio

$$\frac{Z}{\sqrt{W/r}}$$

has the Student's t-distribution with r degrees of freedom.

Proof: See Mood *et al.* (1974), section 4.5. ■

We will see below that in this case of independent random sampling from Normal data, the sample variance s^2 in fact has a χ^2_{n-1} distribution, so that

the feasible statistic $(\overline{X}_n - \mu)/(s/\sqrt{n})$ is distributed as t_{n-1}. Note again that this result, which gives an exact distribution applicable to any particular sample size, has been obtained under the generally unrealistic assumption that the data that we are sampling themselves have a Normal distribution. In more general circumstances where we do not know this, we will have to rely on an asymptotic approximation to get the distribution of this feasible statistic, as described in the next chapter.

10.4 Using a sampling distribution

Consider then that we have a sample of size n from a population with expectation μ and variance σ^2. What can we deduce from this?

If we take a given population mean, we can answer questions about where the sample mean is likely to be – what is the probability that it lies in a certain interval, for example, or the probability that it lies more than a certain distance away from the population mean. If we determine how tightly concentrated the distribution of the sample mean is around the true expectation for a given sample size, it will be useful in determining what sample size we need to use to get a given degree of precision. In practical sampling problems where we have a sample already, we are interested in the converse: given our estimate of the mean, $\overline{X}_n$, what is the probability that the true expectation lies in a certain interval?

To answer these questions, let's manipulate the expressions above, working with the feasible or operational form of the statistic:

$$\frac{\overline{X}_n - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$

Since the t distribution with $n-1$ degrees of freedom has a known form, and the quantiles and so on have been tabulated, we can compute an interval so that the expression on the left-hand side above has a given probability of lying in that interval. Using the notation q_α for the α-quantile of the relevant distribution (t_{n-1}), we can define the interval such that

$$P\Big(q_{\alpha/2} < \frac{\overline{X}_n - \mu}{s/\sqrt{n}} < q_{1-\alpha/2}\Big) = 1 - \alpha, \tag{10.4.1}$$

where we have used $\alpha/2$ in each case so that we have a probability $\alpha/2$ of the true value of μ lying outside this interval both above and below the interval, adding up to a total probability of α outside the interval, and $1-\alpha$ inside the interval. Often, α is taken to be 5% (0.05), so that the interval spans the interior 95% of the distribution, leaving 2.5% on both the left and right tails.

When working with the Normal distribution, either to describe the infeasible case or in using the approximation from asymptotic theory that we will learn in the next chapter, it is common to use the notation $z_{\alpha/2}$ to describe

the corresponding quantiles from the Normal. This notation is also sometimes used for the t distribution.

Let us now manipulate this expression further. The expression (10.4.1) above contains $\overline{X}_n - \mu$ in the middle: if we know one of these, we will be able to obtain a statement about the other.

If we perform the same operation on each of the quantities in parentheses, we will not change the probability, and so multiplying through by the denominator of the expression in the middle, we can obtain

$$P\Big(q_{\alpha/2}(s/\sqrt{n}) < \overline{X}_n - \mu < q_{1-\alpha/2}(s/\sqrt{n})\Big) = 1 - \alpha.$$

Because the t distribution is symmetric about the origin, we see that $q_{\alpha/2} = -q_{1-\alpha/2}$. If we use this result and subtract $\overline{X}_n$, from each of the three terms, we obtain a statement purely about μ:

$$P\big(-\overline{X}_n - q_{1-\alpha/2}(s/\sqrt{n}) < -\mu < -\overline{X}_n + q_{1-\alpha/2}(s/\sqrt{n})\Big) = 1 - \alpha,$$

and then if we multiply through by -1 (the inequality signs must then be reversed: for example $5 > 4 > 3$ implies that $-5 < -4 < -3$),

$$P\Big(\overline{X}_n + q_{1-\alpha/2}(s/\sqrt{n}) > \mu > \overline{X}_n - q_{1-\alpha/2}(s/\sqrt{n})\Big) = 1 - \alpha.$$

This expression says that the population mean μ lies in the interval $\overline{X}_n \pm q_{1-\alpha/2}(s/\sqrt{n})$ with probability α. Thus we have succeeded in obtaining a probability statement about where the population mean lies, although we only observe the sample. Notice that as n gets larger, this interval gets narrower: more information produces a more precise statement.

For the t distribution with a large number of degrees of freedom $n-1$ (or for the standard Normal distribution), $q_{1-\alpha/2}$ (or $z_{1-\alpha/2}$ in the notation commonly used for the standard Normal) is approximately 1.96: that is, about 2.5% of the distribution lies below -1.96, and about 2.5% of the distribution lies above 1.96.[2] Thus the probability interval just stated is what lies behind the commonly-remembered result that there is a 95% probability that the population mean of something will lie within about $\pm$ two standard errors of the sample mean.

To take a numerical example, consider the population of fish in the lake mentioned earlier. We catch 100 fish, and find an average weight of 746 g, with

[2] The value of $q_{\alpha/2}$ can be obtained from tables for particular values of the degrees of freedom, or from a computer program that computes the inverse of the cumulative distribution function: that is, given a value of the CDF such as 0.99, a program will calculate the corresponding quantile which gives $\text{CDF}(q_{\alpha/2}) = 0.99$.

a standard error of 205 g. As usual in real data, a moment's reflection tells us that these data could not literally be Normal: weight cannot be negative, so the distribution is bounded below, unlike the Normal. As we said, knowing that the data are Normal is generally unrealistic. Let's go on with this example anyway, because it will turn out below that the results that we have just stated will turn out to be a good approximation in a wide range of circumstances even though the data are not Normal.

So using the intervals given above, and using $q_{0.025} = 1.96$, we have

$$P\big(746 + 1.96(205/\sqrt{100}) > \mu > 746 - 1.96(205/\sqrt{100})\big) = 0.95$$

or since $1.96(20.5) = 40.18$,

$$P(786.18 > \mu > 705.82) = 0.95.$$

So, given the conditions assumed to hold in this sampling experiment, we can be 95% sure that the mean weight of the fish in the lake is between about 706 g and 787 g (rounding to three significant digits).

10.5 Simple case continued: distribution of the sample variance

The sample variance is defined as follows:

$$s^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2, \tag{10.5.1}$$

where $\overline{X}_n = n^{-1}\sum_{i-1}^{n} X_i$. Note that this is *not* the variance of the sampling distribution. We can prove the following Theorem.

Theorem 10.3 Suppose that the data are Normally distributed, the observations being mutually independent and Normally distributed, each with expectation μ and variance σ^2. Then $(n-1)s^2/\sigma^2$ has a χ^2 distribution with $n-1$ degrees of freedom.

Proof: Let $W_i = (X_i - \mu)/\sigma,\ i = 1, \ldots, n$. Clearly each W_i has the standard Normal distribution, and they are mutually independent. If we define $\overline{W}_n$ as the average of these centered and rescaled random variables, then

$$\begin{aligned}\sum_{i=1}^{n}(W_i - \overline{W}_n)^2 &= \sum_{i=1}^{n}\frac{1}{\sigma^2}\big(X_i - \mu - n^{-1}\sum_{j=1}^{n}(X_j - \mu)\big)^2 \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\big(X_i - n^{-1}\sum_{j=1}^{n}X_j\big)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = (n-1)s^2/\sigma^2,\end{aligned}$$

by the definition (10.5.1) of s^2.

Consider a set of coefficients $a_{ik},\ i, k = 1, \ldots, n$, that satisfy the relations

$$\sum_{k=1}^{n} a_{ik}a_{jk} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad i, j = 1, \ldots, n. \tag{10.5.2}$$

Let $Y_i = \sum_{k=1}^{n} a_{ik}W_k,\ i = 1, \ldots, n$. By Theorem 10.1, linear combinations of Normal random variables are Normal, The expectation of Y_i is obviously zero, and its variance is $\sum_{k=1}^{n} a_{ik}^2 = 1$. It follows that Y_i has the standard Normal distribution for each $i = 1, \ldots, n$. If $i \neq j$, the covariance of Y_i and Y_j is

$$\operatorname{cov}(Y_i, Y_j) = \mathrm{E}(Y_iY_j) = \sum_{k=1}^{n}\sum_{l=1}^{n} a_{ik}a_{jl}\mathrm{E}(W_kW_l).$$

But the covariance of the independent random variables W_k and W_l is zero if $k \neq l$, and so only the terms in the double sum above for which $k = l$ are nonzero. Therefore, if $i \neq j$,

$$\operatorname{cov}(Y_i, Y_j) = \sum_{k=1}^{n} a_{ik}a_{jk}\mathrm{E}(W_k^2) = \sum_{k=1}^{n} a_{ik}a_{jk} = 0,$$

by (10.5.2).

Now $Y_i^2 = \sum_{k=1}^{n}\sum_{l=1}^{n} a_{ik}a_{il}W_kW_l$. Consequently,

$$\sum_{i=1}^{n} Y_i^2 = \sum_{k=1}^{n}\sum_{l=1}^{n} W_kW_l\sum_{i=1}^{n} a_{ik}a_{il} = \sum_{k=1}^{n} W_k^2,$$

since $\sum_{i=1}^{n} a_{ik}a_{il} = 1$ only if $k = l$, and is zero otherwise.

Next, let $a_{nk} = n^{-1/2},\ k = 1, \ldots, n$. We see immediately that $\sum_{k=1}^{n} a_{nk}a_{nk} = 1$, as required by (10.5.2). Then $Y_n \equiv \sum_{k=1}^{n} a_{nk}W_k = n^{-1/2}\sum_{k=1}^{n} W_k = n^{1/2}\overline{W}_n$.

Observe next that

$$\begin{aligned}\sum_{i=1}^{n}(W_i - \overline{W}_n)^2 &= \sum_{i=1}^{n}\big(W_i^2 - 2\overline{W}_nW_i + \overline{W}_n^2\big) \\ &= \sum_{i=1}^{n} W_i^2 - 2n\overline{W}_n^2 + n\overline{W}_n^2 = \sum_{i=1}^{n} W_i^2 - n\overline{W}_n^2.\end{aligned}$$

From what we have just seen, the last expression here is $\sum_{i=1}^{n} Y_i^2 - Y_n^2 = \sum_{i=1}^{n-1} Y_i^2$. Consequently, $(n-1)s^2/\sigma^2$ is the sum of $n-1$ squared independent standard Normal random variables, and so it has a χ^2 distribution with $n-1$ degrees of freedom. This completes the proof. ■

It is a legitimate question to ask where the coefficients a_{ik} come from. In the proof, we gave an explicit definition only of $a_{nk},\ k = 1, \ldots n$. For the other

coefficients, there is no unique definition, but here is one set of coefficients that satisfies the requirements. For $i = 1, \ldots, n-1$, and $k = 1, \ldots, n$,

$$a_{ik} = \begin{cases} 0 & \text{for } k < i \\ [(n-i)/(n-i+1)]^{1/2} & \text{for } k = i \\ -[(n-i)(n-i+1)]^{-1/2} & \text{for } k > i \end{cases}$$

The proof that these coefficients do indeed satisfy the requirements is left as a (tedious) exercise.

We have said several times that this case is unrealistic, because it assumes that the data are Normal and that they are known to be Normal. In a typical problem, we do not know the distribution from which the data come. How then will we find the distribution that results when we perform some operation such as taking the sample mean or variance, when we don't even know the distribution of the input data?

Perhaps surprisingly, it is possible to answer questions under these circumstances, using an invariance principle. An invariance principle states that, for any (input) distribution that has certain characteristics, performing some operation on the data will tend to produce, as sample size grows, a particular (output) distribution. The Central Limit Theorem, which we will discuss next, is an example of an invariance principle and states that the distribution of the standardized sample mean of data that have a few simple characteristics will converge toward the standard Normal distribution. With this result, it is not necessary to make unfounded assumptions about the nature of the data that we are analyzing.

Chapter 11

Laws of Large Numbers and Central Limit Theorems

Since the distribution of our data is in general unknown, we need statistical results that do not depend on knowledge of this type. This chapter describes two general classes of result that allow us to draw conclusions about date of unknown form, although of course they do require that some conditions hold.

Before describing these general classes of result, laws of large numbers and central limit theorems, we need to discuss what we mean by convergence in these stochastic contexts, and do some formal definitions of concepts that will turn out to describe types of convergence that can arise. These stochastic convergence concepts are distinct from deterministic convergence. For example, the sequence $\{1/n, n = 1, 2, 3, \ldots\}$ converges deterministically to zero; for any value of n, we can state exactly how close to zero the value in the sequence will be. To take another common example, the limit as the number of terms tends to infinity of the sum $1+1/2+1/2^2+1/2^3+\ldots$ is 2. By contrast, in the case of stochastic convergence, we only know that as some index increases, we will tend to move closer to some limit, in a sense that can be stated precisely using probabilities.

11.1 Some preliminary asymptotic theory

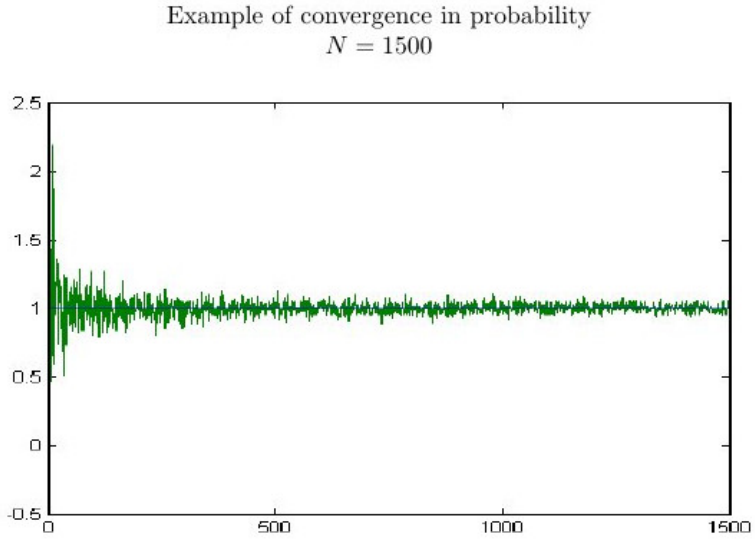
We begin therefore by defining convergence in probability and convergence in distribution.

D11.1 Convergence in probability: A sequence of random variables $\{X_n\}_{n=1,2\ldots}$ is said to *converge in probability* to a value X if

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1 \quad \text{for any } \varepsilon > 0.$$

This is typically denoted either by $X_n \xrightarrow{p} X$ or by $\text{plim}_{n\to\infty}\{X_n\} = X$; note that the value X may be a constant, or a random variable. Figure 11.1.1 provides an example of a sequence of 1500 observations on a random variable which is converging in probability to a probability limit of one, but sometimes moves closer, and sometimes moves farther away, from one. We observe that the range of its fluctuations around the probability limit tends to diminish as

FIGURE 11.1.1



sample size increases (in this example, it is proportional to the square root of sample size).

We will also often meet with cases in which a random variable does not converge to any particular value (random or fixed), but instead has a distribution which converges to another distribution.

D11.2 Convergence in distribution: A sequence of random variables $\{X_n\}_{n=1,2,\ldots}$ which have cumulative distribution functions $\{F_n(x)\}_{n=1,2,\ldots}$ is said to *converge in distribution* to a cumulative distribution $F(x)$ if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

at all points of continuity of the cumulative distribution function $F(x)$.

If there is a random variable X that has the CDF $F(\cdot)$, we can write $X_n \xrightarrow{D} X$. But there is no necessary connection between X and the variables X_n.

An illustration is provided by convergence of the t_k (k degrees of freedom) density to the N(0,1) as $k \to \infty$ recall Figure 9.2.3 of Chapter 9

11.2 Laws of Large Numbers

One important class of asymptotic result concerns convergence of a sequence of sample estimates to the true expectation of the process. The following theorem states the *weak law of large numbers*.

Theorem 11.1: (WLLN) Let $\{X_i\}$, $i = 1, \ldots, n$, be independent random draws from a distribution with cumulative distribution function $F_X(\cdot)$, with expectation μ and variance $\sigma^2 > 0$. Then the sample mean converges in probability to the true expectation, so that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P\big[|\overline{X_n} - \mu| < \varepsilon\big] = 1.$$

For a finite value of n we can specify a further parameter δ to describe the tradeoff between precision in the interval around μ and our degree of confidence in the statement. Let ε and δ be such that $\varepsilon > 0$, $0 < \delta < 1$ and let $n > \sigma^2/(\varepsilon^2\delta)$. Then

$$P\big[|\overline{X} - \mu| < \varepsilon\big] \geq 1 - \delta \quad \forall\ n > \sigma^2/(\varepsilon^2\delta). \tag{11.2.1}$$

As ε and δ become closer to zero, we are stating a more precise interval and higher probability of being in that interval. We can choose these parameters in order to determine which statement we wish to make, constrained by the necessity that the sample size should be large enough to make the statement valid (*i.e.* we must have $n > \sigma^2/(\varepsilon^2\delta)$. The larger is the sample size, the more precise the statement that we can legitimately make.

11.3 Central Limit Theorems

As we have said, in most cases we do not know the distribution of the data that we are analyzing. We might expect it to follow therefore that the distributions of statistics that we compute from these data will also be unknown. However in many cases, particularly involving sums or averages, the distribution of a statistic can be approximated well because we have information about convergence in distribution that applies to the statistic: that is, although its distribution is unknown, it can be shown to converge to a particular distribution $F(\cdot)$. We can therefore take $F(\cdot)$ as an approximation to the true distribution, which will become increasingly precise as sample information accumulates.

Central limit theorems are some of the most important and useful results in statistics, for at least two reasons. First, they apply to sums or means of random variables (data points), and taking the mean of a set of data is one of the most frequently applied operations. Recall for example that the moments and functions of moments that we studied earlier, such as the variance, coefficients of skewness and kurtosis, are based on expectations. There are sample counterparts that are therefore means of some random variable, such as the

sample variance $(n-1)^{-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$, which apart from the degrees of freedom correction is the sample mean of the random variable $(X_i-\overline{X})^2$. A vast set of statistics can be interpreted as sample means of some random variable.

Second, almost invariably we do not know the true distribution function of our data. To assume that we know the distribution when we do not can easily lead to false statements or results, particularly given that differences between distributions that may be critical in practice (for example, tail thickness or relative frequency of extreme events in risk management) can be very difficult to observe precisely in an empirical sample of data. We cannot calculate important quantities such as tail probabilities by the natural method of integrating a particular mathematical form of density function, or directly using the cumulative distribution function, because these functions are unknown.

A Central Limit Theorem (CLT) gives a result of that applies to any input distribution, as long as a few simple conditions are met. In the theorem that we will state here, these conditions are fairly weak (therefore apply fairly widely), but they can be weakened further. For this reason there are many CLTs, as the same result has been obtained under different assumed conditions on the true process.

