Review
Random variables and sampling theory

In the discussion of estimation techniques in this text, much attention is given to the following properties of estimators: unbiasedness, consistency, and efficiency. It is essential that you have a secure understanding of these concepts, and the text assumes that you have taken an introductory statistics course that has treated them in some depth. This chapter offers a brief review.

Discrete random variables

Your intuitive notion of probability is almost certainly perfectly adequate for the purposes of this text, and so we shall skip the traditional section on pure probability theory, fascinating subject though it may be. Many people have direct experience of probability through games of chance and gambling, and their interest in what they are doing results in an amazingly high level of technical competence, usually with no formal training.

We shall begin straight-away with discrete random variables. A random variable is any variable whose value cannot be predicted exactly. A *discrete* random variable is one that has a specific set of possible values. An example is the total score when two dice are thrown. An example of a random variable that is not discrete is the temperature in a room. It can take any one of a continuing range of values and is an example of a *continuous* random variable. We shall come to these later in this review.

Continuing with the example of the two dice, suppose that one of them is green and the other red. When they are thrown, there are 36 possible experimental *outcomes*, since the green one can be any of the numbers from 1 to 6 and the red one likewise. The random variable defined as their sum, which we will denote X, can take only one of 11 *values*—the numbers from 2 to 12. The relationship between the experimental outcomes and the values of this random variable is illustrated in Figure R.1.

Assuming that the dice are fair, we can use Figure R.1 to work out the probability of the occurrence of each value of X. Since there are 36 different combinations of the dice, each outcome has probability 1/36. {Green = 1, red = 1} is the only combination that gives a total of 2, so the probability of $X = 2$ is 1/36. To

red / *green*	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	21

Figure R.1 Outcomes in the example with two dice

Table R.1

Value of X	2	3	4	5	6	7	8	9	10	11	12
Frequency	1	2	3	4	5	6	5	4	3	2	1
Probability	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

obtain $X = 7$, we would need $\{\text{green} = 1, \text{red} = 6\}$ or $\{\text{green} = 2, \text{red} = 5\}$ or $\{\text{green} = 3, \text{red} = 4\}$ or $\{\text{green} = 4, \text{red} = 3\}$ or $\{\text{green} = 5, \text{red} = 2\}$ or {green=6, red=1}. In this case six of the possible outcomes would do, so the probability of throwing 7 is 6/36. All the probabilities are given in Table R.1. If you add all the probabilities together, you get exactly 1. This is because it is 100 percent certain that the value must be one of the numbers from 2 to 12.

The set of all possible values of a random variable is described as the *population* from which it is drawn. In this case, the population is the set of numbers from 2 to 12.

Exercises

R.1 A random variable X is defined to be the difference between the higher value and the lower value when two dice are thrown. If they have the same value, X is defined to be 0. Find the probability distribution for X.

R.2 * A random variable X is defined to be the larger of the two values when two dice are thrown, or the value if the values are the same. Find the probability distribution for X. [*Note*: Answers to exercises marked with an asterisk are provided in the *Student Guide*.]

Expected values of discrete random variables

The expected value of a discrete random variable is the weighted average of all its possible values, taking the probability of each outcome as its weight. You calculate it by multiplying each possible value of the random variable by its

probability and adding. In mathematical terms, if the random variable is denoted X, its expected value is denoted $E(X)$.

Let us suppose that X can take n particular values $x_1, x_2, \ldots, x_n$ and that the probability of x_i is p_i. Then

$$E(X) = x_1 p_1 + \cdots + x_n p_n = \sum_{i=1}^{n} x_i p_i. \tag{R.1}$$

(Appendix R.1 provides an explanation of Σ notation for those who would like to review its use.)

In the case of the two dice, the values x_1 to x_n were the numbers 2 to 12: $x_1 = 2$, $x_2 = 3, \ldots, x_{11} = 12$, and $p_1 = 1/36$, $p_2 = 2/36, \ldots, p_{11} = 1/36$. The easiest and neatest way to calculate an expected value is to use a spreadsheet. The left half of Table R.2 shows the working in abstract. The right half shows the working for the present example. As you can see from the table, the expected value is equal to 7.

Before going any further, let us consider an even simpler example of a random variable, the number obtained when you throw just one die. (*Pedantic note*: This is the singular of the word whose plural is dice. Two dice, one die. Like two mice, one mie.) (Well, two mice, one mouse. Like two hice, one house. Peculiar language, English.)

There are six possible outcomes: $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, x_5 = 5, x_6 = 6$. Each has probability 1/6. Using these data to compute the expected value, you

Table R.2 Expected value of X, example with two dice

X	p	X_p	X	p	X_p
x_1	p_1	$x_1 p_1$	2	1/36	2/36
x_2	p_2	$x_2 p_2$	3	2/36	6/36
x_3	p_3	$x_3 p_3$	4	3/36	12/36
…	…	…	5	4/36	20/36
…	…	…	6	5/36	30/36
…	…	…	7	6/36	42/36
…	…	…	8	5/36	40/36
…	…	…	9	4/36	36/36
…	…	…	10	3/36	30/36
…	…	…	11	2/36	22/36
x_n	p_n	$x_n p_n$	12	1/36	12/36
Total		$E(X) = \sum_{i=1}^{n} x_i p_i$			252/36=7

find that it is equal to 3.5. Thus in this case the expected value of the random variable is a number you could not obtain at all.

The expected value of a random variable is frequently described as its population mean. In the case of a random variable X, the population mean is often denoted by μ_X, or just μ, if there is no ambiguity.

Exercises

R.3 Find the expected value of X in Exercise R.1.
R.4* Find the expected value of X in Exercise R.2.

Expected values of functions of discrete random variables

Let $g(X)$ be any function of X. Then $E[g(X)]$, the expected value of $g(X)$, is given by

$$E[g(X)] = g(x_1)p_1 + \cdots + g(x_n)p_n = \sum_{i=1}^{n} g(x_i)p_i, \tag{R.2}$$

where the summation is taken over all possible values of X.

The left half of Table R.3 illustrates the calculation of the expected value of a function of X. Suppose that X can take the n different values x_1 to x_n, with associated probabilities p_1 to p_n. In the first column, you write down all the values that X can take. In the second, you write down the corresponding probabilities. In the third, you calculate the value of the function for the corresponding value of X. In the fourth, you multiply columns 2 and 3. The answer is given by the total of column 4.

The right half of Table R.3 shows the calculation of the expected value of X^2 for the example with two dice. You might be tempted to think that this is equal to μ^2, but this is not correct. $E(X^2)$ is 54.83. The expected value of X was shown in Table R.2 to be equal to 7. Thus it is not true that $E(X^2)$ is equal to μ^2,which means that you have to be careful to distinguish between $E(X^2)$ and $[E(X)]^2$ (the latter being $E(X)$ multiplied by $E(X)$, that is, μ^2).

Exercises

R.5 If X is a random variable with mean μ, and λ is a constant, prove that the expected value of λX is $\lambda\mu$.
R.6 Calculate $E(X^2)$ for X defined in Exercise R.1.
R.7* Calculate $E(X^2)$ for X defined in Exercise R.2.

