17 Binomial distributions

This chapter deals with binomial probability. It moves quickly through calculations of individual probabilities in Section 17A, to binomial distributions in Section 17B. These two sections combine ideas from all of Chapters 12–15 of the Year 11 book, and this may be an ideal time to revise that material.

The last two sections introduce normal approximation to a binomial distribution. Section 17C does this in the obvious way, expressing a question about a binomial distribution in terms of the normal approximation.

The final section introduces the sample proportions that binomial distributions produce, and uses the normal distribution to approximate the probabilities of the sample proportions rather than of the original binomial distribution. This approach may seem conceptually rather demanding at first, but the calculations involved turn out to be simple variations of those introduced in Section 17C.

Plenty of examples of binomial data are presented in the text and in the exercises, together with simulations of binomial experiments using technology.

Digital Resources are available for this chapter in the **Interactive Textbook** and **Online Teaching Suite**. See the *overview* at the front of the textbook for details.

ISBN 978-1-108-76630-2

Cambridge University Press

17A Binomial probability

Section 17A is about individual binomial probabilities, in preparation for the later sections on binomial probability distributions. The discussion combines probability with the expansion of the binomial $(x + y)^n$, showing their close relationship.

We will be concerned with multi-stage experiments of a rather special form. The stages are independent, each stage has just two outcomes, ‘success’ and ‘failure’, with the same probabilities at each stage, and the random variable for the whole experiment records only the number of successes, not their order.

Bernoulli trials

It is convenient to have a name for the independent stages and describe them carefully. A *Bernoulli trial* or *Bernoulli experiment* is a single-stage random experiment with just two outcomes, conventionally called ‘success’ and ‘failure’, to which we usually assign the probabilities p and $q = 1 - p$.

The classic example of a Bernoulli trial is tossing a coin, where ‘success’ is heads and ‘failure’ is tails. The probability of success is then $p = \frac{1}{2}$, and the probability of failure is $q = 1 - p = \frac{1}{2}$.

The other classic example is throwing a die, provided that we define ‘success’ — if we define ‘success’ to be ‘throwing a six’, and record only that, then we have a Bernoulli trial with $p = \frac{1}{6}$.

A Bernoulli trial is completely determined by just one parameter — the probability p of ‘success’.

Binomial experiments

A *binomial experiment* is an n-stage experiment in which:

- each stage is a Bernoulli trial with the same probability p of success,
- the stages are independent — no stage affects any other stage,
- the random variable X is the number of successes — order is irrelevant.

Think now about tossing a coin 12 times and counting the number of heads. Or think about throwing four dice and counting the number of sixes.

A binomial experiment is thus completely determined by just two parameters — the number n of trials, and the probability p of success at each stage.

Each stage of a binomial experiment is trivially a binomial experiment with just one stage. Thus a Bernoulli trial is a special case of a binomial experiment.

Example 1 17A

Identify some further examples of Bernoulli trials.

SOLUTION

- Choose an adult Australian — did they vote in the last election?
- Choose a shopper in the street — have they visited Wooldi today?
- Ask a whale-spotting boat returning to port — did you spot a whale?
- Visit your letter-box — is there mail in it or not?
- Is there intelligent life elsewhere in the universe — yes or no?
- Choose a person at random — were they offered the last job they applied for?

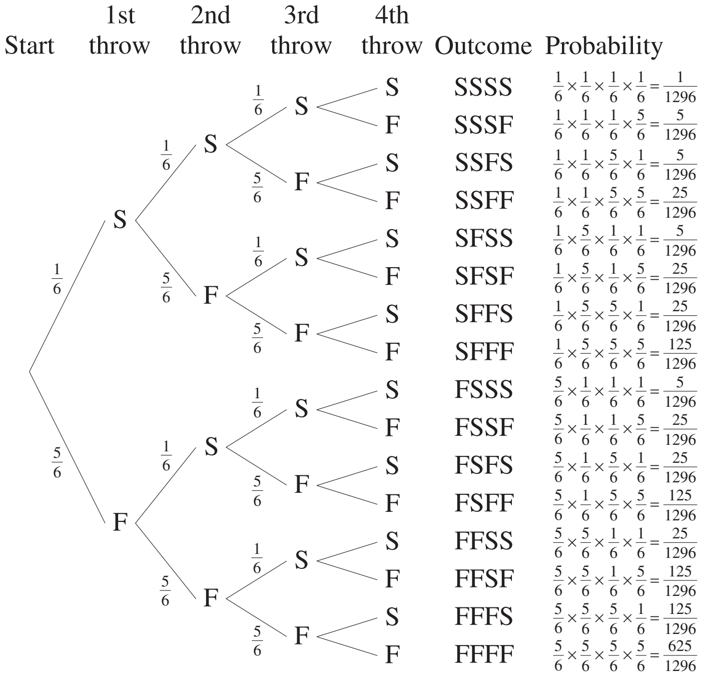
Example — repeatedly attempting to throw a six on a die

We will develop Bernoulli trials further in the next section. Let us now turn to the problem of calculating individual probabilities in a binomial distribution. Here is a classic example.

A die is thrown four times. Find the probabilities of getting $0,\ 1,\ 2,\ 3$ or 4 sixes.

Let S (success) be 'throwing a six' and F (failure) be 'not throwing a six', so that $p = \frac{1}{6}$ and $q = 1 - p = \frac{5}{6}$.

Here is the probability tree diagram showing the sixteen possible outcomes taking account of order, and their respective probabilities.


Start
1st throw
2nd throw
3rd throw
4th throw
Outcome
Probability
S
F
$\frac{1}{6}$
$\frac{5}{6}$
SSSS $\frac{1}{6}\times\frac{1}{6}\times\frac{1}{6}\times\frac{1}{6}=\frac{1}{1296}$
SSSF $\frac{1}{6}\times\frac{1}{6}\times\frac{1}{6}\times\frac{5}{6}=\frac{5}{1296}$
SSFS $\frac{1}{6}\times\frac{1}{6}\times\frac{5}{6}\times\frac{1}{6}=\frac{5}{1296}$
SSFF $\frac{1}{6}\times\frac{1}{6}\times\frac{5}{6}\times\frac{5}{6}=\frac{25}{1296}$
SFSS $\frac{1}{6}\times\frac{5}{6}\times\frac{1}{6}\times\frac{1}{6}=\frac{5}{1296}$
SFSF $\frac{1}{6}\times\frac{5}{6}\times\frac{1}{6}\times\frac{5}{6}=\frac{25}{1296}$
SFFS $\frac{1}{6}\times\frac{5}{6}\times\frac{5}{6}\times\frac{1}{6}=\frac{25}{1296}$
SFFF $\frac{1}{6}\times\frac{5}{6}\times\frac{5}{6}\times\frac{5}{6}=\frac{125}{1296}$
FSSS $\frac{5}{6}\times\frac{1}{6}\times\frac{1}{6}\times\frac{1}{6}=\frac{5}{1296}$
FSSF $\frac{5}{6}\times\frac{1}{6}\times\frac{1}{6}\times\frac{5}{6}=\frac{25}{1296}$
FSFS $\frac{5}{6}\times\frac{1}{6}\times\frac{5}{6}\times\frac{1}{6}=\frac{25}{1296}$
FSFF $\frac{5}{6}\times\frac{1}{6}\times\frac{5}{6}\times\frac{5}{6}=\frac{125}{1296}$
FFSS $\frac{5}{6}\times\frac{5}{6}\times\frac{1}{6}\times\frac{1}{6}=\frac{25}{1296}$
FFSF $\frac{5}{6}\times\frac{5}{6}\times\frac{1}{6}\times\frac{5}{6}=\frac{125}{1296}$
FFFS $\frac{5}{6}\times\frac{5}{6}\times\frac{5}{6}\times\frac{1}{6}=\frac{125}{1296}$
FFFF $\frac{5}{6}\times\frac{5}{6}\times\frac{5}{6}\times\frac{5}{6}=\frac{625}{1296}$

When we ignore order, these 16 outcomes collapse to five, and the resulting binomial random variable X has five possible values, $0,\ 1,\ 2,\ 3$ and 4.

The outcome 'two sixes', for example, can be obtained in ${}^4C_2 = 6$ different ways: SSFF, SFSF, SFFS, FSSF, FSFS, FFSS, because there are $\dfrac{4!}{2! \times 2!} = {}^4C_2$ ways of arranging two Ss and two Fs.

Each of these six outcomes has the same probability $\left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^2$, so

$$P(X = 2) \qquad = {}^4C_2 \times \left(\tfrac{1}{6}\right)^2 \times \left(\tfrac{5}{6}\right)^2.$$

Similar arguments apply to the probabilities of getting $0,\ 1,\ 3$ and 4 sixes:

Result	Probability	Approximation
0 sixes	${}^4C_0 \times \left(\frac{1}{6}\right)^0 \times \left(\frac{5}{6}\right)^4$	0.482 25
1 six	${}^4C_1 \times \left(\frac{1}{6}\right)^1 \times \left(\frac{5}{6}\right)^3$	0.385 80
2 sixes	${}^4C_2 \times \left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^2$	0.115 74
3 sixes	${}^4C_3 \times \left(\frac{1}{6}\right)^3 \times \left(\frac{5}{6}\right)^1$	0.015 43
4 sixes	${}^4C_3 \times \left(\frac{1}{6}\right)^3 \times \left(\frac{5}{6}\right)^1$	0.000 77
—	—	Sum $= 1$

ISBN 978-1-108-76630-2

Cambridge University Press

The five probabilities of course add up to 1, because no other outcomes are possible. This is also clear because the five probabilities are the successive terms in the binomial expansion of $\left(\frac{1}{6}+\frac{5}{6}\right)^4 = 1^4 = 1$,

$$\begin{aligned}\left(\tfrac{1}{6}+\tfrac{5}{6}\right)^4 = {}^4C_0 \times \left(\tfrac{1}{6}\right)^0 \times \left(\tfrac{5}{6}\right)^4 + {}^4C_1 \times \left(\tfrac{1}{6}\right)^1 \times \left(\tfrac{5}{6}\right)^3 + {}^4C_2 \times \left(\tfrac{1}{6}\right)^2 \times \left(\tfrac{5}{6}\right)^2 \\ + {}^4C_3 \times \left(\tfrac{1}{6}\right)^3 \times \left(\tfrac{5}{6}\right)^1 + {}^4C_4 \times \left(\tfrac{1}{6}\right)^4 \times \left(\tfrac{5}{6}\right)^0.\end{aligned}$$

Remember that

$$(x+y)^4 = {}^4C_0\,x^0y^4 + {}^4C_1\,x^1y^3 + {}^4C_2\,x^2y^2 + {}^4C_3\,x^3y^1 + {}^4C_4\,x^4y^0.$$

Binomial probability — the general case

Suppose that a multi-stage experiment consists of n identical stages, and at each stage the probability of 'success' is p and of 'failure' is q, where $p + q = 1$. Then

$$P(x \text{ successes and } n - x \text{ failures in that order}) = p^x q^{n-x}.$$

But there are ${}^nC_x$ ways of ordering x successes and $(n - x)$ failures, so

$$P(x \text{ successes and } n - x \text{ failures in any order}) = {}^nC_x\,p^x q^{n-x}.$$

This is the term in p^xq^{n-x} in the expansion of the binomial $(p + q)^n$.

1 BERNOULLI TRIALS AND BINOMIAL PROBABILITY

- A *Bernoulli trial* or *Bernoulli experiment* is an experiment with two outcomes, 'success' and 'failure', usually assigned the probabilities p and $q = 1 - p$.
- A *binomial random variable* is the outcomes of an n-stage experiment in which:
  — each stage is a Bernoulli trial with the same probability p of success,
  — the stages are independent,
  — the random variable X is the number of successes, not their order.
- Suppose that the probabilities of 'success' and 'failure' in any stage of an n-stage binomial experiment are p and $q = 1 - p$ respectively. Then

  $$P(x \text{ successes}) = {}^nC_x\,p^x q^{n-x}.$$
- This probability is the term in p^xq^{n-x} in the expansion of $(p + q)^n = 1^n = 1$, and the sum of the $n + 1$ binomial probabilities is 1.
- In particular, if $p = q = \frac{1}{2}$, then the formula simplifies to

  $$P\left(x \text{ successes}\right) = {}^nC_x \left(\tfrac{1}{2}\right)^n.$$

The pronumeral x is usually used for the values of a discrete random variable. In this chapter we will therefore mostly use x in place of the r that was used in Chapters 14–15 of the Year 11 book.

The next worked Example shows how to use complementary events and cases to answer questions.

Example 2 17A

Six cards are drawn at random from a pack of 52 playing cards. Each card is replaced and the pack is shuffled before the next card is drawn. Find, as fractions with denominator 4^6, the probability that:

a two are clubs,
b one is a club,
c at least one is a club,
d at least four are clubs.

SOLUTION

There are 13 clubs in the pack, so at each stage the probability of drawing a club is $\frac{1}{4}$. Applying the formula with $p = \frac{1}{4}$ and $q = \frac{3}{4}$:

a
$$\begin{aligned} P\left(\text{two are clubs}\right) &= {}^6C_2 \times \left(\tfrac{1}{4}\right)^2 \times \left(\tfrac{3}{4}\right)^4 \\ &= \frac{15 \times 3^4}{4^6} \\ &= \frac{1215}{4^6} \end{aligned}$$

b
$$\begin{aligned} P\left(\text{one is a club}\right) &= {}^6C_1 \times \left(\tfrac{1}{4}\right)^1 \times \left(\tfrac{3}{4}\right)^5 \\ &= \frac{6 \times 3^5}{4^6} \\ &= \frac{1458}{4^6} \end{aligned}$$

c
$$\begin{aligned} P\left(\text{at least one is a club}\right) &= 1 - P\left(\text{all are non-clubs}\right) \\ &= 1 - \left(\tfrac{3}{4}\right)^6 \quad \left(\text{or } 1 - {}^6C_0 \times \left(\tfrac{1}{4}\right)^0 \times \left(\tfrac{3}{4}\right)^6\right) \\ &= \frac{3367}{4^6} \end{aligned}$$

d $P\left(\text{at least four are clubs}\right)$
$$\begin{aligned} &= P\left(\text{four are clubs}\right) + P\left(\text{five are clubs}\right) + P\left(\text{six are clubs}\right) \\ &= {}^6C_4 \times \left(\tfrac{1}{4}\right)^4 \times \left(\tfrac{3}{4}\right)^2 + {}^6C_5 \times \left(\tfrac{1}{4}\right)^5 \times \left(\tfrac{3}{4}\right)^1 + {}^6C_6 \times \left(\tfrac{1}{4}\right)^6 \times \left(\tfrac{3}{4}\right)^0 \\ &= \frac{15 \times 3^2 + 6 \times 3 + 1}{4^6} \\ &= \frac{154}{4^6} \end{aligned}$$

An example where $p = q = \dfrac{1}{2}$

A particular case of binomial probability is when the probabilities p and q of 'success' and 'failure' are both $\frac{1}{2}$.

Example 3 17A

If a coin is tossed 100 times, what is the probability that it comes up heads exactly 50 times (correct to four significant figures)?

SOLUTION

Taking $p = q = \frac{1}{2}$,
$$\begin{aligned} P\left(50 \text{ heads}\right) &= {}^{100}C_{50} \times \left(\tfrac{1}{2}\right)^{50} \times \left(\tfrac{1}{2}\right)^{50} \\ &= {}^{100}C_{50} \times \left(\tfrac{1}{2}\right)^{100} \\ &\doteqdot 0.0796. \end{aligned}$$

Note: This is a fairly low probability. Should we have expected a higher probability than this? Hardly, because any result from about 45 to 55 heads would be unlikely to surprise us. In general, as already stated in Box 1 above,

$$P\big(x \text{ heads in } n \text{ tosses of a coin}\big) = {}^{n}\mathrm{C}_{x} \times \left(\tfrac{1}{2}\right)^{n}.$$

Experimental probabilities and binomial probability

Some of the most straightforward and important applications of binomial theory arise in situations where the probabilities of 'success' and 'failure' are determined experimentally.

