Review: Random Variables, Sampling, and Estimation

R.1 Introduction

A course in basic mathematical statistics is a non-negotiable prerequisite for any serious course in econometrics. The reason for this is that econometrics courses have two objectives. One is to show how various quantitative techniques can be used to fit models given suitable data. This is relatively easy. The other is to develop an understanding of the statistical properties of these techniques and hence an understanding of why they work satisfactorily in certain contexts and not in others. This is much more demanding and it does require a good basic knowledge of statistical theory.

This review chapter is not intended to serve as an accelerated substitute for a statistics course. A common problem with statistics courses is that they attempt to cover some of the material too fast. Any abbreviated introduction to the discipline would be even more susceptible to such problems. The present treatment therefore assumes that a first course has already been completed and it is intended to provide an opportunity to revisit the material, with three objectives:

1. To provide an opportunity for revisiting all of the basic elements of statistical theory that are used in the remainder of the text. This is the function of Sections R.2–R.8, which cover:
   - random variables and expectations,
   - variance, covariance, and correlation,
   - sampling and estimators,
   - unbiasedness and efficiency, and
   - the normal distribution.

   Unless your statistics course was taken some time ago, it should be possible to skim through these sections fairly rapidly.
2. To reinforce the understanding of some standard topics that often require a second visit before being fully understood. This is the function of Sections R.9–R.13, which cover
   - hypothesis testing,
   - Type I error, Type II error, and the significance level (size) and power of a test,

- t tests,
- confidence intervals,
- one-sided t tests.

The discussion of the double structure of a sampled random variable and the distinction between its potential sample values and a realization in Section R.5 may also fall into this category.

3. To provide a treatment of some important topics on asymptotics (large-sample theory) that are often neglected in statistics courses:
   - probability limits and consistency,
   - convergence in distribution and central limit theorems.

   These topics are covered in Sections R.14 and R.15.

As far as Section R.13 is concerned, the material should be mostly revision, perhaps with some upgrading of understanding. However, the material on asymptotics in Sections R.14 and R.15 may be largely or wholly new. Despite the fact that asymptotic theory is of great importance in econometrics, it is usually covered inadequately in basic statistics courses. Typically, it is omitted entirely, apart from a brief reference to consistency and 'the' central limit theorem, often with no explanation of why these topics are useful. The reason for this neglect is that basic statistics courses tend to be service courses catering to students from many disciplines with a wide variety of priorities and interests and as a consequence they usually cover some topics that are of little relevance to future students of econometrics and they give insufficient attention to others. Apart from the core theory on sampling, estimation, and inference, the topics that are relevant to business studies or psychology are quite different from those that are relevant to econometrics. For most students of statistics, the topics in Sections R.14 and R.15 are not of great interest. However, for students of econometrics, they are crucial.

R.2 Discrete random variables and expectations

Discrete random variables

A simple notion of probability is adequate for the purposes of this text. We shall begin 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 continuous 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 obtain $X = 7$, we would need {green = 1, red = 6} or {green = 2, red = 5} or {green = 3, red = 4} or {green = 4, red = 3} or {green = 5, 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.

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	12

Figure R.1 Outcomes in the example with two dice

Table R.1 Frequencies and probability distribution, example with two dice

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

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.

Expected values of random variables

The expected value (sometimes described as expectation) 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 $x_1, x_2, ..., x_n$ and that the probability of x_i is p_i. Then

$$E(X) = x_1p_1 + \cdots + x_np_n = \sum_{i=1}^{n} x_ip_i. \qquad \textbf{(R.1)}$$

In the case of the two dice, the values x_1 to x_n were the numbers 2 to 12: $x_1 = 2, x_2 = 3, ..., x_{11} = 12$, and $p_1 = 1/36, p_2 = 2/36, ..., p_{11} = 1/36$. 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.

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

X	p	Xp	X	p	Xp
x_1	p_1	x_1p_1	2	1/36	2/36
x_2	p_2	x_2p_2	3	2/36	6/36
x_3	p_3	x_3p_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_np_n	12	1/36	12/36
Total		$E(X) = \sum_{i=1}^{n} x_ip_i$			252/36 = 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. 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 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.

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

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

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

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

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_X^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_X^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_X^2).

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.

Expected value 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,

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

Expected value 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,

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

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

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

The proof of rule 2 is left as an exercise (Exercise R.5). The proof of 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

$$Y = b_1 + b_2X \tag{R.6}$$

and b_1 and b_2 are constants. Then,

$$\begin{aligned} E(Y) &= E(b_1 + b_2X) \\ &= E(b_1) + E(b_2X) \quad \text{using rule 1} \\ &= b_1 + b_2E(X) \quad \text{using rules 2 and 3.} \end{aligned} \tag{R.7}$$

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

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, var(X), 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 $(X-\mu_X)^2$, where μ_X is the population mean. In equations it is usually denoted σ_X^2, with the subscript being dropped when it is obvious that it is referring to a particular variable:

$$\begin{aligned}\text{var}(X) = \sigma_X^2 &= E\left\{(X-\mu_X)^2\right\} \\ &= (x_1-\mu_X)^2 p_1 + \cdots + (x_n-\mu_X)^2 p_n = \sum_{i=1}^{n}(x_i-\mu_X)^2 p_i. \qquad \textbf{(R.8)}\end{aligned}$$

From σ_X^2 one obtains σ_X, the 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 $\mu_X = E(X) = 7$, $(X-\mu_X)^2$ is $(X-7)^2$ in this case. The expected value of $(X-7)^2$ is calculated in Table R.4 using Table R.3 as a pattern. An extra column, $(X-\mu_X)$, has been introduced as a step in the calculation of $(X-\mu_X)^2$. 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.

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

X	p	$X-\mu_X$	$(X-\mu_X)^2$	$(X-\mu_X)^2 p$
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

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

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

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

$$\begin{aligned}
\sigma_X^2 &= E\left\{(X - \mu_X)^2\right\} \\
&= E(X^2 - 2\mu_X X + \mu_X^2) \\
&= E(X^2) + E(-2\mu_X X) + E(\mu_X^2) \\
&= E(X^2) - 2\mu_X E(X) + \mu_X^2 \\
&= E(X^2) - 2\mu_X^2 + \mu_X^2 \\
&= E(X^2) - \mu_X^2 .
\end{aligned} \tag{R.10}$$

Line 3 uses expected value rule 1. Line 4 uses rules 2 and 3 (μ_X is a constant). Line 5 uses the fact that μ_X is just another way of writing $E(X)$.

Fixed and random components of a random variable

Instead of regarding a random variable as a single entity, it is often 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_X its population mean, one may make the following decomposition:

$$X = \mu_X + u, \tag{R.11}$$

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_X:

$$u = X - \mu_X. \tag{R.12}$$

It follows from its definition that the expected value of u is zero. From equation (R.12),

$$E(u) = E(X - \mu_X) = E(X) + E(-\mu_X) = \mu_X - \mu_X = 0. \tag{R.13}$$

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,

$$\sigma_X^2 = E\{(X-\mu_X)^2\} = E(u^2) \tag{R.14}$$

and

$$\begin{aligned}\sigma_u^2 &= E\{(u-\text{mean of } u)^2\} \\ &= E\{(u-0)^2\} = E(u^2).\end{aligned} \tag{R.15}$$

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.11), where μ_X is a fixed number and u is a random component, with mean zero and population variance σ^2, then X has population mean μ_X and population variance σ^2.

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 zero. 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 *Study Guide*.]

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

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

R.5 If X is a random variable with mean μ_X, and λ is a constant, prove that the expected value of λX is $\lambda\mu_X$.

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.

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 $2E(X) + 3$.

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_X 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_X in Exercise R.4 and $E(X^2)$ in Exercise R.7.)

R.3 Continuous random variables

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, the sum of the probabilities being 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. If you know the size of these packets, you can calculate the population mean and variance in a straightforward fashion.

However, 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 initially assume that this varies within the limits of 55 to 75°F, and we will suppose that it is equally likely to be anywhere within this range.

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 59–60°F is 0.05 since this range is one-twentieth of the complete range 55–75°F. 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 1, its height must be 0.05. The same is true for all the other one-degree intervals in the range that X can take.

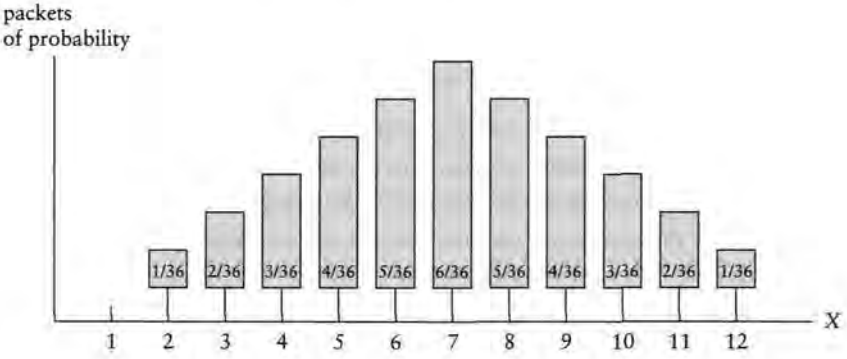


Figure R.2 Discrete probabilities (example with two dice)

Having found the height at all points in the range, we can answer other such questions relating to probabilities. For example, we can determine the probability that the temperature lies between 65 and 70°F. This is given by the area in the interval 65–70°F, 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 65–70°F is a quarter of the whole range.

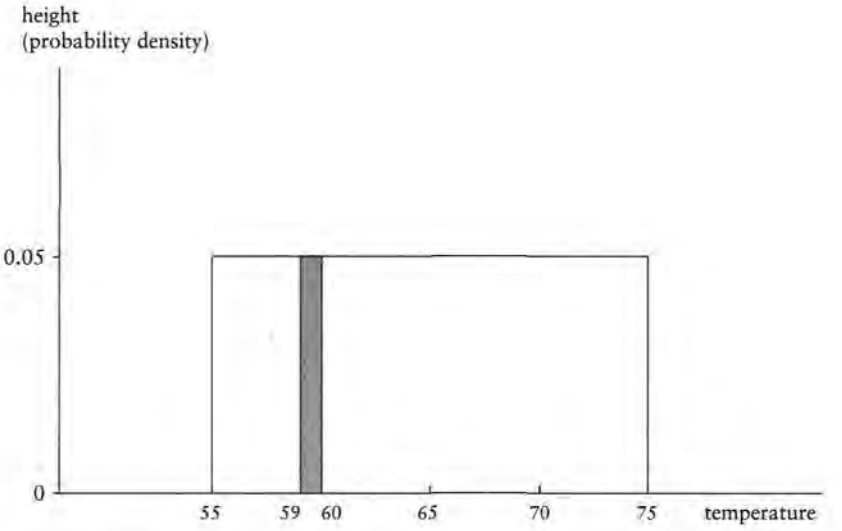


Figure R.3 Probability of the temperature lying in the interval 59–60°F

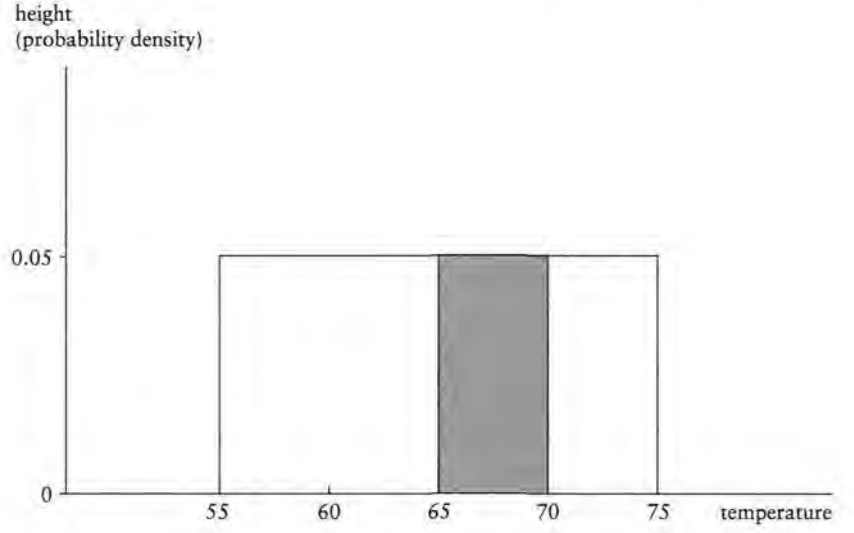


Figure R.4 Probability of the temperature lying in the interval 65–70°F

The height at any point is formally described as the probability density at that point. If the probability density can be written as a function of the random variable, that function is known as the probability density function. In this case it is given by $f(X)$, where X is the temperature and

$$f(X) = 0.05 \quad \text{for } 55 \leq X \leq 75$$

$$f(X) = 0 \quad \text{for } X < 55 \text{ or } X > 75. \tag{R.16}$$

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 zero 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 ½ × base × height, so one has

$$\frac{1}{2} \times 10 \times \text{height} = 1, \tag{R.17}$$

and so 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.

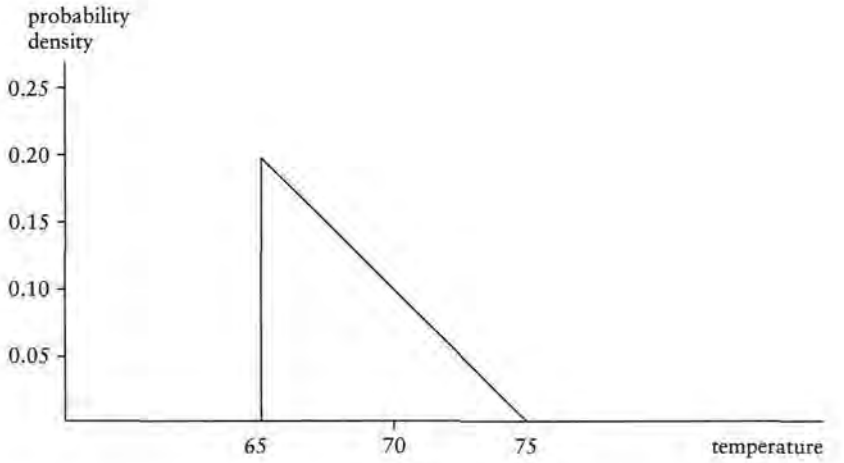


Figure R.5 Triangular density function, 65–75°F

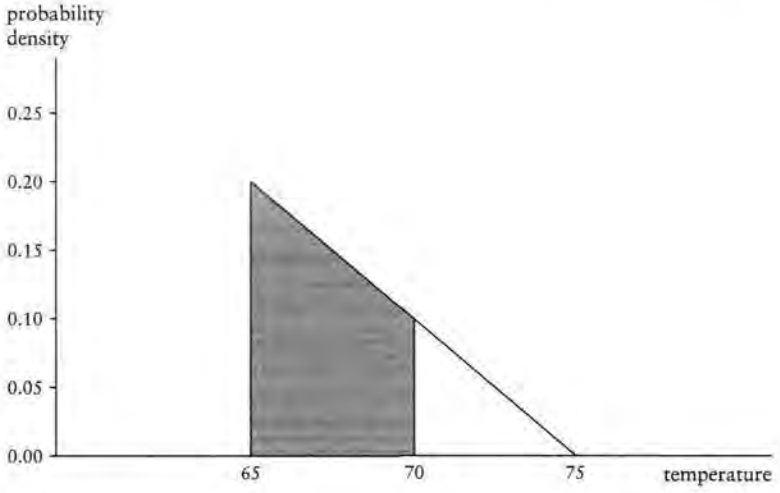


Figure R.6 Probability of the temperature lying in the interval 65–70°F

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

$$f(X) = 1.5 - 0.02X \quad \text{for } 65 \leq X \leq 75$$

$$f(X) = 0 \quad \text{for } X < 65 \text{ or } X > 75. \tag{R.18}$$

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

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. Fortunately, specialized probability tables do exist for all the functions that are going to interest us in practice. Integral calculus is also used in the definitions of the expected value and variance of a continuous random variable. These have much the same meaning for continuous random variables that they have for discrete ones (formal definitions are given in Box R.1), and the expected value rules work in exactly the same way.

R.4 Population covariance, covariance and variance rules, and correlation

Covariance

We come now to some concepts relating to two random variables. The first is covariance. If we have two random variables, X and Y, their population covariance, cov(X, Y), usually written σ_{XY} in equations, is defined to be the expected value of the product of their deviations from their means:

$$\text{cov}(X,Y) = \sigma_{XY} = E\{(X - \mu_X)(Y - \mu_Y)\}, \tag{R.19}$$

BOX R.1 Expected value and variance of a continuous random variable

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

$$E(X) = \int X f(X)\, dX,$$

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 is done on a packet-by-packet basis over all the possible values of X. In the continuous case, it is done on a continuous basis, which means summation by integration instead of Σ summation, with the probability density function $f(X)$ replacing the packets of probability p_i. However, the principle is the same. In the case of the discrete random variable, $E(X)$ is equal to $\sum_{i=1}^{n} x_i p_i$, with the summation taken over $x_1, \ldots, x_n$, the set of specific discrete values that X can take. In the continuous case, it is defined by

$$E(X) = \int X f(X)\, dX,$$

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

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

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 again exactly the same for a continuous random variable, except that it is done on a continuous basis, with integration replacing Σ summation.

Discrete $\sum_{i=1}^{n} g(x_i) p_i$

Continuous $E[g(X)] = \int g(X) f(X)\, dX.$

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 $(X - \mu_X)^2$, where $\mu_X = E(X)$ is the population mean. The variance is the summation of $(X - \mu_X)^2$,

weighted by the appropriate probability, over all the possible values of X. In the case of a discrete random variable, this means that you have to evaluate

$$\sigma_X^2 = E\left\{(X - \mu_X)^2\right\} = \sum_{i=1}^{n} (X_i - \mu_X)^2 p_i.$$

The counterpart for the continuous random variable is

$$\sigma_X^2 = E\left\{(X - \mu_X)^2\right\} = \int (X - \mu_X)^2 f(X) dX.$$

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

where μ_X and μ_Y are the population means of X and Y, respectively. It is a measure of association, but not as useful as correlation, to be discussed shortly. We are mostly interested in covariance as an ingredient in some of our analysis of the properties of estimators.

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 $h(Y)$. In particular, if X and Y are independent, $E(XY)$ is equal to $E(X)E(Y)$. If X and Y are independent, their population covariance is zero, since then

$$E\left\{(X - \mu_X)(Y - \mu_Y)\right\} = E(X - \mu_X)E(Y - \mu_Y) = 0 \times 0 \qquad \textbf{(R.20)}$$

by virtue of the fact that $E(X)$ and $E(Y)$ are equal to μ_X and μ_Y, respectively.

Covariance rules

There are some rules that follow in a straightforward way from the definition of covariance, and since they are going to be used frequently in later chapters, it is worthwhile establishing them immediately:

Covariance rule 1 If $Y = V + W$, $\text{cov}(X, Y) = \text{cov}(X, V) + \text{cov}(X, W)$.

Covariance rule 2 If $Y = bZ$, where b is a constant and Z is a variable, $\text{cov}(X, Y) = b\text{cov}(X, Z)$.

Covariance rule 3 If $Y = b$, where b is a constant, $\text{cov}(X, Y) = 0$.

Proof of covariance rule 1

Since $Y = V + W$, $\mu_Y = \mu_V + \mu_W$ by virtue of expected value rule 1. Hence,

$$\begin{aligned}\text{cov}(X,Y) &= E\{(X-\mu_X)(Y-\mu_Y)\}\\ &= E\{(X-\mu_X)([V+W]-[\mu_V+\mu_W])\}\\ &= E\{(X-\mu_X)(V-\mu_V)+(X-\mu_X)(W-\mu_W)\}\\ &= \text{cov}(X,V)+\text{cov}(X,W).\end{aligned} \tag{R.21}$$

Proof of covariance rule 2

If $Y = bZ$, $\mu_Y = b\mu_Z$. Hence,

$$\begin{aligned}\text{cov}(X,Y) &= E\{(X-\mu_X)(Y-\mu_Y)\}\\ &= E\{(X-\mu_X)(bZ-b\mu_Z)\}\\ &= bE\{(X-\mu_X)(Z-\mu_Z)\}\\ &= b\,\text{cov}(X,Z).\end{aligned} \tag{R.22}$$

Proof of covariance rule 3

This is trivial. If $Y = b$, $\mu_Y = b$ and

$$\begin{aligned}\text{cov}(X,Y) &= E\{(X-\mu_X)(Y-\mu_Y)\}\\ &= E\{(X-\mu_X)(b-b)\}\\ &= E\{0\}=0.\end{aligned} \tag{R.23}$$

Further developments

With these basic rules, you can simplify more complicated covariance expressions. For example, if a variable Y is equal to the sum of three variables U, V, and W,

$$\text{cov}(X, Y) = \text{cov}(X, [U + V + W]) = \text{cov}(X, U) + \text{cov}(X, [V + W]), \tag{R.24}$$

using rule 1 and breaking up Y into two parts, U and $V+W$. Hence,

$$\text{cov}(X, Y) = \text{cov}(X, U) + \text{cov}(X, V) + \text{cov}(X, W), \tag{R.25}$$

using rule 1 again.

As another example, suppose $Y = b_1 + b_2Z$, where b_1 and b_2 are constants and Z is a variable. Then

$$\begin{aligned}\text{cov}(X, Y) &= \text{cov}(X, [b_1 + b_2Z])\\ &= \text{cov}(X, b_1) + \text{cov}(X, b_2Z) && \text{using rule 1}\\ &= 0 + \text{cov}(X, b_2Z) && \text{using rule 3}\\ &= b_2\text{cov}(X, Z) && \text{using rule 2.}\end{aligned} \tag{R.26}$$

Variance rules

There are some straightforward rules for variances, the first three of which are counterparts of those for covariance:

Variance rule 1 If $Y = V + W$, $\text{var}(Y) = \text{var}(V) + \text{var}(W) + 2\text{cov}(V, W)$.
Variance rule 2 If $Y = bZ$, where b is a constant, $\text{var}(Y) = b^2\text{var}(Z)$.
Variance rule 3 If $Y = b$, where b is a constant, $\text{var}(Y) = 0$.
Variance rule 4 If $Y = V + b$, where b is a constant, $\text{var}(Y) = \text{var}(V)$.