Theorem 11.2: (CLT) Let $\{x_i\}$, $i=1,\ldots,n$, be independent random draws from a distribution with cumulative distribution function $F_X(\cdot)$ with expectation μ and variance $\sigma^2>0$. Then the standardized sample mean has a limiting standard Normal distribution: that is,

$$\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}} \xrightarrow{D} \mathrm{N}(0,1).$$

The notation N(0,1) denotes the Normal distribution with expectation zero and variance one: substituting into the Normal density function

$$f_X(x)=(2\pi\sigma^2)^{-1/2}\exp\Big\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\Big\}$$

the values $\mu=0,\sigma^2=1$, we have that the density of the N(0,1) distribution is

$$f_X(x)=(2\pi)^{-1/2}\exp\big\{-\frac{1}{2}x^2\big\}.$$

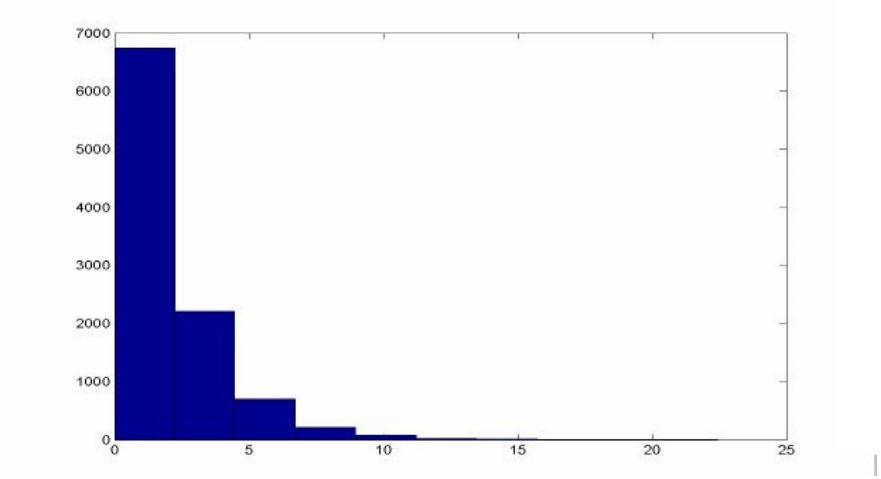
Theorem 11.2 could also be written in terms of the sum of the data points rather than the sample mean, because the sum $\sum_{i=1}^{n}x_i$ differs from the sample mean $n^{-1}\sum_{i=1}^{n}x_i$ only by the constant factor n; being a constant, this factor does not affect the shape of the distribution. So multiplying through top and bottom by n in Theorem 11.2, we have

$$\frac{\sum_{i=1}^{n}x_i-n\mu}{\sigma\sqrt{n}} \xrightarrow{D} \mathrm{N}(0,1).$$

The next figures illustrate convergence to the Normal. We take data points from a heavily right-skewed distribution, a typical sample of which is illustrated in Figure 11.3.1.[1] Numerous samples are taken from this skewed distribution, and each one is averaged; the density of the average is then estimated using kernel smoothing methods, and these densities are illustrated in Figures 11.3.2 and 11.3.3. In each of the latter figures, we plot the density of the sample mean for each of three different sample sizes. These six cases are separated into two figures, because if we placed them all on the same figure, the different widths and heights of the densities would make it difficult to observe the shapes of each one (for example,the largest sample size would appear as a thin spike if placed on the graph with the scale of Figure 11.3.2): note that the horizontal and vertical scales of the two figures differ. They otherwise have the same meaning.

FIGURE 11.3.1

Distribution of input data:
histogram of a single empirical sample

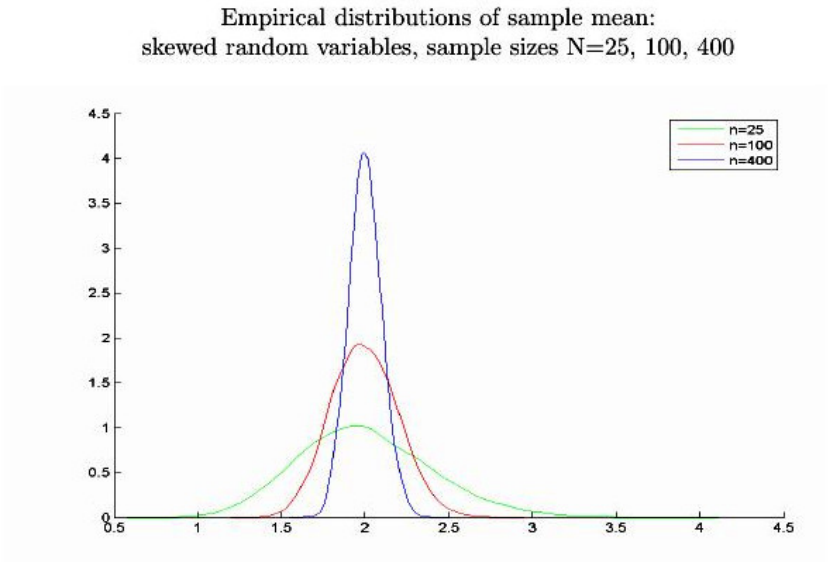


We observe that although this is an asymptotic result, conformity of the mean with the Normal distribution is quite good even at the smallest sample size, and there is almost no visible deviation from symmetry remaining. This should not be taken to indicate that a CLT is always a good approximation even at a very small sample size; we are treating here cases of independent random sampling, which is a relatively straightforward case. Other CLTs can be proven for data which have some dependence, but larger sample sizes will typically be required for this degree of conformity with the asymptotic distribution to appear.

[1] These pseudo-random data are in fact generated from the χ^2_1 distribution.

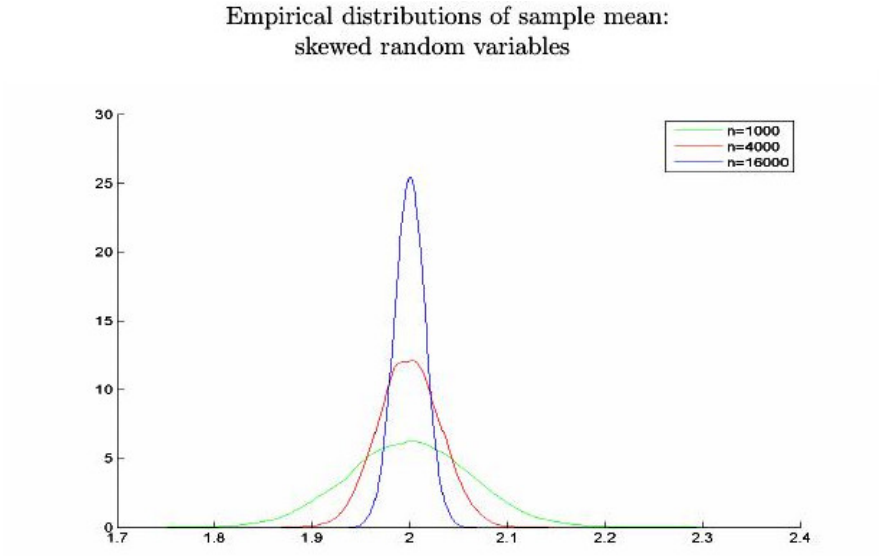
We observe also that as we move to higher and higher sample sizes, the densities become ever more concentrated around the true mean, in conformity with the WLLN as well as the CLT. In fact if we take a range containing any given proportion of the data (for example, imagine marking points on the axes that contain about 99% of the area under each of these densities), the range required to contain the given proportion shrinks as the sample size increases note again the difference in scales between the two figures. As sample size increases by a factor of four, the range containing the given proportion of the data shrinks by a factor of about two: that is, our estimates become more precise at a rate equal to the square root of the rate of increase of sample size. This is an example of 'root-n convergence',

FIGURE 11.3.2



which appears frequently in simple parametric estimation problems such as this, and which can also be read from the left-hand side of the result in Theorem 11.2: this ratio converges to a constant distribution, and so the numerator must on average shrink at the same rate as the denominator, which, because σ is a constant, necessarily shrinks at the same rate as the other term, $\sqrt{n}$. This is also a reflection of the result proven earlier that the variance of the sample mean in an independently and identically distributed random sample is σ^2/n, which means that the standard deviation of the sample mean is $\sigma/\sqrt{n}$ and therefore declines at a rate proportional to the square root of the sample size.

FIGURE 11.3.3



Note again that the existence of first and second moments is needed, and that this is a condition that can fail, particularly in financial data containing many extremes.

It is also crucial to remember that a CLT describes the distribution of a *statistic calculated from the data* and not the data themselves. Unless data arise from a process of summing or averaging, a central limit theorem does not provide any reason to suppose that data should be approximately Normal.

11.4 Application to the distribution of sample proportions

It is straightforward to show that this result applies as well to the distribution of a proportion. Define a 0/1 variable such that the variable takes the value one if a certain condition holds, zero otherwise. Let p be the proportion of cases in the population for which the condition is true. For example, let p represent the proportion of the population who would vote 'yes' to a referendum question, and for every person x_i from that population who is sampled, let the individual be coded as 0 if he or she will vote 'no', and 1 if he or she will vote 'yes'. The number of people in a sample drawn from the population who say that they will vote 'yes' can be denoted by n_y and the sample size by n, so that the sample proportion $\hat{p}$ who say that they will vote 'yes' is n_y/n. But n_y is also the mean of the 0/1 random variable X_i: $n^{-1}\sum_{i=1}^{n} X_i = 1/n$ times the number of 1's in the sample $= n_y/n$.

Now $\hat{p}$ is the sample proportion, and is also the sample mean of a 0/1 variable indicating that the condition (vote 'yes' in the referendum) holds. Therefore results pertaining to a sample mean also apply to $\hat{p}$.

The variance of the random variable $\hat{p}$ takes a simple form, since each X_i is a Bernoulli random variable, and the number of 'successes' (here, 'yes' votes) has a binomial distribution. We can work out the variance (or standard deviation) of $\hat{p}$ as a function of the population value p, so that p is the only unknown value in the sampling distribution.

Note that $\mathrm{E}(X_i) = 0.(1-p) + 1.p = p$, and $\mathrm{E}(X_i^2) = \mathrm{E}(X_i)$, since $0^2 = 0$ and $1^2 = 1$. Thus $\mathrm{Var}(X_i) = \mathrm{E}(X_i^2) - \big(\mathrm{E}(X_i)\big)^2 = p - p^2 = p(1-p)$. If we assume random sampling, the X_i are mutually indepeendent. Then since $\hat{p} = n^{-1}\sum_i X_i$, we have $\mathrm{E}(\hat{p}) = p$, and and

$$\mathrm{Var}(\hat{p}) = n^{-2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = n^{-1}p(1-p).$$

Since the sample proportion is a sample mean of the 0/1 random variables, and since the mean and variance exist as we have just shown, we therefore can obtain the asymptotic distribution using Theorem 11.2. Substituting $\hat{p}$ for the generic expression $\overline{X}$, p for the generic symbol μ, and the standard error of the proportion for the general expression for the standard error of $\overline{X}$, we have

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \xrightarrow{D} \mathrm{N}(0,1).$$

This result can be used to obtain confidence intervals for proportions, as in the confidence interval computations that we saw in Chapter 10.

Proof of the weak law of large numbers given above.[2]

Recall the Markov inequality (7.3.2): let X be a random variable and $g(\cdot)$ a non-negative function on $\mathcal{R}$ such that $\mathrm{E}\big(g(X)\big)$ exists. Then

$$P\big(g(X) \geq k\big) \leq \frac{\mathrm{E}\big(g(X)\big)}{k} \;\forall\, k > 0. \tag{A11.1}$$

As with the proof of the Chebychev inequality given earlier, we make some specific choices. The random variable X in the general definition (7.3.2) is replaced here by the sample mean $\overline{X}$. Then we set $g(\overline{X}) = (\overline{X} - \mu)^2$ and $k = \varepsilon^2$, where μ and σ^2 are the mean and variance of X (*i.e.* of the data) and ε is the bound on the discrepancy between the true (population) mean μ and estimated (sample) mean $\overline{X}$.

We can re-write the statement (A11.1) above in equivalent form as

$$P\big(g(X) < k\big) \geq 1 - \frac{\mathrm{E}(g(X))}{k} \;\forall\, k > 0.$$

Making the substitution of our choices for $g(X)$ and k, we find

$$P((\overline{X} - \mu)^2 < \varepsilon^2) \geq 1 - \frac{\mathrm{E}((\overline{X} - \mu)^2)}{\varepsilon^2} \;\forall\, \varepsilon > 0.$$

Now since we are dealing with a simple case of independently distributed data, we have $\mathrm{E}((\overline{X} - \mu)^2) = \sigma^2/n$, as shown earlier. Therefore

$$P((\overline{X} - \mu)^2 < \varepsilon^2) \geq 1 - \sigma^2/(n\varepsilon^2).$$

Taking square roots of the quantities inside the probability on the left side does not change the statement, and so if we define $\delta = \sigma^2/(n\varepsilon^2)$, we are left with the statement

$$P(|\overline{X} - \mu| < \varepsilon) \geq 1 - \delta, \tag{A11.2}$$

or, re-arranging, $n = \sigma^2/(\varepsilon^2\delta)$, as required by (11.2.1). The statement remains true for larger δ (if the probability is $\geq 1 - \delta$, it is also $\geq 1 - \delta'$ for $\delta' > \delta$ since the latter is a lower probability). So (A11.2) holds for any $n \geq \sigma^2/(\varepsilon^2\delta)$ (integer constraints on sample size may make it impossible to find an n that makes this hold with equality.) ■

[2] This proof is based on that given by Mood *et al.* (1974).

Chapter 12

Sampling distributions revisited

In Chapter 10 we saw how we can construct confidence intervals for an estimated parameter, given knowledge of the distribution of that estimate. In the case is considered there, we were able to determine that distribution using theoretical results on the relations between distributions, but only by assuming that we knew the distribution of the input data. In typical practical cases, we don't of course know the distribution of the input data; we obtain it in the form of a matrix or table, and set out to analyze some question of interest. Nothing is in general known about the distribution of any quantity in the data matrix; we cannot work out these distributions mathematically without knowledge of the inputs. We therefore need to have ways of approximating the unknown distributions to an acceptable degree of accuracy. There are two broad classes of approximation which are commonly used: approximations based on asymptotic theory, where we take the distribution to which another distribution converges asymptotically as its approximation, and simulation-based approximation, where we use computer-generated random numbers to emulate a problem and attempt to approximate the relevant distributions. Simulation-based methods such as Monte Carlo tests and bootstrap tests have wide applicability, but require different forms for different types of problem and therefore require some sophistication in their implementation. These simulation-based methods are beyond the scope of this book, at least at the time of writing. Although we will discuss asymptotic approximations, it's worth underlining that simulation-based methods may also be used in this context, in part to check on the accuracy of asymptotic approximations in different cases.

In Chapter 11 we saw that we can obtain asymptotic distributions for estimates in some such cases, using other theoretical results, and in particular central limit theorems. Without knowing the distribution of the input data, we can determine nonetheless that the sample mean has an asymptotically Normal distribution, given some simple conditions which will often apply. The asymptotic distribution does not correspond perfectly to the finite-sample distribution, but will provide a good approximation in a wide range of circumstances.

In the present chapter we will see how to apply central limit theorem results to obtain confidence intervals for estimators. We will therefore have

moved to a set of methods that allow us to handle the realistic problem in which we begin with nothing more than a set of numbers of unknown distribution, and compute confidence intervals for estimates that can be expressed as the mean of something measurable. This kind of argument is widely applicable, and forms the basis for a great deal of applied statistical inference, giving approximate confidence intervals with a reliable justification. After we go through the method in the next section, we'll step back to look at an example where we start with a set of numbers that we know very little about, and get some confidence intervals for estimates.

12.1 Sampling distributions based on a CLT

Consider the following problem. We want to know whether two random variables have the same mean, and we have a random sample from each. Let these two random samples be $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, and let the means of the underlying random variables X and Y be μ_x and μ_y. If $\mu_x = \mu_y$ then the mean of $(X - Y)$ is zero, so we can look at the question by considering the difference $X - Y$.

We will put a label on $X - Y$, and call it D for difference. Then we have a sample $\{d_i\} = \{x_i\} - \{y_i\}$ for $i = 1, \ldots, n$, and $\overline{d} = n^{-1} \sum_{i=1}^n (x_i - y_i)$. We do not know the distribution of X, Y, or D, but we will make the weak assumption that its mean and variance exist.[1] Then we can get a confidence interval for μ_d from a central limit theorem: we know

$$\frac{(\overline{d} - \mu_d)}{\sqrt{\sigma_d^2/n}} \xrightarrow{D} \mathrm{N}(0, 1), \tag{12.1.1}$$

and that replacing σ_d^2 with $s_d^2 = (n-1)^{-1} \sum (d_i - \overline{d})^2$ leaves this asymptotic distribution result intact. We can then obtain an approximate (asymptotic) confidence interval for the true μ_d using the same manipulations that we saw earlier:

$$P\left(z_{\alpha/2} < \frac{(\overline{d} - \mu_d)}{\sqrt{s_d^2/n}} < z_{1-\alpha/2} \right) \simeq 1 - \alpha,$$

which implies $P(z_{\alpha/2}\sqrt{s_d^2/n} < (\overline{d} - \mu_d) < z_{1-\alpha/2}\sqrt{s_d^2/n}) \simeq 1 - \alpha$ and so

$$P(-\overline{d} + z_{\alpha/2}\sqrt{s_d^2/n} < -\mu_d < -\overline{d} + z_{1-\alpha/2}\sqrt{s_d^2/n}) \simeq 1 - \alpha$$

which, reversing the inequalities as we change sign, implies

$$P\left(\overline{d} - z_{\alpha/2}\sqrt{s_d^2/n} > \mu_d > \overline{d} - z_{1-\alpha/2}\sqrt{s_d^2/n} \right) \simeq 1 - \alpha. \tag{12.1.2}$$

[1] We may know or observe sufficient conditions for this, such as that the two sequences are bounded, which guarantees the existence of all finite moments.

If we recall that for a distribution symmetric about the origin, $q_{\alpha/2} = -q_{1-\alpha/2}$, it follows that μ_d is in the interval $\overline{d} \pm z_{1-\alpha/2}(\sqrt{s_d^2/n})$ with probability approximately equal to $1 - \alpha$. Note again that we say 'approximately' because this is not an exact finite-sample confidence interval like what we could obtain if we somehow knew for example that the original data were Normal. Instead this is an approximation based on the asymptotic results, which will tend to become more and more accurate as sample size increases.

We have stated the problem here in the form of a difference between two series, which is a type of problem that is often interesting (are starting salaries in some job the same for men and women with the same qualifications? Are accident rates for first-year drivers the same for those who completed a driver education course and those who did not?) However the same method can be applied whenever we are interested in the mean of a random variable with an unknown distribution, as long as that distribution possesses the first two moments, which will very commonly be the case and can sometimes be verified unambiguously.

12.2 Example

Consider in more detail the example of the number of accidents that newly licensed drivers have in their first years of driving.[2] Note that this random variable is discrete $(0, 1, 2, \ldots)$ and is bounded on one side. It certainly cannot be Normally distributed.

We might look at whether the true mean number of accidents per year is some round number such as 1, for example, but that doesn't sound very interesting: the true answer is almost certainly going to be some fraction. A more interesting question is the male-female difference.

Consider that a driving school tracks data on its students for one year after licensing. Over a certain period, the school has 422 male graduates and 378 female graduates.

The numbers are not the same, so we would not simply take the difference between pairs of students as above. To illustrate a point, however, imagine for a moment that we have 378 observations on each group. We can then take pairs and construct the difference $d_i = x_i - y_i,\ i = 1, \ldots, 378$, that is, the difference between the number of accidents that the male driver has and that the female driver has in each pair; this can of course be positive, negative or zero in each case. Then, a confidence interval for the difference exactly fits the pattern above.

[2] We of course need to make a precise definition of the term 'accident': for example, an accident might be defined to be an event leading to damage which is reported to an insurance company. This definition would exclude many small incidents, which are not reported because the cost of repair is less than, or not much more than, the insurance deductible.

While this test would be straightforward, it fails to use all of the available information; the additional 44 male drivers are excluded from the sample, and of course the results of the test will depend upon which 44 drivers are excluded. In general, sample sizes from two groups may be unequal, and so we would like to be able to derive a test which does not depend on the assumption of equal sample sizes.

Bearing in mind that the number of accidents in a group can reasonably be supposed to possess the same mean and variance (this would be guaranteed if the number of accidents is bounded above, for example if access to a vehicle is removed for anyone who has more than some number of accidents), we will be able to use a central limit theorem and therefore an asymptotic normal approximation for this problem. Let's say that the quantity of interest to us is the mean number of accidents for each of the two groups, male and female drivers in the first year of licensing. (We could by contrast look at a different quantity such as the probability of having at least one accident.) The question of interest to us is whether the mean on each of these groups, μ_1 and μ_2, is the same, and one way to approach that is to construct a confidence interval for the difference in the mean number of accidents in the two groups. We can do that along the lines indicated above, computing $\overline{d}$ this time as $\overline{d}_1 - \overline{d}_2$. The problem now is to compute the standard error of this difference, which at first sight may look tricky because we do not have a single set of differences and so cannot compute it directly as the standard error of a particular data series. However, the quantity that we need can be computed straightforwardly using the expression for the variance of a linear combination of two random variables X and Y, that is $\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{cov}(X, Y)$, for weights a and b. Here, $a = 1, b = -1$ and it is reasonable to suppose that the covariance is approximately 0.

