The Normal Approximation to the Binomial Distribution 39.2

Introduction

We have already seen that the Poisson distribution can be used to approximate the binomial distribution for large values of n and small values of p provided that the correct conditions exist. The approximation is only of practical use if just a few terms of the Poisson distribution need be calculated. In cases where many - sometimes several hundred - terms need to be calculated the arithmetic involved becomes very tedious indeed and we turn to the normal distribution for help. It is possible, of course, to use high-speed computers to do the arithmetic but the normal approximation to the binomial distribution negates this in a fairly elegant way. In the problem situations following this introduction the normal distribution is used to avoid very tedious arithmetic while at the same time giving a very good approximate solution.

Prerequisites

Before starting this Section you should ...

① be familiar with the normal distribution and the standard normal distribution

② be able to calculate probabilities using the standard normal distribution

Learning Outcomes

After completing this Section you should be able to ...

✓ recognise when it is appropriate to use the normal approximation to the binomial distribution

✓ solve problems using the normal approximation to the binomial distribution.

✓ interpret the answer in terms of the original problem.

1. The Normal Approximation to the Binomial Distribution

Problem

An engineering professional body estimates that 75% of the students taking undergraduate engineering courses are in favour of studying of statistics as part of their studies. If this estimate is correct, what is the probability that more than 780 undergraduate engineers out of a random sample of 1000 will be in favour of studying statistics?

Discussion

The problem involves a binomial distribution with a large value of n and so very tedious arithmetic may be expected. This can be avoided by using the normal distribution to approximate the binomial distribution underpinning the problem.
If X represents the number of engineering students in favour of studying statistics, then

$$X \sim B(1000, 0.75)$$

Essentially we are asked to find the probability that X is greater than 780, that is $P(X > 780)$. The calculation is represented by the following statement

$$P(X > 780) = P(X = 781) + P(X = 782) + P(X = 783) + \cdots + P(X = 1000)$$

In order to complete this calculation we have to find all 220 terms on the right hand side of the expression.
To get some idea of just how big a task this is when the binomial distribution is used, imagine applying the formula

$$P(X = r) = \frac{n(n-1)(n-2)\ldots(n-r+1)p^r(1-p)^{n-r}}{r(r-1)(r-2)\ldots 3.2.1}$$

220 times! You would have to take $n = 1000,\ p = 0.75$ and vary r from 781 to 1000. Clearly, the task is enormous.
Fortunately, we can approximate the answer very closely by using the normal distribution with the *same mean and standard deviation* as $X \sim B(1000, 0.75)$. Applying the usual formulae for μ and σ we obtain the values $\mu = 750$ and $\sigma = 13.7$ from the binomial distribution.

We now have two distributions, $X \sim B(1000, 0.75)$ and (say) $Y \sim N(750, 13.7^2)$. Remember that the second parameter in this series of workbooks represents the variance.
By doing the appropriate calculations, (this is extremely tedious even for one term!) it can be shown that

$$P(X = 781) \approx P(780.5 \leq Y \leq 781.5)$$

This statement means that the probability that $X = 781$ calculated from the binomial distribution $X \sim B(1000, 0.75)$ can be very closely approximated by the area under the normal curve $Y \sim N(750, 13.7^2)$ between 780.5 and 781.5. This relationship is then applied to all 220 terms involved in the calculation.

The result is summarised below:

$$\begin{aligned} P(X = 781) &\approx P(780.5 \leq Y \leq 781.5) \\ P(X = 782) &\approx P(781.5 \leq Y \leq 782.5) \\ &\vdots \\ P(X = 999) &\approx P(998.5 \leq Y \leq 999.5) \\ P(X = 1000) &\approx P(999.5 \leq Y \leq 1000.5) \end{aligned}$$

By adding these probabilities together we get

$$\begin{aligned} P(X > 780) &= P(X = 781) + P(X = 782) + \cdots + P(X = 1000) \\ &\approx P(780.5 \leq Y \leq 1000.5) \end{aligned}$$

To complete the calculation we need only to find the area under the curve $Y \sim N(750, 13.7^2)$ between the values 780.5 and 1000.5. This is far easier than completing the 220 calculations suggested by the use of the binomial distribution.

