Part III: Statistical methods

Chapter 10

Introduction to sampling and sampling distributions

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

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

Once we are clear about what the population being studied is, we want to know

how the quantities to be obtained from the sample are related to the true characteristics of the population that are of interest to us. They will not of course be identical in general, but we hope that they will be close and will tend to get closer as we take larger and larger samples. The purpose of this and the following chapters is to characterize what is known about the relation between sample quantities and population quantities: since we learn about the population quantity from the sample, we need to understand how close it is likely to be in different circumstances, and how great the probabilities are of making errors of different magnitudes.

10.1 Sample and population

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

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

Normally we would prefer a 'random' sample of the entire population.

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

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

In some cases, it is difficult to achieve the goal of a random sample, because some members of the population are more difficult to observe, or less likely to be observed by simple methods, than others. Researchers may therefore sometimes use a 'stratified sample'. A stratified sample is one in which the population is divided into mutually exclusive and exhaustive classes, and the final sample is designed to have the same proportion of each class as does the population. Simple random sampling may be used within each class, with the goal of obtaining an overall sample which is representative of the population.

For example, we may have 8% of a particular population which is elderly (let's say, 70 or over). If we try to sample randomly from the population, however, we may find that we are getting in touch with elderly people less often, either because

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

When we have a sample, we will want to use it to learn about some characteristic of the population. Often, we will start by estimating the mean of some characteristic in the population, for example, the mean weight of fish in a lake. But we know that the mean weight in our sample will not, except by outrageous coincidence, be the same to the nearest gram as in the population. So what does the mean on the sample tell us about the population mean? We can hope that they are close, but that is not very useful. In order to answer the question well, we would like to be able to characterize the entire distribution of the sample mean, given some population mean and size of sample. If the mean weight of a trout in the lake is 746 grams, and if we catch, measure and release a sample of 100 trout, what is the distribution of possible sample means that we could find? We can answer this question, at least approximately (and with an approximation error that declines to zero as sample size increases) in a very wide variety of cases. Once we do, we can make valid statements about where the population mean (not just the sample mean) lies. This in turn might allow us to compare different lakes, and draw legitimate conclusions about, for example, whether the fish are on average bigger in one than in the other.

10.2 Sampling and distributions of samples

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

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

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

distribution is zero.

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

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

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

In each case we take samples of size N from the distribution, and take the mean of each sample. We do this 100,000 times for each sample size, so that we have many examples of sample means, and then we can actually estimate the density that applies to the sample mean. We do that using a kernel density estimator (which is at present not described in this book, but can be thought of for now as a development or refinement of the idea of the histogram, producing a smooth curve instead of a set of bars).[2] There are three sample sizes, so that we can observe something about the way in which the sampling distribution changes as the sample size changes.

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

---

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

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

[2] See Silverman (1986) for an exposition.

FIGURE 10.2.1
Empirical distributions of sample mean:
$U[0,2]$ random variables, N=10, 100, 500

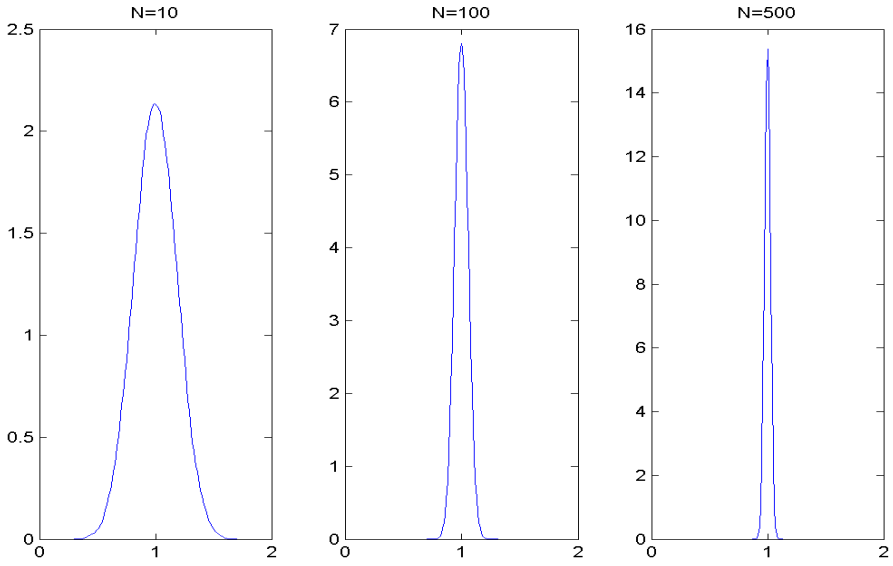
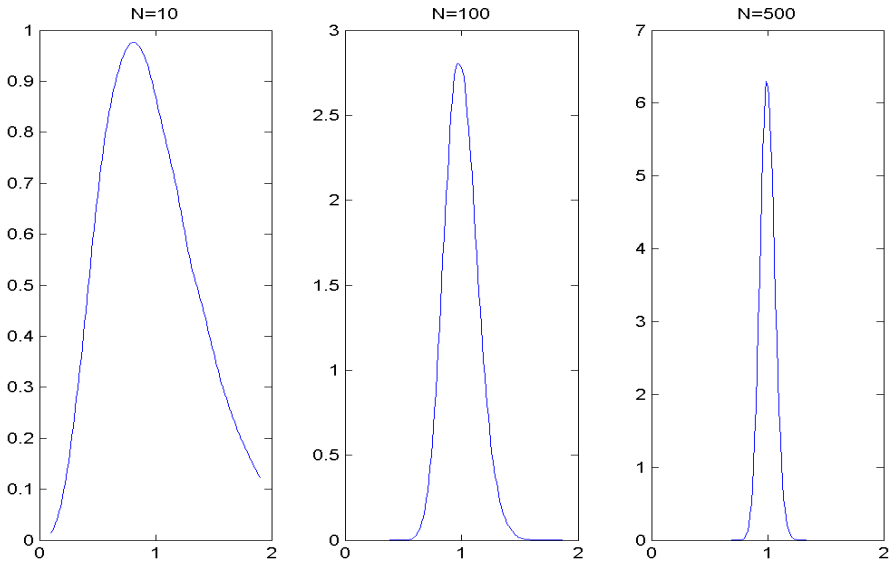


FIGURE 10.2.2
Empirical distributions of sample mean:
χ^2_1 random variables, N=10, 100, 500



The top panel shows results from Uniform random variables. The Uniform distribution is symmetric, and the sampling distributions are apparently symmetric as well. In the case of the input data which are χ^2_1, shown in the bottom panel, the sampling distribution for $N = 10$ is noticeably skewed; in fact if one looks very closely (compare the heights of the density function around 0.5 and 1.5), a tiny degree of skewing is visible at $N = 100$ as well. In the largest sample size, no skewing is visible to the naked eye.

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

10.3 A simple, if unrealistic, case

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

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

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

Theorem 10.1: (Linear combinations of Normal random variables are Normal.) Let $z_1, z_2, \ldots z_N$ be independent Normal variables each of which has mean 0 and variance σ_i^2. Then the linear combination $a_1 z_1 + a_2 z_2 + \ldots + a_n z_N$ has the distribution $N(0, \sum_{i=1}^{N} a_i^2 \sigma_i^2)$.

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

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

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

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

$$E\left(\sum_{i=1}^{N} \frac{1}{N} x_i\right) = \frac{1}{N} \sum_{i=1}^{N} E(x_i) = \frac{1}{N}(N\mu) = \mu.$$

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

$$var\left(\sum_{i=1}^{N} \frac{1}{N} x_i\right) = var\left(\frac{1}{N} x_1 + \frac{1}{N} x_2 + \ldots + \frac{1}{N} x_N\right)$$

$$= \left(\sum_{i=1}^{N} \frac{1}{N^2} var(x_i)\right) = \frac{1}{N^2}\left(\sum_{i=1}^{N} \sigma^2\right) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}.$$

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

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

- We can standardize the sample mean to obtain a distribution which does not change with the sample size: subtracting the mean and dividing by the square root of the variance, we find $\frac{\overline{X}_N - \mu}{\sigma/\sqrt{N}} \sim N(0,1)$, the standard Normal distribution.
- Multiplying numerator and denominator by the square root of sample size N in the distribution just given, the result may be rewritten as $\sqrt{N}\left(\frac{\overline{X}_N - \mu}{\sigma}\right) \sim N(0,1)$. This implies that scaling up the discrepancy in the estimate of the mean by the square root of the sample size leads to a fixed, non-degenerate distribution. It follows therefore that the discrepancy itself is declining at the rate of the square root of sample size. This is an example of 'square root-N' convergence, which appears in many standard parametric problems.

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

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

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

Given the sampling conditions assumed, s will converge probabilistically to σ in a sense that we will define precisely in the next chapter. So in large samples, using 'SE' for standard error, the sample quantity, and 'SD' for standard deviation, the population quantity:

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

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

But in fact, for this case, the exact distribution that applies for any given sample size N (not just the asymptotic result) has been worked out, and we do not need to use an approximation.[3]

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

[3]In 1908, by William S. Gosset, 1876-1937. Because Gossett used the pseudonym Student, the distribution is often called Student's $t-$ distribution.

independent, the ratio

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

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

Proof: See e.g. Mood et al. (1974), section 4.5.

We will see below that in this case of independent random sampling from Normal data, the sample variance s^2 in fact has a χ^2_{N-1} distribution, so that the feasible statistic $\frac{(\overline{X}_N-\mu)}{s/\sqrt{N}}$ will be distributed as t_{N-1}. Note again that this result, which gives an exact distribution applicable to any particular sample size, has been obtained under the generally-unrealistic assumption that the data that we are sampling themselves have a Normal distribution. In more general circumstances were we do not know this, we will have to rely on an asymptotic approximation to get the distribution of this feasible statistic, as described in the next chapter.[4]

10.4 Using a sampling distribution

Consider then a situation in which we have a sample of size N from a population with known mean μ and variance σ^2. What can we deduce from this?

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

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

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

Since the $t-$ distribution with $N-1$ degrees of freedom has a known form, and the quantiles and so on have been tabulated, we can compute an interval such

[4]The quantity $\frac{(\overline{X}_N-\mu)}{s/\sqrt{N}}$ is distributed as t_{N-1} under these conditions, and more generally is asymptotically $N(0,1)$. These results may appear to conflict, but do not, because the t_{N-1} converges in distribution to the $N(0,1)$ as sample size increases without bound.

that the expression on the left-hand side above has a given probability of lying in that interval. Using the notation q_α for the $\alpha-$ quantile of the relevant distribution (t_{N-1}), we can define the interval such that

$$P\left(-q_{\alpha/2} < \frac{\overline{X}_N - \mu}{s/\sqrt{N}} < q_{\alpha/2}\right) = 1 - \alpha, \tag{10.a}$$

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

When working with the Normal distribution, either to describe the infeasible case or in using the approximation from asymptotic theory that we will learn in the next chapter, it is common to use the notation $z_{\alpha/2}$ to describe the corresponding quantiles from the Normal. This notation is also sometimes used for the $t-$ distribution.

Let us now manipulate this expression further. Our goal is to get a statement about where the true mean is likely to lie, when we observe only the sample quantities. The expression (10.a) above contains $\overline{X}_N - \mu$ in the middle: if we know one of these, we will be able to obtain a statement about the other.

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

$$P\left(-q_{\alpha/2}(s/\sqrt{N}) < \overline{X}_N - \mu < q_{\alpha/2}(s/\sqrt{N})\right) = 1 - \alpha.$$

If we now subtract $\overline{X}_N$ from each of the three terms, we obtain a statement purely about μ:

$$P\left(-\overline{X}_N - q_{\alpha/2}(s/\sqrt{N}) < -\mu < -\overline{X}_N + q_{\alpha/2}(s/\sqrt{N})\right) = 1 - \alpha,$$

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

$$P\left(\overline{X}_N + q_{\alpha/2}(s/\sqrt{N}) > \mu > \overline{X}_N - q_{\alpha/2}(s/\sqrt{N})\right) = 1 - \alpha.$$

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

precise statement. The bounds are divided by $\sqrt{N}$, so the intervals get narrower at a rate proportional to the square root of sample size increase.

For the $t-$ distribution with a large number of degrees of freedom $N-1$ (or for the standard Normal distribution), $q_{\alpha/2}$ (or $z_{\alpha/2}$ in the notation commonly used for the standard Normal) is approximately 1.96: that is, about 2.5% of the distribution lies below -1.96, and about 2.5 % of the distribution lies above 1.96.[5] So the probability interval just stated is what lies behind the commonly-remembered result that there is a 95% probability that the population mean of something will lie within about $\pm$ two standard errors of the sample mean.

To take a numerical example, consider the population of fish in the lake mentioned earlier. We catch 100 fish, and find an average weight of 746 g, with a standard error of 205 g. As usual in real data, a moment's reflection tells us that these data could not literally be Normal: weight cannot be negative, so the distribution is bounded below, unlike the Normal. As we said, knowing that the data are Normal is generally unrealistic. Let's go on with this example anyway, because it will turn out below that the results that we have just stated will turn out to be a good approximation in a wide range of cases, even though the data are not Normal.

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

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

or since $1.96 \times 20.5 = 40.18$,

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

So, given the conditions assumed to hold in this sampling experiment, we can be 95% sure that the mean weight of the fish in the lake is between about 706 g and 787 g (rounding to three significant digits). This might be of interest in itself, but it might also be useful in comparing two lakes. We might measure average weight of fish in a second lake as 921 g, for example. Are the fish really bigger on average in lake 2, or are we just seeing sampling variation? As we check more and more fish in each lake, we can narrow down the intervals for the average weight of each, and if these intervals become clearly separated, then we would conclude that the fish are on average heavier (healthier?) in one lake than the other. Alternatively, we might find that the intervals for the population means in the two lakes overlap even in large samples, so that we cannot be confident that there is any genuine difference. This is analogous to a test of the hypothesis that mean weight is the same in each lake. Below we will study ways to perform these tests precisely.

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

10.5 Simple case continued: distribution of the sample variance

It may seem strange to say that the sample variance has a variance. But of course an estimate of the population variance of the data, as with any estimate based on a sample of data, depends upon random elements in the data and has a distribution. That distribution, and moments of it that exist, can be estimated. In particular, we can estimate features of it such as the (true, or population) mean of the sample variance and the (true, or population) variance of the sample variance. To say this slightly differently, whenever we have an estimate of something we can try to characterize the moments or distribution of that estimate, for example to get confidence intervals. The estimate of the variance of a sample of data is just one example of this: once we've estimated the population variance using a sample, we might want to ask, what are the uncertainty bands for this estimate? To do that, we'll need the variance of the estimate, that is, the variance of the sample variance.

This problem has a simple analytic solution in the case where data come from a Normal distribution. We will work through that now to illustrate the way in which the calculation can be done in a simple case, and also to illustrate one of the connections between the Normal distribution and the χ^2. To begin, let's continue down the road of unrealistic cases for a while, because it will help to clarify several things later.

Consider then that the mean of a random variable X is known. Then on an independent and identically distributed sample of size N, $\{x_1, x_2, \ldots x_N\}$, we would estimate the population variance $\sigma^2 = E(X-\mu)^2$ by its sample analogue, $\tilde{\sigma}^2 = N^{-1}\Sigma(x_i - \mu)^2$. (Note that we don't have to use $\overline{X}$ because we're considering an unrealistic case in which we know the mean, μ.)

We can therefore divide both sides of this expression by σ^2, to obtain $\frac{\tilde{\sigma}^2}{\sigma^2} = N^{-1}\Sigma(\frac{x_i-\mu}{\sigma})^2$.

Now let us assume further that we are dealing with Normally distributed data, so that for any sample point x_i, $x_i \sim N(\mu, \sigma^2)$ or (standardizing) $\left(\frac{x_i-\mu}{\sigma}\right) \sim N(0,1)$. From the result stated earlier that sums of squared independent standard Normal random variables have χ^2 distributions, it follows that

$$\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2 \sim \chi^2_N.$$

Now $\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2 = \frac{N}{\sigma^2}\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{N} = \frac{N}{\sigma^2}\tilde{\sigma}^2$. Taking expectations of the first and last terms, and bearing in mind the fact that the mean of a χ^2_N random variable is N, we have $E\left[N\frac{\tilde{\sigma}^2}{\sigma^2}\right] = N$ and so $E\left[\frac{\tilde{\sigma}^2}{\sigma^2}\right] = 1$, or $E(\tilde{\sigma}^2) = \sigma^2$.