This is an example of a case in which we may substitute a value for some quantity based on reasoning about the way the sample is constructed. Here, we might have samples of male and female drivers who don't know each other and live in different places, so that we feel confident that the driving experiences of individuals are independent of each other, so that their covariance must be zero. We may also be aware that this is not a literal truth; for example, imagine that one of the male drivers and one of the female drivers are a couple who like to send texts to each other, and drive different cars. If they do this while driving – one of the forms of behavior that clearly seems to raise accident rates – then they may each have elevated probabilities of an accident, and may even text each other while both parties are driving simultaneously, so that the male and female accident rate data are no longer independent and the covariance would be positive. To the extent that this is possible, it seems likely to be a very small effect, so that in practice we would probably continue to use the approximation that the covariance between the two series is zero, being aware nonetheless that this is an approximation and may not be a literal truth.

We therefore have $\mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. We can compute the estimated variance s^2 for each of the series $\{x_i\}$ and $\{y_j\}$ recording the number of accidents for each male driver indexed i and each female driver indexed j, and the variances of the *means* of the two series are then s_x^2/n_x and s_y^2/n_y. The estimated variance of the difference is then $s_x^2/n_x + s_y^2/n_y$ and the standard error of the difference $s_d = \sqrt{s_d^2}$ is the square root of that quantity. We can then compute a confidence interval using the expression (12.1.2) above.

The problem discussed here is related to what is called the *Behrens-Fisher problem*, after two statisticians, Waldemar Behrens and Ronald A. Fisher, who each contributed to the study of the problem. The problem is to develop a test for the null hypothesis that the expectations of two populations are equal, on the basis of two random samples, generally of different sizes, drawn from the two populations, *without* assuming that the population variances are the same. However this classic problem, discussed in Lehmann (1986), assumes that the population distributions are Normally distributed.

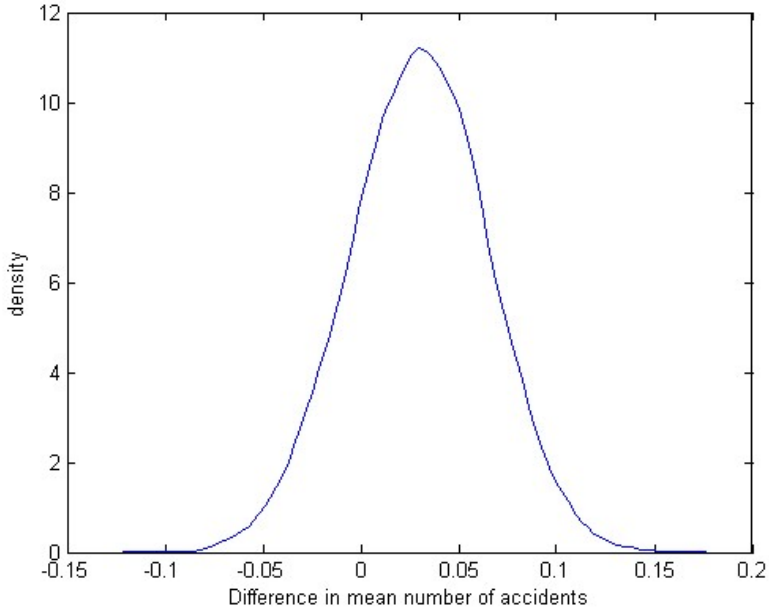
To illustrate further, here is an example with some numbers computed on a simulated data set constructed with known probabilities of accidents. The sample sizes for male and female drivers are 422 and 378, and data series of these lengths were constructed such that each data point is the number of accidents for that driver, 0, 1 or 2. The data were constructed in such a way that the accident probabilities are not in fact exactly the same (the distribution will be described below). The statistics computed on the two samples were: $\mu_x = 0.4123$; $\mu_y = 0.3783$; $s_x^2 = 0.2429$; $s_y^2 = 0.2358$; therefore $\overline{d} = \mu_x - \mu_y = 0.0340$; $s_{\mu_x}^2 = s_x^2/422 = 5.7556 \times 10^{-4}$; $s_{\mu_y}^2 = s_y^2/378 = 6.2385 \times 10^{-4}$; finally using the expression for the variance of the difference as above, the standard error of the difference is $s_d = [s_{\mu_x}^2 + s_{\mu_y}^2]^{1/2} = 0.0346$. This is quite close to the difference itself, meaning that the difference is about one standard error away from zero; clearly therefore we do not have very strong evidence against zero being the correct value for the difference. The 95% confidence interval for the difference, using the percentage points of the asymptotic standard Normal distribution, is $0.0340 \pm 1.96(0.0346)$ or [-0.0338, 0.1018], so that zero is well inside the interval.

Thus these data do not show clearly that there is any difference in average number of accidents between male and female drivers, although the sample average computed is somewhat higher for male drivers. In fact, for this example the data were generated to be such that male drivers do have a higher mean number of accidents; why then does the sample not give us a clearer result? Intuitively, it must be because the sampling variation is large enough to make any difference difficult to discern; the signal is obscured by noise. We can examine this more clearly by repeating the experiment many times, to see what happens. It's possible in this case because these data were constructed for this example rather than observed empirically, and so instead of constructing one example we are free to construct many, and observe the outcomes in a large number of similar cases. In the simulations that will be recorded in

the next two density functions, there were 10,000 experiments each with the same sample sizes of 422 and 378. The value on the lower axis in each of these figures indicates a male-female difference, and the density indicates how likely it is to observe differences of that magnitude.

What we see in the first figure (apart from the fact that conformity with the shape of the asymptotic normal distribution is quite good, as we expect from a central limit theorem in a case with independent observations) is that the range of outcomes observed is quite broad, and that although these data are constructed in such a way that male drivers do have a slightly higher accident rate, samples in which the difference is zero or negative, that is that on a particular sample the female drivers had more accidents, are by no means uncommon. Many samples of these sizes, in other words, would lead to observations where the female drivers have a higher mean accident rate; the degree of random variation is very substantial relative to the quantity that we are trying to estimate. In the next example, we do the same thing but with much larger sample sizes: each of the previous sample sizes is multiplied by 10. We again take 10,000 examples on these larger sample sizes, and plot the density over each of these 10,000 estimated differences.

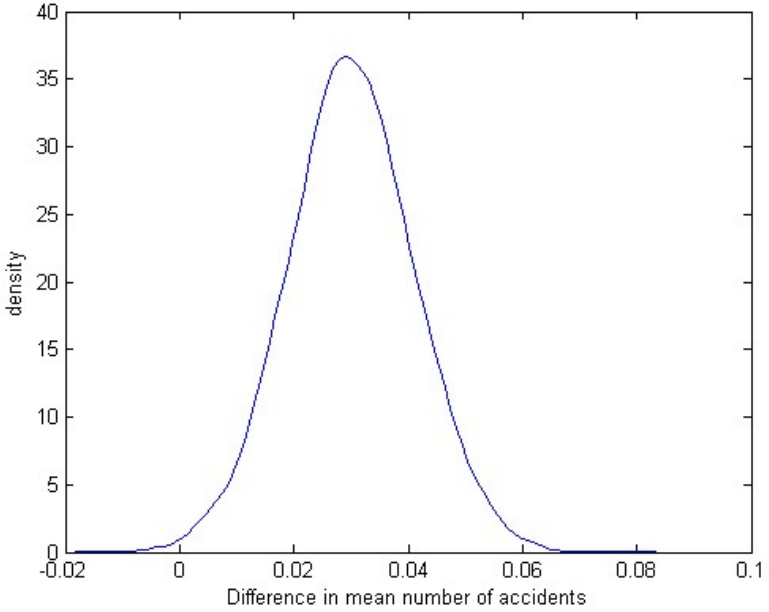
FIGURE 12.2.1



Although the shape of the distribution is the same in the second case, we see that the numbers on the lower axis are less spread out (they're scaled

down by $\sqrt{(10)}$, of course). Now, observing a male-female difference that is near zero or even negative is quite a rare event; the mean differences about the same, around 0.02, but there is much less sampling noise on these larger sample sizes and the results are more reliable in the sense that the estimated difference comes out to be positive, as it in fact is, over 95% of the time.

FIGURE 12.2.2



In our single original example, zero was in the confidence interval. We could not reliably conclude that the mean difference in accidents between male and female drivers was a positive number. We now see that with a larger sample from the same process, we would have concluded that zero is outside even a 95% confidence interval: a process with a zero difference is unlikely to have generated these data. [3]

This illustrates a general feature about drawing, or failing to draw, a conclusion from data: it may well be that we will fail to see a difference

[3] Data were generated by independent random draws such that for male drivers, the probability of no accident was 0.6, of one accident was 0.395, and of two accidents was 0.005. For female drivers, the probability of no accident was 0.65, of one accident was 0.345, and of two accidents was 0.005. Male drivers therefore have higher probability of at least one accident (0.65 vs 0.60) and higher mean number of accidents.

between two things because our sample does not contain enough information (contains too much sampling noise). When therefore we fail to find a clear difference (or in the language of hypothesis testing that we will soon use, when we fail to reject the null hypothesis), it does not mean no difference is there (it does not mean that the hypothesis is true). It could simply be that our data were insufficiently informative to allow us to conclude that. Failing to show that something is false doesn't prove that it's true.

This example might also help us to recall that the exact distributional result for the t statistic depends on quite strong assumptions about the distribution of the input data. But here that is clearly not the case; the input data are the numbers of accidents for samples of different people, which are integers with a distribution that is bounded and skewed right, and therefore nothing like what would follow from a Normal distribution.

All we can rely on to obtain a distribution of the test statistic in this case is a central limit theorem, which tells us that the mean will have an asymptotic Normal distribution. If we are relying on our asymptotic approximation from the central limit theorem, then our asymptotic distribution is Normal.

Is it wrong therefore to use a t distribution rather than a Normal to get the confidence interval, that is to use $t_{k,\alpha/2}$ for k degrees of freedom, rather than $z_{\alpha/2}$, in the confidence interval expression?

Bear in mind that the t_k distribution converges to $\mathrm{N}(0,1)$ as k increases. In other words, these distributions are asymptotically the same. The size of the confidence intervals given by each of these distributions becomes arbitrarily close as the sample size increases. For example with a sample size of 60 we have, for a 95% confidence interval, $t_{60,0.975} = 2.0003$ and $z_{.975} = 1.960$ if we are using the standard Normal. The size of the confidence intervals that we obtain will therefore differ by about 2% (i.e. about 0.040/2). Of course, the distributions differ by greater percentages as we go farther out into the tails, so that if we wanted a 99% confidence interval or a 99.99% confidence interval, the percentage difference would be larger. Nonetheless, all of these differences decline with sample size; for a sample size of 400 and a 95% confidence interval, we have $t_{400,0.975} = 1.9659$ and of course $z_{.975} = 1.96$ still.

Notice that, although the values are typically close and certainly converging as sample size (degrees of freedom) increases, the value for the t distribution is never less than for the Normal. Therefore confidence intervals computed using the t distribution will be slightly wider, or at least no smaller. This in turn means that we are making a slightly less strong, that is a slightly more conservative, statement if we compute those values using the t distribution. Given that this inference is approximate, and given that we would rather err on the side of weaker statements rather than statements which are excessively strong (in other words we would rather not exaggerate the strength of the conclusions that we can draw from the data), many statistical workers will tend to prefer the slightly more conservative confidence intervals produced

by use of the t distribution. This is a perfectly sensible practice as long no one is deluded into thinking that the confidence intervals are exact because a t distribution is used rather than an 'asymptotic' Normal. Whichever of these distributions is applied here, we are relying on the asymptotic approximation provided by the standard Normal, and justified by the central limit theorem.

Chapter 13

Point estimators

Much of the time, the quantities that we want to estimate are scalar values, or sets of scalar values. These might be responses of one variable to another such as elasticities, probabilities, moments of the distribution such as the mean or variance, and so on.

For example, we might want to estimate the price elasticity of demand (proportionate change in demand divided by proportionate change in price) for gasoline, in order to reduce consumption and environmental damage, using a Pigovian tax.[1] This example also illustrates some of the ways in which we approximate an entire time sequence with a subset of the values; in fact, the introduction of a tax on gasoline will have an immediate effect but also a changing effect over time, as people adapt their behaviour to the new tax. The immediate impact might arise only through a reduction in optional or recreational trips in the car, but over time other adaptations become possible, such as buying a smaller car than when gasoline prices were higher, building (in response to demand conditions) more apartments near areas where many people work, rather than houses in the suburbs, and so on. Typically, we do not attempt to estimate the entire dynamic path of the response of gasoline demand to a change in price over the long horizon until it stabilizes at some value; instead we usually approximate the information in this path with a short-term elasticity (the immediate effect that we observe in the first weeks or months after the introduction of a tax, before capital expenditures have adjusted) and a long-term elasticity (the value to which the elasticity settles down after people have had time to adjust fully their stocks of capital, including cars and housing units). In this case we would have two price elasticities of demand for gasoline to estimate, the short-term and the long-term.[2]

[1] This term is a reference to the classic work of A.C. Pigou (1920) on the use of taxes to reduce negative externalities.

[2] Of course, these values are different in different places and at different times, so that we constantly need to be updating our statistical information about these values. Nonetheless virtually all published estimates are less than one in absolute value, indicating inelastic demand: a given percentage change in gasoline prices leads to a smaller percentage change in demand. Typical values are around -0.4 for the short-term elasticity and around -0.6 for the long-term

Another example, which illustrates a different set of difficulties that we face, would be estimation of the average effect of completing a university degree on the annual income of an individual, at different points in the lifespan; say at ages 30, 40, and 50. In using the word 'effect' in the previous sentence we seem to imply a causal relationship, *i.e.* we imply that the fact of completing a university degree itself raises the income of the individual. It's possible of course that people who complete university degrees will have higher incomes on average at each of these ages, simply because people who complete a degree will on average have other qualities that will tend to lead them to higher incomes, such as ability, persistence, capacity for a high workload, and so on. The fact that a higher income is associated with completing a degree does not necessarily imply that the degree itself was the cause of the higher income. Attempting to distinguish causality from association is the subject of a large literature in statistics and econometrics, to which we will refer only briefly later in this book. For our present purposes, we simply need to note that this value, the increase in income associated with the degree at each age, is a set of scalar numbers that we want to estimate.

13.1 Estimators

D13.1 An *estimator* is a function of the data that provides an estimate of a population ('true') quantity.

For example, earlier we defined the sample mean $n^{-1}\sum_{i=1}^{n} x_i$ as an estimator of the true mean of the distribution of a random variable X. An alternative estimator that we defined was the trimmed mean, $(n-2k)^{-1}\sum_{i=k+1}^{n-k} x_i$, estimating the same quantity, but with a different trade-off between efficiency and robustness. We saw therefore that there could be more than one estimator of the same quantity, with different properties.

D13.2 A *point estimator* is an estimator that produces a scalar value, or a vector of values.

In the next chapter we will consider interval estimators, which produce estimates of an interval within which some value lies, or is likely to fall.

If we use the label θ for the quantity of interest in the population, then we will typically write the estimate as $\hat{\theta}$, which is some function of a data set X so that we can write $\hat{\theta} = g(X)$.

elasticity. This information helps to choose the tax rate that will reduce demand by a particular proportion.

There are typically numerous estimators that we might think of for a particular problem. In order to choose among them, we need some objective criteria that we consider desirable.

First define the distribution of an estimator, $F(\hat{\theta})$, with density (if it exists) $f(\hat{\theta})$ and mean $\mathrm{E}(\hat{\theta})$.[3]

The bias of an estimator does not have quite the same meaning as in normal speech, where it usually refers to an attitude on the part of a conscious human being which will tend to push the individual toward one or another conclusion, independent of evidence. The meaning here is related but does not imply any unfairness or poor decisions; there are some contexts in which it makes sense to use a biased estimator (see for example Chapter ––, time series).

D13.3 The bias of an estimator $\hat{\theta}$ of a parameter θ is $\mathrm{E}(\hat{\theta}) - \theta$.

An unbiased estimator has mean equal to the true value; a biased estimator does not.

D13.4 An *unbiased* estimator is one for which bias$(\hat{\theta})$=0.

An efficient estimator is one that uses information as completely as possible to pin down the true value to as narrow an interval as possible (recall that although we are talking about the value itself in this chapter, we will discuss estimating intervals in the next chapter); $\hat{\theta}_1$ is said to be more efficient than $\hat{\theta}_2$ if the variance of $\hat{\theta}_1$ is less than the variance of $\hat{\theta}_2$. Of course, if $\hat{\theta}_1$ has a larger bias, it may be worse for our purposes in spite of having lower variance, so that we may need to trade off the two qualities of low bias and low variance.

D13.5 An estimator $\hat{\theta}_1$ is more efficient than another, $\hat{\theta}_2$, if $\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$.

In order to make a trade-off in some formal way, it's useful to define some function of the bias, variance, or other features of the distribution that describe how bad the outcome is considered to be when $\hat{\theta}$ misses θ by a particular amount. We will often then want to estimate the expected value of this 'badness' or loss over the distribution of the estimate.

[3] There may be points where the density does not exist because, for example, the distribution function jumps at a certain point, in which case the slope is non-finite and there is no finite value for the derivative, or density.

D13.6 A *loss function* $\ell(\hat{\theta}, \theta)$ is a real-valued function that describes the loss associated with an estimate $\hat{\theta}$ when the true value is θ.

The loss may be purely a function of the estimation error, $\hat{\theta} - \theta$, in which case the loss function can be written as $\ell(\hat{\theta} - \theta)$.

In order that this function can represent loss, it must be true that $\ell(\hat{\theta}, \theta) \geq 0$ and $\ell(\hat{\theta}, \theta) = 0$ where $\hat{\theta} = \theta$.

D13.7 A *risk function* $L(\hat{\theta}, \theta)$ gives the expectation of the loss associated with an estimate $\hat{\theta}$ when the true value is θ.

Although loss and risk are clearly distinct concepts, it is commonplace to refer to either of them as a loss function (that is, either the loss function or the expectation of that function over the density of the estimate).

Here are several examples of risk functions; in each case.strictly speaking, the corresponding loss function is defined at a single value, rather than as an expectation over the distribution of $\hat{\theta}$; for example, the squared-error or quadratic loss is $\ell(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$. However, again following common usage, we will generally use the term loss function for either loss or risk, relying on context to distinguish a single value from the expectation.

Mean squared error: $L(\hat{\theta}, \theta) = \mathrm{E}[(\hat{\theta} - \theta)^2]$.

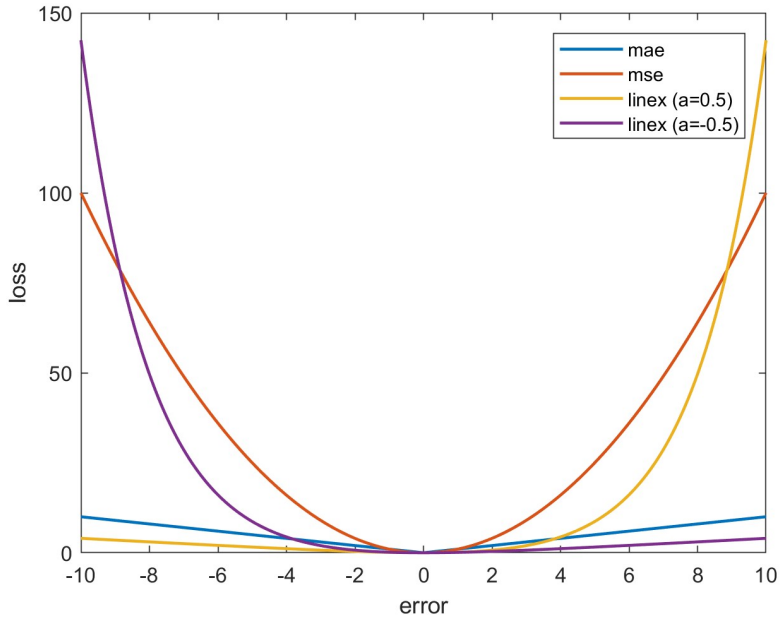
Mean absolute error: $L(\hat{\theta}, \theta) = \mathrm{E}[|\hat{\theta} - \theta|]$.

Linear-exponential ('linex') loss: $L(\hat{\theta}, \theta) = \mathrm{E}\left[\exp a(\hat{\theta}_i - \theta_i) - a(\hat{\theta}_i - \theta_i) - 1\right]$, where $\exp(\cdot)$ is the exponential function, *i.e.* $\exp(z) \equiv \mathrm{e}^z$, and a is a shape parameter that determines whether errors increase exponentially on the positive or negative side.