Example 4 17A

A light bulb is classed as 'defective' if it burns out in under 1000 hours. A company making light bulbs finds, after careful testing, that 1% of its bulbs are defective. If it packs its bulbs in boxes of 50, find, correct to three significant figures:

a the probability that a box contains no defective bulbs,

b the probability that at least two bulbs in a box are defective.

SOLUTION

In this case, $p = 0.01$ and $q = 0.99$. Let X be the number of defective bulbs in the box. Then:

a $P(X = 0) = 0.99^{50} \doteqdot 0.605$

b
$$\begin{aligned} P(X \geq 2) &= 1 - \big(P(X = 0) + P(X = 1)\big) \\ &= 1 - \big(0.99^{50} + {}^{50}\mathrm{C}_1 \times 0.01^1 \times 0.99^{49}\big) \\ &\doteqdot 0.0894 \end{aligned}$$

An example where each stage is a compound event

Sometimes, when each stage of the experiment is itself a compound event, it may take some work to find the probability of success at each stage.

Example 5 17A

Joe King and his sister Fay make shirts for a living. Joe works more slowly, but more accurately, making 20 shirts a day, of which 2% are defective. Fay works faster, making 30 shirts a day, of which 4% are defective. If they send out their shirts in randomly mixed parcels of 30 shirts, what is the probability (correct to three significant figures) that no more than two shirts in a box are defective?

SOLUTION

If a shirt is chosen at random from one parcel, then using the product rule and the addition rule, the probability p that the shirt is defective is

$$\begin{aligned} p &= P\,(\text{Joe made it, and it is defective}) + P\,(\text{Fay made it, and it is defective}) \\ &= \frac{20}{50} \times \frac{2}{100} + \frac{30}{50} \times \frac{4}{100} \\ &= \frac{4}{125}, \end{aligned}$$

so $p = \frac{4}{125}$ and $q = \frac{121}{125}$.

Let X be the number of defective shirts.

Then $P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2)$

$$= \left(\tfrac{121}{125}\right)^{30} + {}^{30}C_1 \times \left(\tfrac{121}{125}\right)^{29} \times \tfrac{4}{125} + {}^{30}C_2 \times \left(\tfrac{121}{125}\right)^{28} \times \left(\tfrac{4}{125}\right)^2$$

$$\doteqdot 0.930.$$

Exercise 17A

FOUNDATION

Unless otherwise specified, leave your answers in unsimplified form.

1 Assume that the probability that a child is female is $\frac{1}{2}$, that sex is independent from child to child, and that there are only two sexes. Giving your answers as fractions in simplest form, find the probability that in a family of five children:

a all are boys,

b there are two girls and three boys,

c there are four boys and one girl,

d at least one will be a girl.

2 In a one-day cricket game, a batsman has a chance of $\frac{1}{5}$ of hitting a boundary every time he faces a ball. If he faces all six balls in an over, what is the probability that he will hit exactly two boundaries, assuming that successive strikes are independent?

3 A jury roll contains 2000 names, 700 of females and 1300 of males. Twelve jurors are randomly selected.

a Explain why it is reasonable to make the approximation that the probability of selecting a male does not change with each selection.

b What is the probability of ending up with an all-male jury?

4 A die is rolled twelve times. Find the probability that 5 appears on the uppermost face:

a exactly three times,

b exactly eight times,

c ten or more times (that is ten, eleven or twelve times).

5 A die is rolled six times. Let N denote the number of times that the number 3 is shown on the uppermost face. Find, correct to four decimal places:

a $P(N = 2)$

b $P(N < 2)$

c $P(N \ge 2)$

6 An archer finds that on average he hits the bulls-eye nine times out of ten. Assuming that successive attempts are independent, find the probability that in twenty attempts:

a he scores at least eighteen hits,

b he misses at least once.

7 A torch manufacturer finds that on average 9% of the bulbs are defective. What is the probability that in a randomly selected batch of ten one-bulb torches:

a there will be no more than two with defective bulbs,

b there will be at least two with defective bulbs.

ISBN 978-1-108-76630-2

Cambridge University Press

DEVELOPMENT

8 A coins is tossed four times and the result is recorded. Janice wins if there are exactly two heads.

a List all the ways that this could occur, that is, in any order.

b By counting, verify that there are exactly 6 cases where Janice wins and find the probability of this outcome.

c Explain why this is the same as the number of ways of ordering the word HHTT, and use combinatorics to count how many such arrangements exist. Does your answer agree with part **b**?

d Show that this is equivalent to choosing two of the coins and placing them heads up, while placing the other two coins tails up.

9 A poll indicates that 55% of people support the policies of the Working Together Party. If five people are selected at random, what is the probability that a majority of them will support WTP party policies? Give your answer correct to three decimal places.

10 The probability that a small earthquake occurs somewhere in the world on any one day is 0.95. Assuming that earthquake frequencies on successive days are independent (this assumption is probably false), what is the probability that a small earthquake occurs somewhere in the world on exactly 28 of January's 31 days? Leave your answer in index form.

11 The probability that a jackpot prize will be won in a given lottery is 0.012.

a Find, correct to five decimal places, the probability that the jackpot prize will be won:

i exactly once in ten independent lottery draws,

ii at least once in ten independent lottery draws.

b The jackpot prize is initially \$10 000 and increases by \$10 000 each time the prize is not won. Find, correct to five decimal places, the probability that the jackpot prize will exceed \$200 000 when it is finally won.

12 a How many times must a die be rolled so that the probability of rolling at least one six is greater than 95%?

b How many times must a coin be tossed so that the probability of tossing at least one tail is greater than 99%?

13 Five families have three children each.

a Find, correct to three decimal places, the probability that:

i at least one of these families has three boys,

ii each family has more boys than girls.

b What assumptions have been made in arriving at your answer?

14 Comment on the validity of the following arguments:

a 'In the McLaughlin Library, 10% of the books are mathematics books. Hence if I go to a shelf and choose five books from that shelf, then the probability that all five books are mathematics books is 10^{-5}.'

b 'During an election, 45% of voters voted for party A. Hence if I select a street at random, and then select a voter from each of four houses in the street, the probability that exactly two of those voters voted for party A is ${}^4C_2 \times (0.45)^2 \times (0.55)^2$.'

ISBN 978-1-108-76630-2 Cambridge University Press

15 During winter it rains on average 18 out of 30 days. Five winter days are selected at random. Find, correct to four decimal places, the probability that:

a the first two days chosen will be fine and the remainder wet,

b more rainy days than fine days have been chosen.

16 A tennis player finds that, on average, he gets his serve in in eight out of every ten attempts, and that he serves an ace (a serve which is in the boundaries and not touched by his opponent) once every fifteen serves. He serves four times. Assuming that successive serves are independent events, find, correct to six decimal places, the probability that:

a all four serves are in,

b he hits at least three aces,

c he hits exactly three aces and the other serve lands in.

17 A man is restoring ten old cars, six of them manufactured in 1955 and four of them manufactured in 1962. When he tries to start them, on average the 1955 models will start 65% of the time and the 1962 models will start 80% of the time. Find, correct to four decimal places, the probability that at any time:

a exactly three of the 1955 models and one of the 1962 models will start,

b exactly four of the cars will start. (Hint: You will need to consider five cases.)

18 An apple exporter deals in two types of apples, Red Delicious and Golden Delicious. The ratio of Red Delicious to Golden Delicious is 4:1. The apples are randomly mixed together before they are boxed. One in every fifty Golden Delicious and one in every one hundred Red Delicious apples will need to be discarded because they are undersized.

a What is the probability that an apple selected from a box will need to discarded?

b If ten apples are randomly selected from a box, find the probability that:

i all the apples will have to be discarded,

ii half of the apples will have to be discarded,

iii less than two apples will be discarded.

19 One bag contains three red and five white balls, and another bag contains four red and four white balls.

a One bag is chosen at random, a ball is selected from that bag, its colour is noted, and then it is replaced. Find the probability that the ball chosen is red.

b If the operation in part **a** is carried out eight times, find the probability that:

i exactly three red balls are drawn,

ii at least three red balls are drawn.

20 a If six dice are rolled one hundred times, how many times would you expect the number of even numbers showing to exceed the number of odd numbers showing?

b If eight coins are tossed sixty times, how many times would you expect the number of heads to exceed the number of tails?

ENRICHMENT

21 A game is played using a barrel containing twenty similar balls numbered 1 to 20. The game consists of drawing four balls, without replacement, from the twenty balls in the barrel. Thus the probability that any particular number is drawn in any game is 0.2.

a Find as a decimal the probability that the number 19 is drawn in exactly two of the next five games played.

b Find as a decimal the probability that the number 19 is drawn in at least two of the next five games played.

c Let n be an integer, where $4 \le n \le 20$.

i What is the probability that, in any one game, all four selected numbers are less than or equal to n?

ii Show that the probability that, in any one game, n is the largest of the four numbers drawn is $\dfrac{{}^{n-1}C_3}{{}^{20}C_4}$.

22 a Expand $(a + b + c)^3$.

b In a survey of football supporters, 65% supported Hawthorn, 24% followed Collingwood and 11% followed Sydney. Use the expansion in part a to find, correct to five decimal places, the probability that if three people are randomly selected:

i one supports Hawthorn, one supports Collingwood and one supports Sydney,

ii exactly two of them support Collingwood,

iii at least two of them support the same team.

17B Binomial distributions

Now that we have dealt with individual binomial probabilities, we can look at the whole distribution. The binomial distribution is a discrete distribution, and Chapter 13 of the Year 11 book was devoted to discrete distributions. As we have done before, we will draw its graph and work with its mean and variance.

Notation for binomial distributions

Some notation is convenient. The binomial distribution consisting of n independent trials, each with probability p of success, is denoted by $B(n, p)$ or $\text{Bin}(n, p)$. In particular, a Bernoulli experiment with probability p is a special case of a binomial experiment, so is denoted by $B(1, p)$.

To say that X is a binomial random variable for the distribution $B(n, p)$, we can use the symbolic notation

$$X \sim B(n,p) \qquad \text{OR} \qquad X \sim \text{Bin}(n,p).$$

In particular, $X \sim B(1, p)$ means that X is a Bernoulli random variable (there is only one stage) with probability p of success.

2 BINOMIAL AND BERNOULLI DISTRIBUTIONS

- The binomial distribution with n independent Bernoulli stages, each with probability p of success, is denoted by $B(n, p)$ or $\text{Bin}(n, p)$.
- The Bernoulli distribution with probability p is therefore $B(1, p)$.
- $X \sim B(n, p)$ means that X is a random variable for the distribution $B(n, p)$.

The mean and variance of a Bernoulli distribution

A Bernoulli trial with probabilities p of 'success' and $q = 1 - p$ of 'failure' has mean p and variance pq,

$$\mu = p \qquad \text{and} \qquad \sigma^2 = pq,$$

so that its standard deviation is $\sigma = \sqrt{pq}$.

These formulae are easy to prove because there are only two outcomes.

We use the formula $\mu = \sum xP(X = x)$,

where the sum is taken over the distribution, meaning for $x = 0$ and $x = 1$,

so
$$\begin{aligned} \mu &= 0 \times P(X = 0) + 1 \times P(X = 1) \\ &= 0 \times q + 1 \times p \\ &= p. \end{aligned}$$

Similarly,
$$\begin{aligned} \sigma^2 &= \sum x^2 P(X = x) - \mu^2, \text{ summing over the distribution,} \\ &= (0^2 \times q + 1^2 \times p) - p^2 \\ &= p - p^2 \\ &= p(1 - p) \\ &= pq \quad \text{(either form is an appropriate answer)}. \end{aligned}$$

ISBN 978-1-108-76630-2

Cambridge University Press

The mean and variance of a binomial distribution

The mean and variance of the binomial random variable $X \sim B(n, p)$ with n trials and probabilities p and q of 'success' and 'failure' are

$$\mu = np \qquad \text{and} \qquad \sigma^2 = npq.$$

so that the standard deviation is $\sigma = \sqrt{npq}$.

These two results $\mu = np$ and $\sigma^2 = npq$ seem to follow immediately from the results $\mu = p$ and $\sigma^2 = pq$ for Bernoulli trials just by multiplying by n. And they do — it is true in general that if we have a number of independent random variables, then the mean of the sum is the sum of the means, and the variance of the sum is the sum of the variances.

Unfortunately, however, those two very general theorems are too difficult to prove, and instead we must prove the result for binomial distributions directly. The proofs below are not conceptually difficult, but they do require a sequence of computations. We begin with a lemma about binomial coefficients.

Lemma: Let n and x be whole numbers with $x \le n$.

a If $x \ge 1$, then $\quad x \times {}^{n}\mathrm{C}_{x} = n \times {}^{n-1}\mathrm{C}_{x-1}$.

b If $x \ge 2$, then $\quad x(x-1) \times {}^{n}\mathrm{C}_{x} = n(n-2) \times {}^{n-2}\mathrm{C}_{x-2}$.

Proof:

a
$$\begin{aligned} x \times {}^{n}\mathrm{C}_{x} &= x \times \frac{n!}{x! \times (n-x)!} \\ &= \frac{x \times n \times (n-1)!}{x \times (x-1)! \times (n-x)!} \\ &= n \times \frac{(n-1)!}{(x-1)! \times (n-x)!} \\ &= \text{RHS} \end{aligned}$$

b
$$\begin{aligned} & x(x-1) \times {}^{n}\mathrm{C}_{x} \\ &= x(x-1) \times \frac{n!}{x! \times (n-x)!} \\ &= \frac{x(x-1) \times n(n-1) \times (n-2)!}{x(x-1) \times (x-2)! \times (n-x)!} \\ &= n(n-1) \times \frac{(n-2)!}{(x-2)! \times (n-x)!} \\ &= \text{RHS} \end{aligned}$$

Proving that the mean is *np*

This proof requires binomial expansions, together with part **a** of the lemma above.

$$\begin{aligned} \mathrm{E}(X) &= \sum xP(X = x), \text{ summed over the distribution,} \\ &= 0 \times {}^{n}\mathrm{C}_{0} \times p^{0}q^{n} + 1 \times {}^{n}\mathrm{C}_{1} \times p^{1}q^{n-1} + 2 \times {}^{n}\mathrm{C}_{2} \times p^{2}q^{n-2} + \cdots \\ &\qquad + n \times {}^{n}\mathrm{C}_{n} \times p^{n}q^{0}. \end{aligned}$$

The first term is 0, then we apply part **a** of the lemma to the remaining terms,

$$\begin{aligned} \mathrm{E}(X) &= n \times {}^{n-1}\mathrm{C}_{0}\, p^{1}q^{n-1} + n \times {}^{n-1}\mathrm{C}_{1}\, p^{2}q^{n-2} + \cdots + n \times {}^{n-1}\mathrm{C}_{n-1}\, p^{n}q^{0} \\ &= np\left({}^{n-1}\mathrm{C}_{0}\, p^{0}q^{n-1} + {}^{n-1}\mathrm{C}_{1}\, p^{1}q^{n-2} + \cdots + {}^{n-1}\mathrm{C}_{n-1}\, p^{n-1}q^{0}\right). \end{aligned}$$

The bit in brackets is the binomial expansion of $(p+q)^{n-1}$, where $p + q = 1$, so

$$\begin{aligned} \mathrm{E}(X) &= np(p+q)^{n-1} \\ &= np \times 1 \\ &= np. \end{aligned}$$

Proving that the variance is *npq*

This proof again requires binomial expansions. First, however, we need to develop yet another formula for the variance.