Table R.3 Expected value of $g(X)$, example with two dice

Expected value of $g(X)$				Expected value of X^2			
X	p	$g(X)$	$g(X)p$	X	p	X^2	X^2p
x_1	p_1	$g(x_1)$	$g(x_1)p_1$	2	1/36	4	0.11
x_2	p_2	$g(x_2)$	$g(x_2)p_2$	3	2/36	9	0.50
x_3	p_3	$g(x_3)$	$g(x_3)p_3$	4	3/36	16	1.33
...	...	...	...	5	4/36	25	2.78
...	...	...	...	6	5/36	36	5.00
...	...	...	...	7	6/36	49	8.17
...	...	...	...	8	5/36	64	8.89
...	...	...	...	9	4/36	81	9.00
...	...	...	...	10	3/36	100	8.83
...	...	...	...	11	2/36	121	6.72
x_n	p_n	$g(x_n)$	$g(x_n)p_n$	12	1/36	144	4.00
Total			$E[g(X)] = \sum_{i=1}^{n} g(x_i)p_i$				54.83

Expected value rules

There are three rules that we are going to use over and over again. They are virtually self-evident, and they are equally valid for discrete and continuous random variables.

Rule 1 The expected value of the sum of several variables is equal to the sum of their expected values. For example, if you have three random variables X, Y, and Z,

$$E(X + Y + Z) = E(X) + E(Y) + E(Z). \tag{R.3}$$

Rule 2 If you multiply a random variable by a constant, you multiply its expected value by the same constant. If X is a random variable and b is a constant,

$$E(bX) = bE(X). \tag{R.4}$$

Rule 3 The expected value of a constant is that constant. For example, if b is a constant,

$$E(b) = b. \tag{R.5}$$

Rule 2 has already been proved as Exercise R.5. Rule 3 is trivial in that it follows from the definition of a constant. Although the proof of Rule 1 is quite easy, we will omit it here.

Putting the three rules together, you can simplify more complicated expressions. For example, suppose you wish to calculate $E(Y)$, where

$$Y = b_1 + b_2 X. \tag{R.6}$$

and b_1 and b_2 are constants. Then,

$$\begin{aligned} E(Y) &= E(b_1 + b_2 X) \\ &= E(b_1) + E(b_2 X) \quad \text{using Rule 1} \\ &= b_1 + b_2 E(X) \qquad \text{using Rules 2 and 3.} \end{aligned} \tag{R.7}$$

Therefore, instead of calculating $E(Y)$ directly, you could calculate $E(X)$ and obtain $E(Y)$ from equation (R.7).

Exercises

R.8 Let X be the total when two dice are thrown. Calculate the possible values of Y, where Y is given by

$$Y = 2X + 3.$$

and hence calculate $E(Y)$. Show that this is equal to $2E(X) + 3$.

Independence of random variables

Two random variables X and Y are said to be independent if $E[g(X)h(Y)]$ is equal to $E[g(X)]E[h(Y)]$ for any functions $g(X)$ and $h(Y)$. Independence implies, as an important special case, that $E(XY)$ is equal to $E(X)E(Y)$.

Population variance of a discrete random variable

In this text there is only one function of X in which we shall take much interest, and that is its population variance, a useful measure of the dispersion of its probability distribution. It is defined as the expected value of the square of the difference between X and its mean, that is, of $(X-\mu)^2$, where μ is the population mean. It is usually denoted σ_X^2, with the subscript being dropped when it is

Table R.4 Population variance of X, example with two dice

X	P	$X-\mu$	$(X-\mu)^2$	$(X-\mu)^2p$
2	1/36	−5	25	0.69
3	2/36	−4	16	0.89
4	3/36	−3	9	0.75
5	4/36	−2	4	0.44
6	5/36	−1	1	0.14
7	6/36	0	0	0.00
8	5/36	1	1	0.14
9	4/36	2	4	0.44
10	3/36	3	9	0.75
11	2/36	4	16	0.89
12	1/36	5	25	0.69
Total				5.83

obvious that it is referring to a particular variable.

$$\begin{aligned}\sigma_X^2 &= E[(X-\mu)^2] \\ &= (x_1-\mu)^2p_1+\cdots+(x_n-\mu)^2p_n=\sum_{i=1}^{n}(x_i-\mu)^2p_i. \end{aligned} \tag{R.8}$$

From σ_X^2 one obtains σ_X, the population standard deviation, an equally popular measure of the dispersion of the probability distribution; the standard deviation of a random variable is the square root of its variance.

We will illustrate the calculation of population variance with the example of the two dice. Since $\mu = E(X) = 7, (X-\mu)^2$ is $(X-7)^2$ in this case. We shall calculate the expected value of $(X-7)^2$ using Table R.3 as a pattern. An extra column, $(X-\mu)$, has been introduced as a step in the calculation of $(X-\mu)^2$. By summing the last column in Table R.4, one finds that σ_X^2 is equal to 5.83. Hence σ_X, the standard deviation, is equal to $\sqrt{5.83}$, which is 2.41.

One particular use of the expected value rules that is quite important is to show that the population variance of a random variable can be written

$$\sigma_X^2 = E(X^2)-\mu^2, \tag{R.9}$$

an expression that is sometimes more convenient than the original definition. The proof is a good example of the use of the expected value rules. From its

definition,

$$\begin{aligned}\sigma_X^2 &= E[(X-\mu)^2] \\ &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) + E(-2\mu X) + E(\mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2. \end{aligned} \tag{R.10}$$

Thus, if you wish to calculate the population variance of X, you can calculate the expected value of X^2 and subtract μ^2.

Exercises

R.9 Calculate the population variance and the standard deviation of X as defined in Exercise R.1, using the definition given by equation (R.8).

R.10* Calculate the population variance and the standard deviation of X as defined in Exercise R.2, using the definition given by equation (R.8).

R.11 Using equation (R.9), find the variance of the random variable X defined in Exercise R.1 and show that the answer is the same as that obtained in Exercise R.9. (*Note*: You have already calculated μ in Exercise R.3 and $E(X^2)$ in Exercise R.6.)

R.12* Using equation (R.9), find the variance of the random variable X defined in Exercise R.2 and show that the answer is the same as that obtained in Exercise R.10.(*Note*: You have already calculated μ in Exercise R.4 and $E(X^2)$ in Exercise R.7.)

Probability density

Discrete random variables are very easy to handle in that, by definition, they can take only a finite set of values. Each of these values has a 'packet' of probability associated with it, and, if you know the size of these packets, you can calculate the population mean and variance with no trouble. The sum of the probabilities is equal to 1. This is illustrated in Figure R.2 for the example with two dice. X can take values from 2 to 12 and the associated probabilities are as shown.

Unfortunately, the analysis in this text usually deals with continuous random variables, which can take an infinite number of values. The discussion will be illustrated with the example of the temperature in a room. For the sake of argument, we will assume that this varies within the limits of 55 to 75 °F, and initially we will suppose that it is equally likely to be anywhere within this range.

