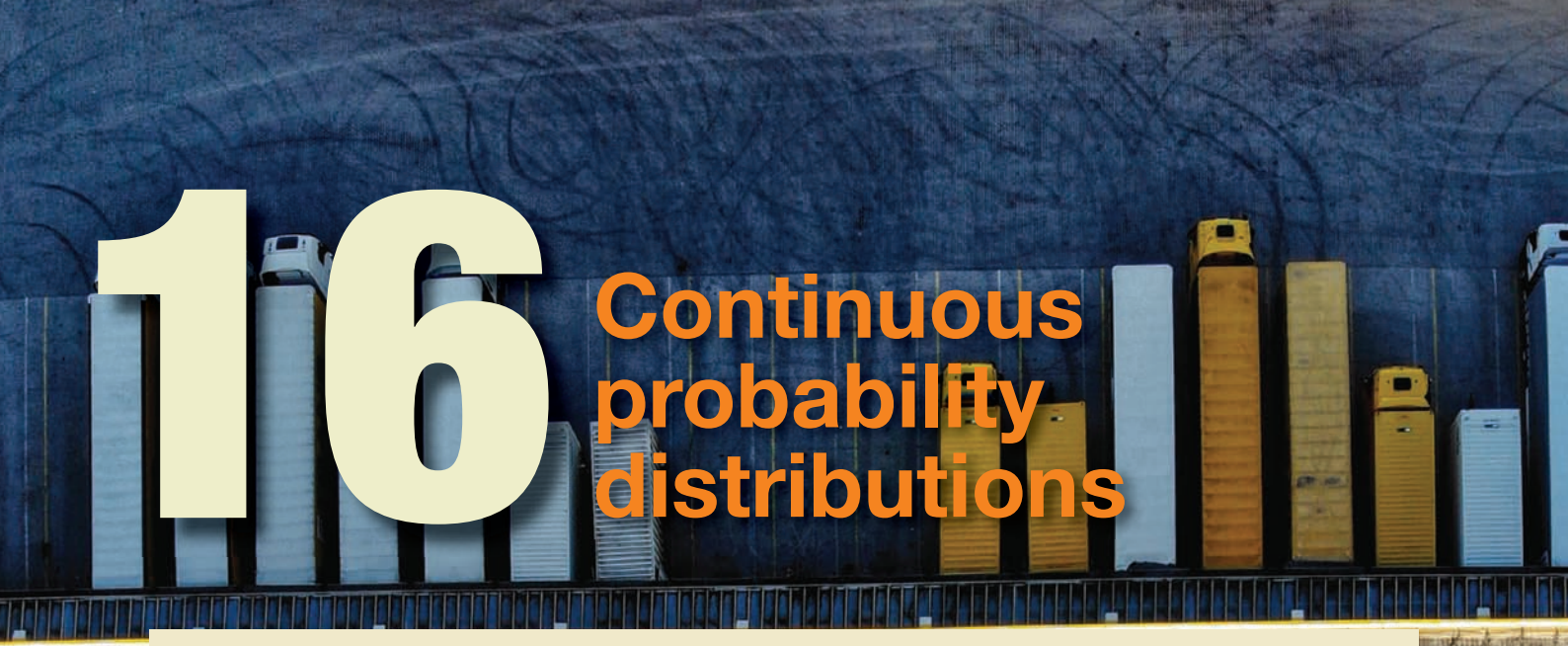


16 Continuous probability distributions

Last year we moved from calculating individual probabilities in Chapter 12 to calculating in Chapter 13 a whole set of closely related probabilities grouped together as a *discrete probability distribution*. *Random variables* became part of our machinery, and we calculated expected values (or means) and standard deviations of these random variables in theoretical probability distributions.

The final section of that chapter, Section 13D, only just began to combine the theoretical probabilities in a discrete probability distribution with the data collected from experiments. This present chapter now builds on the techniques of organising and displaying data in Chapter 15 to bring probabilities and data into a closer relationship. The key to this is *relative frequency*, because relative frequencies obtained from data are estimates of theoretical probabilities.

These methods, combined with grouping, allow us finally to give a coherent account of a *continuous random variable*, defining it in terms of a *probability density function*. Integration is needed to understand this material, because a probability is now interpreted as an area under a curve — a dramatic and unexpected idea.

The *normal distribution* is the most important of all continuous distributions, and indeed of all distributions. Sections 16D–16G develop the basic theory of normal distributions, and explain how to apply them to practical problems.

There are many calculations in this chapter, as in Chapter 15. Your pen-and-paper work can be assisted in several ways, all of which can be found in the interactive textbook or online:

- a scientific calculator and a table of values of the standard normal,
- the Desmos graphing calculator or a statistics calculator,
- a spreadsheet,
- specialised statistics software.

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.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

16A Relative frequency

Relative frequencies were introduced in Chapter 13 of the Year 11 book because they are estimates of the probabilities of the outcomes of an experiment. They are needed again in this chapter, where probability theory is central once more.

A brief review of the mean and variance of a discrete distribution

In Year 11, Sections 13B and 13C introduced the expected value and the variance of a discrete random variable X. Let $p(x) = P(X = x)$. Then the expected value $\mu = \mathrm{E}(X)$ is

$$\mu = \mathrm{E}(X) = \sum x\,p(x), \quad \text{summing over all values of the distribution.}$$

This is the *weighted mean of the values, weighted according to their probabilities*.

The variance $\mathrm{Var}(X) = \sigma^2$ is the square of the standard deviation σ. It is the expected value of the squared deviation from the mean,

$$\mathrm{Var}(X) = \mathrm{E}\left((X - \mu)^2\right) = \sum (x - \mu)^2 p(x).$$

We proved also that the variance has an alternative form that is preferable when the mean is not an integer, and is therefore particularly suited to data,

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2 = \sum x^2 p(x) - \mu^2.$$

Setting out the calculations

Our model experiment in the Year 11 book was, 'throw four coins and record the number of heads'. The second rows in the tables below show the theoretical probabilities of obtaining 0, 1, 2, 3 or 4 heads.

A Here is the way we set out the calculations of mean and variance when we use the definition of variance as $\mathrm{E}(X - \mu)^2$.

x	0	1	2	3	4	Sum	
$p(x)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$	1	(a check)
$x\,p(x)$	0	$\frac{4}{16}$	$\frac{12}{16}$	$\frac{12}{16}$	$\frac{4}{16}$	2	(the mean μ)
$(x - \mu)$	-2	-1	0	1	2	—	
$(x - \mu)^2$	4	1	0	1	4	—	
$(x - \mu)^2 p(x)$	$\frac{4}{16}$	$\frac{4}{16}$	0	$\frac{4}{16}$	$\frac{4}{16}$	1	(the variance)

The mean is $\mu = 2$, the variance is $\sigma^2 = 1$, and the standard deviation is $\sigma = 1$.

B Here is our setting-out using the other variance formula $E(X^2) - \mu^2$. The calculations are more straightforward, whether or not μ is a whole number.

x	0	1	2	3	4	Sum	
$p(x)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$	1	(a check)
$x\,p(x)$	0	$\frac{4}{16}$	$\frac{12}{16}$	$\frac{12}{16}$	$\frac{4}{16}$	2	(the mean μ)
$x^2 p(x)$	0	$\frac{4}{16}$	$\frac{24}{16}$	$\frac{36}{16}$	$\frac{16}{16}$	5	(this is $\mathrm{E}(X^2)$)

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

From the third row, $\mathrm{E}(X) = 2$, which is the mean μ.
From the last row, $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2$
$= 5 - 2^2$
$= 1$, which is σ^2.

Cumulative distribution function

In Year 11 we did not discuss the cumulative distribution function, but it is easily defined. With the four coins, it is the function $F(x)$ obtained by adding all the probabilities up to a certain point:

x	0	1	2	3	4
$p(x)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$
$F(x)$	$\frac{1}{16}$	$\frac{5}{16}$	$\frac{11}{16}$	$\frac{15}{16}$	1

For example, $F(3) = p(0) + p(1) + p(2) + p(3)$,
and in general, $F(x) = p(0) + p(1) + \cdots + p(x)$, for $x = 0, 1, 2, 3, 4$.

Relative frequencies and the histogram and polygon

Relative frequency is the key connection between the probabilities in the theoretical distributions studied in Year 11 and the analysis of the datasets in Chapter 15 this year. In Year 11 we performed the experiment of tossing four coins 100 times, and obtained these results.

x	0	1	2	3	4	Sum
f	7	29	34	21	9	100
f_r	0.07	0.29	0.34	0.21	0.09	1

- The first line lists the *values* — the number x of heads can be 0, 1, 2, 3 or 4.
- The second line lists the *frequencies* f — the experiment was run 100 times.
- The third line lists the *relative frequencies* $f_r = \dfrac{f}{100}$ — divide by 100 trials.

These relative frequencies are *estimates of the probabilities* of tossing 0, 1, 2, 3 or 4 heads — they are often referred to as 'experimental probabilities'. Unless the experiments are biased in some way, these estimates will almost certainly be closer and closer to the theoretical probabilities as the number of trials increases.

