Economic Statistics

J.W. Galbraith
McGill University

Version of: September 2014
Typeset on: 23. 11. 2022

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 IV: Surveys of more advanced topics

Economic Statistics

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, following Frank Knight (1921), between 'risk'

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

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 ('Let's 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;

[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.

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 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.

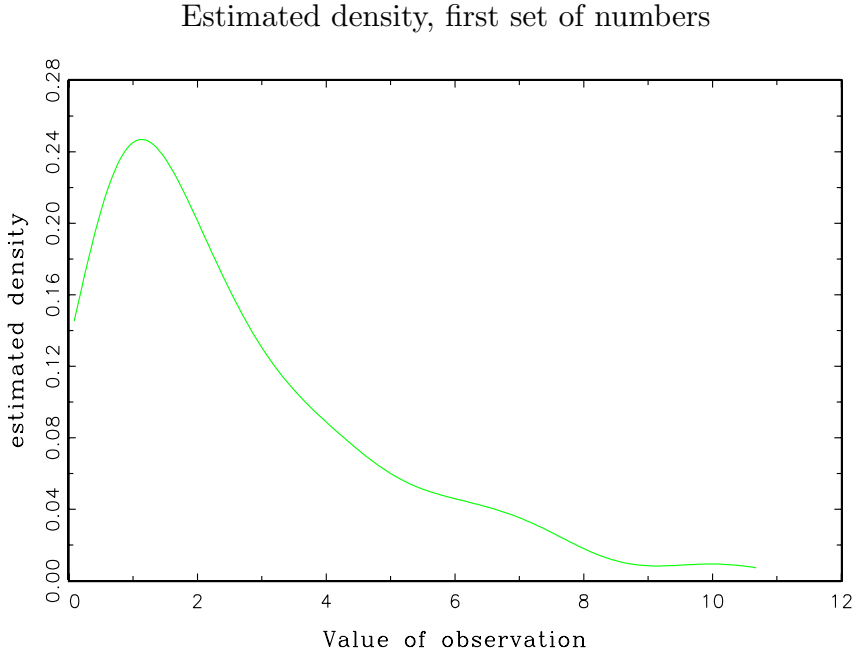
Example 3: Description

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.

FIGURE 1.1



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 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 likely–it's still 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 15-16 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: 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 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. What's 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

[4]The Dow Jones Industrial Average and related indices are described at the site www.djindexes.com.

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.

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.[5]

[5]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
Canadian dollar / US dollar exchange rate,
$C per $US, monthly average of noon rates

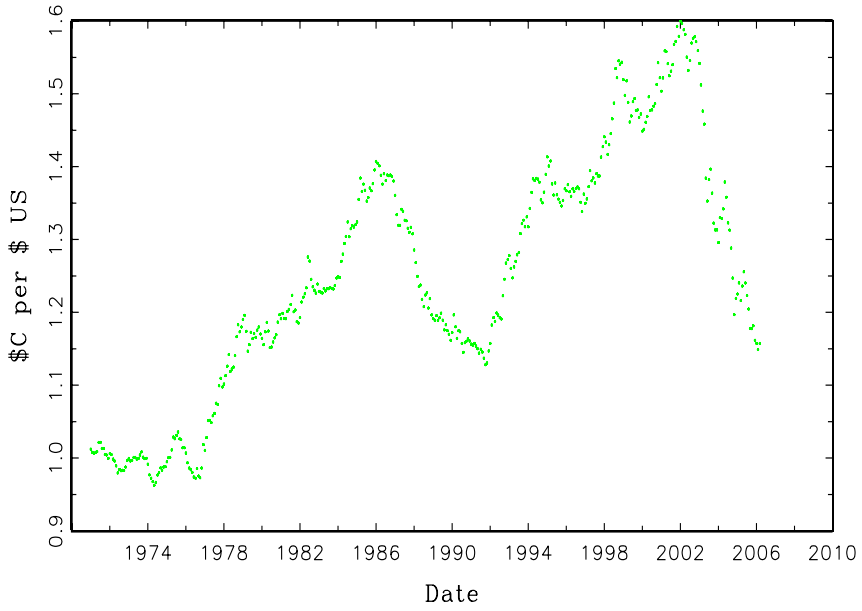
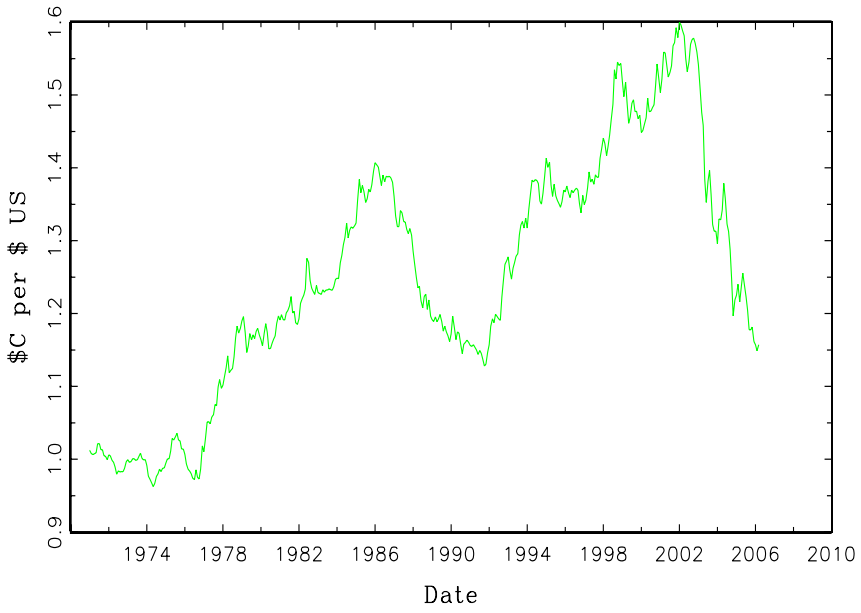
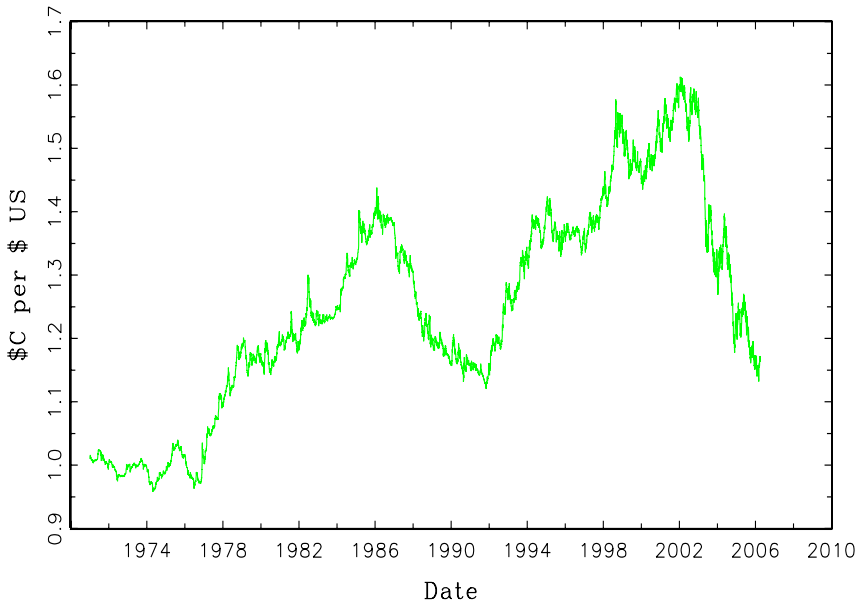


FIGURE 2.1.2
Canadian dollar / US dollar exchange rate,
$C per $US, monthly average of noon rates



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 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.

FIGURE 2.1.3
Canadian dollar / US dollar exchange rate,
$C per $US, daily noon rate



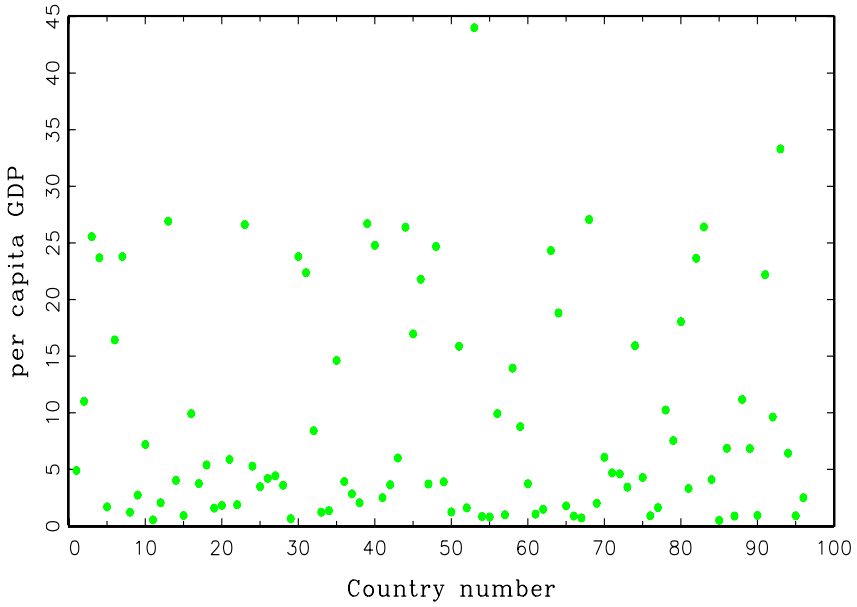
Note that while the data behind Figures 2.1.1 and 2.1.2 are measured at a lower frequency than those yielding Figure 2.1.3, each series represents measurements on the same underlying economic process (currency exchange between Canada and the U.S.).

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.[6]

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

FIGURE 2.1.4A
GDP per capita in a sample of 96 countries, year 2000
Thousands of US dollars (Penn World Data)



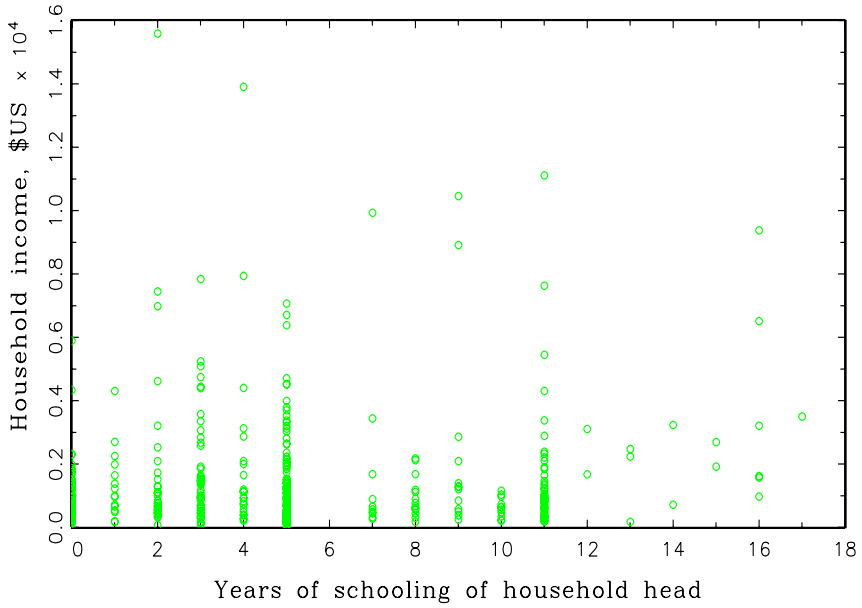
We see immediately many countries with low income, a smaller stratum of relatively high income countries with GDP around $25,000 US dollars per capita, and one substantially higher value near the middle of the set (Luxembourg, with annual per capita GDP of nearly $45,000 per capita in 2000). This set 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 sample–almost 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.[7]

[7]The Peruvian Living Standards Survey is one of the World Bank's Living Standards Measurement surveys.

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.

FIGURE 2.1.4B
Household income by educational attainment of head
Peruvian Living Standards Survey

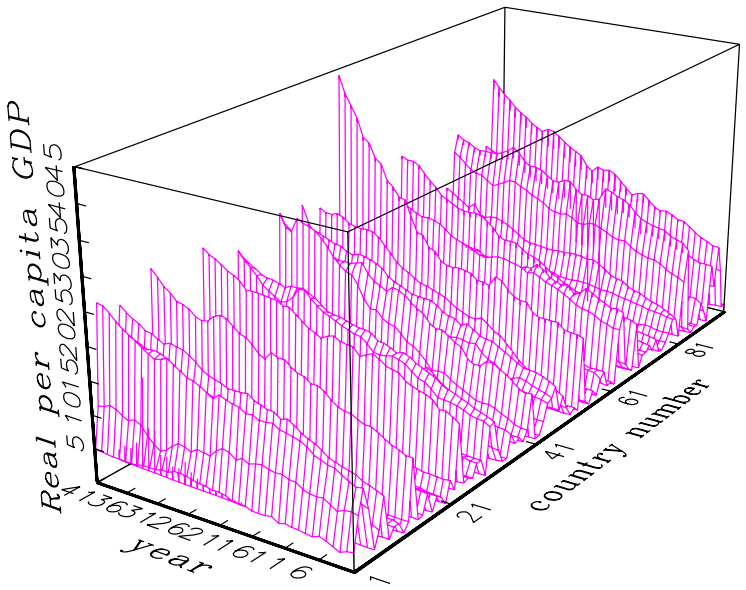


Panel data

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, PSID, or Canadian Survey of Labour

and Income Dynamics, SLID)[8] 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).

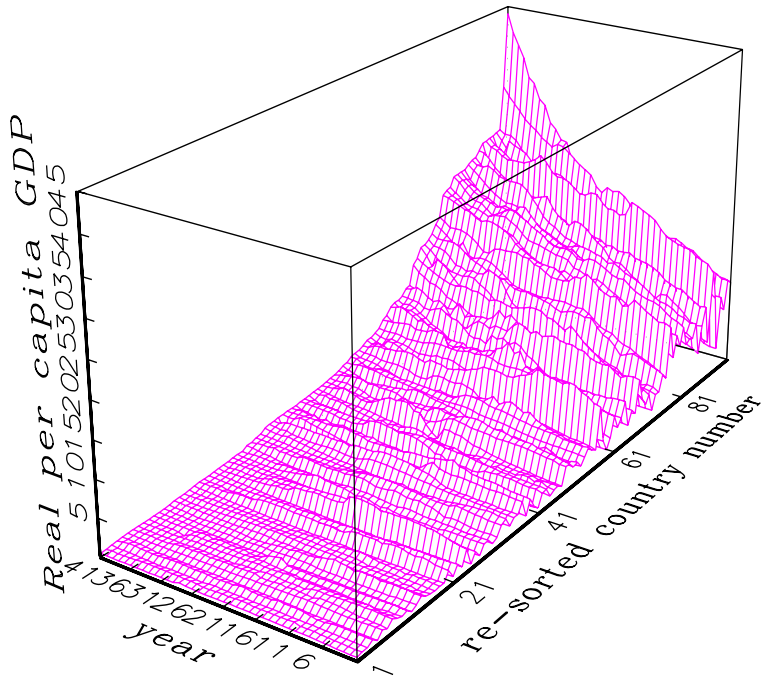
FIGURE 2.1.5
Real GDP per capita, 96 countries, years 1960–2000
Thousands of year-2000 US dollars (Penn World Data)



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:

[8]See http://psidonline.isr.umich.edu and http://www.statcan.ca/english/sdds /3889.htm respectively.