It is useful to note that the variance of a variable X can be thought of as the covariance of X with itself:

$$\begin{aligned}\text{var}(X) &= E\left\{(X-\mu_X)^2\right\} \\ &= E\left\{(X-\mu_X)(X-\mu_X)\right\} \\ &= \text{cov}(X,X). \end{aligned} \tag{R.27}$$

In view of this equivalence, we can make use of the covariance rules to establish the variance rules.

Proof of variance rule 1

If $Y = V + W$,

$$\begin{aligned}\text{var}(Y) &= \text{cov}(Y,Y) = \text{cov}(Y,[V+W]) = \text{cov}(Y,V)+\text{cov}(Y,W) \text{ using covariance rule 1} \\ &= \text{cov}([V+W],V)+\text{cov}([V+W],W) \\ &= \text{cov}(V,V)+\text{cov}(W,V)+\text{cov}(V,W)+\text{cov}(W,W) \text{ using covariance rule 1 again} \\ &= \text{var}(V)+\text{var}(W)+2\text{cov}(V,W). \end{aligned} \tag{R.28}$$

Note that $\text{cov}(W, V)$ and $\text{cov}(V, W)$ are the same. The order of the variables makes no difference in the definition of covariance (R.19).

Proof of variance rule 2

If $Y = bZ$, where b is a constant, using covariance rule 2 twice,

$$\begin{aligned}\text{var}(Y) &= \text{cov}(Y,Y) = \text{cov}(bZ,Y) = b\text{cov}(Z,Y) \\ &= b\text{cov}(Z,bZ) = b^2\text{cov}(Z,Z) = b^2\text{var}(Z). \end{aligned} \tag{R.29}$$

Proof of variance rule 3

If $Y = b$, where b is a constant, using covariance rule 3,

$$\text{var}(Y) = \text{cov}(b, b) = 0. \tag{R.30}$$

This is trivial. If Y is a constant, its expected value is the same constant and $(Y - \mu_Y) = 0$. Hence $\text{var}(Y) = 0$.

Proof of variance rule 4

If $Y = V + b$, where V is a variable and b is a constant, using variance rule 1,

$$\begin{aligned} \text{var}(Y) &= \text{var}(V + b) = \text{var}(V) + \text{var}(b) + 2\text{cov}(V, b) \\ &= \text{var}(V). \end{aligned} \tag{R.31}$$

Correlation

As a measure of association between two variables X and Y, $\text{cov}(X, Y)$ is unsatisfactory because it depends on the units of measurement of X and Y. It is the expected value of the product of the deviation of X from its population mean and the deviation of Y from its population mean, $E\{(X - \mu_X)(Y - \mu_Y)\}$. The first deviation is measured in units of X and the second in units of Y. Change the units of measurement and you change the covariance. A better measure of association is the population correlation coefficient because it is dimensionless and therefore invariant to changes in the units of measure. It is traditionally denoted ρ, the Greek letter that is the equivalent of 'r', and pronounced 'row', as in 'row a boat'. For variables X and Y it is defined by

$$\rho_{XY} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}. \tag{R.32}$$

The numerator possesses the units of measurement of both X and Y. The variances of X and Y in the denominator possess the squared units of measurement of those variables. However, once the square root has been taken into account, the units of measurement are the same as those of the numerator, and the expression as a whole is unit free. It is left as an exercise to show that replacing X or Y by a linear function of itself (which is what happens when one changes units) has no effect on the correlation.

If X and Y are independent, ρ_{XY} will be equal to zero because σ_{XY} will be zero. If there is a positive association between them, σ_{XY}, and hence ρ_{XY}, will be positive. If there is an exact positive linear relationship, ρ_{XY} will assume its maximum value of 1. Similarly, if there is a negative relationship, ρ_{XY} will be negative, with a minimum value of -1.

EXERCISES

R.13 Suppose a variable Y is an exact linear function of X:

$$Y = \lambda + \mu X,$$

where λ and μ are constants. Demonstrate that the correlation between X and Y is equal to 1 or -1, according to the sign of μ.

R.14* Suppose a variable Y is an exact linear function of X:

$$Y = \lambda + \mu X,$$

where λ and μ are constants, and suppose that Z is a third variable. Show that $\rho_{XZ} = \rho_{YZ}$.

R.5 Samples, the double structure of a variable, and 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 derive the population mean and variance and any other population characteristics in which we might be interested.

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 observations and derive an estimate of the population characteristic using some appropriate formula. It is important to be very clear conceptually about what this involves and we will take it one step at a time.

Sampling

We will suppose that we have a random variable X and that we take a sample of n observations with the intention of obtaining information about the distribution of X. We might, for example, wish to estimate its population mean. Before we consider devising estimators, it is useful to make a distinction between the way we think about the sample *before* it has actually been taken and *after* we have taken it.

Once the sample has been generated, the observations are just specific numbers. A statistician would refer to this as a realization.

However, before the sample is generated, the potential observations $\{X_1, X_2, \ldots, X_n\}$ themselves may be thought of as a set of random numbers. Let us focus on the first observation, X_1. Before we take the sample, we do not know what the value of X_1 will be. All we know is that it will be generated randomly from the distribution for X. It is itself, therefore, a random variable. Being generated randomly from the distribution for X means that its potential distribution, before the sample is generated, is that of X.

The same is true for all the other observations in the sample, when we are thinking about their potential distribution before the sample is generated.

After the sample has been taken, we have a specific realization and would denote it as $\{x_1, x_2, \ldots, x_n\}$, the lower case indicating that the values are specific numbers.

So we are now thinking about the variable on two levels: the X variable that is the subject of attention, and the X_i components in a potential sample. It is essential to be clear about the double structure of a variable. It is the key to understanding the analysis of the properties of estimators based on the sample of observations.

Estimators

An estimator is a general rule, usually just a formula, for estimating an unknown population parameter given the sample of data. It is defined in terms of the $\{X_1, X_2,\ldots,X_n\}$. Once we have obtained a specific sample $\{x_1, x_2,\ldots,x_n\}$ we use it to obtain a specific number that we describe as the estimate. To repeat, the estimator is a formula, whereas the estimate is a number. If we take repeated samples, the estimator will be the same, but the estimate will vary from sample to sample.

An estimator is a special case of a random variable. This is because it is a combination of the $\{X_1, X_2,\ldots,X_n\}$ and, since the $\{X_1, X_2,\ldots,X_n\}$ are random quantities, a combination of them must also be a random variable.

The sample mean $\bar{X}$, the usual estimator of the population mean, provides a simple example since it is just the average of the X_i in the sample:

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i. \qquad \textbf{(R.33)}$$

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

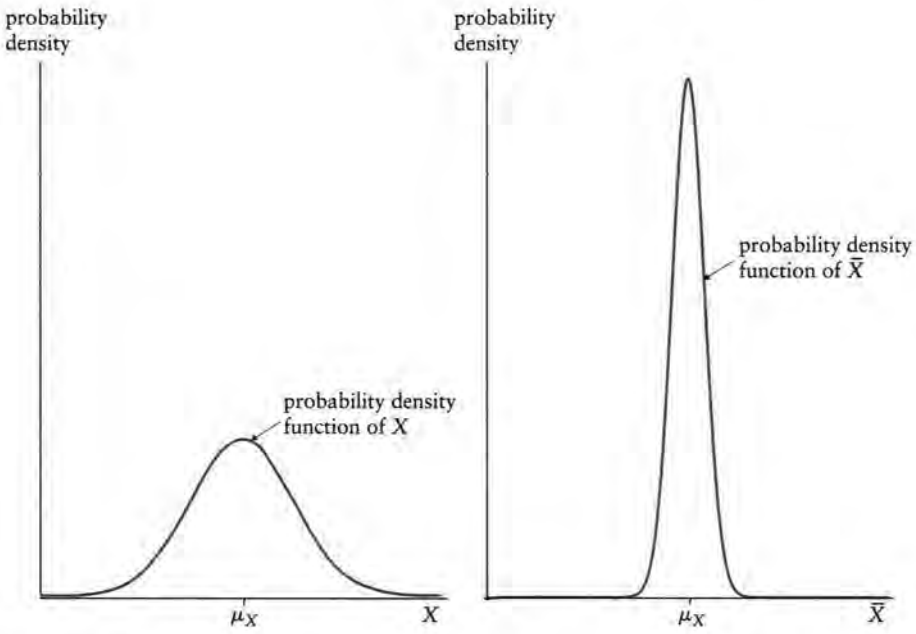


Figure R.7 Comparison of the probability density functions of a single observation and the mean of a sample

normal distribution. You will see that the distributions of both X and $\bar{X}$ are symmetrical about μ_X. The difference between them is that the distribution for $\bar{X}$ is narrower and taller. $\bar{X}$ tends to be closer to μ_X than a single observation on X because it is an average. Some of the X_i in the sample will be greater than the population mean, some will be smaller, and the positive deviations and the negative deviations will to some extent cancel each other out when the average is taken. We will demonstrate that if the distribution of X has variance σ_X^2, the sample mean has variance σ_X^2 / n:

$$\begin{aligned}
\sigma_{\bar{X}}^2 &= \operatorname{var}\left\{\frac{1}{n}(X_1 + \cdots + X_n)\right\} \\
&= \frac{1}{n^2}\operatorname{var}(X_1 + \cdots + X_n) \\
&= \frac{1}{n^2}\{\operatorname{var}(X_1) + \cdots + \operatorname{var}(X_n)\} \\
&= \frac{1}{n^2}(\sigma_X^2 + \cdots + \sigma_X^2) \\
&= \frac{1}{n^2}(n\sigma_X^2) = \frac{\sigma_X^2}{n}.
\end{aligned} \tag{R.34}$$

It may be helpful to have some further explanation of equation (R.34). There are some important conceptual issues at stake that are crucial to an understanding of basic statistical theory. We will go through the equation line by line.

The first line simply states what is meant by $\sigma_{\bar{X}}^2$.

The second uses variance rule 2 to take the factor $1/n$ out of the expression. The factor must be squared when it is taken out, as shown in the derivation of the rule in Section R.4.

The third and fourth lines are the ones that sometimes give rise to trouble. The third line uses variance rule 1.

$$\operatorname{var}(X_1 + \cdots + X_n) = \operatorname{var}(X_1) + \cdots + \operatorname{var}(X_n). \tag{R.35}$$

There are no population covariance terms on the right side of the equation because the observations are assumed to be generated independently. The fourth line reads as

$$\operatorname{var}(X_i) = \sigma_X^2 \quad \text{for all } i. \tag{R.36}$$

What is going on? X has a specific value in observation i, so how can it have a population variance?

The key to this is the double structure. We need to make a distinction between the *potential* distribution of the observations in the sample, as individual random variables, and the sample mean, *before* the sample is generated, and the *actual* realization *after* the sample has been generated.

To illustrate the distinction, we will suppose that X has a normal distribution (the bell-shaped distribution discussed in Section R.8) with population mean 5 and variance 1, and that there are 10 observations in the sample.

The first line of Table R.5 (Sample 1) shows the result of randomly drawing numbers for X_1 to X_{10} from this distribution. The sample mean is 5.04. The numbers that appear in this line are the realization for this sample. The remaining 19 rows are the same as the first, but with different sets of randomly generated numbers.

Next we will focus on the observation X_1. Looking at its values in the 20 samples, we obtain an insight into its *potential* distribution. We see that its average value is 4.95 and its estimated variance (see Section R.7) is 1.11. This is not a surprise, because X_1 has been generated randomly from a normal distribution with population mean 5 and population variance 1. The same is true

Table R.5

sample	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	$\bar{X}$
1	6.25	4.35	5.30	6.44	4.63	3.13	5.97	5.21	3.88	5.28	5.04
2	3.81	5.19	4.49	5.51	4.41	4.39	5.43	4.75	5.39	5.63	4.90
3	5.65	4.88	6.86	6.42	6.98	4.50	4.92	7.04	5.32	5.77	5.83
4	5.78	4.15	3.99	5.86	6.27	6.32	3.80	4.78	3.67	4.83	4.95
5	2.92	5.48	5.07	4.75	4.73	5.09	5.50	4.46	3.50	4.76	4.63
6	4.82	3.01	5.59	5.02	5.37	4.06	6.04	5.21	6.17	4.59	4.99
7	5.84	4.30	4.69	3.82	5.21	5.74	6.05	7.29	3.77	5.13	5.18
8	5.13	5.02	5.35	4.03	4.90	5.42	4.90	4.21	4.41	5.50	4.89
9	4.13	5.16	5.85	6.11	7.12	5.77	3.91	6.30	3.88	2.81	5.10
10	5.21	4.91	4.01	4.45	5.75	3.20	3.84	3.93	4.08	3.88	4.33
11	7.32	3.96	2.75	5.69	4.60	7.90	3.61	5.88	5.47	3.34	5.05
12	6.52	5.51	5.34	5.47	4.51	5.72	2.78	4.40	4.55	4.80	4.96
13	4.71	5.06	6.22	5.99	4.62	5.00	5.38	3.56	3.90	5.35	4.98
14	4.59	4.54	4.63	4.84	6.38	5.62	4.75	5.86	4.57	4.64	5.04
15	5.13	4.99	7.36	4.60	3.85	5.26	6.13	5.26	5.83	4.83	5.32
16	4.26	4.99	4.49	4.48	4.76	3.77	5.49	5.31	6.66	6.44	5.07
17	4.07	5.55	4.26	5.07	4.96	4.38	5.85	5.51	4.21	5.12	4.90
18	3.83	5.14	5.69	5.24	3.41	4.24	5.00	4.99	5.40	4.09	4.70
19	4.38	5.54	3.70	5.06	5.59	4.00	5.16	4.64	6.25	6.03	5.03
20	4.74	4.45	4.69	5.67	5.51	3.84	5.14	4.89	4.16	4.89	4.80
mean	4.95	4.81	5.02	5.23	5.18	4.87	4.98	5.17	4.75	4.89	4.98
estimated variance	1.11	0.43	1.19	0.64	0.98	1.33	0.89	0.96	0.99	0.77	0.09

for all the other X observations. In each case, their average over the 20 samples is approximately 5, and their variances are approximately 1, and the approximations would have been closer to the population values if we had had more samples. This is what we are referring to when we write

$$\frac{1}{n^2}\left(\sigma_{X_1}^2 + \cdots + \sigma_{X_n}^2\right) = \frac{1}{n^2}\left(\sigma_X^2 + \cdots + \sigma_X^2\right) \tag{R.37}$$

in equation (R.34). We are saying that the variance of the *potential* distribution of X_1, before a sample is generated, is σ_X^2, because X_1 is drawn randomly from the distribution for X. The same is true for all the other observations.

Of course, we are not interested in the distributions of the individual observations, but in the distribution of the estimator, in this case $\bar{X}$. What we are saying is that the variance of the potential distribution of $\bar{X}$, before the sample is generated, is σ_X^2/n.

R.6 Unbiasedness and efficiency

Much of the analysis in later chapters will be concerned with three properties of estimators: unbiasedness, efficiency, and consistency. The first two, treated in this section, relate to finite sample analysis: analysis where the samples have a finite number of observations. Consistency, a property that relates to analysis when the sample size tends to infinity, is treated in Section R.14.

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 not lead us astray by having a tendency either to overestimate or to underestimate the population characteristic. 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 value of the population characteristic is described as the bias.

Let us start with the sample mean. We will show that its expected value is equal to μ_X and that it is therefore an unbiased estimator of the population mean:

$$\begin{aligned} E(\bar{X}) &= E\left\{\frac{1}{n}(X_1 + \cdots + X_n)\right\} = \frac{1}{n}E(X_1 + \cdots + X_n) \\ &= \frac{1}{n}\left\{E(X_1) + \cdots + E(X_n)\right\} \\ &= \frac{1}{n}(\mu_X + \cdots + \mu_X) = \frac{1}{n}(n\mu_X) = \mu_X. \end{aligned} \tag{R.38}$$

Note that when we make the step

$$\frac{1}{n}\{E(X_1) + \cdots + E(X_n)\} = \frac{1}{n}(\mu_X + \cdots + \mu_X), \tag{R.39}$$

we are referring to the fact that the potential distribution of each X_i, before the sample is actually generated, has population mean μ_X.

We have shown that the sample mean is an unbiased estimator of the population mean μ_X. However, it is not the only unbiased estimator 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:

$$Z = \lambda_1 X_1 + \lambda_2 X_2. \tag{R.40}$$

The expected value of Z is given by

$$\begin{aligned} E(Z) &= E(\lambda_1 X_1 + \lambda_2 X_2) = E(\lambda_1 X_1) + E(\lambda_2 X_2) \\ &= \lambda_1 E(X_1) + \lambda_2 E(X_2) = \lambda_1 \mu_X + \lambda_2 \mu_X \\ &= (\lambda_1 + \lambda_2)\mu_X. \end{aligned} \tag{R.41}$$

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

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. There is a more compelling reason: efficiency.

Efficiency

Unbiasedness is one desirable feature of an estimator, but it is not the only one. Another important consideration is its reliability. 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.

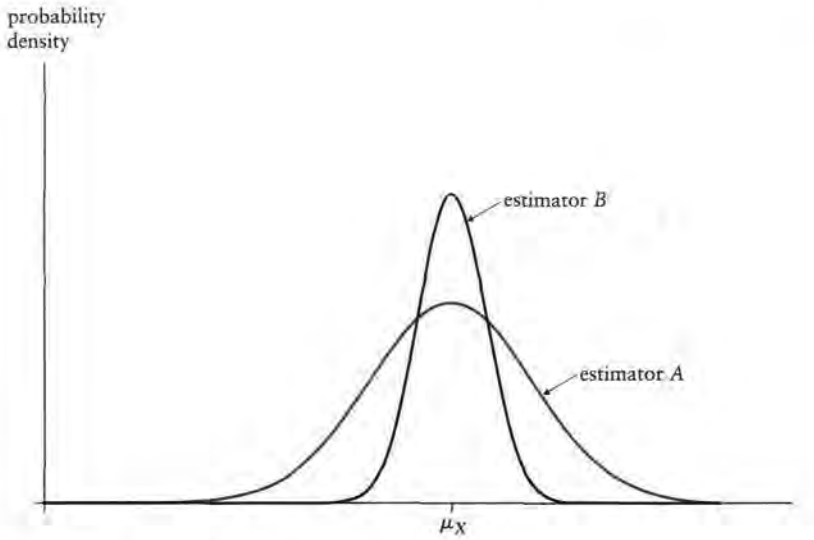


Figure R.8 Efficient and inefficient estimators

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. Sometimes it will have a bad day, and estimator *A* will have a good day, and *A* will be closer to the true value. But as a matter of probability, *B* will tend to be more accurate than *A*.

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 might 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.

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.

The population variance of the generalized estimator is given by

$$\begin{aligned}\sigma_Z^2 &= \operatorname{var}(\lambda_1 X_1 + \lambda_2 X_2) \\ &= \operatorname{var}(\lambda_1 X_1) + \operatorname{var}(\lambda_2 X_2) + 2\operatorname{cov}(\lambda_1 X_1, \lambda_2 X_2) \\ &= \lambda_1^2 \sigma_{X_1}^2 + \lambda_2^2 \sigma_{X_2}^2 + 2\lambda_1 \lambda_2 \sigma_{X_1 X_2} \\ &= (\lambda_1^2 + \lambda_2^2)\sigma_X^2. \end{aligned} \tag{R.42}$$

We are assuming that X_1 and X_2 are generated independently and hence that $\sigma_{X_1X_2}$ is zero.

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 unbiasedness, $\lambda_2 = 1 - \lambda_1$ and

$$\lambda_1^2 + \lambda_2^2 = \lambda_1^2 + (1-\lambda_1)^2 = 2\lambda_1^2 - 2\lambda_1 + 1. \quad \textbf{(R.43)}$$

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. The first-order condition is

$$4\lambda_1 - 2 = 0. \quad \textbf{(R.44)}$$

Thus, the minimum value is reached when λ_1 is equal to 0.5. Hence λ_2 is also equal to 0.5. (We should check the second derivative. This is equal to 4, which is positive, confirming that we have found a minimum rather than a maximum.)

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 Section R.14, 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.

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*.

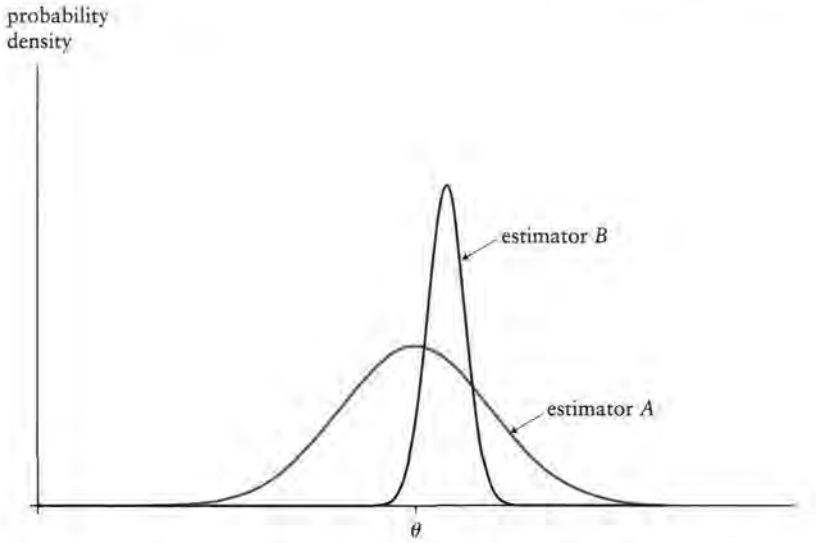


Figure R.9 Which estimator is to be preferred? *A* is unbiased but *B* has smaller variance

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 is known as the mean square error (MSE):

$$\text{MSE of estimator} = E\left\{(Z-\theta)^2\right\}, \tag{R.45}$$

where Z is the estimator and θ is the value of the population characteristic being estimated.

