Contents

1 Fundamental Concepts

CHAPTER CONTENTS

Introduction

This chapter outlines some of the fundamental concepts that lie behind much of the rest of this book, including the ideas of a population distribution and a sampling distribution, the importance of random sampling, the law of large numbers and the central limit theorem. It then goes on to show how these ideas underpin the conventional approach to testing hypotheses and constructing confidence intervals.

Econometrics has a number of roles in terms of forecasting and analysing real data and problems. At the core of these roles, however, is the desire to pin down the magnitudes of effects and test their significance. Economic theory often points to the direction of a causal relationship (if income rises we may expect consumption to rise), but theory rarely suggests an exact magnitude. Yet, in a policy or business context, having a clear idea of the magnitude of an effect may be extremely important, and this is the realm of econometrics.

The aim of this chapter is to clarify some basic definitions and ideas in order to give the student an intuitive understanding of these underlying concepts. The account given here will therefore deliberately be less formal than much of the material later in the book.

A simple example

Consider a very simple example to illustrate the idea we are putting forward here. Table 1.1 shows the average age at death for both men and women in the 15 European countries that made up the European Union (EU) before its enlargement.

Simply looking at these figures makes it fairly obvious that women can expect to live longer than men in each of these countries, and if we take the average across all

Table 1.1 Average age at death for the EU15 countries (2002)

	Women	*Men*
Austria	81.2	75.4
Belgium	81.4	75.1
Denmark	79.2	74.5
Finland	81.5	74.6
France	83.0	75.5
Germany	80.8	74.8
Greece	80.7	75.4
Ireland	78.5	73.0
Italy	82.9	76.7
Luxembourg	81.3	74.9
Netherlands	80.6	75.5
Portugal	79.4	72.4
Spain	82.9	75.6
Sweden	82.1	77.5
UK	79.7	75.0
beginlwmath26pt] Mean	81.0	75.1
Standard deviation	1.3886616	1.2391241

countries we can clearly see that again, on a Europe-wide basis, women tend to live longer than men. However, there is quite considerable variation between the countries, and it might be reasonable to ask whether in general, in the world population, we would expect women to live longer than men.

A natural way to approach this would be to look at the difference in the mean life expectancy between men and women for the whole of Europe and to ask whether this is significantly different from zero. This involves a number of fundamental steps: first the difference in average life expectancy has to be estimated, then a measure of its uncertainty must be constructed, and finally the hypothesis that the difference is zero needs to be tested.

Table 1.1 gives the average (or mean) life expectancy for men and women for the EU as a whole, simply defined as:

$$\bar{Y}_w = \frac{1}{15}\sum_{i=1}^{15} Y_{wi} \qquad \bar{Y}_m = \frac{1}{15}\sum_{i=1}^{15} Y_{mi} \tag{1.1}$$

where $\bar{Y}_w$ is the EU average life expectancy for women and $\bar{Y}_m$ is the EU average life expectancy for men. A natural estimate of the difference between the two means is $(\bar{Y}_w - \bar{Y}_m)$. Table 1.1 also gives the average dispersion for each of these means, defined as the standard deviation, which is given by:

$$S.D._j = \sqrt{\sum_{i=1}^{15} (Y_{ji} - \bar{Y}_j)^2} \quad j = w, m \tag{1.2}$$

As we have an estimate of the difference and an estimate of the uncertainty of our measures, we can now construct a formal hypothesis test. The test for the difference between two means is:

$$t = \frac{\bar{Y}_w - \bar{Y}_m}{\sqrt{\dfrac{s_w^2}{15} + \dfrac{s_m^2}{15}}} = \frac{81 - 75.1}{\sqrt{\dfrac{1.389^2}{15} + \dfrac{1.24^2}{15}}} = 12.27 \tag{1.3}$$

The *t*-statistic of 12.27 is greater than 1.96, which means that there is less than a 5% chance of finding a *t*-statistic of 12.27 purely by chance when the true difference is zero. Hence we can conclude that there is a significant difference between the life expectancies of men and women.

Although this appears very intuitive and simple, there are some underlying subtleties, and these are the subject of this chapter. The questions to be explored are: what theoretical framework justifies all this? Why is the difference in means a good estimate of extra length of life for women? Is this a good estimate for the world as a whole? What is the measure of uncertainty captured by the standard deviation, and what does it really mean? In essence, what is the underlying theoretical framework that justifies what happened?