These are theoretical values of the functions, defined using the (population) expectation. If one is estimating the loss (risk) function on empirical data, for example on a series of forecasts and outcomes, then the expectation would be replaced by an estimator such as the sample mean. For example if $\hat{y}$ is a forecast of a random variable y, and if a sequence of N forecasts of it becomes is available, then we could estimate the mean squared error from this estimator or the future value as $\sum_{i=1}^{N} (\hat{y}_i - y_i)^2$.

Figure 13.2.1 illustrates the three functions of a given error. Notice that the linex functions are individually asymmetrical around zero. In the pair illustrated here the left-hand side of one resembles the right-hand side of the other because the shape parameters are equal and opposite; however each is linear on one side of zero and exponential on the other. Notice also that the growth rate of the exponential function eventually comes to dominate the squared-error function as the error increases in magnitude.

FIGURE 13.2.1
Loss as a function of $\hat{\theta} - \theta$:
Absolute error, squared error and Linex loss functions



Each of these functions has other interesting properties. It's straightforward to show, although one needs to be very careful to keep track of parentheses and expectation operators, that the mean squared error is equal to the squared bias plus the variance, so that if the bias is zero, the mean squared error is equal to the variance of the estimator. Formally:

Theorem 13.2.1: (Mean squared error = squared bias plus variance.) Let $\hat{\theta}$ be an estimator of a parameter θ. Then

$$\mathrm{E}[(\hat{\theta} - \theta)^2] = \left(\mathrm{E}[\hat{\theta} - \theta]\right)^2 + \mathrm{E}[(\hat{\theta} - \mathrm{E}[\hat{\theta}])^2].$$

Proof: See the Appendix. ■

The Mean Absolute Error function is non-differentiable at zero: although this is difficult to see in the figure above, the two straight lines meet at zero at a fixed angle rather than in a smoothly curved transition. This means that there is no unique tangent to the function at zero, and therefore no unique derivative. One result is that the MAE is more difficult to use in theoretical

work; one cannot use the derivative of the function at all points. Proving results that can be proven straightforwardly for the MSE often requires much more mathematical sophistication. This is one of the reasons that the MSE is widely used; nonetheless it is often argued that symmetry is an unrealistic description of a user's loss for many problems. The linex loss function is one way to allow a departure from symmetry, describing either positive or negative errors as causing exponential loss, depending upon the sign of the parameter embodied in the function. However, symmetry may be a reasonable approximation for many problems, or at least a reasonable starting point if one does not know what asymmetries may be present in users' loss functions.

Among other properties that we would like an estimator to have, consistency (*i.e.* with enough data, the estimator will converge probabilistically to the true value) and asymptotic normality are two of the most commonly investigated. It is often impossible to establish these properties by applying laws of large numbers and central limit theorems to functions of the data that the estimator is based upon.

D13.8 An estimator $\hat{\theta}$ is *consistent* for a parameter θ if $\hat{\theta} \xrightarrow{p} \theta$ as the sample size $n \to \infty$.

Defining asymptotic Normality requires a little care because it may apply even in cases in which the density of the estimator does not exist at any finite sample size; we cannot therefore state that the density of the estimator must converge on the Normal density, as the density of the estimator may not exist at finite sample sizes. Nonetheless the distribution of the estimator may converge to the Normal distribution. The following is a simple definition which refers to the Normal distribution function, although there is no closed form for that function, and to the concept of convergence in distribution introduced earlier.

D13.9 An estimator $\hat{\theta}$ is *asymptotically Normal* if the distribution function of the estimator converges to the Normal distribution function.

13.3 Principles and methods for defining estimators

An estimator is a function of observable data. What function of the data should one choose for a given problem?

Sometimes this question is answered by specifying an arbitrary loss function for estimates, so that the estimator is chosen by minimizing this loss function. In some cases this can be done analytically, often by taking a derivative of the loss and minimizing it, but even in more difficult cases the function can usually be minimized using numerical methods; numerous algorithms are available to minimize arbitrary functions, although in many cases these will only find a local, rather than a global, optimum. For example, in cases in

which we are trying to fit or 'explain' a large number of data points with a model having a few parameters, we might take as our function to be minimized either the sum of squared discrepancies between the predicted and actual values for each data point, or the sum of the absolute values of these discrepancies.

However, there are also a number of general principles available for defining and computing estimators, and there are reasons for relying on such principles when we can. First, although in some cases it may be easy to come up with a sensible estimator by simple reasoning, in trickier circumstances having principles for defining estimators may be helpful, because simple reasoning does not point to any obvious estimator. As well, general properties of estimators defined according to a principle can sometimes be established, so that any example of an estimator defined according to such a principle will be known immediately to have certain features. Moreover, knowing that an estimator was defined according to some principle may make clear that there is a common structure with estimators used in another type of problem, possibly allowing the investigator to benefit from experience and knowledge that has arisen in other contexts.

In the rest of this section we will give a brief introduction to several of these principles, and we will see more examples of their application in later chapters, particularly when we review regression models.

13.4 Least squares (LS)

To understand least-squares estimation, it may be useful to return to the problem that originally motivated it: that of finding an approximate solution to a system of equations.

Consider the system

$$\begin{aligned} ax + by &= f \\ cx + dy &= g. \end{aligned}$$

Recall that a system of linear equations such as this may have no solution, for example, $2x + 3y = 5$; $4x + 6y = 11$: these statements cannot both be true because the latter statement implies, dividing by 2, that $2x + 3y = 5.5$, which contradicts the former; or one unique solution, for example, $2x+3y = 5$; $4x+5y = 11$). From the first equation we have $2x = 5-3y$ and so $4x = 10-6y$; substituting the latter into the second equation we have $(10 - 6y) + 5y = 11$, or $y = -1$; substituting this back into one of the original equations, we have $2x = 8$ or $4x = 16$, *i.e.* $x = 4$, or an infinite number of solutions (for example $2x + 3y = 5$; $4x + 6y = 10$: in this case the two equations contain the same information (*i.e.* are linearly dependent), so any pair (x, y) that solves the first will also solve the second.

Now consider the case in which there are more equations, still with two unknowns in each. There will now be no solution to the system, unless some of the equations are redundant (i.e. linearly dependent with other equations

in the system). For example, the system

$$\begin{aligned} 2x + 3y &= 5 \\ 4x + 5y &= 11 \\ 7x + 9y &= 15 \\ -x + 10y &= -11 \\ 7x - 7y &= 30 \end{aligned}$$

has no solution, which we can easily see because the solution to the first pair of equations, $(x, y) = (4, -1)$, does not solve any of the other equations.

Since no solution exists, we might decide to look for an approximate solution, using some criterion to define what constitutes a good approximation. We will begin by explicitly recognizing that the equations cannot be solved exactly, by writing in a set of terms to describe the discrepancy (or 'deviation' or 'error' or 'residual') in each equation:

$$\begin{aligned} 2x + 3y &= 5 + \varepsilon_1 \\ 4x + 5y &= 11 + \varepsilon_2 \\ 7x + 9y &= 15 + \varepsilon_3 \\ -x + 10y &= -11 + \varepsilon_4 \\ 7x - 7y &= 30 + \varepsilon_5. \end{aligned}$$

We have used the common notation ε_i to represent the discrepancy in the ith equation. Our aim now is to find a good approximation, which in general entails keeping the values of the ε_i terms as small as possible, but bearing in mind that a change in a value of x or y that lowers one of these discrepancy terms will in general raise another.

In order to come up with some solution, we might define a good approximation as follows: define the best approximating solution $(\hat{x}, \hat{y})$ as the pair of values that minimizes the sum of squared discrepancies, $\sum_{i=1}^{5} \varepsilon_i^2$. This leads to an estimator:

$$(\hat{x}, \hat{y}) = \operatorname{argmin} \left[\sum_{i=1}^{5} (\varepsilon_i(x, y))^2 \right],$$

where argmin $f(\theta)$ means 'the value of the argument, θ, at which the function $f(\theta)$ is minimized'. Here the argument is the pair (x, y), and in the last expression we have written ε_i as an explicit function of this argument.

In the example just given, the least-squares approximate solution (to five significant digits) is $(\hat{x}, \hat{y}) = (3.4005, -0.8220)$, and the reader may easily verify that other pairs of values lead to a higher sum of squared discrepancies. The method by which this solution was computed is given below, in Chapter 18, on linear regression. In a regression problem, the values x and

y are estimated weights on data series given by the vectors (2,4,7,-1,7) and (3,5,9,10, -7).

Minimizing the sum of squared discrepancies (errors) is a very widely applied technique. Because the criterion is quadratic, its derivative is linear, leading to a linear rule for finding the optimum (minimum); again see Chapter. This linear rule is not only convenient, but in some circumstances coincides with the form of estimator implied by other principles which have desirable general properties, in particular Maximum Likelihood.

13.5 Least absolute deviation (LAD)

LAD can be used to treat the same problem as above, but in this case, we replace the quadratic criterion with

$$(\hat{x}, \hat{y}) = \operatorname{argmin} \left[\sum_{i=1}^{5} |\varepsilon_i(x, y)| \right].$$

Changing the criterion in this way of course changes the relative weight of small and large deviations in defining the best approximation. With the LAD criterion, the 'badness' of a discrepancy changes linearly rather than quadratically with its magnitude, so that, for example, an error of 10 is only 5 times as bad as an error of 2, rather than 25 times as bad as the least-squares criterion would imply. Which of these is preferable will of course depend upon the problem and the person using the method. However, there is an additional important difference: the LAD criterion function, which is the absolute value function, is non–differentiable at zero (that is, the absolute value function has a corner or kink at zero, so that there is no unique tangent line. The minimum has to be found numerically rather than by deriving a simple equation for the estimator, and the fact that the criterion function is not everywhere differentiable means that more sophisticated mathematics is typically required in order to prove results concerning the properties of the estimator.[4]

13.6 Method of Moments (MoM)

A Method-of-Moments estimator is one in which unknown population moments are replaced, and estimated, by corresponding sample moments. and the descriptive statistics as examples of this

When we studied descriptive statistics in Chapter 3, we computed a number of functions of the data that we later saw as analogous to moments of the

[4] Notice by the way that the criterion argmin $\left[\sum_{i=1}^{n} \varepsilon_i(x, y)\right]$, without the absolute value, would not lead to good results: this implies that negative and positive errors cancel each other out, so that for example an estimator which leads to equal and opposite errors would be deemed just as desirable as one that picks the correct answer.

distribution. For example, the sample mean $\overline{X} = \sum_{i=1}^{n} x_i$ is the sample analogue of the population mean $\mathrm{E}(X)$, and can be taken as an estimate of the population mean, although we saw that other estimators such as a trimmed mean are also available. As the analogue of the population quantity, $\overline{X}$ is a Method of Moments estimator.

Some other descriptive statistics that we saw differ from the method-of-moments estimator. For example, the population variance is defined as $\mathrm{E}(X-\mu)^2$, and the Method of Moments estimator takes the analogous form, replacing μ with its sample analogue, $\overline{X}$. The Method-of-Moments estimator of the variance is therefore $n^{-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$, which differs from the unbiased estimator of the variance, which uses the factor $(n-1)$ in the denominator.

Method-of-Moments estimators (and a generalized version of the estimator) have frequently been used for estimating economic models in cases where the theory model suggests that a certain moment should take a value such as zero, in the population; for example, an economic theory may suggest that two quantities are independent, so that the expectation of the product of the two quantities should be zero. Imposing the condition that the expectation be zero ensures that the estimated model conforms with the assumed economic theory, and if this condition is then used to estimate other parameters, then those results also have been obtained by using and imposing the information implied by the economic theory. Whether this is desirable or not depends upon what one is trying to achieve: exploring the implications of the economic theory, versus exploring the content of the data set with minimal constraint.

13.7 Maximum Likelihood (ML)

Like LS, Maximum Likelihood is a very widely applied principle, and in fact in some important cases they lead to the same estimator.

Consider the density function (or probability function if discrete) corresponding with a random variable X. Although we often suppress the explicit dependence of a density function on the parameter vector, to write $f(X)$, a more complete notation for the density function is $f(X,\theta)$, where θ is the vector of parameters of the distribution; for example, if the distribution is Normal, the parameter vector $\theta = (\mu, \sigma^2)$ consists of the mean and variance parameters that characterize this distribution.

Earlier, we thought of this density, given knowledge of θ, as describing to us the values of X that are relatively likely or unlikely to arise. For example, if the distribution is Normal and $\theta = (\mu, \sigma^2) = (2, 10)$, then approximately 95% of the values of Xthat we will observe live in the interval $2 \pm 1.96\sqrt{10}$, since $\sqrt{10}$ is the standard deviation.

In using the density function in this way, we are taking the parameters as given, and asking where the observations are likely to arise. Another way of using the same information from the density would be to take a set of X values as given, and ask where θ is likely to lie: that is, we could try to deduce what the parameters must be, given a set of observations. More precisely, we

could ask which parameter values in a given density are most likely to have led to the observed sample of data.

Recall that if we have a density function, then we can look for the most likely region in which to find data points by looking for the interval where the density is highest. Analogously, if we take the data as given and look for the most likely region in which to find a parameter, we can again look for the interval in which the density – called likelihood when we view the data as fixed and the parameter vector as changeable – is highest. That is, we can maximize this likelihood.

The ML estimator is then.

$$\hat{\theta} = \text{argmax } \mathcal{L}(X,\theta),$$

where $\mathcal{L}(X,\theta) \equiv f(X,\theta)$, but now interpreted in such a way that X represents a fixed set of n observations, and θ is varied.

Actually obtaining an estimator from the ML principle involves, conceptually, two steps. First, one needs to determine the likelihood function of the observations, using the assumed likelihood, which is a function of unknown parameters. This needs to be maximized over values of the parameters, in order to obtain estimates of them corresponding to the maximum of this likelihood. Sometimes this can be done analytically, but often the optimization is numerical, using a one of a number of well-known computational algorithms.

An important case in which the optimization may be straightforward analytically is that of a Normal likelihood function, so that a quadratic function of the observations arises, leading to a linear derivative. Setting a linear derivative to zero gives a linear formula for computing the ML parameter estimates. In some cases, as for example in obtaining parameter estimates of a linear regression model (again Chapter—- below), the ML and LS criteria result in the same linear expression for estimates, and so identical estimated parameters are given by the two methods.

Of course, in maximizing a likelihood function, we assume that we know what that function is, i.e., we assume that we know the true density of the data. In practical examples that will typically not be the case, but we may decide to use this estimation method with an assumed likelihood function which is believed to be a good approximation to the true likelihood. Estimation using the computational method of Maximum Likelihood, but where the assumed likelihood function is not identical to the true likelihood, is commonly called quasi-Maximum Likelihood (QML). For example, we may use the Normal density as a likelihood function, in a case where the unknown true density is also symmetric, but has thicker tails than the Normal. In this case we would obtain QML estimates.

Proof of Theorem 13.2.1.

Let the true value of a parameter θ for a given DGP be denoted by θ_0. For an estimator $\hat{\theta}$ of θ, denote $\mathrm{E}[\hat{\theta}]$ by $\overline{\theta}$. Then we wish to show that

$$\mathrm{E}[(\hat{\theta} - \theta_0)^2] = \big(\mathrm{E}[\hat{\theta} - \theta_0]\big)^2 + \mathrm{E}\big[(\hat{\theta} - \overline{\theta})^2\big]. \tag{13.8.1}$$

By expanding the square, we see that

$$\begin{aligned}(\hat{\theta} - \theta_0)^2 &= \big(\hat{\theta} - \overline{\theta} - (\theta_0 - \overline{\theta})\big)^2 \\ &= (\hat{\theta} - \overline{\theta})^2 + (\overline{\theta} - \theta_0)^2 - 2(\hat{\theta} - \overline{\theta})(\overline{\theta} - \theta_0).\end{aligned} \tag{13.8.2}$$

The expectation of the last term above is zero, because both $\overline{\theta}$ and θ_0 are non-random, and by definition $\mathrm{E}(\hat{\theta} - \overline{\theta}) = \overline{\theta} - \overline{\theta} = 0$. By taking the expectation of (13.8.2), it follows that (13.8.1) holds. ∎

Chapter 14

Interval estimators and confidence intervals

Statistical answers to empirical questions do not involve certainties: they are estimates, probabilities, ranges within which the true answer might lie, and so on. When we obtain a point estimate of something, we generally want further information to go along with it: how accurate is this estimate likely to be? For example, if we estimate a price elasticity of demand for gasoline to be -0.5 (in a particular place, at a particular historical time), we will generally also want to know whether that estimate is likely to be accurate to within, say, ± 0.1 or ± 0.4. The former interval gives us much clearer information about what is likely to happen in response to an increase in the gasoline tax. Similarly, if we estimate that 48% of voters will vote 'yes' to a referendum question, we will have a much better idea of the likely outcome if we are highly confident that the answer will lie within ± 0.01 of this value than if we can only be confident of lying within ± 0.05. Of course, in a case like this we can directly estimate the probability that the referendum result will be positive using the known (binomial, asymptotically Normal) distribution of the point estimate.

In general, a complete answer to a point estimation problem involves not only the estimate, but a measure of the uncertainty associated with that estimate, or alternatively a range within which the correct answer to the problem will probably (according to a formal computation) lie.

We have seen examples in previous chapters in which we could calculate the probability that a true value lies within a certain interval around an estimate. The present chapter will apply methods given in previous chapters on sampling distributions, and will review and extend our treatment of the computation of confidence intervals for standard point estimation problems, where we can work with standard distributions given to us by statistical results such as a central limit theorem. We will also give further examples of computations of confidence intervals from empirical distributions of data or simulated statistics.

14.1 Definitions

D14.1 An *interval estimator* is a function of the data that provides estimates of lower and upper bounds A and B such that a quantity of interest (a parameter) lies in the interval $[A, B]$ with a given probability p.

D14.2 A *confidence interval* for a parameter β is an interval $[A, B]$ on the real line such that the probability that $[A, B]$ contains β is $1 - \alpha$.

Typical values of α are small, such as 0.01, so that the probability that the interval contains the true value is close to 1.

Where we have more than one parameter to describe, we may estimate a confidence region: that is, a region of possibly more than one dimension such that the probability that each of a set of parameters lies within the region is $1 - \alpha$.

14.2 Obtaining confidence intervals: two examples

We can think of ourselves as following a few simple steps in order to obtain a confidence interval for a given value. First, we need to obtain an estimate, a point estimate in this case, for the particular value. Next, we need to know the distribution that applies, at least approximately, to this estimate: this will typically be the step that requires the most sophistication. For many problems, however, we will be able to interpret our estimate as the mean of something that possesses at least two moments, so that we will be able to rely on a central limit theorem for this distribution. (In other cases, another distribution may apply, or the form of distribution may be unknown and we will have to use a computer simulation.) Finally, we need to know the parameters of this distribution: for example, if we are dealing with an asymptotically normal distribution, we will need to know the variance; if we are dealing with a χ^2 distribution, we will need to know the degrees of freedom, which will again allow us to determine how much of the distribution lies between particular bounds.

With a point estimate, a distribution and estimated values of the parameters of this distribution, we typically have the information that we need to determine how far on either side of the point estimate we need to draw our boundaries in order to have a given probability, such as 95% or 99%, that the true value lies in our interval.

Next we have two examples of confidence intervals for the difference in means of two random variables.

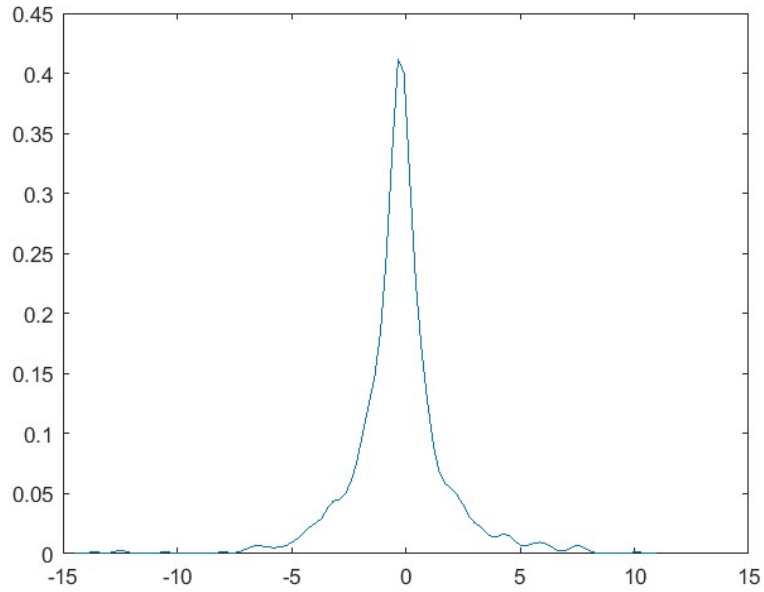
14.2.1 *The difference of two means: matched pairs*

One of the most common problems that we encounter is determining whether two variables are genuinely different from one another, and in particular, whether the expectations of the two variables are the same or not. Two

medical treatments for a given condition, two different types of education or training program, salaries offered to two different types of worker: are they genuinely different on average, or are the differences that we see just sampling error? We will now look at this problem by constructing a confidence interval for the difference between two variables.

