The Normal Distribution as an Approximation to the Binomial Distribution

Prerequisites

You should be familiar with the use of both the binomial and normal distributions.

Example (1)

Using tables or otherwise find to 4 decimal places

(a) Given $X \sim B(16, 0.5)$, find $P(X = 8)$

(b) Given $X \sim N(8, 4)$, find $P(7.5 < X < 8.5)$

Solution

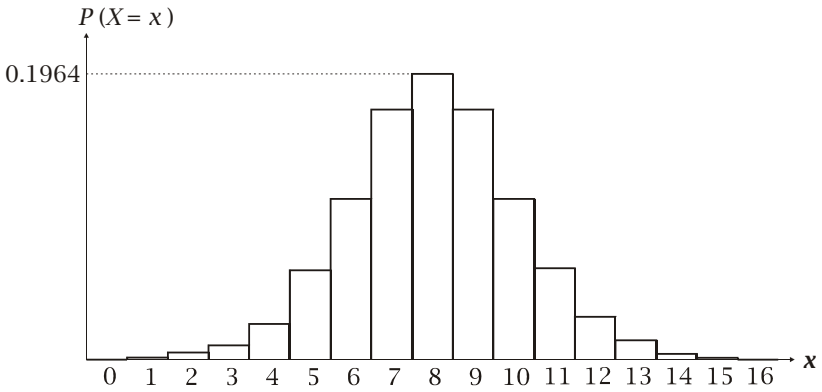
(a) $X \sim B(16, 0.5)$ $\qquad P(X = 8) = \binom{16}{8}(0.5)^{16} = 0.1964$

(b) $X \sim N(8, 4)$

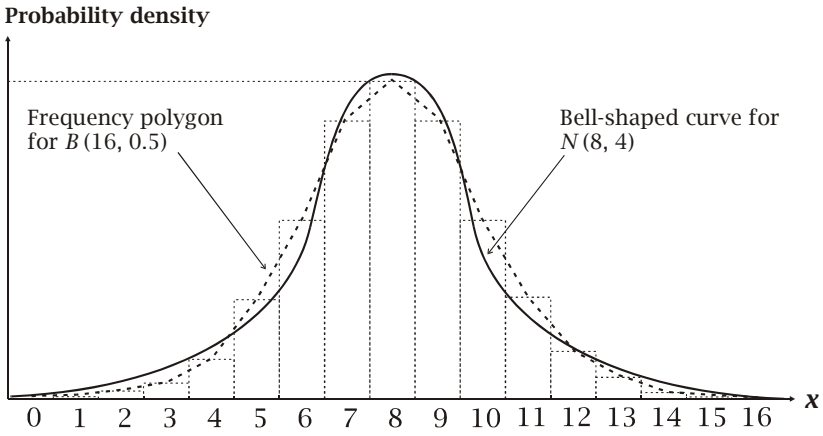
$x = 7.5 \qquad \Rightarrow \qquad z = \dfrac{7.5 - 8}{\sqrt{4}} = -0.25 \qquad \Phi(0.25) = 0.5987$

$P(7.5 < X < 8.5) = P(-0.25 < Z < 0.25) = 2 \times 0.5987 - 1 = 0.1974$

When we introduced the normal distribution it was by showing that it arises naturally from the binomial distribution as the number of trials becomes greater and greater. The following diagram shows the binomial distribution for $n = 16$ and $p = 0.5$.



This is a histogram, with the **heights** of the rectangles representing the probabilities that X takes a given value. Let us change the vertical axis to be probability density and thus equate the probability that X takes a given value with the **area** of the rectangle. We shall treat the probability densities as frequencies and join the midpoints of the rectangles to form a frequency polygon. We see that the shape of the frequency polygon is close to a smooth continuous bell-shaped curve.



This smooth continuous bell-shaped curve is the normal distribution given by $X \sim N(8, 4)$. We see that we could use a normal distribution to find approximate values to probabilities drawn from the binomial distribution.

Example (2)

In example (1) we found

$X \sim B(16, 0.5) \quad \Rightarrow \quad P(X = 8) = 0.1964$

$X \sim N(8, 4) \quad \Rightarrow \quad P(7.5 < X < 8.5) = 0.1974$

In this question we shall discuss the implications of using the value 0.1974 arising from $X \sim N(8, 4)$ as an approximation to the true value of $P(X = 8) = 0.1964$ when $X \sim B(16, 0.5)$.