A statistical framework

The statistical framework that underlies the approach above rests on a number of key concepts, the first of which is the population. We assume that there is a population of events or entities that we are interested in. This population is assumed to be infinitely large and comprises all the outcomes that concern us. The data in Table 1.1 are for the EU15 countries for the year 2002. If we were interested only in this one year for this one set of countries, then there would be no statistical question to be asked. According to the data, women lived longer than men in that year in that area. That is simply a fact. But the population is much larger; it comprises all men and women in all periods, and to make an inference about this population we need some statistical framework. It might, for example, just be chance that women lived longer than men in that one year. How can we determine this?

The next important concepts are random variables and the population distribution. A random variable is simply a measurement of any event that occurs in an uncertain way. So, for example, the age at which a person dies is uncertain, and therefore the age of an individual at death is a random variable. Once a person dies, the age at death ceases to be a random variable and simply becomes an observation or a number. The population distribution defines the probability of a certain event happening; for example, the population distribution would define the probability of a man dying before he is 60 ($\Pr(Y_m < 60)$). The population distribution has various moments that define its shape. The first two moments are the mean (sometimes called the expected value, $E(Y_m) = \mu_{Y_m}$, or the average) and the variance ($E(Y_m - \mu_{Y_m})^2$, which is the square of the standard deviation and is often defined as $\sigma^2_{Y_m}$).

The moments described above are sometimes referred to as the unconditional moments; that is to say, they apply to the whole population distribution. But we can also condition the distribution and the moments on a particular piece of information. To make this clear, consider the life expectancy of a man living in the UK. Table 1.1 tells us that this is 75 years. What, then, is the life expectancy of a man living in the UK who is already 80? Clearly not 75! An unconditional moment is the moment for the complete distribution under consideration; a conditional moment is the moment for those members of the population who fulfil some condition, in this case being 80. We can consider a conditional mean $E(Y_m|Y_{im} = 80)$, in this case the mean of men aged 80, or conditional higher moments such as the conditional variance, which will be the subject of a later chapter. This is another way of thinking of subgroups of the population: we could think of the population as consisting of all people, or we could think of the distribution of the population of men and women separately. What we would like to know about is the distribution of the population we are interested in, that is, the mean of the life expectancy of all men and all women. If we could measure this, again there would be no statistical issue to address; we would simply know whether, on average, women live longer than men. Unfortunately, typically we can only ever have direct measures on a sample drawn from the population, and we have to use this sample to draw some inference about the population.

If the sample obeys some basic properties we can proceed to construct a method of deriving inference. The first key idea is that of random sampling: the individuals who make up our sample should be drawn at random from the population. The life expectancy of a man is a random variable; that is to say, the age at death of any individual is uncertain. Once we have observed the age at death and the observation becomes part of our sample it ceases to be a random variable. The data set then comprises a set of individual observations, each of which has been drawn at random from the population. So our sample of ages at death for men becomes $Y_m = (Y_{1m}, Y_{2m}, \ldots, Y_{nm})$. The idea of random sampling has some strong implications: because any two individuals are drawn at random from the population they should be **independent** of each other; that is to say, knowing the age at death of one man tells us nothing about the age at death of the other man. Also, as both individuals have been drawn from the same population, they should have an **identical distribution**. So, based on the assumption of random sampling, we can assert that each of the observations in our sample should have an independent and identical distribution; this is often expressed as iid.