Setting out the calculations using relative frequencies

Calculating the sample mean and variance of this dataset was explained in Section 13D of the Year 11 book.

x	0	1	2	3	4	Sum
f_r	0.07	0.29	0.34	0.21	0.09	1
xf_r	0	0.29	0.68	0.63	0.36	1.96
$x^2 f_r$	0	0.29	1.36	1.89	1.44	4.98

The sample mean $\bar{x}$, variance s^2 and standard deviation s can then be calculated from the table as follows. These are *estimates* of the expected value $\mu = E(X)$, variance $\sigma^2 = \mathrm{Var}(X)$ and standard deviation σ of the probability distribution.

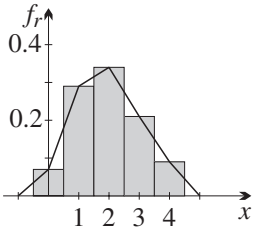
$\bar{x} = \sum xf_r$
$= 1.96$

$s^2 = \sum x^2 f_r - \bar{x}^2$
$= 4.98 - (1.96)^2$
$\fallingdotseq 1.14$ (compare with $\sigma^2 = 1$)

$s = \sqrt{s^2}$
$\fallingdotseq 1.07$ heads (compare with $\sigma = 1$)

ISBN 978-1-108-76630-2

Histograms and polygons using relative frequencies



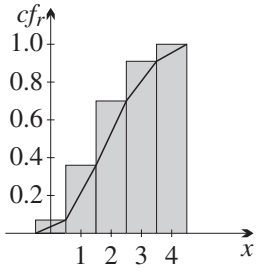
We can graph the relative frequencies in a *relative frequency histogram* and a *relative frequency polygon*. Look at the total area of the histogram rectangles, and the area under the polygon — the two areas are equal because each interval of the polygon cuts a triangle off one rectangle, and adds a triangle of the same area to an adjacent rectangle.

This particular histogram has a significant property — *the sum of the areas of the rectangles is 1*. This will always happen when the rectangles have width 1, because the sum of probabilities is 1. The Enrichment Question 12 in Exercise 16A deals with the situation when the rectangles have width different from 1.

These remarks about areas are made in preparation for using integration in Section 16B.

Cumulative relative frequencies

In the fourth line of the table below, we have calculated the *cumulative relative frequencies*.



x	0	1	2	3	4	Sum
f	7	29	34	21	9	100
f_r	0.07	0.29	0.34	0.21	0.09	1
cf_r	0.07	0.36	0.70	0.91	1	

The cumulative relative frequencies have been graphed in a *cumulative relative frequency histogram*. A *cumulative relative frequency polygon* or *ogive* can also be drawn, starting with accumulation zero at $x = -\frac{1}{2}$, and finishing with accumulation 1 at $x = 4\frac{1}{2}$.

Each cumulative frequency estimates the probabilities of obtaining a particular number of heads or fewer. For example,

estimated probability of tossing 3 or fewer heads $= 0.91$.

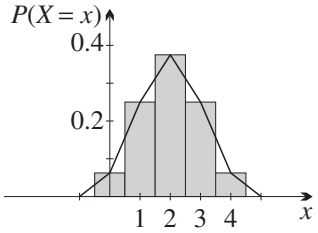
Discrete probability distributions and estimates from data

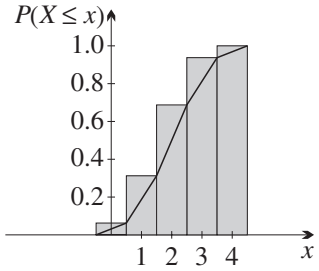
Relative frequencies obtained from data are estimates of probabilities. If we also know the theoretical probabilities, as we do for the four tossed coins, we can draw the same histograms and polygons as we have just done for data, but using probabilities and cumulative probabilities instead.

Here again is the theoretical probability distribution for the four tossed coins, and the two histograms and polygons — using decimals for easy comparison with the 100-trial data above. Compare these two diagrams with the diagrams above.

x	0	1	2	3	4
$P(X = x)$	0.0625	0.25	0.375	0.25	0.0625
$P(X \leq x)$	0.0625	0.3125	0.6875	0.9375	1

ISBN 978-1-108-76630-2





Probability estimates obtained from data are rarely exactly the same as the theoretical probabilities.

1 RELATIVE FREQUENCIES AND CUMULATIVE RELATIVE FREQUENCIES

- For a dataset, the relative frequencies and cumulative relative frequencies are obtained by dividing through by the total frequency.
- The relative frequencies of a dataset are estimates of probabilities. For this reason, they are often referred to as 'experimental probabilities'.
- The cumulative distribution function $F(x)$ of any numeric probability distribution is the probability that the score is less than or equal to x,

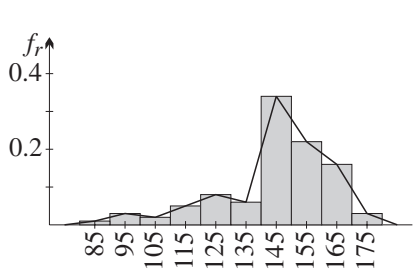
$$F(x) = P(X \leq x), \text{ for all } x \text{ in the domain.}$$

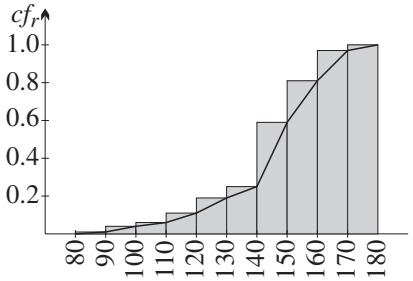
- The cumulative relative frequencies are estimates of the cumulative distribution function.
- The histograms and polygons of these relative frequencies and cumulative relative frequencies are drawn in the usual way.

Grouping data from a continuous random variable

When the underlying random variable is continuous, grouping is usually required. Here are the histograms and polygons for the heights x of 100 people from Section 15B, drawn this time using relative frequencies. The values of x are the class centres of the 10 cm intervals that were used in the grouping.

Class	80–90	90–100	100–110	110–120	120–130	130–140	140–150	150–160	160–170	170–180
x	85	95	105	115	125	135	145	155	165	175
f	1	3	2	5	8	6	34	22	16	3
f_r	0.01	0.03	0.02	0.05	0.08	0.06	0.34	0.22	0.16	0.03
F_r	0.01	0.04	0.06	0.11	0.19	0.25	0.59	0.81	0.97	1.00





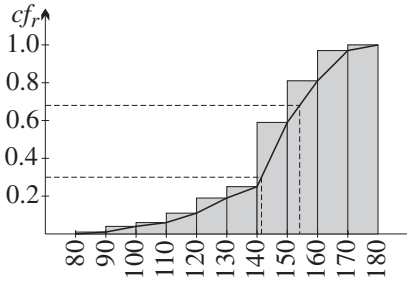
Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

Deciles and percentiles

We have seen that quartiles divide the scores into four equal lists. They can be read approximately from the cumulative relative frequency polygon by drawing horizontal lines at height 0.25 for the lower quartile Q_1, height 0.5 for the median Q_2, and height 0.75 for the upper quartile Q_3.

The graph to the right shows how *deciles* and *percentiles* can be found similarly.

- To find the 3rd decile, draw a horizontal line of height 0.3.
- To find the 68th percentile, draw a horizontal line of height 0.68.

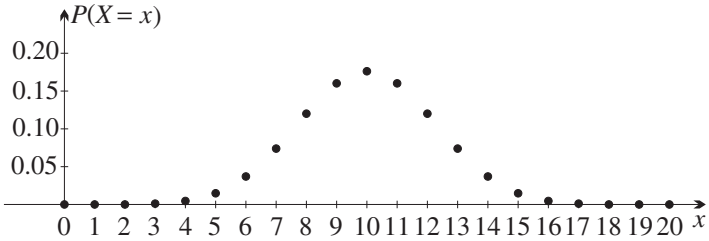


From the graph to the right, the 3rd decile is about 142 and the 68th percentile is about 154.

Note: This graphical method of finding medians and quartiles will give slightly different results from the method used in Sections 15A and 15C. Once the cumulative histogram is drawn, it may seem more natural to use the graph, and graphical methods fit better with the integrals that are needed in the continuous distributions of this chapter.

When does a discrete distribution begin to look continuous?

As the number of values of a discrete distribution increases, the graph of the distribution may suggest a curve. For example, when 20 coins are thrown and the number of heads recorded, the diagram below shows the graph of the resulting probability distribution (the calculations are omitted). There definitely seems to be a curve involved here.



This chapter is about continuous distributions. The more coins there are, the more difficult the calculations become, and the more attractive it is to work out some way to approximate the discrete distribution by a continuous distribution. This is one of many ways in which continuous distributions are useful.

Probability and area

Here is a rather simple probability problem that requires area and cannot possibly be reduced to a discrete probability distribution.