ISBN 978-1-108-76630-2 Cambridge University Press

$$\begin{aligned}\operatorname{Var}(X) &= \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2.\\ &= \mathrm{E}\big(X(X-1)\big) + \mathrm{E}(X) - \big(\mathrm{E}(X)\big)^2.\\ &= \mathrm{E}\big(X(X-1)\big) + np - n^2p^2,\ \text{ because } \mathrm{E}(X) = np.\end{aligned}$$

We can find $\mathrm{E}\big(X(X-1)\big)$ using binomial expansions and part **b** of the lemma,

$$\begin{aligned}\mathrm{E}\big(X(X-1)\big) &= \sum x(x-1)P(X=x), \text{summed over the distribution,}\\ &= 0 \times (-1) \times {}^{n}\mathrm{C}_0 \times p^0q^n + 1 \times 0 \times {}^{n}\mathrm{C}_1 \times p^1q^{n-1} + 2 \times 1 \times {}^{n}\mathrm{C}_2 \times p^2q^{n-2}\\ &\quad + 3 \times 2 \times {}^{n}\mathrm{C}_3 \times p^3q^{n-3} + \cdots + n(n-1) \times {}^{n}\mathrm{C}_n \times p^nq^0.\end{aligned}$$

The first two terms are 0, then applying part **b** of the lemma,

$$\begin{aligned}&= n(n-1) \times {}^{n-2}\mathrm{C}_0 \times p^2q^{n-2} + n(n-1) \times {}^{n-2}\mathrm{C}_1 \times p^3q^{n-3}\\ &\quad + n(n-1) \times {}^{n-2}\mathrm{C}_2 \times p^4q^{n-4} + \cdots + n(n-1) \times {}^{n-2}\mathrm{C}_{n-2} \times p^nq^0.\\ &= n(n-1)p^2({}^{n-2}\mathrm{C}_0\, p^0q^{n-2} + {}^{n-2}\mathrm{C}_1\, p^1q^{n-3} + + \cdots + {}^{n-2}\mathrm{C}_{n-2}\, p^{n-2}q^0).\end{aligned}$$

The bit in brackets is the binomial expansion of $(p+q)^{n-2}$, where $p+q=1$, so

$$\begin{aligned}&= n(n-1)p^2(p+q)^{n-2}\\ &= n^2p^2 - np^2.\end{aligned}$$

$$\begin{aligned}\text{Hence } \operatorname{Var}(X) &= \mathrm{E}\big(X(X-1)\big) + np - n^2p^2\\ &= n^2p^2 - np^2 + np - n^2p^2\\ &= np(1-p)\\ &= npq.\end{aligned}$$

3 MEAN AND VARIANCE OF A BINOMIAL DISTRIBUTION

- For a binomial random variable $X \sim B(n,p)$, where $q = 1 - p$,
  $\mu = np$ and $\sigma^2 = npq$ and $\sigma = \sqrt{npq}$.
- In particular, for a Bernoulli random variable $X \sim B(1,p)$,
  $\mu = p$ and $\sigma^2 = pq$ and $\sigma = \sqrt{pq}$.

The symbols $\mathrm{E}(X)$ and μ are interchangeable — μ is more concise, but the notation $\mathrm{E}(X)$ indicates that when we run the experiment, there is a sense in which we are 'expecting' to get $\mathrm{E}(X)$. Similarly, σ^2 and $\operatorname{Var}(X)$ are interchangeable.

Example 6 17B

A binomial random variable has parameters $n = 20$ and $p = 0.1$.

a Find the mean, variance and standard deviation.

b What is the probability of getting the mean when the experiment is run?

c What is the probability that the result is within one standard deviation of the mean?

d Give an example that this distribution could model.

SOLUTION

a $\mu = np$ $\qquad$ $\sigma^2 = npq \;\; (\text{where } q = 1 - p = 0.9)$ $\qquad$ $\sigma = \sqrt{1.8}$

$= 20 \times 0.1$ $\qquad$ $= 20 \times 0.1 \times 0.9$ $\qquad$ $\doteqdot 1.342$

$= 2$ $\qquad$ $= 1.8$

b
$$\begin{aligned} P(X = 2) &= {}^{20}C_2\, p^2 q^{18} \\ &= {}^{20}C_2 \times (0.1)^2 \times (0.9)^{18} \\ &\doteqdot 0.285. \end{aligned}$$

c
$$\begin{aligned} P(\mu - \sigma \le X \le \mu + \sigma) &= P(0.658 \le X \le 3.342) \\ &= P(X = 1 \text{ or } X = 2 \text{ or } X = 3) \\ &= {}^{20}C_1\, p^1 q^{19} + {}^{20}C_2\, p^2 q^{18} + {}^{20}C_3\, p^3 q^{17} \\ &= 20 \times (0.1) \times (0.9)^{19} + 190 \times (0.1)^2 \times (0.9)^{18} \\ &\qquad + 1140 \times (0.1)^3 \times (0.9)^{17} \\ &\doteqdot 0.745. \end{aligned}$$

d Choose a busy intersection with traffic lights on the way to work. On 20 mornings, when crossing at the lights, look at the number plate of the left-most front vehicle stopped at the lights, and record whether the last digit-character is a 7. In a small country town, it is just possible that the events may not be independent, but in a city, independence is virtually certain.

Example 7 17B

A binomial random variable has parameters $n = 100$ and $p = \frac{1}{5}$.

a Find the mean and standard deviation.

b Keeping $p = \frac{1}{5}$, what would n need to be increased to so that the standard deviation is less than 1% of the number n of trials?

c Keeping $n = 100$, what must p be decreased to so that the standard deviation is less than 1? Can p be increased to give a standard deviation less than 1?

SOLUTION

a $\mu = np$ $\qquad$ $\sigma^2 = npq \;\; (\text{where } q = 1 - p = \frac{4}{5})$ $\qquad$ $\sigma = \sqrt{16}$

$= 100 \times \frac{1}{5}$ $\qquad$ $= 100 \times \frac{1}{5} \times \frac{4}{5}$ $\qquad$ $= 4.$

$= 20,$ $\qquad$ $= 16,$

b
$$\begin{aligned} \text{Put} \qquad \sigma &< \frac{n}{100}. \\ \text{Squaring,} \quad npq &< \frac{n^2}{10\,000} \\ n^2 &> 10\,000 \times \tfrac{4}{25} \times n, \quad \text{because } pq = \tfrac{1}{5} \times \tfrac{4}{5} = \tfrac{4}{25}. \\ \text{Hence} \qquad n &> 1600, \quad \text{because } n \text{ is positive.} \end{aligned}$$

c
$$\begin{aligned} \text{Put} \qquad \sigma &< 1. \\ \text{Squaring,} \qquad npq &< 1 \\ 100p(1 - p) &< 1 \\ 100p^2 - 100p + 1 &> 0. \end{aligned}$$

ISBN 978-1-108-76630-2

The quadratic $100p^2 - 100p + 1 = 0$, with $a = 100,\ b = -100$ and $c = 1$, has axis of symmetry $p = \frac{1}{2}$ and discriminant $\Delta = 9600 = 40^2 \times 6$,

$$\text{so its roots are } p = \frac{100 - 40\sqrt{6}}{200} \quad \text{and} \quad p = \frac{100 + 40\sqrt{6}}{200}$$

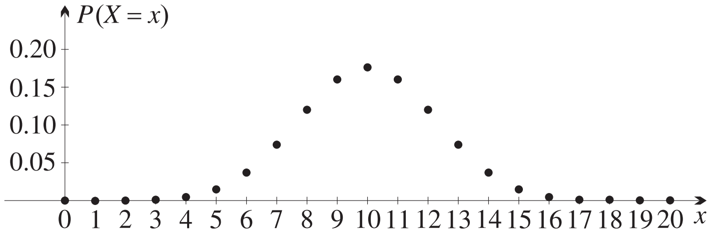
$$= \tfrac{1}{2} - \tfrac{1}{5}\sqrt{6} \qquad\qquad = \tfrac{1}{2} + \tfrac{1}{5}\sqrt{6}$$

$$\doteqdot 0.0101 \qquad\qquad \doteqdot 0.9899,$$

so p should be decreased to less than $\frac{1}{2} - \frac{1}{5}\sqrt{6}$ (roughly, less than 0.01), or increased to more than $\frac{1}{2} + \frac{1}{5}\sqrt{6}$ (roughly, greater than 9.99).

Skewed and non-skewed binomial distributions

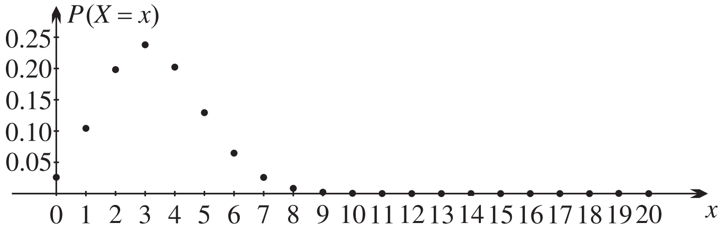
If $p = q = \frac{1}{2}$, such as when we toss a coin a number of times, the distribution is symmetric. For example, in Section 15E we graphed the distribution when a coin is tossed 20 times and the number of heads is recorded.



The mean is $\mu = np = 10$, and the distribution is symmetric about $x = 10$, with standard deviation $\sigma = \sqrt{npq} = \sqrt{5} \doteqdot 2.236$. This symmetry is easily seen from the symmetry of the binomial coefficient. For example, ${}^{20}C_7 = {}^{20}C_{13}$, so

$$P(X = 7) = {}^{20}C_7 \times \left(\tfrac{1}{2}\right)^7 \times \left(\tfrac{1}{2}\right)^{13} = {}^{20}C_{13} \times \left(\tfrac{1}{2}\right)^{13} \times \left(\tfrac{1}{2}\right)^7 = P(X = 13).$$

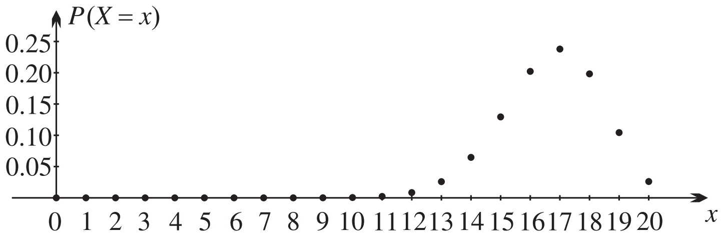
When, however, we throw a die 20 times, recording ‘success’ when the result is 6, the distribution is decidedly skewed to the right (positive skewness). Here $p = \frac{1}{6}$ and $q = \frac{5}{6}$, with mean $\mu = 3\frac{1}{3} \doteqdot 3.333$ and standard deviation $\sigma = \frac{5}{3} \doteqdot 1.667$.



Remember that *skewed to the right* means that there is a *longer tail on the right-hand side*, and the peak is on the left-hand side.

On the other hand, when we throw the die 20 times but record ‘success’ as ‘not getting 6’, then the distribution is skewed to the left (negative skewness). Here $p = \frac{5}{6}$ and $q = \frac{1}{6}$, with mean $\mu = 16\frac{2}{3} \doteqdot 16.667$ and the same standard deviation $\sigma = \frac{5}{3} \doteqdot 1.667$.

ISBN 978-1-108-76630-2 © Bill Pender et al. 2019 Cambridge University Press



Again, remember that *skewed to the left* means that there is a *longer tail on the left-hand side*, and the peak is on the right-hand side.

The graphs are reflections of each other in $x = 10$. For example,

$P(7\text{ heads}) = {}^{20}C_7 \times \left(\frac{1}{6}\right)^7 \times \left(\frac{5}{6}\right)^{13} = {}^{20}C_{13} \times \left(\frac{5}{6}\right)^{13} \times \left(\frac{1}{6}\right)^7 = P(13\text{ non-heads}).$

Question 11 in Exercise 17B shows how, for a fixed number of trials, the standard deviation decreases as the distribution becomes more skewed. Here is a simple example.

Example 8 17B

Two binomial random variables X and Y both consist of 100 Bernoulli trials, For X, $p = q = \frac{1}{2}$, and for Y, $p = \frac{1}{10}$. Find the ratio of their standard deviations.

SOLUTION

For X, $\sigma_X{}^2 = npq$
$= 100 \times \frac{1}{2} \times \frac{1}{2}$
$= 25,$
so $\sigma_X = 5.$

For Y, $\sigma_Y{}^2 = npq$
$= 100 \times \frac{1}{10} \times \frac{9}{10}$
$= 9,$
so $\sigma_X = 3.$

Hence the ratio of the two standard deviations is $\sigma_X : \sigma_Y = 5:3$.

4 SKEWED BINOMIAL DISTRIBUTIONS

Suppose that a binomial distribution consists of n independent Bernoulli trials, each with probability p of success and probability $q = 1 - p$ of failure.

- Reversing 'success' and 'failure' reverses the probabilities p and q, and reflects the graph in $x = \frac{1}{2}n$.
- If $p = q = \frac{1}{2}$, then the distribution is symmetric about $x = \frac{1}{2}$.
- If $p < \frac{1}{2}$, the distribution is skewed to the right (positive skewness).
  If $p > \frac{1}{2}$, the distribution is skewed to the left (negative skewness).
- For distributions with the same number of trials, the standard deviation gets closer to 0 or 1 as the skewness becomes more pronounced.

Simulating the experiment and graphing the data

The probabilities in a binomial distribution are estimates of what will happen when the experiment is run. Let us see what happens in simulations of two binomial experiments. The first experiment is symmetric, the second is skewed to the right.

Calculation and simulation in both worked examples should be done in a spreadsheet (the binomial function is currently `BINOM.DIST` in Excel).

Example 9 17B

The experiment is, 'toss 10 coins and count the number of heads'.

a Find the mean and standard deviation, and complete a table of the probability distribution.

b Simulate the experiment 100 times, complete the table of relative frequencies, and find the mean and standard deviation of the data.

c Draw the theoretical frequency histograms and polygon from part **a**. Then draw the frequency histograms and polygon of the simulation in part **b**.

SOLUTION

a Here $p = q = \frac{1}{2}$ and $n = 10$, so

$$\mu = np = 5, \qquad \sigma^2 = npq = 2.5, \qquad \sigma = \sqrt{2.5} \doteqdot 1.58.$$

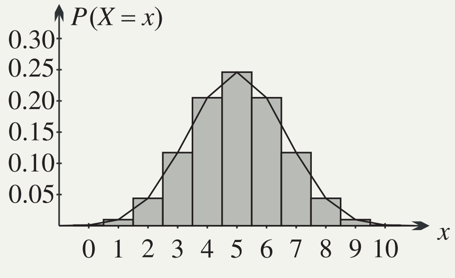
x	0	1	2	3	4	5	6	7	8	9	10
$P(X = x)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

b After running the experiment 100 times,

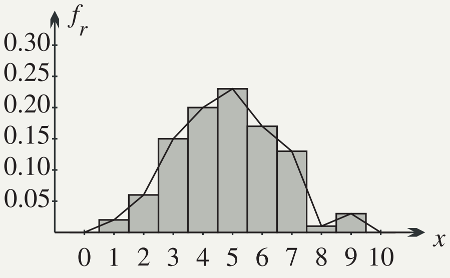
x	0	1	2	3	4	5	6	7	8	9	10
f	0	2	6	15	20	23	17	13	1	3	0
f_r	0	0.02	0.06	0.15	0.2	0.23	0.17	0.13	0.01	0.03	0

Calculation gives $\bar{x} = 4.82$ and $s \doteqdot 1.700$.

c



The theoretical distribution



The simulation

ISBN 978-1-108-76630-2

Cambridge University Press

Example 10 17B

The experiment is, ‘roll 10 dice and count the number of sixes’.

a Find the mean and standard deviation, and complete a table of the probability distribution.

b Simulate the experiment 100 times, complete the table of relative frequencies, and find the mean and standard deviation of the data.

c Draw the theoretical frequency histograms and polygon from part **a**. Then draw the frequency histograms and polygon of the simulation in part **b**.