We are now in a position to begin to construct a statistical framework. We want to make some inference about a population distribution from which only a sample has been observed. How can we know whether the method we choose to analyse the sample is a good one or not? The answer to this question lies in another concept, called the sampling distribution. If we draw a sample from our population, let's suppose we have a method for analysing that sample. It could be anything; for example, take the odd-numbered observations and sum them and divide by 20. This will give us an estimate. If we had another sample this would give us another estimate, and if we kept drawing samples this would give us a whole sequence of estimates based on this technique. We could then look at the distribution of all these estimates, and this would be the sampling distribution of this particular technique. Suppose the estimation procedure produces an estimate of the population mean which we call $\tilde{Y}_m$, then the sampling distribution will have a mean and a variance $E(\tilde{Y}_m)$ and $E(\tilde{Y}_m - E(\tilde{Y}_m))^2$; in essence, the sampling distribution of a particular technique tells us most of what we need to know about the technique. A good estimator will generally have the property of **unbiasedness**, which implies that its mean value is equal to the population feature we want to estimate. That is, $E(\tilde{Y}_m) = \eta$, where η is the feature of the population we wish to measure. In the case of unbiasedness, even in a small sample we expect the estimator to get the right answer on average. A slightly weaker requirement is **consistency**; here we only expect the estimator to get the answer correct if we have an infinitely large sample, $\lim_{n\to\infty} E(\tilde{Y}_m) = \eta$. A good estimator will be either unbiased or consistent, but there may be more than one possible procedure which has this property. In this case we can choose between a number of estimators on the basis of **efficiency**; this is simply given by the variance of the sampling distribution. Suppose we have another estimation technique, which gives rise to $\overleftrightarrow{Y}$, which is also unbiased; then we would prefer $\tilde{Y}$ to this procedure if $\text{var}(\tilde{Y}) < \text{var}(\overleftrightarrow{Y})$. This simply means that, on average, both techniques get the answer right, but the errors made by the first technique are, on average, smaller.

Properties of the sampling distribution of the mean

In the example above, based on Table 1.1, we calculated the mean life expectancy of men and women. Why is this a good idea? The answer lies in the sampling distribution of the mean as an estimate of the population mean. The mean of the sampling distribution of the mean is given by:

$$E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \mu_Y = \mu_Y \tag{1.4}$$

So the expected value of the mean of a sample is equal to the population mean, and hence the mean of a sample is an unbiased estimate of the mean of the population distribution. The mean thus fulfils our first criterion for being a good estimator. But what about the variance of the mean?

$$\mathrm{var}(\bar{Y}) = E(\bar{Y} - \mu_Y)^2 = E\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(Y_i - \mu_Y)(Y_j - \mu_Y)\right)$$

$$beginlwmath28pt] = \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathrm{var}(Y_i) + \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n}\mathrm{cov}(Y_i, Y_j)\right) = \frac{\sigma_Y^2}{n} \tag{1.5}$$

So the variance of the mean around the true population mean is related to the sample size that is used to construct the mean and the variance of the population distribution. As the sample size increases, the variance in the population shrinks, which is quite intuitive, as a large sample gives rise to a better estimate of the population mean. If the true population distribution has a smaller mean the sampling distribution will also have a smaller mean. Again, this is very intuitive; if everyone died at exactly the same age the population variance would be zero, and any sample we drew from the population would have a mean exactly the same as the true population mean.

Hypothesis testing and the central limit theorem

It would seem that the mean fulfils our two criteria for being a good estimate of the population as a whole: it is unbiased and its efficiency increases with the sample size. However, before we can begin to test a hypothesis about this mean, we need some idea of the shape of the whole sampling distribution. Unfortunately, while we have derived a simple expression for the mean and the variance, it is not in general possible to derive the shape of the complete sampling distribution. A hypothesis test proceeds by making an assumption about the truth; we call this the null hypothesis, often referred to as H. We then set up a specific alternative hypothesis, typically called H_a. The test consists of calculating the probability that the observed value of the statistic could have arisen purely by chance, assuming that the null hypothesis is true. Suppose that our null hypothesis is that the true population mean for age at death for men is 70, H: $E(\bar{Y}_m) = 70$. Having observed a mean of 75.1, we might then test the alternative that the mean is greater than 70. We would do this by calculating the probability that 75.1 could arise purely by chance when the true value of the population mean is 70. With a continuous distribution the probability of any exact point coming up is zero, so strictly what we are calculating is the probability of drawing any value for the mean that is greater than 75.1. We can then compare this probability with a predetermined value, which we call the **significance level** of the test. If the probability is less than the significance level, we reject the null hypothesis in favour of the alternative. In traditional statistics the significance level is usually set at 1%, 5% or 10%. If we were using a 5% significance level and we found that the probability of observing a mean greater than 75.1 was 0.01, as $0.01 < 0.05$ we would reject the hypothesis that the true value of the population mean is 70 against the alternative that it is greater than 70.