The MSE can be decomposed into the variance of Z plus the square of the bias:

$$\text{MSE of estimator} = \text{variance of estimator} + \text{bias}^2. \tag{R.46}$$

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

$$\begin{aligned} E\left\{(Z-\theta)^2\right\} &= E\left\{([Z-\mu_Z]+[\mu_Z-\theta])^2\right\} \\ &= E\left\{(Z-\mu_Z)^2 + 2(Z-\mu_Z)(\mu_Z-\theta) + (\mu_Z-\theta)^2\right\} \\ &= E\left\{(Z-\mu_Z)^2\right\} + 2(\mu_Z-\theta)E(Z-\mu_Z) + E\left\{(\mu_Z-\theta)^2\right\} \end{aligned} \tag{R.47}$$

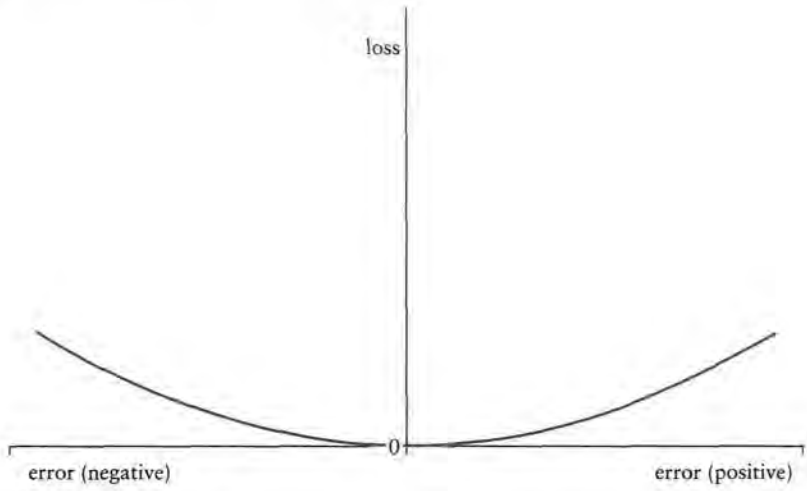


Figure R.10 Loss function

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

$$E(Z-\mu_Z)=E(Z)+E(-\mu_Z)=\mu_Z-\mu_Z=0. \tag{R.48}$$

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 the decomposition.

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 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 (R.43) 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.16* 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_X is $\lambda_1 + \cdots + \lambda_n = 1$.

R.17 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.18 Draw a loss function for getting to an airport later (or earlier) than the official check-in time.

R.19* 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?

R.20 If you have two estimators of an unknown population parameter, is the one with the smaller variance necessarily more efficient?

R.7 Estimators of variance, covariance, and correlation

The concepts of population variance and covariance were defined in Sections R.2 and R.4. For a random variable X, the population variance σ_X^2 is

$$\text{var}(X) = \sigma_X^2 = E\left\{\left(X - \mu_X\right)^2\right\}. \tag{R.49}$$

Given a sample of n observations, the usual estimator of σ_X^2 is the sum of the squared deviations around the sample mean divided by $n - 1$, typically denoted s_X^2:

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2. \tag{R.50}$$

Since the population variance is the expected value of the squared deviation of X about its mean, it makes intuitive sense to use the average of the sample squared deviations as an estimator. But why divide by $n - 1$ rather than by n? The reason is that the sample mean is by definition in the middle of the sample, while the unknown population mean is not, except by coincidence. As a consequence, the sum of the squared deviations from the sample mean tends to be slightly smaller than the sum of the squared deviations from the population mean. As a consequence, a simple average of the squared sample deviations is a downwards biased estimator of the population variance. However, the bias can be shown to be a factor of $(n - 1)/n$. Thus, one can allow for the bias by dividing the sum of the squared deviations by $n - 1$ instead of n. A formal proof of the unbiasedness of s_X^2 is given in Appendix R.1.

A similar adjustment has to be made when estimating a population covariance. For two random variables, X and Y, the population covariance σ_{XY} is

$$\text{cov}(X,Y) = \sigma_{XY} = E\left\{\left(X - \mu_X\right)\left(Y - \mu_Y\right)\right\}. \tag{R.51}$$

An unbiased estimator of σ_{XY} is given by the sum of the products of the deviations around the sample means divided by $n - 1$, typically denoted s_{XY}:

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right). \tag{R.52}$$

Again, for a formal proof of the unbiasedness of s_{XY}, see Appendix R.1.

The population correlation coefficient ρ_{XY} was defined in Section R.4 as

$$\rho_{XY} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}. \tag{R.53}$$

The sample correlation coefficient, r_{XY} is obtained from this by replacing σ_{XY}, σ_X^2, and σ_Y^2 by their estimators:

$$\begin{aligned} r_{XY} &= \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} = \frac{\frac{1}{n-1}\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n-1}\sum (X_i - \bar{X})^2 \frac{1}{n-1}\sum (Y_i - \bar{Y})^2}} \\ &= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}. \end{aligned} \tag{R.54}$$

EXERCISE

R.21* Suppose that you have observations on three variables X, Y, and Z, and suppose that Y is an exact linear function of Z:

$$Y = a + bZ$$

where a and b are constants. Show that $r_{XZ} = r_{XY}$. (This is the counterpart of Exercise R.14.)

R.8 The normal distribution

In the analysis so far, we have discussed the mean and the variance of a distribution of a random variable, but we have not said anything specific about the actual shape of the distribution. It is time to do that. There are only four distributions, all of them continuous, that are going to be of importance to us: the normal distribution, the t distribution, the F distribution, and the chi-squared (χ^2) distribution. We will consider the normal distribution here, the t distribution in Section R.11, the F distribution in Chapter 2, and the chi-squared distribution in Chapter 8.

The normal distribution has the graceful, bell-shaped form shown in Figure R.11. The probability density function for a normally distributed random variable X is

$$f(X) = \frac{1}{\alpha\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\beta}{\alpha}\right)^2} \tag{R.55}$$

where α and β are parameters. It is in fact an infinite family of distributions since β can be any finite real number and α any finite positive real number. The expression may seem somewhat forbidding at first, but we can make an immediate improvement. It can be shown that the expected value of the distribution, μ,

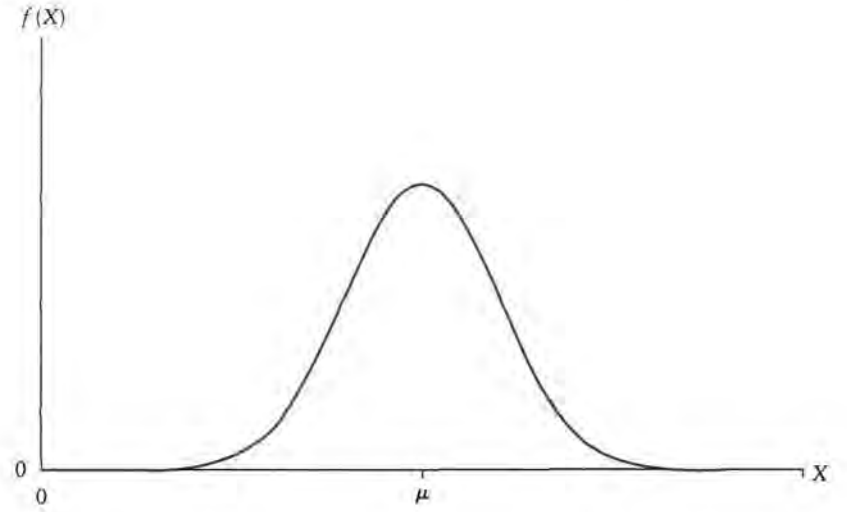


Figure R.11 Normal distribution

is equal to β and its variance, σ^2, is equal to α^2. Thus, it is natural to write the probability density function in the form

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} \tag{R.56}$$

The distribution is symmetric, so it automatically follows that the mean and the mode coincide in the middle of the distribution. Further, its shape is fixed when expressed in terms of standard deviations, so all normal distributions look the same when expressed in terms of μ and σ. This is shown in Figure R.12.

As a matter of mathematical shorthand, if a variable X is normally distributed with mean μ and variance σ^2, this is written

$$X \sim N(\mu, \sigma^2) \tag{R.57}$$

(the symbol ~ means 'is distributed as'). The first argument in the parentheses refers to the mean and the second to the variance. This, of course, is the general expression. If you had a specific normal distribution, you would replace the arguments with the actual numerical values.

An important special case is the standardized normal distribution, where $\mu = 0$ and $\sigma = 1$. This is shown in Figure R.13.

EXERCISE

R.22 A scalar multiple of a normally distributed random variable also has a normal distribution. A random variable X has a normal distribution with mean 5 and variance 10. Sketch the distribution of $Z = X/2$.

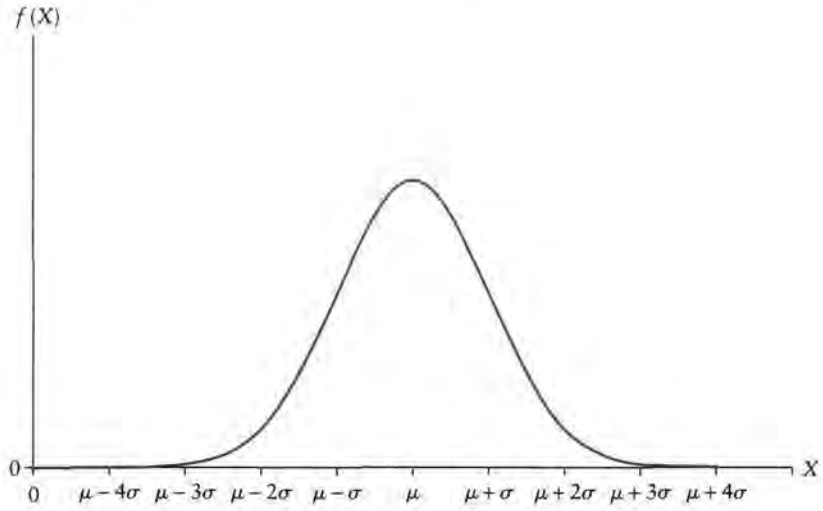


Figure R.12 Structure of the normal distribution in terms of μ and σ

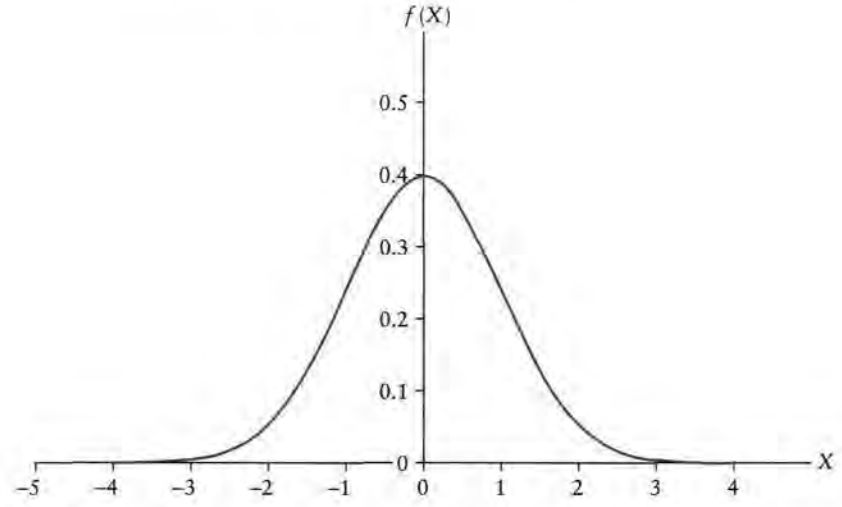


Figure R.13 Standardized normal distribution

R.9 Hypothesis testing

Which comes first, theoretical hypothesizing or empirical research? Perhaps, at the outset, theory might be the starting point, but in practice, theorizing and experimentation rapidly become interactive activities. For this reason, we will approach the topic of hypothesis testing from both directions. On the one hand, we may suppose that the theory has come first and that the purpose of the experiment is to evaluate its plausibility. This will lead to the execution of significance tests. Alternatively, we may perform the experiment first and then

consider what theoretical hypotheses would be consistent with the results. This will lead to the construction of confidence intervals.

Formulation of a null hypothesis and development of its implications

Suppose that a random variable X is assumed to have a normal distribution with mean μ and variance σ^2. We will start by assuming that the theory precedes the experiment and that we hypothesize that μ is equal to some specific value μ_0. We then describe

$$H_0: \mu = \mu_0 \tag{R.58}$$

as our null hypothesis. We also define an alternative hypothesis, denoted H_1, which represents our conclusion if the evidence indicates that H_0 is false. In the present case, H_1 is simply that μ is not equal to μ_0:

$$H_1: \mu \neq \mu_0. \tag{R.59}$$

Our test strategy consists of generating a sample of n independent observations of X and calculating the sample mean, $\bar{X}$. If the null hypothesis is true, values of $\bar{X}$ obtained in repeated samples will be normally distributed with mean μ_0 and variance σ^2/n. Since the variance of the distribution is σ^2/n, the standard deviation is $\sigma/\sqrt{n}$. The potential distribution of $\bar{X}$, conditional on H_0 being true, is shown in Figure R.14. To draw this figure, we have to know the standard deviation of $\bar{X}$, $\sigma/\sqrt{n}$, which means that we have to know the value of σ. For the time being, to simplify the discussion, we shall assume that we do. In practice, we have to estimate it, so we will eventually need to relax this assumption.

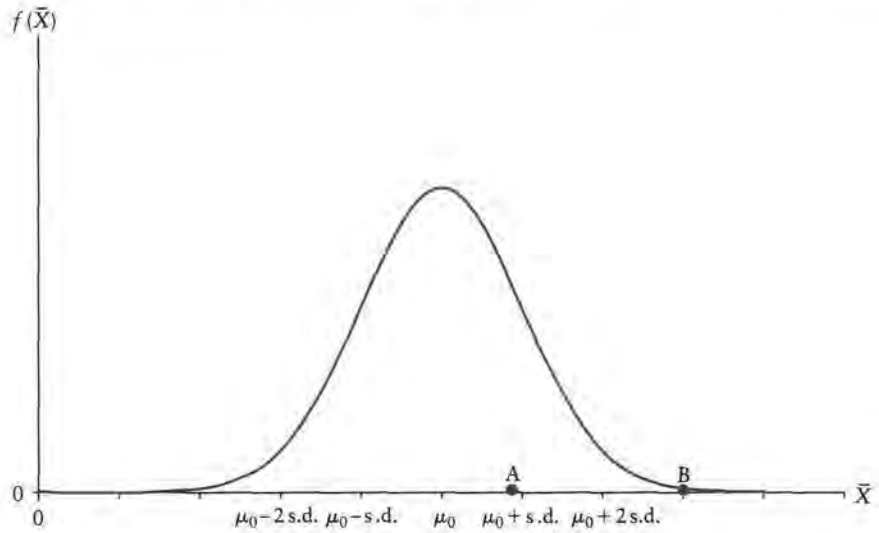


Figure R.14 Distribution of $\bar{X}$ if H_0: $\mu = \mu_0$ is true. s.d. = standard deviation of $\bar{X}$

Compatibility, freakiness, and the significance level

Now we come to the crunch. Suppose that we take an actual sample of observations and calculate $\bar{X}$. Suppose that it lies within one standard deviation of μ_0. The point A in Figure R.14 is an example of such an outcome. It is perfectly compatible with the null hypothesis. We do not anticipate $\bar{X}$ being exactly equal to μ_0, except by freak coincidence, because it has a random component, but we would expect it to be somewhere 'close' to μ_0, in the sense of occurring reasonably frequently if H_0 is true. That is the case with A.

Suppose, on the other hand, that $\bar{X}$ is located at the point B, three standard deviations above μ_0. If the null hypothesis is correct, the probability of being three standard deviations away from the mean, positive or negative, is only 0.0027, which is very low. You could come to either of two conclusions about this problematic result:

1. You could continue to maintain that your null hypothesis H_0: $\mu = \mu_0$ is correct, and that the experiment has given a freak result. You concede that the probability of such a low value of $\bar{X}$ is very small, but nevertheless it does occur 0.27 percent of the time and you reckon that this is one of those times.
2. You could conclude that the hypothesis is contradicted by the sample result. You are not convinced by the explanation in (1) because the probability is so small and you think that a much more likely explanation is that μ is not really equal to μ_0. In other words, you adopt the alternative hypothesis H_1: $\mu \neq \mu_0$ instead.

How do you decide when to choose (1) and when to choose (2)? There is no error-free procedure. The best you can do is to establish some decision rule, but there is no guaranteed way of avoiding mistakes. Any decision rule will sometimes lead to the rejection of the null hypothesis when it is in fact true. This is known as Type I error. It will also sometimes lead to the non-rejection of the null hypothesis when it is in fact false. This is known as Type II error.

The usual procedure is to reject the null hypothesis if it implies that the probability of getting such an extreme estimate is less than some (small) probability p. For example, we might choose to reject the null hypothesis if it implies that the probability of getting such an extreme estimate is less than 0.05 (5 percent). According to this decision rule, we would reject the null hypothesis H_0: $\mu = \mu_0$ if $\bar{X}$ fell in the upper or lower 2.5 percent tails of the distribution shown in Figure R.14. This occurs when $\bar{X}$ is more than 1.96 standard deviations from μ_0. If you look up the normal distribution table, Table A.1 in Appendix A, you will see that the probability of $\bar{X}$ being more than 1.96 standard deviations above its mean is 2.5 percent, and similarly the probability of it being more than 1.96 standard deviations below its mean is 2.5 percent. The total probability of it being more than 1.96 standard deviations away is thus 5 percent.

Figure R.15 shows the rejection regions, thus defined, for the null hypothesis H_0: $\mu = \mu_0$. It also shows the points A and B in Figure R.14. If $\bar{X}$ were equal to the

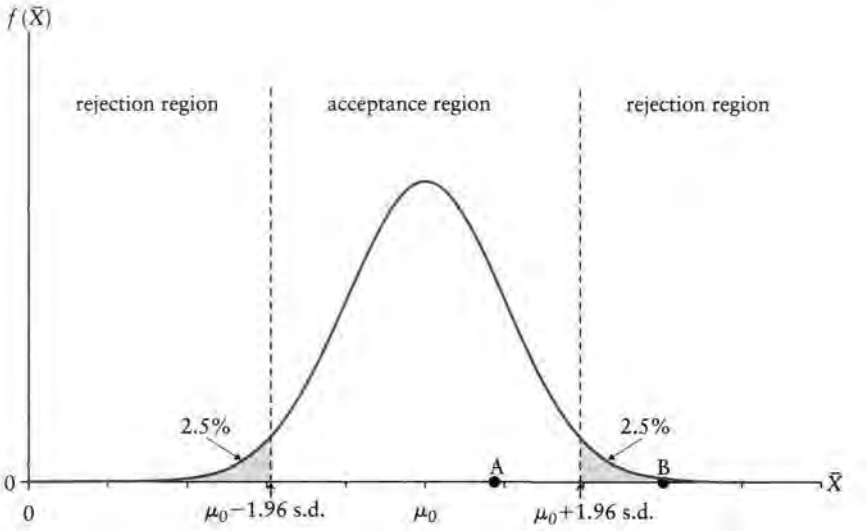


Figure R.15 Rejection regions, conditional on H_0: $\mu = \mu_0$, 5 percent test

value indicated by the point B, H_0 would be rejected. The value indicated by the point A would not lead to rejection. Generalizing, we can see from the figure that we would reject H_0: $\mu = \mu_0$ if

$$(1) \quad \bar{X} > \mu_0 + 1.96 \text{ s.d.}(\bar{X})$$

or

$$(2) \quad \bar{X} < \mu_0 - 1.96 \text{ s.d.}(\bar{X}). \tag{R.60}$$

Rewriting the inequality, we would reject H_0 if

$$(1) \quad \bar{X} - \mu_0 > 1.96 \text{ s.d.}(\bar{X})$$

or

$$(2) \quad \bar{X} - \mu_0 < -1.96 \text{ s.d.}(\bar{X}). \tag{R.61}$$

Dividing through by the standard deviation of $\bar{X}$, we reject H_0 if

$$(1) \quad \frac{\bar{X} - \mu_0}{\text{s.d.}(\bar{X})} > 1.96$$

or

$$(2) \quad \frac{\bar{X} - \mu_0}{\text{s.d.}(\bar{X})} < -1.96. \tag{R.62}$$

It is convenient to define a z statistic that is equal to the discrepancy between $\bar{X}$ and μ_0 measured in terms of standard deviations:

$$z = \frac{\bar{X} - \mu_0}{\text{s.d.}(\bar{X})}. \tag{R.63}$$

Then the decision rule is to reject H_0 if

$$(1) \quad z > 1.96$$

or

$$(2) \quad z < -1.96, \tag{R.64}$$

that is, if z is greater than 1.96 in absolute terms.

The inequality

$$\mu_0 - 1.96 \text{ s.d.}\left(\bar{X}\right) \leq \bar{X} \leq \mu_0 + 1.96 \text{ s.d.}\left(\bar{X}\right) \tag{R.65}$$

gives the set of values of $\bar{X}$ that will not lead to the rejection of a specific null hypothesis $\mu = \mu_0$. It is known as the acceptance region for $\bar{X}$, at the 5 percent significance level. 'Acceptance' is an unfortunate term, because it misleadingly suggests that we should conclude that the null hypothesis is true, if $\bar{X}$ lies within it. In fact, there will in general be a whole range of null hypotheses not contradicted by the sample result, so it is too strong to talk of 'accepting' H_0. It would be better to talk of the 'fail-to-reject' region, but it is too late to change the terminology now.

The decision procedure described above is not foolproof. Suppose that the null hypothesis H_0: $\mu = \mu_0$ is true. Then there is a 2.5 percent probability that $\bar{X}$ will be so large that it lies in the right rejection region and we decide to reject H_0. Likewise, there is a 2.5 percent probability that it will be so large and negative that it lies in the left rejection region. So there is a 5 percent chance of the occurrence of a Type I error. The significance level of a test is the term used to describe the risk of a Type I error if the null hypothesis is true. Another, equivalent, term is the size of a test.