Written slightly differently, $E\left[\frac{\tilde{\sigma}^2}{\sigma^2}\right] = N^{-1}E\left[\sum(\frac{(x_i-\mu)^2}{\sigma^2})\right] = N^{-1}\frac{N\sigma^2}{\sigma^2}$ and cancelling σ^2 from both sides of the equality, $E(\tilde{\sigma}^2) = \sigma^2$.

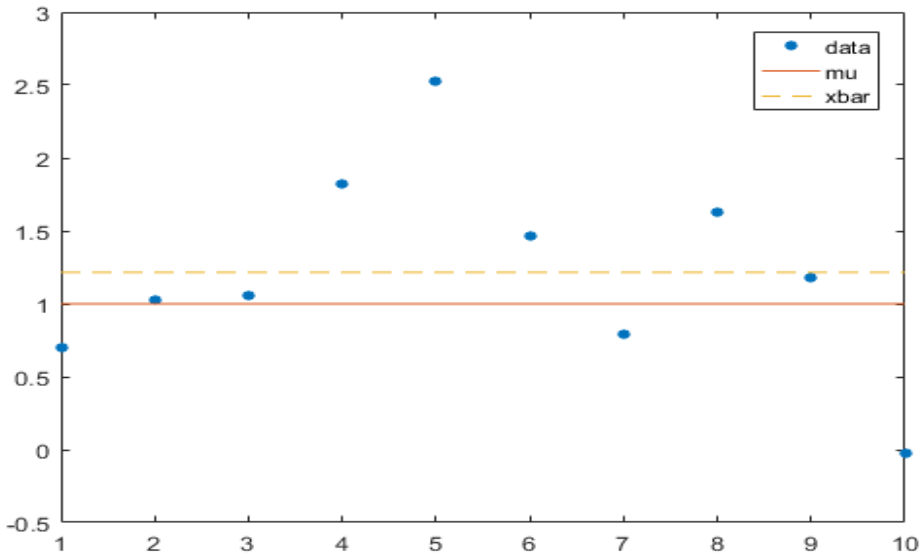
We therefore have an unbiased estimator of the population variance in this infeasible example (infeasible because the estimator uses the value of μ, which is not known in practice); we can obtain the unbiased estimator of σ^2 by taking the sample analogue of the expression for variance, which is the sample mean of $(x_i - \mu)^2$, i.e. $(1/N)(x_i - \mu)^2$.

When μ is not known, things are not quite as simple. The reason is that, while substituting $\overline{X}$ for μ seems straightforward, and although $\overline{X}$ is indeed an unbiased estimator of μ, nonetheless this substitution creates a bias in the variance estimator: $(1/N)(x_i - \overline{X})^2$ is not an unbiased estimator of σ^2 in the way that $(1/N)(x_i - \mu)^2$ would be if it were available. This point deserves emphasis: substituting an unbiased estimator of an unknown parameter into an expression which gives another unbiased estimator when that parameter is known, nonetheless leads to a biased estimator.

In this case the reason, intuitively, is that we are trying to measure the typical squared deviation of a point from the mean of the distribution. When we use $\overline{X}$ to measure that mean of the distribution, our sum of squared deviations comes out slightly too small, because $\overline{X}$ is defined to be right in the middle of (at the mean of) the sample, so that squared deviations of the data points from $\overline{X}$ tend to add up to a lower value than squared deviations from μ would: μ is in the middle of the population, not the sample.

Figure 10.5.1 illustrates this on a small sample of data from a random variable with a true (population) mean of one, but where the sample mean differs by a degree that is large enough to be clear visually. Taking the sum of squared deviations to the red line, the population mean, would by the argument above give on average the correct value for the variance. The deviations to the yellow dotted line will on average be smaller, and so the estimate of the variance resulting from taking squared deviations to that line will tend to be too small. It turns out as we will show in the appendix that dividing by $N - 1$ rather than N precisely corrects for this effect, so that the feasible estimator $s^2 = (N-1)^{-1}\Sigma((x_i - \overline{X})^2)$ is also unbiased.

FIGURE 10.5.1
Population and sample mean on a small random sample with true mean = 1



The formal result is given in Theorem 10.3.

Theorem 10.3: Let $\{x_i\}$, $i = 1, \ldots N$ be a set of N independent draws from a random variable X which has a Normal distribution with unknown mean and variance (μ, σ^2). Define $s^2 = (N-1)^{-1}\Sigma((x_i - \overline{X})^2)$. Then $E(s^2) = \sigma^2$ and

$$(N-1)\frac{s^2}{\sigma^2} \sim \chi^2_{N-1}.$$

Proof: See the appendix.

Once we have a known distribution, confidence intervals can be obtained mechanically, by manipulating a statement about the probability that some quantity lies in a given interval. We turn the statement, as we did earlier, into one that directly concerns the quantity that interests us, by steps that preserve the truth of the statement. In this case, however, we are dealing with an asymmetric distribution, and the confidence bounds will lie at unequal distances from the mean.

From the lines above we know that $(N-1)\frac{s^2}{\sigma^2} \sim \chi^2_{N-1}$ gives the distribution of the sample variance estimator s^2 under these restrictive circumstances; note that on the left-hand side of this expression there is only one random variable, s^2. Label the quantiles of the distribution that leave $(\alpha/2)\%$ of the probability in each tail of the χ^2_{N-1} distribution as $c_{\alpha/2}$ and $c_{1-\alpha/2}$. Then we have

$$P[c_{\alpha/2} < (N-1)\frac{s^2}{\sigma^2} < c_{1-\alpha/2}] = 1 - \alpha.$$

Therefore, dividing all terms by $(N-1)s^2$,

$$P\left[\frac{c_{\alpha/2}}{(N-1)s^2} < \frac{1}{\sigma^2} < \frac{c_{1-\alpha/2}}{(N-1)s^2}\right] = 1-\alpha.$$

We want a statement about σ^2, so we will need to invert these terms, bearing in mind that we will then need to switch the inequalities to keep the statement true; for example, $5 > 4 > 3 \Leftrightarrow (1/5) < (1/4) < (1/3)$. Therefore

$$P\left[\frac{(N-1)s^2}{c_{\alpha/2}} > \sigma^2 > \frac{(N-1)s^2}{c_{1-\alpha/2}}\right] = P\left[\frac{(N-1)s^2}{c_{1-\alpha/2}} < \sigma^2 < \frac{(N-1)s^2}{c_{\alpha/2}}\right] = 1-\alpha$$

gives the required $(1-\alpha)\%$ confidence interval.

Consider for example a random sample of 210 points, for which the variance is estimated to be s^2 1.234. The relevant distribution is the χ^2_{209}, which can be approximated from tables, or for which quantiles can be obtained computationally. In this case, $c_{.025} = 170.86$ and $c_{.975} = 250.93$ from a computer algorithm, and the confidence interval for the true variance is then 209(1.234)/250.93 to 209(1.234)/170.86 or [1.028, 1.509]. Notice that 1.234 is not the middle of this interval; the bounds are unequal distances from that point.

. . .

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

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

Chapter 11

Laws of Large Numbers and Central Limit Theorems

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

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

11.1 Some preliminary asymptotic theory

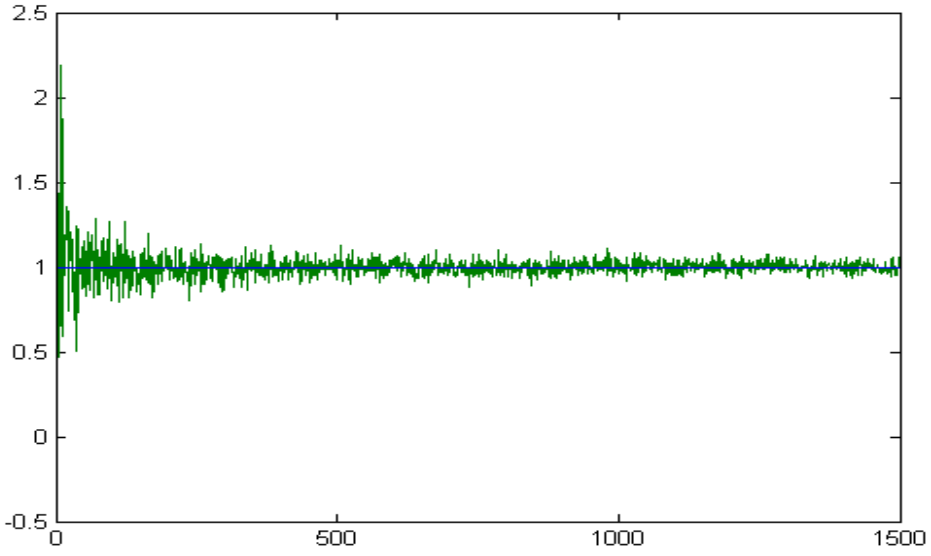
We begin therefore by defining convergence in probability and convergence in distribution.

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

$$\lim_{n\rightarrow\infty} P\left(|X_n - x| < \epsilon\right) = 1 \text{ for any } \epsilon > 0.$$

This is typically denoted either by $X_n \xrightarrow{p} x$ or by $plim(X_n) = x$; note that the value x may be a constant, or a random variable. The following graphic provides an example of a sequence of 1500 observations on a random variable which is converging in probability to a probability limit of one, but sometimes moves closer, and sometimes moves farther away, from one. We observe that the range of its fluctuations around the probability limit tends to diminish as sample size increases (in this example, in proportion to the square root of sample size).

FIGURE 11.1.1
Example of convergence in probability
$N = 1500$



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

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

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

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

This is usually written as $X_n \xrightarrow{D} X$.

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

11.2 LAWS OF LARGE NUMBERS

One important class of asymptotic result concerns convergence of a sequence of sample estimates to the true mean of the process.

Theorem 11.1: (Weak Law of Large Numbers– WLLN). Let $\{x_i\},\ i = 1, \ldots n$, be independent random draws from a distribution with cumulative distribution function $F_X(x)$, such that the distribution has a mean μ and variance $\sigma^2 > 0$. Then the sample mean converges in probability to the true mean, such that for any $\varepsilon > 0$,

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

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

$$P\left[|\overline{X} - \mu| < \varepsilon\right] \geq 1 - \delta.$$

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

11.3 Central Limit Theorems

As we have said, in most cases we do not know the distribution of the data that we are analyzing. We might therefore expect it to follow that the distributions of statistics that we compute from these data will also be unknown. However, in many cases particularly involving sums or averages, the distribution of a statistic can be approximated well because we have information about convergence in distribution that applies to be statistic: that is, its finite-sample distribution is unknown, but can be shown to be converging to a particular distribution $F(.)$. We can therefore take $F(.)$ as an approximation to the true distribution, which will become increasingly precise as sample information accumulates.

Central limit theorems are some of the most important and useful results and statistics, for at least two reasons. First, they apply to sums or means of random variables (data points), and taking the mean of something is one of the most frequently applied operations. Recall for example that the moments and functions of moments that we studied earlier, such as the variance, coefficient of skewness and kurtosis, are based on expectations. Their sample counterparts are therefore means of some random variable, such as the sample variance $(n-1)^{-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$, which apart from the degrees of freedom correction is the sample mean of the random variable $(x_i - \overline{X})^2$. A vast set of statistics can be interpreted as sample means of some random variable.

Second, almost invariably we do not know the true distribution function of our data. To assume that we know the distribution when we do not can easily lead to

false statements or results, particularly given that differences between distributions that may be critical in practice (for example, tail thickness or relative frequency of extreme events in risk management) can be very difficult to observe precisely in an empirical sample of data. We cannot calculate important quantities such as tail probabilities by the natural method of integrating under a particular mathematical form of density function, or directly using the cumulative distribution function, because these functions are unknown.

A Central Limit Theorem (CLT) gives a result that applies to any input distribution, as long as a few simple conditions are met. In the theorem that we will state here, these conditions are fairly weak (therefore apply fairly widely), but they can be weakened further. For this reason there are many CLT's, as similar results have been obtained under different assumed conditions on the true process.

Theorem 11.2: (Central Limit Theorem– CLT.) Let $\{x_i\},\ i = 1, \ldots n$, the independent random draws from a distribution with cumulative distribution function $F_X(x)$, such that the distribution has a mean μ and variance $\sigma^2 > 0$. Then the standardized sample mean has a limiting standard Normal distribution: that is,

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \quad \stackrel{D}{\rightarrow} \quad N(0,1).$$

The notation $N(0,1)$ denotes the Normal distribution with mean zero and variance one: substituting into the Normal density function

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

the values $\mu = 0,\ \sigma^2 = 1$, we therefore have that $N(0,1)$ represents the density

$$f_X(x) = (2\pi)^{-\frac{1}{2}} exp\{-\frac{1}{2}x^2\}.$$

(We might find it more natural to do without standardization in Theorem 11.2, and write $\overline{X}_n - \mu \quad \stackrel{D}{\rightarrow} \quad N(0, \frac{\sigma^2}{n})$ or $\overline{X}_n \quad \stackrel{D}{\rightarrow} \quad N(\mu, \frac{\sigma^2}{n})$. The problem with writing this is that we would now be talking about convergence to some distribution which isn't constant, since its variance is declining with sample size n, and in fact isn't even well defined in the limit as the variance would be zero, and we divide by the variance in the expression for the Normal density. So this representation might look more natural, but isn't strictly valid. Therefore the standardization is applied in order to have a limiting distribution which is *non-degenerate*.)

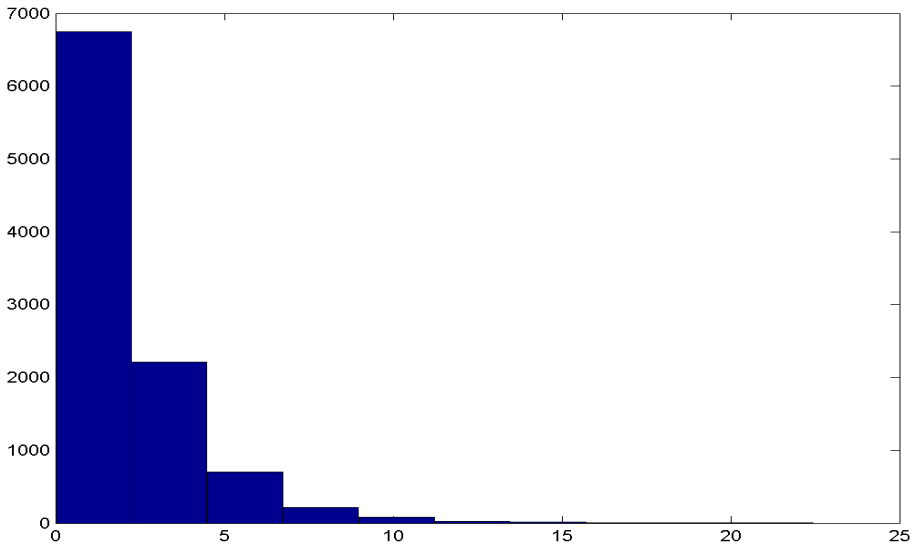
Theorem 11.2 could also be written in terms of the sum of the data points rather than the sample mean, because the sum $\sum_{i=1}^{n} x_i$ differs from the sample mean

$n^{-1}\sum_{i=1}^{n} x_i$ only by the constant factor n : being a constant, this factor does not affect the shape of the distribution. So multiplying through top and bottom by n in Theorem 11.2, we have

$$\frac{\sum_{i=1}^{n} x_i - n\mu}{\sigma\sqrt{n}} \quad \xrightarrow{D} \quad N(0,1).$$

The next figures illustrate convergence to the Normal. We take data points from a heavily right-skewed distribution, a typical sample of which is illustrated in Figure 11.3.1.[6] Numerous samples are taken from this skewed distribution, and in each one the mean is computed; the density of the sample mean is then estimated using kernel smoothing methods, and these densities are illustrated in Figures 11.3.2 and 11.3.3.[7] In each of the latter figures, we plot the density of the sample mean, standardized as in the theorem, for each of three different sample sizes. These six cases are separated into two figures, because if we placed them all on the same figure, the different widths and heights of the densities would make it difficult to observe the shapes of each one (for example,the largest sample size would appear as a thin spike if placed on the graph with the scale of Figure 11.3.2): note that the horizontal and vertical scales of the two figures differ. They otherwise have the same meaning.