(*a*) In the above we have approximated the probability of a **discrete value** $X = 8$ under the binomial distribution $X \sim B(16, 0.5)$ by the probability of an **interval** $7.5 < X < 8.5$ under the normal distribution $X \sim N(8, 4)$. Explain why we have done this.

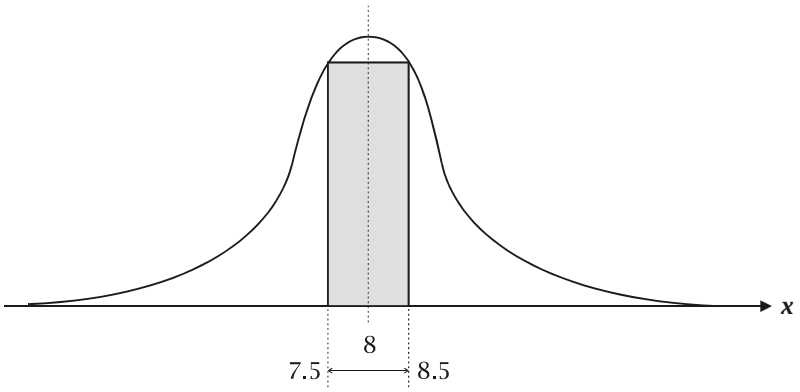
(*b*) Find to 2 significant figures the percentage relative error in using the normal distribution to approximate $P(X = 8) = 0.1964$ when $X \sim B(16, 0.5)$.

(*c*) Discuss whether the approximation is sufficiently accurate.

Solution

(*a*) We are approximating a discrete probability distribution $X \sim B(16, 0.5)$ by a continuous probability distribution $X \sim N(8, 4)$. By drawing the discrete distribution as a histogram and changing the vertical axis to probability density we have allowed the area of each rectangle in the histogram to equal the probability of that the variable X takes a certain value. When using the normal distribution $X \sim N(8, 4)$ we are fitting a curve approximately to match these areas. The area under a curve is given by an integral. The area corresponding to the histogram at $X = 8$ is the integral of the probability density function for $X \sim N(8, 4)$ between 7.5 and 8.5.

$P(X = 8)\left[\text{discrete value}\right] \longleftrightarrow P(7.5 \le X \le 8.5)\left[\text{continuous interval}\right]$



$P(7.5 \le X \le 8.5) = \int_{7.5}^{8.5} p(x)\,dx$

For a continuous distribution the probability of a point value is always 0.

$P(X = 8) = \int_{8}^{8} p(x)\,dx = 0$

(*b*) $\text{Percentage relative error} = \dfrac{\text{Absolute error}}{\text{True value}} \times 100\,\%$

$= \dfrac{0.1974 - 0.19638...}{0.19638...} \times 100 = 0.52\%\ (2\text{ s.f.})$

(*c*) When using an approximation rather than a true value there will always be error. The question of how much error is acceptable depends on the purpose to which you are using the approximation. If you are planning an expedition to Mars a percentage relative error of 0.52% is likely to be unacceptable. As a general rule statisticians would not accept 0.52% error because for a large number of applications that approximation would be insufficiently accurate. Furthermore, we have here only examined the relative error involved in approximating the probability for one discrete value, and the percentage relative error may be greater when approximating another value.

The normal approximation to the binomial distribution

Let us now examine in more detail how a binomial distribution can be approximated by a normal distribution. The result that we require is as follows.

Normal approximation to the binomial distribution

$X \sim B(n,p)$ can be approximated by $X \sim N(\mu,\sigma^2)$ where $\mu = np$ and $\sigma^2 = npq$ provided that

(1)	$n > 30$	the number of trials > 30
(2)	$npq > 5$	variance > 5
(3)	$np > 5$	mean > 5
(4)	$p \approx 0.5$	probability of a "success" ≈ 0.5

The conditions arise from the way in which the normal distribution arises as the limit of the binomial distribution as the number of trials, *n*, tends to infinity $n \to \infty$. As *n* gets larger, the distribution gets closer and closer to a normal distribution, and the condition $n > 30$ is roughly the minimum requirement to make the difference between the original binomial distribution and its normal approximation an acceptable margin of error. Other textbooks give different values of *n* or just state that the approximation is valid when *n* is large. The point of using the approximation is to make calculations easier for large values of *n*. The other conditions also specify the minimum requirements to make the approximation feasible. If $p \approx 0.5$, the binomial distribution centres around its mean, $\mu = np$; however, the approximation can still be valid if *p* drifts away from 0.5, except that in such a case, the number of trials, *n*, should be larger than 30.