SOLUTION

a Here $p = \frac{1}{6}$ and $q = \frac{5}{6}$ and $n = 10$, so

$\mu = np = 1.667, \qquad \sigma^2 = npq = 1.389, \qquad \sigma \doteqdot 1.179.$

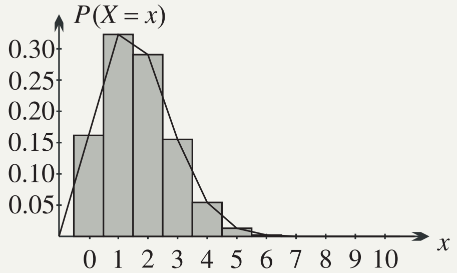
x	0	1	2	3	4	5	6	7	8	9	10
$P(X = x)$	0.162	0.323	0.291	0.155	0.054	0.013	0.002	0.0002	0.00002	0	0

b After running the experiment 100 times,

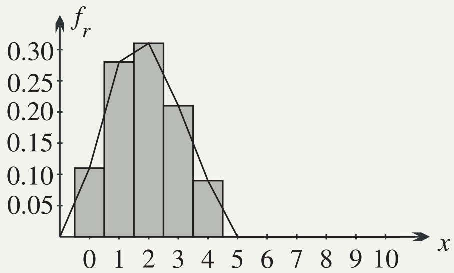
x	0	1	2	3	4	5	6	7	8	9	10
f	11	28	31	21	9	0	0	0	0	0	0
f_r	0.11	0.28	0.31	0.21	0.09	0.00	0.00	0.00	0.00	0.00	0.00

Calculation gives $\bar{x} = 1.89$ and $s \doteqdot 1.13$.

c



The theoretical distribution



The simulation

Exercise 17B FOUNDATION

1 Identify which of the following experiments may be modelled using a binomial random variable. If so, identify the variable.

- **a** It rains four days in a row. The number of rainy days is recorded.
- **b** A die is thrown ten times, and after each throw it is recorded whether the result is less than five.
- **c** Maddy is playing a simple game of cards. She turns up a card from the top of the pack. If it is an ace of spades she wins. Otherwise the card is returned to the pack, which is shuffled. The number of plays until she wins is recorded.
- **d** The probability of a head on tossing an unfair coin is 0.4. A coin is tossed twenty times and the number of tails is recorded.
- **e** Quality control testers have been given a random sample of 20 pens from a batch. It is known that 3% of the pens in the batch are defective. The testers record the number of faulty pens in the sample.

f A pupil records the time for his journey to school over a thirty-day period.

g At stage n an experimenter selects one number from 1–100 and records a success if the number matches n. This experiment is repeated 50 times (that is, up to $n = 50$) and the experimenter wishes to know how many successes occur.

2 In a simple experiment, 6 fair coins are tossed and the number of heads is recorded. A student wishes to compare the results with theoretical predictions.

a Copy and complete the table below using the formula $P(X = x) = {}^nC_x p^x q^{n-x}$ for binomial probability.

Number of heads	0	1	2	3	4	5	6	Total
Number of ways can occur probability								

b Read off the mode (the most common result).

c Use this table to determine the expected value and variance of this discrete probability distribution.

d Compare your results to those obtained using the two formulae $\text{E}(X) = np$ and $\text{Var}(X) = npq$, where $q = 1 - p$.

e Explain, in your own words, why the result for the expected value is not a surprise.

3 For each situation, construct a table for the theoretical distribution associated with the given random variable and calculate the mode, mean and standard deviation.

a A coin is tossed five times and X is the number of heads that are face up.

b A die is thrown five times and X is the number of times 5 or 6 occurs.

c Five cards are drawn in turn with replacement from a standard pack of 52 cards and X is the number of court cards (jack, queen, king) that are drawn. Give your result correct to 3 decimal places.

4 At the School fête, few people have entered any of the 24 meat raffles, so to support the school, Larry buys 8 of the 20 tickets in each raffle.

a Calculate the expected value and standard deviation for the random variable of the number of Larry's wins.

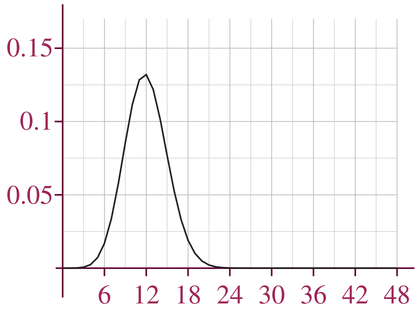
b If Larry wins 6 raffles, calculate how many standard deviations Larry's result is below the mean, as a measure of his poor luck.

5 **a** If a player throws 6 dice, what is the probability of getting at least two sixes? Let the random variable X record the number of sixes. What is the expected value and standard deviation for X?

b Repeat the question for:

i 12 dice, **ii** 24 dice.

6 The graph to the right shows the relative frequency polygon for the binomial distribution with $n = 48$ and $p = 0.25$.



a Is the distribution skewed to the left (negatively) or to the right (positively) or not at all?

b Find the mean and standard deviation for the distribution.

c What is the mode of the distribution? Estimate the probability of this most likely outcome.

d Copy the graph and shade the region no more than two standard deviations from the mean.

e Use the given sketch for $p = 0.25$ to assist in sketching the graph of the distribution for $p = 0.75$.

7 [Technology]
Large binomial calculations can be done in a spreadsheet, as in the following question. This question extends over two exercises — it continues in Exercise 17C — so be careful to save your spreadsheet for later use.

A student wishes to use Excel to explore the theoretical probabilities that occur when a coin is tossed 100 times.

- **a** In cell `F1` enter 100 (for the number of trials), and in cell `F2` enter 0.5 (for the probability p of obtaining a head). Add suitable labels in `E1` and `E2`.
- **b** Enter the numbers 0–100 in cells `A0:A101`.
- **c** In cell `B1` enter the formula = `BINOM.DIST($A1,$F$1,$F$2,FALSE)` to calculate the binomial probability of `A1` heads in 100 tosses for $p = 0.5$.
- **d** Use `Fill Down` to fill `B1` down to the first 101 rows in the second column.
- **e** What is the probability of obtaining at least 60 heads? (Select `B61:B101` and your spreadsheet program may show the sum on a bottom status line. Otherwise use a `SUM` command).
- **f** What is the probability of obtaining between 30 and 55 heads inclusive?
- **g** What is the mode?
- **h** Find the smallest integer i such that more than 50% of the data lies in the range $[50 - i, 50 + i]$.
- **i** Draw a histogram of the data in column `B`, with column `A` as the x-axis labels (this may well be the default). What famous shape does it remind you of?
- **j** Save all your work for future use in Exercise 17C.

DEVELOPMENT

8 Ms Taylor sets her class a test of 48 multiple-choice questions, each with four options A–D. Let the random variable X be the number of questions a person gets correct.

- **a** Calculate the expected value $\text{E}(X)$. How can this value be understood in this context?
- **b** What is the standard deviation σ of the random variable?
- **c** Ms Taylor is annoyed to discover that Fayola, one of her students, claims to have got 24 just by guessing. How many standard deviations is Fayola's score above the mean?

Ms Taylor writes a new test, with 100 questions, each with five options A–E.

- **d** What is the expected value and standard deviation of this new distribution with random variable Y?
- **e** Fayola gets 40 questions right this time, and again she claims to have achieved this result just by guessing. If true, would this be more or less unusual than her previous result?
- **f** Idette gets 75% in the first test and 60% in the second. Which is the more unusual of her results, if she is guessing?

9 A certain drug is found to be helpful for 70% of patients who take it. Two research teams are attempting to improve its effectiveness by adjusting the delivery system for the drug. Both conduct random trials to test the effectiveness of their changes. Team A runs a trial with 50 patients, of whom 45 show improvement using the drug. Team B runs a trial with 90 patients, of whom 74 show improvement. Use the mean and standard deviation of the two trials to decide which adjustment shows stronger evidence of an improvement.

10 It is known that 15% of a large constituency voted for the Working Together Party (WTP) at the last election. A poll of 100 people is taken and they are asked whether they voted WTP.

a What is the mean and standard deviation for this poll?

b What is the probability that in the sample, the number of WTP voters lies within half a standard deviation of the mean?

11 In this question, take the number n of stages of a binomial variable $X \sim B(n, p)$ to be fixed, and allow p to vary.

a Find σ^2 as a quadratic in p, and graph it.

b Explain the symmetry of the graph.

c Show that $\sigma^2 < np$ and $\sigma^2 < nq$.

d Show that the maximum value of σ^2 is $\dfrac{n}{4}$.

e Let n be fixed. Show that $\sigma \to 0$ as $p \to 0^+$ and as $p \to 1^-$.

12 A fair coin is tossed n times. The histogram for the resulting binomial distribution is labelled $0,\ 1,\ 2, \ldots,\ n$ on the horizontal axis, and each column is 1 unit wide. How many columns are entirely contained in the interval one standard deviation or less from the mean, when n is:

a 16, **b** 36, **c** 64?

ISBN 978-1-108-76630-2

Cambridge University Press

ENRICHMENT

13 Continuing with Question 12 above, as p moves away from 0.5, the standard deviation, and hence the number of column contained in 1 standard deviation, decreases. Find the limit of ratio of the number of columns entirely contained in 1 standard deviation for $p = 0.5$ to the number contained for $p = 0.25$, as $n \to \infty$.

14 In experimental trials, it may be useful to ensure that there is at least one positive result. For example, the trial may be difficult or expensive to run, and at least one successful result may be required. Suppose that in a certain set of n independent Bernoulli trials, the probability of success at each stage is p.

a Show that the probability of obtaining at least one success is $P(X > 1) = 1 - (1 - p)^n$.

b Show that the number of trials required to ensure that the probability of obtaining a success is at least 95% is $n = \dfrac{\log(0.05)}{\log(1 - p)}$.

c Copy and complete the table below to show the number of trials required to ensure at least 95% probability of success for the given value of p.

p	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.5
n									

d A certain journal refuses to publish experimental studies that do not yield a positive result, meaning a result agreeing with the experimental hypothesis. Explain why this is a dangerous practice.

15 If a Bernoulli trial occurs a fixed number n of times, and if the stages are independent, then the resulting distribution is a *binomial distribution*. If instead the experiment continues until a success is obtained and the number of trials is recorded, the resulting distribution is called a *geometric distribution* (and is also called a *discrete waiting-time distribution*).

Suppose that at each stage the probability of success is p and the probability of failure is $q = 1 - p$, and that X is the first trial producing a success. Thus the possible values of X are $1, 2, 3, \ldots$.

a For this geometric distribution:

i Show that $P(X = x) = pq^{x-1}$.

ii Show that $\mu = p + 2pq + 3pq^2 + \cdots$.

iii Show that $\mu - q\mu = p + pq + pq^2 + \cdots$, and hence calculate μ.

b During the day at a certain medical practice, the waiting room is constantly at capacity, with 20 patients waiting to be seen by a doctor. Every 5 minutes, a new patient will be chosen at random to see a doctor and a new patient will arrive at the practice. What is the mean waiting time to see a doctor?

17C Normal approximations to a binomial

A pollster asks a sample of 10 000 people whether they intend to vote for the Working Together Party in the next election. The WTP won $\frac{1}{3}$ of the vote in the last election, and using the assumption that this will be the case in the next election, he wants to know the probability that his survey will give him a result within [3300, 3400].

Using binomial probability, the calculation is

$$\begin{aligned} P(3300 \le X \le 3400) &= {}^{10\,000}C_{3300}\left(\tfrac{1}{3}\right)^{3300}\left(\tfrac{2}{3}\right)^{6700} + {}^{10\,000}C_{3301}\left(\tfrac{1}{3}\right)^{3301}\left(\tfrac{2}{3}\right)^{6699} \\ &\quad + \cdots + {}^{10\,000}C_{3399}\left(\tfrac{1}{3}\right)^{3399}\left(\tfrac{2}{3}\right)^{6601} + {}^{10\,000}C_{3400}\left(\tfrac{1}{3}\right)^{3400}\left(\tfrac{2}{3}\right)^{6600}. \end{aligned}$$

This is a dreadful calculation because of all the huge factorials and massive sizes of the numbers. Even with a computer, there are serious problems about controlling all the errors in the calculation. And if the calculation involves the whole population of Australia — over 25 000 000 — the situation is even more difficult.

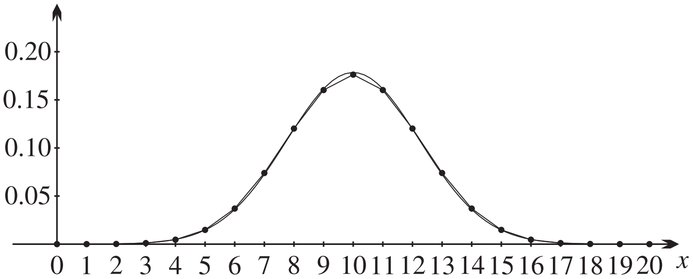
Fortunately, binomial distributions can be approximated very well by the normal distribution, and the larger the numbers are, the better the approximation. This section is about approximating a binomial distribution by a normal distribution.

An example of approximations

In Section 16E, we took the binomial distribution of the number of heads in 20 coin tosses, where $n = 20$ and $p = q = \frac{1}{2}$, so that

$$\mu = np = 20 \times \tfrac{1}{2} = 10 \quad \text{and} \quad \sigma^2 = npq = 20 \times \tfrac{1}{2} \times \tfrac{1}{2} = 5,$$

and $\sigma = \sqrt{5}$. We graphed the frequency polygon, then we superposed a normal curve with the same mean $\mu = 10$ and the same standard deviation $\sigma = \sqrt{5}$. The result looks reasonably good,



Thus the normal PDF approximates the frequency polygon of the binomial, and taking cumulative frequencies, it follows that the normal CDF approximates the ogive of the binomial. We cannot in this course give theoretical arguments why such approximations work — we can only look at the pictures and confirm the results. Then we can use these results in problems.

For example, the standard deviation is $\sqrt{5} \doteqdot 2.236$. Suppose that we want to find the probability that when we toss 20 coins, the number of heads will be within one standard deviation of the mean. These numbers are small enough for us to calculate it two ways, as in the next worked example.

Notation for normal distributions

The notation $N(\mu, \sigma^2)$ is used for the normal distribution with mean μ and variance σ^2. Thus in the diagram above, we are claiming that the binomial distribution $B\left(20, \frac{1}{2}\right)$ is approximated by the normal distribution $N(10, 5)$. In general, $B(n, p)$ is approximated by $N(np, npq)$, where $q = 1 - p$.

Example 11 17C

A coin is tossed 20 times, and X is the number of heads. Find the probability that the number of heads is greater than 12:

a using a binomial distribution,

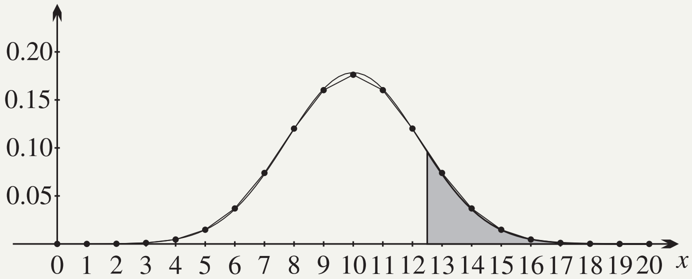
b using the normal approximation.