FIGURE 11.3.1
Distribution of the input data:
histogram of realizations of a single random sample



[6]These pseudo-random data are in fact generated from the χ^2 distribution.
[7]We could have used histograms instead, but would give up the smooth curve estimated by the kernel method.

FIGURE 11.3.2
Empirical distributions of the sample mean of N realizations:
skewed random variables, sample sizes N=25, 100, 400

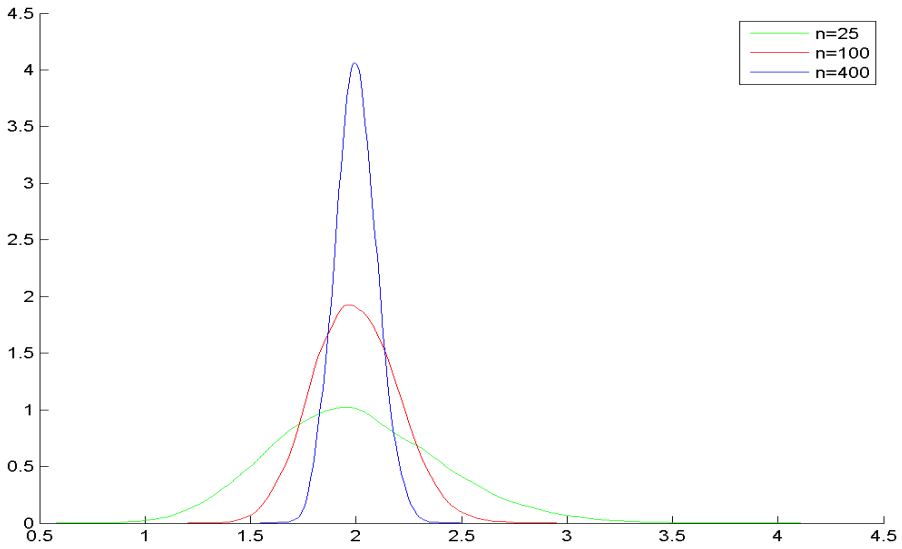
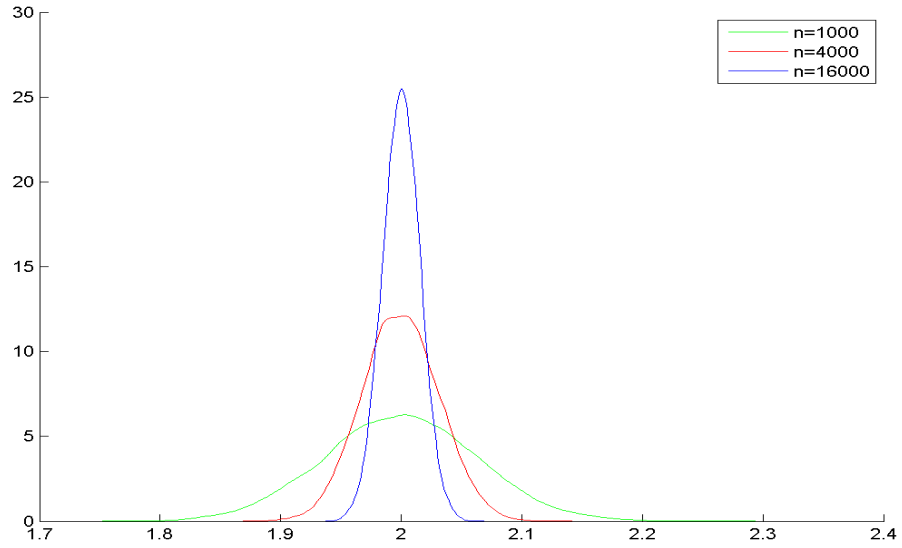


FIGURE 11.3.3
Empirical distributions of the sample mean of N realizations:
skewed random variables, sample sizes N=1000, 4000, 16000



We observe that although this is an asymptotic result, conformity of the mean with the Normal distribution is quite good even at the smallest sample size, and there is almost no visible deviation from symmetry remaining. This should not be

taken to indicate that a CLT is always a good approximation even at a very small sample size; here we are treating cases of independent random sampling, which is a relatively straightforward case. Other CLT's can be proven for data which have some dependence, but larger sample sizes will typically be required for this degree of conformity with the asymptotic distribution to appear.

We observe also that as we move to higher and higher sample sizes, the densities become ever more concentrated around the true mean, in conformity with the WLLN as well as the CLT. In fact if we take a range containing any given proportion of the data (for example, imagine marking points on the axes that contain about 99% of the area under each of these densities), the range required to contain the given proportion shrinks as the sample size increases– note again the difference in scales between the two figures. As sample size increases by a factor of four, the range containing the given proportion of the data shrinks by a factor of about two: that is, our estimates become more precise at a rate equal to the square root of the rate of increase of sample size. This is an example of 'root-N convergence', which appears frequently in simple parametric estimation problems such as this, and which can also be read from the left-hand side of the result in Theorem 11.2: this ratio converges to a constant distribution, and so the numerator must on average be shrinking at the same rate as the denominator, which (because σ is a constant) is necessarily shrinking at the rate of the term that it is divided by, $\sqrt{n}$. This is also a reflection of the result proven earlier that the variance of the sample mean in an independently and identically distributed random sample is σ^2/n, which means that the standard deviation of the sample mean is $\sigma/\sqrt{n}$ and is therefore declining in proportion to the square root of the sample size.

Note again that the existence of first and second moments is needed, and that this is a condition that can fail, particularly in financial data containing many extremes. Nonetheless, in most circumstances, the existence of a the first two moments is a safe assumption.

It is also crucial to remember that a CLT describes the distribution of a *statistic calculated from the data* and not the data themselves. Unless data arise from a process of summing or averaging, a central limit theorem does *not* provide any reason to suppose that the raw data should be approximately Normal.

11.4 Application to the distribution of sample proportions

It is straightforward to show that this result applies as well to the distribution of a proportion. Define a 0/1 variable such that the variable takes the value 1 if a certain condition holds, 0 otherwise. Let p be the proportion of cases in the population for which the condition is true. For example, let p represent the proportion of the population who would vote 'yes' to a referendum question, and for every person x_i from that population who is sampled, let the individual be coded as 0 if he or she will vote 'no', and 1 if he or she will vote 'yes'. The number of people who say that they will vote 'yes' can be denoted by N_y and the sample size by N, so that the sample proportion $\hat{p}$ who say that they will vote 'yes' is $\frac{N_y}{N}$. But $\frac{N_y}{N}$ is also the mean of

the 0/1 random variable x_i: $\frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}$ times the number of 1's in the sample, which equals $\frac{N_y}{N}$.

That is, $\hat{p}$ is the sample proportion, and is also the sample mean value of a 0/1 dummy variable indicating that the condition (vote 'yes' in the referendum) holds. Therefore results pertaining to a sample mean also apply to $\hat{p}$.

The variance of the random variable $\hat{p}$ takes a simple form, since this is a Bernoulli random variable, and the number of 'successes' (here, 'yes' votes) has a binomial distribution. We can work out the variance (or standard deviation) of $\hat{p}$ as a function of the population value p, so that p is the only unknown value in the sampling distribution.

Given random sampling from a population with true proportion p,

$$E(\hat{p}) = E\left(\frac{N_y}{N}\right) = \frac{1}{N}(Np) = p,$$

and

$$var(\hat{p}) = E(\hat{p} - E(\hat{p}))^2 = E(\hat{p} - p)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - p)^2 = \frac{1}{N}\sum_{k=0,1} p_k(x_k - p)^2$$

$$= \frac{1}{N}(p(1-p)^2 + (1-p)(-p)^2) = \frac{p(1-p)}{N},$$

where $p_0 = 1 - p$ and $p_1 = p$. Note that we obtained the variance here directly from the probability function: for any given sample point there is a probability p of obtaining a 1 when the random variable is sampled, in which case x_i exceeds the mean (p) by $(1 - p)$, and a probability $(1 - p)$ of obtaining a 0, in which case the random variable is below the mean by p. We have used the assumption that the sample points are independent to write the step $E(\hat{p} - p)^2 = \frac{1}{N}E(x_i - p)^2$.

Since the sample proportion is a sample mean of the 0/1 random variables, and since the mean and variance exist as we have just shown, we therefore can obtain the asymptotic distribution using Theorem 11.2. Substituting $\hat{p}$ for the generic expression $\overline{X}$ and p for the generic symbol μ, and the standard error of the proportion for the general expression for the standard deviation of $\overline{X}$ (that is, $\frac{\sigma}{\sqrt{(N}}$), we have

$$\frac{(\hat{p} - p)}{\sqrt{p(1-p)/n}} \quad \xrightarrow{D} \quad N(0, 1).$$

This result can be used to obtain confidence intervals for proportions, as in the confidence interval computations that we saw in Chapter 10.

Proof of the weak law of large numbers given above.[8]

Recall the Markov inequality: let X be a random variable and $g(.)$ a non-negative function on $\mathcal{R}$ such that $E(g(X))$ exists. Then

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

As with the proof of the Chebychev inequality given earlier, we make a choice of function $g(X)$ and of k; here $g(x) = (\overline{X} - \mu)^2,\ k = \epsilon^2$, where μ and σ^2 are the mean and variance of X (i.e. of the data) and ϵ will be the bound on the discrepancy between the true (population) mean μ and estimated (sample) mean $\overline{X}$. Note that the function $g(X)$ here contains the sample mean $\overline{X}$ rather than the value of the variable itself, X, which occurs in the Chebychev inequality.

We can re-write the statement above in equivalent form as

$$P(g(X) < k) \geq 1 - \frac{E(g(X))}{k} \ \forall k > 0.$$

Making the substitution of our choices for $g(X)$ and k,

$$P((\overline{X} - \mu)^2 < \epsilon^2) \geq 1 - \frac{E((\overline{X} - \mu)^2)}{\epsilon^2} \ \forall \epsilon > 0.$$

Now since we are dealing with a simple case of independently distributed data, we have $E((\overline{X} - \mu)^2) = \sigma^2/n$, as shown earlier. Therefore

$$P((\overline{X} - \mu)^2 < \epsilon^2) \geq 1 - \frac{\sigma^2}{n\epsilon^2}.$$

Taking square roots of the quantities inside the probability on the left side does not change the statement, and we will define $\delta = \frac{\sigma^2}{n\epsilon^2}$, leaving us with the statement

$$P(|\overline{X} - \mu| < \epsilon) \geq 1 - \delta, \qquad (A11.1)$$

for $\delta = \frac{\sigma^2}{n\epsilon^2}$ or re-arranging, $n = \frac{\sigma^2}{\epsilon^2 \delta.}$ The statement will remain true for larger δ (if the probability is $\geq 1 - \delta$, it is also $\geq 1 - \delta'$ for $\delta' > \delta$ since the latter is a lower probability). So (A11.1) holds for any $n \geq \frac{\sigma^2}{\epsilon^2 \delta}$ (integer constraints on sample size may make it impossible to find an n that makes this hold with equality.)

[8]This proof is based on that given by Mood et al. (1974).

Chapter 12

Sampling distributions revisited

In Chapter 10 we saw how we can construct confidence intervals for an estimated parameter, given knowledge of the distribution of that estimate. In the case is considered there, we were able to determine that distribution using theoretical results on the relations between distributions, but only by assuming that we knew the distribution of the input data. In typical practical cases, we don't of course know the distribution of the input data; we obtain in the form of a matrix or table, and set out to analyze some question of interest. Nothing is in general known about the distribution of any quantity in the data matrix; we cannot work out these distributions mathematically without knowledge of the inputs. We therefore need to have ways of approximating the unknown distributions to an acceptable degree of accuracy. There are two broad classes of approximation which are commonly used: approximations based on asymptotic theory, where we take the distribution to which another distribution will converge asymptotically as its approximation, and simulation-based approximation, where we use computer-generated random numbers to emulate a problem and attempt to approximate the relevant distributions. Simulation-based methods such as Monte Carlo tests and bootstrap tests have wide applicability, but require different forms for different types of problem and therefore requires some sophistication in their implementation. These simulation-based methods are beyond the scope of this book, at least at the time of writing. Although therefore we will discuss asymptotic approximations, it's worth underlining that simulation-based methods may also be used in this context, in part to check on the accuracy of asymptotic approximations in different cases.

In Chapter 11 we saw that we can obtain asymptotic distributions for estimates in some such cases, using other theoretical results, and in particular central limit theorems. Without knowing the distribution of the input data, we can determine nonetheless that the sample mean will have an asymptotically Normal distribution, given some simple conditions which will often apply. The asymptotic distribution will not correspond perfectly with the finite-sample distribution, but will provide a good approximation in a wide range of circumstances.

In the present chapter we will see how to apply central limit theorem results to obtain confidence intervals for estimators. We will therefore have moved to a set of methods that allow us to handle the realistic problem in which we begin with nothing more than a set of numbers of unknown distribution, and compute confidence intervals for estimates that can be expressed as the mean of something measurable. This kind of argument is widely applicable, and forms the basis for a great deal of applied statistical inference, giving approximate confidence intervals with a reliable justification. After we go through the method in the next section, we'll step back to

look at an example where we start with a set of numbers that we know very little about, and get some confidence intervals for estimates.

12.1 Sampling distributions based on a CLT

Consider the following problem. We want to know whether two random variables have the same mean, and we have a random sample from each. Let these two random samples be $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, and let the means of the underlying random variables X and Y be μ_x and μ_y. If $\mu_x = \mu_y$ then the mean of $(X - Y)$ is zero, so we can look at the question by considering the difference $X - Y$.

We will put a label on $X - Y$, and call it d for difference. Then we have a sample $\{d_i\} = \{x_i\} - \{y_i\}$ for $i = 1, \ldots n$, and $\overline{d} = \sum_{i=1}^n \frac{(x_i - y_i)}{n}$. We do not know the distribution of X, Y or d, but we will make the weak assumption that its mean and variance exist.[9] Then we can get a confidence interval for d from a central limit theorem: we know

$$\frac{(\overline{d} - \mu_d)}{\sqrt{\sigma_d^2/n}} \xrightarrow{D} N(0,1), \tag{12.1}$$

and that replacing σ_d^2 with $s_d^2 = \frac{\sum (d_i - \overline{d})^2}{(n-1)}$ leaves this asymptotic distribution result intact. We can then obtain an approximate (asymptotic) confidence interval for the true μ_d using the same manipulations that we saw earlier:

$$P\left(-z_{\alpha/2} < \frac{(\overline{d} - \mu_d)}{\sqrt{s_d^2/n}} < z_{\alpha/2}\right) \simeq 1 - \alpha,$$

which implies $P(-z_{\alpha/2}\sqrt{s_d^2/n} < (\overline{d} - \mu_d) < z_{\alpha_2}\sqrt{s_d^2/n}) \simeq 1 - \alpha$ and so

$$P(-\overline{d} - z_{\alpha/2}\sqrt{s_d^2/n} < -\mu_d < -\overline{d} + z_{\alpha/2}\sqrt{s_d^2/n}) \simeq 1 - \alpha$$

which, reversing the inequalities as we change sign, implies

$$P\left(\overline{d} + z_{\alpha/2}\sqrt{s_d^2/n} > \mu_d > \overline{d} - z_{\alpha/2}\sqrt{s_d^2/n}\right) \simeq 1 - \alpha, \tag{12.2}$$

or that d is in the interval $\overline{d} \pm z_{\alpha/2}(\sqrt{s_d^2/n})$ with probability approximately equal to $1 - \alpha$. (Note again that we say 'approximately' because this is not a finite-sample-exact confidence interval as we could obtain if we somehow knew for example that the original data were Normal. Instead this is an approximation based on the asymptotic results, which will tend to become more and more accurate as sample size increases.)