The alternative hypothesis can typically be specified in two ways, which give rise to either a one-sided test or a two-sided test. The example above is a one-sided test, as the alternative was that the age at death was greater than 70, but we could equally have tested the possibility that the true mean was either greater or less than 70, in which case we would have been conducting a two-sided test. In the case of a two-sided test we would be calculating the probability that a value either greater than 75.1 or less than $70 - (75.1 - 70) = 64.9$ could occur by chance. Clearly this probability would be higher than in the one-sided test.

Figure 1.1 shows the basic idea of hypothesis testing. It illustrates a possible sampling distribution for the mean life expectancy of men under the null hypothesis that the population mean is 70. It is an unlikely shape, being effectively a triangle, but we will discuss this later; for the moment, simply assume that this is the shape of the distribution. By definition, the complete area under the triangle sums to 1. This simply means that with probability 1 (certainty) the mean will lie between 62 and 78 and that it is centred on 70. We actually observe a mean of 75.1, and if we wish to test the hypothesis that the true mean is 70 against the alternative that it is greater than 70 (a one-sided test) we calculate the probability of observing a value of 75.1 or greater. This is given by area C in the figure. If we wished to conduct the two-sided test, that the alternative is either greater than 75.1 or less than 64.9, we would calculate the sum of areas A and C, which is clearly greater than C. If we adopted a 5% critical value and if $C < 0.05$, we would reject the null on a one-sided test. If $C + A < 0.05$, we would reject the null at a 5% level on the two-sided test.

As noted above, while we have calculated the mean and the variance of the sampling distribution in the case of the mean, it is not generally possible to calculate the

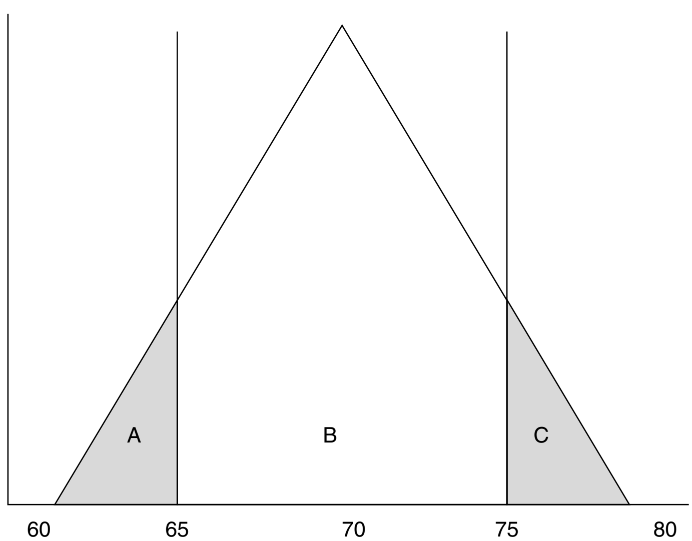

Figure 1.1 A possible distribution for life expectancy

shape of the complete distribution. However, there is a remarkable theorem which does generally allow us to do this as the sample size grows large. This is the central limit theorem.

Central limit theorem

If a set of data is iid with n observations, $(Y_1, Y_2, \ldots Y_n)$, and with a finite variance then as n goes to infinity the distribution of $\bar{Y}$ becomes normal. So as long as n is reasonably large we can think of the distribution of the mean as being approximately normal.

This is a remarkable result; what it says is that, regardless of the form of the population distribution, the sampling distribution will be normal as long as it is based on a large enough sample. To take an extreme example, suppose we think of a lottery which pays out one winning ticket for every 100 tickets sold. If the prize for a winning ticket is \$100 and the cost of each ticket is \$1, then, on average, we would expect to earn \$1 per ticket bought. But the population distribution would look very strange; 99 out of every 100 tickets would have a return of zero and one ticket would have a return of \$100. If we tried to graph the distribution of returns it would have a huge spike at zero and a small spike at \$100 and no observations anywhere else. But, as long as we draw a reasonably large sample, when we calculate the mean return over the sample it will be centred on \$1 with a normal distribution around 1.

The importance of the central limit theorem is that it allows us to know what the sampling distribution of the mean should look like as long as the mean is based on a reasonably large sample. So we can now replace the arbitrary triangular distribution in Figure 1.1 with a much more reasonable one, the normal distribution.