For the first example, let us consider obtaining a confidence interval for the difference in the expectations of two variables. We will begin with the relatively straightforward case of pairs of observations, so that we can directly compute the difference in each case. For example, pairs of plants in a greenhouse may be given one of two fertilizers, and their growth measured over the following months. For each pair, we can directly measure the difference in growth. The following example is based on simulated data on amounts of growth for each of 1000 pairs of plants. Note that the growth is non-negative, and strongly right-skewed, so that the data are clearly not Normal. Figure 14.4.1 plots the differences in growth between the two plants in each of 1000 cases.

Figure 14.4.1
Estimated density of differences $X_1 - X_2$
1000 data points



One question that we would naturally want to answer is whether there is any difference in the efficacy of the two types of fertilizer, and if so we would

like to have an idea of how big that difference is; in other words, we want to obtain a confidence interval for the difference in growth for plants treated with fertilizer 1 versus those treated with 2.

Using these thousand data points, we compute a mean difference of -0.281, with a standard error of this estimated mean (the square root of the estimated variance of the differences divided by n) of 0.0657. We do not know the distribution of the data in this case. However, we are looking for the distribution of the mean of the data, which we can approximate using a central limit theorem. Given the conditions required to apply a central limit theorem, and in particular that the first two moments exist,[1] we have

$$\overline{d} \quad \xrightarrow{D} \quad N(\mu, \sigma^2/n),$$

where $\overline{d}$ is the estimated (or sample) mean of the difference; $n = 1000$; μ is the true mean of the difference, and σ^2 is the true variance of the difference (so that σ^2/n is the variance of the mean of the difference, where n is the sample size). Given this approximate distribution from a central limit theorem, we can construct an approximate confidence interval using the Normal distribution as in earlier chapters: using the familiar value $z_{\alpha/2} = 1.96$ from the standard Normal distribution and replacing σ by its estimate, we have

$$P(\mu - 1.96(0.0657) < \overline{d} < \mu + 1.96(0.0657)) \simeq 0.95,$$

or, re-arranging and substituting $\overline{d} = -0.281$,

$$P(-0.281 - 1.96(0.0657) < \mu < -0.281 + 1.96(0.0657)) \simeq 0.95,$$

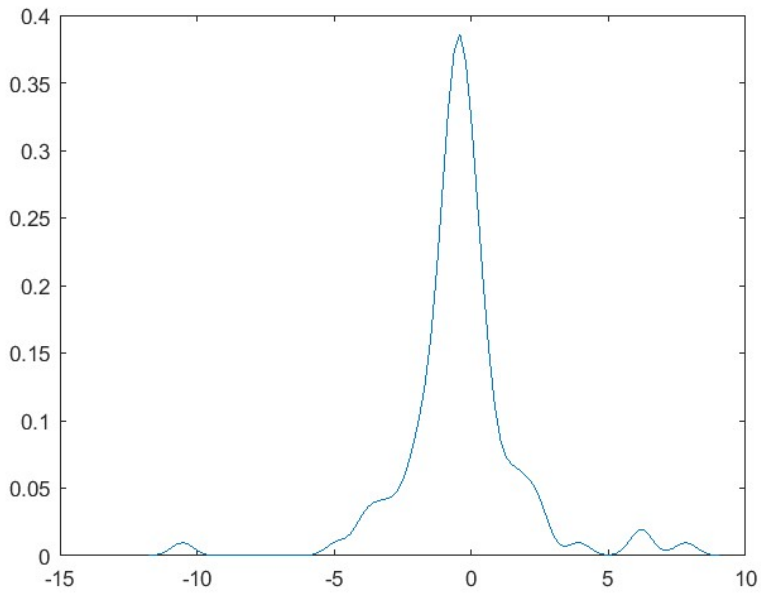
or $P(-0.410 < \mu < -0.152) \simeq 0.95$.

Notice that the value 0 is not in this confidence interval; we can be about 95% confident (in fact more than that, since 0 is well away from the boundary) that 0 is not the average difference between these two series. We have established in other words that we can be quite confident that there is a genuine difference in the means and that the observed difference is not simply due to sampling error. Moreover, it is the second random variable (treatment with fertilizer 2) that tends to be higher: $X_1 - X_2$ is on average negative. (This is the same type of computation that we will perform later in testing hypotheses: here, we might test the hypothesis that the true difference is zero; it turns out that we can be at least 95% confident that that is not so.

To illustrate the effect of sample size and also to make a point about what we can conclude when a value of interest *is* in a confidence interval, let's

[1] We can in principle check this by estimating the tail index, but it is in most cases simply assumed to be true because it is true for a very wide range of distributions.

FIGURE 14.4.2
Estimated density of differences $X_1 - X_2$
Subsample of 100 data points



perform this same exercise again, but with only the first 100 sample points from this data set. The estimated density of the difference $X_1 - X_2$ from the first hundred sample points follows.

The estimated density is similar, but of course with only 100 sample points, fewer extreme values appear and so the tails of the estimated density are not as wide. The corresponding estimates on the $n = 100$ sample are $\overline{d} = -0.366$ and $\sqrt{s^2/n} = 0.214$. The estimated confidence interval becomes

$$P(-0.366 - 1.96(0.214) < \mu < -0.366 + 1.96(0.214)) \simeq 0.95,$$

or $P(-0.785 < \mu < 0.053) \simeq 0.95$.

Although on the smaller sample the estimated mean difference was actually greater, 0 would not have been in the confidence interval on that sample: the estimated variance of the sample mean is larger (the square root of the sample size differs by the square root of 10, or about 3.16, so apart from sampling variation the standard error of the mean is bigger by this factor). On a sample of 100 points we would not have been highly confident that the difference is genuinely non-zero; on a sample of 1000 points, we could be.

This example illustrates a couple of points that are applicable widely. First, if a difference or other effect is present, it may nonetheless not be detectable in a small sample size. Because this is always a possibility, failing to detect an effect such as a difference does not prove that there is no effect. If someone had looked at the smaller sample and said, 'See? There is no mean difference between the two samples', that would be incorrect reasoning. It would be true to say that we cannot be confident that there is a difference, or to say that a zero difference is in our confidence interval, but it would not have been correct to conclude that the data have established that the difference is zero. (Of course, there is a whole continuum of values that are in the confidence interval.) That one cannot find strong evidence against a thing does not prove that the thing is true. We will return to this point below when we study hypothesis testing.

Second, the strength of an effect, or in this case the size of the difference, affects the number of sample points that we will typically require to detect it. In this example, the difference in mean between the two samples was small, and we could only be confident that the difference is non-zero once we had obtained a fairly a large number of sample points. Had the effect been much larger, we would have been able to detect it in a smaller sample size. In general, the more subtle the effect that we are trying to detect, the more sample points we will typically need in order to be confident that it is present.

14.2.2 *The difference of two means: independent samples*

In the example given above, we could compute directly a set of differences between the two random variables: the differences become a new random variable, and we simply computed the sample mean and standard error of that random variable as inputs to a formula obtained from a central limit theorem, in order to obtain a confidence interval for the mean of the differences. In many cases however we will have two samples of data which are not matched, and not even of the same size, and so it will not be possible to compute a series of differences to make calculations on directly. We can nonetheless again compute a confidence interval for the difference in the means of the two samples, using the expression for the variance of the linear combination of random variables, as long as it is reasonable to suppose that the two samples are statistically independent.

For a test statistic based on the central limit theorem, we need to estimate the mean and variance of the mean. Recall that $E(X_1 - X_2) = E(X_1) - E(X_2)$: the mean difference between the two can be computed by taking the difference of the two means individually. In the previous section we were able to create the random variable $(X_1 - X_2)$ and estimated sample mean directly; with unmatched samples, we can nonetheless estimate the sample mean for each random variable and subtract one from the other. So computing the

estimate $\overline{d}$ (as $\overline{X}_1 - \overline{X}_2$) is still straightforward, even though we don't have a sequence of differences.

Obtaining the estimated variance or standard error of the difference requires a little more reasoning. Recall however that for two random variables X_1 and X_2, we can compute the variance of a linear function of the two variables as $\mathrm{Var}(aX_1+bX_2) = a^2\,\mathrm{Var}(X_1)+b^2\,\mathrm{Var}(X_2)+2ab\,\mathrm{cov}(X_1, X_2)$. Without a sequence of pairs from the two distributions, it may not be possible to estimate the covariance term. But if the circumstances are such that it seems reasonable to assume that the two variables are independent, then this covariance is zero. In that case, specializing the formula to the difference of the means, we have $\mathrm{Var}(\overline{X}_1 - \overline{X}_2) = \mathrm{Var}(\overline{X}_1) + \mathrm{Var}(\overline{X}_2) = \mathrm{Var}(X_1)/n_1 + \mathrm{Var}(X_2)/n_2$, where n_i is the sample size for variable $i, i = 1, 2$.

Here is an example on simulated data with $n_1 = 100$ and $n_2 = 1000$. We compute $\overline{X}_1 = 0.969$, $\overline{X}_2 = 1.192$, and so $\overline{d} = -0.223$; next, $\mathrm{Var}(\overline{X}_1) = 0.0135, \mathrm{Var}(\overline{X}_2) = 0.0021$ and so the standard error of $\overline{d}$, that is the square root of the expression given above for the variance of the difference, is 0.125. The corresponding 95% confidence interval for the difference in the means of the two series is then $-0.223 \pm 1.96 \times 0.125$, or [-0.468, 0.022].

Notice that the variable for which the sample size is smaller makes the larger contribution to the variance, or standard error of the difference: that variable is less precisely estimated, so more of the uncertainty about the difference stems from that variable. Imagine for example that we increased the second sample size from 1000 to 1 million or 1 billion: we would get increasingly precise estimates of the mean of the second random variable, but after a while increasing the sample size will have very little effect on the variance of the difference: in the limit, where we know the mean of the second random variable exactly, there will still be uncertainty about the mean of the difference, because the mean of the first random variable is uncertain.

Chapter 15

Hypothesis testing

In the previous chapter we discussed confidence intervals, and mentioned a link to *hypothesis testing*, which is the main topic of this chapter.

15.1 Hypothesis tests

When we conduct hypothesis tests, we must do so in the context of a model. The hypotheses considered in this chapter are about a parameter or parameters of the model. As with the parameter estimators considered in Chapter 13, we work with a data set, consisting of observations on a dependent variable and possibly some explanatory variables.

The very simplest sort of hypothesis test concerns the (population) mean from which a *random sample* has been drawn. By saying that the sample is random, we mean that the observations are independent and identically distributed (IID), and are realizations drawn from some underlying distribution. The term "population mean", borrowed from biostatistics, here refers simply to the expectation of that distribution.

Suppose that we wish to test the hypothesis that the expectation is equal to some value that we specify. A suitable model for this test is the following regression model

$$y_t = \beta + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \tag{15.1.1}$$

where y_t is an observation on the dependent variable, β is the expectation of each of the y_t, and σ^2 is the variance of the disturbance term u_t. Let β_0 be the specified value of the expectation, so that we can express the hypothesis to be tested as $\beta = \beta_0$.

According to the idea behind least squares estimation, the estimator $\hat{\beta}$ minimises the sum of the squared discrepancies $\sum_{i=1}^n (y_t - \beta)^2$ with respect to β. It is easy to see that the least-squares estimator of β is just the sample mean. If we denote it by $\hat{\beta}$, then it follows that, for a sample of size n,

$$\hat{\beta} = \frac{1}{n}\sum_{t=1}^{n} y_t \qquad \text{and} \qquad \text{Var}(\hat{\beta}) = \frac{1}{n}\sigma^2. \tag{15.1.2}$$

The hypothesis to be tested is called, for historical reasons, the *null hypothesis*. It is often given the label H_0 for short. In order to test H_0, we

need a *test statistic*, which is a random variable that has a known distribution when the null hypothesis is true and some other distribution when the null hypothesis is false. If the value of this test statistic is one that might frequently be encountered by chance under the null hypothesis, then the test provides no evidence against the null. On the other hand, if the value of the test statistic is an extreme one that would rarely be encountered by chance under the null, then the test does provide evidence against the null. If this evidence is sufficiently convincing, we may decide to **reject** the null hypothesis that $\beta = \beta_0$.

Analogously to (12.1.1), we can use as our test statistic the expression

$$\tau = \frac{\hat{\beta} - \beta_0}{\sqrt{s^2/n}}, \tag{15.1.3}$$

where, as in Chapter 3,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{\beta})^2.$$

We have seen that, asymptotically, τ has the N(0,1) distribution: $\tau \xrightarrow{D} \text{N}(0,1)$. This is exactly how we derived the confidence interval (12.1.2) for a sample mean.

For every null hypothesis there is, at least implicitly, an *alternative hypothesis*, which is often given the label H_1. The alternative hypothesis is what we are testing the null against. Note that, if we consider the model that results from imposing the condition of the null hypothesis on the model (15.1.1), we get

$$y_t - \beta_0 = u_t, \quad u_t \sim \text{IID}(0, \sigma^2).$$

The parameter β does not appear in this model; rather it appears only in the model (15.1.1), in which the null hypothesis is not imposed. In this case, the model (15.1.1) represents the alternative hypothesis, which can be thought of as providing a framework in which the null hypothesis can be expressed as a restriction on one or more parameters, here just $\beta = \beta_0$. As important as the fact that τ has as asymptotic $\text{N}(0,1)$ distribution under the null is the fact that τ does *not* have this distribution under the alternative. Suppose that β takes on some other value, say β_1. Then it is clear that $\hat{\beta} = \beta_1 + \hat{\gamma}$, where $\hat{\gamma}$ has expectation 0 and variance σ^2/n, and is asymptotically Normal. We find from (15.1.3) that

$$\tau \xrightarrow{D} \text{N}(\lambda, 1), \quad \text{with} \quad \lambda = \frac{n^{1/2}}{\sigma}(\beta_1 - \beta_0). \tag{15.1.4}$$

The expectation λ is called the *non-centrality parameter*, or NCP, of the distribution of τ. Provided n is large enough, we would expect λ to be large

and positive if $\beta_1 > \beta_0$ and large and negative if $\beta_1 < \beta_0$. Thus we reject the null hypothesis whenever τ is sufficiently far from 0. Just how we can decide what "sufficiently far" means will be discussed shortly.

If we want to test the null that $\beta = \beta_0$ against the alternative that $\beta \neq \beta_0$, we must perform a *two-tailed test* and reject the null whenever the absolute value of τ is sufficiently large. If instead we were interested in testing the null hypothesis that $\beta \leq \beta_0$ against the alternative that $\beta > \beta_0$, we would perform a *one-tailed test* and reject the null whenever τ was sufficiently large and positive. In general, tests of equality restrictions are two-tailed tests, and tests of inequality restrictions are one-tailed tests.

15.2 Rejection by a test

Since τ is a random variable that can, in principle, take on any value on the real line, no value of τ is absolutely incompatible with the null hypothesis, and so we can never be absolutely certain that the null hypothesis is false. One way to deal with this situation is to decide in advance on a *rejection rule*, according to which we choose to reject the null hypothesis if and only if the value of τ falls into the *rejection region* of the rule. For two-tailed tests, the appropriate rejection region is the union of two sets, one containing all values of τ greater than some positive value, the other all values of τ less than some negative value. For a one-tailed test, the rejection region would consist of just one set, containing either sufficiently positive or sufficiently negative values of τ, according to the sign of the inequality we wish to test.

A test statistic combined with a rejection rule is generally simply called a *test*. A test returns a binary result, namely, reject or do-not-reject. We can never reach a conclusion that a null hypothesis is true on the basis of statistical evidence, and so our conclusion if a test fails to reject must simply be that the test provides no evidence against the null. Other tests, or other data sets, may well provide strong evidence against it.

If the test incorrectly leads us to reject a null hypothesis that is true, we are said to make a *Type I error*. The probability of making such an error is, by construction, the probability, *under the null hypothesis*, that τ falls into the rejection region. A property of any given test is its *significance level*, or just *level*, and it is defined as the probability, under the null, of making a Type I error, that is, the probability of rejecting the null when it is true. A common notation for this is α. Like all probabilities, α is a number between 0 and 1, although, in practice, it is generally chosen to be much closer to 0 than 1. Popular values of α include .05 and .01.

In order to construct the rejection region for a test at level α based on the test statistic τ, the first step is to calculate the *critical value* associated with the level α. We begin with the simplest case, which is when we want to test an inequality restriction of the form $\beta \geq \beta_0$. Evidence against this null is provided by a value of τ that is negative and large enough in absolute value. The rejection region is thus an infinite interval containing everything to the

left of the critical value appropriate for level α, say c_α. In order to attain this level, the probability under the null of a realization in this interval must be α. We continue to suppose that τ is asymptotically Normal under the null, and so the critical value c_α has to satisfy the equation

$$\Phi(c_\alpha) = \alpha; \tag{15.2.1}$$

recall that Φ denotes the CDF of the standard Normal distribution. We can solve (15.2.1) in terms of the inverse function Φ^{-1}, and we find that

$$c_\alpha = \Phi^{-1}(\alpha).$$

This means, of course, that c_α is just the α-quantile of the standard Normal distribution. Note that, for $\alpha < 1/2$, c_α, being in the left-hand tail of the distribution, is negative.

In order to test an equality restriction, we use a two-tailed test. This means that we need both a negative and a positive critical value. In this case, the commonest sort of test is an *equal-tail test*, with the same probability mass in the rejection regions on the left and on the right. For level α, then, that means that we want a probabilty mass of $\alpha/2$ in both tails. For the left-hand tail, we must have

$$\Phi(-c_\alpha) = \alpha/2,$$

We know that the left-tail critical value is negative, hence the minus sign. On account of the symmetry of the N(0,1) distribution, the right-tail critical value is just $+c_\alpha$. We could equally well have defined c_α by the equation

$$\Phi(c_\alpha) = 1 - \alpha/2, \tag{15.2.2}$$

which allocates a probability mass of $\alpha/2$ to the right of c_α. Solving equation (15.2.2) for c_α gives

$$c_\alpha = \Phi^{-1}(1 - \alpha/2). \tag{15.2.3}$$

Clearly, the critical value c_α increases as α approaches 0. As an example, when $\alpha = .05$, we see from equation (15.2.3) that the critical value for a two-tailed test is $\Phi^{-1}(.975) = 1.96$. We would reject the null at the .05 level whenever the observed absolute value of the test statistic exceeds 1.96.

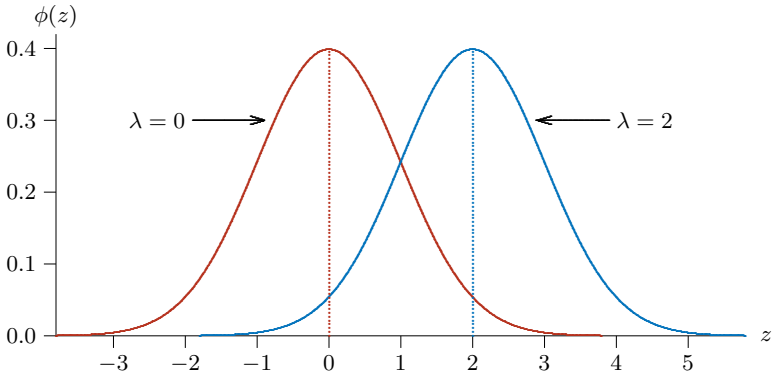
As with the confidence intervals treated in the previous chapter, the rejection probabilities of the tests we have looked at in this chapter are approximate, since they are based on the asymptotic distribution of the statistic τ, not its distribution for a finite sample size n. Thus we need to draw a distinction between the *nominal level* of the test, that is, the probability of making a Type I error according to whatever approximate distribution we are using to determine the rejection region, and the actual *rejection probability*, which may differ greatly from the nominal level. The rejection probability is generally unknowable in practice, because it typically depends on unknown features of the DGP.[1]

[1] Another term that often arises in the discussion of hypothesis testing is the *size* of a test. Technically, this is the supremum of the rejection probability over all

15.2.1 *Power of a test*

The probability that a test rejects the null is called the *power* of the test. If the data are generated by a DGP that satisfies the null hypothesis, the power of an exact test is equal to its level. In general, power depends on precisely how the data were generated and on the sample size. We can see from (15.1.4) that the distribution of τ is entirely determined by the value of the non-centrality parameter λ, with $\lambda = 0$ under the null, and that the value of λ depends on the parameters of the DGP. In this example, λ is proportional to $\beta_1 - \beta_0$ and to the square root of the sample size, and it is inversely proportional to σ.