Example 1 16A

A point-chook is wandering randomly around a 20 m $\times$ 20 m square enclosure. It is just as likely to be at any one place in the enclosure as any other. A circle 10 metres in radius has been inscribed in the square. If Farmer Brown looks out at the enclosure, what is the probability that the chook is inside the circle?

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

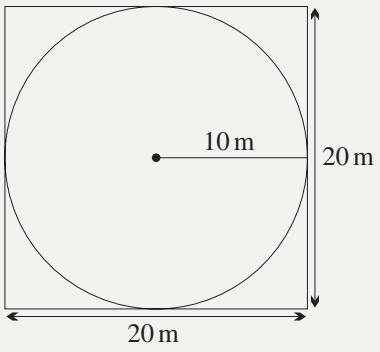
SOLUTION

Here area of enclosure $= 20^2$
$= 400 \text{ m}^2$,

and area of circle $= \pi r^2$
$= 100\pi \text{ m}^2$,

so $P(\text{chook is inside the circle}) = \dfrac{\text{area of circle}}{\text{area of square}}$

$= \dfrac{\pi}{4}$.



In this problem it is completely obvious that we take the ratio of areas. Yet the calculations have nothing to do with the discrete sample spaces that we have spent so much time analysing. The answer $\frac{\pi}{4}$ is not even a rational number. The association of probability with area is fundamental to the way we shall deal with continuous probability distributions.

Exercise 16A

FOUNDATION

Questions 1–5 are a short review of ideas from Chapter 13 of the Year 11 book

1 The probabilities in a discrete probability distribution must all be non-negative and add to 1. Which of the following are valid discrete probability distributions?

A

x	1	2	3	4
$P(X = x)$	0.2	0.3	0.3	0.2

B

x	1	2	3	4
$P(X = x)$	1.4	0	−0.5	0.1

C

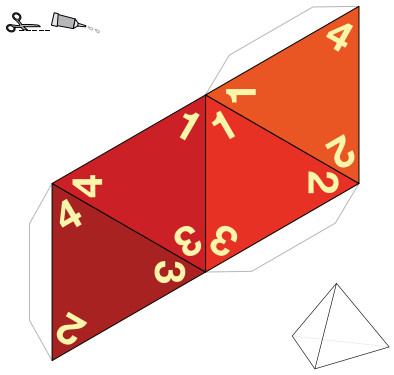
x	1	2	3	4
$P(X = x)$	0.15	0.2	0.4	0.25

D

x	1	2	3	4
$P(X = x)$	0.35	0.2	0.3	0.1

2 Two four-sided tetrahedral dice are thrown. The apex number on each die is read, and the sum of these two numbers is recorded. Let the random variable X be the outcome of this experiment.

Tetrahedral dice



a Record the possible outcomes and their probabilities in a probability table.

b Find:

i $P(X < 5)$ **ii** $P(X > 7)$

iii $P(X < 2)$ **iv** $P(X \leq 10)$

c Find the probability:

i that the sum is less than 4,

ii that the sum is odd,

iii that the sum is not 2,

iv that the sum is at least 6.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

3 A certain weighted spinner has five outcomes 1, 2, 3, 4 and 5. The spinner is known to be biased, and after a large number of trials, the relative frequencies of each outcome x were:

Score x	1	2	3	4	5	Total
f_r	0.1	0.2	0.45	0.15	0.1	
xf_r						
$x^2 f_r$						

a Complete the table.

b What is the significance of your total of row 2?

c Sum row 3, then use the formula $\bar{x} = \sum xf_r$ to calculate the sample mean $\bar{x}$.

d In your own words, what is the sample mean $\bar{x}$ measuring?

e Sum row 4, then use the formula $s^2 = \sum x^2 f_r - \bar{x}^2$ to calculate the sample variance.

f Hence calculate the sample standard deviation s, correct to two decimal places.

g In your own words, what is the sample standard deviation s measuring?

h What are the sample mean $\bar{x}$, and the sample standard deviation s, estimates of in this experiment?

i The spinner is thrown 100 times, and the outcome is recorded for each throw. State a reasonable estimate for the sum of these outcomes.

4 Data are recorded in the following table.

score x	3	4	5	6	7	Total
relative frequency f_r	0.04	0.21	0.35	0.25	0.15	

a Complete the table as in Question **3** to calculate the sample mean and sample standard deviation, correct to two decimal places.

b There are 5 values in this dataset and in the dataset in Question **3**. Comment on the centre and spread of the data in this set, compared with Question **3**.

5 These are the results from a class quiz where the maximum possible score is six marks.

score x	1	2	3	4	5	6	Total
frequency f	2	4	4	8	2	0	
$P(X = x)$							
$x \times P(X = x)$							
$x^2 \times P(X = x)$							

a Find the median and mode.

b Use the relative frequencies as probabilities to fill in the row for $P(X = x)$. Then complete the table.

c Use the formula $\mathrm{E}(X) = \sum x\, P(X = x)$ to estimate the expected value (also called the mean).

d Use the formula $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \sum x^2 P(X = x) - \mu^2$ to estimate the variance.

e Find the standard deviation.

f Comment on the class's performance in this quiz by reference to the distribution of these quiz scores.

g The teacher likes to record all quiz results out of 30, so he multiplies all these results by 5 before recording them in his markbook. What will be the mean and standard deviation of this new set of marks? You may find it helpful to remember the formulae

$$\mathrm{E}(aX + b) = a\mathrm{E}(X) + b \quad \text{and} \quad \mathrm{Var}(aX + b) = a^2\mathrm{Var}(X).$$

ISBN 978-1-108-76630-2

Start of Foundation for Section 16A of this chapter

6 A simple experiment has generated the following table of discrete data:

score x	1	2	3
frequency f	2	5	3

a i Construct a frequency histogram for the data. Add the frequency polygon to your diagram by joining the centres of the data points. Remember to join the ends of the polygon back to the horizontal axis.

ii Calculate the total area of the histogram rectangles.

iii Calculate the area under the frequency polygon, bounded by the horizontal axis.

iv What do you notice?

b i Copy the table, and add a row showing the relative frequency, obtained by dividing the frequencies by the total number of scores, which is 10.

ii Construct a relative frequency histogram for the data, including the relative frequency polygon.

iii Calculate the total area of the histogram rectangles.

iv Calculate the area under the relative frequency polygon, bounded by the horizontal axis.

v What do you notice?

vi What is the relationship between the relative frequencies and the probabilities $P(X = x)$ of the experiment's probability distribution?

7 a Copy and complete the following table by filling in the relative frequencies, cumulative frequencies and cumulative relative frequencies.

x	1	2	3	4	5	6	7	Total
f	3	1	4	3	1	3	1	
f_r								
cf								—
cf_r								—

b Construct a cumulative relative frequency histogram and polygon (ogive). Mark your vertical axis in divisions of $\frac{1}{8} = 0.125$.

c Use the ogive to read off the three quartiles Q_1, Q_2 and Q_3.

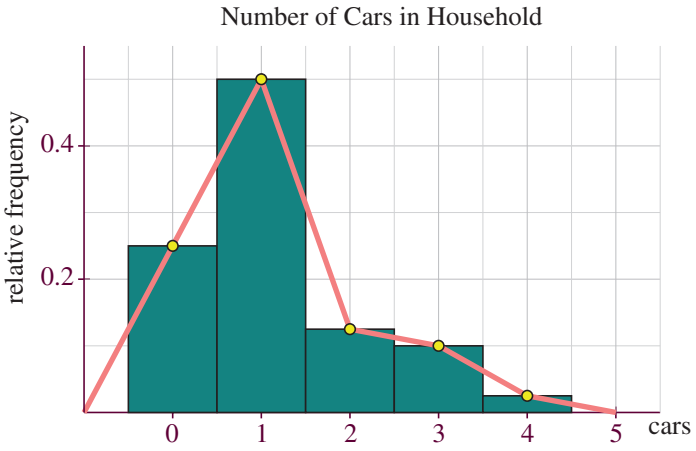
8 Repeat question 7 for the following dataset. Mark your vertical axis in divisions of 0.1.

x	5	6	7	8	9	10	11
f	5	4	1	1	1	3	5

ISBN 978-1-108-76630-2

DEVELOPMENT

9 For town-planning purposes, the number of cars owned by each household in a suburb was recorded from census data. The results are displayed in the relative frequency histogram and polygon below. (This is a population, so the relative frequencies are probabilities.)



a What fraction of the households have no cars?

b What fraction of the households have fewer than 2 cars?

c What is the probability that a household chosen at random has three cars?

d Town planners will advise that additional on-street parking be provided if more than 40% of the households have 3 or more cars. Will they be advising that additional parking be provided? Explain your answer.

e Copy and complete the following table for this probability distribution.

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

f Show that the sum of the probabilities is 1. How is this related to the area of the rectangles of the histogram?

g Explain in your own words, and with reference to the graph above, why the area bounded by the relative frequency polygon and the horizontal axis will be the same as the area of the relative frequency histogram.