A final small piece of our statistical framework is the **law of large numbers**. This simply states that if a sample $(Y_1, Y_2, \ldots Y_n)$ is IID with a finite variance then $\bar{Y}$ is a consistent estimator of μ, the true population mean. This can be formally stated as $\Pr(|\bar{Y}-\mu| < \varepsilon) \to 1$ as $n \to \infty$, meaning that the probability that the absolute difference between the mean estimate and the true population mean will be less than a small positive number tends to one as the sample size tends to infinity. This can be proved straightforwardly, since, as we have seen, the variance of the sampling distribution of the mean is inversely proportional to n; hence as n goes to infinity the variance of the sampling distribution goes to zero and the mean is forced to the true population mean.

We can now summarize: $\bar{Y}$ is an unbiased and consistent estimate of the true population mean μ; it is approximately distributed as a normal distribution with a variance which is inversely proportional to n; this may be expressed as $N(\mu, \sigma^2/n)$. So if we subtract the population mean from $\bar{Y}$ and divide by its standard deviation we will create a variable which has a mean of zero and a unit variance. This is called standardizing the variable.

$$\frac{\bar{Y}-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1) \tag{1.6}$$

One small problem with this formula, however, is that it involves σ^2. This is the population variance, which is unknown, and we need to derive an estimate of it. We may estimate the population variance by:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{1.7}$$

Here we divide by $n-1$ because we effectively lose one observation when we estimate the mean. Consider what happens when we have a sample of one. The estimate of the mean would be identical to the one observation, and if we divided by $n = 1$ we would estimate a variance of zero. By dividing by $n - 1$ the variance is undefined for a sample of one. Why is S^2 a good estimate of the population variance? The answer is that it is essentially simply another average; hence the law of large numbers applies and it will be a consistent estimate of the true population variance.

Now we are finally in a position to construct a formal hypothesis test. The basic test is known as the student 't' test and is given by:

$$t = \frac{\bar{Y} - \mu}{\sqrt{S^2/n}} \tag{1.8}$$

When the sample is small this will follow a student t-distribution, which can be looked up in any standard set of statistical tables. In practice, however, once the sample is larger than 30 or 40, the t-distribution is almost identical to the standard normal distribution, and in econometrics it is common practice simply to use the normal distribution. The value of the normal distribution that implies 0.025 in each tail of the distribution is 1.96. This is the critical value that goes with a two-tailed test at a 5% significance level. So if we want to test the hypothesis that our estimate of the life expectancy of men of 75.1 actually is a random draw from a population with a mean of 70, then the test would be:

$$t = \frac{75.1 - 70}{\sqrt{S^2}/3.87} = \frac{5.1}{.355} = 14.37$$

This is greater than the 5% significance level of 1.96, and so we would reject the null hypothesis that the true population mean is 70. Equivalently, we could evaluate the proportion of the distribution that is associated with an absolute t-value greater than 4.1, which would then be the probability value discussed above. Formally, the probability, or p-value, is given by:

$$p\text{-value} = \Pr_{H_0}(|Y - \mu| > |\bar{Y}^{act} - \mu|) = \Pr_{H_0}(|t| > |t^{act}|)$$

So if the t-value is exactly 1.96 the p-value will be 0.05, and when the t-value is greater than 1.96 the p-value will be less than 0.05. The two values contain exactly the same information, simply expressed in a different way. The p-value is useful in other circumstances, however, as it can be calculated for a range of different distributions and can avoid the need to consult statistical tables, as its interpretation is always straightforward.