FIGURE 15.2.1
THE NORMAL DISTRIBUTION CENTERED AND UNCENTERED



Values of λ different from 0 move the probability mass of the $N(\lambda, 1)$ distribution away from the centre of the $N(0, 1)$ distribution and into its tails. This can be seen in Figure 15.2.1, which graphs the $N(0, 1)$ density and the $N(\lambda, 1)$ density for $\lambda = 2$. The second density places much more probability than the first on values of τ greater than 2. Thus, if the rejection region for our test were the interval from 2 to $+\infty$, there would be a much higher probability in that region for $\lambda = 2$ than for $\lambda = 0$. Therefore, we would reject the null hypothesis more often when the null hypothesis is false, with $\lambda = 2$, than when it is true, with $\lambda = 0$.

Mistakenly failing to reject a false null hypothesis is called making a *Type II error*. The probability of making such a mistake is equal to 1 minus the power of the test. It is not hard to see that, quite generally, the probability of rejecting the null with a two-tailed test based on τ increases with the absolute value of λ. Consequently, the power of such a test increases as $\beta_1 - \beta_0$ increases, as σ decreases, and as the sample size increases.

DGPs that satisfy the null hypothesis. For an exact test, the size equals the level. For an approximate test, the size is typically difficult or impossible to calculate. It is often, but by no means always, greater than the nominal level of the test.

15.3 *P*–VALUES

As we have defined it, the result of a test is yes or no: Reject or do not reject. The result depends on the chosen level, and it is to that extent subjective: different people can be expected to have different tolerances for Type I error. A more sophisticated approach to deciding whether or not to reject the null hypothesis is to calculate the *P-value*, or *marginal significance level*, associated with a test statistic. The P-value for the statistic τ is defined as the greatest level for which a test based on τ fails to reject the null. Equivalently, at least if the statistic τ has a continuous distribution, it is the smallest level for which the test rejects. Thus, the test rejects for all levels greater than the P-value, and it fails to reject for all levels smaller than the P-value. The P-value is given as a deterministic function of the (random) statistic τ by finding the level for which τ is equal to the critical value for that level. Therefore, if the P-value determined by τ is denoted $p(\tau)$, we must be prepared to accept a probability $p(\tau)$ of Type I error if we choose to reject the null. But the P-value itself, being a purely objective quantity, is the same for everyone, and it allows different people to draw their own subjective conclusions.

The consequences of the definition of the P-value are a little trickier for the equal-tail test we have been discussing. We find that

$$p(\tau) = 2\big(1 - \Phi(|\tau|)\big). \tag{15.3.1}$$

To see this, note that the test based on τ rejects at level α if and only if $|\tau| > c_\alpha$. This inequality is equivalent to $\Phi(|\tau|) > \Phi(c_\alpha)$, because $\Phi(\cdot)$ is a strictly increasing function. Further, for this equal-tail test, $\Phi(c_\alpha) = 1 - \alpha/2$, by equation (15.3.1). The smallest value of α for which the inequality holds is thus obtained by solving the equation

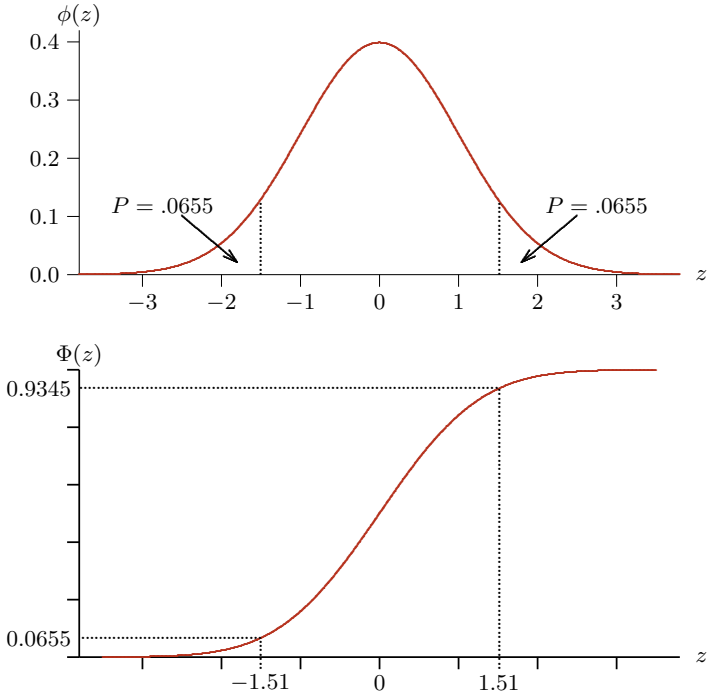
$$\Phi(|\tau|) = 1 - \alpha/2,$$

and the solution is easily seen to be the right-hand side of equation (15.3.1).

One advantage of using P-values is that they preserve all the information conveyed by a test statistic, while presenting it in a way that is directly interpretable. For example, the test statistics 2.02 and 5.77 would both lead us to reject the null at the .05 level using a two-tailed test. The second of these obviously provides more evidence against the null than does the first, but it is only after they are converted to P-values that the magnitude of the difference becomes apparent. The P-value for the first test statistic is .0434, while the P-value for the second is 7.93×10^{-9}, an extremely small number.

Computing a P-value transforms τ from a random variable with the $N(0, 1)$ distribution into a new random variable $p(\tau)$ with the uniform $U(0, 1)$ distribution. Readers are invited to prove this fact. It is quite possible to think

Figure 15.3.1
P-values for a two-tailed test



of $p(\tau)$ as a test statistic, of which the observed realization is $p(\hat{\tau})$ whenever the realization of τ is $\hat{\tau}$. A test at level α rejects whenever $p(\hat{\tau}) < \alpha$. Note that the sign of this inequality is the opposite of that in the condition $|\hat{\tau}| > c_\alpha$. Generally, one rejects for *large* values of test statistics, but for *small* P-values.

Figure 15.3.1 illustrates how the test statistic τ is related to its P-value $p(\tau)$. Suppose that the value of the test statistic is 1.51. Then

$$\Pr(z > 1.51) = \Pr(z < -1.51) = .0655. \tag{15.3.2}$$

This implies, by equation (15.3.1), that the P-value for an equal-tail test based on z is .1310. The top panel of the figure illustrates (15.3.2) in terms of the standard normal density, and the bottom panel illustrates it in terms of the CDF. To avoid clutter, no critical values are shown on the figure, but it is clear that a test based on z does not reject at any level smaller than .131. From the figure, it is also easy to see that the P-value for a one-tailed test of the hypothesis that $\beta \leq \beta_0$ is .0655. This is just $\Pr(z > 1.51)$. Similarly, the P-value for a one-tailed test of the hypothesis that $\beta \geq \beta_0$ is $\Pr(z < 1.51) = .9345$.

Chapter 16

Matrix algebra

It is impossible to study econometrics beyond the most elementary level without using matrix algebra. Most readers are probably already quite familiar with matrix algebra. This section reviews some basic results that will be used throughout the book. It also shows how regression models can be written very compactly using matrix notation. More advanced material will be discussed in later chapters, as it is needed.

16.1 Basic definitions

An $n \times m$ *matrix* $\boldsymbol{A}$ is a rectangular array that consists of nm elements arranged in n rows and m columns. The name of the matrix is conventionally shown in boldface. A typical element of $\boldsymbol{A}$ might be denoted by either A_{ij} or a_{ij}, where $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The first subscript always indicates the row, and the second always indicates the column. It is sometimes necessary to show the elements of a matrix explicitly, in which case they are arrayed in rows and columns and surrounded by large brackets, as in

$$\boldsymbol{B} = \begin{bmatrix} 2 & 3 & 6 \\ 4 & 5 & 8 \end{bmatrix}.$$

Here $\boldsymbol{B}$ is a 2×3 matrix.

If a matrix has only one column or only one row, it is called a *vector*. There are two types of vectors, *column vectors* and *row vectors*. Since column vectors are more common than row vectors, a vector that is not specified to be a row vector is normally treated as a column vector. If a column vector has n elements, it may be referred to as an n-vector. Boldface is used to denote vectors as well as matrices. It is conventional to use uppercase letters for matrices and lowercase letters for column vectors. However, it is sometimes necessary to ignore this convention.

If a matrix has the same number of columns and rows, it is said to be *square*. A square matrix $\boldsymbol{A}$ is *symmetric* if $A_{ij} = A_{ji}$ for all i and j. Symmetric matrices occur very frequently in econometrics. A square matrix is said to be *diagonal* if $A_{ij} = 0$ for all $i \neq j$; in this case, the only nonzero entries are those on what is called the *principal diagonal*. Sometimes a square matrix has all zeros above or below the principal diagonal. Such a matrix is said to be *triangular*. If the nonzero elements are all above the diagonal, it is

said to be *upper-triangular*; if the nonzero elements are all below the diagonal, it is said to be *lower-triangular*. Here are some examples:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 4 \\ 2 & 3 & 6 \\ 4 & 6 & 5 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix} \qquad \boldsymbol{C} = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 2 & 0 \\ 5 & 2 & 6 \end{bmatrix}.$$

In this case, $\boldsymbol{A}$ is symmetric, $\boldsymbol{B}$ is diagonal, and $\boldsymbol{C}$ is lower-triangular.

The *transpose* of a matrix is obtained by interchanging its row and column subscripts. Thus the ij^{th} element of $\boldsymbol{A}$ becomes the ji^{th} element of its transpose, which is denoted $\boldsymbol{A}^\top$. Note that many authors use $\boldsymbol{A}'$ rather than $\boldsymbol{A}^\top$ to denote the transpose of $\boldsymbol{A}$. The transpose of a symmetric matrix is equal to the matrix itself. The transpose of a column vector is a row vector, and vice versa. Here are some examples:

$$\boldsymbol{A} = \begin{bmatrix} 2 & 5 & 7 \\ 3 & 8 & 4 \end{bmatrix} \quad \boldsymbol{A}^\top = \begin{bmatrix} 2 & 3 \\ 5 & 8 \\ 7 & 4 \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} \quad \boldsymbol{b}^\top = [\,2 \quad 4 \quad 6\,].$$

Note that a matrix $\boldsymbol{A}$ is symmetric if and only if $\boldsymbol{A} = \boldsymbol{A}^\top$.

16.2 Arithmetic Operations on Matrices

Addition and *subtraction* of matrices works exactly the way it does for scalars, with the proviso that matrices can be added or subtracted only if they are *conformable*. In the case of addition and subtraction, this just means that they must have the same dimensions, that is, the same number of rows and the same number of columns. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable, then a typical element of $\boldsymbol{A} + \boldsymbol{B}$ is simply $A_{ij} + B_{ij}$, and a typical element of $\boldsymbol{A} - \boldsymbol{B}$ is $A_{ij} - B_{ij}$.

Matrix multiplication actually involves both additions and multiplications. It is based on what is called the *inner product*, or *scalar product*, or sometimes *dot product* of two vectors. Suppose that $\boldsymbol{a}$ and $\boldsymbol{b}$ are n-vectors. Then their inner product is

$$\boldsymbol{a}^\top\boldsymbol{b} = \boldsymbol{b}^\top\boldsymbol{a} = \sum_{i=1}^{n} a_i b_i.$$

As the name suggests, this is just a scalar.

When two matrices are multiplied together, the ij^{th} element of the result is equal to the inner product of the i^{th} row of the first matrix with the j^{th} column of the second matrix. Thus, if $\boldsymbol{C} = \boldsymbol{AB}$,

$$C_{ij} = \sum_{k=1}^{m} A_{ik} B_{kj}. \tag{16.2.1}$$

For (16.2.1) to make sense, we must assume that $\boldsymbol{A}$ has m columns and that $\boldsymbol{B}$ has m rows. In general, if two matrices are to be conformable for multiplication, the first matrix must have as many columns as the second has rows. Further, as is clear from (16.2.1), the result has as many rows as the first matrix and as many columns as the second. One way to make this explicit is to write something like

$$\underset{n\times m}{\boldsymbol{A}} \quad \underset{m\times l}{\boldsymbol{B}} = \underset{n\times l}{\boldsymbol{C}}.$$

One rarely sees this type of notation in a book or journal article. However, it is often useful to employ it when doing calculations, in order to verify that the matrices being multiplied are indeed conformable and to derive the dimensions of their product.

The rules for multiplying matrices and vectors together are the same as the rules for multiplying matrices with each other; vectors are simply treated as matrices that have only one column or only one row. For instance, if we multiply an n-vector $\boldsymbol{a}$ by the transpose of an n-vector $\boldsymbol{b}$, we obtain what is called the *outer product* of the two vectors. The result, written as $\boldsymbol{ab}^\top$, is an $n \times n$ matrix with typical element $a_i b_j$.

Matrix multiplication is, in general, not commutative. The fact that it is possible to *premultiply* $\boldsymbol{B}$ by $\boldsymbol{A}$ does not imply that it is possible to *postmultiply* $\boldsymbol{B}$ by $\boldsymbol{A}$. In fact, it is easy to see that both operations are possible if and only if one of the matrix products is square, in which case the other matrix product is square also, although generally with different dimensions. Even when both operations are possible, $\boldsymbol{AB} \neq \boldsymbol{BA}$ except in special cases.

A special matrix that econometricians frequently make use of is $\mathbf{I}$, which denotes the *identity matrix*. It is a diagonal matrix with every diagonal element equal to 1. A subscript is sometimes used to indicate the number of rows and columns. Thus

$$\mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The identity matrix is so called because when it is either premultiplied or postmultiplied by any matrix, it leaves the latter unchanged. Thus, for any matrix $\boldsymbol{A}$, $\boldsymbol{AI} = \mathbf{I}\boldsymbol{A} = \boldsymbol{A}$, provided, of course, that the matrices are conformable for multiplication. It is easy to see why the identity matrix has this property. Recall that the only nonzero elements of $\mathbf{I}$ are equal to 1 and are on the principal diagonal. This fact can be expressed simply with the help of the symbol known as the *Kronecker delta*, written as δ_{ij}. The definition is

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{16.2.2}$$

The ij^{th} element of $\mathbf{I}$ is just δ_{ij}. By (16.2.1), the ij^{th} element of $\boldsymbol{AI}$ is

$$\sum_{k=1}^{m} A_{ik}\mathrm{I}_{kj} = \sum_{k=1}^{m} A_{ik}\delta_{kj} = A_{ij},$$

since all the terms in the sum over k vanish except that for which $k = j$.

A special vector that we frequently use in this book is $\boldsymbol{\iota}$. It denotes a column vector every element of which is 1. This special vector comes in handy whenever one wishes to sum the elements of another vector, because, for any n-vector $\boldsymbol{b}$,

$$\boldsymbol{\iota}^\top \boldsymbol{b} = \sum_{i=1}^{n} b_i. \tag{16.2.3}$$

Matrix multiplication and matrix addition interact in an intuitive way. It is easy to check from the definitions of the respective operations that the *distributive* properties hold. That is, assuming that the dimensions of the matrices are conformable for the various operations,

$$\begin{aligned} \boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) &= \boldsymbol{AB}+\boldsymbol{AC}, \text{ and} \\ (\boldsymbol{B}+\boldsymbol{C})\boldsymbol{A} &= \boldsymbol{BA}+\boldsymbol{CA}. \end{aligned}$$

In addition, both operations are *associative*, which means that

$$\begin{aligned} (\boldsymbol{A}+\boldsymbol{B})+\boldsymbol{C} &= \boldsymbol{A}+(\boldsymbol{B}+\boldsymbol{C}), \text{ and} \\ (\boldsymbol{AB})\boldsymbol{C} &= \boldsymbol{A}(\boldsymbol{BC}). \end{aligned}$$

The transpose of the product of two matrices is the product of the transposes of the matrices with the order reversed. Thus

$$(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top. \tag{16.2.4}$$

The reversal of the order is necessary if the transposed matrices are to be conformable for multiplication. The result (16.2.4) can be proved immediately by writing out the typical entries of both sides and checking that

$$(\boldsymbol{AB})^\top_{ij} = (\boldsymbol{AB})_{ji} = \sum_{k=1}^{m} A_{jk}B_{ki} = \sum_{k=1}^{m} (\boldsymbol{B}^\top)_{ik}(\boldsymbol{A}^\top)_{kj} = (\boldsymbol{B}^\top\boldsymbol{A}^\top)_{ij},$$

where m is the number of columns of $\boldsymbol{A}$ and the number of rows of $\boldsymbol{B}$. It is always possible to multiply a matrix by its own transpose: If $\boldsymbol{A}$ is $n \times m$, then $\boldsymbol{A}^\top$ is $m \times n$, $\boldsymbol{A}^\top\boldsymbol{A}$ is $m \times m$, and $\boldsymbol{AA}^\top$ is $n \times n$. It follows directly from (16.2.4) that both of these matrix products are symmetric:

$$\boldsymbol{A}^\top\boldsymbol{A} = (\boldsymbol{A}^\top\boldsymbol{A})^\top \quad \text{and} \quad \boldsymbol{AA}^\top = (\boldsymbol{AA}^\top)^\top.$$

It is frequently necessary to multiply a matrix, say $\boldsymbol{B}$, by a scalar, say α. *Multiplication by a scalar* works exactly the way one would expect: Every element of $\boldsymbol{B}$ is multiplied by α. Since multiplication by a scalar is commutative, we can write this either as $\alpha\boldsymbol{B}$ or as $\boldsymbol{B}\alpha$, but $\alpha\boldsymbol{B}$ is the more common notation.

A square matrix may or may not be *invertible*. If $\boldsymbol{A}$ is invertible, then it has an *inverse matrix* $\boldsymbol{A}^{-1}$ with the property that

$$\boldsymbol{AA}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \mathbf{I}.$$

If $\boldsymbol{A}$ is symmetric, then so is $\boldsymbol{A}^{-1}$. If $\boldsymbol{A}$ is triangular, then so is $\boldsymbol{A}^{-1}$. Except in certain special cases, it is not easy to calculate the inverse of a matrix by hand. One such special case is that of a diagonal matrix, say $\boldsymbol{D}$, with typical diagonal element D_{ii}. It is easy to verify that $\boldsymbol{D}^{-1}$ is also a diagonal matrix, with typical diagonal element D_{ii}^{-1}.

If an $n \times n$ square matrix $\boldsymbol{A}$ is invertible, then its *rank* is n. Such a matrix is said to have *full rank*. If a square matrix does not have full rank, and therefore is not invertible, it is said to be *singular*. If a square matrix is singular, its rank must be less than its dimension. If, by omitting j rows and j columns of $\boldsymbol{A}$, we can obtain a matrix $\boldsymbol{A}'$ that is invertible, and if j is the smallest number for which this is true, then the rank of $\boldsymbol{A}$ is $n - j$. More generally, for matrices that are not necessarily square, the rank is the largest number m for which an $m \times m$ nonsingular matrix can be constructed by omitting some rows and some columns from the original matrix.

Chapter 17

Linear regression

In Chapter 8, we introduced the conditional expectation. Since we do not normally know the full joint probability distribution of a set of variables, we cannot directly calculate the conditional expectation using Definition 8.5.1, but we noted that we can approximate it using, for example, a linear regression. We will now learn how to estimate linear regression models to accomplish this.

Of course, conditional expectations may be non-linear, and so linear regression will typically be considered as an approximation (by Taylor's theorem, a straight line provides a local approximation to a function, but whether this approximation will be adequate depends on the purpose for which it is used). We can to some extent handle non-linearity by using a linear function with non-linear terms, for example by including squared terms or logarithms of conditioning variables in a linear form.

This chapter deals with only the simplest case of a regression model in which a parametric linear form is assumed, and we need only estimate the parameters of this form to specify the conditional expectation function completely.