Of course, we can reduce the risk of making a Type I error by reducing the size of the rejection region. For example, we could change the decision rule to reject the null hypothesis only if it implies that the probability of getting the sample value is less than 0.01 (1 percent). The rejection region now becomes the upper and lower 0.5 percent tails. Looking at the normal distribution table again, you will see that the 0.5 percent tails of a normal distribution start 2.58 standard deviations from the mean, as shown in Figure R.16. Since the probability of making a Type I error, if the null hypothesis is true, is now only 1 percent, the test is said to be a 1 percent significance test.

By definition, the lower is your critical probability, the smaller is the risk of a Type I error. If your significance level is 5 percent, you will reject a true hypothesis 5 percent of the time. If it is 1 percent, you will make a Type I error 1 percent of the time. Thus, the 1 percent significance level is safer in this respect. If you reject the hypothesis at this level, you are almost certainly right to do so. For this reason, the 1 percent significance level is described as *higher* than the 5 percent level.

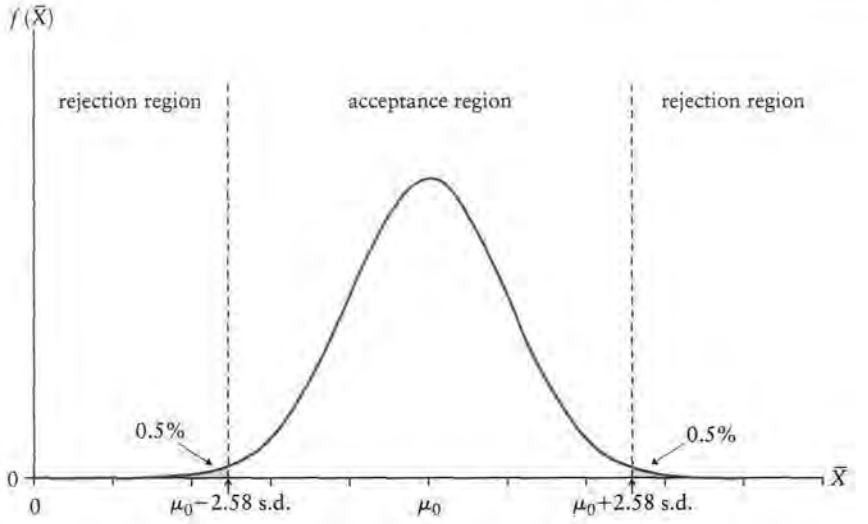


Figure R.16 Rejection regions, conditional on H_0: $\mu = \mu_0$, 1 percent test

Note that the acceptance region for the 5 percent test lies entirely inside the acceptance region for the 1 percent test. This has important implications for the *reporting of tests:*

1. If you reject the null hypothesis at the 1 percent level, you must also reject it at the 5 percent level. There is no need to mention the 5 percent test. Indeed, you look ignorant if you do.
2. If you do not reject at the 5 percent level, you cannot reject at the 1 percent level. There is no need to mention the 1 percent test, and, again, you look ignorant if you do.
3. You should mention both tests only when you reject at 5 percent but do not reject at 1 percent.

Example

Suppose that a random variable X may be assumed to have a normal distribution with variance 4. It is hypothesized that the unknown mean μ is equal to 10. Given a sample of 25 observations, suppose that we wish to determine the acceptance and rejection regions for $\bar{X}$ under H_0: $\mu = 10$ (a) using a 5 percent significance test, (b) using a 1 percent test.

If the variance of X is 4, its standard deviation is 2 and the standard deviation of $\bar{X}$ is $2/\sqrt{25} = 0.4$. The acceptance region, at the 5 percent significance level, is therefore

$$10 - 1.96 \times 0.4 \leq \bar{X} \leq 10 + 1.96 \times 0.4, \qquad \textbf{(R.66)}$$

which is

$$9.22 \leq \bar{X} \leq 10.78. \qquad \textbf{(R.67)}$$

If the sample mean lies in this range, the null hypothesis will not be rejected at the 5 percent significance level. If it is greater than 10.78 or less than 9.22, we reject H_0. Replacing 1.96 by 2.58, the acceptance region for a 1 percent test is

$$8.97 \leq \bar{X} \leq 11.03. \tag{R.68}$$

EXERCISE

R.23 Suppose that a random variable with unknown mean may be assumed to have a normal distribution with variance 25. Given a sample of 100 observations, derive the acceptance and rejection regions for $\bar{X}$ (a) using a 5 percent significance test, (b) using a 1 percent test.

R.10 Type II error and the power of a test

A Type I error occurs when the null hypothesis is rejected when it is in fact true. A Type II error occurs when the null hypothesis is not rejected when it is in fact false. We will see that, in general, there is a trade-off between the risk of making a Type I error and the risk of making a Type II error.

BOX R.2 Type I and Type II errors in everyday life

The problem of trying to avoid Type I and Type II errors is pervasive in everyday life. A criminal trial provides a particularly acute example. Taking as the null hypothesis that the defendant is innocent, a Type I error occurs when the jury wrongly decides that the defendant is guilty. A Type II error occurs when the jury wrongly acquits the defendant.

We will consider the case where the null hypothesis, H_0: $\mu = \mu_0$ is false and the actual value of μ is μ_1. This is shown in Figure R.17. If the null hypothesis is tested, it will be rejected only if $\bar{X}$ lies in one of the rejection regions associated with it. To determine the rejection regions, we draw the distribution of $\bar{X}$ conditional on H_0 being true. The distribution is marked with a dashed curve to emphasize that H_0 is not actually true. The rejection regions for a 5 percent test, given this distribution, are marked on the diagram.

If $\bar{X}$ lies in the acceptance region, H_0 will not be rejected, and so a Type II error will occur. What is the probability of this happening? To determine this, we now turn to the actual distribution of $\bar{X}$, given that $\mu = \mu_1$. This is the solid curve on the right. The probability of $\bar{X}$ lying in the acceptance region for H_0 is the area under this curve in the acceptance region. It is the shaded area in Figure R.18. In this particular case, the probability of $\bar{X}$ lying within the acceptance region for H_0, thus causing a Type II error, is 0.15.

The probability of rejecting the null hypothesis, when it is false, is known as the power of a test. By definition, it is equal to 1 minus the probability of making a Type II error. It is therefore 0.85 in this example.

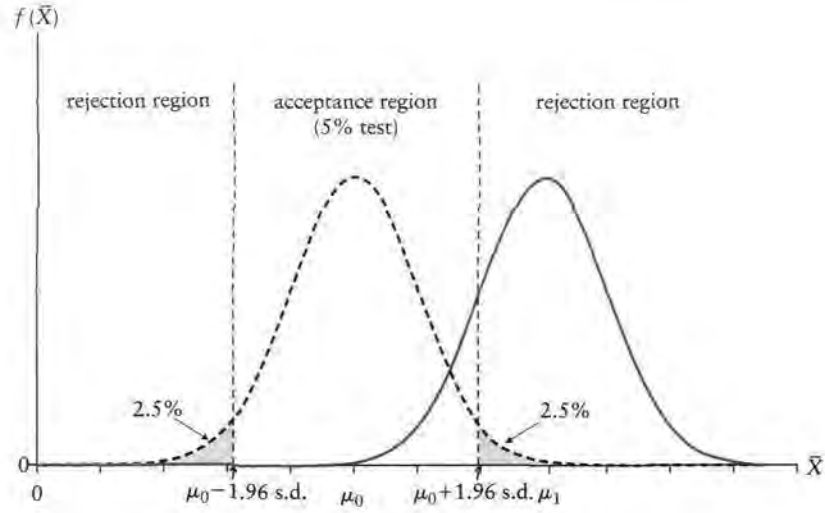


Figure R.17 Acceptance and rejection regions, conditional on H_0: $\mu = \mu_0$, 5 percent test

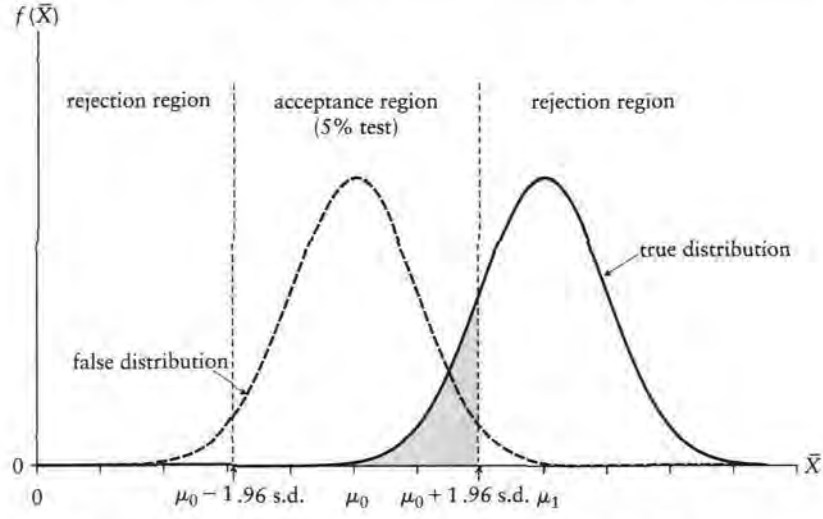


Figure R.18 Probability of making a Type II error if $\mu = \mu_1$, 5 percent test

The power depends on the distance between the value of μ under the false null hypothesis and its actual value. The closer that the actual value is to μ_0, the harder it is to demonstrate that H_0: $\mu = \mu_0$ is false. This is illustrated in Figure R.19. μ_0 is the same as in Figure R.18, and so the acceptance region and rejection regions for the test of H_0: $\mu = \mu_0$ are the same as in Figure R.18. As in Figure R.18, H_0 is false, but now the true value is μ_2, and μ_2 is closer to μ_0. As a consequence, the probability of $\bar{X}$ lying in the acceptance region for H_0 is much greater, 0.68 instead of 0.15, and so the power of the test, 0.32, is much lower.

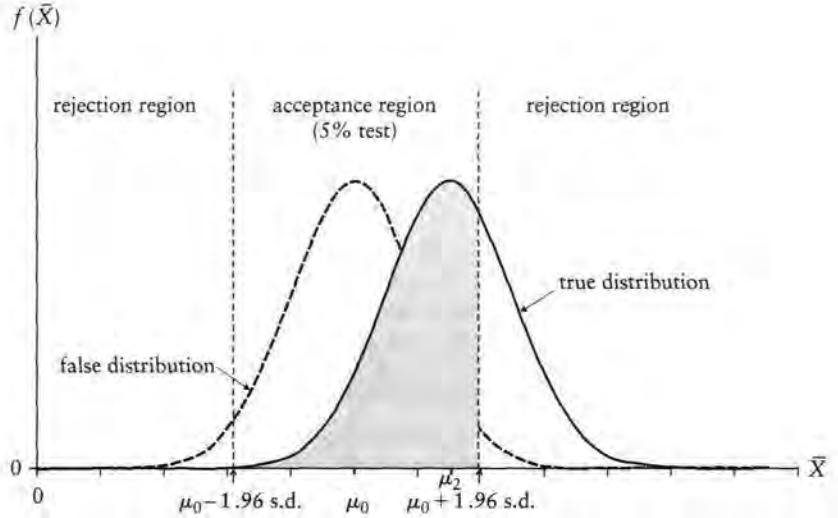


Figure R.19 Probability of making a Type II error if $\mu = \mu_2$, 5 percent test

Figure R.20 plots the power of a 5 percent significance test as a function of the distance separating the actual value of μ and μ_0, measured in terms of the standard deviation of the distribution of $\bar{X}$. As is intuitively obvious, the greater is the discrepancy, the greater is the probability of H_0: $\mu = \mu_0$ being rejected.

We now return to the original value of μ_1, shown in Figure R.17, and again consider the case where H_0: $\mu = \mu_0$ is false and H_1: $\mu = \mu_1$ is true. What difference does it make if we perform a 1 percent test, instead of a 5 percent test? The acceptance region is as shown in Figure R.21. The probability of $\bar{X}$ lying in this region, given that it is actually distributed with mean μ_1, is shown in Figure R.22.

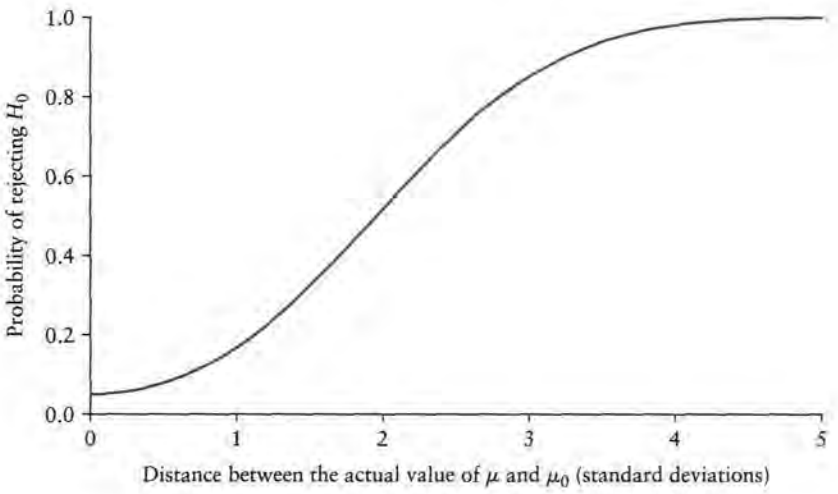


Figure R.20 Power function, 5% significance test

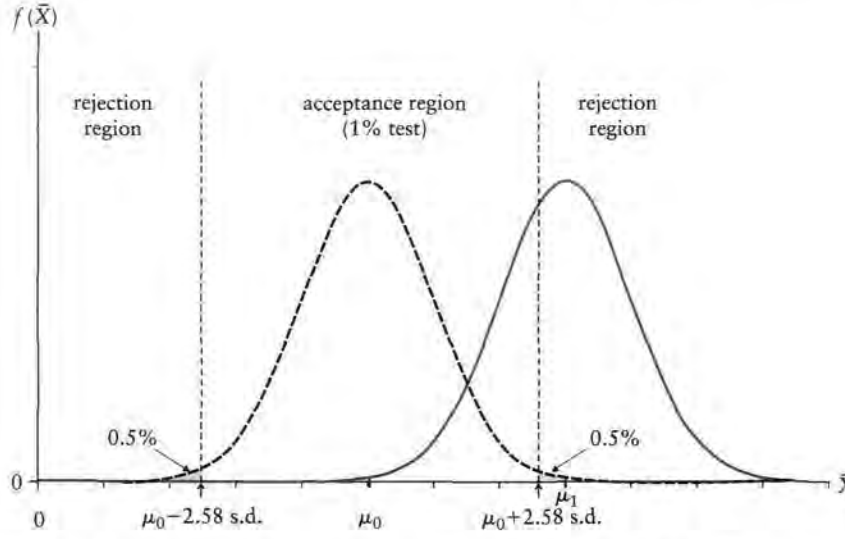


Figure R.21 Acceptance and rejection regions, conditional on H_0: $\mu = \mu_0$, 1 percent test

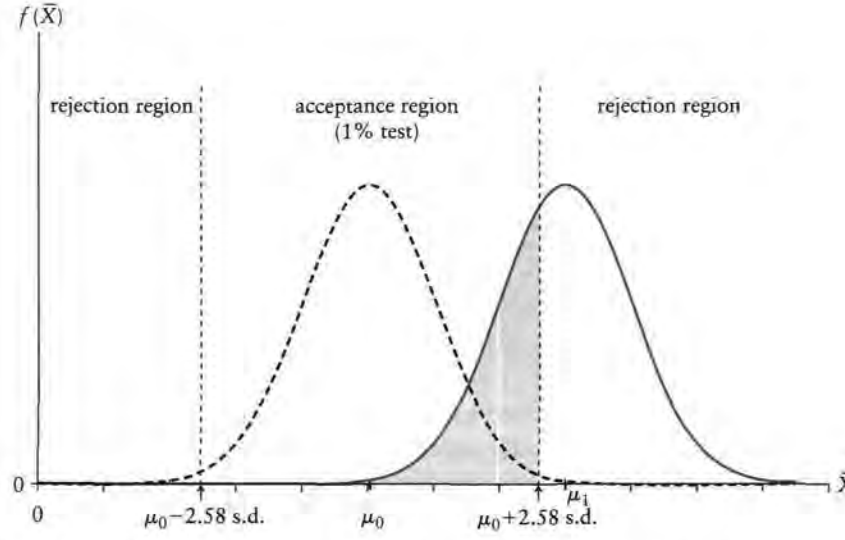


Figure R.22 Probability of making a Type II error if $\mu = \mu_1$, 1 percent test

It is larger than before, 0.34. The vertical white line within it shows the limit of the area in the case of the 5 percent test in Figure R.18.

Thus, we see that there is a trade-off between risk of Type I error and risk of Type II error. If we perform a 1 percent test instead of a 5 percent test, and H_0 is true, the risk of mistakenly rejecting it (and therefore committing a Type

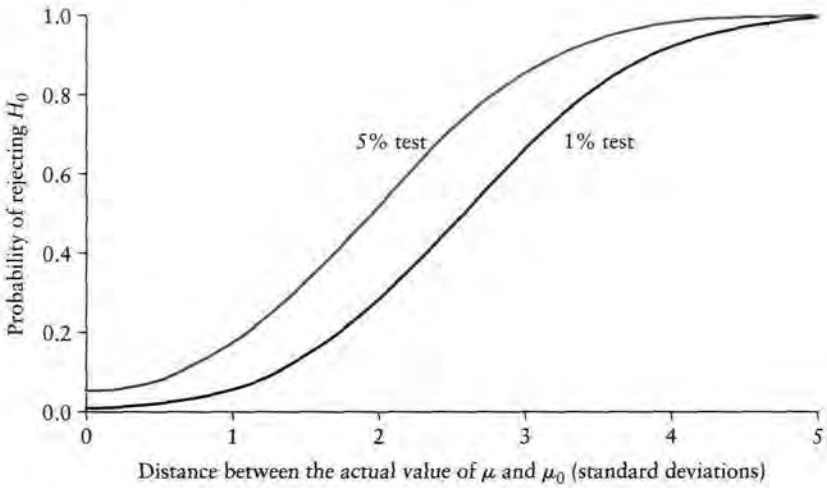


Figure R.23 Comparison of the power of 5 percent and 1 percent tests

I error) is only 1 percent instead of 5 percent. However, if H_0 happens to be false, the probability of not rejecting it (and therefore committing a Type II error) is larger.

How much larger? This is not fixed. It depends on the distance between μ_0 and μ_1, measured in terms of standard deviations. In this particular case, comparing Figures R.18 and R.22, it has increased from 0.15 to 0.34, so it has about doubled. To generalize, we plot the power functions for the 5 percent and 1 percent tests, shown in Figure R.23.

In applied economics, tests are typically performed at either the 5 percent or the 1 percent level. How do we choose between them? Compared with a 5 percent test, a 1 percent test involves a lower risk of a Type I error if the null hypothesis is true, but it has a greater risk of a Type II error if the null hypothesis is false. It is common to take out an insurance policy and perform the test at both of these levels, being prepared to quote the results of each. Actually, as we have already noted, it is frequently superfluous to quote both results explicitly. If you reject at 1 percent, there is no need to mention the 5 percent test. If you do not reject at 5 percent, there is no need to mention the 1 percent test. Both tests should be mentioned only if you reject at 5 percent but not at 1 percent.

EXERCISES

2.24 Give more examples of everyday instances in which decisions involving possible Type I and Type II errors may arise.

2.25 Before beginning a certain course, 36 students are given an aptitude test. The scores and the course results (pass/fail) are given below:

student	*test score*	*course result*	*student*	*test score*	*course result*	*student*	*test score*	*course result*
1	30	fail	13	26	fail	25	9	fail
2	29	pass	14	43	pass	26	36	pass
3	33	fail	15	43	fail	27	61	pass
4	62	pass	16	68	pass	28	79	fail
5	59	fail	17	63	pass	29	57	fail
6	63	pass	18	42	fail	30	46	pass
7	80	pass	19	51	fail	31	70	fail
8	32	fail	20	45	fail	32	31	pass
9	60	pass	21	22	fail	33	68	pass
10	76	pass	22	30	pass	34	62	pass
11	13	fail	23	40	fail	35	56	pass
12	41	pass	24	26	fail	36	36	pass

Do you think that the aptitude test is useful for selecting students for admission to the course, and if so, how would you determine the pass mark? (Discuss the trade-off between Type I and Type II errors associated with the choice of pass mark.)

R.26* Show that, in Figures R.18 and R.22, the probabilities of a Type II error are 0.15 in the case of a 5 percent significance test and 0.34 in the case of a 1 percent test. Note that the distance between μ_0 and μ_1 is three standard deviations. Hence the right-hand 5 percent rejection region begins 1.96 standard deviations to the right of μ_0. This means that it is located 1.04 standard deviations to the left of μ_1. Similarly, for a 1 percent test, the right-hand rejection region starts 2.58 standard deviations to the right of μ_0, which is 0.42 standard deviations to the left of μ_1.

R.27* Explain why the difference in the power of a 5 percent test and a 1 percent test becomes small when the distance between μ_0 and μ_1 becomes large.

R.11 *t* tests

Thus far, we have assumed that the standard deviation of $\bar{X}$ is known, which is most unlikely in practice. It has to be estimated. If the variance of the probability distribution of X is σ^2, the variance of the probability distribution of $\bar{X}$ is

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \qquad \textbf{(R.69)}$$

for a sample of size n. We estimate this variance as s_X^2/n, where

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2. \qquad \textbf{(R.70)}$$

Our estimator of the standard deviation of the probability distribution of $\bar{X}$ is therefore

$$\text{s.e.}(\bar{X}) = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2}. \tag{R.71}$$

We describe this as the standard error of $\bar{X}$, for short. The fact that we have to estimate the standard deviation with the standard error causes two modifications to the test procedure. First, $s_{\bar{X}}$ replaces $\sigma_{\bar{X}}$ in the test statistic. The test statistic

$$z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} \tag{R.72}$$

becomes the t statistic

$$t = \frac{\bar{X} - \mu_0}{\text{s.e.}(\bar{X})}. \tag{R.73}$$

Second, the t statistic has a distribution that is more complex than the normal distribution of the z statistic. Like the z statistic, the t statistic has a random component, $\bar{X}$, in the numerator. It also has a random component in the denominator. $s_{\bar{X}}^2$ provides only an estimate of $\sigma_{\bar{X}}^2$ and, as is obvious from (R.70), it depends on the actual values of the X_i in the sample and so will vary from sample to sample. Hence $\text{s.e.}(\bar{X})$ will vary from sample to sample. Both the numerator and the denominator of the test statistic contain random elements and as a consequence its distribution is no longer normal, even if X itself has a normal distribution. Its distribution is known as a t distribution.