Example (3)

(*a*) Using tables or otherwise find to 4 decimal places

(i)	$X \sim B(30, 0.5)$	$P(X = 15)$
(ii)	$X \sim N(15, 7.5)$	$P(14.5 < X < 15.5)$
(iii)	$X \sim B(30, 0.3)$	$P(X = 15)$
(iv)	$X \sim N(9, 6.3)$	$P(14.5 < X < 15.5)$

(*b*) Find the percentage relative error to 2 significant figures

(i) When approximating $P(X = 15)$ under $X \sim B(30, 0.5)$ by $P(14.5 < X < 15.5)$ under $X \sim N(15, 7.5)$

(ii) When approximating $P(X = 15)$ under $X \sim B(30, 0.3)$ by $P(14.5 < X < 15.5)$ under $X \sim N(9, 6.3)$.

(*c*) Discuss the relevance of the conditions given above for using the normal distribution as an approximation to the binomial distribution in the light of your answers to parts (*a*) and (*b*).

Solution

(*a*) (i) $X \sim B(30, 0.5)$

$$P(X = 15) = \binom{30}{15}(0.5)^{30} = 0.1445$$

(ii) $X \sim N(15, 7.5)$

$$x = 14.5 \quad \Rightarrow \quad z = \frac{14.5 - 15}{\sqrt{7.5}} = -0.1826 \quad \Phi(0.1826) = 0.5724$$

$$P(14.5 < X < 15.5) = P(-0.1826 < Z < 0.1826) = 2 \times 0.5724 - 1 = 0.1448$$

(iii) $X \sim B(30, 0.3)$

$$P(X = 15) = \binom{30}{15}(0.3)^{15}(0.7)^{15} = 0.0106$$

(iv) $X \sim N(9, 6.3)$

$$x = 14.5 \quad \Rightarrow \quad z = \frac{14.5 - 9}{\sqrt{6.3}} = 2.1913 \qquad \Phi(2.1913) = 0.9858$$

$$x = 15.5 \quad \Rightarrow \quad z = \frac{15.5 - 9}{\sqrt{6.3}} = 2.5897 \qquad \Phi(2.5897) = 0.9952$$

$$P(14.5 < X < 15.5) = P(2.1913 < Z < 2.5897) = 0.9952 - 0.9857 = 0.0095$$

(*b*) (i) Percentage relative error $= \dfrac{0.1448 - 0.1445}{0.1445} \times 100 = 0.21\%$ (2 s.f.)

(ii) Percentage relative error $= \dfrac{0.0106 - 0.0095}{0.0106} \times 100 = 10\%$ (2 s.f.)

(*c*) By increasing the size of n from 16 in example (2) to 30 in example (3) we have increased the accuracy of the approximation. The error in the approximation of the mid-value has decreased from 0.51% to 0.21%. It is still a matter of judgement as to whether this is sufficiently accurate. For most applications an error margin of 0.21% would be acceptable, hence the criterion $n > 30$ for the validity of the approximation. However, when we took the probability of a "success" to be $p = 0.3$ we found that there was a significant relative error of 10%. Clearly the requirement $p \approx 0.5$ is important, and in this case $(p = 0.3)$ if we wish to use the normal approximation to $X \sim B(n, p)$ we would require a value of n much larger than 30.

In the solution to example (3) we have not exactly compared like with like. The effect of changing the probability from $p = 0.5$ to $p = 0.3$ is to skew the histogram for the binomial distribution to the left so that the mean becomes $\mu = np = 30 \times 0.3 = 9$.

Example (4)

The following table compares the percentage relative errors

(i) When approximating $X \sim B(30, 0.5)$ by $X \sim N(15, 7.5)$

(ii) When approximating $X \sim B(30, 0.3)$ by $X \sim N(9, 6.3)$.

Discuss the general features of both approximations.