[9] We may know or observe sufficient conditions for this, such as that the two sequences are bounded, which guarantees the existence of all finite moments.

We have stated the problem here in the form of a difference between two series, which is a type of problem that is often interesting (are starting salaries in some job the same for men and women with the same qualifications? Are accident rates for first-year drivers the same for those who completed a driver education course and those who did not?) However the same method can be applied whenever we are interested in the mean of a random variable with an unknown distribution, as long as that distribution possesses the first two moments, which will very commonly be the case and can sometimes be verified unambiguously.

12.2 Example

Consider in more detail the example of the number of accidents that newly licensed drivers have in their first years of driving.[10] Note that this random variable is discrete $(0, 1, 2, \ldots)$ and is bounded on one side. It certainly cannot be Normally distributed.

We might look at whether the true mean number of accidents per year is some round number such as 1, for example, but that doesn't sound very interesting: the true answer is almost certainly going to be some fraction. A more interesting question is the male-female difference.

Consider that a driving school tracks data on its students for one year after licensing. Over a certain period, the school has 422 male graduates and 378 female graduates.

The numbers are not the same, so we would not simply take the difference between pairs of students as above. To illustrate a point, however, imagine for a moment that we have 378 observations on each group. We can then take pairs and construct the difference $d_i = x_i - y_i,\ i = 1, \ldots, 378$, that is, the difference between the number of accidents that the male driver has and that the female driver has in each pair; this can of course be positive, negative or zero in each case. Then, a confidence interval for the difference exactly fits the pattern above.

While this test would be straightforward, it fails to use all of the available information; the additional 44 male drivers are excluded from the sample, and of course the results of the test will depend upon which 44 drivers are excluded. In general, sample sizes from two groups may be unequal, and so we would like to be able to derive a test which does not depend on the assumption of equal sample sizes.

Bearing in mind that the number of accidents in a group can reasonably be supposed to possess same mean and variance (this would be guaranteed if the number of accidents is bounded above, for example if access to a vehicle is removed for anyone who has more than some number of accidents), we will be able to use a central limit theorem and therefore an asymptotic normal approximation for this problem. Let's

[10] We of course need to make a precise definition of the term 'accident': for example, an accident might be defined to be an event leading to damage which is reported to an insurance company. This definition would exclude many small incidents, which are not reported because the cost of repair is less than, or not much more than, the insurance deductible.

say that the quantity of interest to us is the mean number of accidents for each of the two groups, male and female drivers in the first year of licensing. (We could by contrast look at a different quantity such as the probability of having at least one accident.) The question of interest to us is whether the mean on each of these groups, μ_1 and μ_2, is the same, and one way to approach that is to construct a confidence interval for the difference in the mean number of accidents in the two groups. We can do that along the lines indicated above, computing $\overline{d}$ this time as $\overline{d}_1 - \overline{d}_2$. The problem now is to compute the standard error of this difference, which at first sight may look tricky because we do not have a single set of differences and so cannot compute it directly as the standard error of a particular data series. However, the quantity that we need can be computed straightforwardly using the expression for the variance of a linear combination of two random variables X and Y, that is $var(aX + bY) = a^2 var(X) + b^2 var(Y) + 2ab\ cov(X, Y)$, for weights a and b. Here, $a = 1, b = -1$ and it is reasonable to suppose that the covariance is approximately 0.[11] We therefore have $var(X - Y) = var(X) + var(Y)$. We can compute the estimated variance s^2 for each of the series $\{x_i\}$ and $\{y_j\}$ recording the number of accidents for each male driver indexed i and each female driver indexed j, and the variances of the *means* of the two series are then s_x^2/n_x and s_y^2/n_y. The estimated variance of the difference is then $s_x^2/n_x + s_y^2/n_y$ and the standard error of the difference $s_d = \sqrt{s_d^2}$ is the square root of that quantity. We can then compute a confidence interval using the expression (12.2) above.

To illustrate further, here is an example with some numbers computed on a simulated data set constructed with known probabilities of accidents. The sample sizes for male and female drivers are 422 and 378, and data series of these lengths were constructed such that each data point is the number of accidents for that driver, 0, 1 or 2. The data were constructed in such a way that the accident probabilities are

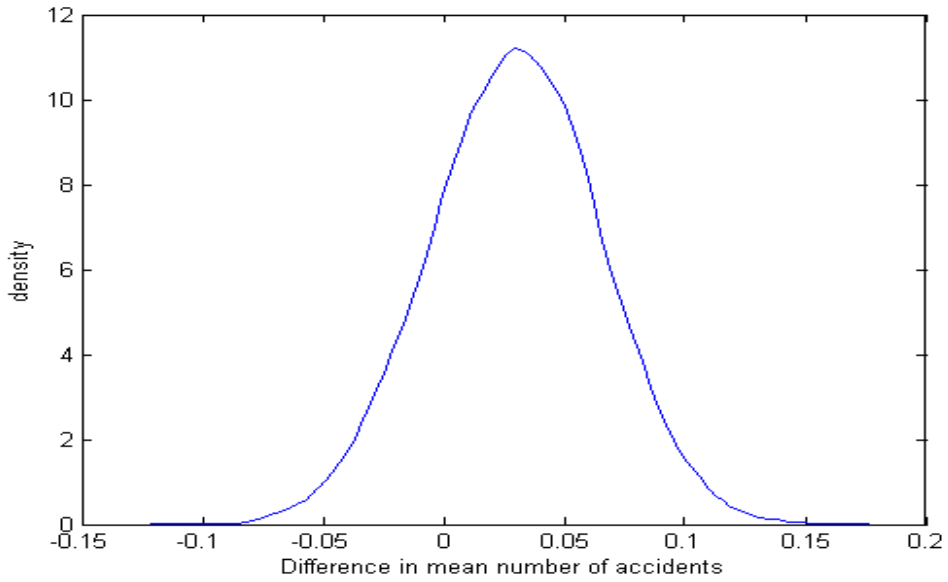
[11] This is an example of a case in which we may substitute a value for some quantity based on reasoning about the way the sample is constructed. Here, we might have samples of male and female drivers who don't know each other and live in different places, so that we feel confident that the driving experiences of individuals are independent of each other, so that their covariance must be zero. We may also be aware that this is not a literal truth; for example, imagine that one of the male drivers and one of the female drivers are a couple who like to send texts to each other, and drive different cars. If they do this while driving–one of the forms of behavior that clearly seems to raise accident rates–then they may each have elevated probabilities of an accident, and may even text each other a while both parties are driving simultaneously, so that the male and female accident rate data are no longer independent and the covariance would be positive. To the extent that this is possible, it seems likely to be a very small effect, so that in practice we would probably continue to use the approximation that the covariance between the two series is zero, being aware nonetheless that this is an approximation and may not be a literal truth.

not in fact exactly the same (the distribution will be described below). The statistics computed on the two samples were: $\mu_x = 0.4123$; $\mu_y = 0.3783; s_x^2 = 0.2429$; $s_y^2 = 0.2358$; therefore $\overline{d} = \mu_x - \mu_y = 0.0340; s_{\mu_x}^2 = s_x^2/422 = 5.7556 \times 10^{-4}; s_{\mu_y}^2 = s_y^2/378 = 6.2385 \times 10^{-4}$; finally using the expression for the variance of the difference as above, the standard error of the difference is $s_d = [s_{\mu_x}^2 + s_{\mu_y}^2]^{\frac{1}{2}} = 0.0346$. This is quite close to the difference itself, meaning that the difference is about one standard error away from zero; clearly therefore we do not have very strong evidence against zero being the correct value for the difference. The 95% confidence interval for the difference, using the percentage points of the asymptotic standard Normal distribution, is 0.0340 $\pm$ 1.96(0.0346) or [-0.0338, 0.1018], so that zero is well inside the interval.

That is, these data do not show clearly that there is any difference in average number of accidents between male and female drivers, although the sample average computed is somewhat higher for male drivers. In fact, for this example the data were generated to be such that male drivers do have a higher mean number of accidents; why then does the sample not give us a clearer result? Intuitively, it must be because the sampling variation is large enough to make any difference difficult to discern; the signal is obscured by noise. We can examine this more clearly by repeating the experiment many times, to see what happens. It's possible in this case because these data were constructed for this example rather than observed empirically, so instead of constructing one example we are free to construct many, and observe the outcomes in a large number of similar cases. In the simulations that will be recorded in the next two density functions, there were 10,000 experiments each with the same sample sizes of 422 and 378. The value on the lower axis in each of these figures indicates a male-female difference, and the density indicates how likely it is to observe differences of that magnitude.

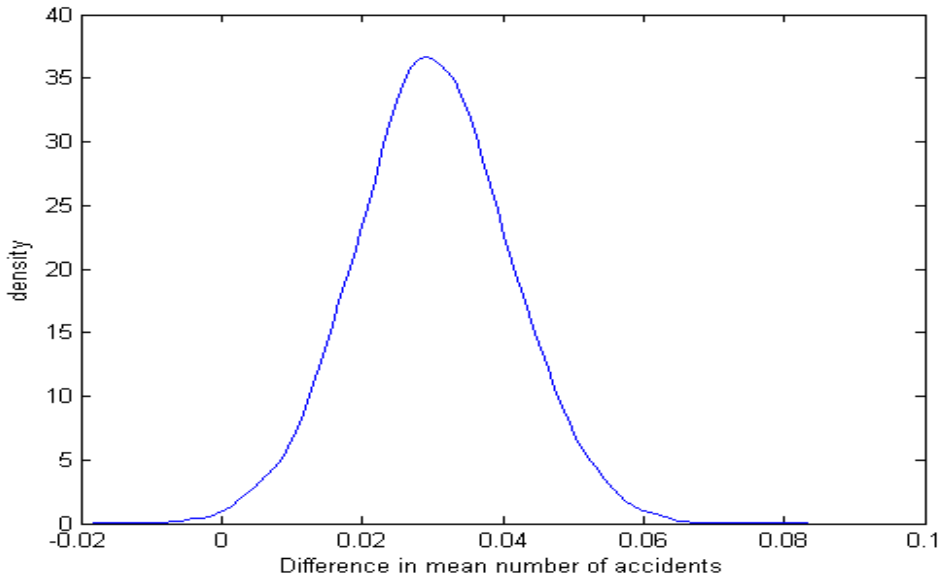
What we see in the first figure (apart from the fact that conformity with the shape of the asymptotic normal distribution is quite good, as we expect from a central limit theorem in a case with independent observations) is that the range of outcomes observed is quite broad, and that although these data are constructed in such a way that male drivers do have a slightly higher accident rate, samples in which the difference is zero or negative, that is that on a particular sample the female drivers had more accidents, are by no means uncommon. Many samples of these sizes, in other words, would lead to observations where the female drivers have a higher mean accident rate; the degree of random variation is very substantial relative to the quantity that we are trying to estimate. In the next example, we do the same thing but with much larger sample sizes: each of the previous sample sizes is multiplied by 10. We again take 10,000 examples on these larger sample sizes, and plot the density over each of these 10,000 estimated differences.

FIGURE 12.2.1
Distribution of the mean difference in number of accidents,
simulated sample of 422 male and 378 female drivers



Although the shape of the distribution is the same in the second case, we see that the numbers on the lower axis are less spread out (they're scaled down by $\sqrt{(10)}$, of course). Now, observing a male-female difference that is near zero or even negative is quite a rare event; the mean differences about the same, around 0.02, but there is much less sampling noise on these larger sample sizes and the results are more reliable in the sense that the estimated difference comes out to be positive, as it in fact is, over 95% of the time.

FIGURE 12.2.2
Distribution of the mean difference in number of accidents,
simulated sample of 4220 male and 3780 female drivers



In our single original example, zero was in the confidence interval. We could not reliably conclude that the mean difference in accidents between male and female drivers was a positive number. We now see that with a larger sample from the same process, we would have concluded that zero is outside even a 95% confidence interval: a process with a zero difference is unlikely to have generated these data.[12]

This illustrates a general feature about drawing, or failing to draw, a conclusion from data: it may well be that we will fail to see a difference between two things because our sample does not contain enough information (contains too much sampling noise). When therefore we fail to find a clear difference (or in the language of hypothesis testing that we will soon use, when we fail to reject the null hypothesis), it does not mean no difference is there (it does not mean that the hypothesis is true). It could simply be that our data were insufficiently informative to allow us to conclude that. Failing to show that something is false doesn't prove that it's true.

12.3 WHY DIDN'T WE USE A $t-$ DISTRIBUTION IN THAT EXAMPLE...

...since the variance was estimated?

[12]Data were generated by independent random draws such that for male drivers, the probability of no accident was 0.6, of one accident was 0.395, and of two accidents was 0.005. For female drivers, the probability of no accident was 0.65, of one accident was 0.345, and of two accidents was 0.005. Male drivers therefore have higher probability of at least one accident (0.65 vs 0.60) and higher mean number of accidents.

This example might also help us to recall that the exact distributional result for the $t-$ depends on quite strong assumptions about the distribution of the input data; normally distributed input data would lead to a statistic that has an exact $t-$ distribution. But that is clearly not the case; the input data are the numbers of accidents for samples of different people, which are integers with a distribution that is bounded and skewed right.

All we can rely on to obtain a distribution of the test statistic in this case is a central limit theorem, which tells us that the mean will have an asymptotic Normal distribution. If we are relying on our asymptotic approximation from the central limit theorem, then our asymptotic distribution is a Normal.

Is it wrong therefore to use a $t-$ distribution rather than a Normal to get the confidence interval, that is to use $t_{k,\alpha/2}$ for k degrees of freedom, rather than $z_{\alpha/2}$, in the confidence interval expression?

Bear in mind that the t_k distribution converges to the $N(0,1)$ as k increases. In other words, these distributions are asymptotically the same. The size of the confidence intervals given by each of these distributions will become arbitrarily close as the sample size increases. For example with a sample size of 60 we have, for a 95% confidence interval, $t_{60,0.975} = 2.0003$ and $z_{.975} = 1.960$ if we are using the standard Normal. The size of the confidence intervals that we obtain will therefore differ by about 2% (i.e. about 0.040/2). Of course, the distributions differ by greater percentages as we go farther out into the tails, so that if we wanted a 99% confidence interval or a 99.99% confidence interval, the percentage difference would be larger. Nonetheless, all of these differences are declining with sample size; for a sample size of 400 and a 95% confidence interval, we have $t_{400,0.975} = 1.9659$ and of course $z_{.975} = 1.96$ still.

Notice that, although the values are typically close and certainly converging as sample size (degrees of freedom) increases, the value for the $t-$ distribution is never less than for the Normal. Therefore confidence intervals computed using the $t-$ distribution will be slightly wider, or at least no smaller. This in turn means that we are making a slightly less strong, that is a slightly more conservative, statement if we compute those values using the $t-$ distribution. Given that this inference is approximate, and given that we would rather err on the side of weaker statements rather than statements which are excessively strong (in other words we would rather not exaggerate the strength of the conclusions that we can draw from the data), many statistical workers will tend to prefer the slightly more conservative confidence intervals produced by use of the $t-$ distribution. This is a perfectly sensible practice as long no one is deluded into thinking that the confidence intervals are exact because a $t-$ distribution is used rather than an 'asymptotic' Normal. Whichever of these distributions is applied here, we are relying on the asymptotic approximation provided by the standard Normal, and justified by the central limit theorem.

Chapter 13
Point estimators

Much of the time, the quantities that we want to estimate are scalar values, or sets of scalar values. These might be responses of one variable to another such as elasticities, probabilities, moments of the distribution such as the mean or variance, and so on.