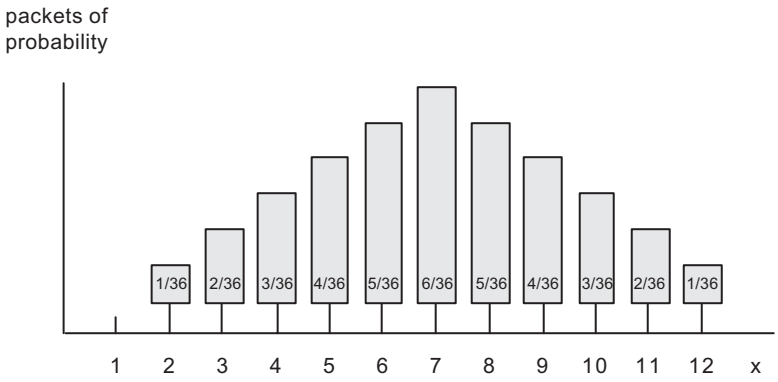

Figure R.2 Discrete probabilities (example with two dice)

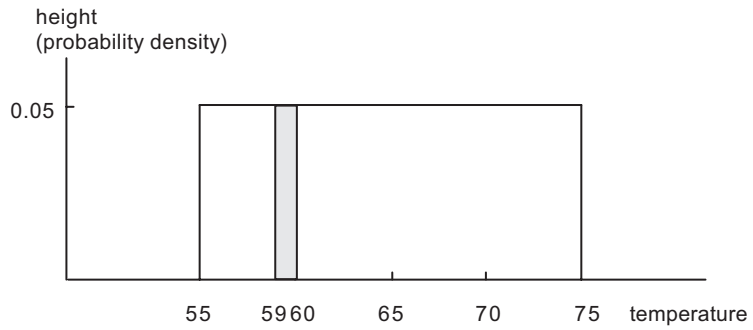

Figure R.3

Since there are an infinite number of different values that the temperature can take, it is useless trying to divide the probability into little packets and we have to adopt a different approach. Instead, we talk about the probability of the random variable lying within a given interval, and we represent the probability graphically as an area within the interval. For example, in the present case, the probability of X lying in the interval 59 to 60 is 0.05 since this range is one twentieth of the complete range 55 to 75. Figure R.3 shows the rectangle depicting the probability of X lying in this interval. Since its area is 0.05 and its base is one, its height must be 0.05. The same is true for all the other one-degree intervals in the range that X can take.

Having found the height at all points in the range, we can answer such questions as 'What is the probability that the temperature lies between 65 and 70 °F?'. The answer is given by the area in the interval 65 to 70, represented by the shaded area in Figure R.4. The base of the shaded area is 5, and its height is 0.05, so the area is 0.25. The probability is a quarter, which is obvious anyway in that 65 to 70 °F is a quarter of the whole range.

The height at any point is formally described as the probability density at that point, and, if it can be written as a function of the random variable, it is known

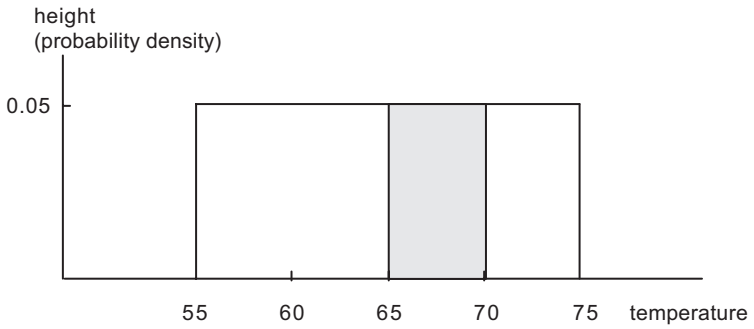

Figure R.4

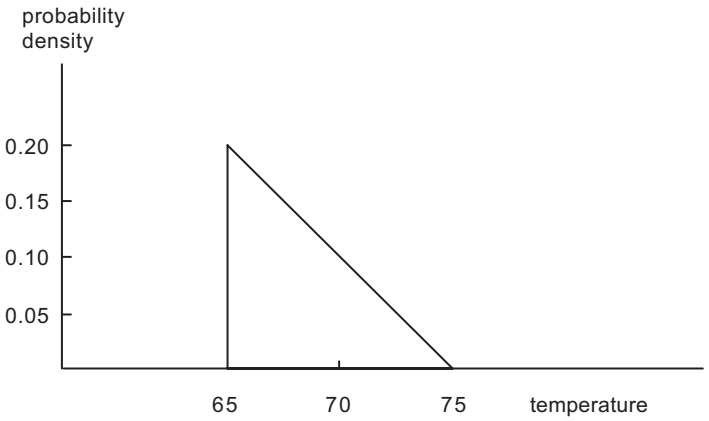

Figure R.5

as the 'probability density function'. In this case it is given by $f(x)$, where X is the temperature and

$$\begin{aligned} f(x) &= 0.05 \quad \text{for } 55 \leq x \leq 75 \\ f(x) &= 0 \qquad \text{for } x < 55 \text{ or } x > 75. \end{aligned} \tag{R.11}$$

The foregoing example was particularly simple to handle because the probability density function was constant over the range of possible values of X. Next we will consider an example in which the function is not constant, because not all temperatures are equally likely. We will suppose that the central heating and air conditioning have been fixed so that the temperature never falls below 65 °F, and that on hot days the temperature will exceed this, with a maximum of 75 °F as before. We will suppose that the probability is greatest at 65 °F and that it decreases evenly to 0 at 75 °F, as shown in Figure R.5.

The total area within the range, as always, is equal to 1, because the total probability is equal to 1. The area of the triangle is 1/2 × base × height, so

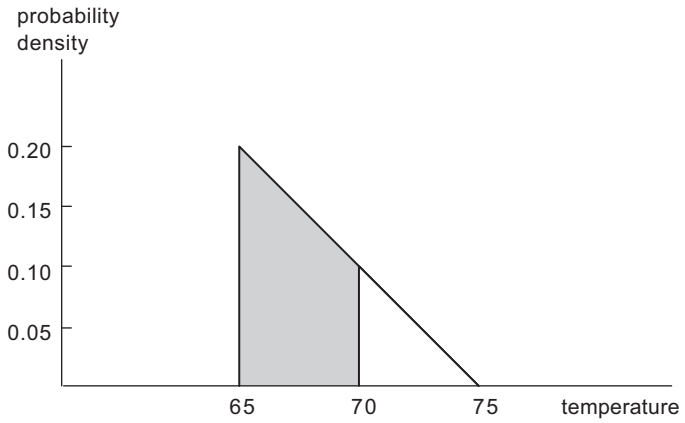

Figure R.6

one has

$$1/2 \times 10 \times \text{height} = 1 \tag{R.12}$$

and the height at 65 °F is equal to 0.20.

Suppose again that we want to know the probability of the temperature lying between 65 and 70 °F. It is given by the shaded area in Figure R.6, and with a little geometry you should be able to verify that it is equal to 0.75. If you prefer to talk in terms of percentages, this means that there is a 75 percent chance that the temperature will lie between 65 and 70 °F, and only a 25 percent chance that it will lie between 70 and 75 °F.

In this case the probability density function is given by $f(x)$, where

$$\begin{aligned} f(x) &= 1.5 - 0.02x \quad \text{for } 65 \leq x \leq 75 \\ f(x) &= 0 \qquad\qquad\quad \text{for } x < 65 \text{ or } x > 75. \end{aligned} \tag{R.13}$$

(You can verify that $f(x)$ gives 0.20 at 65 °F and 0 at 75 °F.)

Now for some good news and some bad news. First, the bad news. If you want to calculate probabilities for more complicated, curved functions, simple geometry will not do. In general you have to use integral calculus or refer to specialized tables, if they exist. Integral calculus is also used in the definitions of the expected value and variance of a continuous random variable.

Now for the good news. First, specialized probability tables do exist for all the functions that are going to interest us in practice. Second, expected values and variances have much the same meaning for continuous random variables that they have for discrete ones (formal definitions will be found in Appendix R.2), and the expected value rules work in exactly the same way.