FIGURE 2.1.6
Real GDP per capita, 96 countries, years 1960–2000, sorted
Thousands of year-2000 US dollars (Penn World Data)



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 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 country's individual time series of output by the value in the first year, 1960. Each country's 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 country's 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
Real GDP per capita, 96 countries, years 1960–2000, scaled (1960=1)
Thousands of year-2000 US dollars (Penn World Data)

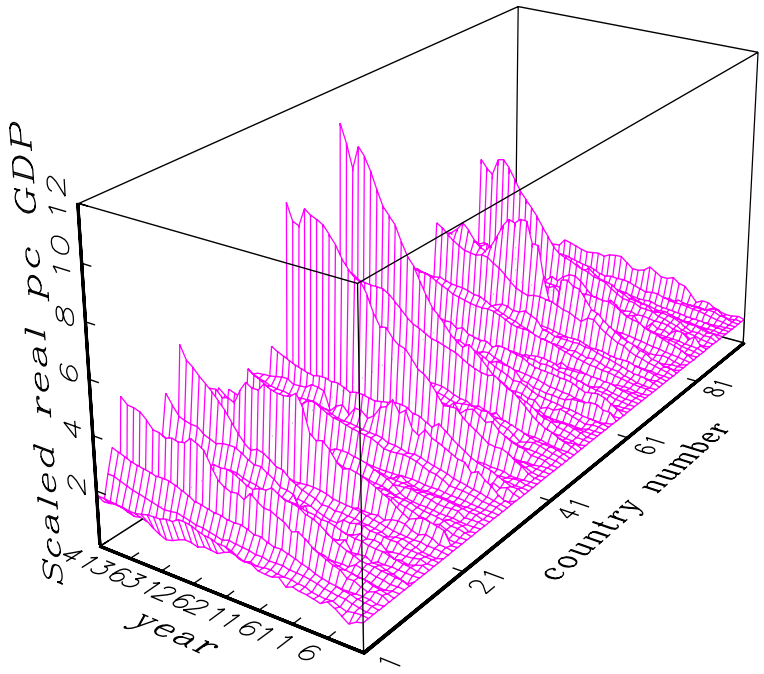
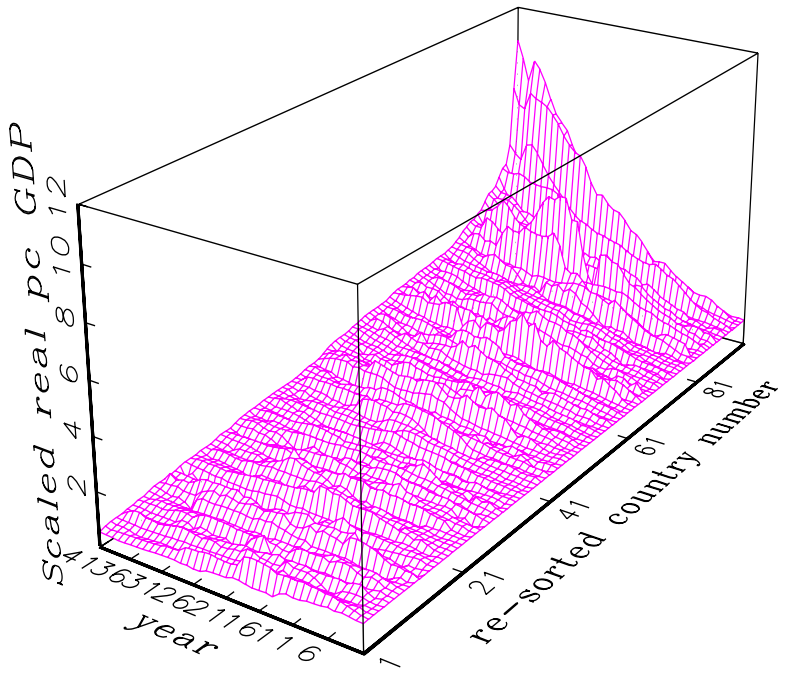


Figure 2.1.7 is again difficult to read because the more slowly-growing countries are obscured. We can again choose an ordering to make the figure easier to inspect, and order by the ratio of real per capita GDP in 2000 to that of 1960, to get Figure 2.1.8.

FIGURE 2.1.8
Real GDP per capita, 96 countries, years 1960–2000, scaled & sorted
Thousands of year-2000 US dollars (Penn World Data)



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.

2.2 TRANSFORMATIONS OF DATA SERIES

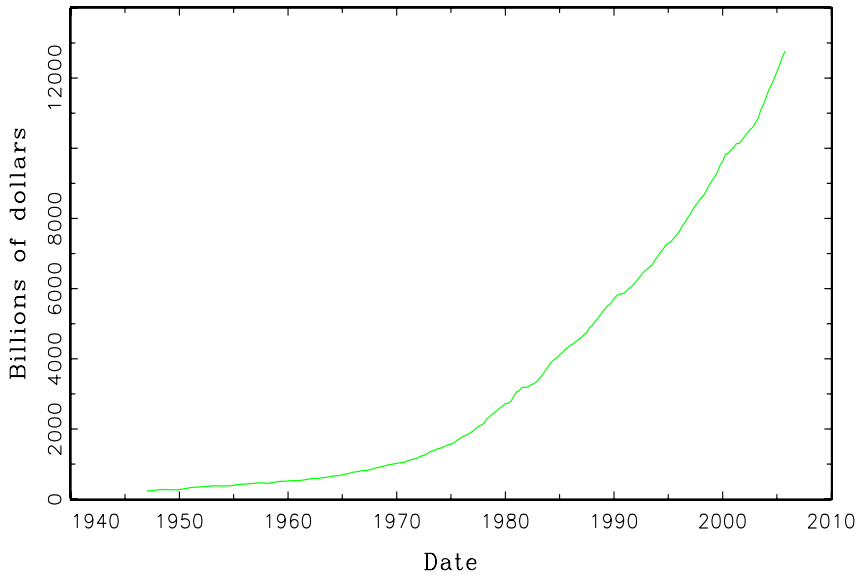
At times it is easier to understand some properties of a data series by doing some simple calculations before graphing the data. We begin by looking at some of the transformations common in time series data, for which transformations are often particularly useful. Consider first of all one of the best-known 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.[9] 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

$$y_t = \alpha(1+\beta)^t, \tag{2.1}$$

[9]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.

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 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.

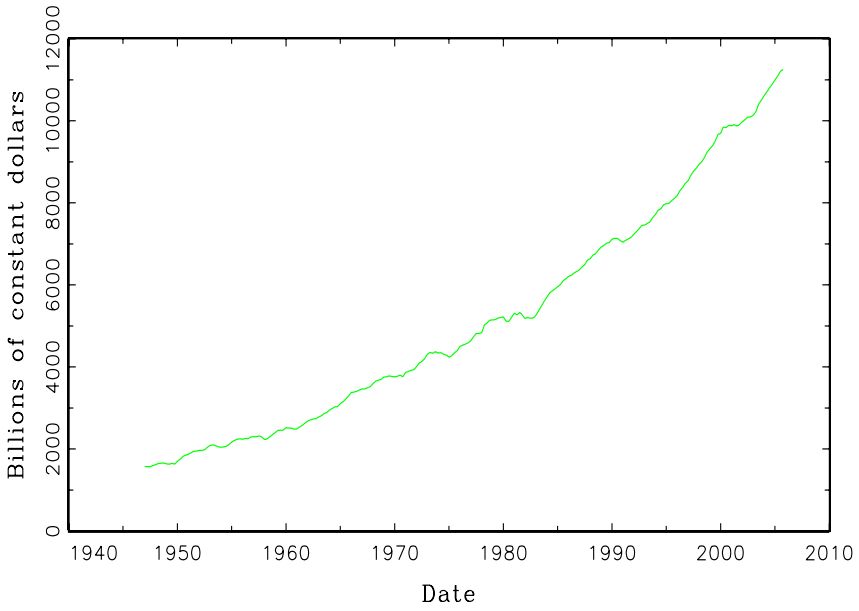
FIGURE 2.2.1
US gross domestic product
billions of dollars, seasonally adjusted



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 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.

FIGURE 2.2.2
US real gross domestic product
billions of (chained) 2000 dollars, seasonally adjusted



To understand how this is routinely handled, take the logarithm of the function (2.1) (the logarithmic function is reviewed in the appendix to this chapter). If $y_t = \alpha(1+\beta)^t$ then

$$log(y_t) = log(\alpha(1+\beta)^t) = log(\alpha) + t \cdot log(1+\beta), \tag{2.2}$$

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 (2.1) has turned a case of constant proportional or percentage growth into one of constant absolute growth. Moreover, $log(1+\beta) \simeq \beta$ for small values of β, so that $log(y_t)$ increases by roughly 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.

FIGURE 2.2.3
logarithm of US gross domestic product
seasonally adjusted

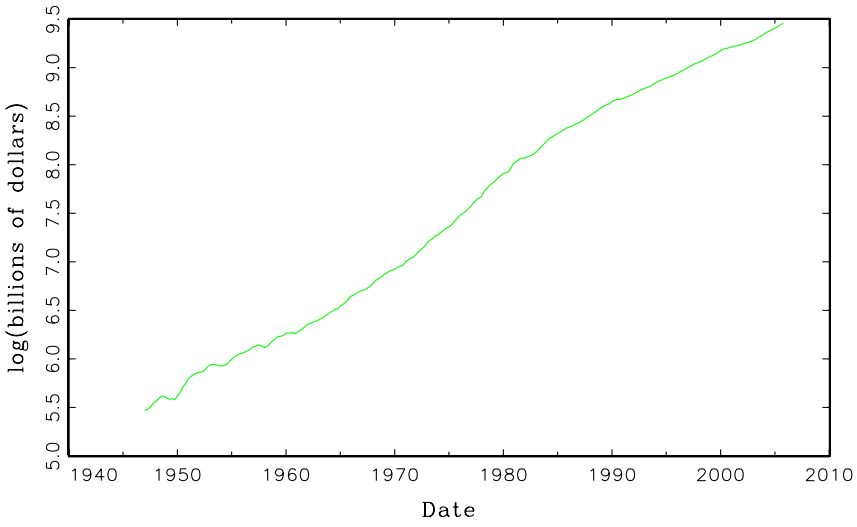
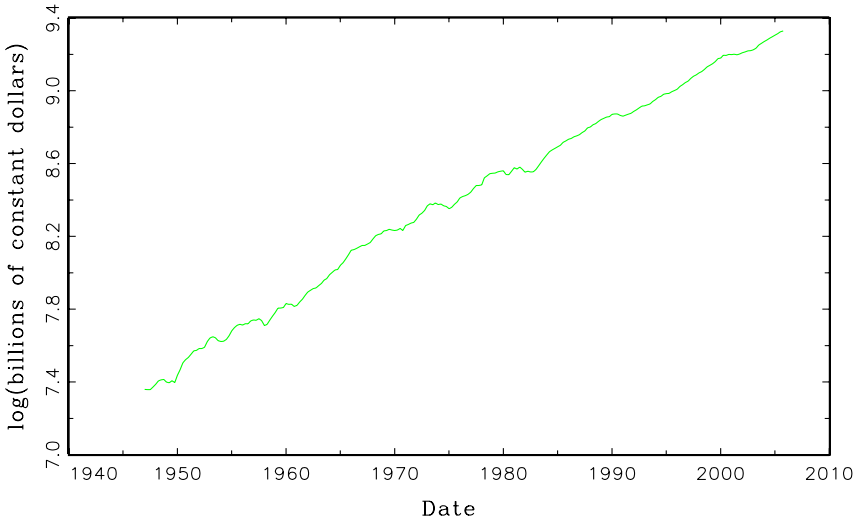


FIGURE 2.2.4
logarithm of US real gross domestic product
seasonally adjusted



A final transformation that is commonly applied has a more dramatic effect.

Consider the change between two neighbouring points,

$$log(y_t) - log(y_{t-1}) = log(1 + \beta), \tag{2.3}$$

by (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 $\Delta log(y_t) \simeq \beta$. Since this relationship is not exact, it is usually preferable to compute the proportionate change directly, however, as

$$\frac{y_t - y_{t-1}}{y_{t-1}}. \tag{2.4}$$

To compare these two computations, note that $\Delta log(y_t) = log(y_t) - log(y_{t-1}) = log(\frac{y_t}{y_{t-1}})$; for growth by the proportion β (100 β %), $\frac{y_t}{y_{t-1}} = 1 + \beta$, and so $\frac{y_t - y_{t-1}}{y_{t-1}} = \frac{y_t}{y_{t-1}} - 1 = \beta$. By contrast, $\Delta log(y_t) = log(1 + \beta) \simeq \beta$, 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 (2.4) gives the exact growth rate, and (2.3) the approximation.

Equation (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 (2.3), 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.