There is a further complication. The t distribution is actually a family of distributions varying with what is known as the number of degrees of freedom in the sample. This is equal to $n - 1$ in the present context. We will not go into the reasons for this, or even describe the t distribution mathematically. Suffice to say that the t distribution is a relative of the normal distribution, its exact shape depending on the number of degrees of freedom in the regression, and that it approximates the normal distribution increasingly closely as the number of degrees of freedom becomes large.

Figure R.24 compares the distributions for 5, 10, and 20 degrees of freedom with that of the normal distribution. It will be seen that differences are noticeable only at the mode and in the tails. The mode of the t distribution is lower than that of the normal distribution and the tails are thicker, the differences being greater, the smaller the number of degrees of freedom. The modal differences are of no consequence. However, the difference in the shape of the tails is important because this is where the rejection regions for tests are defined. If we have decided to perform a two-sided 5 percent significance test, the 2.5 percent tails start 1.96 standard deviations from the mean in the case of a normal distribution. In the case of a t distribution with 10 degrees of freedom, the extra thickness of the tails means that the 2.5 percent tails are reached 2.23 standard

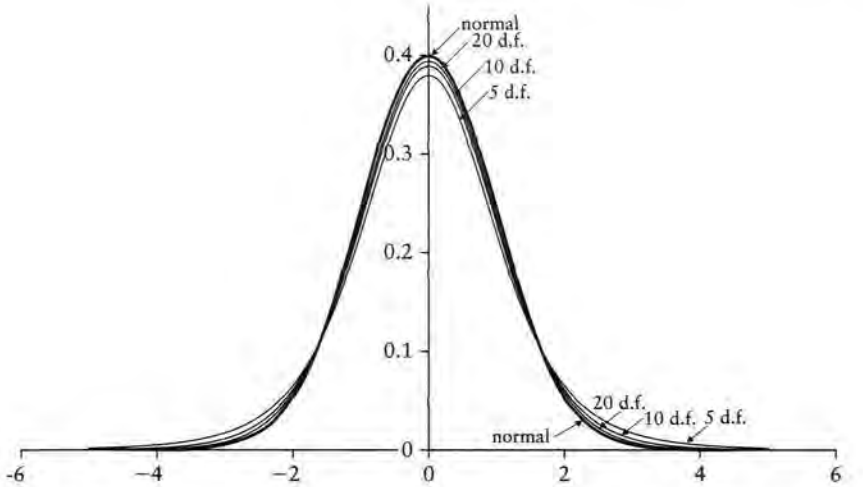


Figure R.24 Normal distribution and *t* distributions with 5, 10, and 20 degrees of freedom

deviations from the mean. This is shown in Figure R.25. The tail of the normal distribution has lighter shading. That of the *t* distribution has darker shading. Of course, they overlap.

Table A.2 in Appendix A gives the critical values of *t* cross-classified by significance level and the number of degrees of freedom. At the top of the table are listed possible significance levels for a test. For the time being, we are performing two-sided tests, so ignore the line for one-sided tests. Thus, the critical values for 5 percent tests are given in the second column and those for 1 percent tests in the fourth. The left-hand vertical column lists degrees of freedom. So, if we were performing a

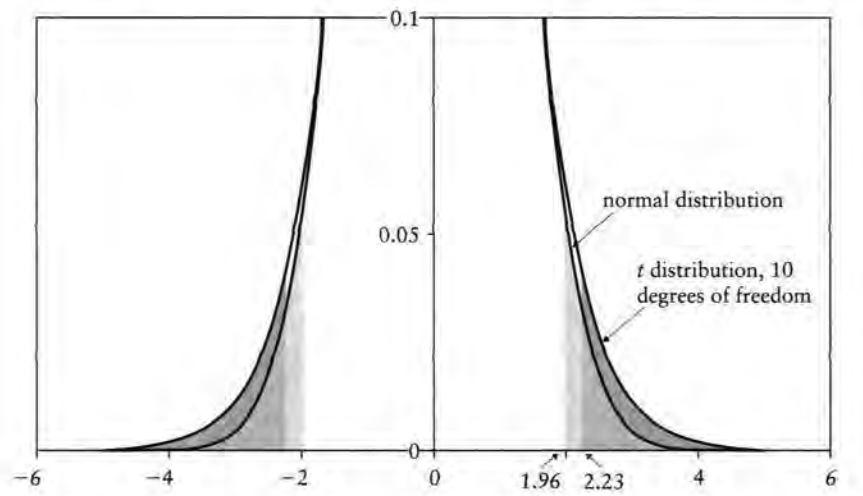


Figure R.25 2.5 percent tails of a normal distribution and a *t* distribution with 10 degrees of freedom

5 percent test with 10 degrees of freedom, the critical value of t, in absolute terms, would be 2.23, as indicated in Figure R.25. The critical value declines as the number of degrees of freedom increases, approaching a limiting value as the t distribution converges to a normal distribution. Thus, in the case of a 5 percent test, the limiting value is 1.96, and in the case of a 1 percent test, 2.58.

The critical value of t, which we will denote t_{crit}, replaces the number 1.96 in the test procedure for a 5 percent test. The t test consists of comparing the t statistic with t_{crit}. The condition that a sample mean should not lead to the rejection of a null hypothesis H_0: $\mu = \mu_0$ is

$$-t_{crit} \leq \frac{\bar{X} - \mu_0}{\text{s.e.}(\bar{X})} \leq t_{crit}. \tag{R.74}$$

Hence we have the decision rule: reject H_0 if $|t| > t_{crit}$, do not reject if $|t| \leq t_{crit}$, where

$$|t| = \left| \frac{\bar{X} - \mu_0}{\text{s.e.}(\bar{X})} \right| \tag{R.75}$$

is the absolute value of t (its numerical value, neglecting the sign). This makes a difference only for relatively small samples, say fewer than 50 observations. If the sample size is larger than 50, the fact that we have to estimate the standard error of $\bar{X}$ makes negligible difference because the t distribution converges on the normal distribution and the critical values of t converge on those for the normal distribution.

Example

A certain city abolishes its local sales tax on consumer expenditure. A survey of 20 households shows that, in the following month, mean household expenditure increased by \$160 and the standard error of the increase was \$60. We wish to determine whether the abolition of the tax had a significant effect on household expenditure. We take as our null hypothesis that there was no effect: H_0: $\mu = 0$. The test statistic is

$$t = \frac{160 - 0}{60} = 2.67. \tag{R.76}$$

The critical values of t with 19 degrees of freedom are 2.09 at the 5 percent significance level and 2.86 at the 1 percent level. Hence, we reject the null hypothesis of no effect at the 5 percent level but not at the 1 percent level.

The reject/fail-to-reject terminology

In this section, it has been shown that you should reject the null hypothesis if the absolute value of the t statistic is greater than t_{crit}, and that you fail to reject

it otherwise. Why 'fail to reject', which is a clumsy expression? Would it not be better just to say that you accept the hypothesis if the absolute value of the t statistic is less than t_{crit}?

The argument against using the term 'accept' is that you might find yourself 'accepting' several mutually exclusive hypotheses at the same time. For instance, in the sales tax example, a null hypothesis H_0: $\mu = 100$ would not be rejected, even at the 5 percent level, because the t statistic

$$t = \frac{200 - 100}{80} = 1.25 \qquad \text{(R.77)}$$

is lower than 2.09. But the same would be true if the null hypothesis were $\mu = 150$ or $\mu = 250$, with corresponding t statistics 0.63 and −0.63. It is logical to say that you would not reject any of these null hypotheses, but it makes little sense to say that you simultaneously accept all three of them. In the next section you will see that one can define a whole range of hypotheses which would not be rejected by a given experimental result, so it would be incautious to pick out one as being 'accepted'.

EXERCISES

R.28* A researcher is evaluating whether an increase in the minimum hourly wage has had an effect on employment in the manufacturing industry in the following three months. Taking a sample of 25 firms, what should she conclude if

(a) the mean decrease in employment is 9 percent, and the standard error of the mean is 5 percent;
(b) the mean decrease is 12 percent, and the standard error is 5 percent;
(c) the mean decrease is 20 percent, and the standard error is 5 percent;
(d) there is a mean *increase* of 10 percent, and the standard error is 5 percent?

R.29 A drug company asserts that its course of treatment will, on average, reduce a person's cholesterol level by 0.8 mmol/L. A researcher undertakes a trial with a sample of 30 individuals. What should he report if he obtains the following results:

(a) a mean increase of 0.6 units, with standard error 0.2 units;
(b) a mean decrease of 0.4 units, with standard error 0.2 units;
(c) a mean increase of 0.4 units, with standard error 0.2 units?

R.12 Confidence intervals

Thus far, we have been assuming that the hypothesis preceded the empirical investigation. In particular, given a random variable X with unknown population mean μ, we set up a null hypothesis H_0: $\mu = \mu_0$, obtained a sample of n observations, calculated $\bar{X}$, and checked whether it caused H_0 to be rejected or not. A little more systematically, we established the range of sample values

of $\bar{X}$ that would not lead to the rejection of H_0. We called this the acceptance region.

In practice, in empirical work, it is much more common to do the opposite. We conduct an experiment and then consider what hypotheses would be compatible with it, in the sense of not being rejected by it at a chosen significance level. For the moment, to be specific, we will adopt the 5 percent significance level. We will generalize from this in due course.

In Figure R.26, we suppose that we have calculated $\bar{X}$ from a sample of observations and we are considering whether a hypothesis $\mu = \mu_0$ would be rejected by it. To determine this, we need to draw the distribution of $\bar{X}$ conditional on H_0 being true, and this is shown. For the moment, we are assuming for simplicity that we know the value of σ and thus $\sigma/\sqrt{n}$, the standard deviation of the distribution. We can see that, in this case, the hypothesis would not be rejected, given the value of $\bar{X}$.

Next we will consider the hypothesis $\mu = \mu_1$. This is shown in Figure R.27, together with the conditional distribution for $\bar{X}$. In this case, $\bar{X}$ lies in the left rejection region of the conditional distribution, so $\bar{X}$ and μ_1 are incompatible and the hypothesis $\mu = \mu_1$ is rejected.

Now we ask ourselves whether we can generalize and establish the entire range of hypotheses that would not be rejected, given the experimental outcome $\bar{X}$, and given our choice of significance level. In particular, we will ask ourselves what is the maximum hypothetical value of μ that would not be rejected, given $\bar{X}$. We will denote it μ^{max}. The answer is shown graphically in Figure R.28, for the case where we have chosen a 5 percent significance level. The figure shows the probability distribution of $\bar{X}$, conditional on $\mu = \mu^{max}$ being true. $\bar{X}$ lies just on the edge of the lower rejection region. Any null hypothesis greater than μ^{max}

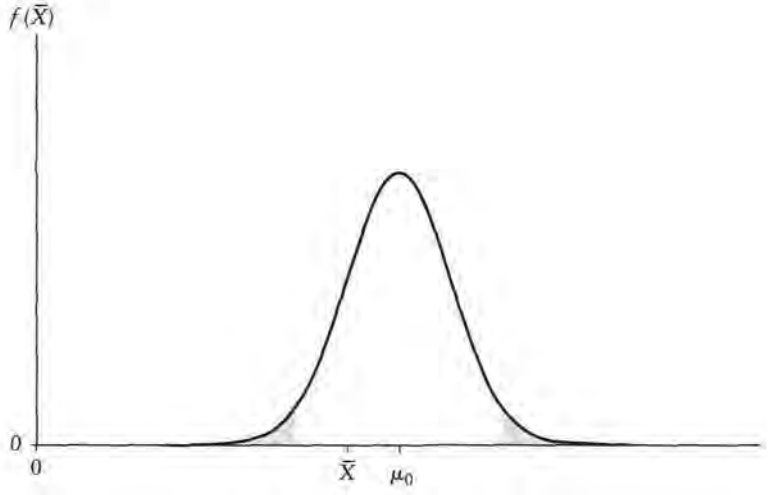


Figure R.26 Distribution of $\bar{X}$ conditional on $\mu = \mu_0$

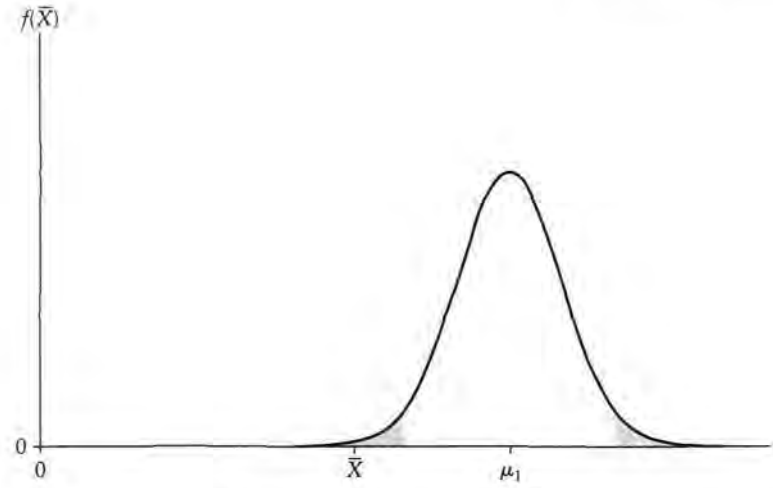


Figure R.27 Distribution of $\bar{X}$ conditional on $\mu = \mu_1$

would be rejected because $\bar{X}$ would lie in the left rejection region for the conditional distribution, as in the case of μ_1 in Figure R.27.

How do we determine μ^{max}? From the geometry of Figure R.28, we can see that the distance from the middle of the distribution, μ^{max}, to the limit of the left rejection region is $1.96\sigma_{\bar{X}}$. Hence,

$$\bar{X} = \mu^{max} - 1.96\sigma_{\bar{X}} \tag{R.78}$$

and so

$$\mu^{max} = \bar{X} + 1.96\sigma_{\bar{X}}. \tag{R.79}$$

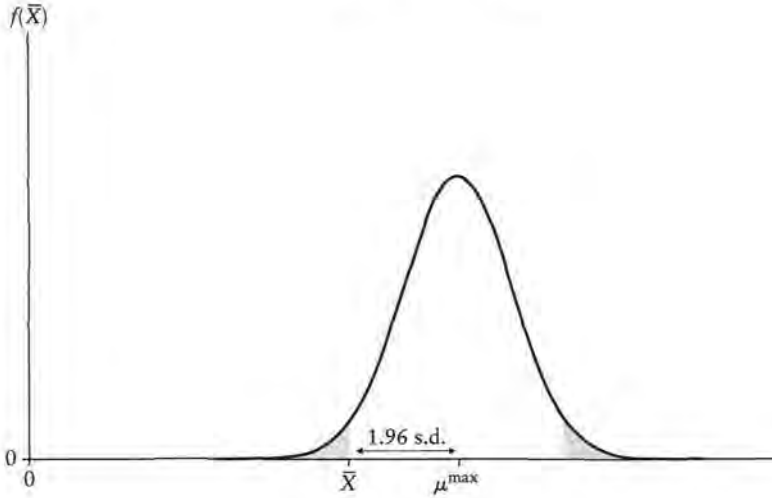


Figure R.28 Distribution of $\bar{X}$ conditional on $\mu = \mu^{max}$

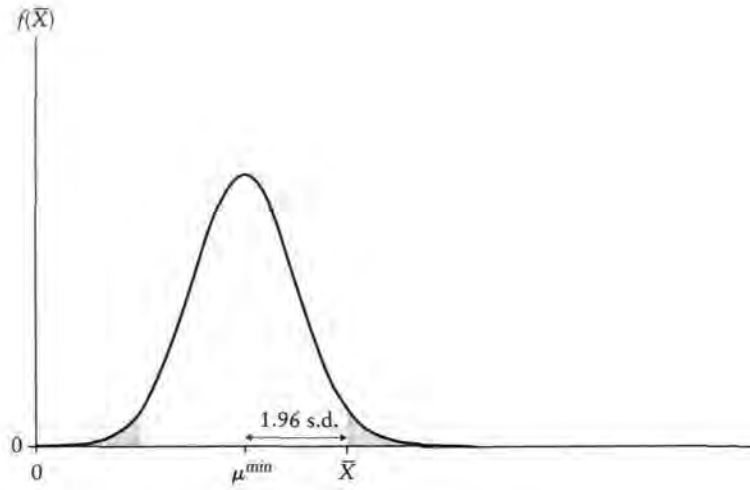


Figure R.29 Distribution of $\bar{X}$ conditional on $\mu = \mu^{min}$

So far we have been considering hypothetical values of μ that are greater than $\bar{X}$. We should also consider hypothetical values of μ lower than $\bar{X}$. Those that are close enough to $\bar{X}$ will not be rejected by it and those that are too far below it will be rejected. In the same way that we found a limiting value for μ above $\bar{X}$, we can find a limiting value below it. We will call this μ^{min}. It is the value which, given its conditional distribution for $\bar{X}$, is just not rejected by $\bar{X}$. This is shown in Figure R.29.

To determine μ^{min}, we note that the distance from the middle of the distribution, μ^{min}, to the limit of the left rejection region is $1.96\sigma_{\bar{X}}$. Hence,

$$\bar{X} = \mu^{min} + 1.96\sigma_{\bar{X}} \tag{R.80}$$

and so

$$\mu^{min} = \bar{X} - 1.96\sigma_{\bar{X}}. \tag{R.81}$$

The set of all hypothetical values of μ not rejected by $\bar{X}$ is known as the confidence interval for μ, in this case the 95 percent confidence interval since we have been using the 5 percent significance level. (It is conventional to designate the confidence interval as 100 minus the significance level. The reason for this is explained in Box R.3.) The confidence interval therefore comprises all values of μ from μ^{min} to μ^{max}. Given equations (R.79) and (R.81), the 95 percent confidence interval is

$$\bar{X} - 1.96\sigma_{\bar{X}} \leq \mu \leq \bar{X} + 1.96\sigma_{\bar{X}}. \tag{R.82}$$

So far, we have assumed that we know $\sigma_{\bar{X}}$, the standard deviation of the distribution of $\bar{X}$. Of course, in practice we have to estimate it. When we do this,

we must replace 1.96, which applies to the normal distribution, with the corresponding critical value of the t distribution. Hence, in practice, the 95 percent confidence interval is computed as

$$\bar{X} - t_{\text{crit, 5\%}} \times \text{s.e.}(\bar{X}) \leq \mu \leq \bar{X} + t_{\text{crit, 5\%}} \times \text{s.e.}(\bar{X}), \qquad \text{(R.83)}$$

where $t_{\text{crit,5\%}}$ is the critical value of t at the 5 percent level, given the number of degrees of freedom.

We can immediately generalize from this to other significance levels. For example, the 99 percent confidence interval is given by

$$\bar{X} - t_{\text{crit, 1\%}} \times \text{s.e.}(\bar{X}) \leq \mu \leq \bar{X} + t_{\text{crit, 1\%}} \times \text{s.e.}(\bar{X}), \qquad \text{(R.84)}$$

where $t_{\text{crit,1\%}}$ is the critical value of t at the 1 percent significance level.

BOX R.3 A second interpretation of a confidence interval

When you construct a confidence interval, the numbers you calculate for its upper and lower limits have random components. For example, in inequality (R.83), the lower and upper limits are

$$\bar{X} - t_{\text{crit, 5\%}} \times \text{s.e.}(\bar{X}) \text{ and } \bar{X} + t_{\text{crit, 5\%}} \times \text{s.e.}(\bar{X}).$$

Both $\bar{X}$ and s.e. $(\bar{X})$ are random quantities that depend on the actual values of the X_i in the observations in the sample. One hopes that the confidence interval will include the true value of μ, but sometimes it will be so distorted by the random element that it will fail to do so. This may happen if the sample contains an unusually large number of high or low values of X.

What is the probability that a confidence interval will capture the true value of μ? It can easily be shown, using elementary probability theory, that, in the case of a 95 percent confidence interval, the probability is 95 percent. Similarly, in the case of a 99 percent confidence interval, the probability is 99 percent.

$\bar{X}$ provides a point estimate of μ, but of course the probability of the true value being exactly equal to this estimate is infinitesimal. The confidence interval provides what is known as an *interval estimate* of μ, that is, a range of values that will include the true value with a high, predetermined probability. It is this interpretation that gives the confidence interval its name. The proof is left as an exercise. It is also the reason why, for example, we talk about a 95 percent confidence interval rather than a 5 percent one.

Example

In an example in Section R.11, when a local sales tax was abolished, a survey of 20 households showed that mean household expenditure increased by \$160 and the standard error of the increase was \$60. The 95 percent confidence interval for the effect is

$$160 - 2.09 \times 60 \leq \mu \leq 160 + 2.09 \times 60 \qquad \text{(R.85)}$$

since the critical value of t with 19 degrees of freedom is 2.09 at the 5 percent significance level. Hence, the interval is

$$35 \leq \mu \leq 285. \tag{R.86}$$

EXERCISES

R.30 Determine the 99 percent confidence interval for the effect of the sales tax in the example.

R.31 Determine the 95 percent confidence interval for the effect of an increase in the minimum wage on employment, given the data in Exercise R.28, for each part of the exercise. How do these confidence intervals relate to the results of the t tests in that exercise?

R.32 Demonstrate that the 95 percent confidence interval defined by equation (R.83) has a 95 percent probability of capturing μ_0 if H_0: $\mu = \mu_0$ is true.