Fixed and random components of a random variable

Instead of regarding a random variable as a single entity, it is often possible and convenient to break it down into a fixed component and a pure random component, the fixed component always being the population mean. If X is a random variable and μ its population mean, one may make the following decomposition:

$$X = \mu + u, \tag{R.14}$$

where u is what will be called the pure random component (in the context of regression analysis, it is usually described as the disturbance term).

You could of course look at it the other way and say that the random component, u, is defined to be the difference between X and μ:

$$u = X - \mu. \tag{R.15}$$

It follows from its definition that the expected value of u is 0. From equation (R.15),

$$E(u_i) = E(x_i - \mu) = E(x_i) + E(-\mu) = \mu - \mu = 0. \tag{R.16}$$

Since all the variation in X is due to u, it is not surprising that the population variance of X is equal to the population variance of u. This is easy to prove. By definition,

$$\sigma_x^2 = E[(X - \mu)^2] = E(u^2) \tag{R.17}$$

and

$$\begin{aligned} \sigma_u^2 &= E[(u - \text{mean of } u)^2] \\ &= E[(u - 0^2] = E(u^2). \end{aligned} \tag{R.18}$$

Hence σ^2 can equivalently be defined to be the variance of X or u.

To summarize, if X is a random variable defined by (R.14), where μ is a fixed number and u is a random component, with mean 0 and population variance σ^2, then X has population mean μ and population variance σ^2.

Estimators

So far we have assumed that we have exact information about the random variable under discussion, in particular that we know the probability distribution, in the case of a discrete random variable, or the probability density function, in the case of a continuous variable. With this information it is possible to work out the population mean and variance and any other population characteristics in which we might be interested.

Table R.5

Population characteristic		Estimator
Mean	μ	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Population variance	σ^2	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{X})^2$

Now, in practice, except for artificially simple random variables such as the numbers on thrown dice, you do not know the exact probability distribution or density function. It follows that you do not know the population mean or variance. However, you would like to obtain an estimate of them or some other population characteristic.

The procedure is always the same. You take a sample of n observations and derive an estimate of the population characteristic using some appropriate formula. You should be careful to make the important distinction that the formula is technically known as an *estimator*; the number that is calculated from the sample using it is known as the *estimate*. The estimator is a general rule or formula, whereas the estimate is a specific number that will vary from sample to sample.

Table R.5 gives the usual estimators for the two most important population characteristics. The sample mean, $\bar{X}$, is the usual estimator of the population mean, and the formula for s^2 given in Table R.5 is the usual estimator of the population variance.

Note that these are the *usual* estimators of the population mean and variance; they are not the only ones. You are probably so accustomed to using the sample mean as an estimator of μ that you are not aware of any alternatives. Of course, not all the estimators you can think of are equally good. The reason that we do in fact use $\bar{X}$ is that it is the best according to two very important criteria, unbiasedness and efficiency. These criteria will be discussed later.

Estimators are random variables

An estimator is a special case of a random variable. This is because it is a combination of the values of X in a sample, and, since X is a random variable, a combination of a set of its values must also be a random variable. For instance, take $\bar{X}$, the estimator of the mean:

$$\bar{X} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{R.19}$$

We have just seen that the value of X in observation i may be decomposed into two parts: the fixed part, μ, and the pure random component, u_i:

$$x_i = \mu + u_i. \tag{R.20}$$

Hence

$$\begin{aligned}\bar{X} &= \frac{1}{n}(\mu + \mu + \cdots + \mu) + \frac{i}{n}(u_1 + u_2 + \cdots + u_n) \\ &= \frac{1}{n}(n\mu) + \bar{u} = \mu + \bar{u},\end{aligned} \tag{R.21}$$

where $\bar{u}$ is the average of u_i in the sample.

From this you can see that $\bar{X}$, like X, has both a fixed component and a pure random component. Its fixed component is μ, the population mean of X, and its pure random component is $\bar{u}$, the average of the pure random components in the sample.

The probability density functions of both X and $\bar{X}$ have been drawn in the same diagram in Figure R.7. By way of illustration, X is assumed to have a normal distribution. You will see that the distributions of both X and $\bar{X}$ are centered over μ, the population mean. The difference between them is that the distribution for $\bar{X}$ is narrower and taller. $\bar{X}$ is likely to be closer to μ than a single observation on X, because its random component $\bar{\mu}$ is an average of the pure random components $u_1, u_2, \ldots, u_n$ in the sample, and these are likely to cancel each other out to some extent when the average is taken. Consequently the population variance of $\bar{u}$ is only a fraction of the population variance of u. It will be shown in Section 1.7 that, if the population variance of u is σ^2, then the population variance of $\bar{u}$ is σ^2/n.

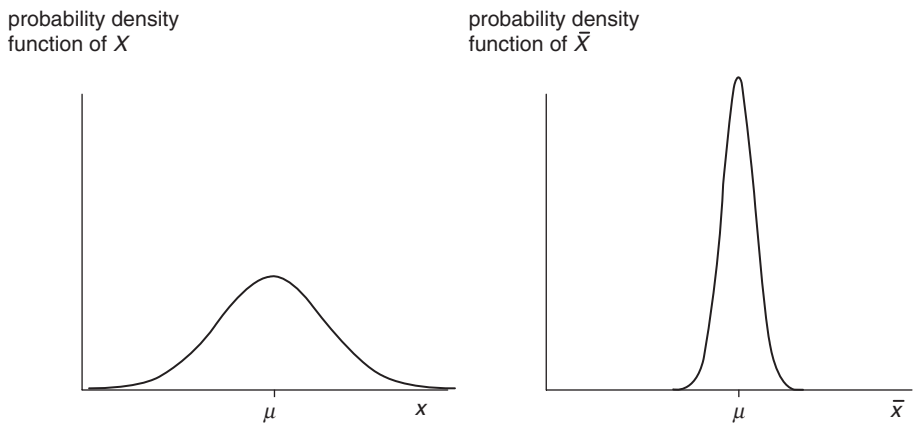

Figure R.7 Comparison of the probability density functions of a single observation and the mean of a sample

s^2, the unbiased estimator of the population variance of X, is also a random variable. Subtracting (R.21) from (R.20),

$$x_i - \bar{X} = u_i - \bar{u}. \tag{R.22}$$

Hence

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}[(x_i - \bar{X})^2] = \frac{1}{n-1}\sum_{i=1}^{n}[(u_i - \bar{u})^2]. \tag{R.23}$$

Thus s^2 depends on (and only on) the pure random components of the observations on X in the sample. Since these change from sample to sample, the value of the estimator s^2 will change from sample to sample.

Unbiasedness

Since estimators are random variables, it follows that only by coincidence will an estimate be exactly equal to the population characteristic. Generally there will be some degree of error, which will be small or large, positive or negative, according to the pure random components of the values of X in the sample.

Although this must be accepted, it is nevertheless desirable that the estimator should be accurate on average in the long run, to put it intuitively. To put it technically, we should like the expected value of the estimator to be equal to the population characteristic. If this is true, the estimator is said to be *unbiased*. If it is not, the estimator is said to be *biased*, and the difference between its expected value and the population characteristic is described as the *bias*.

Let us start with the sample mean. Is this an unbiased estimator of the population mean? Is $E(\bar{X})$ equal to μ? Yes, it is, and it follows immediately from (R.21).