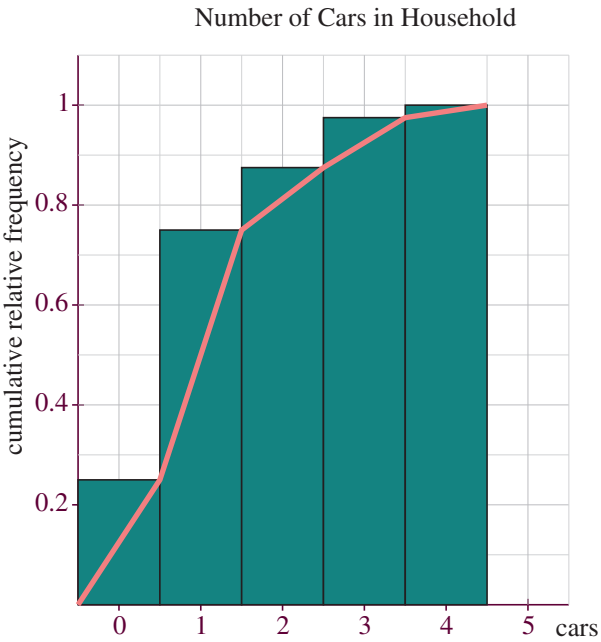
h Use your table to show that the mean number of cars per household is 1.15. What do you understand by this answer — how can a household have a fraction of a car?

i A street in the suburb is selected at random. If there are 100 households in the street, how many cars would you expect to belong to the households in the street in total? Are your assumptions for this estimate reasonable?

j Copy and complete the following table for the cumulative relative frequencies of this probability distribution.

x	0	1	2	3	4
$P(X \leq x)$					

k A town planner constructs a cumulative relative frequency polygon and histogram from these data. His graph is shown to the right. Confirm that your data agree with this graph.



l By drawing horizontal lines at heights 0.25, 0.5 and 0.75, find the three quartiles Q_1, Q_2 and Q_3.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

10 a Construct a cumulative relative frequency histogram and polygon (ogive) for the following data.

x	0	1	2	3	4
f_r	0.1	0.3	0.2	0.1	0.3

b Estimate the 70th percentile (also called the 7th decile) by intersecting the horizontal line at height 0.7 with the ogive.

c Similarly estimate the first quartile Q_1 and the median Q_2 using horizontal lines and the ogive.

d Similarly estimate the third quartile Q_3 using the ogive.

e [Challenge] Use ratios on the last segment of the ogive to calculate the quartile Q_3 using the formula $3.5 + 0.05 \times \dfrac{4.5 - 3.5}{1 - 0.7}$. Compare your answer with part **d**.

11 To raise funds, a school running a musical performance also runs a set of stalls selling cheap items. The total amount spent at the stalls by each person attending was recorded.

Amount spent ($)	0–1	1–2	2–3	3–4	4–5	Total
class centre x	0.50	1.50	2.50	3.50	4.50	—
frequency f	20	5	15	40	20	

(Any value in the range $0 \leq$ price < 1 is recorded in the class 0–1 etc.)

a Find the median and mode.

b Copy the table and add a row for the relative frequency.

c Calculate the expected value $\mathrm{E}(X)$ and the variance $\mathrm{Var}(X)$.

d Construct a relative frequency histogram, including the relative frequency polygon.

e Find the probability that an attendee spends between:

i \$0–\$1 **ii** \$1–\$2 **iii** \$2–\$3 **iv** \$3–\$4 **v** \$4–\$5.

f Find the sum of the probabilities in part **e**. What area does this represent?

g An attendee is chosen at random and asked how much they spent.

i Is the amount spent more likely to be \$0–\$3, or \$3–\$4?

ii Is the amount spent more likely to be \$0–\$1, or \$3–\$4?

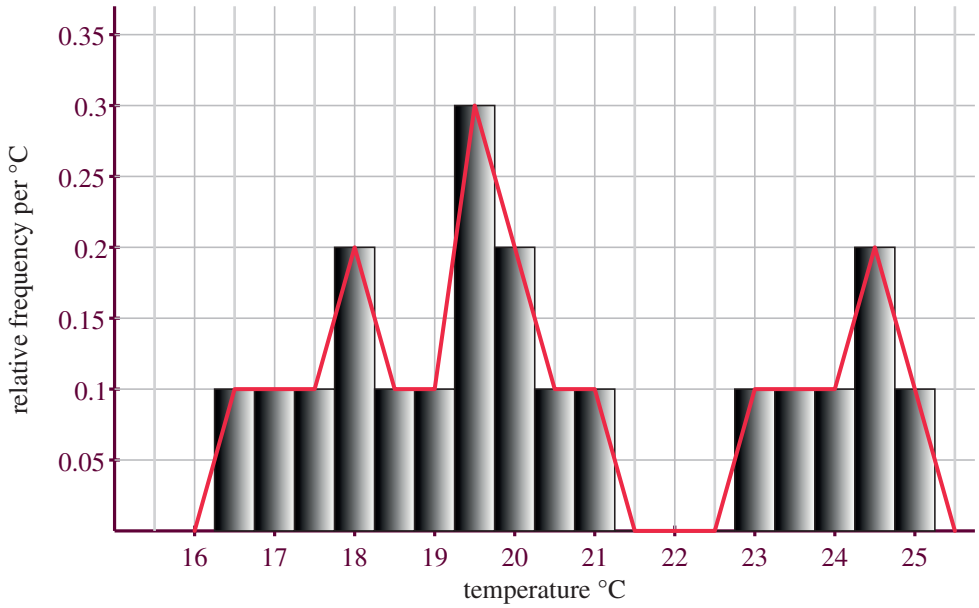
h If the school also charged an entry fee of \$2, find the expected value and variance of this new distribution $Y = X + 2$. What does the value $E(Y)$ represent?

ENRICHMENT

12 We have seen several times that when the width of the rectangles is 1, the rectangles of the relative frequency histogram have total area 1. When the rectangles have a width w different from 1, however, then the total area is w. We can restore the area to 1 by using instead a scale of 'relative frequency per unit' on the vertical axis, as in this question.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

The maximum temperature for the day over a period of twenty days at a local weather station is recorded. These temperatures are displayed in the relative frequency histogram and polygon below, where the rectangles each have width 0.5°C.



In this histogram, temperature (more accurately, the temperature class) is shown on the horizontal axis, and the *relative frequency per unit of temperature* is shown on the vertical axis. Temperature has been grouped in classes of 0.5°.

a Show that the total area under the histogram is 1. (Hint: Each box on the grid has an area of $0.025 = \frac{1}{40}$.)

b With this adjustment, the probability that the temperature will lie in a given class (or classes) is the area of the corresponding rectangle (or rectangles).

- **i** Find the probability that the maximum temperature is between 19.25°C and 19.75°C.
- **ii** Find the probability that the maximum temperature is between 16.25°C and 17.25°C.
- **iii** Find the probability that a day chosen at random is *warm*, if a warm day is defined to be one with a maximum of more than 22°C.

c The probability of a given temperature is proportional to the height of the frequency polygon at that point.

- **i** Estimate the relative likelihood of the maximum temperature being 17°C as compared with 20°C.
- **ii** What is the mode, that is, the most likely maximum temperature?

d The frequency polygon gives an estimate of the shape of the continuous probability distribution that would be obtained by successively grouping the data in narrower and narrower classes of temperatures. Use the area under the frequency polygon to estimate the probability that the maximum temperature on a given day is between:

- **i** 16.5°C and 17.5°C,
- **ii** 19°C and 20.5°C.

e Comment on the validity of using this histogram to decide on the probability of a given temperature at any time of the year.

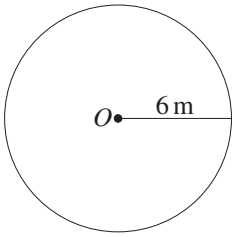
ISBN 978-1-108-76630-2

16B Continuous distributions

In a *continuous probability distribution*, the domain of the values is typically from a closed interval on the number line, such as $[0, 6]$. There are thus infinitely many values, which cannot even be listed. The probability of any one particular value is zero, and we want to talk instead about a probability such as $P(2 \leq X \leq 5)$, which is the probability that the value lies in the subinterval $[2, 5]$ of $[0, 6]$.

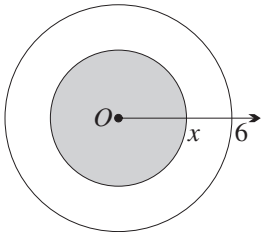
A cumulative distribution function

A point-chook is wandering randomly around a circular enclosure of radius 6 metres. It is just as likely to be in any one place in the enclosure as any other. Farmer Brown wants to know how far the chook is from the water at the centre O of the circle.

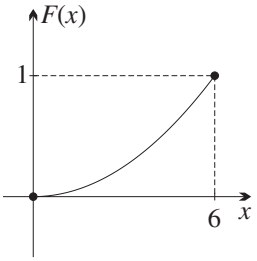


There are infinitely many distances from the centre within the enclosure. The probability that the chook is say exactly 2 metres from the centre is zero. Thus the tabular methods used with discrete probability distributions are useless here.