SOLUTION

a
$$\begin{aligned} P(\text{at least 13 heads}) &= {}^{20}C_{13}\left(\tfrac{1}{2}\right)^{20} + {}^{20}C_{14}\left(\tfrac{1}{2}\right)^{20} + \cdots + {}^{20}C_{20}\left(\tfrac{1}{2}\right)^{20} \\ &= \left(\tfrac{1}{2}\right)^{20}\left({}^{20}C_{13} + {}^{20}C_{14} + \cdots + {}^{20}C_{20}\right) \\ &= \left(\tfrac{1}{2}\right)^{20} \times 137\,980 \\ &\doteqdot 0.132. \end{aligned}$$

b To approximate the binomial by the normal distribution $N(10, 5)$, we treat the discrete random variable *X as if it were a continuous normal variable*. Because we are using a continuous density function, we need to follow the boundaries of the histograms and find the probability, not that $X \geq 13$, but that $X \geq 12.5$, which is the left-hand boundary of the first histogram.

$$P(\text{at least 13 heads}) \doteqdot P(X \geq 12.5).$$



This adjustment from 13 to 12.5 is called the *continuity correction* — it is the correction needed when a discrete distribution is treated as if it were continuous.

Using z-scores,
$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &\doteqdot 1.118, \end{aligned}$$

so
$$\begin{aligned} P(\text{at least 13 heads}) &\doteqdot P(Z > 1.118) \\ &\doteqdot 1 - \phi(1.118), \\ &\doteqdot 1 - 0.881 \\ &\doteqdot 0.132. \end{aligned}$$

Continuity corrections

We are approximating a discrete binomial variable by a continuous normal variable. We thus integrate from halfway between the values of the discrete distribution. In the example above, that means from 12.5.

In the next worked example, we are adding the five values at $x = 8,\ 9,\ 10,\ 11$ and 12, so we integrate over the interval $[7.5, 12.5]$.

Example 12 17C

A coin is tossed 20 times, and X is the number of heads. Find the probability that the number of heads is within one standard deviation of the mean:

a using a binomial distribution,

b using the normal approximation.

Comment on the results of the calculations.

SOLUTION

a $P(X \text{ is within one standard deviation of the mean}) = P(7.764 \le X \le 12.236)$

$$= P(X = 8,\ 9,\ 10,\ 11 \text{ or } 12)$$

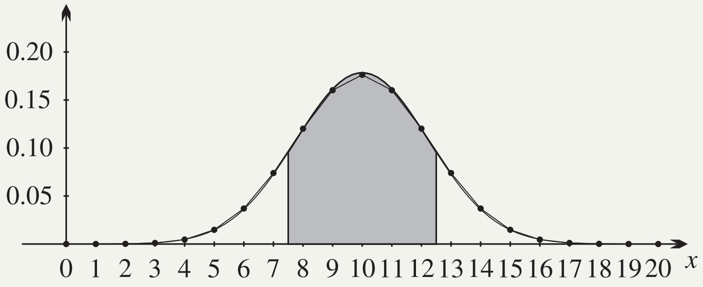
$$= {}^{20}C_8 \left(\tfrac{1}{2}\right)^{20} + {}^{20}C_9 \left(\tfrac{1}{2}\right)^{20} + {}^{20}C_{10} \left(\tfrac{1}{2}\right)^{20} + {}^{20}C_{11} \left(\tfrac{1}{2}\right)^{20} + {}^{20}C_{12} \left(\tfrac{1}{2}\right)^{20}$$

$$= \frac{{}^{20}C_8 + {}^{20}C_9 + {}^{20}C_{10} + {}^{20}C_{11} + {}^{20}C_{12}}{2^{20}}$$

$$= \frac{772\,616}{1024 \times 1024}$$

$$\doteqdot 0.737.$$

b We are approximating the binomial distribution by the normal distribution $N(10, 5)$, so we treat the discrete random X as if it were a continuous normal variable, and calculate $P(7.5 \le X \le 12.5)$.



Using z-scores, the two limits of integration are

$$z = \frac{x - \mu}{\sigma} = \frac{12.5 - 10}{\sqrt{5}} = 1.118, \qquad z = \frac{x - \mu}{\sigma} = \frac{7.5 - 10}{\sqrt{5}} = -1.118,$$

so $P(X \text{ is within one standard deviation of the mean}) \doteqdot P(-1.118 \le Z \le 1.118)$

$$\doteqdot 0.736.$$

The results in parts **a** and **b** are in excellent agreement.

ISBN 978-1-108-76630-2

5 NORMAL APPROXIMATION TO A BINOMIAL DISTRIBUTION

- A binomial distribution can be approximated by the normal distribution with the same mean and standard deviation.
- Thus the binomial distribution $B(n, p)$ or $\text{Bin}(n, p)$ can be approximated by the normal distribution $N(np, npq)$, where $q = 1 - p$.
- $N(\mu, \sigma^2)$ means the normal distribution with mean μ and variance σ^2.
- When approximating, we treat the discrete binomial variable X as if it were a continuous normal variable.
- For small values of n, apply the *continuity correction*. This means integrating between half-intervals, corresponding to the boundaries of the cumulative frequency histogram. For example, to approximate $P\,(X = 8,\ 9,\ 10,\ 11 \text{ or } 12)$, we treat X as a continuous normal variable and find
  $P(7.5 \le X \le 12.5)$.

The continuity correction for large numbers

In the example in the introduction to this chapter, the numbers are large enough so that the continuity correction is negligible, as the next worked example shows.

Example 13 17C

Compute $P(3300 \le X \le 3400)$ in the example in the introduction:

a ignoring the continuity correction, **b** using the continuity correction.

Then comment on the results of the calculations.

SOLUTION

Here $n = 10\,000$ and $p = \frac{1}{3}$, so $\mu = 3333\frac{1}{3}$ and $\sigma^2 = 2222\frac{2}{9}$ and $\sigma \doteqdot 47.1416$.
We will treat the discrete random variable X as if it were a continuous normal variable with the same mean $3333\frac{1}{3}$ and standard deviation 47.1416.

a Without the continuity correction, we are finding $P(3300 \le X \le 3400)$.
The z scores are -0.7071 and 1.4142,

$$\begin{aligned}\text{so}\quad P(3300 \le X \le 3400) &\doteqdot P(-0.7071 \le Z \le 1.4142)\\ &\doteqdot 0.6816.\end{aligned}$$

b With the continuity correction, we are finding $P(3299.5 \le X \le 3400.5)$.
The z scores are -0.7177 and 1.4248,

$$\begin{aligned}\text{so}\quad P(3299.5 \le X \le 3400.5) &\doteqdot P(-0.7177 \le Z \le 1.4248)\\ &\doteqdot 0.6859.\end{aligned}$$

No pollster would ever need the third decimal place of the probabilities, so there is no problem whatsoever ignoring the continuity correction.

ISBN 978-1-108-76630-2

Using the normal approximation to calculate a binomial coefficient

One of the consequences of the normal approximation to the binomial is that for large parameters, an individual binomial coefficient can be approximated. These calculations demonstrate very well how vital the continuous correction can be.

Example 14 17C

Approximate ${}^{100}\mathrm{C}_{53}$ using the normal approximation to the binomial distribution $B\left(100, \frac{1}{2}\right)$.
Check how good the approximation is.

SOLUTION

We know that ${}^{100}\mathrm{C}_{53} \times \left(\frac{1}{2}\right)^{100} = P(X = 53)$.
We treat the discrete binomial variable X as if it were a continuous normal variable,

$$ {}^{100}\mathrm{C}_{53} \times \left(\tfrac{1}{2}\right)^{100} = P(52.5 \le X \le 53.5). $$

The binomial distribution has mean $\mu = np = 50$ and SD $\sigma = \sqrt{npq} = \sqrt{25} = 5$.
So also does the normal approximation, and the limits have z-scores 0.5 and 0.7.

$$\begin{aligned} \text{Hence} \quad {}^{100}\mathrm{C}_{53} \times \left(\tfrac{1}{2}\right)^{100} &\doteqdot P(0.5 \le Z \le 0.7) \\ &\doteqdot 0.066574 \\ {}^{100}\mathrm{C}_{53} &\doteqdot 2^{100} \times 0.066574 \\ &\doteqdot 8.438 \times 10^{28}. \end{aligned}$$

This compares with ${}^{100}\mathrm{C}_{53} \doteqdot 8.441 \times 10^{28}$ (all according to Excel).

Sampling with replacement — when is a survey a binomial experiment?

Suppose that I choose ten cards in succession *without replacement* from a pack of 52 cards and count the number of aces. This is not a binomial experiment because the probabilities of getting an ace change as each card is removed from the pack. If *replace the card each time*, however, then I have a binomial experiment.

More generally, a survey that questions n people chosen at random from N people without replacment is not a binomial experiment. A survey that chooses each person independently of all previous choices, however, thus allowing the same person possibly to be questioned more than once, is a binomial experiment because the stages are identical and independent.

Now suppose that the pool of N people from which the sample is chosen is very large. This allows us to ignore these distinctions, because whether replacement is used or not, the probabilities will be virtually the same in all stages. This is certainly the case with a sample chosen from all Australians, but in a college of say 3000 people and a sample that is small relative to 3000, the effect is also almost negligible.

How good an approximation is the normal?

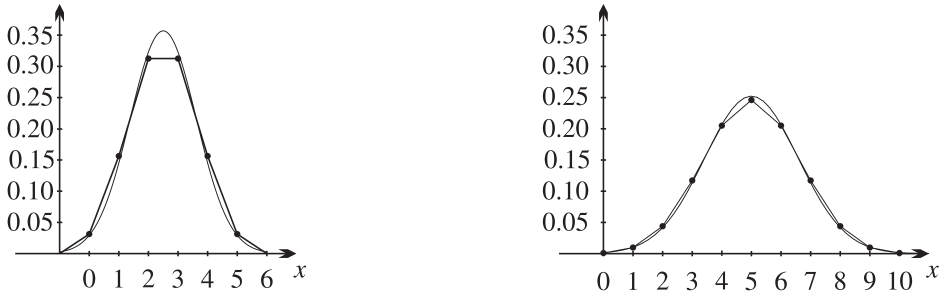
This depends on how much accuracy is needed. The diagrams below show that n should be larger for a skewed distribution, meaning that p is near 0 or 1, than when the distribution is reasonably symmetric, meaning that p is near 0.5. A common rule of thumb therefore is to require

$$np > 5 \qquad \text{and} \qquad nq > 5,$$

but numbers other than 5 are often stated, and there are other tests.

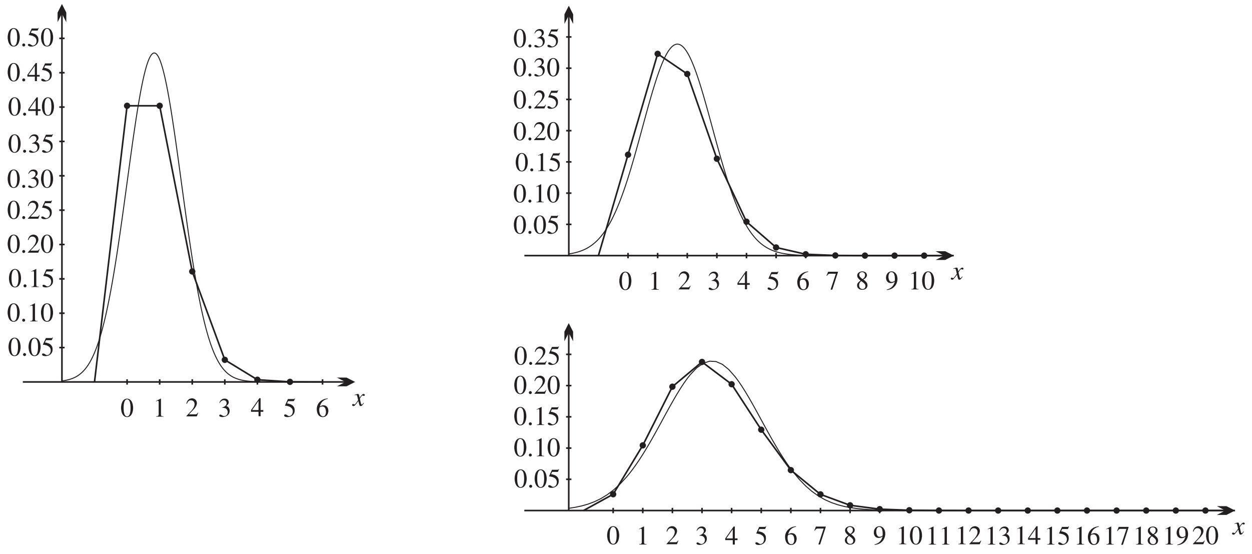
When $p = \frac{1}{2}$, we have already drawn the approximation when $n = 20$. Here are the ogive and the normal approximation, still for $p = \frac{1}{2}$, when $n = 5$ and $n = 10$.

a



When $p = \frac{1}{6}$, which arises when throwing a die and recording if it is a six, here are the graphs for $n = 5$, $n = 10$ and $n = 20$

b



6 A RULE OF THUMB FOR USING THE NORMAL APPROXIMATION

- For a fixed number n of Bernoulli trials, the normal approximation to a binomial distribution is less satisfactory when the distribution is skewed.
- For a fixed probability p, the normal approximation to a binomial distribution is more satisfactory when the number of trials is large.
- A common rule of thumb — one amongst many — for using the normal approximation to a binomial distribution is to require

$$np > 5 \qquad \text{and} \qquad nq > 5,$$

The normal approximation to a binomial distribution is a special case

We remarked at the start of Section 16G in the last chapter that the normal approximation to a binomial distribution was one of the first examples of a very general theorem in statistics called the *central limit theorem*. This would be a good occasion to read again the short account of the central limit limit theorem given there.

Exercise 17C

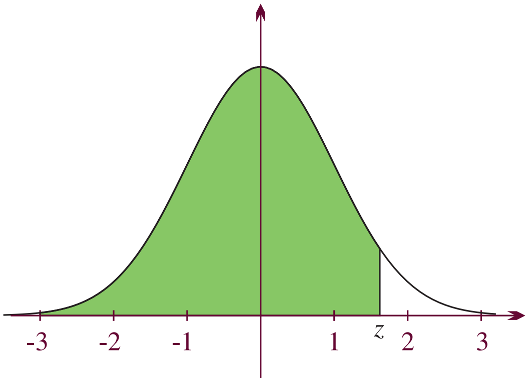
FOUNDATION

The standard normal probability distribution

The shaded area in the graph to the right represents a value of the *cumulative standard normal distribution function*

$$P(Z \le z) = \phi(z) = \int_{-\infty}^{z} \phi(t)\,dt.$$

The table below gives some further values of the probabilities $P(Z \le z) = \phi(z)$, allowing two decimal places for z — after that, use interpolation. For example,



$$P(Z \le 1.627) = \phi(1.627) = \int_{-\infty}^{1.627} \phi(z)\,dz \doteqdot 0.9474 + 0.7\,(0.9484 - 0.9474) \doteqdot 0.9481.$$

	second decimal place									
z	+ .00	+ .01	+ .02	+ .03	+ .04	+ .05	+ .06	+ .07	+ .08	+ .09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916

ISBN 978-1-108-76630-2

Cambridge University Press

	second decimal place									
z	+ .00	+ .01	+ .02	+ .03	+ .04	+ .05	+ .06	+ .07	+ .08	+ .09
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Note: You may find in this exercise and the next that your answers differ slightly from the text, depending whether or not you interpolate, and whether you use the supplied tables or alternatives such as statistical calculators and spreadsheets that provide more accurate values.