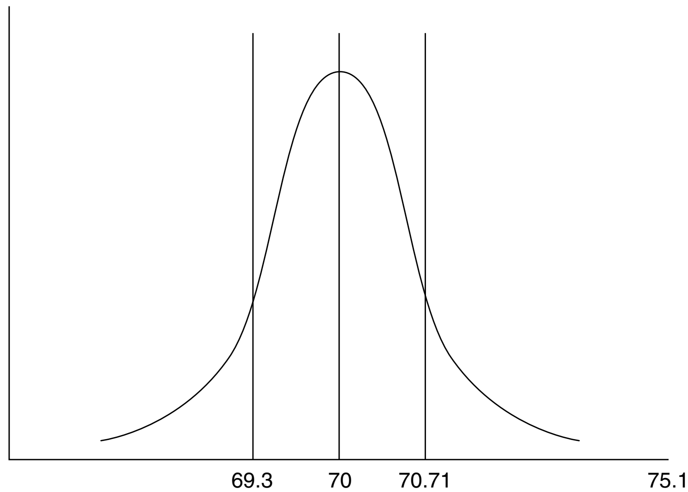

Figure 1.2 A normal distribution for life expectancy around the null

Figure 1.2 illustrates this procedure. It shows an approximately normal distribution centred on the null hypothesis with the two tails of the distribution defined by 69.3 and 70.71. Ninety-five per cent of the area under the distribution lies between these two points. The estimated value of 75.1 lies well outside this central region, and so we can reject the null hypothesis that the true value is 70 and that we observed 75.1 purely by chance. The *p*-value is twice the area under the curve which lies beyond 75.1, and clearly this is very small indeed.

One final way to think about the confidence we have in our estimate is to construct a confidence interval around the estimated parameter. We have an estimated mean value of 75.1, but we know there is some uncertainty as to what the true value is. The law of large numbers tells us that this is a consistent estimate of the true value, so with just this one observation our best guess is that the true value is 75.1. The central limit theorem tells us that the distribution around this value is approximately normal, and we know the variance of this distribution. So we can construct an interval around 75.1 that will contain any required amount of the distribution. The convention again is to use a 95% confidence interval, and this may be constructed as follows:

$$CI_{95\%} = \left\{ \bar{Y} + 1.96\frac{S}{\sqrt{n}},\ \bar{Y} - 1.96\frac{S}{\sqrt{n}} \right\} = \bar{Y} + 0.71,\ \bar{Y} - 0.71$$

So with 95% confidence we can say that the true mean lies between 74.39 and 75.81. This is shown in Figure 1.3; all that has happened here is that the picture has been moved so that it now centres on the estimated value of 75.1 and 95% of the figure lies inside the confidence interval. Clearly the null value of 70 lies way outside this region, and so we can again conclude that the true value of the mean is highly unlikely to be 70.

The same conclusion arises from calculating the formal *t*-test or the *p*-value or considering the confidence interval, because they are all simply different ways of expressing the same underlying distribution.

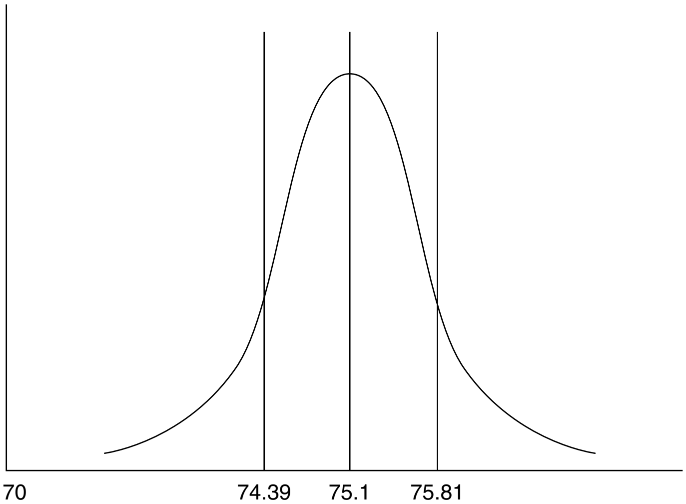

Figure 1.3 A 95% confidence interval around the estimated mean

Conclusion

In this chapter we have outlined the basic steps in constructing a theory of estimation and hypothesis testing. We began from the simple idea of random sampling, which gave rise to the proposition that the elements of a sample will have an iid distribution. From this we were able to define a population distribution and to make some inference about this distribution by constructing the mean and then defining the sampling distribution of the mean. By using the law of large numbers and the central limit theorem we were able to define the shape of the sampling distribution, and finally, given this, we were able to outline the basic testing procedure used in classical econometrics.

While at first sight this may appear to relate specifically to a simple estimation procedure, the mean, the same steps may be applied to almost any estimation procedure, as we will see in later chapters of this book. So when we estimate a parameter in a model from a data set we are essentially following the same steps. Any estimation procedure is essentially just taking a sample of data and averaging it together in some way. We have a sampling distribution for the parameter and we can investigate the unbiasedness and consistency of the estimation procedure. We can go on to apply the central limit theorem, which will establish that this sampling distribution will tend to a normal distribution as the sample size grows. Finally, we can use this result to construct hypothesis tests about the parameters that have been estimated and to calculate p-values and confidence intervals.

Index