FIGURE 2.2.5
Proportionate change in US gross domestic product
seasonally adjusted

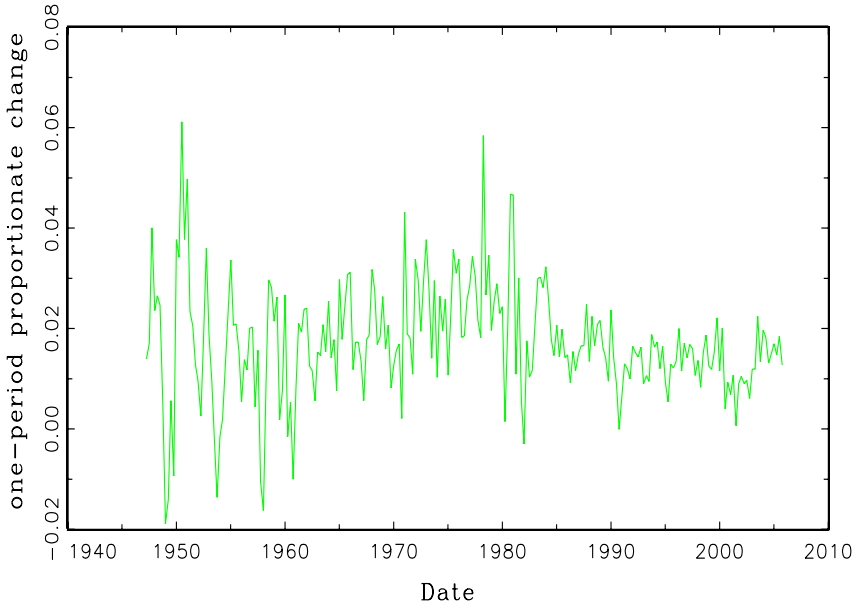
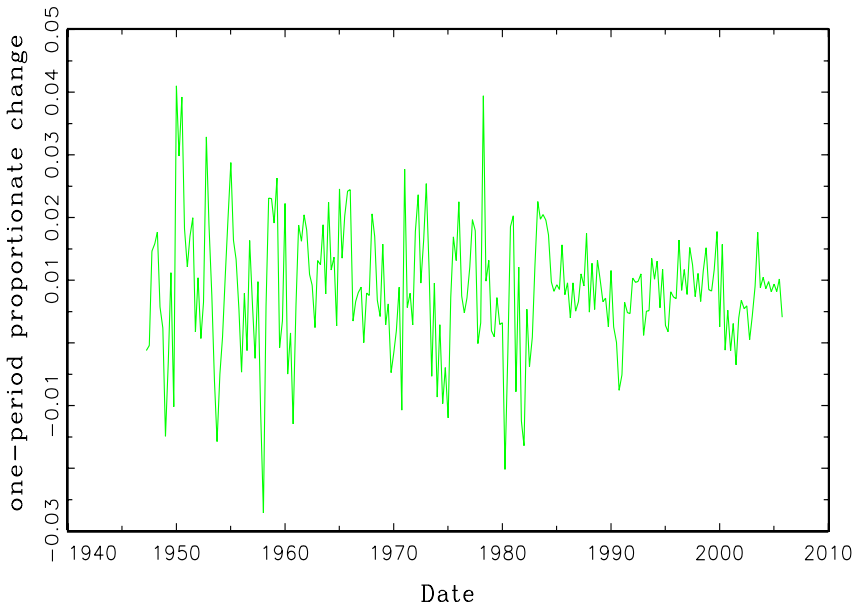


FIGURE 2.2.6
Proportionate change in US real gross domestic product
seasonally adjusted



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 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

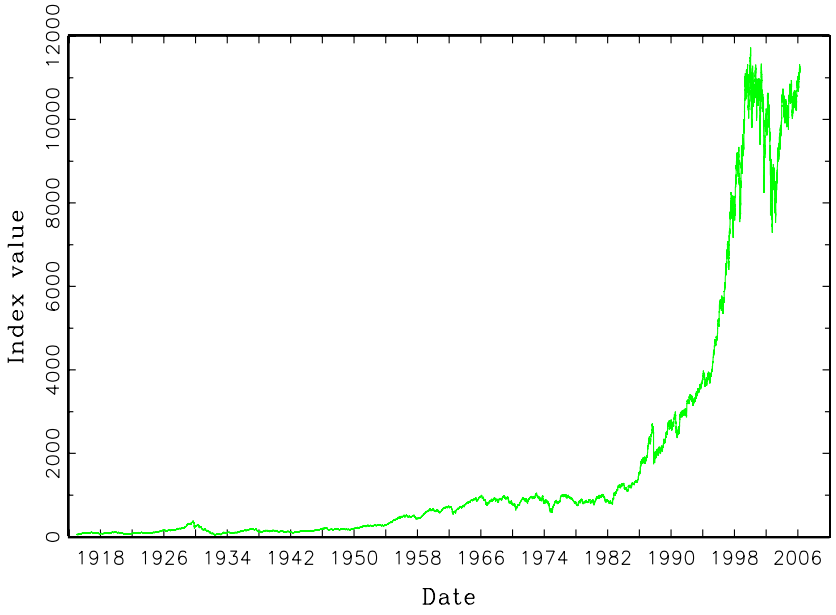
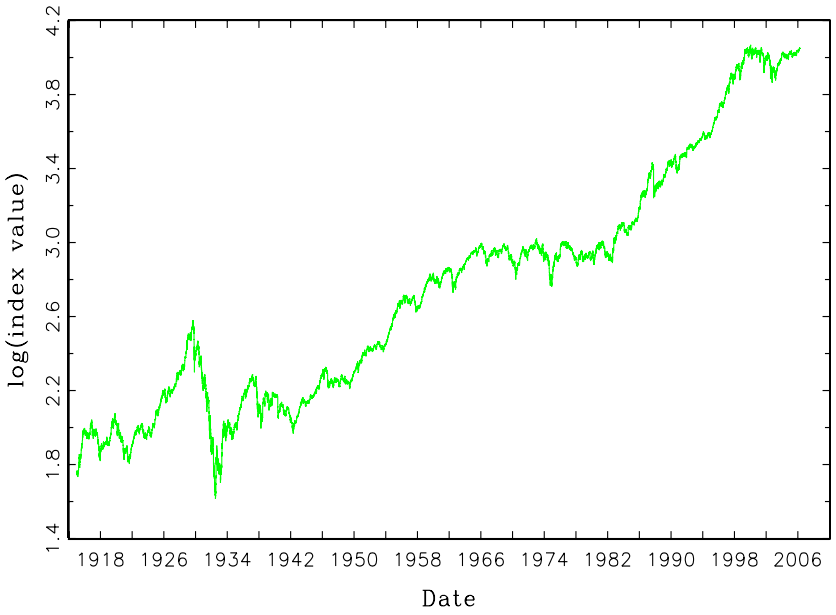


FIGURE 2.2.8
logarithm of Dow Jones Industrial Average



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.

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.[10] 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 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.

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

[10]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.

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.

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.

Appendix to Chapter 2

The logarithmic function

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. $\log_a(a^z) = z$).

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.

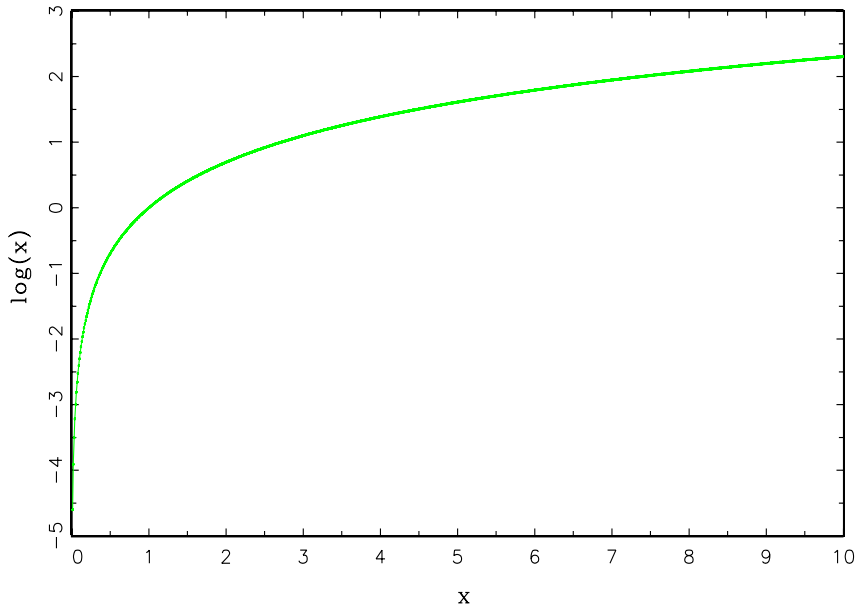
A very commonly used value of the base is the constant e, is defined as

$$e = \lim_{h \to 0} (1 + h)^{\frac{1}{h}}.$$

This constant has the property that the exponential function based on e is its own derivative, i.e. if $y = e^x$, then $\frac{dy}{dx} = e^x$. In general for a base a, $y = a^x$ and $\frac{dy}{dx} = ka^x$, for some constant k which is not equal to unity unless $a = e$.

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.

Figure A2.1
The natural logarithmic function



In addition to the properties given in section 2, note that

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

(therefore $\log(\frac{y}{z}) = \log(y) - \log(z)$). Logarithms taken to different bases differ only by a constant factor: $\log_a(z) = (\log_a(b))\log_b(z)$.

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:[11] $\overline{X} \equiv \frac{1}{N} \sum_{i=1}^{N} x_i$.

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

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:

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

[11] This is the sample *arithmetic* mean. The sample *geometric* mean is

$$\left(\prod_{i=1}^{N} x_i \right)^{\frac{1}{N}},$$

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

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 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.*[12]

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 $\frac{j}{N} \geq \alpha$.

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 $\hat{q}_{0.682} = 5.12$.[13]

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

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

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;

---

[12]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.*

[13]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).

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.)

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: $s^2 \equiv (N-1)^{-1} \sum_{i=1}^{N} (x_i - \overline{X})^2$.

D3.5 Standard error:[14] $s = +[(N-1)^{-1} \sum_{i=1}^{N} (x_i - \overline{X})^2]^{\frac{1}{2}}$.

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.[15]

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: $q_{0.75} - q_{0.25}$.

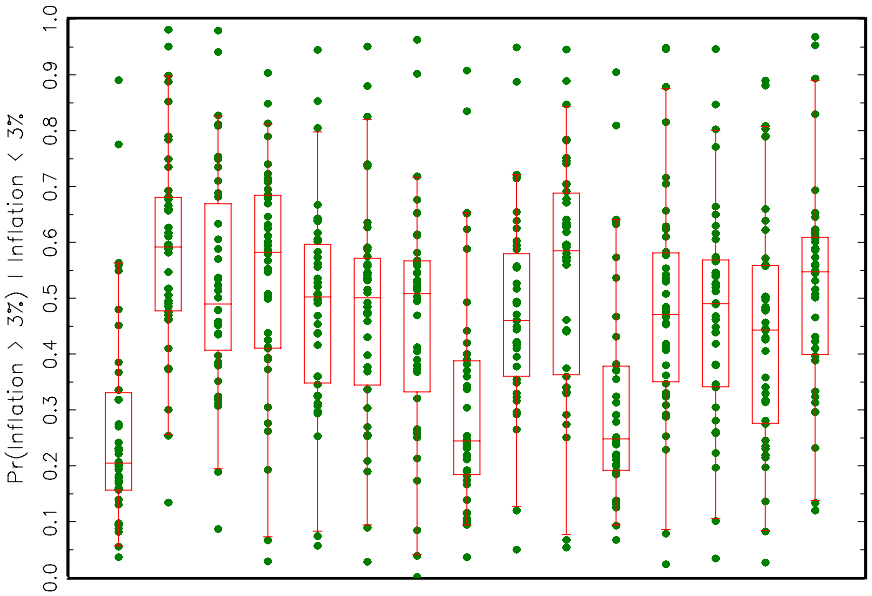
We could of course define other measures, taking differences between any two percentiles, but the interquartile range is commonly used.

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

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

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 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
Box-whisker plots of dispersion of probabilistic forecasts
Fifteen inflation forecasting models, outcomes with $\pi <$ target



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

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 $(x_i - \overline{X})$, 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: $(\overline{X} - \hat{q}_{0.5})/s$.

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 $(x_i - \overline{X})$ 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 standardized fourth power is often used as a measure of the relative frequency of extreme events, or the thickness of tails, as follows:

D3.10 Coefficient of kurtosis: $\frac{(N-1)^{-1}\sum_{i=1}^{N}(x_i-\overline{X})^4}{s^4}$.

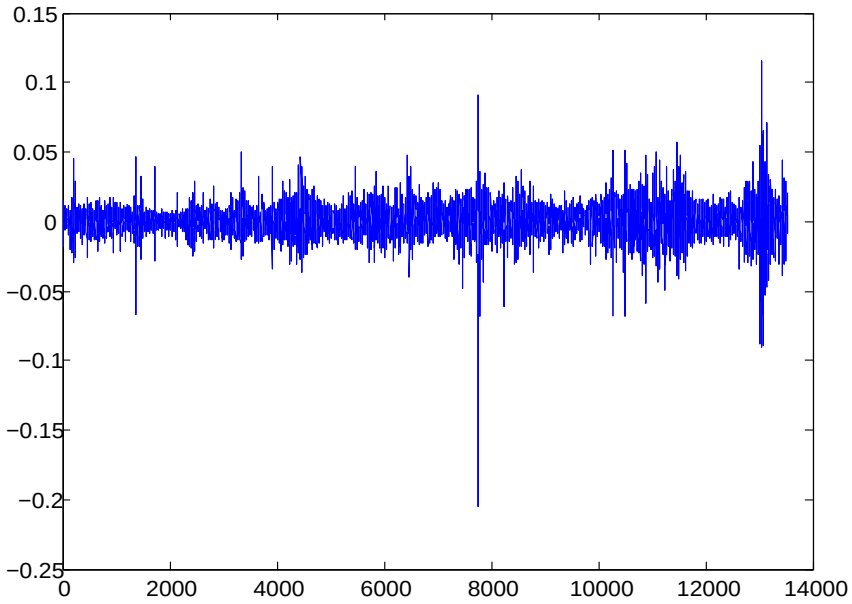
This standardized fourth moment is one natural measure of tail thickness, but many sources will define the coefficient of kurtosis by subtracting three, the reason for this being that three is the value in the population for the *Normal distribution,* (Chapter 7), so that by subtracting three we obtain a positive number if the kurtosis exceeds that of the Normal. Some sources will describe this as the coefficient of *excess* kurtosis.

D3.11 Coefficient of (excess) kurtosis: $\frac{(N-1)^{-1}\sum_{i=1}^{N}(x_i-\overline{X})^4}{s^4} - 3.$

This then is useful as a measure of the frequency of extreme events relative to the Normal distribution, for which, again, the theoretical quantity estimated by D3.10 is equal to three; therefore D3.11 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 S_t 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 $\overline{X}$ 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 day's 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 later chapters, 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.

D3.12 Sample covariance between variables X and Y: $(N-1)^{-1} \sum_{i=1}^{N} (x_i - \overline{X})(y_i - \overline{Y})$.

Note the analogy to the definition of the sample variance above: in that definition the term $(x_i - \overline{X})$ appears multiplied by itself (that is, squared) where is 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.

D3.13 Sample correlation between variables X and Y:

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

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. Often, numerous different explanations can be suggested for an observation, 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 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 ($i = \pi + \overline{r}$, in a standard notation).[16] 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 some elements 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 U.S. real 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

[16] By a 'real' interest rate we mean a return measured in purchasing power, rather than in nominal currency units; for example, if the annual inflation rate is 5% and we obtain a return in currency units of 9% on a one-year bond, then the return in units of purchasing power is $1.09/1.05 - 1 \simeq 1.038$ or about 3.8%. We received 9% more currency on our investment, but currency lost 5% of its purchasing power, so we can only buy about 3.8% more after our one-year investment.

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 or randomness, and does not reflect a genuine relationship. 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, Albert Einstein produced his theories of special relativity and general relativity, which implied that there would be deviations (very small, within the range of size and velocity usually observable) 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 detectable 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 Newton's theory which could be accounted for by the relativistic theories.

In this case it became 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, I've noticed that I'm 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 a given speeding driver being given 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, as the discussion above implies, 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.

This asymmetry between proving things true and proving things false also shows up 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 in the current case, but would provide better predictions in other circumstances. When we fail to reject the hypothesis, we learn 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[17]) writes,[18]

> 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

[17] A footnote is omitted from this quotation.

[18] 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. Wrong.

singular statements. Consequently it is possible 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 Conflating empirical observation and interpretation

A common kind of invalid inference arises from combining an empirical observation with an interpretation, and concluding that the reliability of the empirical observation implies the reliability of the whole package, with the interpretation as well.