1 A certain binomial distribution for the random variable X consists of 20 independent trials, each with probability of success $p = 0.3$.

a Write this binomial distribution in symbolic form.

b Calculate the probability of obtaining 9, 10 or 11 successes.

c Confirm that $np > 5$ and $nq > 5$, suggesting that a normal approximation to the binomial could be used to estimate this probability.

d Calculate the normal approximation by determining $P(8.5 \leq X \leq 11.5)$, treating X as approximately $N(\mu, \sigma^2)$, and using the normal tables at the start of this exercise.

e Find the percentage error in the normal approximation. Does the approximation seems fairly accurate for this value of n and p?

2 Repeat the steps of Question **1** for these cases.

a $n = 50$, $p = 0.5$, find $P(18 \leq X \leq 20)$,

b $n = 20$, $p = 0.4$, find $P(8 \leq X \leq 9)$,

c $n = 30$, $p = 0.3$, find $P(5 \leq X \leq 7)$,

d $n = 40$, $p = 0.2$, find $P(9 \leq X \leq 12)$,

e $n = 22$, $p = 0.6$, find $P(13 \leq X \leq 15)$,

f $n = 80$, $p = 0.1$, find $P(10 \leq X \leq 13)$,

g $n = 500$, $p = 0.25$, find $P(100 \leq X \leq 103)$,

h $n = 200$, $p = 0.9$, find $P(170 \leq X \leq 172)$.

3 A barrel contains 600 pink and 400 blue counters. At each stage of an experiment, the barrel is stirred well, then a counter is removed, its colour is noted, and it is returned to the barrel. This experiment is repeated 20 times.

a Explain why each stage of the experiment is a Bernoulli trial.

b Explain why the full experiment is binomial.

c Is it necessary to return the counter after each draw?

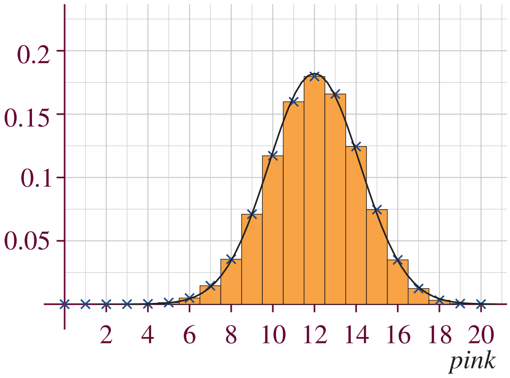
d Write down the probability of drawing a pink counter, and find the mean and standard deviation for this binomial distribution.

e Find the probability of drawing exactly 14 pink counters, correct to 3 decimal places.

f A student constructs a histogram for this distribution and overlays the normal distribution with the same mean and standard deviation. He notes the strong agreement, and to test this he uses his standard normal tables to calculate the area corresponding to drawing 14 pink counters.

i Explain why the required area for the normal distribution with random variable X is $P(13.5 < X < 14.5)$.

ii Hence find this area. Is it in strong agreement with your answer to part **e**?



4 In a college of 3000 pupils, 1320 are girls. A teacher selects a group of 15 pupils at random from the college rolls, without revealing the result, and asks her students to determine the probability that the group has more than 8 girls in it.

a Write down the probability p of selecting a girl from the population of 3000.

b Write down the mean and standard deviation of the binomial distribution obtained by selecting 15 pupils from the population.

c Use the exact binomial distribution to determine the probability of obtaining 9 or 10 girls.

d Is the sample size big enough to use the normal approximation to the binomial? Use the criterion $np > 5$ and $n(1 - p) > 5$.

e Use the normal approximation to the binomial to estimate the probability in part **c**.

f Find the percentage error in the estimation.

5 A commonly used rule of thumb states that the normal approximation to a binomial distribution will be reasonable if $np > 5$ and $n(1 - p) > 5$. This means that if p is further away from 0.5, the sample size needs to be bigger to get a reasonable approximation.

According to this rule, how big does the sample need to be if:

a $p = 0.5$, b $p = 0.25$, c $p = 0.125$, d $p = 0.01$,

e $p = 0.75$, f $p = 0.875$, g $p = 0.9$, h $p = 0.55$?

6 According to some estimates, eight per cent of males in the world are colour-blind. A representative random sample includes 854 people, and a statistician wishes to determine the probability that between 7% and 9% of the people in the sample are colour-blind.

a Give a reason why the researcher might want to use a normal approximation to the binomial to calculate this probability.

b Comment on the assumption that the sample is representative and random.

ISBN 978-1-108-76630-2

c Calculate the required probability, using a normal approximation without a continuity correction.

d As a measure of the accuracy of ignoring any continuity correction, calculate the probability $P(76 < X < 76.5)$ associated with the boundary of 76 successes.

DEVELOPMENT

7 A horticulturalist is attempting to cross two species of flowers to strengthen certain characteristics. One of the plants has red flowers and the other's flowers are white, but the horticulturalist wishes to retain the strong red colour in the offspring. According to Mendel's theory of inheritance, there is a 25% chance that the flowers of the offspring will be red. The horticulturalist crosses 15 pairs of parent plants and notes the colour of their offspring plant. Find the probability that of the offspring:

a none will have red flowers,

b there will be at least one with red flowers,

c at least twenty per cent will have red flowers (use a normal approximation here).

8 Suppose that it is known that 45% of eighteen-year-olds in a particular city do not have a driver's licence. If a random sample of 20 eighteen-year-olds is taken in the city, what is the probability that more than half of them will not have a driver's licence?

9 Long-term studies show that 60% of the residents and visitors to Nashville Tennessee prefer Country music to Western. The local council provides Country music for those eating their lunch in the park to listen to. How confident can they be that in a group of thirty, more than twenty of them prefer Country? Comment on the assumption of independence in this question.

10 [Technology]

This question continues from Question **7** of Exercise 17B, and you will need to retrieve your saved spreadsheet. Do not overwrite this spreadsheet — make a copy and work only with the copy.

The student is continuing to explore the theoretical probabilities that occur if a coin is tossed 100 times, but this time he wonders what happens if $p \neq 0.5$. He previously drew a histogram of the data in column `B`, with column `A` as the x-axis labels (this may well be the default) and noted the visual similarity with a bell-shaped curve, like a normal distribution.

a Change the probability to $p = 0.1$ in cell `F2` and let the spreadsheet recalculate.

b What binomial situation could your new spreadsheet be modelling?

c Comment on the shape of the curve for different values of p. Try $p = 0.1, 0.4, 0.5, 0.7, 0.95$. In particular, comment on the centre and spread of the distribution for the different values of p.

11 [Technology simulation]

In this question we simulate the tossing of twenty coins. This binomial experiment is repeated 100 times to generate data approximating the distribution $B(20, 0.5)$. The authors have used a recent version of Excel for this exercise, and your spreadsheet program may use different commands for some of the required functions.

a Open a new spreadsheet.

b In cell `B2` enter the code `=RANDBETWEEN(0,1)`. This will generate 1 if the toss is a success (head) and a 0 otherwise (failure).

ISBN 978-1-108-76630-2 Cambridge University Press

c Select the 20 cells `B2:U2`, and `Fill Right` using `Ctrl+R` or some other method. This represents the twenty tosses of the coin.

d In cell `W1` add the heading ‘Heads’, and in cell `W2` add the code `=SUM(B2:U2)`. This counts the number of heads in the 20 tosses.

e Select the 22 cells `B2:W2`, then extend this selection to a block from row `2` to row `101`, and `Fill Down` to cells `B101:W101` using `Ctrl+D` or some other method. This represents the 100 repetitions of the experiment, ‘toss the coin 20 times’.

f In cell `AB32` enter the formula `=AVERAGE(W2:W101)`, and in cell `AB33` enter the formula `=STDEV.P(W2:W101)`. Add suitable headings in the preceding cells. Confirm that the results for the mean and standard deviation agree with the values predicted by the formulae $\mu = np$ and $\sigma = \sqrt{npq}$.

g Press the `Calculate now` option (Excel Formulas tab, or try function key `F9` or `CTRL+ALT+SHIFT+F9`), and see the data updating as the spreadsheet tosses another 100 lots of twenty coins.

h Enter the following values and formula in cells `Z1:AB3`.

	···	Z	AA	AB
1	···	Num.heads	`f_r`	`cf_r`
2	···	`0`	`=COUNTIF($W$2:$W$101,$Z2/100)`	`=AA2`
3	···	`=1+Z2`	`=COUNTIF($W$2:$W$101,$Z3/100)`	`=AA3+AB2`

Select cells `Z3:AB3`, then extend this selection to a block from row `3` to row `22`, and `Fill Down` to cells `Z22:AB22`. We now have the results of the simulation experiment. Graph the results in cells `AA2-AA22`, using the entries in cells `Z2-Z22` as horizontal axis (category) labels in a bar graph. Be careful to adjust the gap between the bars to zero so that it is a histogram.

i Add a second graph on separate axes using the cumulative frequency entries in column `AB`.

j We want to add the theoretical normal distribution to these results as a comparison. Add the following entries:

	···	AC	AD
1	···	theoretical f_r	theoretical cf_r
2	···	`=BINOM.DIST(Z2,20,0.5,FALSE)`	`AC2`
3	···	`=BINOM.DIST(Z3,20,0.5,FALSE)`	`=AC3+AD2`

k Again, select cells `AC3:AD3`, extend the selection to a block from row `3` to row `22`, and `Fill Down` to cells `AC22:AD22`. This gives us the theoretical results. Add the series in column `AC` to the first graph and the series in column `AD` to the second graph (the cumulative graph) — this is done by the `Select Data` option when you right-click on the graph area. Click each graph and format them as line graphs (polygons).

l Recalculate a few times and notice how good the agreement of the experiment and theoretical graphs is in each case. This is particularly true for the cumulative graph, which has a smoothing effect on the data.

ENRICHMENT

12 Berry punnets distributed by a certain grower are exported in large batches. The receiver selects 10 punnets at random from each batch and tests whether any must be rejected because of rotten berries. The fraction p of punnets that are not of acceptable standard in the entire batch is unknown. The receiver needs a reasonable method of determining the quality of the batch without testing the entire batch, because the berries are destroyed in the process.

a One possible method is to reject the entire batch if any punnets in the sample are rejected.

i Suppose that $p = 0.05$, and determine the probability that the batch will be rejected.

ii Similarly determine and draw up a table showing the probability of rejection for $p = 0, 0.1, 0.2, 0.3, 0.4$. Include the case $p = 0.05$ in your table.

iii Plot a graph of p (on the horizontal axis) and the probability of rejecting the batch (on the vertical axis). This is called the *operating characteristic curve* (OC).

b Repeat part a if samples of 15 punnets are taken, but the batch is only rejected if two or more punnets are rejected. Include both plots on the same axes.

c Comment on whether part a or part b is the better method of ensuring the standard of the whole batch.

ISBN 978-1-108-76630-2

Cambridge University Press

17D Sample proportions

So far, we have recorded the result of a binomial experiment as the number of successes. Often, however, it is more interesting to record instead the *proportion of the Bernoulli trials that were successes* — this is called the *sample proportion*.

Sample proportions have many purposes, but one important purpose is to estimate the probability p of each Bernoulli trial in the binomial experiment. For example, before every election, surveys estimate the proportion of Australians voting for the Working Together Party. If a survey of 2000 finds 700 WTP voters (the number of successes), then the the sample proportion is 0.35 (the proportion of successes), so the survey has estimated the WTP vote to be 35% of the population.

Population proportions

The probability p that a voter chosen at random will vote for the Working Together Party is a *population proportion*,

$$p = \frac{\text{number of Australians voting WTP}}{\text{number of Australian voters}}.$$

Many probabilities arise in this way, as we have often seen. Sample proportions apply to all binomial situations, but may be intuitively easier to understand if we think of the probability p arising as a population proportion, as in this voting example.

Sample proportions

Let $X \sim B(n, p)$ be a binomial variable consisting of n independent Bernoulli trials, each with probability p of success. The *sample proportion* $\hat{p}$ is the proportion of successes when the experiment is run. Thus *the sample proportion is a new random variable denoted by* $\hat{p}$ *and defined by*

$$\hat{p} = \frac{X}{n}.$$

For example, I survey 20 voters, and 7 say that they will vote WTP. This is a binomial experiment with 20 independent Bernoulli stages, and random variable X with 21 possible values $0, 1, 2, 3, \dots, 20$. I can record my result as

$$X = 7 \qquad \text{or as} \qquad \hat{p} = \frac{7}{20}.$$

The binomial random variable X has 21 possible values $0, 1, 2, 3, \dots, 20$, and the new random variable $\hat{p}$ has 21 possible values $0 = \frac{0}{20}, \frac{1}{20}, \frac{2}{20}, \dots, 1 = \frac{20}{20}$.

The sample proportion distribution — mean and variance

A sample proportion random variable $\hat{p}$ is not a binomial random variable X because its values are fractions between 0 and 1 rather than whole numbers from 0 to n. But it is closely related to the corresponding binomial variable — the probability of each value of $\hat{p}$ is the same as the probability of the corresponding value of X,

$$P\left(\hat{p} = \frac{x}{n}\right) = P(X = x) = {}^nC_x\, p^x q^{n-x},$$

so the various graphs of a sample proportion distribution are just the graphs of the corresponding binomial distribution stretched horizontally by a factor of $\frac{1}{n}$.

ISBN 978-1-108-76630-2

Cambridge University Press

The mean and variance of $\hat{p}$ come directly from the mean and variance of X,

$$\begin{aligned} \mathrm{E}(\hat{p}) &= \mathrm{E}\left(\frac{X}{n}\right) \\ &= \frac{E(X)}{n} \\ &= \frac{np}{n} \\ &= p, \end{aligned} \qquad \begin{aligned} \mathrm{Var}(\hat{p}) &= \mathrm{Var}\left(\frac{X}{n}\right) \\ &= \frac{E(X)}{n^2} \\ &= \frac{npq}{n^2} \\ &= \frac{pq}{n}, \end{aligned} \qquad \begin{aligned} \mathrm{SD} &= \sqrt{\frac{npq}{n^2}} \\ &= \frac{\sqrt{npq}}{n}. \end{aligned}$$

The first formula $\mathrm{E}(\hat{p}) = p$ proves what we said in the introduction to this section — running a binomial experiment and recording the sample proportion $\hat{p}$ gives an estimate of the Bernoulli probability p.

7 THE SAMPLE PROPORTION $\hat{p}$

Suppose that a binomial experiment with random variable X consists of n independent Bernoulli trials, each with probability p.

- The *sample proportion* is the random variable $\hat{p} = \dfrac{X}{n}$.
- The values of $\hat{p}$ are the $n + 1$ values $0, 1, 2, \ldots, n$ of X divided by n. Thus the values of $\hat{p}$ are $0 = \frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \ldots, 1 = \frac{n}{n}$.
- The probability of each value of $\hat{p}$ equals the probability of the corresponding value of X,
$$P\left(\hat{p} = \frac{x}{n}\right) = P(X = x) = {}^{n}\mathrm{C}_x\, p^x q^{n-x}.$$
- The mean, variance and standard deviation of $\hat{p}$ are
$$\mathrm{E}(\hat{p}) = p, \qquad \mathrm{Var}(\hat{p}) = \frac{pq}{n}, \qquad \text{standard deviation} = \frac{\sqrt{npq}}{n}.$$
- The sample proportion $\hat{p}$ gives an estimate of the probability p.