For example, we might want to estimate the price elasticity of demand (proportionate change in demand divided by proportionate change in price) for gasoline, in order to reduce consumption and environmental damage, using a Pigovian tax.[13] This example also illustrates some of the ways in which we approximate an entire time sequence with a subset of the values; in fact, the introduction of a tax on gasoline will have an immediate effect but also a changing effect over time, as people adapt their behaviour to the new tax. The immediate impact might arise only through a reduction in optional or recreational trips in the car, but over time other adaptations become possible, such as buying a smaller car than when gasoline prices were higher, building (in response to demand conditions) more apartments near areas where many people work, rather than houses in the suburbs, and so on. Typically, we do not attempt to estimate the entire dynamic path of the response of gasoline demand to a change in price over the long horizon until it stabilizes at some value; instead we usually approximate the information in this path with a short-term elasticity (the immediate effect that we observe in the first weeks or months after the introduction of a tax, before capital expenditures have adjusted) and a long-term elasticity (the value to which the elasticity settles down after people have had time to adjust fully their stocks of capital, including cars and housing units). In this case we would have two price elasticities of demand for gasoline to estimate, the short-term and the long-term.[14]

Another example, which illustrates a different set of difficulties that we face, would be estimation of the average effect of completing a university degree on the

[13]This term is a reference to the classic work of A.C. Pigou (1920) on the use of taxes to reduce negative externalities.

[14]Of course, these values are different in different places and at different times, so that we constantly need to be updating our statistical information about these values. Nonetheless virtually all published estimates are less than one an absolute value, indicating inelastic demand: a given percentage change in gasoline prices leads to a smaller percentage change in demand. Typical values are around -0.4 for the short-term elasticity and around -0.6 for the long-term elasticity. This information helps to choose the tax rate that will reduce demand by a particular proportion.

annual income of an individual, at different points in the lifespan; say at ages 30, 40, and 50. In using the word 'effect' in the previous sentence we seem to imply a causal relationship, i.e. we imply that the fact of completing a university degree itself raises the income of the individual. It's possible of course that people who complete university degrees will have higher incomes on average at each of these ages, simply because people who complete a degree will on average have other qualities that will tend to lead them to higher incomes, such as ability, persistence, capacity for a high workload, and so on. The fact that a higher income is associated with completing a degree does not necessarily imply that the degree itself was the cause of the higher income. Attempting to distinguish causality from association is the subject of a large literature in statistics and econometrics, to which we will refer only briefly later in this book. For our present purposes, we simply need to note that this value, the increase in income associated with the degree at each age, is a set of scalar numbers that we want to estimate.

In order to obtain an estimate, we use an *estimator*.

Estimators

D13.1 An *estimator* is a function of the data that provides an estimate of a population ('true') quantity.

For example, earlier we defined the sample mean $N^{-1}\sum_{i=1}^{N} x_i$ as an estimator of the true mean of the distribution of a random variable X. An alternative estimator that we defined was the trimmed mean, $(N-2k)^{-1}\sum_{i=k+1}^{N-k} x_i$, estimating the same quantity, but with a different trade-off between efficiency and robustness. We saw therefore that there could be more than one estimator of the same quantity, with different properties.

D13.2 A *point estimator* is an estimator that produces a scalar value, or a vector of values.

In the next chapter we will consider interval estimators, which produce estimates of an interval within which some value lies, or is likely to fall.

If we use the label θ for the quantity of interest in the population, then we will typically write the estimate as $\hat{\theta}$, which is some function of a data set X so that we can write $\hat{\theta} = g(X)$.

Properties of estimators

There are typically numerous estimators that we might think of for a particular problem. In order to choose among them, we need some objective criteria that we consider desirable.

First define the distribution of an estimator, $F(\hat{\theta})$, with density (if it exists) $f(\hat{\theta})$ and mean $E(\hat{\theta})$.[15]

The bias of an estimator does not have quite the same meaning as in normal speech, where it usually refers to an attitude on the part of a conscious human being which will tend to push the individual toward one or another conclusion, independent of evidence. The meaning here is related but does not imply any unfairness or poor decisions; there are some contexts in which it makes sense to use a biased estimator (see for example Chapter $- - - - - - - - - - - - - - - - - - - -$, time series).

D13.3 The bias of an estimator $\hat{\theta}$ of a parameter θ is $E(\hat{\theta}) - \theta$.

An unbiased estimator has mean equal to the true value; a biased estimator does not.

D13.4 An *unbiased* estimator is one for which bias($\hat{\theta}$)=0.

An efficient estimator is one that uses information as completely as possible to pin down the true value to as narrow an interval as possible (recall that although we are talking about the value itself in this chapter, we will discuss estimating intervals in the next chapter); $\hat{\theta}_1$ is said to be more efficient than $\hat{\theta}_2$ if the variance of $\hat{\theta}_1$ is less than the variance of $\hat{\theta}_2$. Of course, if $\hat{\theta}_1$ has a larger bias, it may be worse for our purposes in spite of having lower variance, so that we may need to trade off the two qualities of low bias and low variance.

D13.5 An estimator $\hat{\theta}_1$ is more efficient than another, $\hat{\theta}_2$, if $var(\hat{\theta}_1) < var(\hat{\theta}_2)$.

In order to make a trade-off in some formal way, it's useful to define some function of the bias, variance, or other features of the distribution that describe how bad the outcome is considered to be when $\hat{\theta}$ misses θ by a particular amount. We will often then want to estimate the expected value of this 'badness' or loss over the distribution of the estimate.

D13.6 A *loss function* $\ell(\hat{\theta}, \theta)$ is a real-valued function that describes the loss associated with an estimate $\hat{\theta}$ when the true value is θ.

The loss may be purely a function of the estimation error, $\hat{\theta} - \theta$, in which case the loss function can be written as $\ell(\hat{\theta} - \theta)$.

[15]There may be points where the density does not exist because, for example, the distribution function jumps at a certain point, in which case the slope is non-finite and there is no finite value for the derivative, or density.

In order that this function can represent loss, it must be true that $\ell(\hat{\theta}, \theta) \geq 0$ and $\ell(\hat{\theta}, \theta) = 0$ where $\hat{\theta} = \theta$.

D13.7 A *risk function* $L(\hat{\theta}, \theta)$ gives the expectation of the loss associated with an estimate $\hat{\theta}$ when the true value is θ.

Although loss and risk are clearly distinct concepts, it is commonplace to refer to either of them as a loss function (that is, either the loss function or the expectation of that function over the density of the estimate).

Here are several examples of risk functions; in each case.strictly speaking, the corresponding loss function is defined at a single value, rather than as an expectation over the distribution of $\hat{\theta}$; for example, the squared-error or quadratic loss is $\ell(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$. However, again following common usage, we will generally use the term loss function for either loss or risk, relying on context to distinguish a single value from the expectation.

Mean squared error: $L(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2]$.

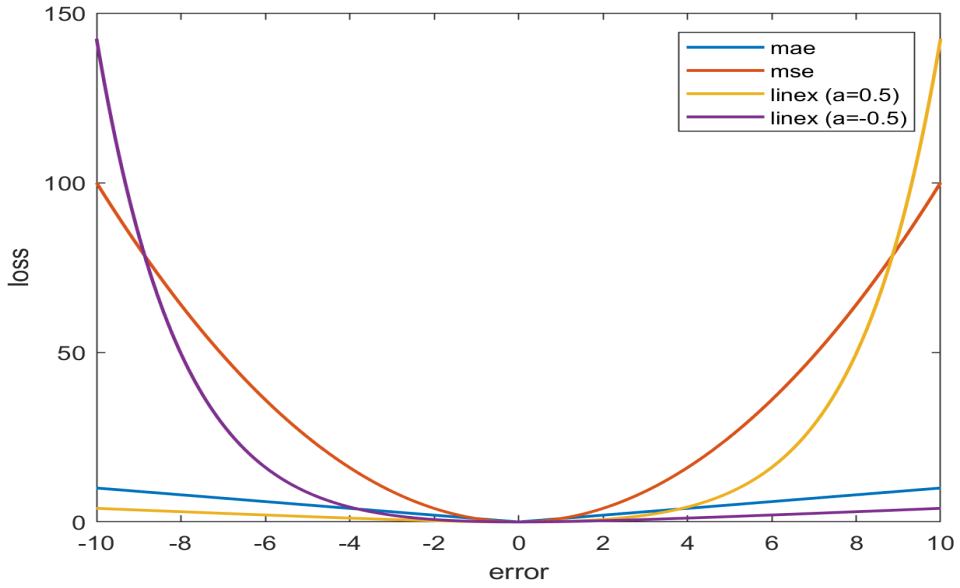
Mean absolute error: $L(\hat{\theta}, \theta) = E[|\hat{\theta} - \theta|]$.

Linear-exponential ('linex') loss: $L(\hat{\theta}, \theta) = E\left[\exp a(\hat{\theta}_i - \theta_i) - a(\hat{\theta}_i - \theta_i) - 1\right]$,
where $exp(.)$ is the exponential function, i.e. $exp(z) \equiv e^z$, and a is a shape parameter that determines whether errors increase exponentially on the positive or negative side.

These are theoretical values of the functions, defined using the (population) expectation. If one is estimating the loss (risk) function on empirical data, for example on a series of forecasts and outcomes, then the expectation would be replaced by an estimator such as the sample mean. For example if $\hat{y}$ is a forecast of a random variable y, and if a sequence of N forecasts of it becomes is available, then we could estimate the mean squared error from this estimator or the future value as $\sum_{i=1}^{N} (\hat{y}_i - y_i)^2$.

The next figure illustrates the three functions of a given error. Notice that the linex functions are individually asymmetrical around zero. In the pair illustrated here the left-hand side of one resembles the right-hand side of the other because the shape parameters are equal and opposite; however each is linear on one side of zero and exponential on the other. Notice also that the growth rate of the exponential function eventually comes to dominate the squared-error function as the error increases in magnitude.

FIGURE 13.2.1
Loss as a function of $\hat{\theta} - \theta$:
Absolute error, squared error and Linex loss functions



Each of these functions has other interesting properties. It's straightforward to show, although one needs to be very careful to keep track of parentheses and expectation operators, that the mean squared error is equal to the squared bias plus the variance, so that if the bias is zero, the mean squared error is equal to the variance of the estimator. Formally:

Theorem 13.1: (Mean squared error = squared bias plus variance.) Let $\hat{\theta}$ be an estimator of a parameter θ. Then

$$E[(\hat{\theta} - \theta)^2] = (E[\hat{\theta} - \theta])^2 + E[(\hat{\theta} - E[\hat{\theta}])^2].$$

Proof: See the Appendix.

The Mean Absolute Error function is non-differentiable at zero: although this is difficult to see in the figure above, the two straight lines meet at zero at a fixed angle rather than in a smoothly curved transition. This means that there is no unique tangent to the function at zero, and therefore no unique derivative. One result is that the MAE is more difficult to use in theoretical work; one cannot use the derivative of the function at all points. Proving results that can be proven straightforwardly for the MSE often requires much more mathematical sophistication. This is one of the reasons that the MSE is widely used; nonetheless it is often argued that symmetry is an unrealistic description of a user's loss for many problems. The

linex loss function is one way to allow a departure from symmetry, describing either positive or negative errors as causing exponential loss, depending upon the sign of the parameter embodied in the function. However, symmetry may be a reasonable approximation for many problems, or at least a reasonable starting point if one does not know what asymmetries may be present in users' loss functions.

Among other properties that we would like an estimator to have, consistency (i.e. with enough data, the estimator will converge probabilistically to the true value) and asymptotic normality are two of the most commonly investigated. It is often impossible to establish these properties by applying laws of large numbers and central limit theorems to functions of the data that the estimator is based upon.

D13.8 An estimator $\hat{\theta}$ is *consistent* for a parameter θ if $\hat{\theta} \xrightarrow{p} \theta$ as the sample size $N \to \infty$.

Defining asymptotic normality requires a little care because it may apply even in cases in which the density of the estimator does not exist at any finite sample size; we cannot therefore state that the density of the estimator must converge on the Normal density, as the density of the estimator may not exist at finite sample sizes. Nonetheless the distribution of the estimator may converge to the Normal distribution. The following is a simple definition which refers to the Normal distribution function, although there is no closed form for that function, and to the concept of convergence in distribution introduced earlier.

D13.9 An estimator $\hat{\theta}$ is *asymptotically Normal* if the distribution function of the estimator converges on the Normal distribution function.

Principles and methods for defining estimators

An estimator is a function of observable data. What function of the data should one choose for a given problem?

Sometimes this question is answered by specifying an arbitrary loss function for estimates, so that the estimator is chosen by minimizing this loss function. In some cases this can be done analytically, often by taking a derivative of the loss and minimizing it, but even in more difficult cases the function can usually be minimized using numerical methods; numerous algorithms are available to minimize arbitrary functions, although in many cases these will only find a local, rather than a global, optimum. For example, in cases in which we are trying to fit or 'explain' a large number of data points with a model having a few parameters, we might take as our function to be minimized either the sum of squared discrepancies between the predicted and actual values for each data point, or the sum of the absolute values of these discrepancies.

However, there are also a number of general principles available for defining and computing estimators, and there are reasons for relying on such principles when we can. First, although in some cases it may be easy to come up with a sensible

estimator by simple reasoning, in trickier circumstances having principles for defining estimators may be helpful, because simple reasoning does not point to any obvious estimator. As well, general properties of estimators defined according to a principle can sometimes be established, so that any example of an estimator defined according to such a principle will be known immediately to have certain features. Moreover, knowing that an estimator was defined according to some principle may make clear that there is a common structure with estimators used in another type of problem, possibly allowing the investigator to benefit from experience and knowledge that has arisen in other contexts.

In the rest of this section we will give a brief introduction to several of these principles, and we will see more examples of their application in later chapters, particularly when we review regression models.

Least squares (LS)

To understand least-squares estimation, it may be useful to return to the problem that originally motivated it: that of finding an approximate solution to a system of equations.

Consider the system

$$\begin{aligned} ax + by &= f \\ cx + dy &= g. \end{aligned}$$

Recall that a system of linear equations such as this may have no solution (for example, $2x + 3y = 5; 4x + 6y = 11$: these statements cannot both be true because the latter statement implies, dividing by 2, that $2x + 3y = 5.5$, which contradicts the former) or one unique solution (for example, $2x + 3y = 5; 4x + 5y = 11$)[16], or an infinite number of solutions (for example $2x + 3y = 5; 4x + 6y = 10$: in this case the two equations contain the same information (i.e. are linearly dependent), so any pair (x, y) that solves the first will also solve the second).

Expressed in matrix form, the general linear system is

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix},$$

or $Ax = b$. A unique solution to this exists if the matrix A is invertible, in which case the solution is $x = A^{-1}b$. Invertibility of the matrix A requires that the two equations be linearly independent, or equivalently that the matrix A be of full rank, or (again equivalently) that it have a non-zero determinant.

Now consider the case in which there are more equations, still with two unknowns in each. There will now be no solution to the system, unless some of the equations are

[16]From the first equation we have $2x = 5 - 3y$ and so $4x = 10 - 6y$; substituting the latter into the second equation we have $(10 - 6y) + 5y = 11$, or $y = -1$; substituting this back into one of the original equations, we have $2x = 8$ or $4x = 16$, i.e. $x = 4$.

redundant (i.e. linearly dependent with other equations in the system). For example, the system

$$\begin{aligned} 2x + 3y &= 5 \\ 4x + 5y &= 11 \\ 7x + 9y &= 15 \\ -x + 10y &= -11 \\ 7x - 7y &= 30 \end{aligned}$$

has no solution, which we can easily see because the solution to the first pair of equations, $(x, y) = (4, -1)$, does not solve any of the other equations.

Since no solution exists, we might decide to look for an approximate solution, using some criterion to define what constitutes a good approximation. We will begin by explicitly recognizing that the equations will not be solved exactly, by writing in a set of terms to describe the discrepancy (or 'deviation' or 'error' or 'residual') in each equation:

$$\begin{aligned} 2x + 3y &= 5 + \epsilon_1 \\ 4x + 5y &= 11 + \epsilon_2 \\ 7x + 9y &= 15 + \epsilon_3 \\ -x + 10y &= -11 + \epsilon_4 \\ 7x - 7y &= 30 + \epsilon_5. \end{aligned}$$

We have used the common notation ϵ_i to represent the discrepancy in the ith equation. Our aim now is to find a good approximation, which in general entails keeping the values of the ϵ_i terms as small as possible, but bearing in mind that a change in a value of x or y that lowers one of these discrepancy terms will in general raise another.