Here is an example from a recent news story. A photographer has made frequent trips over many years to the Antarctic, documenting snow cover and animal populations. One of the things that the photographer observes is that a penguin of particular species, after falling over, will get back up *and look around.* This is apparently a very common behaviour pattern, and the observation (coming from someone who has visited the Antarctic over many years) is probably reliable.

The photographer then interprets the penguin looking around as the penguin looking to see whether anyone was looking, and further, that the penguin has a sense of humour. So we now have an empirical observation (reliable) coupled with an interpretation (fanciful); of course other interpretations of the penguin's behaviour are possible, such as that the penguin was looking around to see if it had slipped on something, or anything hit it, or perhaps it wanted to take its bearings after becoming disoriented. But we are encouraged (as the story is written, anyway–it's not clear to me how seriously anyone believed this) to think that a reliable observer with many years experience has observed penguins looking over their shoulders to see if any other penguins were looking as they fell, and perhaps even demonstrating a sense of humour, and that this entire package has been reliably established. But this is not so: only the looking around is a reliable empirical observation.

An example from economics might be that of a behavioural or experimental economist who, in laboratory experiments with human subjects, observes a particular pattern repeatedly. He or she then makes an interpretation of that pattern with respect to some cognitive bias in the subjects, for example. The careful researcher will distinguish the pattern observed in experiments, which may be repeatable and reliable, from the interpretation that is made of these experiments; but if the claim is made that the given interpretation has been proven to be correct through repeated experiments, then that is an invalid claim. The existence of the behaviour pattern may have been established empirically; the reasons for the pattern's occurrence have not. So a careful researcher will say something along the lines of: 'We reliably observe that agents display behaviour pattern X. One possible interpretation of this is ...'

Notice that we are in a situation of asymmetry once again, and as usual. In the two cases just noted, an investigator has posited an explanation – a hypothesis – based on an empirical observation. It would be a mistake to conclude that one has proven the hypothesis true based on this set of observations, for reasons discussed earlier; other explanations will in principle be possible. By contrast, if the empirical

observation is used to refute a hypothesis, then the argument is legitimate if indeed the observation is reliable and is incompatible with the hypothesis.

A very well known example is the Allais paradox. In a classic set of experiments, Maurice Allais (c. 1953) checked conformity of human subjects with some predictions of expected utility theory, by offering participants a choice of two different gambles with known probabilities associated with each of the outcomes, in each of two different experiments or settings.[19] It is straightforward to show that, although utility functions differ across individuals and so the preferred gamble may differ, an expected-utility-maximizing agent who prefers one of the gambles in the first experiment should prefer the corresponding gamble in the second experiment. (In other words, while we don't know agents' utility functions, choosing gamble A (alternatively, B) in the first experiment gives us enough information about the utility function to conclude that the agent should choose gamble A (alternatively, B) in the second experiment as well; the two experiments differ in such a way that the difference between them would not change the choice of an agent who maximizes expected utility.) While there is of course variation across individuals, this experiment has been repeated many, many times in many different places, and it is a reliable observation that a majority of subjects do not conform with this pattern, that is, they do not make the choice of gamble in the second experiment which must correspond with their first-experiment choice, if they are maximizers of expected utility.

In this case the set of experiments is a valid refutation of the idea that agents conform precisely with the predictions of expected utility theory. It is a rejection rather than an affirmation that a theory is true; the empirical observation that people do not conform with the predicted pattern is sufficient to reject the (literal)[20] truth of the hypothesis.

4.5 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.

[19]The form of these experiments has been described in so many places that I won't write out the structure here.

[20]Expected utility maximization does seem to be quite a reasonable approximation in many contexts; the existence of anomalies such as these tells us however that the theory cannot be literally true, just as Newtonian mechanics provides an excellent approximation to the behaviour of physical objects through a wide range of speeds and scales, but not for all cases.

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[21] and equally probable. If n_A of these outcomes have a characteristic A, the *classical probability* of an outcome with characteristic A is $\frac{n_A}{n}$.

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 $\frac{1}{2}$. If we roll a six-sided die and each of the integers $1, 2, \ldots, 6$ is equally likely to arise, then the probability of any one of these outcomes is $\frac{1}{6}$; the probability of an even number arising is $\frac{3}{6}$ or $\frac{1}{2}$; of a number greater than 4 arising is $\frac{2}{6} = \frac{1}{3}$, and so on. In picking a card from a standard deck at random, the probability of a club is $\frac{13}{52} = \frac{1}{4}$; the probability of picking out a Queen is $\frac{4}{52} = \frac{1}{13}$.[22]

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 $2, 3, \ldots, 12$. 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

[21] Mutually exclusive: any one outcome excludes the possibility of any other outcome.
[22] 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.

probability of a 2 is $\frac{1}{36}$. 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 $\frac{1}{6}$. Similarly, imagine that a couple decides to have two children. What is the probability that they will both be girls? The possible 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 $\frac{1}{3}$). 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 $\frac{1}{4}$.

Now what if the probabilities of a boy or a girl being born are not exactly $\frac{1}{2}$ 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 5.3.1: 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 $\frac{n_A}{n}$.

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.[23] The proportion of newborns who 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

[23]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 $\frac{1}{2}$, although the child's sex was already determined with certainty and would have been observable to someone with the right technology.

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'. 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 $B \supset A$ 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 $\overline{A}$, consists of all elements of the space 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 $\overline{(A \cap B \cap C)} = \overline{A} \cup \overline{B} \cup \overline{C}$

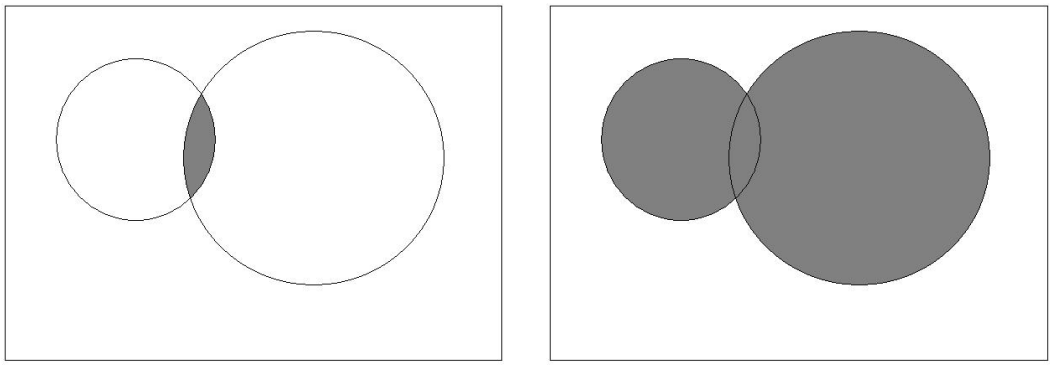
While we have used only the sets A, B, C in T5.3.1, we could extend the theorem to cover any number of sets. For example, the first part of T5.3.1 (iv) can be written for ℓ sets, $A_1, A_2, \ldots, A_\ell$ as

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

These theorems and others are often illustrated and visualized using *Venn diagrams.*[24] 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

$A \cap B$ $\qquad\qquad$ $A \cup B$



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

FIGURE 5.3.2 A/B
Intersection and union of three sets
$A \cap B \cap C$ $A \cup B \cup C$

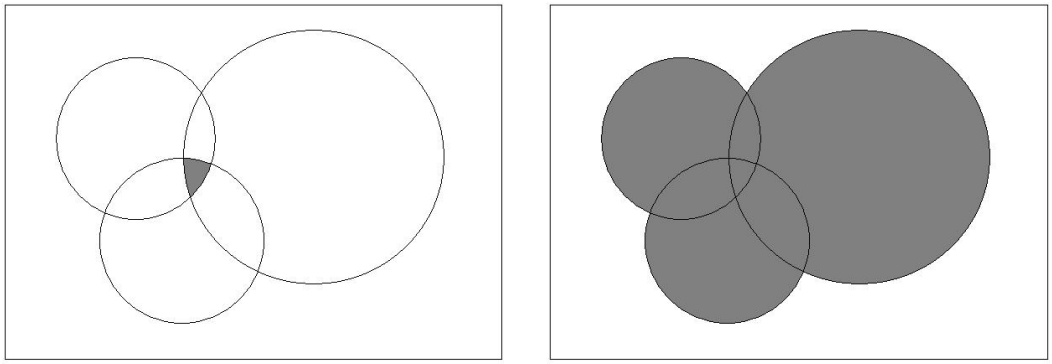
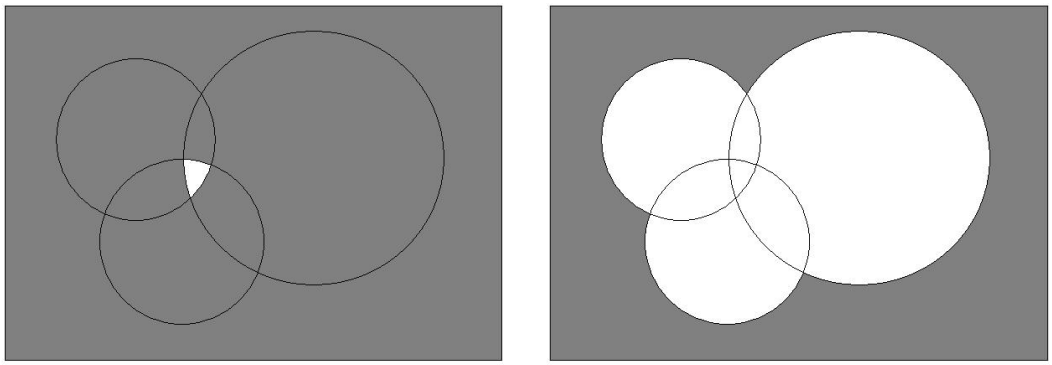


FIGURE 5.3.3 A/B
Complements of intersection and union of three sets
$\overline{(A \cap B \cap C)}$ $\overline{(A \cup B \cup C)}$



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.[25] First we will need to make further definitions

[25] 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.

of concepts directly related to probability, and will need to write down axioms which correspond with our concept of probability.

Using the set-theoretic definitions above, we can define events, sample spaces, probability 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, 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 sample 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. $A \cap B = \phi$.

Definition 5.4.4: 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 ϕ 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, $A_i \cap A_j = \phi \; \forall i \neq j$. (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, from which we will later derive rules for manipulating probabilities that will allow us to solve specific problems (that is, will allow us to compute probabilities from a probability function, given specific conditions).

Definition 5.4.5: A *probability function* $P(.)$ is a function defined on a *field*, $\mathcal{F}$, of events, 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 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 *field* or *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 $\mathcal{F}$. If we want to allow a potentially infinite set of events, then we can define instead a *sigma-field* or σ-field, in which we have properties (i) and (iii) above and also

$$\text{if } A_1, A_2, \ldots \in \mathcal{F} \text{ then } \mathcal{B} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \ldots \in \mathcal{F}.$$

In the rest of this chapter, we will assume that any events discussed come from a probability space, defined on a field or 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.

We will now treat the three parts of D5.4.5 as the axioms of a formal system, from which we can derive theorems. By applying valid reasoning (logical or mathematical operations) to the axioms, new statements that we derive from the axioms (the theorems of the system) are logically implied by the axioms, and are therefore as reliable as the axioms themselves, and may be used in proving further theorems without returning to the axioms alone in each case. A full set of theorems of the system provides, in this case, a set of rules for defining and determining probabilities which assumes no more than the axioms stated in D5.4.5. These theorems can then be used directly in solving problems involving probabilities, since they provide additional valid statements concerning probability functions.

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 $P(A) = \frac{n_A}{n}$.

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 a mutually exclusive and exhaustive set of events. Then $P(A) + P(B) = 1$.

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

vii $P(\phi) = 0$.

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 $m \cdot (m-1) \cdot \ldots \cdot 1 = \Pi_{j=0}^{m-1}(m-j) = \Pi_{i=1}^{m} i$, 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 ${}_nP_k = \frac{n!}{(n-k)!}$.

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.

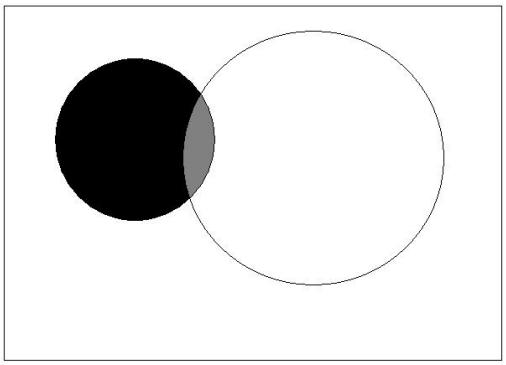
The definition of conditional probability is given in D5.5.1 and is easily illustrated with a Venn diagram, as in Figure 5.5.1, if we think of areas as representing probabilities.

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[26] 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:

Theorem 5.5.1 Bayes' Theorem. Let A and B be two events in 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.)

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

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 $\frac{1}{200}$: 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, $\frac{1}{200}$. 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 $\frac{1}{200}$.

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.[27] 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) Where $P(A) > 0$, $P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$.

[27] 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.

ii For $P(A_1 \cap A_2 \cap \ldots \cap A_\ell) > 0$, we have $P(A_1 \cap A_2 \cap \ldots \cap A_\ell) = P(A_\ell | A_1 \cap A_2 \cap \ldots \cap A_{\ell-1}) \cdot P(A_{\ell-1} | A_1 \cap A_2 \cap \ldots \cap A_{\ell-2}) \cdot \ldots P(A_3 | A_1 \cap A_2) \cdot P(A_2 | A_1) \cdot P(A_1)$. 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

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

Appendix to Chapter 5

Examples of proofs of some simple theorems of the system

Recall:

Definition 5.4.5: A *probability function* $P(.)$ is a function defined on a *field*, $\mathcal{F}$, of events, 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 $P(B) = \sum_{i=1}^{\ell} P(A_i)$.

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) 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 consist 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 $P(A) = \frac{n_A}{n}$.

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 5.5.2, (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. $\gamma \in \overline{(A \cup B)}$. Then $\gamma \in \overline{(A \cup B)} \Rightarrow \gamma \notin (A \cup B)$. 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 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; recall the definitions of sample space, probability function and field given in Definitions 5.4.1, 5.4.5 and 5.4.6. The following definition makes D6.1.1 somewhat more precise, and uses the definition of a probability space, composed of sample space, field of events and probability function, $(\Omega, \mathcal{F}, \mathcal{P})$. Recall that any ω is an outcome from the sample space Ω.

D6.1.2 A real-valued *random variable* X is a (real-valued) function on a probability space, $X : \Omega \to \mathcal{R}$, such that the set $\{\omega : X(\omega) \leq \ell\}$ belongs to $\mathcal{F}$ for all $\ell \in \mathcal{R}$.