Here is a simple example comparing a binomial distribution with the corresponding sample proportion distribution.

Example 15 17D

A coin is tossed six times. Let X be the binomial variable and $\hat{p}$ the corresponding sample proportion random variable.

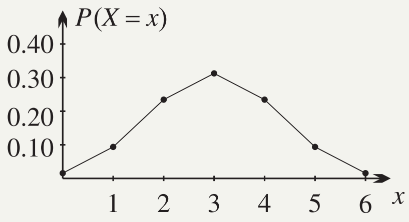
a Tabulate the probabilities of X, find the mean and variance, and draw the frequency polygon.

b Tabulate the probabilities of $\hat{p}$, find the mean and variance, and draw the frequency polygon.

SOLUTION

a Using the formula $P\ (x \text{ heads}) = {}^{6}\mathrm{C}_x \left(\frac{1}{2}\right)^x$,

x	0	1	2	3	4	5	6
prob	$\frac{1}{64}$	$\frac{6}{64}$	$\frac{15}{64}$	$\frac{20}{64}$	$\frac{15}{64}$	$\frac{6}{64}$	$\frac{1}{64}$



For the binomial variable X,

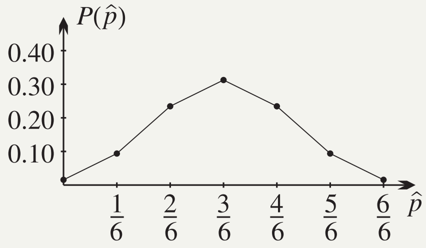
$$\mathrm{E}(X) = np = 3 \quad \text{and} \quad \mathrm{Var}(X) = npq = 1\tfrac{1}{2}.$$

ISBN 978-1-108-76630-2

Cambridge University Press

b The sample proportion is $\hat{p} = \dfrac{X}{n}$, and the corresponding probabilities are the same, so

$\hat{p}$	0	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1
prob	$\frac{1}{64}$	$\frac{6}{64}$	$\frac{15}{64}$	$\frac{20}{64}$	$\frac{15}{64}$	$\frac{6}{64}$	$\frac{1}{64}$



For the sample proportion p,

$$\text{E}(X) = p = \tfrac{1}{2} \text{ and } \text{Var}(X) = \frac{pq}{n} = \tfrac{2}{3}.$$

Why all this machinery of sample proportions?

Marketers and public opinion pollsters are very keen to find sample proportions for the Australian population. But surveys are expensive, and pollsters need to know how effective they can be with the small samples that they can afford or have time to carry out.

These things are also crucial for health research and public policy. Estimating a quantity such as the monthly unemployment statistics, or the number of Australians who require disability assistance services, is dependent on sample survey methods.

Working with the sample proportion not only provides an immediate estimate of p, but it also provides a straightforward approach to assessing how accurate the estimate is likely to be.

For example, the next worked example calculates the probability that a pollster will get within 5% and 10% of the correct result when using a sample of 20 voters to estimate the proportion voting for the WTP.

Example 16 17D

A binomial experiment with 20 Bernoulli trials is being used to estimate the probability p of success in the Bernoulli experiment. If the actual value of p is $\frac{1}{3}$, find the probability that the experiment will yield an estimate within:

a 0.05 of the actual value,

b 0.1 of the actual value.

SOLUTION

Notice first that the estimate cannot be exactly $p = \frac{1}{3} \doteqdot 0.333$ because the size of the sample is 20, which is not a multiple of 3.

a To obtain an estimate within 0.05 of the actual value $p \doteqdot 0.333$, the sample proportion would need to be $\hat{p} = \frac{6}{20} = 0.30$ or $\hat{p} = \frac{7}{20} = 0.35$.

$$P\left(\hat{p} = \tfrac{6}{20} \text{ or } \hat{p} = \tfrac{7}{20}\right) = {}^{20}C_6 \left(\tfrac{1}{3}\right)^6 \left(\tfrac{2}{3}\right)^{14} + {}^{20}C_7 \left(\tfrac{1}{3}\right)^7 \left(\tfrac{2}{3}\right)^{13}$$
$$= 0.36425 \ldots \left(\text{save this value in the memory}\right).$$

Hence the probability that the estimate is within 0.05 of p is about 0.364.

ISBN 978-1-108-76630-2

Cambridge University Press

b To obtain an estimate within 0.1 of the actual value $p \doteqdot 0.333$, the sample proportion would need to be $\hat{p} = \frac{5}{20} = 0.25$ or $\frac{6}{20}$ or $\frac{7}{20}$ or $\frac{8}{20} = 0.40$.

$$\begin{aligned} P\left(p = \tfrac{5}{20} \text{ or } p = \tfrac{8}{20}\right) &= {}^{20}C_5 \left(\tfrac{1}{3}\right)^5 \left(\tfrac{2}{3}\right)^{15} + {}^{20}C_8 \left(\tfrac{1}{3}\right)^8 \left(\tfrac{2}{3}\right)^{12} \\ &= 0.29368\ldots, \end{aligned}$$

Adding the answer to part **a**, the probability that the estimate is within 0.1 of the actual value is about 0.658.

8 USING THE SAMPLE PROPORTION

The sample proportion $\hat{p}$ can often be used to estimate the probability of obtaining an acceptable estimate to the probability p of a Bernoulli experiment.

Using the normal approximation to the sample proportion

We discussed in Section 17C how to use normal approximations to a binomial distribution to obtain reasonable approximations to the binomial distribution far more quickly. The probabilities of the sample proportions are just the probabilities of the corresponding values of the binomial variable, so the same approximation methods can be used here. We cannot prove in this course that the normal distribution approximates $\hat{p}$ — it is another example of the central limit theorem discussed at the start of Section 16G in the last chapter.

The normal approximation to the sample proportion thus has the same mean p as the sample proportion distribution, and the same variance $\dfrac{pq}{n}$, so we can approximate it using the normal distribution $N\left(p, \dfrac{pq}{n}\right)$.

As an example, let us increase the numbers in the previous worked example back up to realistic levels, and use the normal approximation to obtain the result. Political surveys before elections attempt to reduce the margin of error in their estimates to 1%–2%, that is, they attempt to estimate p correct to within 0.01–0.02.

Example 17 17D

A survey of 2000 Australian voters is being used to estimate the probability p that a random voter will vote for the Working Together Party. Assuming that the actual value of p is $\frac{1}{3}$, use the normal approximation to the sample proportion distribution to find the probability that the experiment will yield an estimate:

a within 0.01 of the actual value,

b within 0.02 of the actual value.

SOLUTION

The numbers in the sample are large enough for us to ignore continuity corrections.
The normal approximation to $\hat{p}$ has

$$\text{mean} = \tfrac{1}{3} \doteqdot 0.3333 \qquad \text{and} \qquad \text{variance} = \frac{pq}{n} = \frac{2}{9} \times \frac{1}{2000} = \frac{1}{9000},$$

so taking the square root, the standard deviation is about 0.01054.

ISBN 978-1-108-76630-2

Cambridge University Press

a We need to find $P(0.3233 \le \hat{p} \le 0.3433)$, so we find z-scores.

For 0.3233, $z = \dfrac{-0.01}{0.01054}$ $\doteqdot -0.9487.$ For 0.3433, $z = \dfrac{0.01}{0.01054}$ $\doteqdot 0.9487.$

Hence the probability that the estimate is within 0.01 of the actual value is about

$$P(-0.9487 \le Z \le 0.9487) \doteqdot 0.66$$

b We need to find $P(0.3133 \le \hat{p} \le 0.3533)$. The z-scores are now;

For 0.3233, $z = \dfrac{-0.02}{0.01054}$ $\doteqdot -1.8974.$ For 0.3433, $z = \dfrac{0.02}{1.01054}$ $\doteqdot 1.8974.$

Hence the probability that the estimate is within 0.02 of the actual value is about

$$P(-1.8974 \le Z \le 1.8974) \doteqdot 0.94$$

Example 18 17D

Election surveys usually claim that their margin of error is about 2%, and the last worked example has addressed this. What are some other issues that pollsters need to take into account?

SOLUTION

- Can I be sure that the sample is unbiased?
- Are people answering honestly?
- Are opinions changing as the election approaches?
- What is to be done with people who refuse to answer or ‘don’t know’?
- Were there language or cultural problems that interfered with the interview?

The data, the sample proportion, and the normal approximation

The situation has now become quite complicated because we have:

1 values of the sample proportion $\hat{p}$ obtained from surveys,

2 the sample proportion distribution, if we assume a value for $\hat{p}$,

3 the normal approximation to the sample proportion distribution, also if we assume a value for p.

We conclude this chapter by drawing one more picture, using the dataset that we gained by simulation of a binomial experiment in worked Example 9 of Section 16B. We are interested in the distribution of the sample proportion $\hat{p}$, that is, the distribution of the estimate of the probability, and we will use the cumulative distributions in our comparisons.

Notice that the density function of the normal approximation to $\hat{p}$ is obtained from the density function of the approximation to X by stretching horizontally by a factor of $\dfrac{1}{n}$, and then stretching vertically by a factor of n so that the area under the curve remains 1. Therefore we can’t superpose the graph of that normal density function on the graph of the frequency polygon because the heights don’t match. Instead, we will compare the cumulative graphs to see the agreement.

ISBN 978-1-108-76630-2 Cambridge University Press

Example 19 17D

In worked Example 9 of Section 17B, we ran 100 simulations of the binomial experiment of tossing 10 coins and recording the number of heads. Now imagine that the purpose of each simulation was to produce an estimate of the probability of tossing heads, that is, a value of the random variable $\hat{p}$. Note that this random variable $\hat{p}$ has the 11 values $0 = \frac{0}{10}, \frac{1}{10}, \frac{2}{10}, \dots, 1 = \frac{10}{10}$.

a From the data, produce a table showing the cumulative relative frequency of obtaining each value of $\hat{p}$, and find the mean and standard deviation.

b Use binomial probability to calculate the cumulative probabilities of obtaining the values of $\hat{p}$, and find the mean and standard deviation.

c Use the normal approximation to the sample proportion distribution to calculate the cumulative probabilities of values of $\hat{p}$, and find the mean and standard deviation. Apply the continuous correction, because $n = 10$ is small.

d On one graph, draw the cumulative relative frequency polygon of the data, the cumulative probability polygon of the distribution, and the CDF of the normal approximation.

SOLUTION

The calculations in parts **a** and **b** were done in worked Example 9, and we only need to divide the values through by 10.

a After running the experiment 100 times, each relative frequency is the estimated probability of obtaining each estimate of $\hat{p}$.

$\hat{p}$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
f_r	0.00	0.02	0.06	0.15	0.2	0.23	0.17	0.13	0.01	0.03	0.00
cf_r	0.00	0.02	0.08	0.23	0.43	0.66	0.83	0.96	0.97	1.00	1.00

Calculation gives $\bar{x} = 0.482$ and $s \doteqdot 0.170$.

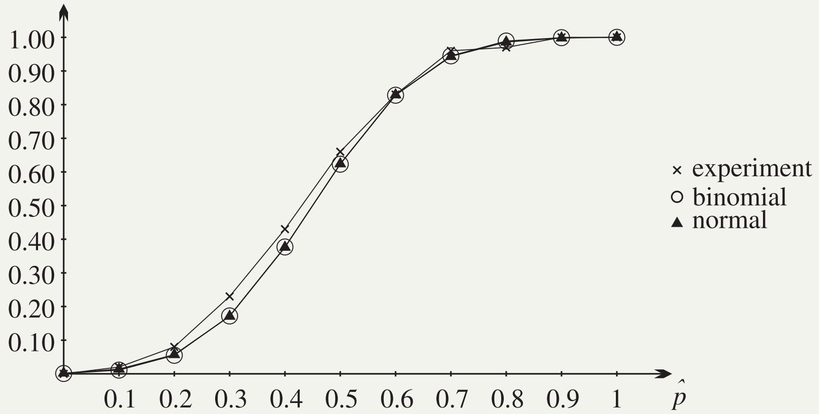
b From previous calculations, $\mu = 0.5$ and $\sigma = \frac{1}{10}\sqrt{2.5} \doteqdot 0.158$.

$\hat{p}$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$P(\hat{p})$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001
Cmlve	0.001	0.011	0.055	0.172	0.377	0.623	0.828	0.945	0.989	0.999	1.000

c The mean and standard deviation of the normal approximation are the same as for the sample proportion distribution. We need z-scores first, and we apply the continuity correction with 0.1 as the gaps in the values for $\hat{p}$. For example, for $\hat{p} = 0.7$ we find the z-score corresponding to $\hat{p} = 0.75$.

$\hat{p}$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
z	−2.85	−2.21	−1.581	−0.949	−0.316	0.316	0.949	1.581	2.21	2.85	3.48
$\phi(z)$	0.002	0.014	0.057	0.171	0.376	0.624	0.829	0.943	0.986	0.998	1.000

d



At the resolution of this graph, the theoretical probability distribution curve and the normal approximation to it are not separated. Look at the tables.

Exercise 17D

FOUNDATION

Note: In the previous exercise we used continuity corrections when we approximated a discrete binomial distribution by a continuous normal distribution. For large values of n this correction is less necessary. In this section, however, using the continuity correction when using sample proportions leads to much messier calculations, and the numbers are mostly large, so it will seldom be applied.

1 a A student tosses five coins and the results are: H T H T T. What is the sample proportion $\hat{p}$ of heads in this experiment?

b A student selects ten cards from a standard pack with replacement, and records the suit: ♡♣♠◇◇♠♡♠♠◇. What is the sample proportion $\hat{p}$ of spades in this sample?

c A manufacturer takes a sample of 12 items from their recent batch of gizmos, testing each item to see if it passes quality control (P) or not (F). The results were: P P P F F P P P P F P P. What is the proportion $\hat{p}$ of items that pass?

2 A single fair coin is tossed five times, and the number x of heads is recorded.

a Copy and complete the upper table to the right, using the binomial probability formula.

x	0	1	2	3	4	5
$P(X = x)$						

b Copy and complete the lower table to the right to convert part a to a table of probabilities of the sample proportions.

$\hat{p}$	
$P(\hat{p})$	

You will need to divide each score x by 5 to obtain the corresponding sample proportion.

c Calculate the mean of the second table.

d How would you interpret this mean?

3 Every Saturday for 20 weeks, a marketer surveyed five people chosen at random in a suburban shopping centre. The first question is, ‘Do you live in the suburb?’, and the weekly frequency of the number x of ‘Yes’ answers is given in the upper table.

x	0	1	2	3	4	5
f	1	1	3	2	6	7

ISBN 978-1-108-76630-2

Cambridge University Press

$\hat{p}$	
f_r	

a Copy the table to the right, and complete it to show a table in which the two rows are the sample proportions $\hat{p}$ and the relative frequencies f_r.

b Calculate the mean of this second table.

c What is this mean an estimate of?

4 A coin is tossed 10 times, and a student is asked to calculate the probability that more than 75% of the coins will show heads.

a If more than 75% of the coins show heads, how many coins would this be?

b Use the binomial probability formula to determine the probability of obtaining more than 75% heads. (On this occasion the calculation is easily done without a normal approximation.)

5 A die is thrown 50 times. What is the probability that less than 9% of the time the result will be a head? Do this:

a using an exact binomial calculation, finding the probability of 0, 1, 2, 3 or 4 heads;

b using a normal approximation to the sample proportion, and finding the probability $P(\hat{p} \le 0.09)$.