17.1 The least-squares (LS) criterion

Recall that in Chapter 13 we discussed some criteria by which to define a 'good' estimator, so that optimizing a criterion can give us a rule that will allow us in principle to pick the best estimator within some class. One of these criteria leads to the least-squares estimator. To draw the analogy to estimating a linear conditional expectation model, let us return to the example given in Chapter 8, of a model having the form

$$\mathrm{E}(Y|X_1, X_2) = a + bX_1 + cX_2,$$

where Y is income, X_1 is age and X_2 is years of formal education. The parameters of this linear model, a, b, c, are to be estimated.

Recognizing that this form will not fit the data perfectly, that is, that Y will not be exactly equal to its conditional expectation for each (or any) observation, we usually introduce a term describing the discrepancy (or error, or,

better, disturbance) between the observed Y and its conditional expectation as given in the equation above. We therefore write

$$Y = a + bX_1 + cX_2 + \varepsilon,$$

where ε is the symbol used here for this disturbance. Referring to each data point individually, instead of in this vector form, we could write

$$Y_i = a + bX_{1,i} + cX_{2,i} + \varepsilon_i, \quad i = 1, ...n,$$

for each observation i in a sample of n observations.

We would like to pick parameters, that is, values that can be adapted to fit the data while remaining in the context of the model that we have specified, so that the model fits well. One criterion by which to define what it means to fit well is that the ε_i's are as small as possible, in the sense of minimizing their sum of squares, $\sum_{i=1}^{n} \varepsilon_i^2$. Using a power of 2 (a quadratic) has two advantages: all discrepancies count as positive (we wouldn't consider that large negative errors in a model somehow offset large positive errors, and make it a good model: instead we'd like errors to be small in magnitude, regardless of whether they are positive or negative); and the derivative of the square gives a linear function, leading to a linear rule for minimizing this quantity. We could also use the sum of the absolute values of the errors as our criterion, leading to the least absolute deviations or LAD estimator, but this requires more sophisticated mathematics since the absolute value function is not differentiable at zero, and so we cannot use elementary calculus to obtain a simple formula for the estimator, as we can with least squares.

17.2 A simple regression with one or two variables

We consider estimation of a linear model of a variable Y, in order to obtain a conditional expectation $\mathrm{E}(Y|X)$, where X is a matrix of conditioning variables. We have n observations, and there are k separate variables in X, each of which also has n observations available. We could write the model as:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots \beta_k X_{k,i} + \varepsilon_i,$$

where the β_i, $i = 1, \ldots, k$, are parameters to be estimated, and ε_i is an unobservable disturbance (error) term, allowing for the fact that the model does not fit perfectly. In matrix notation, we can write this as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \tag{17.2.1}$$

where $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$ are of dimensions $n \times 1$, $n \times k$, $k \times 1$, and $n \times 1$ respectively. Note then that $\boldsymbol{X\beta}$ becomes of dimension $n \times 1$ also (an $n \times k$ matrix multiplied by a $k \times 1$ vector). For example, the vector of parameters is

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}.$$

We assume that $\boldsymbol{\varepsilon}$ has an expectation of zero, which in practice we can assure by including a constant (intercept) in the model, as we will see later.

Each column of the $\boldsymbol{X}$ matrix represents a different data series, and the n elements in that column are the observations on the data series. For example, in a model involving individual human subjects, the columns might represent age, sex, years of formal education, *etc*. Each row of the matrix would represent a particular individual, so reading across the row we have that individual's age, sex, years of formal education ...

Our aim again is to estimate this model in order to obtain a conditional expectation of $\boldsymbol{y}$ given the available $\boldsymbol{X}$ variables, recognizing that there may be no causal link between $\boldsymbol{X}$ and $\boldsymbol{y}$, and that we could also condition on other variables.

If we estimate the parameters $\boldsymbol{\beta}$, then the estimated conditional expectation becomes: $\widehat{\mathrm{E}(\boldsymbol{y}|\boldsymbol{X})} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, since $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$. (If we drop the matrix notation for a minute, this is equivalent to $\widehat{\mathrm{E}(y_i|X_i)} = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} + \ldots \hat{\beta}_k X_{k,i}$). We use the circumflex, $\widehat{\mathrm{E}(y_i|X_i)}$, to indicate that this is an estimated conditional expectation (although in practice people often omit this symbol).

Now we can ask how we should estimate the parameters $\boldsymbol{\beta}$, that is, the unknown weights in this linear approximation to the conditional expectation function. The most commonly applied method for this simple model is that of least squares, *i.e.*, one minimizes the sum of squared residuals (the estimated disturbances) by choice of $\boldsymbol{\beta}$.

Consider first a simple case that we can handle without matrices. Let $y_i = \beta x_i + e_i$, and let the estimated version of the model be $y_i = \hat{\beta} x_i + \hat{e}_i$. The 'residuals' are the $\hat{e}_i$, and the sum of squared residuals is $\sum_i (y_i - \hat{\beta}_i x_i)^2 = \sum_i [y_i^2 - 2\hat{\beta} x_i y_i + (\hat{\beta}_i x_i)^2]$. Taking the derivative with respect to $\hat{\beta}$ and setting to zero for an optimum, we have $-2\sum_i (x_i y_i) + 2\sum_i \hat{\beta} x_i^2 = 0$, or

$$\hat{\beta} = \sum_i (x_i y_i) / \sum_i (x_i^2). \tag{17.2.2}$$

Notice that we began with a quadratic criterion– the sum of *squared* errors– and so by taking a derivative, the square becomes a linear rule ($\mathrm{d}(x^2)/\mathrm{d}x = 2x$). Equation (17.2.2) is a linear rule for computing the coefficients that minimize the sum of squared residuals: the (ordinary) least squares estimator, or OLS.

The case above is a simple illustration, but we would rarely want to run a regression without a constant, since we will not want to force weight onto explanatory variables (that is, change the coefficients $\hat{\beta}_i$) simply to fit the mean of y. So let us consider now the slightly more elaborate case in which we have a constant as well as a single additional explanatory variable X.

Our model is now $y_i = \alpha + \beta x_i + e_i$, and the estimated version of the model is $y_i = \hat{\alpha} + \hat{\beta} x_i + \hat{e}_i$, and so we can write the residuals as $\hat{\varepsilon}_i = y_i - \hat{\alpha} - \hat{\beta} X_{1,i}$,

with sum of squared residuals

$$\begin{aligned} SSR &= \sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n \left[(y_i - \hat{\alpha} - \hat{\beta} X_{1,i}) \right]^2 \\ &= \sum_{i=1}^n y_i^2 - 2\hat{\alpha} \sum_{i=1}^n y_i - 2\hat{\beta} \sum_{i=1}^n x_i y_i + 2\hat{\alpha}\hat{\beta} \sum_{i=1}^n x_i + \sum_{i=1}^n \hat{\alpha}^2 + \hat{\beta}^2 \sum_{i=1}^n x_i^2. \end{aligned}$$

Now notice that $\sum_{i=1}^n x_i = n\overline{X}$ and $\sum_{i=1}^n y_i = n\overline{Y}$, and because $\hat{\alpha}^2$ is a fixed number (not indexed by i), $\sum_{i=1}^n \hat{\alpha}^2 = n\hat{\alpha}^2$. Therefore,

$$SSR = \sum_{i=1}^n y_i^2 - 2\hat{\alpha}(n\overline{Y}) - 2\hat{\beta} \sum_{i=1}^n x_i y_i + 2\hat{\alpha}\hat{\beta}(n\overline{X}) + n\hat{\alpha}^2 + \hat{\beta}^2 \sum_{i=1}^n x_i^2. \tag{17.2.3}$$

Next we find the partial derivatives so that we can set them to zero and find the optimum points.

$$\begin{aligned} 0 &= \partial SSR/\partial\hat{\alpha} = -2n\overline{Y} + 2n\hat{\beta}\overline{X} + 2n\hat{\alpha} \\ &\Rightarrow -\overline{Y} + \hat{\beta}\overline{X} + \hat{\alpha} = 0, \quad \text{and so} \\ \hat{\alpha} &= \overline{Y} - \hat{\beta}\overline{X}. \end{aligned} \tag{17.2.4}$$

It's slightly more work to get a formula for $\hat{\beta}$. Again we're going to find a partial derivative and set it to zero. We have

$$\begin{aligned} 0 =& \partial SSR/\partial\hat{\beta} = -2\sum_{i=1}^n x_i y_i + 2n\hat{\alpha}\overline{X} + 2\hat{\beta}\sum_{i=1}^n x_i^2 \\ &\Rightarrow \hat{\beta}\sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i - n\hat{\alpha}\overline{X}, \quad \text{and so} \\ \hat{\beta} &= \frac{\sum_{i=1}^n x_i y_i - n\hat{\alpha}\overline{X}}{\sum_{i=1}^n x_i^2}. \end{aligned} \tag{17.2.5}$$

Solve the two equations (17.2.4) and (17.2.5) in the two unknowns $\hat{\alpha}$ and $\hat{\beta}$ to get each as a function that involves only observables (functions of x_i and y_i).

As we move to three regressors, we will get more equations to solve, and the process quickly becomes too cumbersome. However we can derive the more general solution for the SSR-minimizing coefficients in equation (17.2.1) using matrix differentiation, yielding the solution

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{17.2.6}$$

Notice that equation (17.2.2) is a special case of this, that applies when $\boldsymbol{X}$ has only one column. In the next section we will derive this general result.

17.3 Multiple regression using matrix algebra

In order to derive the estimator just mentioned (for any number of regressors), it's important to begin by representing the data in a standard form; the answers will correspond with this standard form and will be readily interpretable. To return to equation (17.2.1) above, we represent the data on the 'dependent' variable $\boldsymbol{y}$ as:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

where again n is the sample size. Note that the data are ordered from first to last observation; in cross-sectional data the order may be of no importance, but in time series data the order is a crucial element of the data set and must be preserved.

The matrix of variables on which we are conditioning (the 'independent' variables) is $n \times k$, with elements:

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_{1,1} & X_{2,1} & \dots & X_{k-1,1} \\ 1 & X_{1,2} & X_{2,2} & \dots & X_{k-1,2} \\ 1 & X_{1,3} & X_{2,3} & \dots & X_{k-1,3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \dots & X_{k-1,n} \end{bmatrix},$$

where in this example we have labelled the individual observations with the variable number first, and observation number second. This labelling convention could be changed, but we do need the observations (rows) in order from first to last, and the individual variables (columns) in whatever order we want; the order of the estimated weights $\hat{\beta}_i$ will correspond with this order of the variables. The first column, of ones, is a constant intercept in the model. Its presence guarantees that the sum of the estimated errors in model (17.2.1) is zero.

We will again choose parameter estimates $\hat{\beta}_i$ in order to minimize the sum of squared residuals: that is, we minimize $\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} = \sum_{i=1}^{n} \varepsilon_i^2$, or

$$\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} (\boldsymbol{y}^\top\boldsymbol{y} - \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{y} - \boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}).$$

The two central terms on the right-hand side are equal, since one is the transpose of the other and they are both scalars, therefore the same. Thus we can write the minimization as

$$\min_{\boldsymbol{\beta}} (\boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}).$$

Taking the derivative with respect to $\boldsymbol{\beta}$, using the rules given in the Appendix, setting the result to zero for an optimum, and using the symbol $\hat{\boldsymbol{\beta}}$ now to denote the value that solves the equation, we have

$$-2\boldsymbol{X}^\top\boldsymbol{y} + 2\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0 \Longrightarrow \boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\boldsymbol{y} \Longrightarrow \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}.$$

Note that this assumes invertibility of the matrix $\boldsymbol{X}^\top\boldsymbol{X}$, which is equivalent to the assumption that $\boldsymbol{X}^\top\boldsymbol{X}$ is of full rank, which in turn implies that none of the rows or columns of $\boldsymbol{X}$ is a linear combination of any other row or column. This essentially rules out redundant regressors, which would be impossible to distinguish.

17.4 Computing standard errors of parameter estimates

The estimated parameter vector, $\hat{\boldsymbol{\beta}}$, is of course a random variable, being a function of the data $\boldsymbol{y}$, and so has a distribution around the 'true' values $\boldsymbol{\beta}$. The way in which we estimate the variances of these estimates (and therefore their standard errors) depends upon what we can take to be true about the process, and there are many techniques for obtaining these estimates, including simulation-based techniques such as the bootstrap. Here, we will continue to consider only the simplest case, with some strong assumptions on features of the process. In particular, for the regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, we will assume that:

(i) $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$,
(ii) $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2\mathbf{I}_n$,
(iii) $\mathrm{E}(\boldsymbol{X}^\top\boldsymbol{\varepsilon}) = 0$,
(iv) $\mathrm{rank}(\boldsymbol{X}^\top\boldsymbol{X}) = \mathrm{rank}(\boldsymbol{X}) = k$.

Assumption (i) is not restrictive since we can place a constant into the regression model to account for any non-zero intercept, and this will also guarantee that the sum of the residuals, $\hat{\boldsymbol{\varepsilon}}$, is exactly zero.[1] The second assumption indicates that each one of the disturbances has equal variance, and so in this sense each observation is equally reliable and should get equal weight; if this assumption does not hold, we can instead compute *generalized least squares* estimates. Assumption (iii) is critical, since if it does not hold, the unobservable disturbances will project on to the space spanned by the $\boldsymbol{X}$, thereby changing the estimated coefficients. Finally the linearly independent regressors assumption (iv) is necessary in order to invert the matrix $\boldsymbol{X}^\top\boldsymbol{X}$, and so it will be obvious if this assumption fails, since it will not be possible to compute the coefficients from the formula $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$. In a case like this, we say that the parameter vector $\boldsymbol{\beta}$ is not *identified*.

Now let us compute the variance-covariance matrix of the parameter estimates. We begin with an additional assumption which is not necessary for

[1] You can prove this as an exercise.

regression in general, but simplifies this computation by giving us a case of unbiased parameter estimates. We assume that the regressors can be treated as non-stochastic, as when they are chosen values for an experiment.

We begin by writing

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) \\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon} \\ &= \boldsymbol{\beta} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}. \end{aligned} \tag{17.4.1}$$

Thus $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathrm{E}[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}] = \boldsymbol{\beta} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\mathrm{E}[\boldsymbol{X}^\top\boldsymbol{\varepsilon}] = \boldsymbol{\beta}$, since the last term is zero by (iii) above. Thus $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, and the estimator is unbiased in this non-stochastic regressor case. More generally, we could obtain the weaker result that the probability limit of $\hat{\boldsymbol{\beta}}$ is $\boldsymbol{\beta}$, with stochastic regressors.

Next, we compute $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathrm{E}\big((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\big)$. Using the results just obtained above, $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}$, and so

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathrm{E}\big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon})^\top\big) \\ &= \mathrm{E}\big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\big).\end{aligned}$$

Finally, assuming the $\boldsymbol{X}$ variables are non-stochastic means that they can be taken out of the expectation (*e.g.* $\mathrm{E}(cZ) = c\mathrm{E}(Z)$ where c is non-stochastic), so that we have

$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top[\sigma^2\mathbf{I}_n]\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ by assumption (ii) above. Finally, moving the scalar σ^2 outside the matrix product, the identity matrix becomes redundant (like multiplying by 1, we don't need to write it explicitly), and we obtain

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}. \tag{17.4.2}$$

We can estimate the error variance σ^2 as $s^2 = \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}/(n-k)$, and we have the fully operational estimator $s^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ for the variance-covariance matrix of the estimated parameters, $\hat{\boldsymbol{\beta}}$. The square roots of the diagonal elements of this matrix are then the standard errors of the individual parameter estimates $\hat{\beta}_i,\ i = 1,\ldots,k$.

17.5 Tests on linear combinations of parameters

Let $\boldsymbol{a}$ be a $k\times 1$ vector of non-stochastic constants. Then $\boldsymbol{a}^\top\hat{\boldsymbol{\beta}}$ is a (scalar) *linear combination* of the parameter estimates. By use of the result (17.4.2), we see that

$$\mathrm{Var}(\boldsymbol{a}^\top\hat{\boldsymbol{\beta}}) = \boldsymbol{a}^\top\mathrm{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{a} = \sigma^2\boldsymbol{a}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a}.$$

Now let $\boldsymbol{A}$ be a $k\times r$ matrix, with $r\geq 1$, all the elements of which are non-random. Then $\boldsymbol{A}^\top\hat{\boldsymbol{\beta}}$ is an $r\times 1$ vector, each element of which is a linear combination of the estimated parameters. The covariance matrix of $\boldsymbol{A}^\top\hat{\boldsymbol{\beta}}$ is the $r\times r$ matrix

$$\mathrm{Var}(\boldsymbol{A}^\top\hat{\boldsymbol{\beta}}) = \boldsymbol{A}^\top\mathrm{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{A} = \sigma^2\boldsymbol{A}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{A}.$$

17.6 Measures of fit

Measures of the goodness of fit of a regression model to data are commonly quoted, and easily misunderstood. We will begin by defining the widely used *coefficient of determination* (R^2) measure, and a common degrees-of-freedom adjustment to it, producing the $\overline{R}^2$. However to understand the limitations of the R^2 and why it does not provide a good guide to the number of variables to include in a model, we need to consider the question of optimal model complexity, and the potential for over-fitting or over-specification. These will be addressed in the next sub-section.

For a given set of regressors, we have so far chosen the parameters by minimizing the sum of squared residuals. With the estimated parameters $\hat{\boldsymbol{\beta}}$ we obtain the estimated conditional expectation for each of the observations in the vector of dependent variables, $\boldsymbol{y}$. Denoting these estimated conditional expectations, or *fitted values*, by $\hat{\boldsymbol{y}}$ we have $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and the errors in fitting the $\boldsymbol{y}$ values, or *residuals*, are then $\boldsymbol{y}-\hat{\boldsymbol{y}} = \boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}} \equiv \hat{\boldsymbol{\varepsilon}}$. Then the sum of squared residuals is $SSR = \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} = \sum_{i=1}^n \varepsilon_i^2$.

The usual way to define the fit of a least-squares regression model is to consider the sum of squared residuals relative to the target to be explained, which is taken to be total variation of $\boldsymbol{y}$ around its mean, $SST = (\boldsymbol{y}-\overline{y}\boldsymbol{\iota})^\top(\boldsymbol{y}-\overline{y}\boldsymbol{\iota}) = \sum_{i=1}^n(y_i-\overline{y})^2$. We can then define

$$R^2 = 1 - SSR/SST = 1 - \frac{\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}}{(\boldsymbol{y}-\overline{y}\boldsymbol{\iota})^\top(\boldsymbol{y}-\overline{y}\boldsymbol{\iota})} = 1 - \frac{\sum_{i=1}^n\varepsilon_i^2}{\sum_{i=1}^n(y_i-\overline{y})^2}.$$

Note a few properties of the R^2 measure. First, the R^2 can never be negative: the worst SSR that the regression can produce would result from fitting a constant (intercept) alone, which is always an option since coefficients on other variables can be set to zero. Coefficients on other variables are included only if they improve the fit beyond the fit with a constant only. But the fit with a constant only just yields an SSR which is equal to SST, because the constant is estimated as $\overline{y}$, the sample mean of $\boldsymbol{y}$. Thus the worst R^2 obtainable in this case (where a constant included) is $1 - SST/SST = 0$. Similarly, a regression cannot produce a better fit than in fitting every data point perfectly, in which case $R^2 = 1 - 0/SST = 1$, and so if there is a constant in a linear regression, $0\leq R^2\leq 1$.

An argument similar to that above shows that R^2 cannot decrease as variables are added. Consider adding one more variable to the regression. The coefficients are chosen to minimize SSR; if placing a non-zero coefficient on the new variable were to make the fit worse (raise the SSR), then the SSR-minimizing estimate of the parameter would be to set it to zero. In other words, since zero is an option for any coefficient, any non-zero coefficient is chosen only if it reduces SSR. If SSR is reduced, R^2 increases; in the limit of a zero coefficient on the newly added variable, R^2 stays the same, and so R^2 can never decrease as a variable is added.

If one were to decide what variables to include by maximizing the R^2, since it would always be possible to increase the R^2, one would always add any new variable that comes along. As we will see in the next section, at some point this makes the model worse as a device for predicting as-yet unseen values. In order to settle on a number of variables to include which gives the best out-of-sample predictions (predictions of values not used in estimation), there has to be some criterion other than finding the highest R^2. R^2 is an indicator of the match between the values of $\boldsymbol{y}$ and their predicted values, but getting the largest possible R^2 is not a sensible way to choose a model.