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 $F_X(x) = P(\{\omega : X(\omega) \leq x\})\ \forall x \in \mathcal{R}$, again making reference to the elements ω of the sample space Ω on which the random variable is defined.[28]

The cdf is sometimes referred to, somewhat informally, as a 'distribution function', omitting 'cumulative.' 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:

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

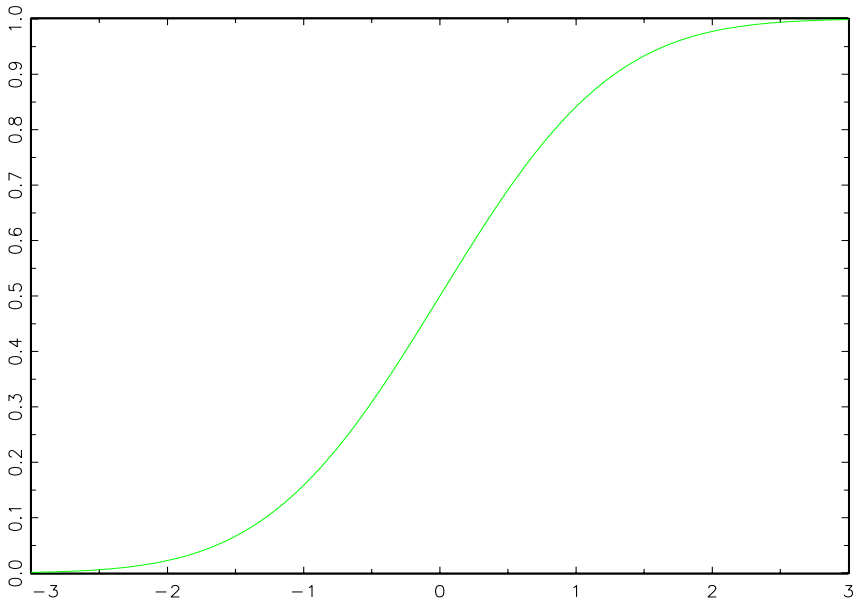
The first statement corresponds with probabilities bounded into the interval $[0, 1]$; no matter how small x is the probability of being at or below it cannot be

[28] 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.

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.

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.[29] (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.

The cdf is a theoretical or population quantity, referring as it does to the true

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

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 $x^{(j+i)} \geq x^{(j)}\ \forall i > 0$. 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.

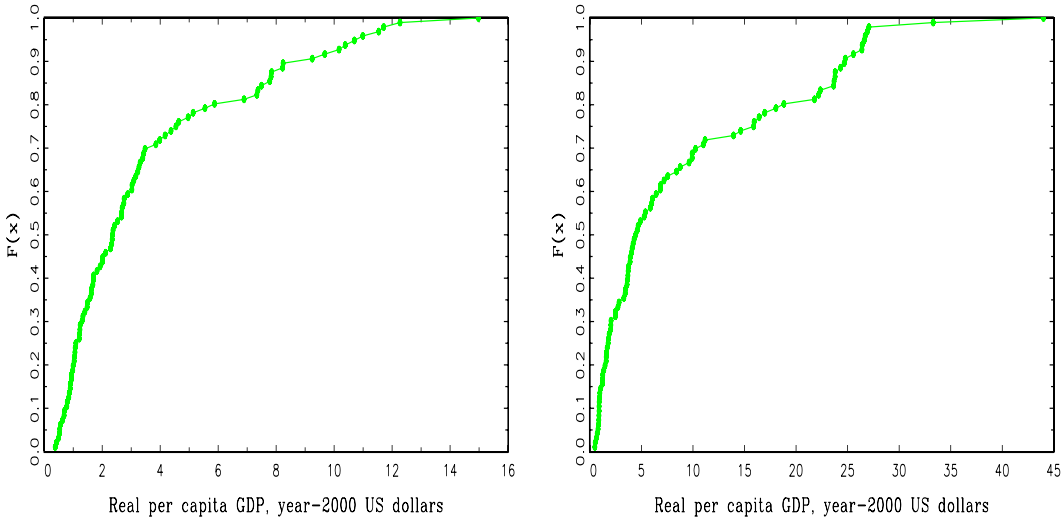
We can now define the empirical cdf.

Definition 6.1.5 The *empirical cumulative distribution function* or *sample cumulative distribution function* is defined at each order statistic as $\hat{F}_X(x^{(i)}) \equiv \frac{n_i}{n}$, 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.

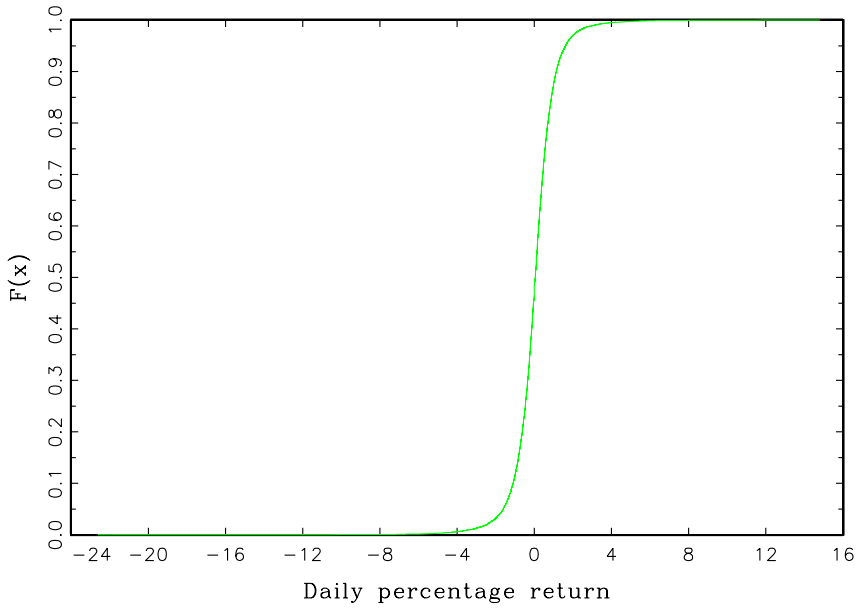
FIGURE 6.1.2
Empirical cdf of real GDP, 96 countries, 1960 and 2000



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.

FIGURE 6.1.3
Empirical cdf of daily percentage returns
Dow Jones Industrial Average, January-1915–March 2006



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: 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 $\{0, 1, 2, \ldots\}$. 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.[30]

The first type of random variable will be called *continuous;* the word 'continuous'

[30]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 $3\frac{1}{2}$. 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.

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 function* $f_X(x)$ 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 a the hypothetical probability that a randomly selected member of a population will be taking a given number of medications at the time of sampling.[31] (Although 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 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.

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

FIGURE 6.2.1A/B
Probability function for number of medications currently used
A: females 12 years + B: males 12 years +

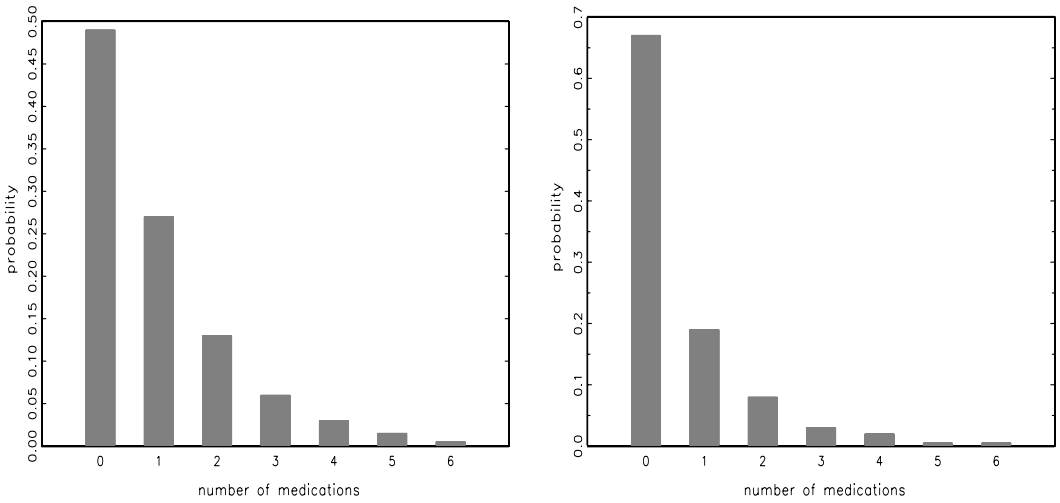
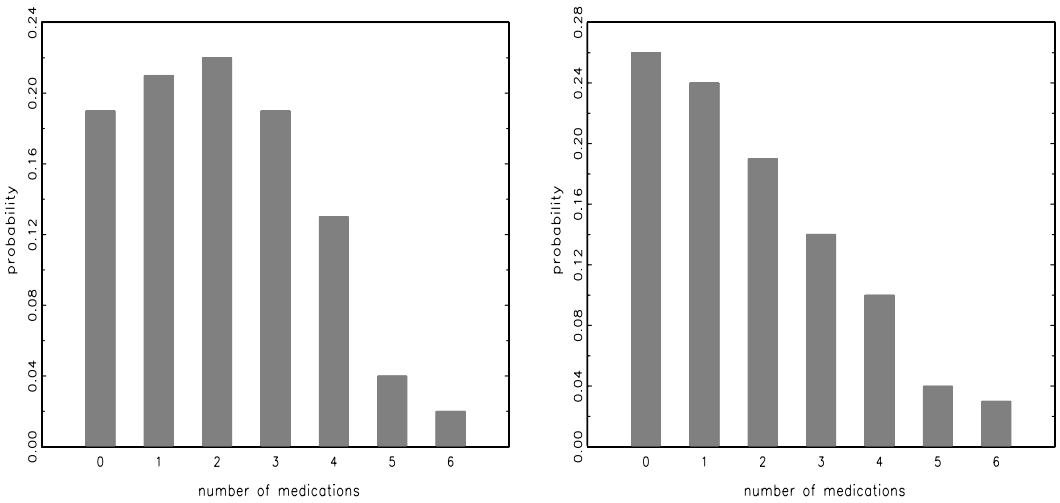


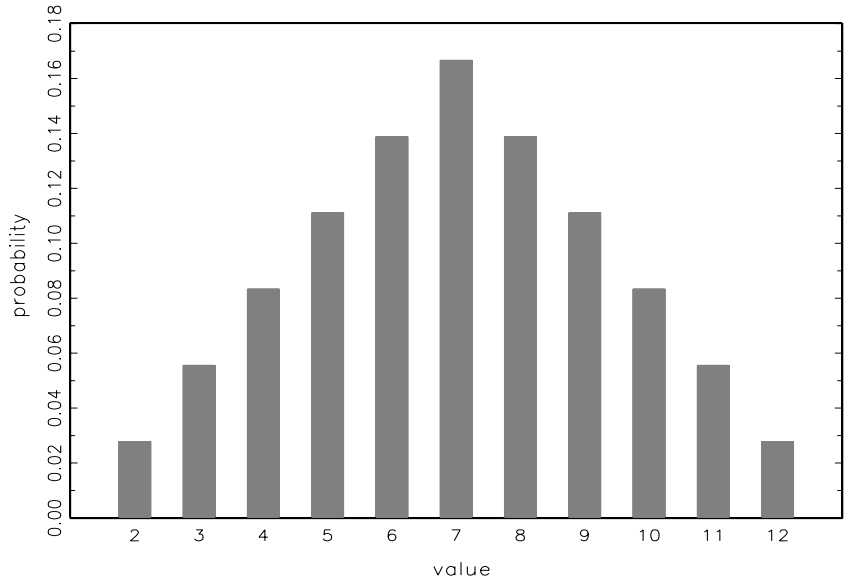
FIGURE 6.2.1C/D
Probability function for number of medications currently used
A: females 75 years + B: males 75 years +



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 function for the sum of the values on two dice, which we worked out in Chapter 5. This probability function is depicted in Figure 6.2.2.

FIGURE 6.2.2
Probability function of sum of values on two standard dice



Notice that the probability 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

$$P(a \leq X < b) = \int_a^b f_X(x)dx,$$

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 $\int_{-\infty}^{x} f_X(x)dx = F_X(x)$, where $F_X(x)$ is the cumulative distribution function of X.

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 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 $x \notin [a, b]$
- $\int_a^b f_X(x)dx = 1$
- $F_X(x) = \int_a^x f_X(x)dx$.

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, $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

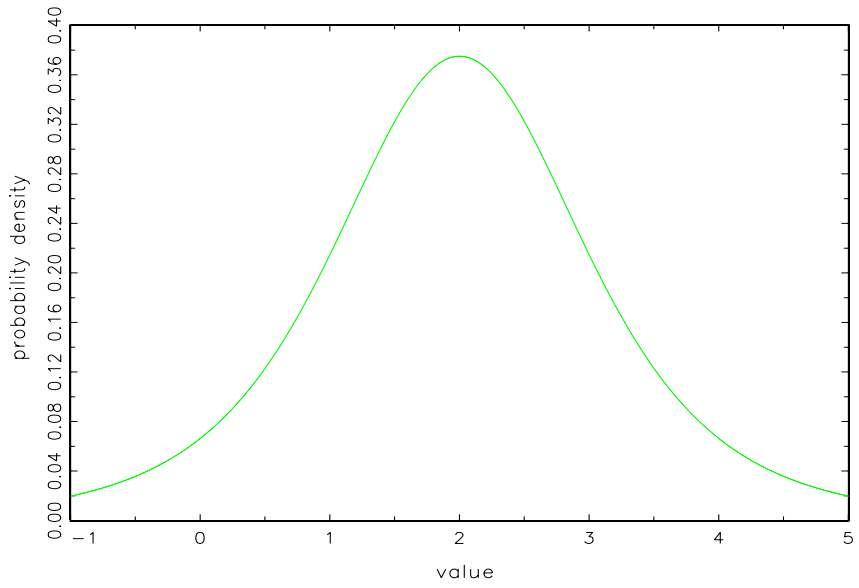
Note that these definitions do not state explicitly, but imply, that $P_X(x_i) \leq 1$ and $\int_c^d f_X(x)dx \leq 1$ for any x_j or interval $[c, d]$.

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 6.2.3
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

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

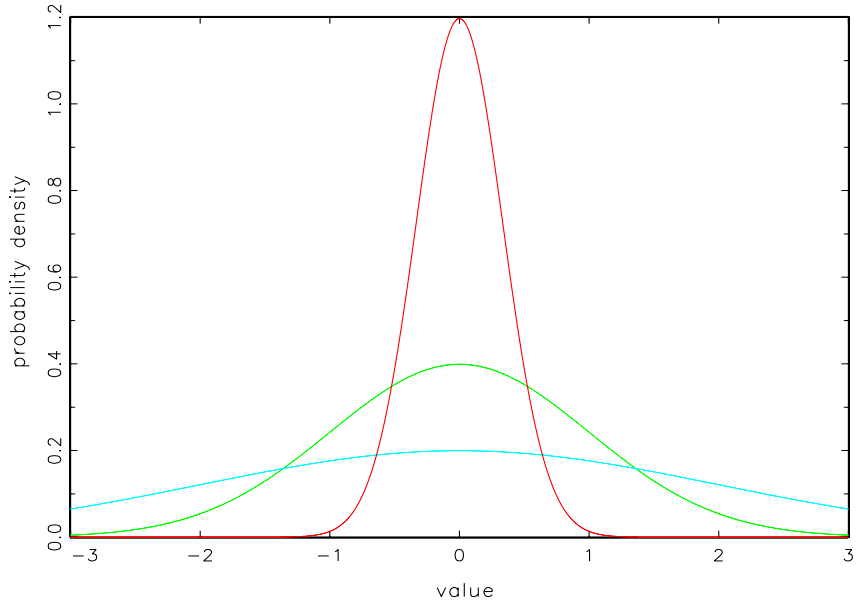
where μ and σ^2 are the mean and variance of the random variable. Any finite mean and any finite positive variance are compatible with a Normal distribution, and when these two values are specified, the particular Normal density is fully specified: that is, different densities can be Normal, because 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 $N(\mu, \sigma^2)$, 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.[32] 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.