6 Information has been recorded about whether the 32 members of a class buy their lunch regularly at the school canteen:

1.	James	N	2.	Kate	N	3.	Xavier	N	4.	Jimmy	N
5.	Clyde	Y	6.	Bob	N	7.	Liam	N	8.	Agata	N
9.	Irene	N	10.	Aqila	Y	11.	Sonny	Y	12.	Andrea	N
13.	Magarida	N	14.	Terry	N	15.	Iman	N	16.	Lucie	Y
17.	Ping	N	18.	Maddy	Y	19.	Kamal	Y	20.	Xue	Y
21.	Odette	N	22.	Billy	N	23.	Chang	N	24.	Nahla	N
25.	Craig	Y	26.	Jerry	N	27.	Zahra	Y	28.	Jun	N
29.	Nara	Y	30.	Dakarai	Y	31.	Lerato	N	32.	Sahar	Y

In order to generate random samples, each student has been given a unique identifying number.

a Calculate the fraction of students who buy their lunch regularly at the canteen. This is called the *population proportion*.

b Kamal generates the five random numbers 12 15 3 30 17, thus generating the sample of 5 students Andrea, Iman, Xavier, Dakarai, Ping. What proportion of these five students buy their lunch regularly at the canteen? This is called a *sample proportion*.

c Copy the table below. Then use the following sets of five random numbers to generate 10 sample proportions for $n = 5$. Enter your results in the table — the first sample was dealt with in part **b** and is already included in the tally. (Notice that repetition is allowed — this is sampling with replacement.)

12 15 3 30 17	3 25 17 17 20	27 7 24 26 2	20 9 21 10 16
26 6 11 5 25	29 24 23 27 3	22 11 25 9 8	27 14 22 11 20
9 28 1 17 1	10 32 24 30 13		

$\hat{p}$	0	0.2	0.4	0.6	0.8	1.0
Tally		|				
Frequency						

d Draw a dot plot for the distribution obtained in part **d**.

DEVELOPMENT

7 In a local election, 20% of the people voted independent. What is the probability that if 500 people are chosen for a random survey, more than 22% of them voted independent? Use a normal approximation to the sample proportion.

8 A card is selected from a standard pack and it is then returned. This experiment is repeated 80 times. What is the probability that a hearts card turns up between 20% and 30% of the time, inclusive? Do this:

a By interpreting it as between 16 to 24 hearts inclusive and using a normal approximation with a continuity correction;

b Using sample proportion without any correction.

9 A farmer knows that a certain type of seed is 70% likely to germinate when planted. He plants 300 seeds at the start of the season. Use the normal approximation for the sample proportion to find the probability that:

a at least 65% will germinate,

b between 65% and 75% will germinate.

10 A medication causes a painful reaction in 5% of users.

a In a group of 100 people, find the probability that:

i no one reacts,

ii less than two per cent of the people react.

b In a larger study into patients' reactions to this medication, 1000 patients are given the medication (to which 5% are known to have a painful reaction). Researchers find that less than 3% of the patients in this study have a reaction.

i Use the normal approximation to the sample proportion to determine the probability of this happening by chance.

ii What should the researchers conclude?

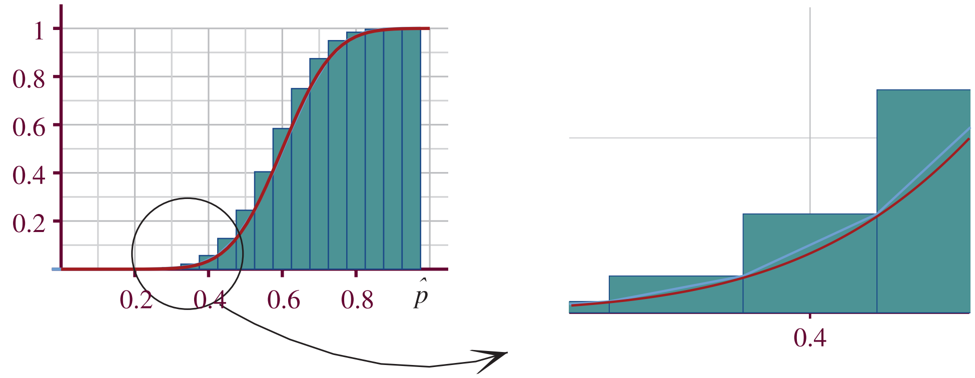
11 A student is asked to redo Question 3 from Exercise 17C, but using sample proportions. A barrel contains 600 pink and 400 blue counters. At each stage of an experiment, the barrel is stirred well, then a counter is removed, its colour is noted, and it is returned to the barrel. This experiment is repeated 20 times.

The random variable for this sample proportion is $\hat{p} = \dfrac{X}{n}$, where X is the corresponding binomial random variable with $n = 20$ trials.

The student redraws his cumulative histogram to explore the idea of *sample proportion*. On the horizontal axis, he divides his results by 20 to record the fraction of pink counters occurring in the draw.

a What is the mean and standard deviation of the sample proportion random variable $\hat{p}$? Recall the formulae $\mu = p$ and $\sigma^2 = \dfrac{pq}{n}$, where $q = 1 - p$.

He overlays the cumulative normal curve with the same mean and standard deviation on top, and finds a very good agreement.



b Explain why the distribution of $\hat{p}$ with values between 0 and 1 is not a continuous distribution.

c To confirm his previous result, he decides to calculate the probability that $\hat{p} \le 0.4$, meaning that no more than 40% of the counters are pink. Assuming that the distribution is approximately that of a normal random variable $N(\mu, \sigma^2)$, calculate the required probability $P(\hat{p} \le 0.4)$. Do *not* attempt any continuity correction.

d Calculate the percentage difference in the exact answer 5.7% obtained by a direct binomial calculation. Is there good agreement?

e Calculate the probability that no more than 75% of the counters are pink, that is, $P(\hat{p} \le 0.75)$.

f Compare the percentage error in this second answer with the exact result 95% and explain why the approximation is better than parts **c**–**d**.

12 In this question we investigate the accuracy of a normal approximation to the sample proportion for various sample sizes n.

A coin is tossed repeatedly and the proportion of heads is recorded. The exact theoretical probability of obtaining at most 52% heads is calculated using the binomial distribution, and recorded in the second row of the table below for differing sample sizes.

n	1000	500	100	50	25
exact	0.9026	0.8262	0.6914	0.6641	0.6550
approx					
% error					

a Use a normal approximation for the sample proportion to fill in the second row of the table.

b In each case calculate the percentage error in the approximation for each of these samples. Record your results in the third row of the table.

c Comment on the accuracy of your approximations for various sample sizes n.

13 A manufacturer distributes tins of pineapple under a recognised brand name, and also as a generic supermarket no-name product. In a trial, 50 customers are given a tin of each and later asked to express a preference. The customers in the trial must choose one product as their favourite. Assuming that there is no difference between the two products, the probability of choosing the branded pineapple should be 0.5. It is found that more than 60% of customers prefer the branded pineapple. Calculate the probability of this using the normal approximation for the sample proportion, then comment.

ISBN 978-1-108-76630-2

Cambridge University Press

14 Long-term trials have showed that 30% of patients with a certain disease respond to treatment by a company's drug. In further trials, 100 patients chosen at random from those with the disease are given a higher than usual dosage of the drug, and 40% respond positively. What is the probability that 40% or more could respond positively purely by chance?

15 The variance of the sample proportion $\hat{p}$ for a binomial distribution is $\sigma^2 = \dfrac{pq}{n}$. The variance measures the spread of the distribution of $\hat{p}$ around the population proportion p, thus it is customary to take n sufficiently large to ensure that σ is small.

a Assume that 70% of residents on a college campus are living at home, and that researchers want to choose a sufficiently large sample to mirror this statistic. How big will a sample need to be to ensure that the standard deviation of $\hat{p}$ is less than:

i 4% **ii** 3% **iii** 2% **iv** 1% **v** k%

b Repeat this question if a new survey finds that the number of residents living at home is 80%.

c Repeat for $p = 50\%$.

16 [Simulation]

a Ten coins are tossed.

i What is the probability of getting exactly six heads?

ii If this experiment is carried out 40 times, use your answer to part **i** to predict how many times you expect to get exactly 6 heads.

b Toss 10 coins and record the number of heads.

i Repeat this experiment 40 times, and copy and complete the table below, filling in the tally and the frequency.

Number of heads	0	1	2	3	4	5	6	7	8	9	10
Proportion of heads $\hat{p}$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.1
Expected frequency	0.0	0.4	1.8	4.7	8.2	9.8	8.2	4.7	1.8	0.4	0.0
Tally											
Frequency											

ii Does your answer to **a ii** agree with the expected frequency of six heads?

iii Draw a dot plot for $\hat{p}$ and note the shape of the curve. For a sufficiently large number of throws, we would expect to get a bell-shaped curve.

ENRICHMENT

17 Two dice are thrown and success is recorded if the sum is at least 10.

a Find the probability of success.

b Use the exact binomial distribution to find the probability of at most four successes on 20 throws.

c Use a normal approximation to the binomial to estimate the probability of at most four successes without continuity correction.

d Repeat part **c** with continuity correction.

e Use sample proportion to estimate the probability of at most 20% successes. Do not use continuity correction. Explain why your result agrees with part **c**.

f Use sample proportion, but this time use an estimate with continuity correction, $P\left(0 \le X \le 0.2 + \frac{1}{40}\right)$. Explain why your result agrees with part **d**.

18 [Confidence intervals]

a Let $X \sim N(\mu, \sigma^2)$.

i Show that 95% of the data lie within 1.96σ of the mean μ, that is,

$$P(\mu - 1.96\sigma < X < \mu + 1.96\sigma) \doteqdot 95\%.$$

ii We interpret this to mean that a randomly chosen value X will lie within the interval $[\mu - 1.96\sigma, \mu + 1.96\sigma]$ ninety-five per cent of the time. Draw a diagram to show that for a given value x of the random variable, μ will lie within the interval $[x - 1.96\sigma, x + 1.96\sigma]$ ninety-five per cent of the time.

b In a sample of 100 school pupils, 67 surf the internet more than 7 hours a week. Thus we might estimate that 67% of all school pupils do the same, but how confident can we be of this estimate?

i Estimate the population standard deviation using this single sample, from the formula $\sigma^2 = p(1 - p)/n$, where we use the estimate $p \doteqdot \hat{p} = 67\%$.

ii Find the margin of error 1.96σ and find the 95% confidence interval, which is $[\hat{p} - 1.96\sigma, \hat{p} + 1.96\sigma]$, indicating that 95% of the time we take such a sample, the population proportion will fall within this interval.

iii Using the estimate $p \doteqdot 67\%$, how large does the sample size n need to be to reduce the margin of error to 1%?

ISBN 978-1-108-76630-2

Cambridge University Press

Chapter 17 Review

Review activity

- Create your own summary of this chapter on paper or in a digital document.

Chapter 17 Multiple-choice quiz

- This automatically-marked quiz is accessed in the Interactive Textbook. A printable PDF worksheet version is also available there.

Chapter review exercise

1 A marksman finds that on average he hits the target five times out of six. Assuming that successive shots are independent events, find the probability that in four shots:

a he has exactly three hits,

b he has exactly two misses.

2 Five out of six people surveyed think that Tasmania is the most beautiful state in Australia. What is the probability that in a group of 15 randomly selected people, at least 13 of them think that Tasmania is the most beautiful state in Australia?

3 There are ten questions in a multiple-choice test, and each question has five possible answers, only one of which is correct. What is the probability of answering exactly seven questions correctly by chance alone? Give your answer correct to three significant figures.

4 A card is selected from a pack, its suit is noted, and it is returned. How many times must this be done so that the probability of a heart is more than 95%?

5 An eight-sided die is inscribed with the digits 1–8.

a What is the probability of obtaining an 8 when the die is thrown?

b Six eight-sided dice are thrown. Construct a table for the distribution of the random variable X that counts the number of eights that occur. Record your results correct to 4 decimal places.

c A player needs to get exactly three eights in order to win. How often would you predict this to occur in 1000 throws of the six dice?

d Repeat part **c** if he needs a throw of three or more eights.

6 Are the following experiments Bernoulli trials? If so, state the probability of success p and failure q.

a A coin is tossed, and it is noted if the result is heads or tails.

b Two dice are thrown, and the player wins if the sum if more than 10.

c Tests show that 4 out of every 1000 items pass quality control. Consider the random variable ‘number of passes’ where an item is selected at random from the manufacturing process.

d A card is drawn from a pack, and its suit is noted.

7 The binomial distribution $B(n, p)$ consists of n independent Bernoulli trials. Find the mean, variance and standard deviation for each distribution.

a $B(20, 0.2)$ b $B(70, 0.5)$ c $B(6, 0.8)$

d $B(120, 0.4)$ e $B(300, 0.1)$ f $B(5, 0.25)$

8 A company manufactures mobile phone cases using a mixture of machinery and traditional techniques. Data shows that the probability that a random case will fail quality control is 5%. An inspector selects a random batch of 60 cases from the warehouse. Let X be the binomial random variable of the number of cases that do not pass inspection.

a What is the mean, variance and standard deviation for this distribution?

b Find the probability that the number of cases that fail to pass lies within one standard deviation of the mean.

c New company standards insist that the number of failures in the batch must be no more that one standard deviation above the mean. Batches that fail to meet this standard are rejected. What is the probability of this?

d Due to the new regulations and the number of rejected batches, the company improves its manufacturing process so that the new experimental probability of failure is reduced to 2%. Repeat part **c** to find the new probability that a batch will be rejected.

9 A coin is tossed 80 times.

a Use the exact binomial formula to find the probability of 38, 39 or 40 heads.

b What is the mean and standard deviation for this binomial distribution?

c Calculate np and nq, and state whether this is a situation where a normal approximation may be used.

d In your own words, explain why we calculate $P(37.5 \le X \le 40.5)$ rather than $P(38 \le X \le 40)$.

e Find the probability of 38, 39 or 40 heads using a normal approximation.

f What is the percentage error in this normal approximation?

g Clearly in the example above, there was no need to use an approximation, because the probability could be calculated directly. Calculate now the probability of at least 50 heads using a normal approximation (but do not estimate the percentage error).

10 The sum of two dice is recorded.

a Find the probability that the sum is at least 10.

b Use a normal approximation to find the probability that the sum is at least 10 in more than 14 out of 80 throws.

11 A sample of 100 voters are asked whether they intend to vote for the Working Together Party at the next election. Thirty-five respond in the affirmative.

a What is the corresponding sample proportion?

b If a further sample is taken, would you expect the same number of respondents to indicate yes?

c What type of distribution will result if further such samples are taken and the sample proportions are recorded?

12 A bag contains 3 red balls and 2 white balls. Five balls are selected in turn, with replacement, and the number of red balls is recorded.

a Construct a table showing the theoretical sample proportions of red balls that are selected.

b Find the probability that the proportion of red balls is less than:

i 40%, ii 50%.

c Find the mean, variance and standard deviation for the random variable $\hat{p}$ tabulated in part **a**.

13 A fair die is to be thrown 500 times.

a Find the mean and standard deviation for the sample proportion of sixes in the theoretical distribution for this experiment.

b In one sample of 500 throws, the number of sixes was 70. How many standard deviations is this result below the mean?

14 Long-term records show that the percentage of male babies born in a large hospital is 53%. A study is carried out on the effect of a high potassium diet (white beans, salmon, avocados, almonds, apples and mushrooms) on increasing the probability that the baby will be male. In the group of 653 births under this study, with the mother following this diet, more than 54% were male. What is the probability of this happening by chance? Use a normal approximation for the sample proportion with no continuity correction.

ISBN 978-1-108-76630-2

Cambridge University Press