R.13 One-sided tests

In Section R.10 we saw that there is a trade-off between the potential for Type I and Type II errors when performing tests of hypotheses. You can reduce the risk of a Type I error, if the null hypothesis is true, by performing a 1 percent test instead of a 5 percent test. However, if the null hypothesis is false, you thereby increase the risk of making a Type II error. In this section, we will see that we can improve the terms of this trade-off if we are in a position to perform a one-sided test instead of a two-sided test. There will still be a trade-off, but it will be a better one. Holding the risk of making a Type I error constant (if the null hypothesis is true), we will have a smaller risk of making a Type II error (if the null hypothesis is false), if we use a one-sided test instead of a two-sided test.

First, we have to explain what is meant by a one-sided test.

In our discussion of t tests, we started out with our null hypothesis H_0: $\mu = \mu_0$ and tested it to see whether we should reject it or not, given the sample value of $\bar{X}$. Thus far, the alternative hypothesis has been merely the negation of the null hypothesis. However, if we are able to be more specific about the alternative hypothesis, we may be able to improve the testing procedure. We will investigate three cases: first, the very special case where, if $\mu \neq \mu_0$, there is only one possible alternative value, which we will denote μ_1; second, where, if μ is not equal to μ_0, it must be greater than μ_0; and third, where, if μ is not equal to μ_0, it must be less than μ_0.

One-sided tests are used very frequently in hypothesis testing. In the context of regression analysis, they are, or they ought to be, more common than the traditional textbook two-sided tests. It is therefore important that you understand the rationale for their use, and this involves a sequence of small analytical steps. None of this should present any difficulty, but you should avoid the

temptation to try to reduce the whole business to the mechanical use of a few formulae.

$\mathbf{H_0}$: $\mu = \mu_0$, $\mathbf{H_1}$: $\mu = \mu_1$

In this case, there are only two possible values of μ, μ_0 and μ_1. For the sake of argument, we will assume for the time being that μ_1 is greater than μ_0.

Suppose that we wish to test H_0 at the 5 percent significance level, and we follow the usual procedure discussed in Section R.9. We locate the limits of the upper and lower 2.5 percent tails under the assumption that H_0 is true, indicated by A and B in Figure R.30, and we reject H_0 if $\bar{X}$ lies to the left of A or to the right of B.

Now, if $\bar{X}$ does lie to the right of B, it is more compatible with H_1 than with H_0; the probability of it lying to the right of B is greater if H_1 is true than if H_0 is true. We should have no hesitation in rejecting H_0. We therefore conclude that H_1 is true.

However, if $\bar{X}$ lies to the left of A, the test procedure will lead us to a perverse conclusion. It tells us to reject H_0, and therefore conclude that H_1 is true, even though the probability of $\bar{X}$ lying to the left of A is smaller (in this case, negligible) if H_1 is true. We have not even drawn the probability density function that far for H_1. If such a value of $\bar{X}$ occurs only once in a million times when H_1 is true, but 2.5 percent of the time when H_0 is true, it is more logical to assume that H_0 is true. Of course, once in a million times you will make a mistake, but the rest of the time you will be right.

Hence we will reject H_0 only if $\bar{X}$ lies in the upper 2.5 percent tail, that is, to the right of B, as in Figure R.31. We are now performing a one-sided test, and we have reduced the probability of making a Type I error to 2.5 percent. Since

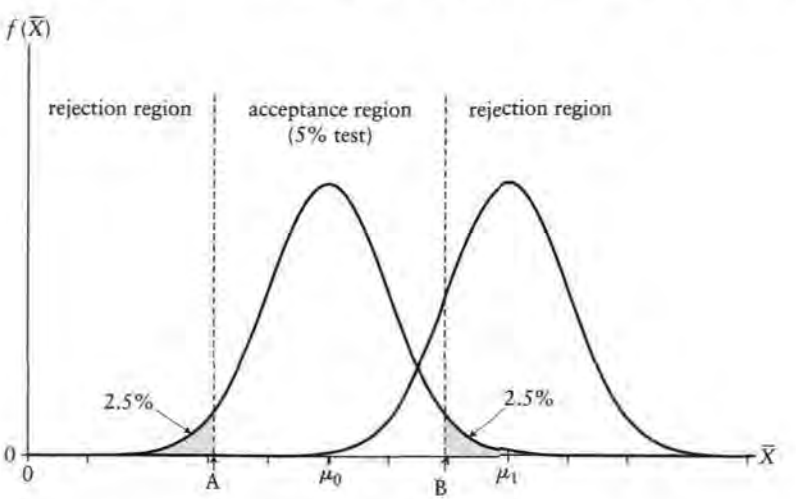


Figure R.30 Distribution of $\bar{X}$ under H_0 and H_1

the significance level is defined to be the probability of making a Type I error, it is now also 2.5 percent.

What difference does this make to the probability of a Type II error? None. A Type II error will occur if the null hypothesis is false (so $\mu = \mu_1$) and $\bar{X}$ lies in the acceptance region for H_0. This is the area under the curve for $\mu = \mu_1$ to the left of the point B. It is shown for the two-sided test in Figure R.32 and for the one-sided test in Figure R.33. Of course, it is the same area. The actual area depends on the distance between μ_0 and μ_1, in terms of standard deviations. In the present example, the probability is 0.15.

The one-sided test is a 2.5 percent significance test because the probability of rejecting H_0: $\mu = \mu_0$, if is true, is 2.5 percent. If we wish to perform a 5 percent

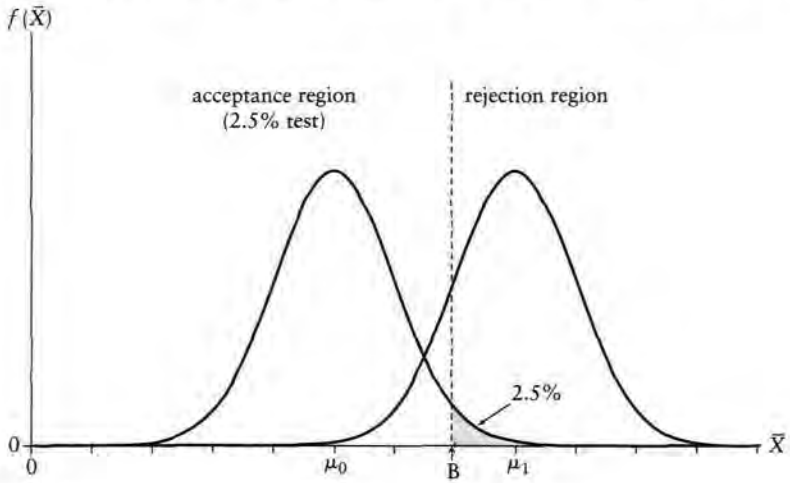


Figure R.31 Rejection region, one-sided test, 2.5% significance level

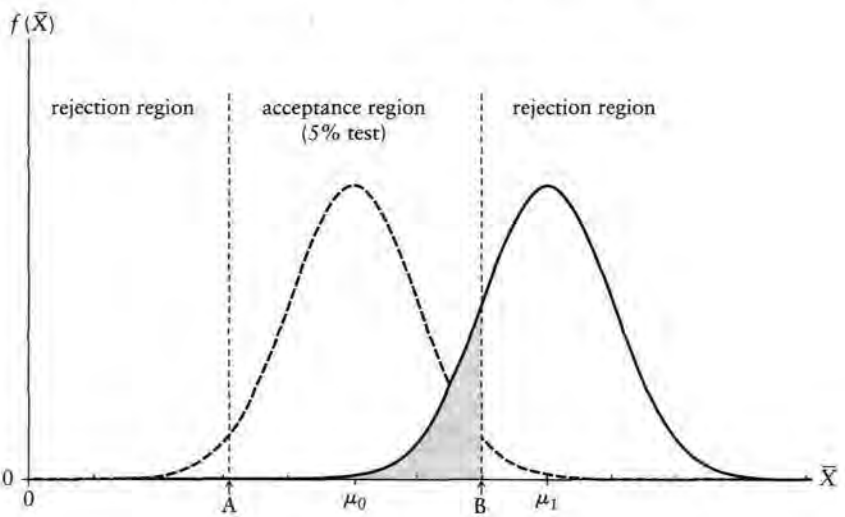


Figure R.32 Probability of Type II error if $\mu = \mu_1$, two-sided test, 5% significance level

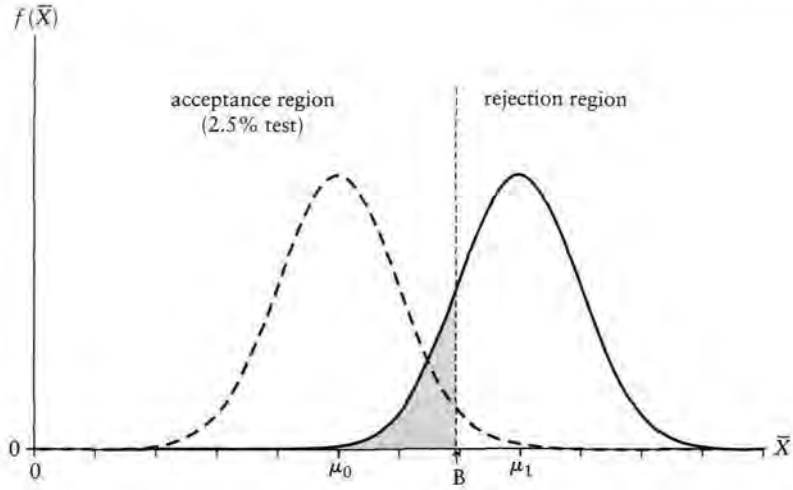


Figure R.33 Probability of Type II error if $\mu = \mu_1$, one-sided test, 2.5% significance level

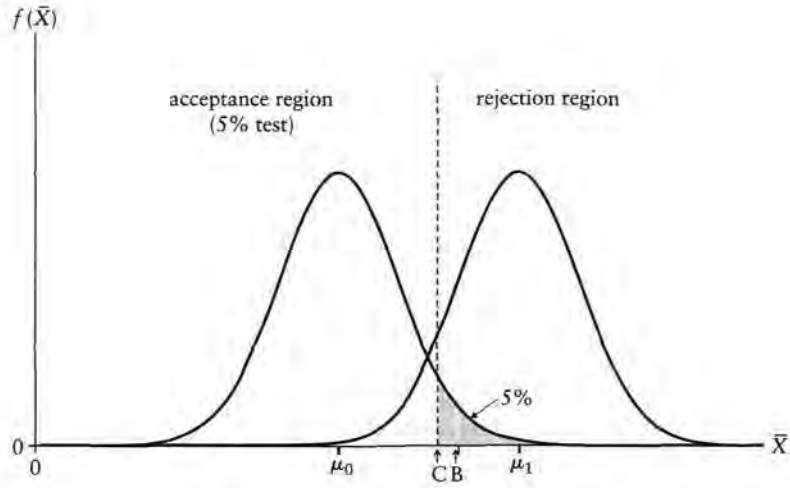


Figure R.34 Rejection region, one-sided test, 5 percent significance level

significance test, we need to increase the right-hand rejection region so that it includes 5 percent of the area under the curve conditional on H_0: $\mu = \mu_0$. This is shown in Figure R.34. The beginning of the rejection region is marked by the point C.

Why should we wish to increase the risk of a Type I error in this way? The answer is that we thereby reduce the probability of a Type II error if H_0 is false.

Suppose that, in fact, $\mu = \mu_1$. We will fail to reject H_0 if $\bar{X}$ lies in the acceptance region for H_0, that is, if it lies to the left of C. The probability of this happening, if $\mu = \mu_1$, is the area under the curve for $\mu = \mu_1$ to the left of the point C. This is

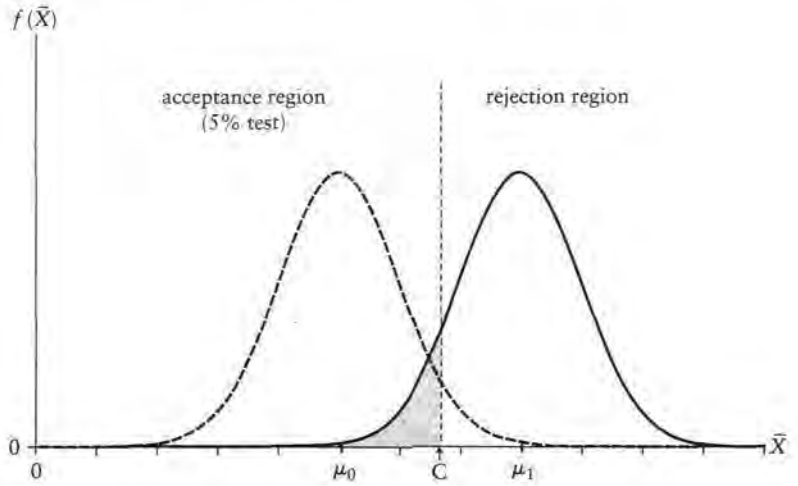


Figure R.35 Probability of Type II error if $\mu = \mu_1$, one-sided test, 5% significance level

shown in Figure R.35. In the present example, it is 0.09. It is smaller than the probability in the case of the two-sided test, 0.15, illustrated in Figure R.33. This is because the point C is to the left of the point B. It is 1.65 standard deviations from μ_0, while B is 1.96 standard deviations from μ_0.

One percent significance test

The discussion thus far has assumed that we are performing a 5 percent significance test. Suppose, instead, that we wished to perform a 1 percent test. What is the advantage of performing a one-sided test in this case, if we know that μ must be equal to μ_1 if the null hypothesis H_0: $\mu = \mu_0$ is false? Figure R.36 shows the probability of failing to reject H_0 when $\mu = \mu_1$ in the case of a two-sided test, where the rejection region is 0.5 percent in both tails. The points A and B mark the limits of the rejection regions. They are 2.58 standard deviations from μ_0. The probability of $\bar{X}$ lying in the acceptance region is the area under the distribution for $\mu = \mu_1$ in the range AB. The probability of a Type II error is 0.34 in the present example.

Figure R.37 shows the probability in the case of a one-sided test. We have eliminated the left tail because, as in the case of the 5 percent test, it is irrational to retain it, and we have increased the right tail to from 0.5 percent to 1 percent. The limit of the right-hand 1 percent tail, marked by the point C, is 2.33 standard deviations from μ_0. The probability of $\bar{X}$ lying in the acceptance region is now the area under the distribution for $\mu = \mu_1$ to the left of C. It is 0.25, in the present example.

Table R.6 summarizes the trade-off between the probabilities of making a Type I error, if H_0 is true, and a Type II error, if H_0 is false, and how the trade-off is altered by performing a one-sided test instead of a two-sided test.

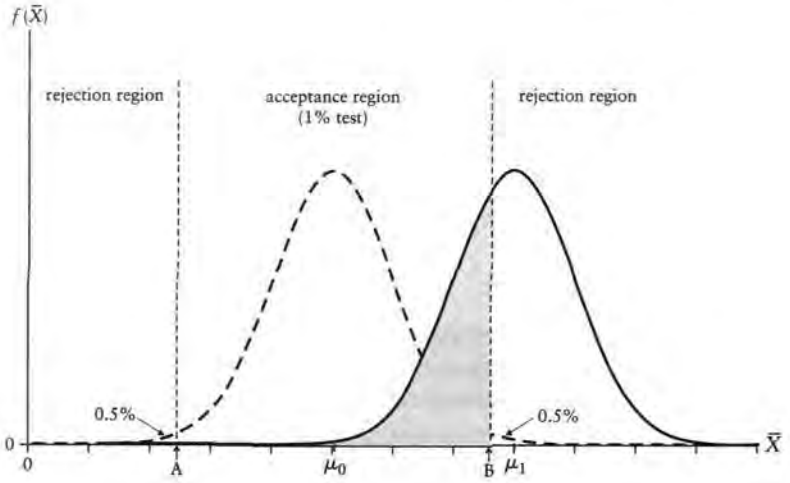


Figure R.36 Probability of Type II error if $\mu = \mu_1$, two-sided test, 1% significance level

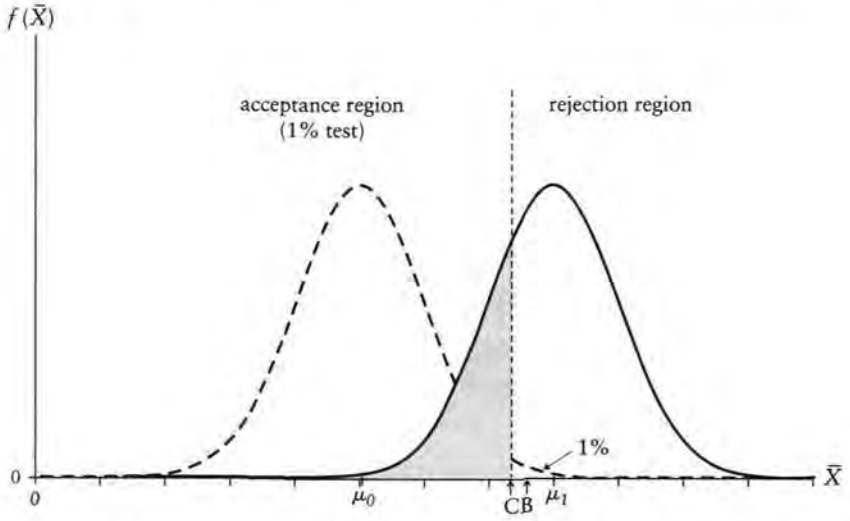


Figure R.37 Probability of Type II error if $\mu = \mu_1$, one-sided test, 1% significance level

Table R.6 Trade-off between Type I and Type II errors, one-sided and two-sided tests

	Probability of Type II error if $\mu = \mu_1$	
	One-sided test	Two-sided test
5 percent significance test	0.09	0.15
2.5 percent significance test	0.15	(not investigated)
1 percent significance test	0.25	0.34

The numbers in the table depend on the distance between μ_0 and μ_1, in terms of standard deviations. In the present example, we have assumed that the distance is three standard deviations. If the distance had been greater or smaller, the numbers would have been different, but the qualitative relationships would have been the same. First, as we saw in Section R.10, for a two-sided test, the smaller is the risk of making a Type I error if H_0 is true, the greater is the risk of a Type II error if H_0 is false. The same is true of one-sided tests. Second, holding constant the risk of a Type I error if H_0 is true, the risk of a Type II error is smaller for a one-sided test than for a two-sided test if H_0 is false. To put it in terms of significance level and power, for any given significance level, a one-sided test is more powerful than a two-sided test. This is illustrated in Figure R.38 for a 5 percent significance test.

Generalizing from H_0: $\mu = \mu_0$, H_1: $\mu = \mu_1$ to H_0: $\mu = \mu_0$, H_1: $\mu > \mu_0$

We have discussed the case in which the alternative hypothesis involved a *specific* hypothetical value μ_1, with μ_1 greater than μ_0. The logic did not depend on how much greater μ_1 was than μ_0. Hence, it also applies to the more general case where we know, or at least believe, that if $\mu \neq \mu_0$, it must be greater than μ_0, but do not know the exact amount. We rule out the possibility that $\mu < \mu_0$ on the basis of logic, theory, or experience. We would still wish to eliminate the left tail from the rejection region because a low value of $\bar{X}$ is more probable under H_0: $\mu = \mu_0$ than under H_1: $\mu > \mu_0$, and this would be evidence in support of H_0, not against it. Therefore, we would still prefer a one-sided test, using the right tail as the rejection region, to a two-sided test.

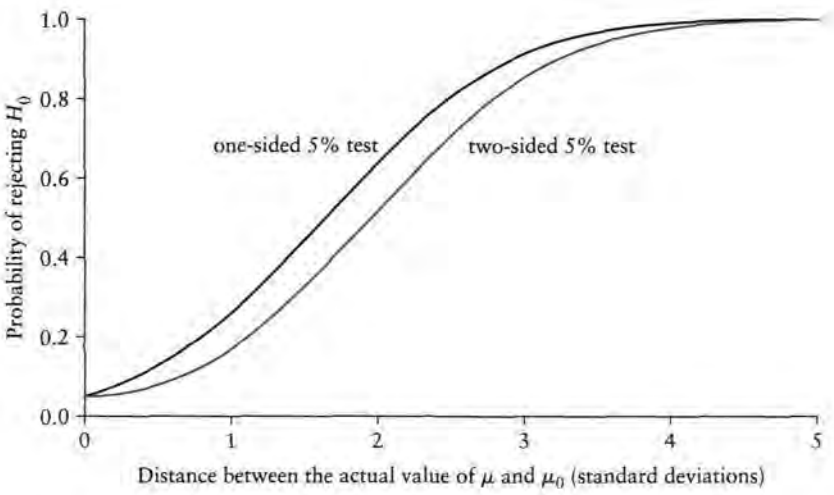


Figure R.38 Comparison of the power of one-sided and two-sided 5 percent tests

Note that, since the true value of μ is not defined specifically under H_1: $\mu > \mu_0$, we now have no way of calculating the probability of a Type II error if H_0 is false. However, we can still be sure that, for any given risk of a Type I error, if H_0 is true, the risk of a Type II error, if H_0 is false, will be lower for a one-sided test than for a two-sided test.

H_0: $\mu = \mu_0$, H_1: $\mu < \mu_0$

Now consider the case where you have two specific alternatives, H_0: $\mu = \mu_0$ and H_1: $\mu = \mu_1$, where μ_1 is less than μ_0. The logic in the first part of this section applies, with the difference that one eliminates the right tail instead of the left tail as a rejection region. If $\bar{X}$ happened to lie in the right tail, this would be even more unlikely under H_1 than under H_0, and so it should not lead to the rejection of H_0. Hence, one should perform a one-sided test, using only the left tail as a rejection region.

Further, as in the previous discussion, the distance between μ_1 and μ_0 makes no difference. It is sufficient that μ_1 should be less than μ_0. Hence, if we are in a position to rule out $\mu > \mu_0$ on the basis of logic, theory, or experience, we should perform a one-sided test in the case of the more general framework H_0: $\mu = \mu_0$, H_1: $\mu < \mu_0$.