[32]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
Probability density functions of Normal random variables, differing in variance



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 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.

6.3 Estimating probability and probability density functions

We saw earlier that the empirical or estimated CDF, $\hat{F}_X(x)$, can be obtained straightforwardly from the order statistics of the data, and the proportion of the sample lying at or below particular points.

Estimation of a probability function is similarly straightforward. Consider a discrete random variable X which can take on any one of k values, or for which k different values are observed in the sample. We can index these by $i = 1, \ldots k$. Let any one of the values that X can take be x_i. Then the sample probability that $X = x_i$ is n_i/N, where there are N sample points in total, and n_i of them for which $X = x_i$. (Note that this is just a re-appearance of one of our first rules for determining probabilities; in the notation of Chapter 5, let the event A be that $X = x_i$; then $p(A) = n_A/N$ where n_A is the same as n_i here, i.e. the number of observations

for which A is true.) The possible values for the discrete random variable form a mutually exclusive and exhaustive set, so the sum of their probabilities is equal to 1; similarly $\sum_{i=1}^{k} n_i = N$ and so $\sum_{i=1}^{k} n_i/N = N/N = 1$ and then $\sum_{i=1}^{k} \hat{p}_i = 1$: the sum of the probabilities of all things that can happen is 1.

Estimation of density functions is not so straightforward, and there is a very large literature on methods for doing this. We will describe one widely used method, kernel density estimation, which, when the choice of bandwidth is made carefully, can produce smooth and efficient (relatively low variance) estimates of the density function.

To begin, recall that the density is the derivative of the CDF, so that estimation of the density is estimation of the slope of the CDF. A simple slope estimator is to take the increase in a function over an interval divided by the width of the interval, and in the limit as width of the interval goes to zero, we obtain the derivative. For a random variable X, a point x where we will evaluate the slope, and an interval of width h centred at the point x, we would have that the slope is

$$f_X(x) = \lim_{h \to 0} (1/h) P(x - h/2 < X < x + h/2), \tag{6.3.1}$$

where $P(x-h/2 < X < x+h/2)$ is the probability that X lies in the interval $x \pm h/2$. Notice that we divide by h to get a slope over an interval: the increase is divided by the width of the interval over which the increase arose.

We want to obtain an estimate of this quantity, $\hat{f}_X(x)$. Note that we are describing an estimate at a single point, x; by re-evaluating the estimator at a sequence of different choices of x, we can plot the full estimated density.

To provide a first estimate, we can use an indicator function. An indicator function takes the value 1 if a condition holds, 0 if it does not. In this case the condition will be that a value lies in the interval $x \pm h/2$, and the indicator function is $\mathcal{I}[x_i \in x \pm h/2]$. So our first density estimator simply takes the proportion of the N observations in the interval as our estimator of the probability of the random variable lying in the interval, to give:

$$\hat{f}_X(x) = (1/Nh) \sum_{i=1}^{N} \mathcal{I}[x_i \in x \pm h/2] = (1/Nh) \sum_{i=1}^{N} \mathcal{I}[x - h/2 < x_i < x + h/2]. \tag{6.3.2}$$

Notice that we now divide by N as well as h because we are taking the proportion of the sample points in the interval, and the indicator function counts those points; in equation (6.3.1), the fact that we refer to a probability means that we have the theoretical analogue of the sample proportion, and in (6.3.2) we take the actual sample proportion, which entails counting and dividing by N.

This can provide a reasonable estimate, but notice that the weight that a point gets in estimating density at point x is either one or zero; if it lies in the interval it counts as 1 regardless of how close it is to the centre, whereas if it falls just slightly

outside the interval it counts as zero. We might think that we can obtain a better estimator by allowing the weight to be at its maximum for a point right near the centre of the interval (the point at which we're evaluating the density, x), with the weight declining gradually as we move away from x. We can do so, and various choices of weighting function have been proposed. These functions are called 'kernels', and the kernel density function estimator replaces the indicator function with a kernel function:

$$\hat{f}_X(x) = (1/Nh) \sum_{i=1}^{N} K\left[(x_i - x)/h\right],$$

where h is the bandwidth (or window-width, or smoothing) parameter and the kernel function $K(.)$ is a real, positive function such that $\int K(z)dz = 1$. Since density functions possess these properties, often a density such as the standard Normal is chosen as a kernel weighting function, but common alternative choices are the Bartlett (triangular, linearly declining weights) or Epanechnikov (inverse parabola). In each of the latter cases the weights decline to zero but cannot become negative. The conditions on $K(.)$ just given also ensure that $\hat{f}_X(x)$ inherits these properties of a density function.

Bandwidth choice is critical in kernel density estimation, and too large a bandwidth ('oversmoothing') will distort the location of the data and shape of the distribution (as well as, in extreme cases, making the estimated density of any data set simply resemble the shape of the kernel). There is a large literature on how to make the choice, but whatever method is chosen, the bandwidth should tend to get smaller as the sample size gets larger; for consistent estimation (that is, the estimated density will converge on the true density as sample size increases), the bandwidth will have to shrink toward zero as sample size increases.[33]

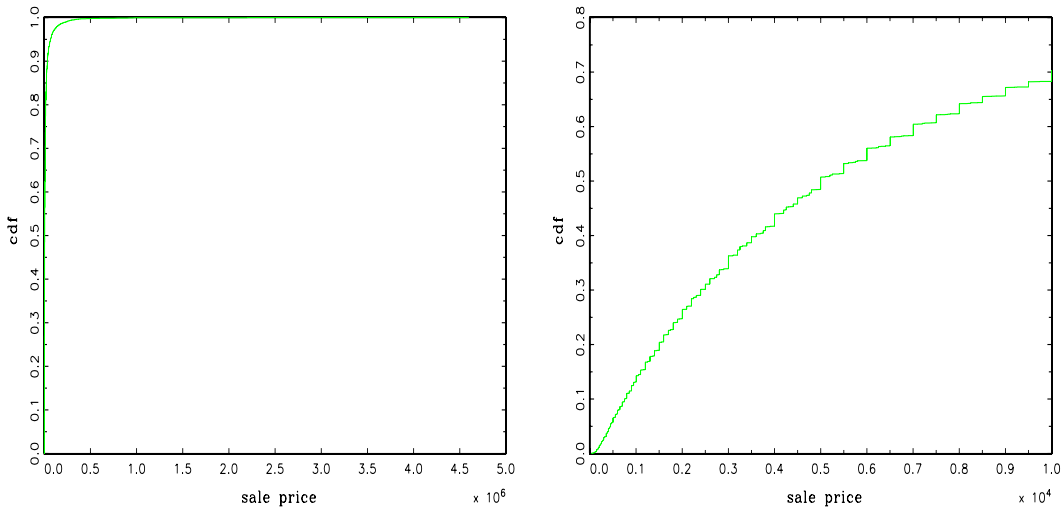
For an empirical example, consider the data represented in Figures 6.2.4A and 6.2.4B, prices paid at selected auctions, 1968–2001, for paintings by any of a set of 152 Canadian artists; there are approximately 8000 observed sale prices in this sample.[34] Figure 6.2.4A gives the full empirical cdf of 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

[33]The property of consistency of an estimator will be defined formally below in Chapter 13.

[34]See Hodgson and Vorkink (2004).

cdf, comprising paintings which sold for $10000 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 $5000. or below ($F(5000) \simeq 0.5$), and about two thirds at $10000 or below.

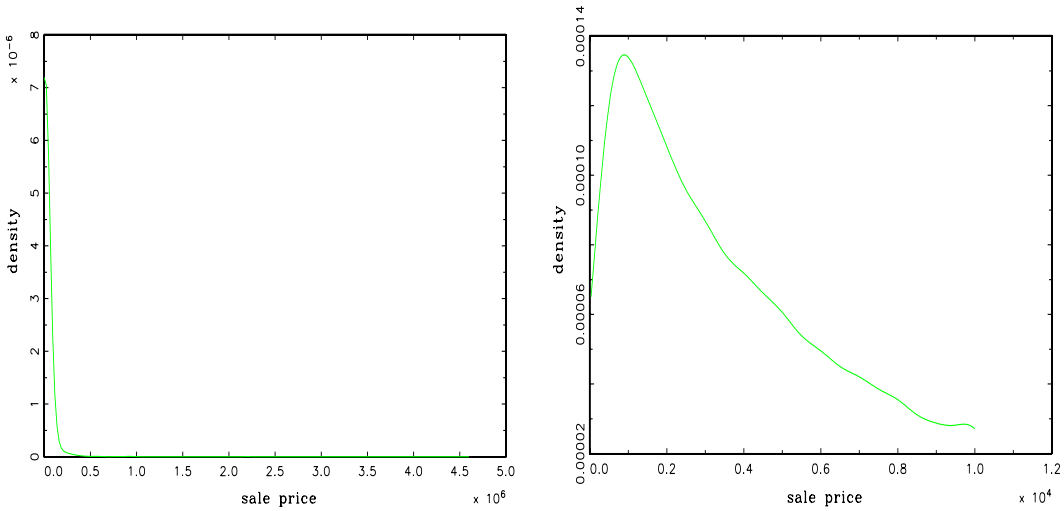
FIGURE 6.2.4A/B
Empirical cdf of sale prices
Canadian paintings sold at selected auctions, 1968-2001
A: full sample B: prices up to $10,000 only



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 prices tend to be at particular numbers. For example, there are many sales at exactly $5000., but none at prices between $5001. and $5099.

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 $1000, and that the density of prices drops steadily from that point.

FIGURE 6.2.4C/D
Estimated pdf of sale prices
Canadian paintings sold at selected auctions, 1968-2001
C: full sample D: prices up to $10,000 only



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.[35]

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

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

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 $2, 3, \ldots 12$ that we

[35] 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.

obtained earlier, is $2(\frac{1}{36}) + 3(\frac{2}{36}) + 4(\frac{3}{36}) + 5(\frac{4}{36}) + 6(\frac{5}{36}) + 7(\frac{6}{36}) + 8(\frac{5}{36}) + 9(\frac{4}{36}) + 10(\frac{3}{36}) + 11(\frac{2}{36}) + 12(\frac{1}{36}) = (\frac{252}{36}) = 7$.

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.

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: $\overline{X} = (\frac{1}{N}) \sum_{i=1}^{N} x_i$. 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 $\frac{3}{2}$ 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

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

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 $E(X)$, 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 $E(X) = \mu_X$, and let a, b be constants. Then:

i $E(a + bX) = a + bE(X) = a + b\mu_X$

ii (Jensen's inequality) Let $g(.)$ be a convex function.[36] Then $E(g(X)) \geq g(E(X))$.

[36] 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.

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. 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 $E(X^2) \geq (E(X))^2$: 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 $E(X) = 0$, so $(E(X))^2 = 0$, while $E(X^2) = 1$ since the mean of X is 0, so its variance is $E(X^2)$, 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 $E(h(X)) = \sum_{i=1}^{k} p_i h(x_i)$ in the discrete case, and as $E(h(X)) = \int_a^b h(x) f_X(x) dx$ 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 $5 \cdot 2(\frac{1}{36}) + 5 \cdot 3(\frac{2}{36}) + 5 \cdot 4(\frac{3}{36}) + 5 \cdot 5(\frac{4}{36}) + 5 \cdot 6(\frac{5}{36}) + 5 \cdot 7(\frac{6}{36}) + 5 \cdot 8(\frac{5}{36}) + 5 \cdot 9(\frac{4}{36}) + 5 \cdot 10(\frac{3}{36}) + 5 \cdot 11(\frac{2}{36}) + 5 \cdot 12(\frac{1}{36}) = (\frac{5 \cdot 252}{36}) = 5 \cdot 7 = 35$.

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.[37] 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 $E(X^\ell),\ \ell = 1, 2, \ldots$.

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

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

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

whereas for a continuous random variable it is

$$\int_a^b x^\ell f_X(x)dx. \tag{7.2.2}$$

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 $E((X-\mu_X)^\ell)$, $\ell = 2, \ldots$. The second central moment is called the variance, $\sigma_X^2 \equiv E((X-\mu_X)^2)$.

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

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

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.[38] To define these functions, let $m_2, m_3, m_4, \ldots$ be the central moments of a distribution. Then the population variance is $m_2 = E(X-\mu_X)^2$, the population coefficient of skewness corresponding to the sample measure in D3.8 is

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

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

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

[38]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.

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.[39]

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 14–16, 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 $E(X-\mu)^2$ 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 $E(X-\mu)^2 = \int_a^b (x-\mu)^2 f_X(x)dx$. Similar expressions for higher-order moments follow also from the definitions of expectation for discrete and continuous random variables.

Recall that we have $E(a+bX) = a+bE(X) = a+b\mu_X$ 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 $\text{var}(X)$ for the variance of X, that $\text{var}(a+bX) = E[a+bX-E(a+bX)]^2 = E[a+bX-(a+b\mu)]^2 = E[bX-b\mu)]^2 = E[b(X-\mu)]^2 = b^2E[X-\mu]^2 = b^2\text{var}(X)$.

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:

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

[39]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.