We can, however, approach the situation using cumulative frequency. Let $F(x)$ be the probability that when Farmer Brown looks out, the chook is no more than x metres from the centre.

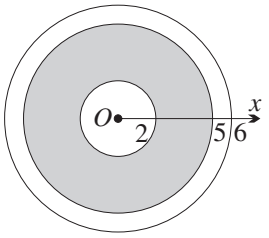


$$\begin{aligned} F(x) &= \frac{\text{area of inner circle}}{\text{area of whole circle}} \\ &= \frac{\pi x^2}{\pi \times 6^2} \\ &= \tfrac{1}{36}x^2, \quad \text{where } 0 \leq x \leq 6. \end{aligned}$$



This function is a *cumulative distribution function* or *CDF*. It is continuous, and increases from $F(0) = 0$ on the left to $F(6) = 1$ on the right. A cumulative function is always non-decreasing. It can also be used to solve many more problems. For example, we can find the probability that the chook is between 2 metres and 5 metres from the centre by subtraction,

$$\begin{aligned} P(\text{chook is 2–5 metres from the centre}) &= F(5) - F(2) \\ &= \tfrac{1}{36}(25 - 4) \\ &= \tfrac{21}{25}. \end{aligned}$$



We can also calculate the median and the quartiles of the probability distribution in the obvious way.

For the first quartile,
put $F(x) = \frac{1}{4}$
$\frac{1}{36}x^2 = \frac{1}{4}$
$x^2 = 9$
$x = 3$.

For the median,
put $F(x) = \frac{1}{2}$
$\frac{1}{36}x^2 = \frac{1}{2}$
$x^2 = 18$
$x \doteq 4.24$.

For the third quartile,
put $F(x) = \frac{3}{4}$
$\frac{1}{36}x^2 = \frac{3}{4}$
$x^2 = 27$
$x \doteq 5.20$.

We still have not precisely defined continuous probability distributions, but let us nevertheless summarise the discussion above.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

2 THE CUMULATIVE DISTRIBUTION FUNCTION

Let a continuous random variable X have values from a closed interval $[a, b]$.

- The *cumulative distribution function* or *CDF* for X is the function

  $F(x) = P(a \leq X \leq x)$, for all x in the interval $[a, b]$.

- The cumulative distribution function is continuous and non-decreasing, with

  $F(a) = 0$ and $F(b) = 1$.

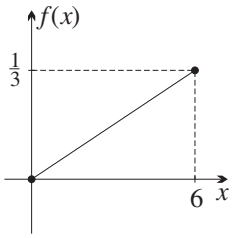
- It can be used to calculate medians, quartiles and percentiles.

A probability density function

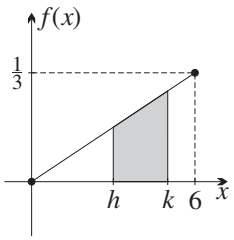
With a discrete distribution, the cumulative frequencies were obtained by adding all the probabilities up to a certain point — the same process produces the cumulative frequencies of a dataset. The continuous analogue of addition is integration, so we should expect the cumulative distribution function $F(x) = \frac{1}{36}x^2$ to be an *integral* over the values up to a certain point.

The fundamental theorem of calculus tells us that $F(x)$ is the integral of its derivative $F'(x)$. So we differentiate $F(x)$ to obtain what is called the *probability density function* $f(x)$,

$$\begin{aligned} f(x) &= \frac{d}{dx}\left(\tfrac{1}{36}x^2\right) \\ &= \tfrac{1}{18}x, \text{ where } 0 \leq x \leq 6. \end{aligned}$$



This linear graph of $f(x)$ is sketched above. It does not tell us the probability that the chook is x metres from the centre, because that probability is zero. Instead, it allows us to find by integration the probability that the chook is in some range of distances from centre.



The probability of the chook being in some range of positions is the area under the curve, which is found by integration.

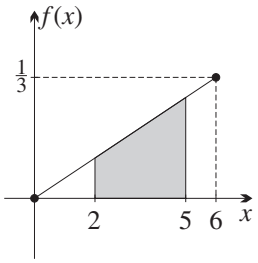
$$P(h \leq x \leq k) = \int_a^b f(x)\,dx.$$

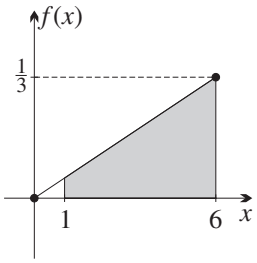
For example,

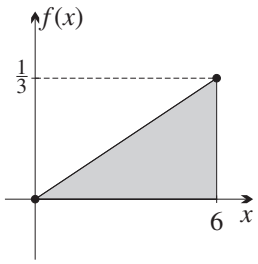
$$\begin{aligned} P(2 \leq x \leq 5) &= \int_2^5 \tfrac{1}{18}x\,dx \\ &= \left[\tfrac{1}{36}x^2\right]_2^5 \\ &= \tfrac{1}{36}(25 - 4) \\ &= \tfrac{21}{25}, \end{aligned}$$

$$\begin{aligned} P(x \geq 1) &= \int_1^6 \tfrac{1}{18}x\,dx \\ &= \left[\tfrac{1}{36}x^2\right]_1^6 \\ &= \tfrac{1}{36}(36 - 1) \\ &= \tfrac{35}{36}, \end{aligned}$$

$$\begin{aligned} P(0 \leq X \leq 6) &= \int_0^6 \tfrac{1}{18}x\,dx \\ &= \left[\tfrac{1}{36}x^2\right]_0^6 \\ &= \tfrac{1}{36}(36 - 0) \\ &= 1. \end{aligned}$$







Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

This probability density function has two important properties:

1 $f(x)$ is never negative, because $F(x)$ is cumulative and never decreases.

2 $\int_0^6 f(x)\,dx = 1$, because the chook is somewhere in the enclosure.

It turns out that the probability density function is more important than the cumulative distribution function when characterising a continuous distribution and working with it. It is also the best way to give a formal definition of a continuous probability distribution.

3 PROBABILITY DENSITY FUNCTIONS

- A *probability density function*, or *PDF*, is a function defined on a closed interval $[a, b]$ and satisfying two properties:
  1 $f(x) \geq 0$, for $a \leq x \leq b$.
  2 $\int_a^b f(x)\,dx = 1$.
- A *continuous probability distribution* is defined to be a probability distribution described by a probability density function.
- A global maximum of the probability density function is called a *mode*.
- Probability is area under the curve. That is, for all closed subintervals $[h, k]$,

$$P(h \leq X \leq k) = \int_h^k f(x)\,dx.$$

- Later, we will allow a to be replaced by $-\infty$, and b to be replaced by ∞.

The probability of any particular value h, however, is always zero. That is,

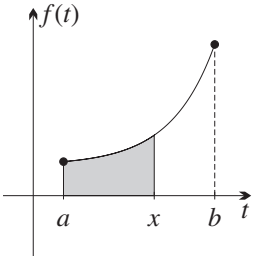
$$P(X = h) = \int_h^h f(x)\,dx = 0.$$

We remarked on such integrals in Box 8 of Section 5C. Because of this, it doesn't really matter whether $\leq$ or $<$ is used for intervals, or whether two adjacent intervals, say $[2, 4]$ and $[4, 5]$, overlap at the endpoints.

The CDF is the signed area function of the PDF

Suppose again that $f(x)$ is a probability density function defined on the closed domain $[a, b]$. Using the language of Section 5D, we can now use integration to define the cumulative distribution function $F(x)$ of $f(x)$ as the signed area function of the PDF,

$$F(x) = \int_a^x f(t)\,dt, \text{ for } a \leq x \leq b.$$



ISBN 978-1-108-76630-2

4 THE CUMULATIVE DISTRIBUTION FUNCTION AS THE SIGNED AREA FUNCTION

The *cumulative distribution function* $F(x)$, or *CDF* for short, of a probability density function, or PDF, is the signed area function

$$F(x) = P(X \leq x) = \int_a^x f(t)\,dt, \quad \text{for } a \leq x \leq b,$$

and conversely $f(x) = F'(x)$ (apart possibly from isolated sharp corners).

Be pedantic and say 'probability density function' and 'cumulative distribution function'. 'Density' means at a point, and 'distribution' means over a range.

Uniform continuous distributions

An important special case of continuous probability distributions is a *uniform continuous distribution*. This is a distribution whose PDF is a constant function.

Example 2 16B

Tran does not know the times when trains leave Lakeside Station, but he does know that they leave precisely every fifteen minutes.

a He wants to know about the probability distribution of his waiting time if he arrives at the station at a random time, and what the PDF and CDF are.

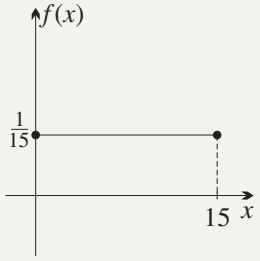
b He also wants to know the median and the 45th percentile, and the probability that he will wait between 5 and 10 minutes.