$\bar{X}$ has two components, μ and $\bar{u}$. $\bar{u}$ is the average of the pure random components of the values of X in the sample, and since the expected value of the pure random component in any observation is 0, the expected value of $\bar{u}$ is 0. Hence

$$E(\bar{X}) = E(\mu + \bar{u}) = E(\mu) + E(\bar{u}) = \mu + 0 = \mu. \tag{R.24}$$

However, this is not the only unbiased estimator of μ that we could construct. To keep the analysis simple, suppose that we have a sample of just two observations, x_1 and x_2. Any weighted average of the observations x_1 and x_2 will be an unbiased estimator, provided that the weights add up to 1. To see this, suppose we construct a generalized estimator:

$$Z = \lambda_1 x_1 + \lambda_2 x_2. \tag{R.25}$$

The expected value of Z is given by

$$\begin{aligned} E(Z) &= E(\lambda_1 x_1 + \lambda_2 x_2) = E(\lambda_1 x_1 + E(\lambda_2 x_2) \\ &= \lambda_1 E(x_1) + \lambda_2 E(x_2) = \lambda_1 \mu + \lambda_2 \mu \\ &= (\lambda_1 + \lambda_2)\mu. \end{aligned} \tag{R.26}$$

If λ_1 and λ_2 add up to 1, we have $E(Z) = \mu$, and Z is an unbiased estimator of μ.

Thus, in principle, we have an infinite number of unbiased estimators. How do we choose among them? Why do we always in fact use the sample average, with $\lambda_1 = \lambda_2 = 0.5$? Perhaps you think that it would be unfair to give the observations different weights, or that asymmetry should be avoided on principle. However, we are not concerned with fairness, or with symmetry for its own sake. We will find in the next section that there is a more compelling reason.

So far we have been discussing only estimators of the population mean. It was asserted that s^2, as defined in Table R.5, is an estimator of the population variance, σ^2. One may show that the expected value of s^2 is σ^2, and hence that it is an unbiased estimator of the population variance, provided that the observations in the sample are generated independently of each another. The proof, though not mathematically difficult, is laborious, and it has been consigned to Appendix R.3 at the end of this review.

Efficiency

Unbiasedness is one desirable feature of an estimator, but it is not the only one. Another important consideration is its reliability. It is all very well for an estimator to be accurate on average in the long run, but, as Keynes once said, in the long run we are all dead. We want the estimator to have as high a probability as possible of giving a close estimate of the population characteristic, which means that we want its probability density function to be as concentrated as possible around the true value. One way of summarizing this is to say that we want its population variance to be as small as possible.

Suppose that we have two estimators of the population mean, that they are calculated using the same information, that they are both unbiased, and that their probability density functions are as shown in Figure R.8. Since the probability density function for estimator B is more highly concentrated than that for estimator A, it is more likely to give an accurate estimate. It is therefore said to be more *efficient*, to use the technical term.

Note carefully that the definition says 'more likely'. Even though estimator B is more efficient, that does not mean that it will always give the more accurate estimate. Some times it will have a bad day, and estimator A will have a good day, and A will be closer to the truth. But the probability of A being more accurate than B will be less than 50 percent.

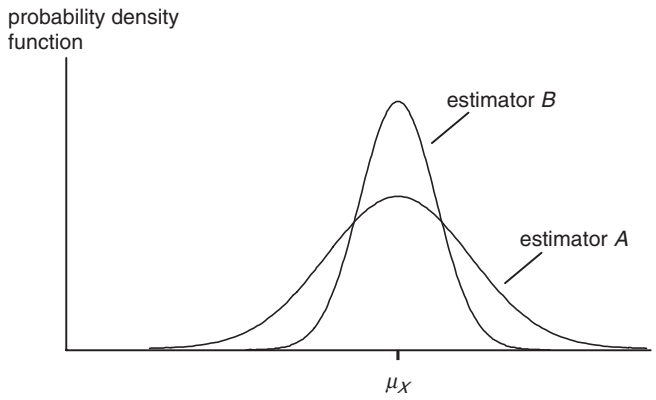

Figure R.8 Efficient and inefficient estimators

It is rather like the issue of whether you should fasten your seat belt when driving a vehicle. A large number of surveys in different countries have shown that you are much less likely to be killed or seriously injured in a road accident if you wear a seat belt, but there are always the odd occasions when individuals not wearing belts have miraculously escaped when they could have been killed, had they been strapped in. The surveys do not deny this. They simply conclude that the odds are on the side of belting up. Similarly, the odds are on the side of the efficient estimator. (*Gruesome comment*: In countries where wearing seat belts has been made compulsory, there has been a fall in the supply of organs from crash victims for transplants.)

We have said that we want the variance of an estimator to be as small as possible, and that the efficient estimator is the one with the smallest variance. We shall now investigate the variance of the generalized estimator of the population mean and show that it is minimized when the two observations are given equal weight.

Provided that x_1 and x_2 are independent observations, the population variance of the generalized estimator is given by

$$\begin{aligned}\sigma_Z^2 &= \text{population variance of } (\lambda_1 x_1 + \lambda_2 x_2)\\ &= \sigma^2_{\lambda_1 x_1} + \sigma^2_{\lambda_2 x_2} + 2\sigma_{\lambda_1 x_1, \lambda_2 x_2}\\ &= \lambda_1^2\sigma^2_{x_1} + \lambda_2^2\sigma^2_{x_2} + 2\lambda_1\lambda_2\sigma_{x_1 x_2} \quad \text{if } x_1 \text{ and } x_2 \text{ are independent}\\ &= (\lambda_1^2 + \lambda_2^2)\sigma^2. \end{aligned} \tag{R.27}$$

(We are anticipating the variance rules discussed in Chapter 1. $\sigma_{x_1 x_2}$, the population covariance of x_1 and x_2, is 0 if x_1 and x_2 are generated independently.)

Now, we have already seen that λ_1 and λ_2 must add up to 1 if the estimator is to be unbiased. Hence for unbiased estimators λ_2 equals $(1 - \lambda_1)$ and

$$\lambda_1^2 + \lambda_2^2 = \lambda_1^2 + (1 - \lambda_1)^2 = 2\lambda_1^2 - 2\lambda_1 + 1. \tag{R.28}$$

Since we want to choose λ_1 in such a way that the variance is minimized, we want to choose it to minimize $(2\lambda_1^2 - 2\lambda_1 + 1)$. You could solve this problem graphically or by using the differential calculus. In either case, the minimum value is reached when λ_1 is equal to 0.5. Hence λ_2 is also equal to 0.5.

We have thus shown that the sample average has the smallest variance of estimators of this kind. This means that it has the most concentrated probability distribution around the true mean, and hence that (in a probabilistic sense) it is the most accurate. To use the correct terminology, of the set of unbiased estimators, it is the most efficient. Of course we have shown this only for the case where the sample consists of just two observations, but the conclusions are valid for samples of any size, provided that the observations are independent of one another.

Two final points. First, efficiency is a *comparative* concept. You should use the term only when comparing alternative estimators. You should not use it to summarize changes in the variance of a single estimator. In particular, as we shall see in the next section, the variance of an estimator generally decreases as the sample size increases, but it would be wrong to say that the estimator is becoming more efficient. You must reserve the term for comparisons of *different* estimators. Second, you can compare the efficiency of alternative estimators only if they are using the same information, for example, the same set of observations on a number of random variables. If the estimators use different information, one may well have a smaller variance, but it would not be correct to describe it as being more efficient.