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

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

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 1 gives the second, fourth and sixth moments of these distributions, where the ℓ^{th} central moment is computed using equation (7.2.1) as $E(V-0)^\ell = \sum_{i=1}^{4} p_i v_i^\ell$ 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

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

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 $P(|X - \mu| < \ell\sigma) \geq (1 - \frac{1}{\ell^2})$ or: the probability that X is less than ℓ standard deviations from its mean is at least $1 - \frac{1}{\ell^2}$. 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, the Markov inequality, which is stated and proved in the Appendix:

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

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

Appendix to Chapter 7

Proof of Theorems 7.3.2 and 7.3.1.[40]

We begin by proving the Markov inequality: X is a random variable and $g(.)$ a non-negative function on $\mathcal{R}$ such that $E(g(X))$ exists. Then

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

We know that

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx = \int_{X:g(X)<k} g(x) f_X(x) dx + \int_{X:g(X)\geq k} g(x) f_X(x) dx,$$

where we have simply split the range over which the function is integrated into two parts, that for which $g(X)$ is less than some value k, and that for which it is greater than k. Since neither of the integrals can be negative, we drop the first part, and we obtain

$$E(g(X)) \geq \int_{X:g(X)\geq k} g(x) f_X(x) dx \Rightarrow E(g(X)) \geq \int_{X:g(X)\geq k} k f_X(x) dx,$$

where the implication follows from the fact that we're integrating over cases where $k \leq g(X)$, so replacing $g(X)$ by k can only lower the integral, so the inequality continues to hold. But

$$\int_{X:g(X)\geq k} k f_X(x) dx = k \int_{X:g(X)\geq k} f_X(x) dx = kP[g(X) \geq k],$$

where the latter equality follows from the fact that we are simply integrating the probability density function over values for which $g(X) \geq k$, so we obtain the probability that $g(X) \geq k$. Therefore

$$\frac{E(g(X))}{k} \geq P[g(X) \geq k].$$

■

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

In order to prove the Chebychev inequality from this statement, we make a choice of function $g(X)$ and of k, and in particular $g(x) = (X-\mu)^2,\ k = \ell^2\sigma^2$, where μ and σ^2 are the mean and variance of X. Substituting into the last line of the proof above, we have

$$\frac{E(X-\mu)^2}{\ell^2\sigma^2} \geq P[(X-\mu)^2 \geq \ell^2\sigma^2].$$

Since $E(X-\mu)^2 = \sigma^2$, σ^2 cancels from top and bottom on the right-hand side of the inequality, leaving us with

$$\frac{1}{\ell^2} \geq P[(X-\mu)^2 \geq \ell^2\sigma^2], \text{ or}$$

$$P[(X-\mu)^2 \geq \ell^2\sigma^2] \leq \frac{1}{\ell^2}, \text{ or}$$

$$P[|X-\mu| \geq \ell\sigma] \leq \frac{1}{\ell^2}.$$

■

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' incomes and net asset values.[41] Then we 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

[41] 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 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.

(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 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 $(x_i, y_j),\ i, j = 1, 2, 3$.

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% ($\frac{0.03}{0.05}$) 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 ($\frac{0.08}{0.20}$): 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

differ 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, $P(Y = 4|X = 0) = \frac{0.03}{0.50} = 0.06$, slightly higher than the unconditional probability that $Y = 4$; $P(Y = 2|X = -1) = \frac{0.16}{0.25} = 0.64$, 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

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

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

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

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 $E(X-\mu)^2$, 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.[42]

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

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

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.[43]

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

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

Recall that we have used the standard notation σ^2 for the variance of a random variable, so that $\sigma_X, \sigma_Y)$ 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.

---

[42] 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

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

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

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

Theorem 8.4.1 Let W and Z be two random variables with $E(W^2)$ and $E(Z^2) < \infty$. Then $(E(WZ))^2 \leq E(W^2)E(Z^2)$; $(E(WZ))^2 = E(W^2)E(Z^2)$ 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.

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).[44] 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.[45] 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.

[44] Don't try this at home.

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

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

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

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 '$E(Y|X = x)$', we also frequently write simply $E(Y|X)$.

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

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

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 out 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

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

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 $-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)$ 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: $\mu_1 = \mu_2 = 2$

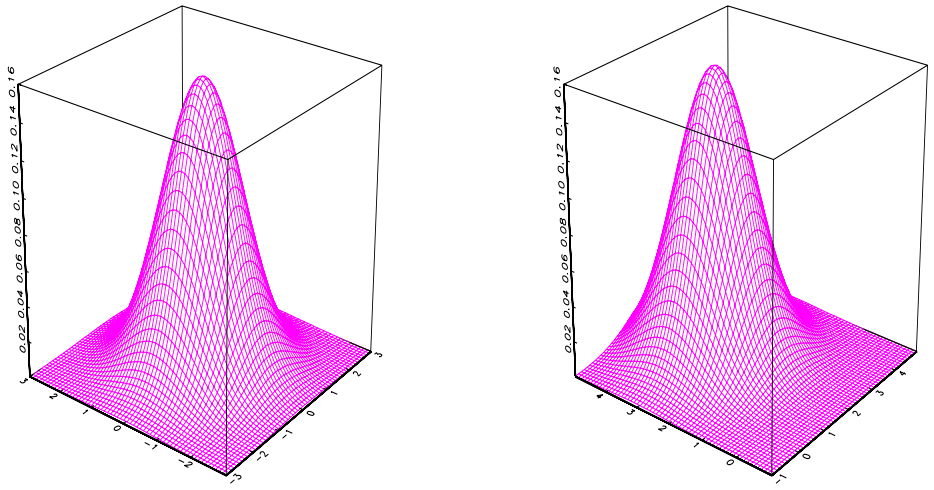


FIGURE 8.1.1.C/D
Bivariate Normal densities
$\mu_1 = \mu_2 = 0;$

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

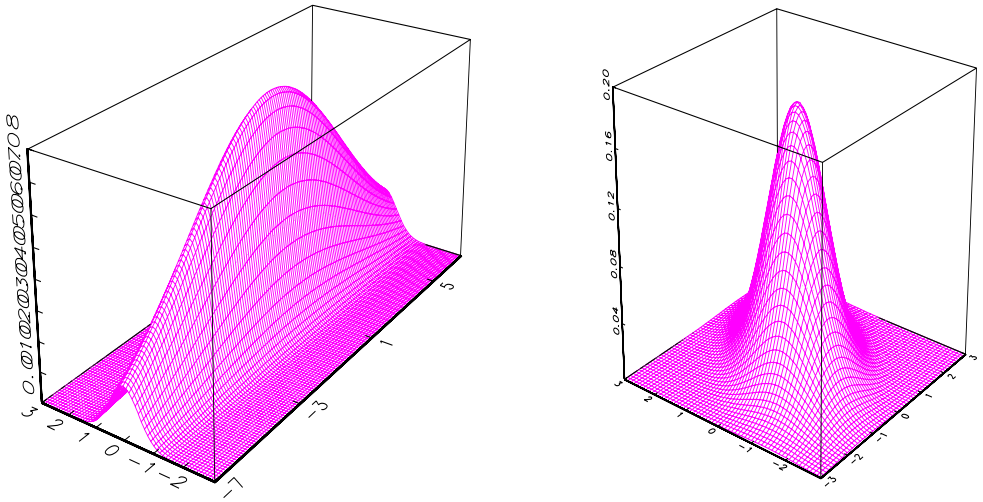
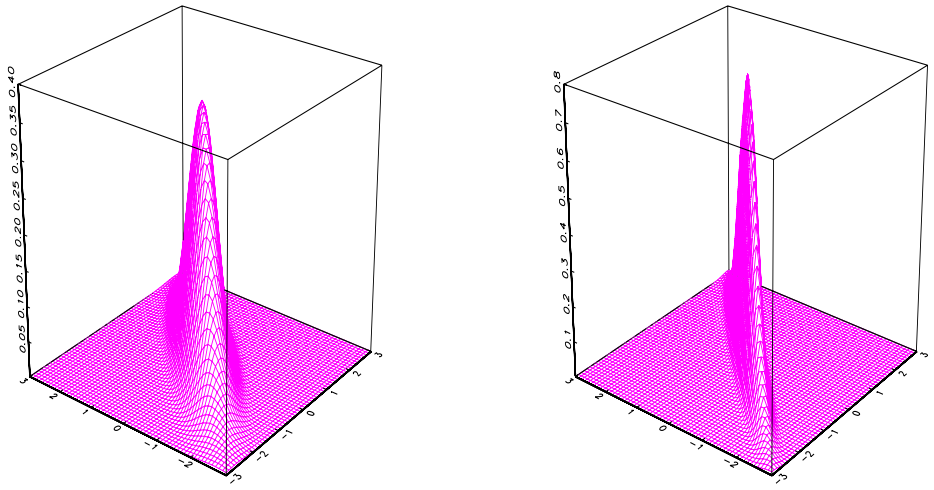


FIGURE 8.1.1.E/F
Bivariate Normal densities
$\mu_1 = \mu_2 = 0;\ \sigma_1^2 = \sigma_2^2 = 1;$
E: $\rho = 0.90$ F: $\rho = 0.98$



joint discrete distribution joint cts: plot joint Normals marginal and conditionals relation to conditional probability conditional probability computations from joint cts distributions

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 CDF's, PDF's or probability functions 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 tells us that, under some conditions, a sample mean will have a distribution that converges toward the Normal as sample size increases; 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 mean and unit variance, will have 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 generalized it to embody information gained from further empirical experience. For example, the $t-$ distribution embodies 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 approx-

imation, there are a number of advantages. We can compute important quantities such as tail area probabilities or probabilities of values occuring 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 mean 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 mean 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, it's 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 Discrete distributions

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

9.1.1 Uniform

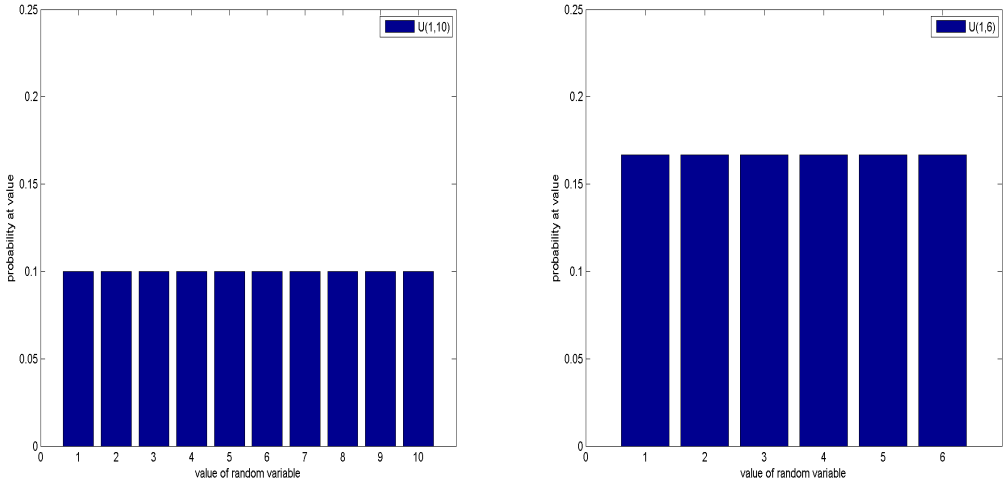
The uniform distribution may describe either discrete or continuous random variables. In the discrete case, it indicates that the probability is the same on each of the finite number of possible outcomes. If we have a uniform distribution on the integers from a to b inclusive, then there are $b-a+1$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

$$p_X(x) = (b-a+1)^{-1},\ x \in I,\ a \leq x \leq b.$$

It is easy to check the condition that $\sum_{x=a}^{x=b} p_X(x) = 1$.

The following figure plots two uniform probability densities, over the integers from 1 to 10 and 1 to 6 respectively. Note that the sum of the probabilities is 1 in each case.

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 $F(.)$. But if we apply $F(.)$ 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 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.

9.1.2 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 it I guess-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 function 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.

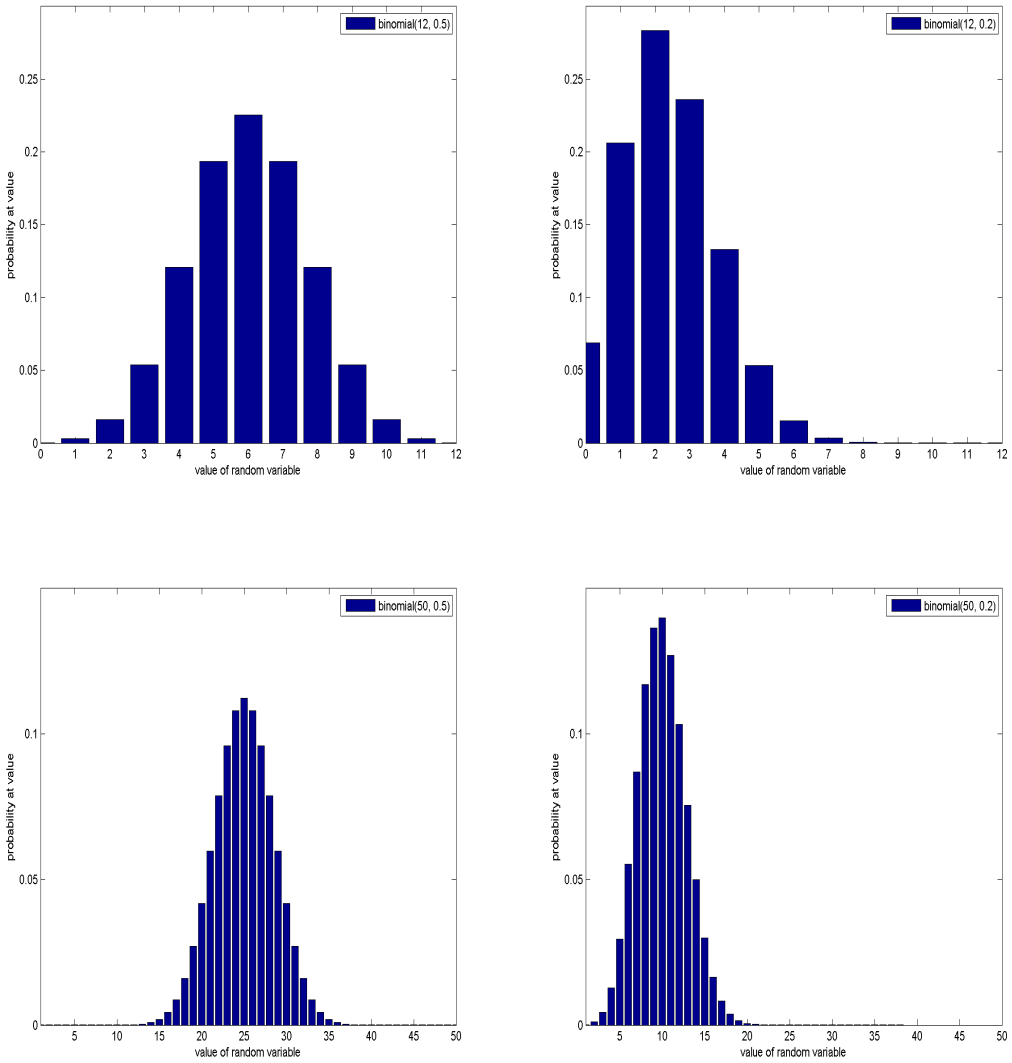
$$P(x) = p^{x}(1-p)^{(1-x)}, \text{ for } x = 0 \text{ or } x = 1,$$

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.

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 second block of graphs, 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.