SOLUTION

a The waiting time x is anything from 0 to 15 minutes, and we have no reason to prefer any waiting time from any other waiting time. This means that the probability density function $f(x)$ is a constant function in the interval $[0, 15]$, and the probability distribution is therefore a *uniform continuous distribution* with values from the closed interval $[0, 15]$.

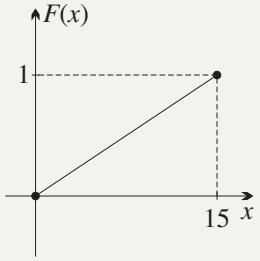
Because the area under the PDF is exactly 1 (the total probability),

$$f(x) = \tfrac{1}{15}, \quad \text{for } 0 \leq x \leq 15.$$



The CDF $F(x)$ is then found by integrating,

$$\begin{aligned} F(x) &= \int_0^x \tfrac{1}{15}\,dt \\ &= \left[\tfrac{1}{15}t\right]_0^x \\ &= \tfrac{1}{15}x. \end{aligned}$$



Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

b To find the probability that he waits between 5 and 10 minutes, either integrate the PDF or use the CDF.

$$P(5 \leq X \leq 10) = \int_5^{10} \tfrac{1}{15}\,dx$$
$$= \left[\tfrac{1}{15}x\right]_5^{10}$$
$$= \tfrac{2}{3} - \tfrac{1}{3}$$
$$= \tfrac{1}{3},$$

OR

$$P(5 \leq X \leq 10) = F(10) - F(5)$$
$$= \tfrac{2}{3} - \tfrac{1}{3}$$
$$= \tfrac{1}{3}.$$

For the median, put $F(x) = \frac{1}{2}$

$$\tfrac{1}{15}x = \tfrac{1}{2}$$
$$x = 7\tfrac{1}{2}.$$

For the 45th percentile, put $F(x) = \frac{45}{100}$

$$\tfrac{1}{15}x = \tfrac{9}{20}$$
$$x = 6\tfrac{3}{4}.$$

5 UNIFORM CONTINUOUS DISTRIBUTIONS

- A continuous distribution is called *uniform* if its density function is constant.
- Because the area under the graph is 1, a uniform continuous distribution defined on an interval $[a, b]$ has probability density function $y = \dfrac{1}{b - a}$.

Example 3 16B

Find the value of k that makes each function a probability density function. Then find the corresponding CDF $F(x)$. Hence find the median and quartiles.

a $f(x) = k$, where $0 \leq x \leq 10$,

b $f(x) = kx$, where $0 \leq x \leq 10$.

SOLUTION

a Put $\displaystyle\int_0^{10} k\,dx = 1$

$$\left[kx\right]_0^{10} = 1$$
$$10k - 0 = 1$$
$$k = \tfrac{1}{10},$$

so $f(x) = \frac{1}{10}$.

Hence $\displaystyle F(x) = \int_0^x \tfrac{1}{10}\,dt$

$$= \left[\tfrac{1}{10}t\right]_0^x$$
$$= \tfrac{1}{10}x.$$

When $F(x) = \frac{1}{2}, x = 5$,

when $F(x) = \frac{1}{4}, x = 2\frac{1}{2}$,

when $F(x) = \frac{3}{4}, x = 7\frac{1}{2}$,

so $Q_1 = 2\frac{1}{2}, Q_2 = 5$ and $Q_3 = 7\frac{1}{2}$.

b Put $\displaystyle\int_0^{10} kx\,dx = 1$

$$\left[\tfrac{1}{2}kx^2\right]_0^{10} = 1$$
$$50k - 0 = 1$$
$$k = \tfrac{1}{50},$$

so $f(x) = \frac{1}{50}x$.

Hence $\displaystyle F(x) = \int_0^x \tfrac{1}{50}t\,dt$

$$= \left[\tfrac{1}{100}t^2\right]_0^x$$
$$= \tfrac{1}{100}x^2.$$

When $F(x) = \frac{1}{2}$, $x = 5\sqrt{2}$,

when $F(x) = \frac{1}{4}$, $x = 5$,

when $F(x) = \frac{3}{4}$, $x = 5\sqrt{3}$,

so $Q_1 = 5, Q_2 = 5\sqrt{2}$ and $Q_3 = 5\sqrt{3}$.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

Piecewise-defined probability density functions

The next worked example shows how to deal with a probability density function that is piecewise defined.

Example 4 10B

A probability density function is defined piecewise by

$$f(x) = \begin{cases} k(4 + x), & \text{for } -4 \leq x \leq 0, \\ k(4 - x), & \text{for } 0 \leq x \leq 4. \end{cases}$$

a Find the value of the constant k. Hence write the equation of $f(x)$ and sketch it.

b What is the probability that $0 \leq X \leq 2$?

c Why is the median zero, and what is the mode?

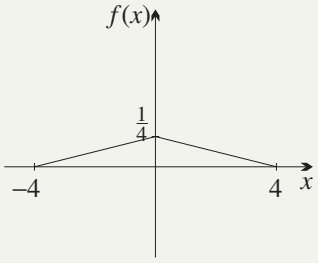
d Find the CDF for $-4 \leq x \leq 0$, and the CDF for $0 \leq x \leq 4$. Then sketch the whole CDF.

SOLUTION

a The integral over the domain $[-4, 4]$ must be 1. Using areas of triangles is easier, but here is how to integrate piecewise over the domain by dissection.

$$\begin{aligned} \int_{-4}^{4} f(x)\,dx &= \int_{-4}^{0} k(4 + x)\,dx + \int_{0}^{4} k(4 - x)\,dx \\ &= k\left[4x + \tfrac{1}{2}x^2\right]_{-4}^{0} + k\left[4x - \tfrac{1}{2}x^2\right]_{0}^{4} \\ &= k\Big((0 + 0) - (-16 + 8) + (16 - 8) - (0 - 0)\Big) \\ &= 16k, \end{aligned}$$

so for the integral to be 1, the value of k must be $k = \frac{1}{16}$.



The function is therefore $f(x) = \begin{cases} \frac{1}{16}(4 + x), & \text{for } -4 \leq x \leq 0, \\ \frac{1}{16}(4 - x), & \text{for } 0 \leq x \leq 4. \end{cases}$

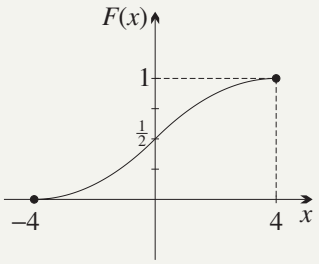
b $$\begin{aligned} P(0 \leq X \leq 2) &= \int_{0}^{2} f(x)\,dx \\ &= \tfrac{1}{16}\int_{0}^{2} (4 - x)\,dx \qquad \text{(only the right-hand branch is relevant)} \\ &= \tfrac{1}{16}\left[4x - \tfrac{1}{2}x^2\right]_{0}^{2} \\ &= \tfrac{1}{16}\Big((8 - 2) - (0 - 0)\Big) \\ &= \tfrac{3}{8} \qquad \text{(or use the area of a trapezium).} \end{aligned}$$

c The areas to the left and right of $x = 0$ are equal, so the median is 0.
The mode is also $x = 0$, because there is a global maximum there.

ISBN 978-1-108-76630-2

d For $-4 \leq x \leq 0$,

$$\begin{aligned} F(x) &= \tfrac{1}{16}\int_{-4}^{x} (4 + t)\,dt \\ &= \tfrac{1}{32}\Big[(4 + t)^2\Big]_{-4}^{x} \\ &= \tfrac{1}{32}\Big((4 + x)^2 - 0\Big) \\ &= \tfrac{1}{32}(4 + x)^2. \end{aligned}$$



Hence $F(0) = \frac{1}{2}$, so for $0 \leq x \leq 4$,

$$\begin{aligned} F(x) &= \tfrac{1}{2} + \tfrac{1}{16}\int_{0}^{x} (4 - t)\,dt \\ &= \tfrac{1}{2} - \tfrac{1}{32}\Big[(4 - t)^2\Big]_{0}^{x} \\ &= \tfrac{1}{2} - \tfrac{1}{32}\Big((4 - x)^2 - 16\Big) \\ &= 1 - \tfrac{1}{32}(4 - x)^2. \end{aligned}$$

Distributions with unbounded domains

We have been using integrals (and sometimes area formulae) to find areas. In many important situations, the probability density function has a horizontal asymptote, however, which means that the possible values extend to infinity. For example, the diagram in Section 16A (page 176) involving 20 tossed coins suggested approximating that discrete distribution by a continuous curve with asymptotes on the left and right.