17.7 Omitted variables bias

Suppose that we estimate the model (17.2.1) when the true DGP is in fact given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{Z}\boldsymbol{\gamma}_0 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma_0^2\mathbf{I}) \tag{17.7.1}$$

for some specific values $\boldsymbol{\beta}_0$, $\boldsymbol{\gamma}_0 \neq 0$, and σ_0^2. In this case, (17.2.1) is an *underspecified model*, and the variables in the matrix $\boldsymbol{Z}$ are *omitted variables* with respect to the model (17.2.1).

It is important to note that the estimator (17.2.6) is biased when the model is underspecified. Instead of the result (17.4.1), we have

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{Z}\boldsymbol{\gamma}_0 + \boldsymbol{u}) \\ &= \boldsymbol{\beta}_0 + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Z}\boldsymbol{\gamma}_0 + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u},\end{aligned}$$

so that

$$\text{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0 + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Z}\boldsymbol{\gamma}_0.$$

The second term in the right-hand side above is equal to zero only when $\boldsymbol{X}^\top\boldsymbol{Z} = 0$ or $\boldsymbol{\gamma}_0 = \boldsymbol{0}$. The first possibility arises when the two sets of regressors are mutually orthogonal, the second when (17.2.1) is not in fact underspecified. Except in these very special cases, $\tilde{\boldsymbol{\beta}}$ is biased. The magnitude of the bias depends on the parameter vector $\boldsymbol{\gamma}_0$ and on the $\boldsymbol{X}$ and $\boldsymbol{Z}$ matrices. Because this bias does not vanish as $n \to \infty$, $\tilde{\boldsymbol{\beta}}$ is also generally inconsistent.

17.8 In-sample and out-of-sample fit

Determining how many (and which) variables to include in a regression model is a perennially tricky problem, for which many solutions have been proposed. In this section we will simply explain the problem and the classic in-sample *vs.* out-of-sample relation between number of included variables and the loss (for example, sum of squared errors) that arises when we use our model to predict (fit) new observations.

When we say 'in sample fit', we mean how well our model based on $\boldsymbol{X}$ fits the data $\boldsymbol{y}$ that we used in estimating the model parameters; we have both $\boldsymbol{y}$ and $\boldsymbol{X}$ available to us to fit our model and estimate the parameters. This is sometimes called a 'training' sample, because we use data on both $\boldsymbol{y}$ and $\boldsymbol{X}$

to 'train' our model to fit $\boldsymbol{y}$ as well as possible given the $\boldsymbol{X}$ data that we have to work with, by estimating parameters as well as we can. When we have a model, we may use it to understand which factors are useful in predicting values for y, and we may also use it to predict some y values when we observe the X values, but we do not observe (at least not immediately) the y. For example, X may be some values observable now, and y may be future values, not observable until next month; we estimate the model parameters on dates for which $\boldsymbol{y}$ can be observed, using $\boldsymbol{X}$ from previous periods. Alternatively, $\boldsymbol{X}$ may be individual characteristics, and y may be predicted behaviour of an individual, such as amount of expenditure on something. We may obtain a (training) data set of data $\boldsymbol{X}$ on individuals whose expenditures $\boldsymbol{y}$ are known, and then we would like to predict expenditures by individuals whose X characteristics are known, but where their expenditures are not known or perhaps have not yet been made.

As we saw above, adding more variables will always improve the fit to the sample (in other words, will always reduce loss in sample), and R^2 will always (weakly) increase.[2] Correspondingly, in Figure 17.8.2, the blue 'in-sample' line decreases monotonically as more variables are added.

However, when we look at the fit to out-of-sample data, for example the out-of-sample sum of squared errors, there are conflicting effects: one effect tends to improve, and one tends to worsen, the out of sample fit.[3] To understand how adding more variables can make things worse, consider the example depicted in Figure 17.8.1.

Figure 17.8.1 gives an example of a variable, say $\boldsymbol{x}_1$, which has some predictive value for the dependent variable $\boldsymbol{y}$; the true coefficient (the best possible weight that we could put on that variable for prediction of $\boldsymbol{y}$) is 0.01. The two densities of the estimated parameter, $f_1(\hat{\beta})$ and $f_2(\hat{\beta})$, are the densities that we would obtain at two different sample sizes. The blue (higher-variance) density would correspond with a lower sample size than the red (lower-variance) density; as we obtain more sample points, variance of the parameter estimates declines.

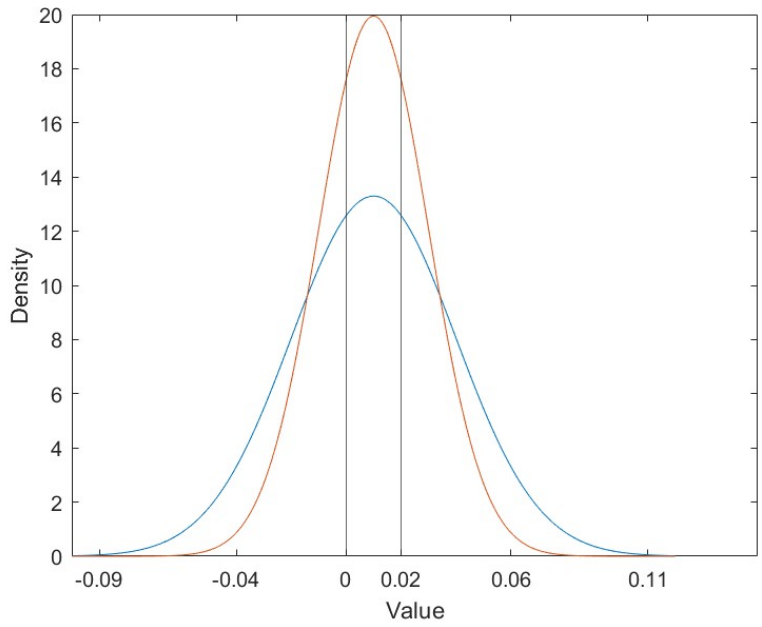
In this example $\boldsymbol{x}_1$ does have some predictive value, but the researcher does not necessarily know that. Consider the following question: imagine that the researcher has left $\boldsymbol{x}_1$ out of the model, but reconsiders and puts it in. What is the probability that the estimated weight on $\boldsymbol{x}_1$ will be worse than what we get by just leaving it out?

The 'true' value of the parameter β on $\boldsymbol{x}_1$ was, we said, 0.01. If we leave it out, equivalent to giving it a weight of zero, our error is 0.01. What is the

[2] By 'weakly' increase or decrease, we allow the possibility that the value will stay the same; so if a quantity weakly increases, the change is ≥ 0.

[3] The out-of-sample fit is $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, but where $\boldsymbol{y}$ and $\boldsymbol{X}$ are data series that were not used to estimate $\boldsymbol{\beta}$.

FIGURE 17.8.1
Probability of worsening an estimate by inclusion



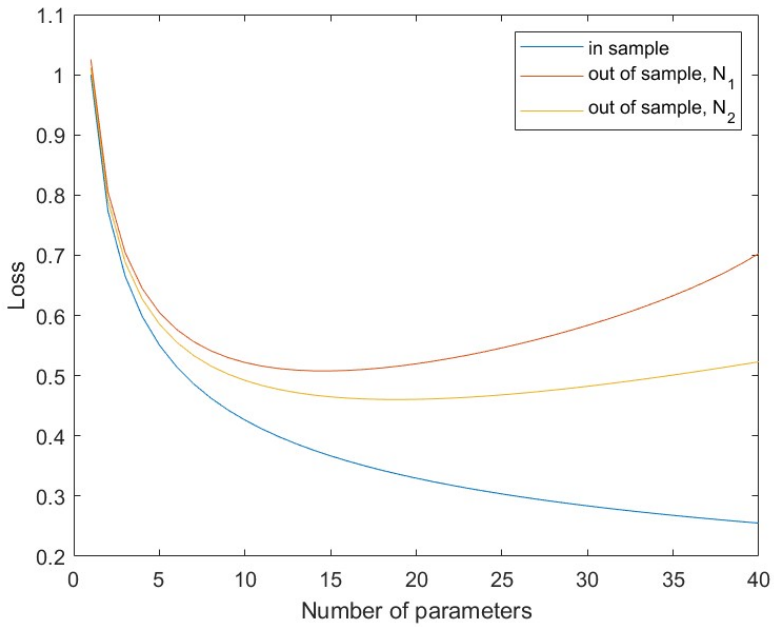
probability that by including it, we will make an error in estimation bigger than the 0.01 that we get by leaving it out?

Start with the lower sample size (blue density). Areas under the density between two points represent probabilities of the parameter estimated lying in that interval.

The region bounded by the two vertical lines is (excluding the bounding lines) the region where we do better by including $\boldsymbol{x}_1$: any outcome between 0 and 0.02 is less of an error than at the boundary line 0. By contrast, an estimate lying outside the bounds means that our parameter error is greater than the error we would have made by dropping the variable. For example, if $\hat{\beta}$ turns out at 0.035 on our sample, our estimation error is 0.035-0.01 = 0.025. By leaving it out, our parameter estimation error was only 0.01.

The probability that we get a worse estimate than by leaving the variable out (setting the parameter to zero) is the area under the blue density, outside the vertical lines. If we knew the standard error of $\hat{\beta}$ we could compute it exactly but, since we don't, looking at the graphic suggests that the probability of making it worse is around 85% (as a rough approximation, around 25% lies between the two vertical lines).

FIGURE 17.8.2
In-sample and out-of-sample loss



As sample size increases, these probabilities shift. The red density again represents the density of the parameter estimate $\hat{\beta}$, but for a larger sample size. The standard error declines at rate $\sqrt{(n)}$, and the density narrows, still centred on 0.01. At the larger sample size, the probability of making things worse by including $\boldsymbol{x}_1$ is lower: the area between the vertical lines under the red density might be around 40%. There is still a 60% chance that our parameter estimate will be worse than if we just set it to 0, but the probabilities are improving. As we consider larger and larger sample sizes, the standard error of the parameter estimate declines, the density of the estimate narrows, and eventually the probability of being between the vertical lines, where we get a better estimate by including the variable, approaches 1.

From this example we see that adding a variable into a regression may produce positive or negative effects; we would expect a worsened parameter estimate to lead to worsened predictions, on average. We also see that the best decision is a function of sample size, for a variable that in principle does have some predictive value: at larger sample sizes the estimation variance declines and the chances of inclusion being a good decision increase. At small sample sizes, adding the variable may make things worse because of the large estimation variance. *Estimation variance can cause our model to deteriorate as we add variables, even if they are in principle potentially useful.*

Figure 17.8.2 is an example of a standard figure showing in-sample loss (*e.g.* sum of squared residuals) and out-of-sample loss, at two sample sizes used for estimation. In the estimation sample, loss declines with additional variables. As more and more variables are added, however, the model is adapting itself increasingly to the random 'noise' $\boldsymbol{\varepsilon}$ in the process, rather than the 'signal', the conditional expectation function. In fitting new observations which were not used in estimation, the model performs less well, and past a certain point performance deteriorates (loss increases) as we add more variables. But as we have just seen, with a larger sample size, parameter estimation variance ($var(\hat{\beta})$) is smaller, and it is more likely to be worthwhile to include an additional variable. The optimal number of variables will tend to increase. In Figure 17.8.2 the minimum point of the function with the larger sample size (yellow line) occurs at a larger number of included variables (and therefore estimated parameters) than for the out-of-sample loss function at the smaller sample size (red line). There is a tradeoff, when we add more variables, between extracting information from the additional regressors (bias reduction) and adding to variance. With larger sample sizes, the variance is reduced and the tradeoff shifts in the direction of favouring additional regressors.

Appendix to Chapter 17

A1 Matrix Differentiation

Differentiation of matrices is essentially the same as scalar differentiation; that is, the principles of calculus applied are identical. The difference is that we may be taking the derivative of one *or more* quantities with respect to one *or more* others. So we need to represent the answers in matrix form. More than one convention for doing so is possible.

Let $\boldsymbol{a}$, $\boldsymbol{A}$, and $\boldsymbol{x}$ be of dimension $n \times 1$, $n \times n$, and $n \times 1$ respectively. Clearly $\boldsymbol{a}^\top \boldsymbol{x} = \sum a_i x_i$ and so

$$\frac{\partial(\boldsymbol{a}^\top \boldsymbol{x})}{\partial x_i} = a_i.$$

This is simply scalar calculus, since $\boldsymbol{a}^\top \boldsymbol{x}$ is a scalar. We can do the same with respect to each element of the vector $\boldsymbol{x}$, however, and put the answers together into a new vector – namely $\boldsymbol{a}$:

$$\frac{\partial(\boldsymbol{a}^\top \boldsymbol{x})}{\partial x} = \left[\frac{\partial(\boldsymbol{a}^\top \boldsymbol{x})}{\partial x_1}, \frac{\partial(\boldsymbol{a}^\top \boldsymbol{x})}{\partial x_2}, \ldots, \frac{\partial(\boldsymbol{a}^\top \boldsymbol{x})}{\partial x_n}\right]^\top = [a_1, a_2, \ldots, a_n]^\top = \boldsymbol{a}.$$

Similarly,

$$\frac{\partial(\boldsymbol{a}^\top \boldsymbol{x})}{\partial \boldsymbol{x}^\top} = \boldsymbol{a}^\top.$$

Some other rules:

$$\begin{aligned}
\frac{\partial(\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x})}{\partial \boldsymbol{x}} &= (\boldsymbol{A} + \boldsymbol{A}^\top)\boldsymbol{x} \\
\frac{\partial^2(\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^\top} &= (\boldsymbol{A} + \boldsymbol{A}^\top) \\
\frac{\partial(\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x})}{\partial \boldsymbol{A}} &= \boldsymbol{x}\boldsymbol{x}^\top \\
\frac{\partial \log(\det(A))}{\partial \boldsymbol{A}} &= (\boldsymbol{A}^\top)^{-1}, \quad \text{if} \quad \det(\boldsymbol{A}) > 0.
\end{aligned}$$

These are the rules which we will be using. Other extensions of scalar rules can of course be derived as well.

A2 Covariance Matrices

Let $\boldsymbol{a}$, $\boldsymbol{V}$, and $\boldsymbol{x}$ be of dimensions $n \times 1$, $n \times n$, and $n \times 1$ respectively. Then $\boldsymbol{a}^\top \boldsymbol{x} = \sum_{i=1}^{n} a_i x_i$, a scalar (1×1) quantity.

Let $\boldsymbol{\mu}$ be the vector of expectations of the elements of the vector $\boldsymbol{x}$, so that $\mathrm{E}(\boldsymbol{x} - \boldsymbol{\mu}) = 0$.

The variance-covariance matrix of $\boldsymbol{x}$, or simply the 'covariance matrix' or 'variance matrix,' is an $n \times n$ matrix such that each element (i, j) represents $\mathrm{E}[(X_i - \mu_i)(X_j - \mu_j)]$; where $i = j$ these terms are variances, and where $i \neq j$

they are covariances. Since $\mathrm{E}[(X_i-\mu_i)(X_j-\mu_j)] = \mathrm{E}[(X_j-\mu_j)(X_i-\mu_i)]$, the $(i,j)^{\text{th}}$ element of the matrix is equal to the $(j,i)^{\text{th}}$ element, and the matrix is therefore symmetric.

We can represent the covariance matrix in vector notation as

$$\mathrm{Var}(x) = \mathrm{E}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top].$$

Notice that the transpose is on the second vector, so that we obtain an $n \times n$ matrix; if the transpose were on the first term, we would obtain the inner product, $\sum_{i=1}^{n}(x_i-\mu_i)^2$, a scalar.

With the covariance matrix, we can obtain the variance of any linear combination of the x_i. Since $\mathrm{Var}(\boldsymbol{x}) = \mathrm{E}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top]$, we have

$$\mathrm{Var}(\boldsymbol{a}^\top\boldsymbol{x}) = \mathrm{E}[\boldsymbol{a}^\top(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{a}] = \boldsymbol{a}^\top\mathrm{Var}(\boldsymbol{x})\boldsymbol{a}.$$

Note that this expression is of dimension $(1 \times n)(n \times n)(n \times 1)$ or 1×1, a scalar.

Let's use this to derive the simple rule for the variance of a linear combination of two random variables that we stated earlier, *i.e.* $\mathrm{Var}(b_1X + b_2Y) = b_1^2\,\mathrm{Var}(X) + b_2^2\,\mathrm{Var}(Y) + 2b_1b_2\mathrm{cov}(X,Y)$. The vector of weights in the linear combination is

$$\boldsymbol{a} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \text{ for } \boldsymbol{x} = \begin{bmatrix} X & Y \end{bmatrix},$$

so that $\boldsymbol{a}^\top\boldsymbol{x} = b_1X + b_2Y$. The covariance matrix is

$$V = \begin{bmatrix} \mathrm{Var}(X) & \mathrm{cov}(X,Y) \\ \mathrm{cov}(X,Y) & \mathrm{Var}(Y) \end{bmatrix}.$$

Finally

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{a}^\top\boldsymbol{x}) = \boldsymbol{a}^\top\mathrm{Var}(\boldsymbol{x})\boldsymbol{a} &= \begin{bmatrix} b_1 & b_2 \end{bmatrix} \begin{bmatrix} \mathrm{Var}(X) & \mathrm{cov}(X,Y) \\ \mathrm{cov}(X,Y) & \mathrm{Var}(Y) \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \\
&= \begin{bmatrix} b_1\,\mathrm{Var}(X) + b_2\mathrm{cov}(X,Y) & b_1 cov(X,Y) + b_2 var(Y) \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \\
&= b_1^2\,\mathrm{Var}(X) + b_1b_2\mathrm{cov}(X,Y) + b_2b_1\mathrm{cov}(X,Y) + b_2^2\,\mathrm{Var}(Y) \\
&= b_1^2\,\mathrm{Var}(X) + b_2^2\,\mathrm{Var}(Y) + 2b_1b_2\mathrm{cov}(X,Y),
\end{aligned}$$

as we wished to show.

References

Cramér, H. (1955) *The Elements of Probability Theory.* Wiley, New York.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*, Springer-Verlag, New York. But see: **http://luc.devroye.org/rnbookindex.html**

Galbraith, J. W. and S. van Norden (2011) 'Kernel-based Calibration Diagnostics for Recession and Inflation Probability Forecasts.' *International Journal of Forecasting* 27, 1041-1057.

Heston, A., R. Summers and B. Aten (2002) Penn World Table v. 6.1. Center for International Comparisons at the University of Pennsylvania.

Hodgson, D. and K. Vorkink (2004) 'Asset pricing theory and the valuation of Canadian paintings.' *Canadian Journal of Economics* 37, 629-655.

Hogg, R. V and A. Craig (1959) *Introduction to Mathematical Statistics.* Macmillan, New York.

Hume, D. (1739) *A Treatise of Human Nature, Book I: Of the Understanding.* London.

Hume, D. (1748) *An Enquiry Concerning Human Understanding.* London.

Johnson, N. L. and S. Kotz (1970) *Distributions in Statistics: Continuous Univariate Distributions-I.* Wiley, New York.

Kendall, M. G., A. Stuart and J. K. Ord (1991) *Kendall's Advanced Theory of Statistics. Oxford University Press, New York.* Fifth edition of: Kendall, M.G. (1946) *The Advanced Theory of Statistics.*

Knight, F. H. (1921) *Risk, Uncertainty and Profit.* Houghton Mifflin, Boston.

Knuth, Donald E. (1998). *The Art of Computer Programming*, Vol. 2, *Seminumerical Algorithms*, third edition, Reading, Mass., Addison-Wesley.

Lehmann. E. L. (1986). *Testing Statistical Hypotheses* (2nd ed.). Springer. - §3.3

Mood, A. M., F. A. Graybill and D. C. Boes (1974) *Introduction to the Theory of Statistics.* McGraw-Hill.

Neyman, J. (1950) *A First Course in Probability and Statistics.* Holt, New York.

Pigou, A. C. (1920) *The Economics of Welfare.* MacMillan, London.

Popper, K. R. (1959) *The Logic of Scientific Discovery.* Hutchinson, London. Translation of the German original *Logik der Forschung,* Vienna, 1935.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Student (W. S. Gosset) (1908) 'The probable error of a mean.' *Biometrika* 6, 1-25.

Todhunter, I. (1865) *A History of the Mathematical Theory of Probability.* Macmillan, Cambridge.

Index