FIGURE 9.1.2 A–D
Binomial probability functions



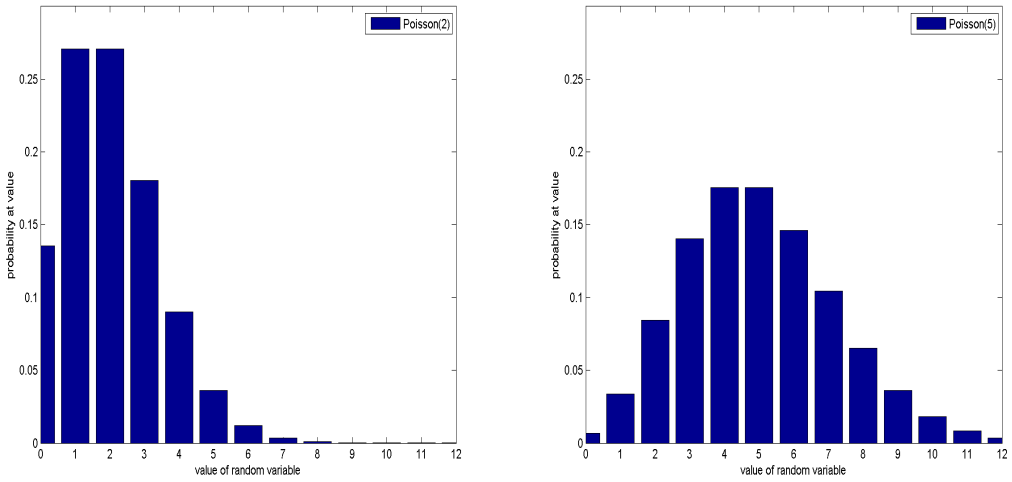
9.1.3 Poisson

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 function for the outcomes $x = 0, 1, 2, \ldots$ is defined as a function of a strictly positive parameter λ :

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

The graphs show probability functions for $\lambda = 2$ and 5.

FIGURE 9.1.3 A/B
Poisson probability functions



9.2 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.

9.2.1 Uniform densities

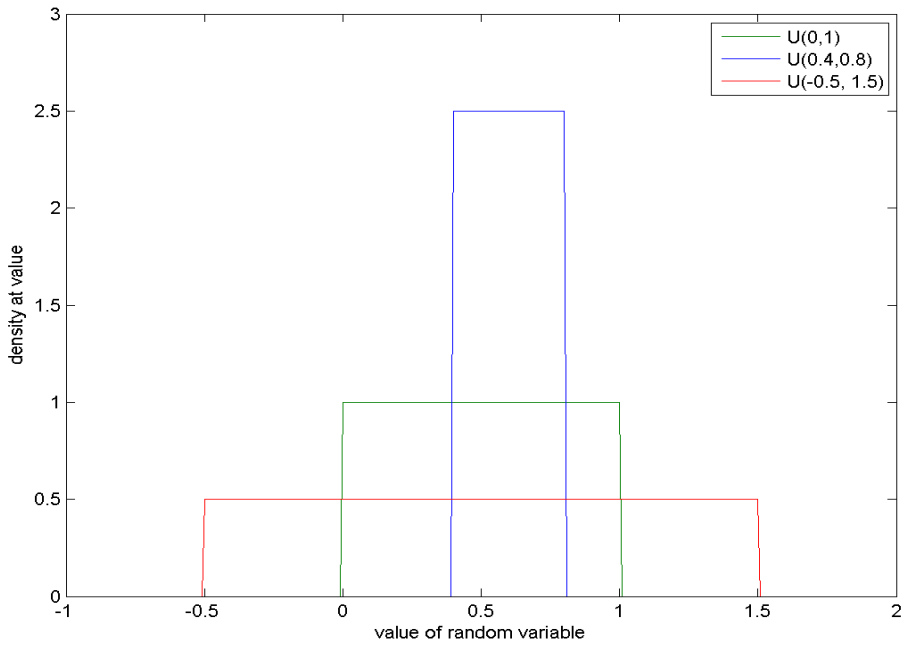
As we said earlier, the Uniform also has a continuous form. Let a random variable X be uniform on the continuous interval $[a, b]$, which we write $U(a, b)$; then its density is

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

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

Notice that in the following graph 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.

FIGURE 9.2.1
Continuous uniform densities



9.2.2 Normal densities

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.

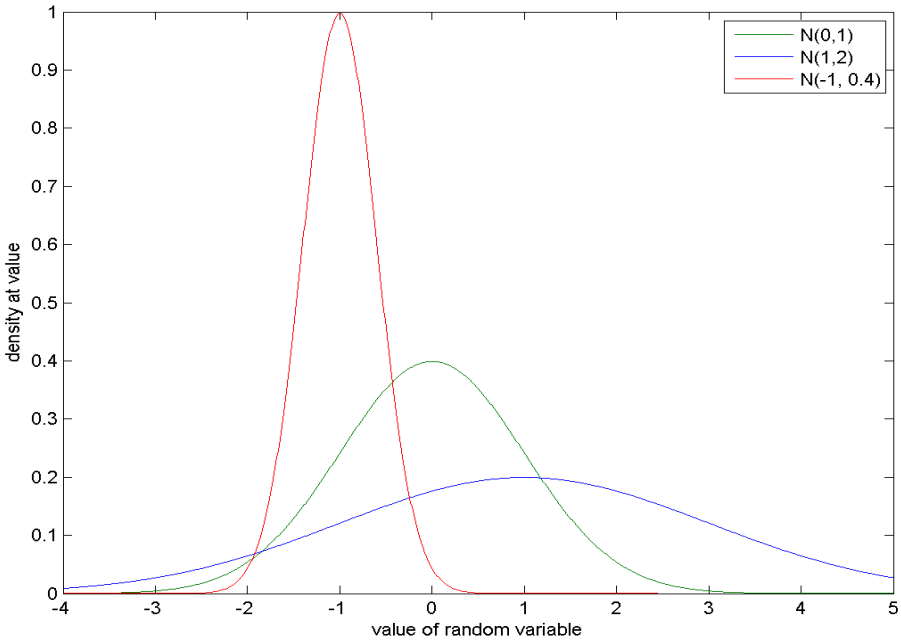
We have already defined the Normal density; again, it is

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

a function of the mean and standard deviation parameters, μ and $\sigma > 0$.

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



9.2.3 (Student's) $t-$ densities

The $t-$ distribution 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 $t_\nu \xrightarrow{D} N(0,1)$ as $\nu \to \infty$.

The $t-$ may be difficult to distinguish visually from the Normal. However, as we have seen elsewhere, the proportion of its probability mass line 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 were 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:

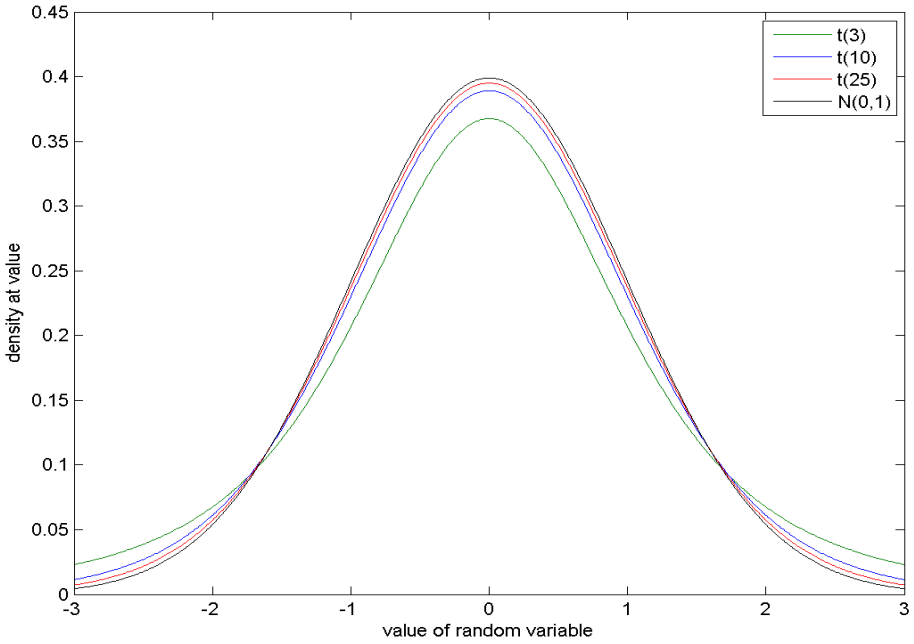
$$f_X(x) = (\nu\pi)^{-\frac{1}{2}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{(-\frac{\nu+1}{2})},$$

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

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

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 $\Gamma(k) = (k-1)!$.

FIGURE 9.2.3
$t-$ densities, varying degrees of freedom, compared with N(0,1)



9.2.4 χ^2 ('Chi-squared') densities

The χ^2 distribution also 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 theorem stew imply asymptotic normality therefore implies a correspondingly wide applicability of joint tests combining several of these variables, yielding χ^2 distributions.

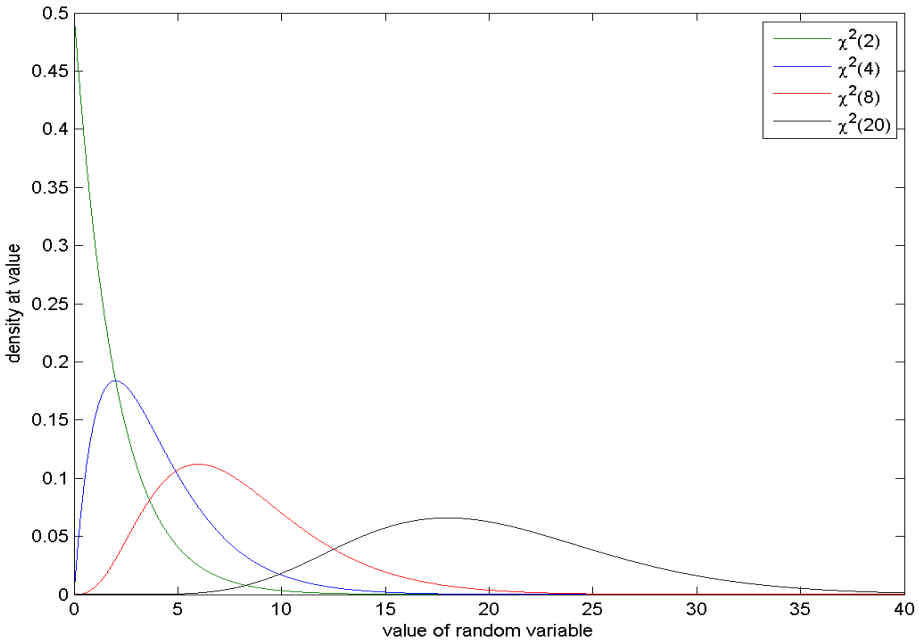
Like the $t-$, the χ^2 distribution is a function of one degrees-of-freedom parameter, ν : its density is

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

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

The mean 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.

FIGURE 9.2.4
χ^2 densities



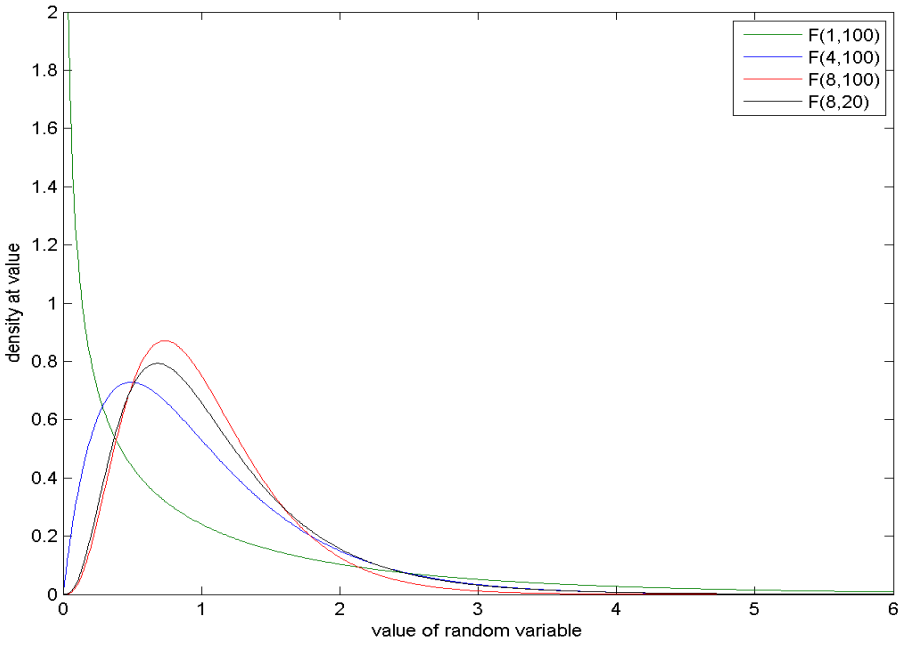
9.2.5 F densities

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 means of numerator and denominator are set to one. Of course, this does not imply that the mean of an F distribution is equal to one, because the ratio is a nonlinear transformation, and the mean of the ratio is not in general equal to the ratio of the means. However, the mean 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:

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

FIGURE 9.2.5
$F-$ densities



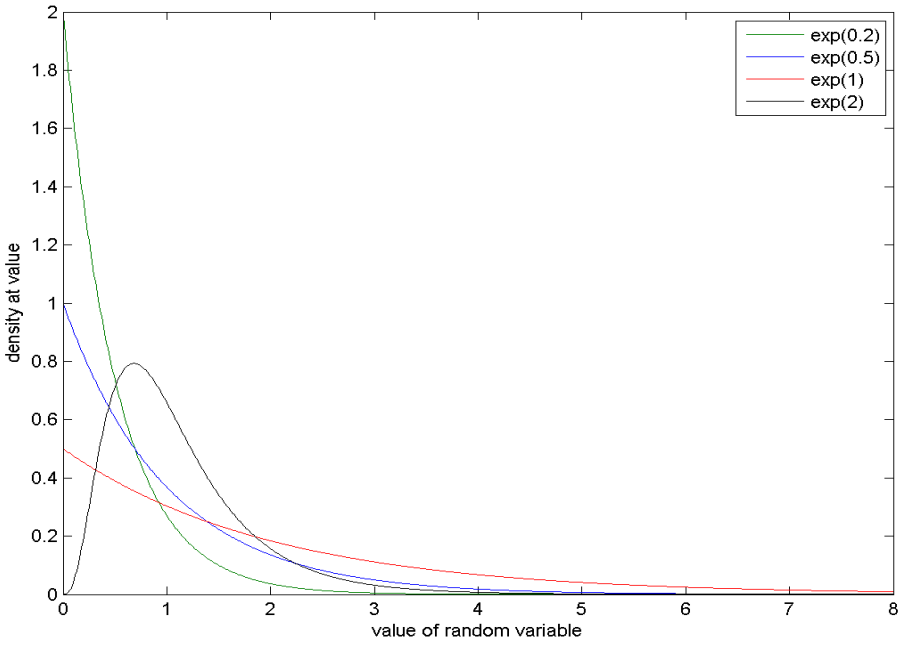
9.2.6 Exponential densities

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

$$f_X(x) = \mu^{-1} e^{\frac{-x}{\mu}},$$

a function of the single parameter μ which is the mean of the distribution.

FIGURE 9.2.6
Exponential densities



9.2.7 Lognormal densities

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:

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

FIGURE 9.2.7
lognormal densities