One-sided *t* tests

In the discussion so far in this section, we have assumed that we know the standard deviation of $\bar{X}$. Of course, in practice, you have to estimate it, as described in Section R.11. This means that in all of the discussion in this section, we ought to be using t distributions rather than normal distributions, and critical values of t instead of critical values of the normal distribution if the sample size is relatively small, say fewer than 50 observations.

Special case: H_0: $\mu = 0$

One-sided tests are often particularly useful where the analysis relates to the evaluation of treatment and effect. Suppose that a number of units of observation receive some type of treatment and X_i is a measure of the effect of the treatment for observation i. To demonstrate that the treatment did have an effect, we set up the null hypothesis H_0: $\mu = 0$ and see if we can reject it, given the sample average $\bar{X}$. The t statistic for testing H_0 is

$$t = \frac{\bar{X} - \mu_0}{\text{s.e.}(\bar{X})} = \frac{\bar{X}}{\text{s.e.}(\bar{X})}. \qquad \textbf{(R.87)}$$

Suppose that we are performing a 5 percent significance test and we have 40 observations. With 39 degrees of freedom, the critical value of t is 2.02, using a

two-sided test. However, if we can rule out the possibility that the treatment can have a negative effect, we can use a one-sided test, for which the critical value is 1.68. This makes it easier to reject the null hypothesis and demonstrate that the treatment has had a significant effect. Similarly, if we were performing a 1 percent test, the critical values using a two-sided test and a one-sided test are 2.71 and 2.43, respectively. Thus, again, it is easier to reject the null hypothesis with a one-sided test. Consider the following cases for the two-sided test H_0: $\mu = 0$, H_1: $\mu \neq 0$ and the one-sided test H_0: $\mu = 0$, H_1: $\mu > 0$:

(a) $t = 1.20$. One-sided and two-sided tests lead to the same conclusion: H_0 is not rejected at any sensible significance level.
(b) $t = 1.80$. H_0 is not rejected at any sensible significance level using a two-sided test. However, if we use a one-sided test, it is rejected at the 5 percent level and we conclude that the treatment did have an effect, after all.
(c) $t = 2.20$. One-sided and two-sided tests lead to the same conclusion: H_0 is rejected at the 5 percent level (2.02, two-sided, 1.68, one-sided) but not at the 1 percent level (2.71 two-sided, 2.43 one-sided).
(d) $t = 2.60$. H_0 is rejected at the 5 percent level using a two-sided test and at the 1 percent level using a one-sided test.
(e) $t = 3.00$. One-sided and two-sided tests lead to the same conclusion: H_0 is rejected at the 1 percent level (but not at the 0.1 percent level).

In three of the five cases, using a one-sided test leads to the same conclusion. But in cases (b) and (d), there is an advantage in using a one-sided test.

Anomalous results

In this discussion, we have assumed that the treatment cannot have a negative effect. Suppose, however, that in our sample $\bar{X}$ turns out to be negative. What should we then conclude? Let us consider two further cases. Since $\bar{X}$ is negative, the t statistic will be negative.

(f) $t = -0.80$. We should maintain the null hypothesis of no effect. If it is true, the sample $\bar{X}$ will be randomly distributed around zero. In repeated samples, t will be positive half the time, and negative half the time. In this particular case, it is one of the negative times.
(g) $t = -2.40$. If we were performing a two-sided test, this would be significant at the 5 percent level. We would reject H_0 and conclude that the true value was negative. However, on the basis of experience or theory, we have excluded this possibility. We are considering only two possibilities: no effect, under H_0, and a positive effect, under H_1. Under H_0, it would be unusual to obtain a negative estimate with such a large negative t statistic. With 39 degrees of freedom, the probability is about 2 percent. However, the probability of obtaining such an anomalous result would be even lower if the true value of μ were positive. So we should stay with H_0 and conclude that we have a somewhat freakish sample.

This said, it would be reasonable to revisit our assumption that the treatment could not possibly have a negative effect. It might be that we have overlooked something and that we should not be so confident about excluding this possibility. If, in the end, we are sure that the effect cannot be negative, then we would stay with H_0. However, if we decide that, perhaps, we should be performing a two-sided test after all, then we could conclude that we have evidence, significant at the 5 percent level, that the effect is, indeed, negative.

Justification of the use of a one-sided test

The use of a one-sided test has to be justified beforehand on the grounds of logic, theory, common sense, or previous experience. When stating the justification, you should be careful not to exclude the possibility that the null hypothesis is true. For example, suppose that you are testing whether a treatment has an effect. To be specific, suppose that you are evaluating the effect of a training course on the productivity of a sample of workers. You would expect a significant positive effect, given a large enough sample. But your justification should not be that, on the basis of theory and common sense, the effect should be positive. This is too strong, for it eliminates the null hypothesis of no effect, and there is nothing left to test. Instead, you should say that, on the basis of theory and common sense, you would exclude the possibility that the course has a *negative* effect. This then leaves the possibility that the effect is zero and the alternative that it is positive.

Example

We return to the example of the evaluation of the effect on consumer expenditure of the abolition of a local sales tax. The survey of 20 households showed that mean household expenditure increased by \$160 and the standard error of the increase was \$60. The test statistic for H_0: $\mu = 0$ is

$$t = \frac{160 - 0}{60} = 2.67. \tag{R.88}$$

For a two-sided test, the critical values of t with 19 degrees of freedom are 2.09 at the 5 percent significance level and 2.86 at the 1 percent level, allowing the null hypothesis of no effect to be rejected at the 5 percent level but not at the 1 percent level. However, we should be able to rule out the possibility of the abolition of the sales tax causing a reduction in consumer expenditure, and so we are in a position to perform a one-sided test instead of a two-sided test. For a one-sided test, the critical values of t with 19 degrees of freedom are 1.73 at the 5 percent significance level and 2.54 at the 1 percent level. Hence, with a one-sided test, we establish that the effect was significant at the 1 percent level.

Variation

Suppose that the survey had shown a mean *reduction* of expenditure of \$130, again with standard error \$60. What should we have concluded then? The t statistic for H_0: $\mu = 0$ is -2.17. Hence, if we had unthinkingly performed a two-sided test, we would have come to the strange conclusion that the abolition of the sales tax had a negative effect significant at the 5 percent level. However, if we have decided that we are justified in performing a one-sided test, what should we conclude? One answer would be to say that, although such a large negative estimate is unlikely under H_0, it is even more unlikely under the alternative H_1: $\mu > 0$. So we stay with H_0, and conclude that we have a somewhat freakish sample. At the same time, we should check whether there might be something that we have overlooked that might have given rise to the unexpected result. In the present case, we should definitely check how consumer expenditure is measured. If it is measured excluding the sales tax, then we have an anomalous result. But if it includes the sales tax, then the abolition of the tax could conceivably account for the reduction in measured expenditure. We should re-estimate the effect, excluding the tax.

EXERCISES

R.33* In Exercise R.28, a researcher was evaluating whether an increase in the minimum hourly wage has had an effect on employment in manufacturing industry. Explain whether she might have been justified in performing one-sided tests in cases (a)–(d), and determine whether her conclusions would have been different.

R.34 In Exercise R.29, a researcher was evaluating the assertion of a drug company that its course of treatment will, on average, reduce a person's cholesterol level by 0.8 mmol/L. Explain whether he might have been justified in performing one-sided tests in cases (a)–(c), and determine whether his conclusions would have been different.

R.35 You wish to test H_0: $\mu = 0$. You believe that μ cannot be negative and so the alternative hypothesis is H_1: $\mu > 0$. Accordingly, you decide to perform a one-sided test. However, you are wrong. μ is actually equal to μ_1, and μ_1 is negative. What are the implications for your test results?

R.14 Probability limits and consistency

The asymptotic properties of estimators are their limiting properties as the number of observations in a sample becomes very large and approaches infinity. We shall be concerned with the concepts of probability limits and consistency, and the central limit theorem. These topics are usually mentioned in standard statistics texts, but with no great seriousness of purpose, and generally without an explanation of why they are relevant and useful. The reason is that most standard introductory statistics courses cater to a wide variety of students, most

of whom will never have any use for these topics. However, for students of econometrics an understanding is essential because the asymptotic properties of estimators lie at the heart of much econometric analysis.

Probability limits

We will start with an abstract definition of a probability limit and then illustrate it with a simple example. A sequence of random quantities Z_n is said to converge in probability to a constant α if

$$\lim_{n\to\infty} P\left(|Z_n - \alpha| > \varepsilon\right) \to 0 \tag{R.89}$$

for any positive ε, however small. As the sample size becomes large, the probability of Z differing from α by any finite amount, however small, tends to zero. The constant α is described as the probability limit of the sequence, usually abbreviated as plim, and the mathematical statement is simplified to

$$\text{plim } Z = \alpha. \tag{R.90}$$

We will take as an example the mean of a sample of observations, $\bar{X}$, generated from a random variable X with unknown population mean μ_X and population variance σ_X^2. We will investigate how $\bar{X}$ behaves as the sample size n becomes large. For convenience, we shall assume that X has a normal distribution, but this does not affect the analysis. If X has a normal distribution with mean μ_X, $\bar{X}$ also has a normal distribution with mean μ_X, the difference being that the distribution for $\bar{X}$ has variance σ_X^2/n, as we saw in Section R.5.

As n increases, the variance decreases. This is illustrated in Figure R.39. We are assuming that X has mean 100 and standard deviation 50. If the sample size is 4, the standard deviation of $\bar{X}$, $\sigma_X/\sqrt{n}$, 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.39 shows the corresponding probability density functions. The larger is the sample size, the narrower and taller is the probability density function of $\bar{X}$. As n tends to infinity, the probability density function will collapse to a vertical spike located at μ_X. Formally,

$$\lim_{n\to\infty} P\left(|\bar{X} - \mu_X| > \varepsilon\right) \to 0. \tag{R.91}$$

The probability of $\bar{X}$ differing from μ_X by any finite amount ε, however small, tends to zero as n becomes large. More simply,

$$\text{plim } \bar{X} = \mu_X. \tag{R.92}$$

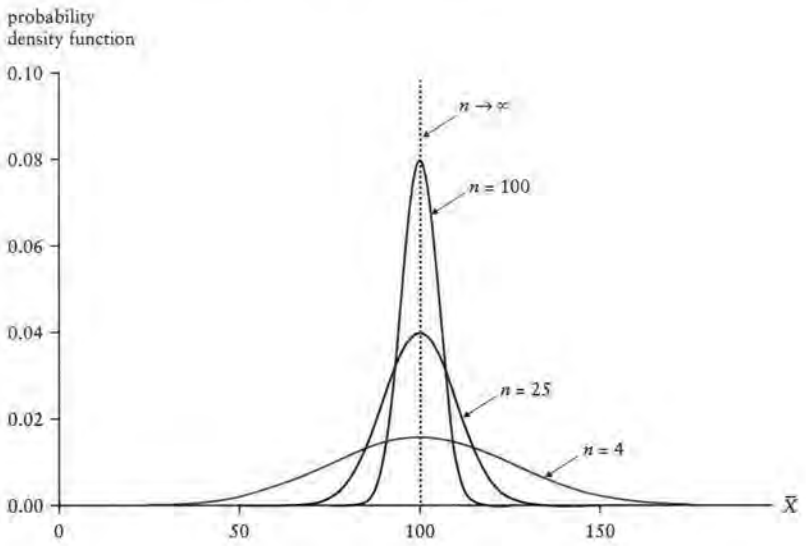


Figure R.39 Effect of increasing the sample size on the distribution of $\bar{X}$

Consistency

An estimator of a population characteristic is said to be consistent if it satisfies two conditions:

1. it possesses a probability limit, and so its distribution collapses to a spike as the sample size becomes large, and
2. the spike is located at the true value of the population characteristic.

The sample mean $\bar{X}$ in our example satisfies both conditions and so it is a consistent estimator of μ_X. Most standard estimators in simple applications satisfy the first condition because their variances are inversely proportional to n and so tend to zero as the sample size becomes large. The only issue then is whether the distribution collapses to a spike at the true value of the population characteristic.

A sufficient condition for consistency is that the estimator should be unbiased and that its variance should tend to zero as n becomes large. It is easy to see why this is a sufficient condition. If the estimator is unbiased for a finite sample, it must stay unbiased as the sample size becomes large. Meanwhile, if the variance of its distribution is inversely proportional to n, its distribution must collapse to a spike. Since the estimator remains unbiased, this spike must be located at the true value. The sample mean is an example of an estimator that satisfies this sufficient condition.

However, the condition is only sufficient, not necessary. It is possible for a biased estimator to be consistent, if the bias vanishes as the sample size becomes large. This is illustrated in principle in Figure R.40. The estimator is biased

upwards for finite samples, but nevertheless it is consistent because (1) the bias attenuates as n becomes large, and (2) its distribution collapses to a spike at the true value.

Consider the following estimator of a sample mean

$$Z = \frac{1}{n+1}\sum_{i=1}^{n} X_i. \tag{R.93}$$

It is biased downwards because

$$E(Z) = \frac{n}{n+1}\mu_X. \tag{R.94}$$

However, the bias will disappear asymptotically because $n/(n+1)$ will tend to 1. The distribution is said to be asymptotically unbiased because the expectation tends to the true value as the sample size becomes large. The variance of the estimator is given by

$$\text{var}(Z) = \frac{n}{(n+1)^2}\sigma_X^2, \tag{R.95}$$

which tends to zero as n becomes large. Thus, Z is consistent because its distribution collapses to a spike at the true value.

An estimator is described as inconsistent if its distribution collapses at a point other than the true value. It is also described as inconsistent if its distribution fails to collapse as the sample size becomes large. See Exercise R.37 for a simple example of a non-collapsing distribution. In practice, the distributions of most

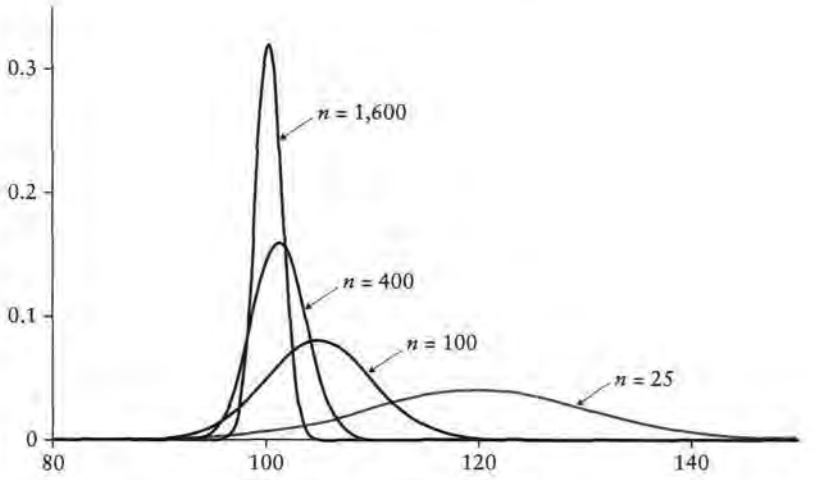


Figure R.40 Estimator that is consistent despite being biased in finite samples

estimators do collapse to a spike and the only issue is whether the spike is at the right value.

Why is consistency of interest?

In practice we deal with finite samples, not infinite ones. So why should we be interested in whether an estimator is consistent? Is this not just an abstract, academic exercise?

One reason for our interest is that estimators of the type shown in Figure R.40 are common in regression analysis, as we shall see later in this text. 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 estimator at all, especially if you are able to assess the direction of the bias in finite samples.

A second reason is that often we are unable to say anything at all about the expectation of an estimator. The expected value rules are weak analytical instruments that can be applied only in relatively simple contexts. In particular, the multiplicative rule $E\{g(X)h(Y)\} = E\{g(X)\}\ E\{h(Y)\}$ applies only when X and Y are independent, and in many situations of interest this will not be the case. By contrast, we have a much more powerful set of rules for plims. The first three are the counterparts of those for expectations. The remainder are new.

Plim rule 1 The plim of the sum of several variables is equal to the sum of their plims. For example, if you have three random variables X, Y, and Z, each possessing a plim,

$$\text{plim } (X + Y + Z) = \text{plim } X + \text{plim } Y + \text{plim } Z. \tag{R.96}$$

Plim rule 2 If you multiply a random variable possessing a plim by a constant, you multiply its plim by the same constant. If X is a random variable and b is a constant,

$$\text{plim } bX = b \text{ plim } X. \tag{R.97}$$

Plim rule 3 The plim of a constant is that constant. For example, if b is a constant,

$$\text{plim } b = b. \tag{R.98}$$

Plim rule 4 The plim of a product is the product of the plims, if they exist. For example, if $Z = XY$, and if X and Y both possess plims,

$$\text{plim } Z = (\text{plim } X)(\text{plim } Y). \tag{R.99}$$

Plim rule 5 The plim of a quotient is the quotient of the plims, if they exist, and provided that the plim of the denominator is not equal to zero. For example, if $Z = X/Y$, and if X and Y both possess plims, and plim Y is not equal to zero,

$$\text{plim } Z = \frac{\text{plim } X}{\text{plim } Y}. \tag{R.100}$$

Plim rule 6 The plim of a function of a variable is equal to the function of the plim of the variable, provided that the variable possesses a plim and provided that the function is continuous at that point,

$$\text{plim } f(X) = f(\text{plim } X). \tag{R.101}$$

To illustrate how the plim rules can lead us to conclusions when the expected value rules do not, consider the following example. Suppose that we hypothesize that a variable Y is a constant multiple of another variable Z:

$$Y = \alpha Z. \tag{R.102}$$

Z is generated randomly from a fixed distribution with population mean μ_Z and variance σ_Z^2. α is unknown and we wish to estimate it. We have a sample of n observations. Y is measured accurately but Z is measured with random error w with population mean zero and constant variance σ_w^2. Thus, in the sample, we have observations on X, where

$$X = Z + w \tag{R.103}$$

rather than Z. One estimator of α (not necessarily the best) is $\bar{Y}/\bar{X}$. Given (R.102) and (R.103),

$$\frac{\bar{Y}}{\bar{X}} = \frac{\alpha\bar{Z}}{\bar{Z}+\bar{w}} = \alpha - \alpha\frac{\bar{w}}{\bar{Z}+\bar{w}}. \tag{R.104}$$

Hence, we have decomposed the estimator into the true value, α, and an error term. To investigate whether the estimator is biased or unbiased, we need to take the expectation of the error term. But we cannot do this. The random quantity $\bar{w}$ appears in both the numerator and the denominator and the expected value rules are too weak to allow us to investigate the expectation analytically. However, we can invoke a law of large numbers that states that, under reasonable assumptions, a sample mean tends to a population mean as the sample size tends to infinity. Hence plim $\bar{w} = 0$ and plim $\bar{Z} = \mu_Z$. Since the plims exist,

$$\text{plim}\left\{\frac{\bar{Y}}{\bar{X}}\right\} = \alpha - \alpha\frac{\text{plim } \bar{w}}{\text{plim } \bar{Z} + \text{plim } \bar{w}} = \alpha - \frac{0}{\mu_Z + 0} = \alpha \tag{R.105}$$

(provided $\mu_Z \neq 0$). Thus, we are able to show that the estimator is consistent, despite the fact that we cannot say anything about its finite sample properties.

This subsection started out by asking why we are interested in consistency. As a first approximation, the answer is that if we can show that an estimator is consistent, then we may be optimistic about its finite sample properties, whereas if the estimator is inconsistent, in the sense of its distribution collapsing to a spike at the wrong value, we know that for finite samples it will definitely be biased. However, there are reasons for being cautious about preferring consistent estimators to inconsistent ones. First, a consistent estimator may also be biased for finite samples. Second, we are usually also interested in variances. If

a consistent estimator has a larger variance than an inconsistent one, the latter might be preferable if judged by the mean square error or a similar criterion that allows a trade-off between bias and variance. How can you resolve these issues? Mathematically they are intractable, otherwise we would not have resorted to large-sample analysis in the first place.

Simulations

The answer is to conduct a simulation, directly investigating the distributions of estimators under controlled conditions. We will do this for the example in the previous subsection. We will generate Z as a random variable with a normal distribution with mean 1 and variance 0.25. We will set α equal to 5, so

$$Y = 5Z. \tag{R.106}$$

We will generate the measurement error as a normally distributed random variable with zero mean and unit variance. The value of X in any observation is equal to the value of Z plus this measurement error. Figure R.41 shows the distributions of $\overline{Y}/\overline{X}$ for sample sizes 25, 100, 400, and 1,600, in each case for 10 million samples.

We can see that the standard deviation of the distribution, and hence its variance, diminishes as the sample size increases and it is reasonable to suppose that if the sample size became very large the distribution would collapse to a spike. Further, it is clear that the mean of the distribution is approaching the true value, 5. The actual numbers are given in Table R.7. Although there is some element of bias for a sample of 25 observations, it has mostly disappeared for

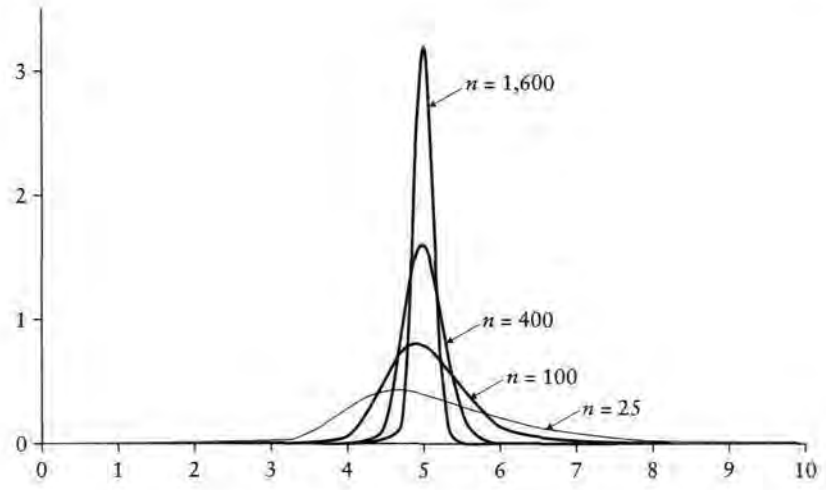


Figure R.41 Distribution of $\overline{Y}/\overline{X}$ for sample sizes 25, 100, 400, and 1,600

Table R.7

Sample size	Mean	Standard deviation
25	5.217	1.181
100	5.052	0.524
400	5.013	0.253
1,600	5.003	0.125

samples of 100 observations and the estimator is virtually unbiased for samples of size 1,600.