In order to come up with some solution, we might define a good approximation as follows: define the best approximating solution $(\hat{x}, \hat{y})$ as the pair of values that minimizes the sum of squared discrepancies, $\sum_{i=1}^{5} \epsilon_i^2$. This leads to an estimator:

$$(\hat{x}, \hat{y}) = argmin \left[\sum_{i=1}^{5} (\epsilon_i(x, y))^2 \right],$$

where $argmin f(\theta)$ means 'the value of the argument, θ, at which the function $f(\theta)$ is minimized'. Here the argument is the pair (x, y), and in the last expression we have written ϵ_i as an explicit function of this argument.

In the example just given, the least-squares approximate solution (to five significant digits) is $(\hat{x}, \hat{y}) = (3.4005, -0.8220)$, and the reader may easily verify that other pairs of values lead to a higher sum of squared discrepancies. The method by which

this solution was computed is given below, in Chapter 18, on linear regression. In a regression problem, the values x and y are estimated weights on data series given by the vectors (2,4,7,-1,7) and (3,5,9,10, -7).

Minimizing the sum of squared discrepancies (errors) is a very widely applied technique. Because the criterion is quadratic, its derivative is linear, leading to a linear rule for finding the optimum (minimum); again see Chapter ——- . This linear rule is not only convenient, but in some circumstances coincides with the form of estimator implied by other principles which have desirable general properties, in particular Maximum Likelihood.

Least absolute deviations (LAD)

LAD can be used to treat the same problem as above, but in this case, we replace the quadratic criterion with

$$(\hat{x}, \hat{y}) = argmin \left[\sum_{i=1}^{5} |\epsilon_i(x, y)| \right].$$

Changing the criterion in this way of course changes the relative weight of small and large deviations in defining the best approximation. With the LAD criterion, the 'badness' of a discrepancy changes linearly rather than quadratically with its magnitude, so that, for example, an error of 10 is only 5 times as bad as an error of 2, rather than 25 times as bad as the least-squares criterion would imply. Which of these is preferable will of course depend upon the problem and the person using the method. However, there is an additional important difference: the LAD criterion function, which is the absolute value function, is non–differentiable at zero (that is, the absolute value function has a corner or kink at zero, so that there is no unique tangent line. The minimum has to be found numerically rather than by deriving a simple equation for the estimator, and the fact that the criterion function is not everywhere differentiable means that more sophisticated mathematics is typically required in order to prove results concerning the properties of the estimator.[17]

Method of Moments (MoM)

A Method of Moments estimator is one in which unknown population moments are replaced, and estimated, by corresponding sample moments.

When we studied descriptive statistics in Chapter 3, we computed a number of functions of the data that we later saw as analogous to moments of the distribution. For example, the sample mean $\overline{X} = \sum_{i=1}^{N} x_i$ is the sample analogue of the population mean $E(X)$, and can be taken as an estimate of the population mean, although we

[17]Notice by the way that the criterion $argmin \left[\sum_{i=1}^{n} \epsilon_i(x, y) \right]$, without the absolute value, would not lead to good results: this implies that negative and positive errors cancel each other out, so that for example an estimator which leads to equal and opposite errors would be deemed just as desirable as one that picks the correct answer.

saw that other estimators such as a trimmed mean are also available. As the analogue of the population quantity, $\overline{X}$ is a Method of Moments estimator.

Some other descriptive statistics that we saw differ from the method of moments estimator. For example, the population variance is defined as $E(X-\mu)^2$, and the Method of Moments estimator takes the analogous form, replacing μ with its sample analogue, $\overline{X}$. The Method of Moments estimator of the variance is therefore $\sum_{i=1}^{N} \frac{(X_i-\overline{X})^2}{N}$, which differs from the unbiased estimator of the variance, which uses the factor $(N-1)$ in the denominator.

Method of Moments estimators (and a generalized version of the estimator) have frequently been used for estimating economic models in cases where the theory model suggests that a certain moment should take a value such as zero, in the population; for example, an economic theory may suggest that two quantities are independent, so that the expectation of the product of the two quantities should be zero. Imposing the condition that the expectation be zero ensures that the estimated model conforms with the assumed economic theory, and if this condition is then used to estimate other parameters, then those results also have been obtained by using and imposing the information implied by the economic theory. Whether this is desirable or not depends upon what one is trying to achieve: exploring the implications of the economic theory, versus exploring the content of the data set with minimal constraint.

Maximum Likelihood (ML)

Like LS, Maximum Likelihood is a very widely applied principle, and in fact in some important cases they lead to the same estimator.

Consider the density function (or probability function if discrete) corresponding with a random variable X. Although we often suppress the explicit dependence of a density function on the parameter vector, to write $f(X)$, a more complete notation for the density function is $f(X,\theta)$, where θ is the vector of parameters of the distribution; for example, if the distribution is Normal, the parameter vector $\theta=(\mu,\sigma^2)$ consists of the mean and variance parameters that characterize this distribution.

Earlier, we thought of this density, given knowledge of θ, as describing to us the values of X that are relatively likely or unlikely to arise. For example, if the distribution is Normal and $\theta=(\mu,\sigma^2)=(2,10)$, then approximately 95% of the values of Xthat we will observe live in the interval $2\pm 1.96\sqrt{10}$, since $\sqrt{10}$ is the standard deviation.

In using the density function in this way, we are taking the parameters as given, and asking where the observations are likely to arise. Another way of using the same information from the density would be to take a set of X values as given, and ask where θ is likely to lie: that is, we could try to deduce what the parameters must be, given a set of observations. More precisely, we could ask which parameter values in a given density are most likely to have led to the observed sample of data.

Recall that if we have a density function, then we can look for the most likely region in which to find data points by looking for the interval where the density

is highest. Analogously, if we take the data as given and look for the most likely region in which to find a parameter, we can again look for the interval in which the density–called likelihood when we view the data as fixed and the parameter vector as changeable–is highest. That is, we can maximize this likelihood.

The ML estimator is then.

$$\hat{\theta} = argmax\ \mathcal{L}(X, \theta),$$

where $\mathcal{L}(X, \theta) \equiv f(X, \theta)$, but now interpreted such that X represents a fixed set of N observations, and θ is varied.

To take a simple example,

Actually obtaining an estimator from the ML principle involves, conceptually, two steps. First, one needs to determine the likelihood function of the observations, using the assumed likelihood, which is a function of unknown parameters. This needs to be maximized over values of the parameters, in order to obtain estimates of them corresponding to the maximum of this likelihood. Sometimes this can be done analytically, but often the optimization is numerical, using a one of a number of well-known computational algorithms.

An important case in which the optimization may be straightforward analytically is that of a Normal likelihood function, so that a quadratic function of the observations arises, leading to a linear derivative. Setting a linear derivative to zero gives a linear formula for computing the ML parameter estimates. In some cases, as for example in obtaining parameter estimates of a linear regression model (again Chapter—- below), the ML and LS criteria result in the same linear expression for estimates, and so identical estimated parameters are given by the two methods.

Of course, in maximizing a likelihood function, we assume that we know what that function is, i.e., we assume that we know the true density of the data. In practical examples that will typically not be the case, but we may decide to use this estimation method with an assumed likelihood function which is believed to be a good approximation to the true likelihood. Estimation using the computational method of Maximum Likelihood, but where the assumed likelihood function is not identical to the true likelihood, is commonly called quasi-Maximum Likelihood (QML). For example, we may use the Normal density as a likelihood function, in a case where the unknown true density is also symmetric, but has thicker tails than the Normal. In this case we would obtain QML estimates.

Chapter 14

Interval estimators and confidence intervals

Statistical answers to empirical questions do not involve certainties: they are estimates, probabilities, ranges within which the true answer might lie, and so on. When we obtain a point estimate of something, we generally want further information to go along with it: how accurate is this estimate likely to be? For example, if we estimate a price elasticity of demand for gasoline to be -0.5 (in a particular place, at a particular historical time), we will generally also want to know whether that estimate is likely to be accurate to within, say, ± 0.1 or ± 0.4. The former interval gives us much clearer information about what is likely to happen in response to an increase in the gasoline tax. Similarly, if we estimate that 48% of voters will vote 'yes' to a referendum question, we will have a much better idea of the likely outcome if we are highly confident that the answer will lie within ± 0.01 of this value than if we can only be confident of lying within ± 0.05. Of course, in a case like this we can directly estimate the probability that the referendum result will be positive using the known (binomial, asymptotically Normal) distribution of the point estimate.

In general, a complete answer to a point estimation problem involves not only the estimate, but a measure of the uncertainty associated with that estimate, or alternatively a range within which the correct answer to the problem will probably (according to a formal computation) lie.

We have seen examples in previous chapters in which we could calculate the probability that a true value lies within a certain interval around an estimate. The present chapter will apply methods given in previous chapters on sampling distributions, and will review and extend our treatment of the computation confidence intervals for standard point estimation problems, where we can work with standard distributions given to us by statistical results such as a central limit theorem. We will also give further examples of computations of confidence intervals from empirical distributions of data or simulated statistics.

Definitions

D14.1 An *interval estimator* is a function of the data that provides estimates of lower and upper bounds A and B such that a quantity of interest (a parameter) lies in the interval $[A, B]$ with a given probability p.

D14.2 A *confidence interval* for a parameter β is an interval $[A, B]$ on the real line such that the probability that β lies in $[A, B]$ is $1 - \alpha$.

Typical values of α are small, such as 0.01, so that the probability that the interval contains the true value is close to 1.

Where we have more than one parameter to describe, we may estimate a confidence region: that is, a region of possibly more than one dimension such that the probability that each of a set of parameters will lie within the region is $1 - \alpha$.

Obtaining confidence intervals: two examples

We can think of ourselves as following a few simple steps in order to obtain a confidence interval for a given value. First, we need to obtain an estimate, a point estimate in this case, for the particular value. Next, we need to know the distribution that applies, at least approximately, to this estimate: this will typically be the step that requires the most sophistication. For many problems, however, we will be able to interpret our estimate as the mean of something that possesses at least two moments, so that we will be able to rely on a central limit theorem for this distribution. (In other cases, another distribution may apply, or the form of distribution may be unknown and we will have to use a computer simulation.) Finally, we need to know the parameters of this distribution: for example, if we are dealing with an asymptotic normal distribution, we will need to know the variance; if we are dealing with a χ^2 distribution, we will need to know the degrees of freedom, which will again allow us to determine how much of the distribution lies between particular bounds.

With a point estimate, a distribution and estimated values of the parameters of this distribution, we typically have the information that we need to determine how far on either side of the point estimate we need to draw our boundaries in order to have a given probability, such as 95% or 99%, that the true value will lie in our interval.

The next section gives two examples of confidence intervals for the difference in means of two random variables.

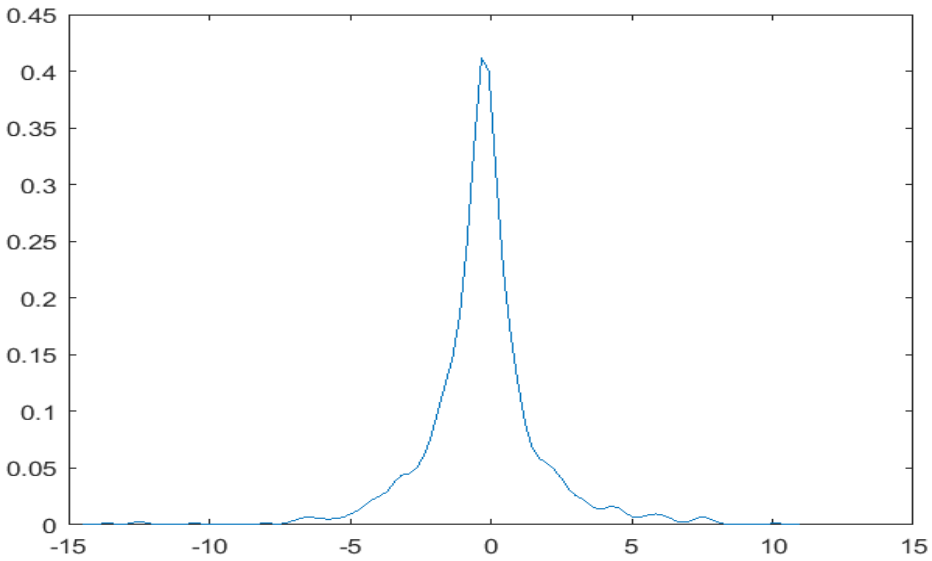
Confidence interval for the difference of two means: matched pairs

One of the most common problems that we see is determining whether two variables are genuinely different from each other, and in particular, whether the means of the two variables are the same or not. Two medical treatments for a given condition, two different types of education or training program, salaries offered to two different types of worker: are they genuinely different on average, or are the differences that we see just sampling error? We will now look at this problem by bounding the difference between two variables into a confidence interval.

For the first example, let us consider obtaining a confidence interval for the difference in the mean of two variables. We will begin with the relatively straightforward case of pairs of observations, so that we can directly compute the difference in each case. For example, pairs of plants in a greenhouse may be given one of two fertilizers, and their growth measured over the following months. For each pair, we can directly measure the difference in growth. The following example is based on simulated data on amounts of growth for each of 1000 pairs of plants. Note that the growth is non-negative, and strongly right-skewed, so that the data are clearly not

Normal. Figure 14.4.1 plots the differences in growth between the two plants in each of 1000 cases.

FIGURE 14.4.1
Estimated density of differences, $X_1 - X_2$
1000 data points



One question that we would naturally want to answer is whether there is any difference in the efficacy of the two types of fertilizer, and if so we would like to have an idea of how big that difference is; in other words, we want to obtain a confidence interval for the difference in growth for plants treated with fertilizer 1 versus those treated with 2.

Using these thousand data points, we compute a mean difference of -0.281, with a standard error of this estimated mean (the square root of the estimated variance of the differences divided by n) of 0.0657. We do not know the distribution of the data in this case. However, we are looking for the distribution of the mean of the data, which we can approximate using a central limit theorem. Given the conditions required to apply a central limit theorem, and in particular that the first two moments exist,[18] we have

$$\overline{d} \quad \stackrel{D}{\rightarrow} \quad N(\mu, \sigma^2/n),$$

where $\overline{d}$ is the estimated (or sample) mean of the difference; $n = 1000$; μ is the true mean of the difference, and σ^2 is the true variance of the difference (so that σ^2/n

[18]We can in principle check this by estimating the tail index, but it is in most cases simply assumed to be true because it is true and a very wide range of distributions.

is the variance of the mean of the difference, where n is the sample size). Given this approximate distribution from a central limit theorem, we can construct an approximate confidence interval using the Normal distribution as in earlier chapters: using the familiar value $z_{\alpha/2} = 1.96$ from the standard Normal distribution and replacing $\sigma$by its estimate, we have

$$P(\mu - 1.96(0.0657) < \overline{d} < \mu + 1.96(0.0657)) \simeq 0.95,$$

or, re-arranging and substituting $\overline{d} = -0.281$,

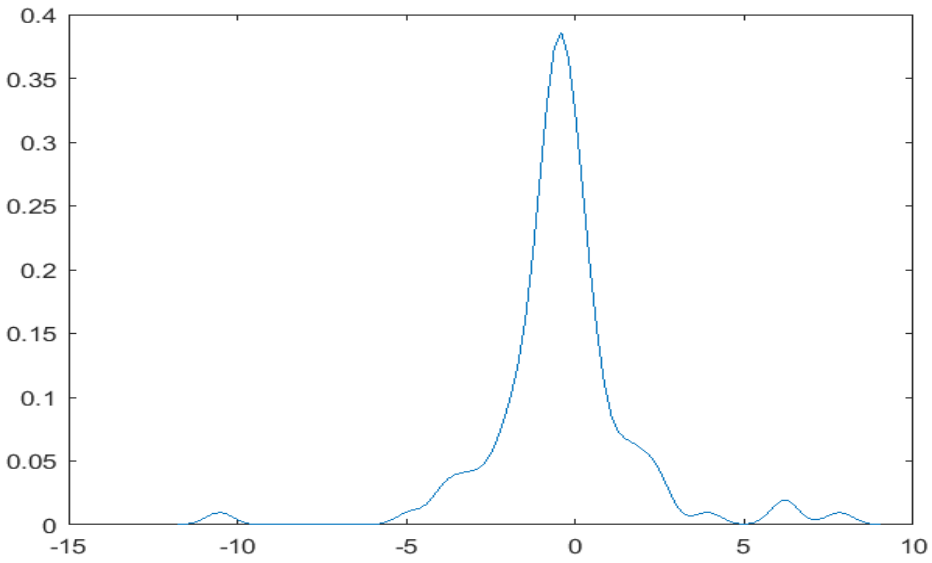
$$P(-0.281 - 1.96(0.0657) < \mu < -0.281 + 1.96(0.0657)) \simeq 0.95,$$

or $P(-0.410 < \mu < -0.152) \simeq 0.95$.