Finding the area under the curve $Y \sim N(750, 13.7^2)$ between the values 780.5 and 1000.5 is easily done by following the procedure used previously.
The calculation, using the tables given previously and working to three decimal places is

$$\begin{aligned} P(X > 780) &\approx P\left(\frac{780.5 - 750}{13.7} \leq Z \leq \frac{1000.5 - 750}{13.7}\right) \\ &= P(2.23 \leq Z \leq 18.28) \\ &= P(Z \geq 2.23) \\ &= 0.013 \end{aligned}$$

Notes:

1. Since values as high as 18.28 effectively tell us to find the area to the right of 2.33 (the area to the right of 18.28 is so close to zero as to make no difference) we have

   $$P(Z \geq 2.23) = 0.0129 \approx 0.013$$

2. The solution given *assumes* that the original binomial distribution can be approximated by a normal distribution. This is not always the case and you must always check that the following conditions are satisfied before you apply a normal approximation. The conditions are:

   - $np > 5$
   - $n(1 - p) > 5$

   You can see that these conditions are satisfied here.

A particular production process used to manufacture ferrite magnets used to operate reed switches in electronic meters is known to give 10% defective magnets on average. If 200 magnets are randomly selected, what is the probability that the number of defective magnets is between 24 and 30?

If X is the number of defective magnets then $X \sim B(200, 0.1)$ and we require

$$P(24 < X < 30) = P(25 \leq X \leq 29)$$

Now,

$$\mu = np = 200 \times 0.1 = 20 \quad \text{and} \quad \sigma = \sqrt{np(1-p)} = \sqrt{200 \times 0.1 \times 0.9} = 4.24$$

Note that $np > 5$ and $n(1-p) > 5$ so that approximating $X \sim B(200, 0.1)$ by $Y \sim N(20, 4.24^2)$ is acceptable. We can approximate $X \sim B(200, 0.1)$ by the normal distribution $Y \sim N(20, 4.24^2)$ and use the transformation

$$Z = \frac{Y - 20}{4.24} \sim N(0, 1)$$

so that

$$\begin{aligned} P(25 \leq X \leq 29) &\approx P(24.5 \leq Y \leq 29.5 \\ &= P(\frac{24.5 - 20}{4.24} \leq Z \leq \frac{29.5 - 20}{4.24} \\ &= P(1.06 \leq Z \leq 2.24) \\ &= 0.4875 - 0.3554 \\ &= 0.1321 \end{aligned}$$

Example Overbooking of passengers on intercontinental flights is a common practice among airlines. Aircraft which are capable of carrying 300 passengers are booked to carry 320 passengers. If 10% of passengers who have a booking fail to turn up for their flights, what is the probability that at least one passenger who has a booking, will end up without a seat on a particular flight?

Solution

Let $p = P(\text{a passenger with a booking, fails to turn up}) = 0.10$.

Then: $q = P(\text{a passenger with a booking, turns up}) = 1 - p = 1 - 0.10 = 0.9$

Let $X =$ number of passengers with a booking who turn up.

As there are 320 bookings, we are dealing with the terms of the binomial expansion of

$$(q+p)^{320} = q^{320} + 320q^{319}p + \frac{320 \times 319}{2!}q^{318}p^2 + \cdots + p^{320}$$

Using this approach is too long to treat by finding the values, term by term. It is easier to switch to the corresponding normal distribution, i.e. that which has the same mean and variance as the binomial distribution above.