The radioactive isotope iodine-131 is often used in medicine for the treatment of thyroid cancer. It has a half-life of about 8 days. Suppose that we isolate a single nucleus of iodine-131, observe it constantly, and record the time X in days before it decays. Then using the fact that the isotope has a half-life of 8 days,

$$P(X > 8) = \tfrac{1}{2}, \qquad P(X > 16) = \tfrac{1}{4}, \qquad P(X > 24) = \tfrac{1}{8}, \qquad \ldots$$

and taking the complementary events,

$$P(X \leq 8) = \tfrac{1}{2}, \qquad P(X \leq 16) = \tfrac{3}{4}, \qquad P(X \leq 24) = \tfrac{7}{8}, \qquad \ldots$$

In general, $P(X \leq 8n) = 1 - 2^{-n}$.

This formula holds for all real values of $n \geq 0$, not just for whole numbers, and to find $P(X \leq x)$, put $x = 8n$,

then $n = \frac{1}{8}x$, giving $P(X \leq x) = 1 - 2^{-\frac{1}{8}x}$.

This last formula is the cumulative distribution function $F(x)$ for the experiment. The next worked example continues the story. We first change to base e, and write

$$F(x) = 1 - e^{-kx}, \qquad \text{where } k = \tfrac{1}{8}\ln 2 = 0.08664\ldots \qquad \text{(store in memory).}$$

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

Example 5 10B

Let $f(x)$ and $F(x) = e^{-kx}$, where $k = \frac{1}{8}\ln 2$, be the PDF and CDF respectively for the experiment described above, observing the time x days that an iodine-131 nucleus survives before decaying.

a Explain why the domain of possible values is the unbounded interval $[0, \infty)$

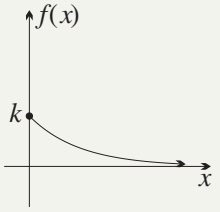
b Find the formula for the PDF, and sketch the CDF and PDF.

c Find the median, and show that it is the half-life.

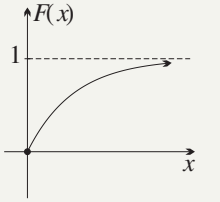
d Find the probabilities that it decays on the first day and after the first day.

SOLUTION

a The experiment is extremely unlikely to last beyond a month or two, but it is minutely possible that it will continue for 10 years or even more. Thus we use the unbounded interval $[0, \infty)$ for the domain of possible values.



b The CDF is $F(x) = 1 - e^{-kx}$.
Differentiating, $f(x) = ke^{-kx}$, which is the PDF.



c To find the median, put $F(x) = 0.5$

$$1 - e^{-kx} = \tfrac{1}{2}$$
$$e^{-kx} = \tfrac{1}{2}$$
$$kx = \ln 2,$$

and using calculator or logs, $x = 8$ days, which is the half-life.

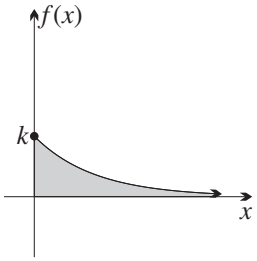
d $P(X \leq 1) = F(1) = 1 - e^{-k} \doteq 0.083$

$P(X > 1) = 1 - P(X \leq 1) = e^{-k} \doteq 0.917$

Improper integrals

Worked Example 5 above is quite sufficient preparation for the normal distribution later in the chapter. Readers may ask, however, how we can reasonably say that the area under the curve in the unbounded interval $[0, \infty)$ is 1 square unit, when it runs off to infinity! The integral involved here is called an *improper integral* — here is how to deal with it (regard this as Enrichment). The PDF is $y = ke^{-kx}$.

$$\text{area under curve over the interval } [0, \infty) = \int_0^\infty ke^{-kx}\,dx$$
$$= \left[-e^{-kx}\right]_0^\infty.$$



Substituting $x = 0$ gives a value -1 for the primitive.
We cannot substitute $x = \infty$ because ∞ is not a number,
but we can take the limit of $-e^{-kx}$ as $x \to \infty$, which is 0,

so $$\int_0^\infty ke^{-kx}\,dx = 0 - (-1)$$
$$= 1.$$

Thus we can reasonably say that the unbounded shaded area is 1 square unit.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

Example 6 10B

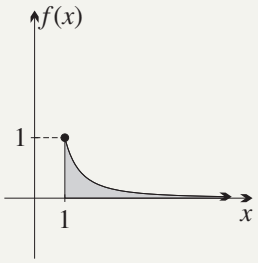
a Find the shaded area under the curve $y = \frac{1}{x^2}$ over the interval $(1, \infty)$.

b Hence show that $y = \frac{1}{x^2}$, for $x \geq 1$, is a PDF, and find the CDF.

SOLUTION

a The improper integral over the closed interval $[1, \infty)$ is

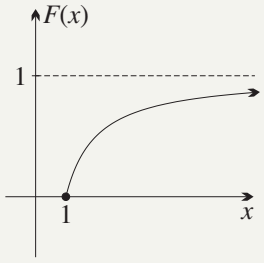
$$\int_1^\infty \frac{1}{x^2}\,dx = \left[-\frac{1}{x}\right]_1^\infty$$



When $x = 1$, the primitive is -1.
We cannot substitute ∞ because ∞ is not a number, but we can take the limit as $x \to \infty$, which is 0.

so $$\int_1^\infty \frac{1}{x^2}\,dx = 0 - (-1)$$
$$= 1.$$

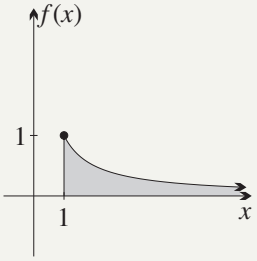
b Thus the area is 1 square unit, and the function $y = \frac{1}{x^2}$ is always positive in the interval $[1, \infty)$, so it is a PDF.



For the CDF, $$F(x) = \int_1^x \frac{1}{t^2}\,dt$$
$$= \left[-\frac{1}{t}\right]_1^x$$
$$= 1 - \frac{1}{x}.$$

Example 7 10B

Show that the improper integral $\int_1^\infty \frac{1}{x}\,dx$ does not converge to a limit.



SOLUTION

Using the same procedure as before,

$$\int_1^\infty \frac{1}{x}\,dx = \left[\log_e x\right]_1^\infty.$$

Substituting $x = 1$ gives a value $\log_e 1 = 0$ for the primitive.
We cannot substitute $x = \infty$ because ∞ is not a number, but neither can we take the limit, because as $x \to \infty$, $\log_e x \to \infty$.
The conclusion is that the region has infinite area, and that the improper integral does not converge.

This example is rather striking because if we look only at the graph, the curves $y = e^{-x}$, $y = \frac{1}{x^2}$ and $y = \frac{1}{x}$ all look so similar in their asymptotic behaviour.

ISBN 978-1-108-76630-2

Exercise 16B

FOUNDATION

1 **a** Sketch $f(x) = \frac{1}{2}$, where $0 \leq x \leq 2$. Then show that it satisfies the two conditions for a probability density function:

i Check from the graph that $f(x) \geq 0$, for all x in the domain.

ii Check that the area under the curve is 1, that is, that $\int_a^b f(x)\,dx = 1$.

b Repeat part **a** for $f(x) = \frac{1}{2}x$, where $0 \leq x \leq 2$.

c Repeat part **a** for $f(x) = \frac{1}{42}x$, where $4 \leq x \leq 10$.

2 Recall that a function $f(x)$ with domain the closed interval $[a, b]$ is called *a probability density function*, or *PDF* for short, if

$$f(x) \geq 0, \text{ for all } x \text{ in the domain} \qquad \text{and} \qquad \int_b^a f(x)\,dx = 1.$$

Determine whether or not each function is a probability density function. If it is a PDF, find its mode (look for global maxima).

a $f(x) = 3x^2$, where $0 \leq x \leq 1$

b $f(x) = \frac{1}{4}x$, where $1 \leq x \leq 5$

c $f(x) = \dfrac{4 - 2x}{3}$, where $0 \leq x \leq 3$

d $f(x) = (n + 1)x^n$, where $0 \leq x \leq 1$

e $f(x) = \frac{1}{2}\sin x$, where $0 \leq x \leq \pi$

f $f(x) = \frac{1}{12}(3x^2 + 2x)$, where $0 \leq x \leq 2$

3 Let $f(x) = \frac{3}{4}(x^2 - 4x + 3)$ be a function defined on the closed interval $[0, 4]$.

a Show that $\int_0^4 f(x)\,dx = 1$.

b Show nevertheless that $f(x)$ is not a valid probability density function. (Hint: Sketch the graph of $y = f(x)$.)

4 For a distribution defined by a probability density function $f(x)$, the probability that x lies in the interval $[h, k]$ is the area given by $P(h \leq X \leq k) = \int_h^k f(x)\,dx$.

a Sketch the uniform probability density function $f(x) = \frac{1}{4}$, where $0 \leq x \leq 4$.

b Confirm that it satisfies the two requirements for a probability density function.

c By calculating areas, find:

i $P(0 \leq X \leq 1)$ **ii** $P(1 \leq X \leq 3)$ **iii** $P(X \leq 2)$

iv $P(X) = 2$ **v** $P(X \leq 3)$ **vi** $P(X \geq 1)$

d Confirm that $P(2 \leq X \leq 3) = P(x \leq 3) - P(x \leq 2)$.