Exercises

R.13 For the special case $\sigma^2 = 1$ and a sample of two observations, calculate the variance of the generalized estimator of the population mean using equation (28) with values of λ_1 from 0 to 1 at steps of 0.1, and plot it in a diagram. Is it important that the weights λ_1 and λ_2 should be exactly equal?

R.14 Show that, when you have n observations, the condition that the generalized estimator $(\lambda_1 x_1 + \cdots + \lambda_n x_n)$ should be an unbiased estimator of μ is $\lambda_1 + \cdots + \lambda_n = 1$.

Conflicts between unbiasedness and minimum variance

We have seen in this review that it is desirable that an estimator be unbiased and that it have the smallest possible variance. These are two quite different criteria and occasionally they conflict with each other. It sometimes happens that one can

construct two estimators of a population characteristic, one of which is unbiased (*A* in Figure R.9), the other being biased but having smaller variance (*B*).

A will be better in the sense that it is unbiased, but *B* is better in the sense that its estimates are always close to the true value. How do you choose between them?

It will depend on the circumstances. If you are not bothered by errors, provided that in the long run they cancel out, you should probably choose *A*. On the other hand, if you can tolerate small errors, but not large ones, you should choose *B*.

Technically speaking, it depends on your loss function, the cost to you of an error as a function of its size. It is usual to choose the estimator that yields the smallest expected loss, which is found by weighting the loss function by the probability density function. (If you are risk-averse, you may wish to take the variance of the loss into account as well.)

A common example of a loss function, illustrated by the quadratic curve in Figure R.10, is the square of the error. The expected value of this, known as the mean square error (MSE), has the simple decomposition:

$$\text{MSE of estimator} = \text{Variance of estimator} + \text{Bias}^2. \tag{R.29}$$

To show this, suppose that you are using an estimator Z to estimate an unknown population parameter θ. Let the expected value of Z be μ_Z. This will be equal to θ only if Z is an unbiased estimator. In general there will be a

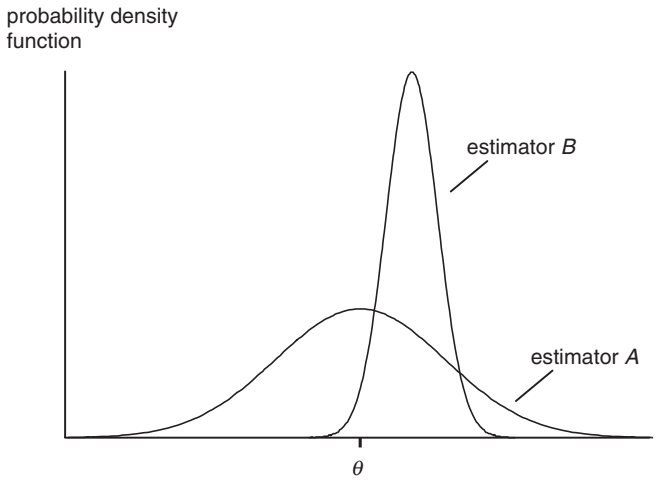

Figure R.9 Which estimator is to be preferred? *A* is unbiased but *B* has smaller variance

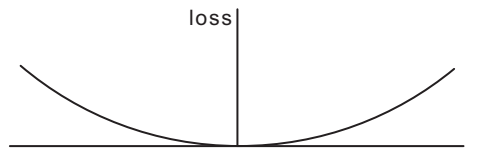

Figure R.10 Loss function

bias, given by $(\mu_Z - \theta)$. The variance of Z is equal to $E[(Z - \mu_Z)^2]$. The MSE of Z can be decomposed as follows:

$$\begin{aligned} E[(Z-\theta)^2] &= E[(\{Z-\mu_Z\}+\{\mu_Z-\theta\})^2] \\ &= E[(Z-\mu_Z)^2 + 2(Z-\mu_Z)(\mu_Z-\theta)+(\mu_Z-\theta)^2] \\ &= E[(Z-\mu_Z)^2] + 2(\mu_Z-\theta)E(Z-\mu_Z) + E[(\mu_Z-\theta)^2]. \end{aligned} \quad \text{(R.30)}$$

The first term is the population variance of Z. The second term is 0 because

$$E(Z-\mu_Z) = E(Z) + E(-\mu_Z) = \mu_Z - \mu_Z = 0. \quad \text{(R.31)}$$

The expected value of the third term is $(\mu_Z - \theta)^2$, the bias squared, since both μ_Z and θ are constants. Hence we have shown that the mean square error of the estimator is equal to the sum of its population variance and the square of its bias.

In Figure R.9, estimator A has no bias component, but it has a much larger variance component than B and therefore could be inferior by this criterion.

The MSE is often used to generalize the concept of efficiency to cover comparisons of biased as well as unbiased estimators. However, in this text, comparisons of efficiency will mostly be confined to unbiased estimators.

Exercises

R.15 Give examples of applications where you might (1) prefer an estimator of type A, (2) prefer one of type B, in Figure R.9.

R.16 Draw a loss function for getting to an airport later (or earlier) than the official check-in time.

R.17 If you have two estimators of an unknown population parameter, is the one with the smaller variance necessarily more efficient?

The effect of increasing the sample size on the accuracy of an estimate

We shall continue to assume that we are investigating a random variable X with unknown mean μ and population variance σ^2, and that we are using $\bar{X}$ to estimate μ. How does the accuracy of $\bar{X}$ depend on the number of observations, n?

Not surprisingly, the answer is that, as you increase n, the more accurate $\bar{X}$ is likely to be. In any single experiment, a bigger sample will not necessarily yield a more accurate estimate than a smaller one—the luck factor is always at work—but as a general tendency it should. Since the population variance of $\bar{X}$ is given by σ^2/n, the bigger the sample, the smaller the variance and hence the more tightly compressed is the probability density function of $\bar{X}$.

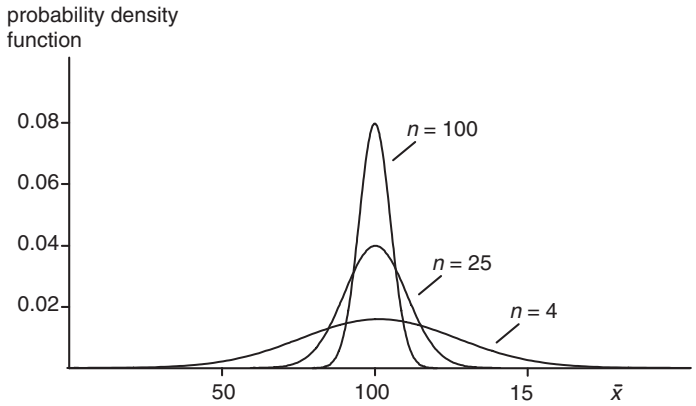

Figure R.11 Effect of increasing the sample size on the distribution of $\bar{X}$

This is illustrated in Figure R.11. We are assuming that X is normally distributed and that it has mean 100 and standard deviation 50. If the sample size is 4, the standard deviation of $\bar{X}, \sigma/\sqrt{n}$, is equal to $50/\sqrt{4} = 25$. If the sample size is 25, the standard deviation is 10. If it is 100, the standard deviation is 5. Figure R.11 shows the corresponding probability density functions. That for $n = 100$ is taller than the others in the vicinity of μ, showing that the probability of it giving an accurate estimate is higher. It is lower elsewhere.