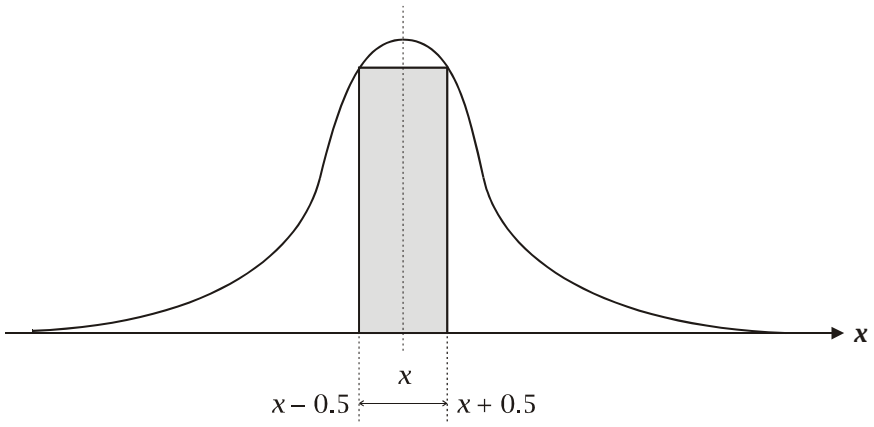
x	$B(30, 0.5)$	$N(15, 7.5)$	% relative error	$B(30, 0.3)$	$N(9, 6.3)$	% relative error
0	0.0000	0.0000	-	0.0000	0.0003	-
1	0.0000	0.0000	-	0.0003	0.0010	67
2	0.0000	0.0000	-	0.0018	0.0034	89
3	0.0000	0.0000	-	0.0072	0.0094	31
4	0.0000	0.0000	-	0.0208	0.0223	7.2
5	0.0001	0.0002	-	0.0464	0.0451	2.8
6	0.0006	0.0007	-	0.0829	0.0781	5.8
7	0.0019	0.0021	-	0.1219	0.1154	5.3
8	0.0055	0.0057	3.6	0.1501	0.1460	2.7
9	0.0133	0.0135	1.5	0.1573	0.1579	0.38
10	0.0280	0.0277	1.1	0.1416	0.1461	3.2
11	0.0509	0.0504	0.98	0.1103	0.1154	4.6
12	0.0806	0.0800	0.74	0.0749	0.0781	4.3
13	0.1115	0.1113	0.18	0.0444	0.0451	1.6
14	0.1354	0.1356	0.15	0.0231	0.0223	3.5
15	0.1445	0.1448	0.21	0.0106	0.0095	10
16	0.1354	0.1356	0.18	0.0043	0.0034	21
17	0.1115	0.1113	0.74	0.0015	0.0010	33
18	0.0806	0.0800	0.98	0.0005	0.0003	40
19	0.0509	0.0504	1.1	0.0001	0.0001	-
20	0.0280	0.0277	1.5	0.0000	0.0000	-
21	0.0133	0.0135	3.6	0.0000	0.0000	-
22	0.0055	0.0057	-	0.0000	0.0000	-
23	0.0019	0.0021	-	0.0000	0.0000	-
24	0.0006	0.0007	-	0.0000	0.0000	-
25	0.0001	0.0002	-	0.0000	0.0000	-
26	0.0000	0.0000	-	0.0000	0.0000	-
27	0.0000	0.0000	-	0.0000	0.0000	-
28	0.0000	0.0000	-	0.0000	0.0000	-
29	0.0000	0.0000	-	0.0000	0.0000	-
30	0.0000	0.0000	-	0.0000	0.0000	-

Solution

The normal distribution sometimes overestimates and sometimes underestimates the binomial distribution. Around the mean values and about the two *tails* (that is where the value is either small or large) the probability drawn from the normal distribution is larger than the corresponding true value from the binomial distribution. Around the central value (mean) the approximation is the best and about the tails the approximation is not so good. However, since the probabilities in the region of the tails are very small this loss of accuracy is not so significant. When the binomial distribution is skewed so that the probability p is not close to 0.5 the normal approximation is not a good fit to the binomial approximation, at least when $n = 30$. In such a case a larger value of n would be required.