Of course, these conclusions are valid only for the particular way in which Z, Y, and X have been generated. However, the conclusions would have been qualitatively the same if we had had a different mean for Z, different standard deviations for Z and the measurement error, or a different value of α. The estimator would have been biased for finite samples, with a skewed distribution, but it would be consistent. We might have found different results for the rate of attenuation of the bias as a function of the sample size. If we were serious about investigating the properties of the estimator, we would perform some further sensitivity analysis. The purpose here was only to show how simulation might in principle shed light where mathematical analysis cannot.

EXERCISES

R.36 A random variable X has unknown population mean μ_X and population variance σ_X^2. A sample of n observations $\{X_1,...,X_n\}$ is generated. The average of the odd-numbered observations is used to estimate μ_X. Determine whether this estimator is consistent.

R.37* A random variable X has unknown population mean μ_X and population variance σ_X^2. A sample of n observations $\{X_1,...,X_n\}$ is generated. Show that

$$Z = \frac{1}{2}X_1 + \frac{1}{4}X_2 + \frac{1}{8}X_3 + \cdots + \frac{1}{2^{n-1}}X_{n-1} + \frac{1}{2^{n-1}}X_n$$

is an unbiased estimator of μ_X. Show that the variance of Z does not tend to zero as n tends to infinity and that therefore Z is an inconsistent estimator, despite being unbiased.

R.38 A random variable X has population mean μ and variance σ^2. Given a sample of n independent observations X_i, $i = 1,..., n$, determine whether the

following estimators of μ are consistent (you may assume that $\bar{X}$ is a consistent estimator):

(a) $\dfrac{n+2}{n^2+3n+1}\sum_{i=1}^{n} X_i$

(b) $\sum_{i=1}^{n} X_i^2 \Big/ \sum_{i=1}^{n} X_i$.

R.15 Convergence in distribution and central limit theorems

If a random variable X has a normal distribution, its sample mean $\bar{X}$ will also have a normal distribution. This fact is useful for the construction of t statistics and confidence intervals if we are employing $\bar{X}$ as an estimator of the population mean. However, what happens if we are *not* able to assume that X is normally distributed?

The standard response is to make use of a central limit theorem. Loosely speaking (we will make a more rigorous statement below), a central limit theorem states that the distribution of $\bar{X}$ will approximate a normal distribution as the sample size becomes large, even if the distribution of X is not normal. There are a number of central limit theorems, differing only in the assumptions that they make in order to obtain this result. Here we shall be content with using the simplest one, the Lindeberg–Levy central limit theorem. It states that, provided that the X_i in the sample are all drawn independently from the same distribution (the distribution of X), and provided that this distribution has finite population mean and variance, the distribution of $\bar{X}$ will converge to a normal distribution. This means that our t statistics and confidence intervals will be approximately valid after all, provided that the sample size is large enough. Of course, we will need to clarify what we mean by 'large enough'.

We will start by looking at two examples. Figure R.42 shows the distribution of $\bar{X}$ for the case where the X has a uniform distribution with range 0 to 1, for 10 million samples. A uniform distribution is one in which all values over the range in question are equally likely. For a sample size of 1, the distribution is the uniform distribution itself, and so it is a horizontal line. The figure also shows the distribution of the sample mean for sample sizes 10, 25, and 100, in each case for 10 million samples. It can be seen that the mean has a distribution very close to a normal distribution even when the sample size is only 10, and for larger sample sizes the approximation is even better.

Figure R.43 shows the corresponding distributions for the case where the underlying distribution is lognormal. A random variable is said to have

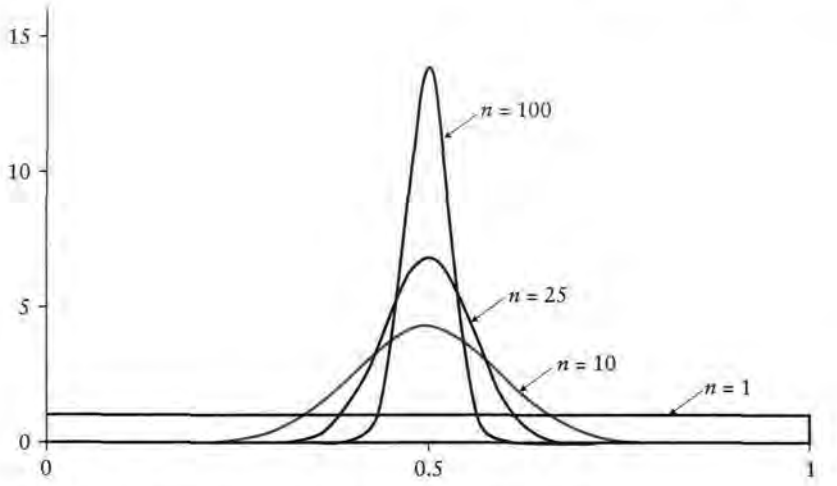


Figure R.42 Distribution of sample mean from a uniform distribution

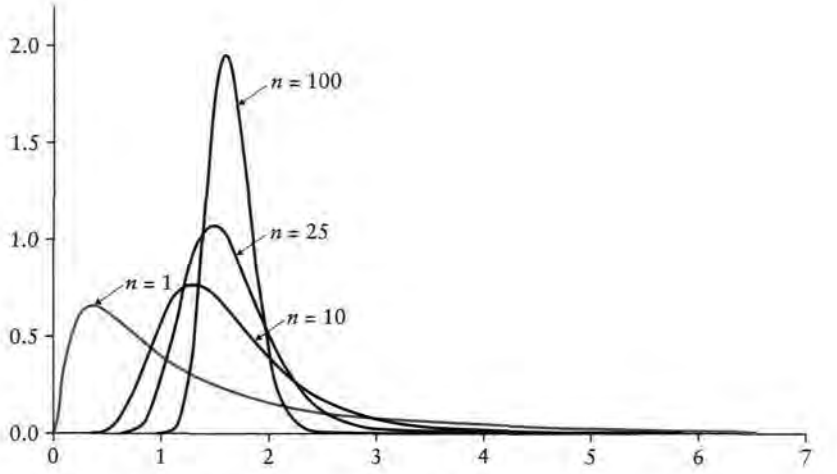


Figure R.43 Distribution of sample mean from a lognormal distribution

a lognormal distribution when its logarithm has a normal distribution. Like the normal distribution, the lognormal distribution is actually a two-parameter family of distributions. The one shown here is based on the standardized normal distribution (log X has a normal distribution with zero mean and unit variance).

As can be seen from the distribution for $n = 1$, the lognormal distribution is highly skewed. Even so, the distribution of $\bar{X}$ is becoming normal-like with sample size 25, and closer still for sample size 100.

In asserting that the distribution of $\bar{X}$ tends to become normal as the sample size increases, we have glossed over an important technical point that needs to

be addressed. The central limit theorem applies only in the limit, as the sample size tends to infinity. However, as the sample size tends to infinity, the distribution of $\bar{X}$ degenerates to a spike located at the population mean. So how can we talk about the limiting distribution being normal?

To answer this question, we transform the estimator in a way that its distribution does not collapse and become degenerate. The variance of $\bar{X}$ is σ^2/n, where σ^2 is the variance of X. So we consider the statistic $\sqrt{n}\bar{X}$. This has variance σ^2 and so the distribution of $\sqrt{n}\bar{X}$ does not collapse. However, $\sqrt{n}\bar{X}$ still does not have a limiting distribution. As n becomes large, $\bar{X}$ tends to μ, and $\sqrt{n}\bar{X}$ tends to $\sqrt{n}\mu$. $\sqrt{n}\mu$ increases without limit as n increases. So, instead, we consider the statistic $\sqrt{n}(\bar{X}-\mu)$. This does the trick. In common with other central limit theorems, the Lindeberg–Levy central limit theorem relates to this transformation, not directly to $\bar{X}$ itself. It tells us that, in the limit as n tends to infinity, $\sqrt{n}(\bar{X}-\mu)$ is normally distributed with mean zero and variance σ^2. To put this in mathematical notation,

$$\sqrt{n}(\bar{X}-\mu) \xrightarrow{d} N(0, \sigma^2). \tag{R.107}$$

(The symbol $\xrightarrow{d}$ means 'has the limiting distribution'.)

Now this relationship is true only as n goes to infinity. However, from the limiting distribution, we can start working back tentatively to finite samples. We can say that, for sufficiently large n, the relationship may hold approximately. Then, dividing the statistic by $\sqrt{n}$, we can say that, as an approximation,

$$(\bar{X}-\mu) \sim N\left(0, \frac{\sigma^2}{n}\right) \tag{R.108}$$

for sufficiently large n. Hence, again as an approximation, we can say that

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right). \tag{R.109}$$

(The symbol ~ means 'is distributed as'.) This is what we have in mind when we say that $\bar{X}$ is asymptotically distributed as a normal random variable, even though X itself is not.

Of course, this begs the question of what might be considered to be 'sufficiently large' n. To answer this question, the analysis must be supplemented by simulation. Figure R.44 shows the distribution of $\sqrt{n}(\bar{X}-\mu)$ for the uniform distribution. It is the counterpart of Figure R.42. You can see that the distributions for $n = 10$, 25, and 100 virtually coincide. This is because the distribution is almost perfectly normal for $n = 10$. Figure R.45 shows this more clearly. It compares the distribution for $n = 10$ with the limiting normal distribution (the dashed curve with a slightly higher mode). Thus, in the case of the uniform distribution, a sample size of 10 is sufficiently large.

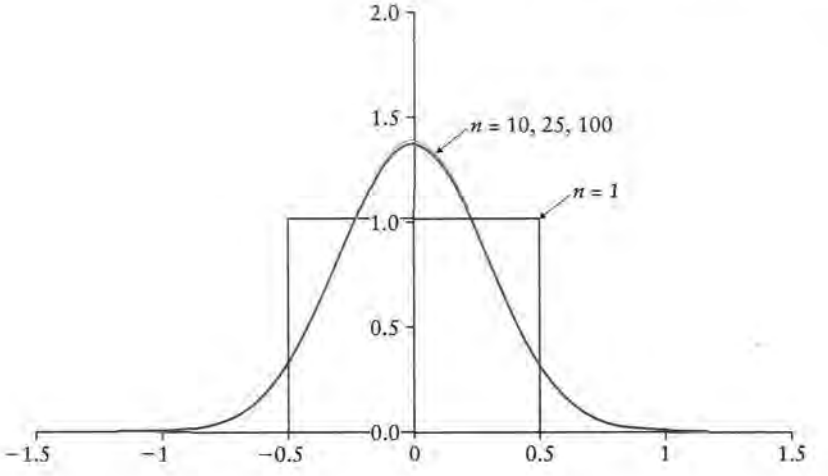


Figure R.44 Distribution of $\sqrt{n}\left(\overline{X}-\mu\right)$ for a uniform distribution

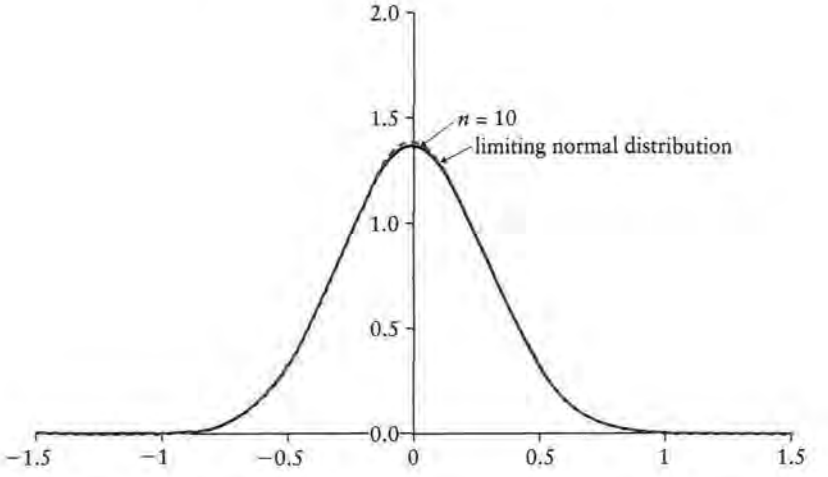


Figure R.45 Distribution of $\sqrt{n}\left(\overline{X}-\mu\right)$ for a uniform distribution, $n = 10$, and limiting normal distribution

Figure R.46 shows the distribution of $\sqrt{n}\left(\overline{X}-\mu\right)$ for the lognormal distribution. Here, we would come to a very different conclusion. In this case, it is evident that a sample size of 100 is not sufficiently large. Even with 100 observations, the distribution of $\sqrt{n}\left(\overline{X}-\mu\right)$ is still clearly different from that of the limiting normal distribution and, as a consequence, the use of conventional test statistics would be likely to lead to misleading results.

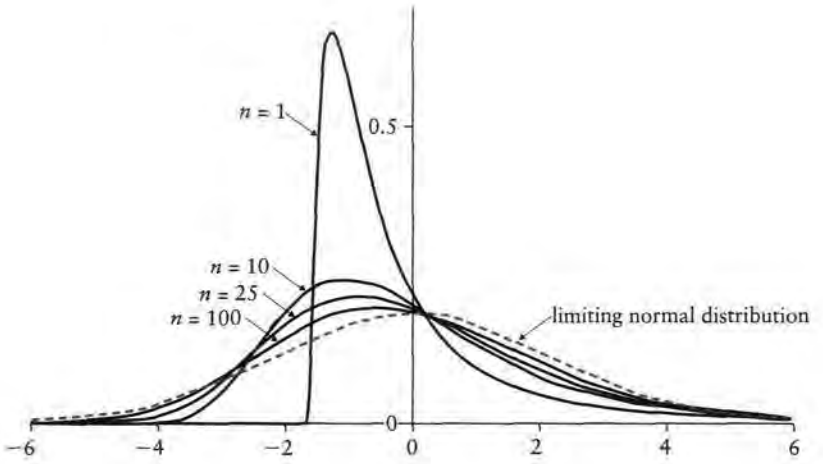
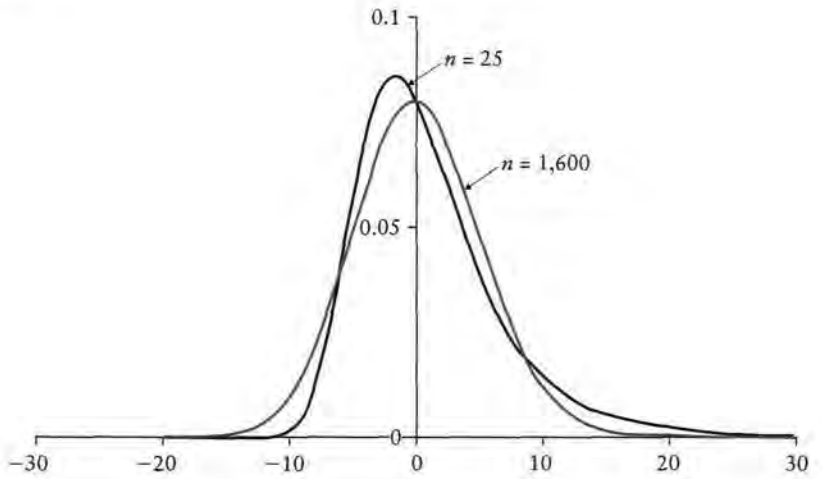


Figure R.46 Distribution of $\sqrt{n}\left(\overline{X}-\mu\right)$ for a lognormal distribution

EXERCISE

R.39



The figure shows the distribution of $\sqrt{n}\left(\frac{\overline{Y}}{\overline{X}}-5\right)$ for the simulation discussed at the end of Section R.14, with $n = 25$ and $n = 1{,}600$. (For clarity, the distributions for $n = 100$ and $n = 400$ have been omitted.) Comment on the relationship between these distributions and the corresponding ones in Figure R.41. Discuss whether one could use $\overline{Y}/\overline{X}$ as a test statistic for hypotheses relating to α.

Key terms

- acceptance region
- alternative hypothesis
- asymptotic properties
- bias, biased
- central limit theorem
- consistent, consistency
- continuous random variable
- degrees of freedom
- discrete random variable
- double structure of a variable
- efficient, efficiency
- estimate
- estimator
- expected value
- inconsistent, inconsistency
- independence
- limiting distribution
- loss function
- mean square error
- normal distribution
- null hypothesis
- one-sided test
- outcome
- plim
- population
- population correlation coefficient
- population covariance
- population mean
- population variance
- power of a test
- probability density function
- probability limit
- realization
- rejection region
- sample correlation coefficient
- significance level of a test
- simulation
- size of a test
- standard error
- standardized normal distribution
- t distribution
- t statistic
- Type I error
- Type II error
- unbiased

Appendix R.1: Unbiased estimators of the population covariance and variance

We will start with the estimator of the population covariance. The proof of the unbiasedness of the estimator of population variance follows immediately if one treats variances as special cases of covariances.

The estimator of the population covariance of X and Y is

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

Rewrite it as

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_X + \mu_X - \bar{X})(Y_i - \mu_Y + \mu_Y - \bar{Y}).$$

Then

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y) + \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_X)(\mu_Y - \bar{Y}) + \frac{1}{n-1}\sum_{i=1}^{n}(\mu_X - \bar{X})(Y_i - \mu_Y) + \frac{1}{n-1}\sum_{i=1}^{n}(\mu_X - \bar{X})(\mu_Y - \bar{Y}).$$

In the second term, $(\mu_Y - \bar{Y})$ is a common factor and can be taken out of the expression. Similarly, in the third term, $(\mu_X - \bar{X})$ is a common factor and can be taken out. The summation in the fourth term consists of n identical products $(\mu_X - \bar{X})(\mu_Y - \bar{Y})$ and is thus equal to $n(\mu_X - \bar{X})(\mu_Y - \bar{Y})$. Hence,

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y) + \frac{1}{n-1}(\mu_Y - \bar{Y})\sum_{i=1}^{n}(X_i - \mu_X) + \frac{1}{n-1}(\mu_X - \bar{X})\sum_{i=1}^{n}(Y_i - \mu_Y) + \frac{n}{n-1}(\mu_X - \bar{X})(\mu_Y - \bar{Y}).$$

Now

$$\sum_{i=1}^{n}(X_i - \mu_X) = \sum_{i=1}^{n}X_i - n\mu_X = n(\bar{X} - \mu_X)$$

and similarly,

$$\sum_{i=1}^{n}(Y_i - \mu_Y) = \sum_{i=1}^{n}Y_i - n\mu_Y = n(\bar{Y} - \mu_Y).$$

Hence,

$$\begin{aligned} s_{XY} &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y) + \frac{n}{n-1}(\mu_Y - \bar{Y})(\bar{X} - \mu_X) \\ &\quad + \frac{n}{n-1}(\mu_X - \bar{X})(\bar{Y} - \mu_Y) + \frac{n}{n-1}(\mu_X - \bar{X})(\mu_Y - \bar{Y}) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y) - \frac{n}{n-1}(\bar{X} - \mu_X)(\bar{Y} - \mu_Y) \\ &\quad - \frac{n}{n-1}(\bar{X} - \mu_X)(\bar{Y} - \mu_Y) + \frac{n}{n-1}(\bar{X} - \mu_X)(\bar{Y} - \mu_Y) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y) - \frac{n}{n-1}(\bar{X} - \mu_X)(\bar{Y} - \mu_Y). \end{aligned}$$

Now

$$(\bar{X}-\mu_X)=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_X) \quad \text{and} \quad (\bar{Y}-\mu_Y)=\frac{1}{n}\sum_{j=1}^{n}(Y_j-\mu_Y)$$

Hence,

$$\begin{aligned} s_{XY} &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\mu_X)(Y_i-\mu_Y) - \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_X)\frac{1}{n}\sum_{j=1}^{n}(Y_j-\mu_Y) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\mu_X)(Y_i-\mu_Y) - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i-\mu_X)(Y_j-\mu_Y). \end{aligned}$$

By definition, the expected value of any component $(X_i-\mu_X)(Y_j-\mu_Y)$ is the population covariance σ_{XY} when j is the same as i. There are n such components in the first term of s_{XY} and n in the second. (There are also $n(n-1)$ components $(X_i-\mu_X)(Y_j-\mu_Y)$ in the second term, with j different from i. These have expected value zero.) Hence,

$$E(s_{XY})=\frac{1}{n-1}n\sigma_{XY}-\frac{1}{n(n-1)}n\sigma_{XY}=\frac{n-1}{n-1}\sigma_{XY}=\sigma_{XY}$$

and so s_{XY} is an unbiased estimator of the population covariance.

In the special case where Y is the same as X, s_{XY} is s_X^2 and σ_{XY} is σ_X^2. Hence, we have also proved that s_X^2 is an unbiased estimator of the population variance of X.

Appendix R.2: Density functions of transformed random variables

Occasionally, we will need to determine the density functions of transformed random variables. Suppose that we have a random variable Y with density function $f(Y)$. Suppose Z is a transformation of Y:

$$Z=h(Y).$$

What can we say about the density function of Z, $g(Z)$? To keep things simple, we shall assume that $f(Y)$ and $h(Y)$ are both continuous and differentiable and that Z is either an increasing function of Y or a decreasing function. Then, if Y lies in the range $[y_1, y_2]$, Z must lie in the range $[z_1, z_2]$, where $z_1 = h(y_1)$ and $z_2 = h(y_2)$. So the probability of Z lying in the range $[z_1, z_2]$ is equal to the probability of Y lying in the range $[y_1, y_2]$. Mathematically,

$$\int_{z_1}^{z_2} g(Z)\,dZ = \int_{y_1}^{y_2} f(Y)\,dY.$$

Thus, at the margin,

$$g(Z)\,dZ = f(Y)\,dY$$

and so

$$g(Z) \approx \frac{f(Y)}{dZ/dY}.$$

For example, suppose that $Z = \log(Y)$. Then

$$g(Z) = \frac{f(Y)}{1/Y} = Y f(Y).$$