The larger the sample size, the narrower and taller will be the probability density function of $\bar{X}$. If n becomes really large, the probability density function will be indistinguishable from a vertical line located at $\bar{X} = \mu$. For such a sample the random component of $\bar{X}$ becomes very small indeed, and so $\bar{X}$ is bound to be very close to μ. This follows from the fact that the standard deviation of $\bar{X}, \sigma/\sqrt{n}$, becomes very small as n becomes large.

In the limit, as n tends to infinity, $\sigma/\sqrt{n}$ tends to 0 and $\bar{X}$ tends to μ exactly. This may be written mathematically

$$\lim_{n\to\infty} \bar{X} = \mu \tag{R.32}$$

An equivalent and more common way of expressing it is to use the term *plim*, where plim means 'probability limit' and emphasizes that the limit is being reached in a probabilistic sense:

$$\text{plim}\, \bar{X} = \mu \tag{R.33}$$

when, for any arbitrarily small numbers ε and δ, the probability of $\bar{X}$ being more than ε different from μ is less than δ, provided that the sample is large enough.

Exercise

R.18* In general, the variance of the distribution of an estimator decreases when the sample size is increased. Is it correct to describe the estimator as becoming more efficient?

Consistency

In general, if the plim of an estimator is equal to the true value of the population characteristic, it is said to be *consistent*. To put it another way, a consistent estimator is one that is bound to give an accurate estimate of the population characteristic if the sample is large enough, regardless of the actual observations in the sample. In most of the contexts considered in this text, an unbiased estimator will also be a consistent one.

It sometimes happens that an estimator that is biased for small samples may be consistent (it is even possible for an estimator that does not have a finite expected value for small samples to be consistent). Figure R.12 illustrates how the probability distribution might look for different sample sizes. The distribution is said to be asymptotically (meaning, in large samples) unbiased because it becomes centered on the true value as the sample size becomes large. It is said to be consistent because it finally collapses to a single point, the true value.

An estimator is described as *inconsistent* either if its distribution fails to collapse as the sample size becomes large or if the distribution collapses at a point other than the true value.

As we shall see later in this text, estimators of the type shown in Figure R.12 are quite important in regression analysis. Sometimes it is impossible to find an estimator that is unbiased for small samples. If you can find one that is at least consistent, that may be better than having no estimate at all, especially if you are able to assess the direction of the bias in small samples. However, it should be borne in mind that a consistent estimator could in principle perform worse (for example, have a larger mean square error) than an inconsistent one in small samples, so you must be on your guard. In the same way that you might prefer a biased estimator to an unbiased one if its variance is smaller, you might prefer

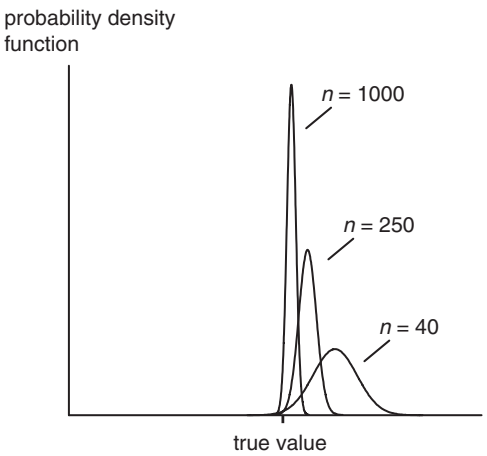

Figure R.12 Estimator that is consistent despite being biased in finite samples

a consistent, but biased, estimator to an unbiased one if its variance is smaller, and an inconsistent one to either if its variance is smaller still.

Two useful rules

Sometimes one has an estimator calculated as the ratio of two quantities that have random components, for example

$$Z = X/Y, \tag{R.34}$$

where X and Y are quantities that have been calculated from a sample. Usually it is difficult to analyze the expected value of Z. In general it is *not* equal to $E(X)$ divided by $E(Y)$. If there is any finite probability that Y may be equal to 0, the expected value of Z will not even be defined. However, if X and Y tend to finite quantities plim X and plim Y in large samples, and plim Y is not equal to 0, the limiting value of Z is given by

$$\text{plim}\, Z = \frac{\text{plim}\, X}{\text{plim}\, Y}. \tag{R.35}$$

Hence, even if we are not in a position to say anything definite about the small sample properties of Z, we may be able to tell whether it is consistent.

For example, suppose that the population means of two random variables X and Y are μ_X and μ_Y, respectively, and that both are subject to random influences, so that

$$X = \mu_X + u_X \tag{R.36}$$

$$Y = \mu_Y + u_Y, \tag{R.37}$$

where u_X and u_Y are random components with 0 means. If we are trying to estimate, the ratio μ_X/μ_Y from sample data, the estimator $Z = \bar{X}/\bar{Y}$ will be consistent, for

$$\text{plim} Z = \frac{\text{plim}\, \bar{X}}{\text{plim}\, \bar{Y}} = \frac{\mu_X}{\mu_Y} \tag{R.38}$$

and we are able to say that Z will be an accurate estimator for large samples, even though we may not be able to say anything about $E(Z)$ for small samples.

There is a counterpart rule for the product of two random variables. Suppose

$$Z = XY. \tag{R.39}$$

Except in the special case where X and Y are distributed independently, it is not true that $E(Z)$ is equal to the product of $E(X)$ and $E(Y)$. However, even if X and

Y are not distributed independently, it is true that

$$\text{plim}Z = \text{plim}X \times \text{plim}Y, \tag{R.40}$$

provided that plim X and plim Y exist.

Exercises

R.19 Is unbiasedness either a necessary or a sufficient condition for consistency?

R.20 A random variable X can take the values 1 and 2 with equal probability. For n equal to 2, demonstrate that $E(\bar{Y})$ is not equal to $1/E(\bar{X})$.

R.21* Repeat Exercise 20 supposing that X takes the values 0 and 1 with equal probability.

Appendix R.1

Σ notation: A review

Σ notation provides a quick way of writing the sum of a series of similar terms. Anyone reading this text ought to be familiar with it, but here is a brief review for those who need a reminder.

We will begin with an example. Suppose that the output of a sawmill, measured in tons, in month i is q_i, with q_1 being the gross output in January, q_2 being the gross output in February, etc. Let output for the year be denoted Z. Then

$$Z = q_1 + q_2 + q_3 + q_4 + q_5 + q_6 + q_7 + q_8 + q_9 + q_{10} + q_{11} + q_{12}.$$

In words, one might summarize this by saying that Z is the sum of the q_i, beginning with q_1 and ending with q_{12}. Obviously there is no need to write down all 12 terms when defining Z. Sometimes you will see it simplified to

$$Z = q_1 + \cdots + q_{12},$$

it being understood that the missing terms are included in the summation.

Σ notation allows you to write down this summary in a tidy symbolic form:

$$Z = \sum_{i=1}^{12} q_i.$$

The expression to the right of the Σ sign tell us what kind of term is going to be summed, in this case, terms of type q_i. Underneath the Σ sign is written the subscript that is going to alter in the summation, in this case i, and its starting point, in this case 1. Hence we know that the first term will be q_1. The $=$ sign reinforces the fact that i should be set equal to 1 for the first term.

Above the Σ sign is written the last value of i, in this case 12, so we know that the last term is q_{12}. It is automatically understood that all the terms between q_1 and q_{12} will also be included in the summation, and so we have effectively rewritten the second definition of Z.