Notice that the value 0 is not in this confidence interval; we can be about 95% confident (in fact more than that, since 0 is well away from the boundary) that 0 is not the average difference between these two series. We have established in other words that we can be quite confident that there is a genuine difference in the means and that the observed difference is not simply due to sampling error. Moreover, it is the second random variable (treatment with fertilizer 2) that tends to be higher: $X_1 - X_2$ is on average negative. (This is the same type of computation that we will perform later in testing hypotheses: here, we might test the hypothesis that the true difference is zero; it turns out that we can be at least 95% confident that that is not so.

To illustrate the effect of sample size and also to make a point about what we can conclude when a value of interest *is* in a confidence interval, let's perform this same exercise again, but with only the first 100 sample points from this data set. The estimated density of the difference $X_1 - X_2$ from the first hundred sample points follows.

FIGURE 14.4.2
Estimated density of differences, $X_1 - X_2$
Sub-sample of 100 data points



The estimated density is similar, but of course with only 100 sample points, fewer extreme values appear and so the tails of the estimated density are not as wide. The corresponding estimates on the $n = 100$ sample are $\overline{d} = -0.366$ and $\sqrt{s^2/n} = 0.214$. The estimated confidence interval becomes

$$P(-0.366 - 1.96(0.214) < \mu < -0.366 + 1.96(0.214)) \simeq 0.95,$$

or $P(-0.785 < \mu < 0.053) \simeq 0.95$.

Although on the smaller sample the estimated mean difference was actually greater, 0 would not have been in the confidence interval on that sample: the estimated variance of the sample mean is larger (the square root of the sample size differs by the square root of 10, or about 3.16, so apart from sampling variation the standard error of the mean is bigger by this factor). On a sample of 100 points we would not have been highly confident that the difference is genuinely non-zero; on a sample of 1000 points, we could be.

This example illustrates a couple of points that are applicable widely. First, if a difference or other effect is present, it may nonetheless not be detectable in a small sample size. Because this is always a possibility, failing to detect an effect such as a difference does not prove that there is no effect. If someone had looked at the smaller sample and said, 'See? There is no mean difference between the two samples', that would be incorrect reasoning. It would be true to say that we cannot be confident that there is a difference, or to say that a zero difference is in our confidence interval,

but it would not have been correct to conclude that the data have established that the difference is zero. (Of course, there is a whole continuum of values that are in the confidence interval.) That one cannot find strong evidence against a thing does not prove that the thing is true. We will return to this point below when we study hypothesis testing.

Second, the strength of an effect, or in this case the size of the difference, will affect the number of sample points that we will typically require to detect it. In this example, the difference in mean between the two samples was small, and we could only be confident that the difference is non-zero once we had obtained a fairly a large number of sample points. Had the effect been much larger, we would have been able to detect it in a smaller sample size. In general, the more subtle the effect that we are trying to detect, the more sample points we will typically need in order to be confident that it is present.

Confidence interval for the difference of two means: independent samples

In the example given above, we could compute directly a set of differences between the two random variables: the differences become a new random variable, and we simply computed the sample mean and standard error of that random variable as inputs to a formula obtained from a central limit theorem, in order to obtain a confidence interval for the mean of the differences. In many cases however we will have two samples of data which are not matched, and not even of the same size, and so it will not be possible to compute a series of differences to make calculations on directly. We can nonetheless again compute a confidence interval for the difference in the means of the two samples, using the expression for the variance of the linear combination of random variables, as long as it is reasonable to suppose that the two samples are statistically independent.

For a test statistic based on the central limit theorem, we need to estimate the mean and variance of the mean. Recall that $E(X_1 - X_2) = E(X_1) - E(X_2)$: the mean difference between the two can be computed by taking the difference of the two means individually. In the previous section we were able to create the random variable $(X_1 - X_2)$ and estimated sample mean directly; with unmatched samples, we can nonetheless estimate the sample mean for each random variable and subtract one from the other. So computing the estimate $\overline{d}$ (as $\overline{X}_1 - \overline{X}_2$) is still straightforward, even though we don't have a sequence of differences.

Obtaining the estimated variance or standard error of the difference requires a little more reasoning. Recall however that for two random variables X_1 and X_2, we can compute the variance of a linear function of the two variables as $var(aX_1+bX_2) = a^2var(X_1) + b^2var(X_2) + 2abcov(X_1, X_2)$. Without a sequence of pairs from the two distributions, it may not be possible to estimate the covariance term. But if the circumstances are such that it seems reasonable to assume that the two variables are independent, then this covariance is zero. In that case, specializing the formula to the difference of the means, we have $var(\overline{X}_1 - \overline{X}_2) = var(\overline{X}_1) + var(\overline{X}_2) = var(X_1)/n_1 + var(X_2)/n_2$, where n_i is the sample size for variable $i, i = 1, 2$.

Here is an example on simulated data with $n_1 = 100$ and $n_2 = 1000$. We compute $\overline{X}_1 = 0.969$, $\overline{X}_2 = 1.192$, and so $\overline{d} = -0.223$; next, $var(\overline{X}_1) = 0.0135, var(\overline{X}_2) = 0.0021$ and so the standard error of $\overline{d}$, that is the square root of the expression given above for the variance of the difference, is 0.125. The corresponding 95% confidence interval for the difference in the means of the two series is then $-0.223 \pm 1.96 \times 0.125$, or [-0.468, 0.022].

Notice that the variable for which the sample size is smaller makes the larger contribution to the variance, or standard error of the difference: that variable is less precisely estimated, so more of the uncertainty about the difference stems from that variable. Imagine for example that we increased the second sample size from 1000 to 1 million or 1 billion: we would get increasingly precise estimates of the mean of the second random variable, but after a while increasing the sample size will have very little effect on the variance of the difference: in the limit, where we know the mean of the second random variable exactly, there will still be uncertainty about the mean of the difference, because the mean of the first random variable is uncertain.

Chapter 15

Statistical inference I: parametric hypothesis tests

Concepts and topics:

p-values

Type I and type II error

Test size

Test power and the power function

Data mining and multiple tests: first look and example

Chapter 16

Statistical inference II: non-parametric hypothesis tests

Chapter 17

The interpretation of statistical tests

Chapter 18
Linear regression

In Chapter 8, we introduced the conditional expectation. Since we do not normally know the full joint probability distribution of a set of variables, we cannot directly calculate the conditional expectation using definition 8.5.1, but we noted that we can approximate it using, for example, a linear regression. We will now learn how to estimate linear regression models to accomplish this.

Of course, conditional expectations may be non-linear, and so linear regression will typically be considered as an approximation (by Taylor's theorem, a straight line provides a local approximation to a function, but whether this approximation will be adequate depends on the purpose for which it will be used). We can to some extent handle non-linearity by using a linear function with non-linear terms, for example by including squared terms or logarithms of conditioning variables in a linear form.

More generally, we can estimate conditional expectations non-parametrically, that is, with flexible shapes which do not impose any fixed functional form, using for example kernel regression methods related to the kernel density estimators discussed earlier. This chapter will however deal with only the simplest case of a regression model in which a parametric linear form is assumed, and we need only estimate the parameters of this form to specify the conditional expectation function completely.

The least-squares (LS) criterion

Recall that in Chapter 13 we discussed some criteria by which to define a 'good' estimator, so that optimizing a criterion can give us a rule that will allow us in principle to pick the best estimator within some class. One of these criteria was the least-squares estimator. To draw the analogy to estimating a linear conditional expectation model, let us return to the example given in Chapter 8, of a model having the form

$$E(Y|X_1, X_2) = a + bX_1 + cX_2,$$

where Y is income, X_1 is age and X_2 is years of formal education. The parameters of this linear model,a, b, c, are to be estimated.

Recognizing that this form will not fit the data perfectly, that is that Y will not be exactly equal to its conditional expectation for each (or any) observation, we will usually introduce a term describing the discrepancy (or error) between the observed Y and its conditional expectation as given in the equation above. We therefore write

$$Y = a + bX_1 + cX_2 + \varepsilon,$$

where ε is the symbol used here for this discrepancy. Referring to each data point individually, instead of in this vector form, we could write

$$Y_i = a + bX_{1,i} + cX_{2,i} + \varepsilon_i, \quad i = 1, ...N,$$

for each observation i in a sample of N observations.

We would like to pick parameters (values that can be adapted to fit the data while remaining in the context of the model that we have specified) so that the model fits well. One criterion by which to define what it means to fit well is that the ε_i's are as small as possible, in the sense of minimizing their sum of squares, $\sum_{i=1}^{N} \varepsilon_i^2$. Using a power of 2 (a quadratic) has two advantages: all discrepancies count as positives (we wouldn't consider that large negative errors in a model somehow offset large positive errors, and make it a good model: instead we'd like errors to be small in magnitude, regardless of whether they are positive or negative); and the derivative of the square gives a linear function, leading to a linear rule for minimizing this quantity. (We could also use the sum of the absolute values of the errors as our criterion, leading to the least absolute deviations or LAD estimator, but this requires more sophisticated mathematics since the absolute value function is not differentiable at zero, so we cannot use elementary calculus to obtain a simple formula for the estimator, as we do with least squares.)

A simple one-variable regression

We consider estimation of a linear model of a variable y, in order to obtain a conditional expectation $E(y|X)$, where X is a matrix of conditioning variables. We have n observations, and there are k separate variables in X, each of which also has n observations available. We could write the model as:

$$y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots \beta_k X_{k,i} + \epsilon_i,$$

where the β's are parameters to be estimated, and ϵ_i is an unobservable error (discrepancy) term, allowing for the fact that the model will not fit perfectly. In matrix notation, we can write this as:

$$y = X\beta + \epsilon, \tag{1}$$

where y, X, β, ϵ are of dimensions $n \times 1, n \times k, k \times 1, n \times 1$ respectively. Note then that $X\beta$ becomes of dimension $n \times 1$ also (an $n \times k$ matrix multiplied by a $k \times 1$ vector). For example, the vector of parameters is

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

We assume that ϵ has a mean of zero, which in practice we can assure by including a constant (intercept) in the model, as we will see later.

Each column of the X matrix represents a different data series, and the n elements in that column are the observations on the data series. For example, in a model involving individual human subjects, the columns might represent age, gender, years of formal education, etc. Each row of the matrix would represent a particular individual, so reading across the row we have that individual's age, gender, years of formal education...

Our aim again is to estimate this model in order to obtain a conditional expectation of y given the available X variables, recognizing that there may be no causal link between X and y, and that we could also condition on other variables.

If we estimate the parameters β, then the estimated conditional expectation becomes: $\widehat{E(y|X)} = X\hat{\beta}$, since $E(\epsilon) = 0$. (Dropping the matrix notation for a minute, this is equivalent to $\widehat{E(y_i|X_i)} = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} + \ldots \hat{\beta}_k X_{k,i}$). We use the circumflex, $\widehat{E(y_i|X_i)}$, to indicate that this is an estimated conditional expectation (although in practice people often omit this symbol).

Now we can ask how we should estimate the parameters β, that is, the unknown weights in this linear approximation to the conditional expectation function. The most commonly applied method for this simple model is that of least squares, i.e., one minimizes the sum of squared residuals (the estimated errors) by choice of β.

Consider first a simple case that we can handle without matrices. Let $y_i = \beta x_i + e_i$, and let the estimated version of the model be $y_i = \hat{\beta} x_i + \hat{e}_i$. The 'residuals' are the $\hat{e}_i$, and the sum of squared residuals is $\sum_i (y_i - \hat{\beta} x_i)^2 = \sum_i [y_i^2 - 2\hat{\beta} x_i y_i + (\hat{\beta} x_i)^2]$. Taking the derivative with respect to $\hat{\beta}$ and setting to zero for an optimum, we have $-2\sum_i (x_i y_i) + 2\sum_i \hat{\beta} x_i^2 = 0$, or

$$\hat{\beta} = \sum_i (x_i y_i) / \sum_i (x_i^2). \tag{2}$$

Notice that we began with a quadratic criterion– the sum of *squared* errors– and so by taking a derivative, the square becomes a linear rule ($d/dx(x^2) = 2x$). Equation (2) is a linear rule for computing the coefficients that minimize the sum of squared residuals: the (ordinary) least squares estimator, or OLS .

We can derive the more general solution for the SSR-minimizing coefficients in equation (1) using matrix differentiation, yielding the solution

$$\hat{\beta} = (X'X)^{-1}X'y. \tag{3}$$

Notice that equation (2) is a special case of this, that applied when X has only one column. In the next section we will derive this result.

Multiple regression

In order to derive the estimator just mentioned (i.e. for any number of regressors), it's important to begin by representing the data in a standard form; the answers will correspond with this standard form and will be readily interpretable. To return to equation (1) above, we represent the data on the 'dependent' variable y as:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

where again n is the sample size. Note that the data are ordered from first to last observation; in cross-sectional data the order may be of no importance, but in time series data the order is a crucial element of the data set and must be preserved.

The matrix of variables on which we are conditioning (the 'independent' variables) is $n \times k$, with elements:

$$X = \begin{pmatrix} 1 & X_{1,1} & X_{2,1} & \dots & X_{k-1,1} \\ 1 & X_{1,2} & X_{2,2} & \dots & X_{k-1,2} \\ 1 & X_{1,3} & X_{2,3} & \dots & X_{k-1,3} \\ \vdots & & & & \\ 1 & X_{n,1} & X_{n,2} & \dots & X_{k-1,n} \end{pmatrix},$$

where in this example we have labelled the individual observations with the variable number first, and observation number second. This labelling convention could be changed, but we do need the observations (rows) in order from first to last, and the individual variables (columns) in whatever order we want; the order of the estimated weights $\hat{\beta}_i$ will correspond with this order of the variables. The first column, of ones, will be a constant intercept in the model. Its presence guarantees the the sum of the estimated errors in model (1) will be zero.

We will again choose parameter estimates $\hat{\beta}_i$ in order to minimize the sum of squared residuals: that is, we minimize $\epsilon'\epsilon = \sum_{i=1}^{n} \epsilon_i^2$, or

$$\min_{\beta}(y - X\beta)'(y - X\beta) = \min_{\beta}(y'y - \beta'X'y - y'X\beta + \beta'X'X\beta).$$

The two central terms on the right-hand side are equal, since one is the transpose of the other and they are both scalars, therefore the same. So we can write the minimization as

$$\min_{\beta}(y'y - 2\beta'X'y + \beta'X'X\beta).$$

Taking the derivative with respect to β using the rules given in the Appendix and setting the result to zero for an optimum, and using the symbol $\hat{\beta}$ now to denote the value that solves the equation, we have

$$-2X'y + 2X'X\hat{\beta} = 0 \longrightarrow X'X\hat{\beta} = X'y \longrightarrow \hat{\beta} = (X'X)^{-1}X'y.$$

Note that this assumes invertibility of the matrix $X'X$, which is equivalent to the assumption that $X'X$ is of full rank, which in turn implies that none of the rows or columns of X is a linear combination of any other row or column. This essentially rules out redundant regressors, which would be impossible to distinguish.