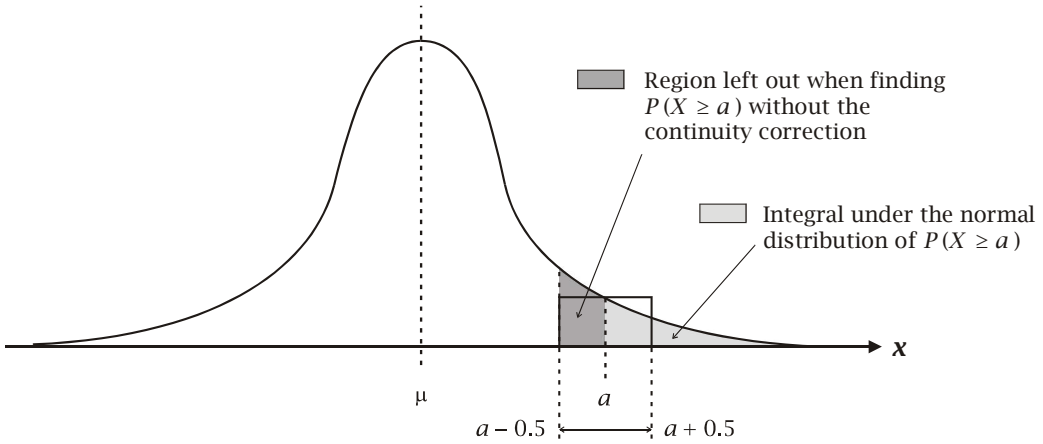
Using the normal approximation and the continuity correction

When the binomial distribution is approximated in this way by the normal distribution, a *continuity correction* must be made. The continuity correction arises because $X \sim B(n,p)$ is a discrete probability distribution and $X \sim N(\mu,\sigma^2) = N(np,npq)$ is a continuous probability distribution. We already encountered this aspect and accounted for it when previously approximating the point value $P(X = x)$ under $X \sim B(n,p)$ by the interval $P(x - 0.5 < X < x + 0.5)$ under $X \sim N(np,npq)$.



Here the continuity correction is found by adding and subtracting 0.5 from the x value to integrate from $x - 0.5$ to $x + 0.5$. In general we will be finding approximations to expressions like

$P(X \geq a)$ where a is an integer. Failure to make a continuity correction would result in part of the rectangle being left out.



The probability $P(X \geq a)$ under the binomial distribution includes the whole of the rectangle from $a-0.5$ to $a+0.5$ so corresponds to the integral $P(X \geq a-0.5)$ under the normal distribution. By a similar argument the probability $P(X > a)$ does **not** include the rectangle from $a-0.5$ to $a+0.5$ so here the continuity correction requires that when using the normal approximation to this binomial distribution we find the probability $P(X \geq a+0.5)$.

The continuity correction

The continuity correction is an addition or subtraction of 0.5 that is introduced when replacing a probability under the binomial distribution by its normal approximation. Whether to add or subtract 0.5 depends on whether the inequality is exact $(\leq \text{ or } \geq)$ or inexact $(< \text{ or } >)$.

The continuity correction adds back (or subtracts) the probability otherwise omitted (or included) as a result of taking the approximation.

Example (5)

Let $X \sim B(50,0.5)$. Use the normal approximation to find $(23 \leq X \leq 28)$.

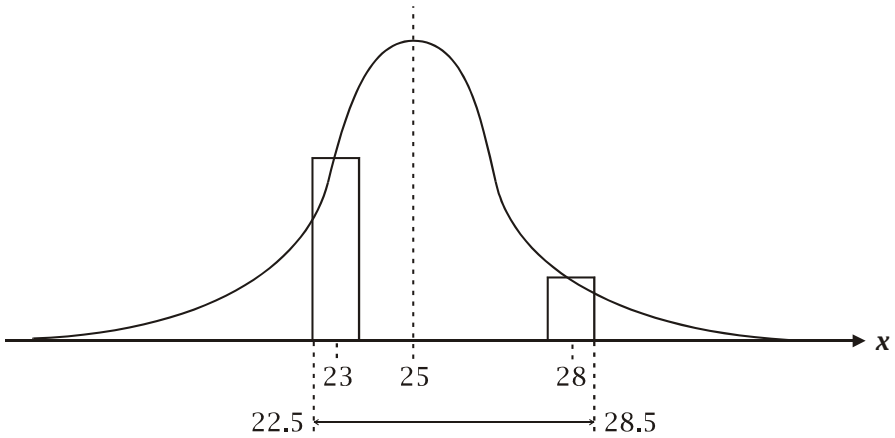
Solution

$X \sim B(50,0.5)$

$\mu = np = 50 \times 0.5 = 25$

$\sigma^2 = npq = 25 \times 0.5 = 12.5$

Then X may be approximated by $X \sim N(25,12.5)$



As 23 and 28 are included in the interval, the entire rectangles containing them are included in the approximation. Thus $P(23 \le X \le 28)$ under $X \sim B(50, 0.5)$ is approximated by $P(22.5 < X < 28.5)$ under $X \sim N(25, 12.5)$.