Suppose that the average price per ton of the output of the mill in month i is p_i. The value of output in month i will be $p_i q_i$, and the total value during the year will be V, where V is given by

$$V = p_1 q_1 + \cdots + p_{12} q_{12}.$$

We are now summing terms of type $p_i q_i$ with the subscript i running from 1 to 12, and using Σ notation this may be written as

$$V = \sum_{i=1}^{12} p_i q_i.$$

If c_i is the total cost of operating the mill in month i, profit in month i will be $(p_i q_i - c_i)$, and hence the total profit over the year, P, will be given by

$$P = (p_1 q_1 - c_1) + \cdots + (p_{12} q_{12} - c_{12}),$$

which may be summarized as

$$P = \sum_{i=1}^{12} (p_i q_i - c_i).$$

Note that the profit expression could also have been written as total revenue minus total costs:

$$P = (p_1 q_1 + \cdots + p_{12} q_{12}) - (c_1 + \cdots + c_{12}),$$

and this can be summarized in Σ notation as

$$P = \sum_{i=1}^{12} p_i q_i - \sum_{i=1}^{12} c_i.$$

If the price of output is constant during the year at level p, the expression for the value of annual output can be simplified:

$$\begin{aligned} V &= pq_1 + \cdots + pq_{12} = p(q_1 + \cdots + q_{12}) \\ &= p \sum_{i=1}^{12} q_i. \end{aligned}$$

Hence

$$\sum_{i=1}^{12} pq_i = p \sum_{i=1}^{12} q_i.$$

If the output in each month is constant at level q, the expression for annual output can also be simplified:

$$Z = q_1 + \cdots + q_{12} = q + \cdots + q = 12q.$$

Hence, in this case,

$$\sum_{i=1}^{12} q_i = 12q.$$

We have illustrated three rules, which can be stated formally:

Σ *Rule* 1 (illustrated by the decomposition of profit into total revenue minus total cost)

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$$

Σ *Rule* 2 (illustrated by the expression for V when the price was constant)

$$\sum_{i=1}^{n} ax_i = a \sum_{i=1}^{n} x_i \quad \text{(if } a \text{ is a constant).}$$

Σ *Rule* 3 (illustrated by the expression for Z when quantity was constant)

$$\sum_{i=1}^{n} a = na \quad \text{(if } a \text{ is a constant).}$$

Often it is obvious from the context what are the initial and final values of the summation. In such cases $\sum_{i=1}^{n} x_i$ is often simplified to $\sum x_i$. Furthermore, it is often equally obvious what subscript is being changed, and the expression is simplified to just $\sum x$.

Appendix R.2

Expected value and variance of a continuous random variable

The definition of the expected value of a continuous random variable is very similar to that for a discrete random variable:

$$E(X) = \int xf(x)\,dx,$$

where $f(x)$ is the probability density function of X, with the integration being performed over the interval for which $f(x)$ is defined.

In both cases the different possible values of X are weighted by the probability attached to them. In the case of the discrete random variable, the summation

Discrete	Continuous
$E(X) = \sum_{i=1}^{n} x_i p_i$	$E(X) = \int x f(x) dx$
(Summation overall all possible values)	(Integration over the range for which $f(x)$ is defined)

is done on a packet-by-packet basis over all the possible values of X. In the continuous case, it is of course done on a continuous basis, integrating replacing summation, and the probability density function $f(x)$ replacing the packets of probability p_i. However, the principle is the same.

In the section on discrete random variables, it was shown how to calculate the expected value of a function of X, $g(X)$. You make a list of all the different values that $g(X)$ can take, weight each of them by the corresponding probability, and sum.

The process is exactly the same for a continuous random variable, except that it is done on a continuous basis, which means summation by integration instead of Σ summation. In the case of the discrete random variable, $E[g(X)]$ is equal to $\sum_{i=1}^{n} g(x_i)p_i$ with the summation taken over all possible values of X. In the continuous case, it is defined by

$$E[g(X)] = \int g(x)f(x)\,dx,$$

with the integration taken over the whole range for which $f(x)$ is defined.

As in the case of discrete random variables, there is only one function in which we have an interest, the population variance, defined as the expected value of $(X - \mu^2)$, where $\mu = E(X)$ is the population mean. To calculate the variance, you have to sum $(X - \mu)^2$, weighted by the appropriate probability, over all the possible values of X. In the case of a continuous random variable, this means that you have to evaluate

$$\sigma_X^2 = E[(X - \mu)^2] = \int (x - \mu)^2 f(x)\,dx.$$

It is instructive to compare this with equation (R.8), the parallel expression for a discrete random variable:

$$\sigma_X^2 = E[(X - \mu)^2] = \sum_{i=1}^{n} (x_i - \mu)^2 p_i.$$

As before, when you have evaluated the population variance, you can calculate the population standard deviation, σ, by taking its square root.

Appendix R.3

Proof that s^2 is an unbiased estimator of the population variance

It was asserted in Table R.5 that an unbiased estimator of σ^2 is given by s^2, where

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

We will begin the proof by rewriting $(x_i - \bar{X})^2$ in a more complicated, but helpful, way:

$$\begin{aligned}(x_i - \bar{X})^2 &= [(x_i - \mu) - (\bar{X} - \mu)]^2 \quad \text{(the } \mu \text{ terms cancel if you expand)} \\ &= (x_i - \mu)^2 - 2(x_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2.\end{aligned}$$

Hence

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

The first term is the sum of the first terms of the previous equation using Σ notation. Similarly the second term is the sum of the second terms of the previous equation using Σ notation and the fact that $(\bar{X} - \mu)$ is a common factor. When we come to sum the third terms of the previous equation they are all equal to $(\bar{X} - \mu)^2$, so their sum is simply $n(\bar{X} - \mu)^2$, with no need for Σ notation.

The second component may be rewritten $-2n(\bar{X} - \mu)^2$ since

$$\sum_{i=1}^{n}(x_i - \mu) = \sum_{i=1}^{n}x_i - n\mu = n\bar{X} - n\mu = n(\bar{X} - \mu),$$

and we have

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{X})^2 &= \sum_{i=1}^{n}(x_i - \mu)^2 - 2n(\bar{X} - \mu)^2 + n(\bar{X} - \mu)^2 \\ &= \sum_{i=1}^{n}(x_i - \mu)^2 - n(\bar{X} - \mu)^2.\end{aligned}$$

Applying expectations to this equation, we have

$$\begin{aligned} E\left[\sum_{i=1}^{n}(x_i-\bar{X})^2\right] &= E\left[\sum_{i=1}^{n}(x_i-\mu)^2\right] - nE[(\bar{X}-\mu)^2] \\ &= E[(x_1-\mu)^2]+\cdots+E[(x_n-\mu)^2] - nE[(\bar{X}-\mu)^2] \\ &= n\sigma^2 - n\sigma_{\bar{X}}^2 \\ &= n\sigma^2 - n(\sigma^2/n) = (n-1)\sigma^2. \end{aligned}$$

using the fact that the population variance of $\bar{X}$ is equal to σ^2/n. This is proved in Section 1.7. Hence

$$\begin{aligned} E(s^2) &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{X})^2\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(x_i-\bar{X})^2\right] \\ &= \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2. \end{aligned}$$

Thus s^2 is an unbiased estimator of σ^2.