Mean $= \mu = 320 \times 0.9 = 288$

Variance $= 320 \times 0.9 \times 0.1 = 28.8$ hence, $\sigma = \sqrt{28.8} = 5.37$

Hence, the corresponding normal distribution is given by $Y \sim N(288, 28.8)$

So that, $P(X \geq 300) \approx P(Y \geq 300.5) = P(Z \geq \dfrac{300.5 - 288}{5.37}) = P(Z \geq 2.33)$

From Z-tables $P(Z \geq 2.33) = 0.0099$.

NB. Continuity correction when changing from the binomial, a discrete distribution, to the normal, a continuous distribution.

Exercises

1. The diameter of an electric cable is normally distributed with mean 0.8cm and variance 0.0004cm^2.

   (a) What is the probability that the diameter will exceed 0.81cm?

   (b) The cable is considered defective if the diameter differs from the mean by more than 0.025cm. What is the probability of obtaining a defective cable?

2. A machine packs sugar in what are nominally 2kg bags. However there is a variation in the actual weight which is described by the normal distribution.

   (a) Previous records indicate that the standard deviation of the distribution is 0.02 kg and the probability that the bag is underweight is 0.01. Find the mean value of the distribution.

   (b) It is hoped that an improvement to the machine will reduce the standard deviation while allowing it to operate with the same mean value. What value standard deviation is needed to ensure that the probability that a bag is underweight is 0.001?

3. Rods are made to a nominal length of 4cm but in fact the length is a normally distributed random variable with mean 4.01cm and standard deviation 0.03. Each rod costs 6p to make and may be used immediately if its length lies between 3.98cm and 4.02cm. If its length is less than 3.98cm the rod cannot be used but has a scrap value of 1p. If the length exceeds 4.02cm it can be shortened and used at a further cost of 2p. Find the average cost per usable rod.

4. A super-market chain sells its 'own-brand' label instant coffee in packets containing 200 gms of coffee granules. The packets are filled by a machine which is set to dispense fills of 200 gms. If fills are normally distributed, about a mean of 200 gms and with a standard deviation of 7 gms, find the number of packets out of a consignment of 1,000 packets that:

   (a) contain more than 215 gms.

   (b) contain less than 195 gms.

   (c) contain between 190 to 210 gms.

   The super-market chain decides to withdraw all packets with less than a certain weight of coffee. As a result, 40 packets which were in the consignment of 1,000 packets are withdrawn. What is the weight at which the 'line has been drawn' ?

5. The time taken by a team to complete the assembly of an electrical component is found to be normally distributed, about a mean of 110 minutes, and with a standard deviation of 10 minutes. Out of a group of 20 teams, how many will complete the assembly:

   i. within 95 minutes.

   ii. in more than 2 hours.

   If the management decides to set a 'cut off' time such that 95% of the teams will have completed the assembly on time, what time limit should be set?

Answers

1. $X \sim N(0.8,\ 0.0004)$ (a)$P(X > 0.81) = P\left(Z > \dfrac{0.81 - 0.8}{0.02}\right)$

$= P(Z > 0.5) = 0.5 - P(0 < Z < 0.5) = 0.5 - 0.1915 = 0.3085$

(b)$P[(X > 0.825) \cup (X < 0.785)] = 2P(X > 0.825)$

$= 2P\left(Z > \dfrac{0.025}{0.02}\right) = 2P(Z > 1.25)$

$= 2[-P(0 < Z < 1.25) + 0.5] = 2[-0.3944 + 0.5] = 0.2112$

2. (a)$\sigma = 0.02$, $P(X < 2) = 0.01$ what is μ?

i.e. $P\left(Z < \dfrac{2-\mu}{0.02}\right) = 0.01$

$\therefore \quad 0.05 - P\left(0 < Z < \dfrac{\mu - 2}{0.02}\right) = 0.01$

i.e. $P\left(0 < Z < \dfrac{\mu - 2}{0.02}\right) = 0.49$

$\therefore \quad \dfrac{\mu - 2}{0.02} = 2.33 \quad \therefore \quad \mu = 2.0466$