$x_1 = 22.5 \qquad z_1 = \dfrac{x-\mu}{\sigma} = \dfrac{22.5-25}{\sqrt{12.5}} = -0.707 \qquad \Phi(0.707) = 0.7601$

$x_2 = 28.5 \qquad z_2 = \dfrac{x-\mu}{\sigma} = \dfrac{28.5-25}{\sqrt{12.5}} = 0.990 \qquad \Phi(0.990) = 0.8389$

$P(23 < X < 28) \approx P(-0.707 < Z < 0.990) = 0.7601 - (1 - 0.8389) = 0.5990 = 0.599 \text{ (3 s.f.)}$

Example (6)

It is known that the probability of a seed of a certain type germinating is 0.42. A gardener plants 400 seeds and finds that only 145 germinate. Using a distributional approximation find the probability that 145 or less seeds germinate. Give your answer to 3 decimal places.

Solution

The phrase "using a distributional approximation" means here that we should use the normal approximation.

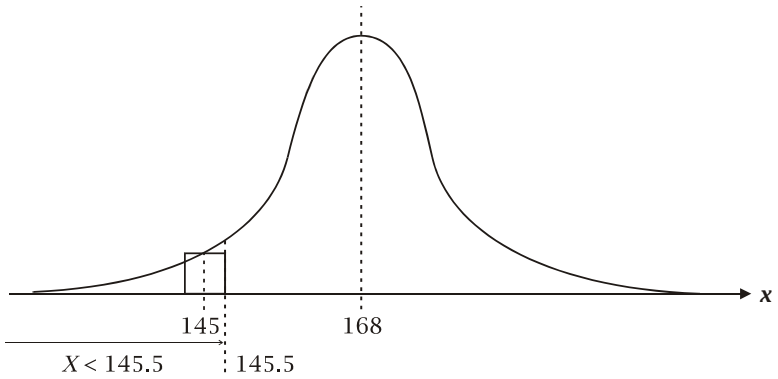
$X \sim B(400, 0.42) \qquad n = 400 \qquad p = 0.42$

$\mu = np = 400 \times 0.42 = 168$

$\sigma^2 = npq = 168 \times 0.58 = 97.44$

Approximate X by $X \sim N(168, 97.44)$

By applying the continuity correction, $P(X \le 145)$ under $X \sim B(400, 0.42)$ corresponds to $P(X \le 145.5)$ under $X \sim N(168, 97.44)$.



The whole of the rectangle enclosing 145 is included. Therefore, using the continuity correction when approximating by the normal distribution we require the entire region up to 145.5.

$$x = 145.5 \qquad z = \frac{x-\mu}{\sigma} = \frac{145.5-168}{\sqrt{97.44}} = -2.279 \quad \Phi(2.279) = 0.9887$$

$$P(X \le 145) \approx P(Z < -2.279) = 1 - 0.9887 = 0.0113 = 0.011 \quad (3 \text{ d.p.})$$

An interesting sequel to this problem is the question – would you advise the gardener to take the packet of seeds back to the shop he bought them from? Assuming that the probability of one of these seeds germinating really is 0.42 our calculations above show that there is only a 1.1% likelihood that only 145 or less will actually germinate. Assuming that the gardener has planted the seeds properly this result is unlikely and suggests that there is something wrong with the packet of seeds. So these reflections lead us into the topic of *hypothesis testing*. The proposition that the probability of the seeds germinating is 0.45 is a hypothesis. On the basis of this hypothesis we have concluded that the probability of only 145 or less seeds germinating is 1.1%. This low value suggests that *if* only 145 seeds germinate that is reason to conclude that the original hypothesis is false. However, this serves as an introduction only to the topic of hypothesis testing which we deal with in another chapter.