Computing standard errors of parameter estimatesn

The estimated parameter vector, $\hat{\beta}$, is of course a random variable (being a function of the data y), and so has a distribution around the 'true' values β. The way in which we estimate the variances of these estimates (and therefore their standard errors) depends upon what we can take to be true about the process, and there are many techniques for obtaining these estimates, including simulation-based techniques such as the bootstrap. Here, we will continue to consider only the simplest case, with some strong assumptions on features of the process. In particular, for the regression model $y = X\beta + \epsilon$, we will assume that:

(i) $E(\epsilon) = 0$
(ii) $E(\epsilon\epsilon') = \sigma^2 I_n$
(iii) $E(X'\epsilon) = 0$
(iv) $\text{rank}(X'X) = \text{rank}(X) = k$.

Assumption (i) is not restrictive since we can place a constant into the regression model to account for any non-zero intercept, which will in fact also guarantee that the sum of the residuals, $\hat{\epsilon}$, will be exactly zero.[19] The second assumption indicates that each one of the errors has equal variance, and so is in this sense each observation is equally reliable and should get equal weight; if this assumption does not hold, we can get instead compute *generalized least squares* estimates. Assumption (iii) is critical, since if this does not hold, the unobservable errors will project onto the space spanned by the X's, changing the estimated coefficients.s Finally the linearly independent regressors assumption (iv) is necessary in order to invert the matrix $X'X$, and so it will be obvious if this assumption fails– it will not be possible to compute the coefficients from the formula $\hat{\beta} = (X'X)^{-1}X'y$.

[19]You can prove this as an exercise.

Now let us compute the variance-covariance matrix of the parameter estimates. We begin with an additional assumption which is not necessary for regression in general, but simplifies this computation by giving us a case of unbiased parameter estimates. That is, we assume that the regressors can be treated as non-stochastic, as when they are chosen values for an experiment.

We begin by writing $\hat{\beta} = (X'X)^{-1}X'y =$

$$(X'X)^{-1}X'(X\beta + \epsilon) = (X'X)^{-1}X'(X\beta) + (X'X)^{-1}X'\epsilon = \beta + (X'X)^{-1}X'\epsilon.$$

So $E(\hat{\beta}) = \beta + E[(X'X)^{-1}X'\epsilon] = \beta + (X'X)^{-1}E[X'\epsilon] = \beta$, since the last term is zero by (iii) above. So $E(\hat{\beta}) = \beta$, and the estimator is unbiased in this non-stochastic regressor case (more generally, we could obtain the weaker result that the probability limit of $\hat{\beta}$ is β, with stochastic regressors).

Next, we compute $var(\hat{\beta}) = E(\hat{\beta} - \beta)(\hat{\beta} - \beta)'$. Using the results just obtained above, $(\hat{\beta} - \beta) = (X'X)^{-1}X'\epsilon$, and so

$$var(\hat{\beta}) = E((X'X)^{-1}X'\epsilon)((X'X)^{-1}X'\epsilon)' = E((X'X)^{-1}X'\epsilon)(\epsilon'X(X'X)^{-1}.)$$

Finally, assuming the X's to be non-stochastic means that they can be taken out of the expectation (e.g. $E(cZ) = cE(Z)$ where c is non-stochastic), so we have

$$var(\hat{\beta}) = (X'X)^{-1}X'E(\epsilon\epsilon')X(X'X)^{-1} = (X'X)^{-1}X'[\sigma^2 I_n]X(X'X)^{-1}$$

by assumption (ii) above. Finally, moving the scalar σ^2, the identity matrix becomes redundant (like multiplying by 1, we don't need to write it explicitly), and we obtain

$$var(\hat{\beta}) = \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}.$$

We can estimate the error variance σ^2 as $s^2 = \hat{\epsilon}'\hat{\epsilon}/(n-k)$, and we have the fully operational estimator $s^2(X'X)^{-1}$ for the variance-covariance matrix of the estimated parameters, $\hat{\beta}$. The square roots of the diagonal elements of this matrix are then the standard errors of the individual parameter estimates $\hat{\beta}_i,\ i = 1, \ldots, k$.

$R^2 = 1 - SSR/SST$
$y - X\hat{\beta} = \hat{\epsilon}$
$SSR = \hat{\epsilon}'\hat{\epsilon}$
$SST = (y - \overline{y})'(y - \overline{y})$

$H_0 : a'\beta = c \Rightarrow a'\beta - c = 0$

Appendix to Chapter 18

A1: Matrix Differentiation

Differentiation of matrices is essentially the same as scalar differentiation; that is, the principles of calculus applied are identical. The difference is that we may be taking the derivative of one *or more* quantities with respect to one *or more* others. So we need to represent the answers in matrix form. More than one convention for doing so is possible.

Let a, A, x be of dimension $n \times 1, n \times n, n \times 1$ respectively. Clearly $a'x = \sum a_i x_i$ so

$$\frac{\partial(a'x)}{\partial x_i} = a_i.$$

This is simply scalar calculus, since $a'x$ is a scalar. We can do the same with respect to each element of the vector x, however, and put the answers together into a new vector– which will be a:

$$\frac{\partial(a'x)}{\partial x} = \left[\frac{\partial(a'x)}{\partial x_1}, \frac{\partial(a'x)}{\partial x_2}, \ldots, \frac{\partial(a'x)}{\partial x_n}\right]' = (a_1, a_2, \ldots, a_n)' = a.$$

Similarly,

$$\frac{\partial(a'x)}{\partial x'} = a'.$$

Some other rules:

$$\frac{\partial(x'Ax)}{\partial x} = (A + A')x$$

$$\frac{\partial(x'Ax)}{\partial x \partial x'} = (A + A')$$

$$\frac{\partial(x'Ax)}{\partial A} = xx'$$

$$\frac{\partial log(det(A))}{\partial A} = (A')^{-1}, \quad \text{if} \quad det(A) > 0.$$

These are the rules which we will be using. Other extensions of scalar rules can of course be derived as well.

A2: Covariance matrices and variance of a linear combination

Let a, V, x be of dimensions $n\times 1, n\times n, n\times 1$ respectively. Then $a'x = \sum_{i=1}^{n} a_i x_i$, a scalar (1×1) quantity.

Let μ be the mean vector of the vector x, so that $E(x - \mu) = 0$.

The variance-covariance matrix of x, or simply the 'covariance matrix' or 'variance matrix,' is an $n \times n$ matrix such that each element (i, j) represents $E[(X_i - \mu_i)(X_j - \mu_j)]$; where $i = j$, these terms are variances and where $i \neq j$, the terms are covariances. Since $E[(X_i - \mu_i)(X_j - \mu_j)] = E[(X_j - \mu_j)(X_i - \mu_i)]$, the (i, j) element of the matrix is equal to the (j, i) element, and the matrix is therefore symmetric.

We can represent the covariance matrix in vector notation as

$$var(x) = E[(x - \mu)(x - \mu)'].$$

Notice that the transpose is on the second vector, so that we obtain an $n \times n$ matrix; if the transpose were on the first term, we would obtain the inner product, $\sum_{i=1}^{n}(x_i - \mu_i)^2$, a scalar.

With the covariance matrix, we can obtain the variance of any linear combination of the x's. Since $var(x) = E[(x - \mu)(x - \mu)']$, we have

$$var(a'x) = E[a'(x - \mu)(x - \mu)'a] = a'var(x)a.$$

Note that this expression is of dimension $(1 \times n)(n \times n)(n \times 1)$ or 1×1, ie a scalar.

Let's use this to derive the simple rule for the variance of a linear combination of two random variables that we stated earlier, ie. $var(b_1X + b_2Y) = b_1^2 var(X) + b_2^2 var(Y) + 2b_1 b_2 cov(X, Y)$.

The vector of weights in the linear combination is

$$a = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \text{ for } x = \begin{pmatrix} X \\ Y \end{pmatrix},$$

so that $a'x = b_1X + b_2Y$. The covariance matrix is

$$V = \begin{pmatrix} var(X) & cov(X, Y) \\ cov(X, Y) & var(Y) \end{pmatrix}$$

.

Finally

$$var(a'x) = a'var(x)a = (b_1 b_2) \begin{pmatrix} var(X) & cov(X,Y) \\ cov(X,Y) & var(Y) \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

$$= \begin{bmatrix} b_1 var(X) + b_2 cov(X,Y) & b_1 cov(X,Y) + b_2 var(Y) \end{bmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

$$= b_1^2 var(X) + b_1 b_2 cov(X,Y) + b_2 b_1 cov(X,Y) + b_2^2 var(Y)$$
$$= b_1^2 var(X) + b_2^2 var(Y) + 2 b_1 b_2 cov(X,Y).$$

Part IV: Surveys of more advanced methods

Chapter 19
Time series and forecasting

One fundamental goal of time series analysis is to measure some properties of time series, and then to define model classes which can mimic these properties as closely as possible with a moderate number of parameters. Estimating a model from one of these classes then gives an estimated representation which should behave similarly to the underlying time series, and so the forecasts from the process should be reasonable forecasts of the time series.

Notes:

Autocovariance at lag 1:

$$E[(y_t - \mu)(y_{t-1} - \mu)]$$

..at any lag k :

$$E[(y_t - \mu)(y_{t-k} - \mu)]$$

... and the autocovariance function (ACVF) is the set of autocovariances at lags $0 \dots \ell$. The autocovariance at lag 0 is just the variance of the process. Note that we have treated the mean as being the same for both the variable and its lag: this is an implication of covariance stationarity. A commonly used notation for the autocovariance function is $\gamma(k), k = 0, \dots \ell$.

As in cross-sectional cases, we can scale the (auto)covariance to obtain the autocorrelation. At lag 1, the autocorrelation is

$$E[(y_t - \mu)(y_{t-1} - \mu)]/\sigma_y^2.$$

Notice that, in the cross sectional case, we divided by the product of the standard deviations of the two variables, $\sigma_x \sigma_y$. We're essentially doing the same here, but in this case the two variables are the same, just measured at different points in time. If the variable is covariance stationary, then the standard deviation of y and of its lag are the same, and so we can just write σ_y^2.

The autocorrelation at lag k is then

$$E[(y_t - \mu)(y_{t-k} - \mu)]/\sigma_y^2,$$

and the autocorrelation function is the set of all autocorrelations from lag 0 (if this is included: it's always equal to one) to some value ℓ. A commonly used notation for this is $\rho(k)$, so the ACF is $\rho(k) = \gamma(k)/\gamma(0)$. As with a regular correlation, the autocorrelation is bounded into the interval $[-1, 1]$, and so is readily interpretable.

For these definitions to be meaningful, the mean and variance of the process must exist. This corresponds with the definition of covariance stationarity (sometimes called 'weak stationarity').

Definition. A time series process is said to be covariance stationary if the first two moments of the process, the mean and variance, exist and are constant.

Another form of stationarity is strict (or 'strong') stationarity:

Definition. A time series process is said to be strictly stationary if the distribution function of the process is constant over time: $F(y_t) = F(y_{t+s})$ for any integer value s.

Clearly, covariance stationarity does not imply strict stationarity: the first two moments might exist and be constant, but the third moment (measuring skewness) might change, violating strict stationarity. It might seem that strict stationarity implies covariance stationarity, but this is also not true: a process might have a constant distribution function, but the second moment might not exist, so it would be strictly stationary but not covariance stationary.

Either definition of stationarity is violated by a process with a trend– a trending process is tending to grow (or decline), and so its mean cannot be constant.

Forecasting

A (time series) forecasting model is a model with links the future with the past: for example,

$$y_{t+1} = \alpha + \gamma_1 y_t + \gamma_2 y_{t-1} + \beta_1 x_t + \beta_2 z_t + \varepsilon_{t+1}. \tag{9.1}$$

(We can also use prediction models for cross-sectional cases, for example predicting whether an individual's prescription drug expenditures given a vector of observable characteristics; in this case observations in a data set would normally be treated as independent, so that there is no information in the ordering, and no 'lagged' values would be relevant.)

In the case of a model such as (9.1), our forecast uses information on values of y, x, z occuring before the date being forecast, in this case dated no later than time t, yielding a forecast

$$\hat{y}_{t+1} = \hat{\alpha} + \hat{\gamma}_1 y_t + \hat{\gamma}_2 y_{t-1} + \hat{\beta}_1 x_t + \hat{\beta}_2 z_t, \tag{9.2}$$

since future values of the error term are of course unobservable.

An even simpler form of model is the pure autoregressive model:

$$y_{t+1} = \alpha + \gamma_1 y_t + \gamma_2 y_{t-1} + \varepsilon_{t+1}, \tag{9.3}$$

in which the process is modelled as being a function of its own past values alone; here we have used two lagged values (equivalently, we could write this as

$$y_t = \alpha + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \varepsilon_t$$

), meaning that we are using an autoregressive model of order 2, or an AR(2). Recall that, in estimating this model, we will lose 2 observations (or k observations, for an AR(k)), since the first observation would be modelled as

$$y_1 = \alpha + \gamma_1 y_0 + \gamma_2 y_{-1} + \varepsilon_1$$

, which refers to the unobservable pre-sample values y_0 and y_{-1}. Since we can't see those, we have to start at observation 3, which refers only to observations 2 and 1 on the right-hand side.

Chapter 20
Models of qualitative dependent variables

References

Cramèr, H. (1955) *The Elements of Probability Theory.* Wiley, New York.

Galbraith, J.W. and S. van Norden (2011) 'Kernel-based Calibration Diagnostics for Recession and Inflation Probability Forecasts.' *International Journal of Forecasting* 27, 1041-1057.

Heston, A., R. Summers and B. Aten (2002) Penn World Table v. 6.1. Center for International Comparisons at the University of Pennsylvania.

Hodgson, D. and K. Vorkink (2004) 'Asset pricing theory and the valuation of Canadian paintings.' *Canadian Journal of Economics* 37, 629-655.

Hogg, R.V and A. Craig (1959) *Introduction to Mathematical Statistics.* Macmillan, New York.

Hume, D. (1739) *A Treatise of Human Nature, Book I: Of the Understanding.* London.

Hume, D. (1748) *An Enquiry Concerning Human Understanding.* London.

Johnson, N.L. and S. Kotz (1970) *Distributions in Statistics: Continuous Univariate Distributions-I.* Wiley, New York.

Kendall, M.G., A. Stuart and J.K. Ord (1991) *Kendall's Advanced Theory of Statistics. Oxford University Press, New York.* Fifth edition of: Kendall, M.G. (1946) *The Advanced Theory of Statistics.*

Knight, F.H. (1921) *Risk, Uncertainty and Profit.* Houghton Mifflin, Boston.

Mood, A.M., F.A. Graybill and D.C. Boes (1974) *Introduction to the Theory of Statistics.* McGraw-Hill.

Neyman, J. (1950) *A First Course in Probability and Statistics.* Holt, New York.

Pigou, A.C. (1920) *The Economics of Welfare.* MacMillan, London.

Popper, K.R. (1959) *The Logic of Scientific Discovery.* Hutchinson, London. Translation of the German original *Logik der Forschung,* Vienna, 1935.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Student (1908) 'The probable error of a mean.' *Biometrika* 6, 1-25.

Todhunter, I. (1865) *A History of the Mathematical Theory of Probability.* Macmillan, Cambridge.

Summary of notation and abbreviations

Abbreviations in text

e.g. *L. exempli gratia:* for example
et al. *L. et alii:* and others
etc. *L. et cetera:* and the others
i.e. *L. id est:* that is
v.i. *L. vide infra:* see below
v.s. *L. vide supra:* see above

Symbols

$\forall$: 'for all' or 'for all values of'
$\Sigma_{i=1}^{n} Z_i$: $Z_1 + Z_2 + \ldots + Z_n$
$\Pi_{i=1}^{n} Z_i$: $Z_1 \cdot Z_2 \cdot \ldots \cdot Z_n$
$\in$: 'is an element of'
$\cap$: intersection (of sets)
$\cup$: union (of sets)