(b)Now require σ such that

$P(X < 2) = 0.001 \quad$ with $\quad \mu = 2.0466$

i.e. $0.5 - P\left(0 < Z < \dfrac{0.0466}{\sigma}\right) = 0.001$

$\therefore \quad P\left(0 < Z < \dfrac{0.0466}{\sigma}\right) = 0.499 \quad \therefore \quad \dfrac{0.0466}{\sigma} = 3.1$

$\therefore \quad \sigma = 0.015$

3. $L \sim N(4.01,\ (0.03)^2)$

Cost has 2 possible values per usable rod, 6p, 8p.

$$P(C = 6) = P(3.98 < L < 4.02) = P\left(0 < Z < \frac{4.01 - 3.98}{0.03}\right)$$

$$+P\left(0 < Z < \frac{4.02 - 4.01}{0.03}\right)$$

$$= P(0 < Z < 1) + P(0 < Z < 0.333) = 0.3413 + 0.1305 = 0.4718$$

$$P(C = 8) = P(L > 4.02) = P(Z > 0.333)$$

$$= 0.5 - P(0 < Z < 0.333) = 0.3695$$

of 100 rods produced.

Total	
usable rods 36.95 cost 8p each	295.6
47.18 cost 6p each	283.08
15.87 cost 5p each	79.35

average cost per usable rod $= \dfrac{283.08 + 295.6 + 79.35}{84.13} = 7.82$

Continued 4. Let X = the amount of coffee in a fill; then $X \sim N(200, 7)$

(a) $P(X > 215) = P(Z > \dfrac{215.0 - 200.0}{7.0}) = P(Z > 2.14) = 0.016$ from Z - tables

Hence, from a consignment of 1,000 packets, the number containing more than 215 gms. $= 1000 \times 0.016 = 16$

(b) $P(X < 195) = P(Z < \dfrac{195.0 - 200.0}{7.0} = P(Z < -0.714) = 0.2389$ from Z - tables

Hence, from a consignment of 1,000 packets, the number containing less than 195 gms. $= 1000 \times 0.2389 = 238.9$

(c)

$$P(190.0 < X < 210.0) = P(\frac{190.0 - 200.0}{7.0} < Z < \frac{210.0 - 200.0}{7}) = P(-1.43 < Z < 1.43) = 0.8472 \quad \text{from } Z \text{ tables}$$

Hence, from a consignment of 1,000 packets, the number containing between 190 gms. and 210 gms. $= 1000 \times 0.8472 = 847$

If 40 out of the 1000 packets withdrawn, then P(sub-standard packet) $= \dfrac{40}{1000} = 0.04$.

Let k be the limit below which, packets are sub-standard, then $P(X < k) = 0.04$

From Z tables, $Z = -1.75$ as we are dealing with 'less than' i.e. the 'left-hand part of the standard normal distribution curve.

Hence, $\dfrac{k - 200.0}{7} = -1.75$ i.e. $k = -1.75(7) + 200.0 = 187.75$

'Line drawn' at 188 gms.; any packet below this value to be withdrawn.

5. Let X be the time taken to assemble the component; then $X \sim N(110, 10)$

(a) $P(X < 95) = P(Z < \dfrac{95.0 - 110.0}{10.0}) = P(Z < -1.5) = 0.3085$ from Z - tables

Hence, from a group of 20 teams, the number completing the assembly within 95 mins. $= 20 \times 0.3085 = 6.17 = 6$

(b) $P(X > 120) = P(Z > \dfrac{120.0 - 110.0}{10.0}) = P(Z > 1.0) = 0.1587$ from Z - tables

Hence, from a group of 20 teams, the number completing the assembly in more than 2 hrs. $= 20 \times 0.1587 = 3.174 = 3$

If 95% of teams are to complete the assembly 'on time', then 5% take longer than k the set time

i.e. $P(X > k) = 0.05$ hence, $Z = 1.64$

Therefore, $\dfrac{k - 110.0}{10.0} = 1.64$

or, $k = 10(1.64) + 110.0 = 126.4$ mins.