5 Recall that for a probability density function, or PDF, defined on the interval $[a, b]$, the cumulative distribution function, or CDF, is $F(x) = \int_a^x f(t)\,dt$, for $a \leq x \leq b$.

For each PDF, calculate the CDF $F(x)$ and confirm that $F(b) = 1$.

a $f(x) = \frac{1}{32}x$, where $0 \leq x \leq 8$. (In this case $a = 0$ and $b = 8$.)

b $f(x) = \frac{3}{16}x^2$, where $-2 \leq x \leq 2$.

c $f(x) = \frac{3}{2}(1 - x^2)$, where $0 \leq x \leq 1$.

d $f(x) = \frac{1}{e}(e^x + 1)$, where $0 \leq x \leq 1$.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

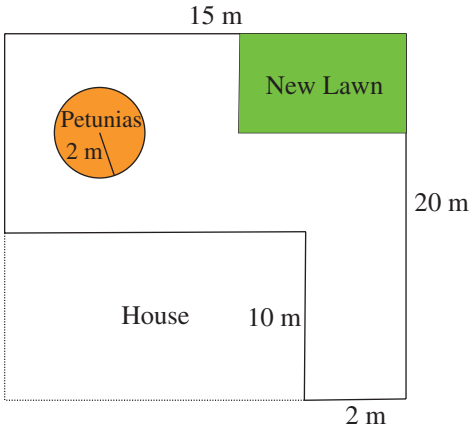
6 For the PDFs in parts **a** and **b** of Question **5**, use the CDF $F(x)$ to calculate:

a the median Q_2, by finding the value x such that $F(x) = 0.5$,

b the quartiles Q_1, by solving $F(x) = 0.25$, and Q_3, by solving $F(x) = 0.75$.

DEVELOPMENT

7 When Jack is at work, he shuts his dog Bud in the L-shaped backyard of his house. This is shown in the diagram to the right. Bud wanders around at random during the day, waiting for Jack to come home. Bud is the only cause of stress in Jack's quiet neighbourhood.



a When Bud is in the area directly to the right of the house, he will be anxious and howl for Jack to come home, to the distress of the neighbours, who shout at him. What is the probability that the neighbours will be stressed?

b If Bud is in the petunia patch, it will stress Jack's mother. What is the probability that Jack's mother will be stressed?

c If the neighbours or Jack's mother are shouting at Bud, Sally the cat cannot have a quiet sleep. What is the probability that Sally will be stressed?

d If the neighbours and Jack's mother aren't complaining, Jack's father is worried that Bud might be digging up his piece of new lawn. What is the probability that Jack's father will be stressed?

8 The function $y = 2x$, for $0 \leq x \leq 1$, is a probability density function.

a Sketch the graph and check that $f(x) \geq 0$.

b Use your diagram to show that the area bounded by the function and the x-axis is 1. Then check your result by integration.

c **i** Mark a point x between 0 and 1 on your diagram, and use area formulae to show that the cumulative distribution function is $P(X \leq x) = x^2$.

ii Confirm your result by calculating the integral $P(X \leq x) = \int_0^x 2t\,dt$.

d Use your expression for the cumulative distribution function to calculate the three quartiles Q_1, Q_2 and Q_3.

9 Find the value of the unknown constant c, given that each function $f(x)$ is a probability density function on the given domain.

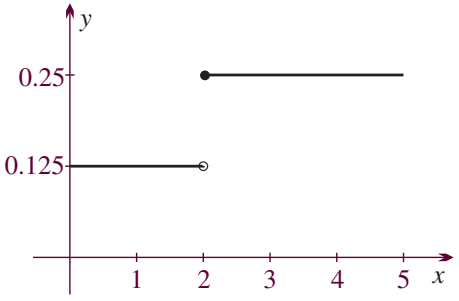
a $y = cx^4$, with domain $[0, 3]$

b $y = c$, with domain $[0, 6]$

c $y = c$, with domain $[-5, 5]$

d $f(x) = \frac{8}{3}(1 - x)$, with domain $[0, c]$

10 A function is graphed to the right.



a Verify that the function forms a valid PDF.

b Fill in the following table of values for the cumulative probabilities $P(X \leq x)$.

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

c Use your table to plot these points. Hence graph the cumulative probability function $P(X \leq x)$ for $0 \leq x \leq 5$.

d Write down a formula for the CDF, writing your answer in piecewise notation.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

11 A probability density function is defined by:

$$f(x) = \begin{cases} c, & \text{for } 0 \leq x \leq 5, \\ 2c, & \text{for } 5 < x \leq 10. \end{cases}$$

a Sketch the probability density function.
b Find the value of c.
c Find an expression for the cumulative distribution function.
d Use your cumulative distribution function to find $P(1 < X < 7)$.

12 Define a probability density function by $f(x) = \frac{3}{32}x(4 - x), 0 \leq x \leq 4$.
a Sketch the probability density function, and state its mode.
b Confirm that $\int_0^4 f(x)\,dx = 1$.
c Write down $P(X \leq 2)$. What other property of the curve enables us to do this without calculating an integral?
d Evaluate $P(X \leq 1)$ and $P(X > 1)$. Explain why your two results add to 1.
e Evaluate $P(X \leq 0.5)$ and $P(X \geq 3.5)$. What do you notice about your answers?
f Determine the cumulative distribution function, defined by $F(x) = P(X \leq x)$, using the formula $F(x) = \int_0^x f(t)\,dt$.
g Use your cumulative distribution function (CDF) to evaluate:
i $P(X < 1.5)$ **ii** $P(1 < X < 1.5) = P(X < 1.5) - P(X < 1)$
iii $P(3 < X < 3.5)$ **iv** $P(2 < X < 2.5)$
h Graph the CDF in your book.
i By evaluating $P(X < 2)$ using your CDF, confirm that 50% of the data lie to the left of the line $x = 2$.
j [Technology] Plot the cumulative distribution function and determine the upper and lower quartiles, defined by $P(X < Q_1) = 0.25$ and $P(X < Q_3) = 0.75$.

13 Define $f(x) = ce^{-x}$, where $0 \leq x \leq 1$.
a Sketch the curve $y = f(x)$.
b Find c, given that $f(x)$ is a probability density function.
c Find the cumulative distribution function.
d Find the quartiles Q_1, Q_2 and Q_3.

14 Grouping approximates a continuous distribution by a discrete distribution. A few trials of an experiment generated data in the interval 1.5– 4.5, and the data were grouped in class intervals of width 1.

Class	1.5– 2.5	2.5– 3.5	3.5– 4.5
Class centre	2	3	4
Relative frequency	0.3	0.4	0.3

a Use this dataset to draw a relative frequency histogram and a relative frequency polygon.
b Find the total area of the histogram and the area under the polygon.
c On a new set of axes, draw the cumulative relative frequency histogram and polygon.
d Estimate the three quartiles Q_1, Q_2 and Q_3 by reading off the corresponding values on the horizontal axis for the relative frequencies 0.25, 0.5 and 0.75.

ISBN 978-1-108-76630-2

e After running more trials and taking finer intervals, the experimenter decides that the data can best be modelled by the curve

$f(x) = \frac{3}{32}(x - 1)(5 - x)$, where $1 \leq x \leq 5$.

i Check that this curve is a probability density function.

ii Tabulate $f(x)$ for $x = 2, 3$ and 4, then graph it on top of the relative frequency polygon and compare the two (this would be easier with suitable technology).

iii Find the cumulative distribution function $F(X) = \int_1^x f(t)\,dt$, for $1 \leq x \leq 5$.

iv Substitute your three estimates for the quartiles into the cumulative distribution function. How close are your answers to 25%, 50% and 75%?

v [Technology] Graph the cumulative distribution function found by integration and read off the resulting estimates for the quartiles.

ENRICHMENT

15 This question and the next both involve improper integrals where the upper limit is ∞. Infinity is not a number, so you cannot substitute ∞. Instead, take the limit of the primitive as $x \to \infty$.

Let $f(x) = \frac{1}{x^2}$, where $x \geq 1$. Notice that this function is defined on an unbounded domain.

a Show that $f(x) > 0$ and $\int_1^\infty f(x)\,dx = 1$.

b Evaluate the cumulative distribution function $F(x)$.

c Confirm that $F(x) \to 1$ as $x \to \infty$. Why is this significant?

d Evaluate the three quartiles Q_1, Q_2 and Q_3.

16 Repeat the previous question for the function $f(x) = e^{-x}$, where $x \geq 0$, changing the limits 1 and ∞ of the integral to 0 and ∞.

17 Monte draws a unit square centred at $(0.5, 0.5)$ and inscribes a circle of radius 0.5 unit. He reasons that the probability of a random point in the square falling within the circle is proportional to the ratio of the areas of the shapes.

a Show that the ratio of the areas is $\frac{\pi}{4}$.

b Monte enters the code `=IF((RAND()-0.5)^2+(RAND()-0.5)^2<0.25,1,0)` in cell `A1` of a spreadsheet.

i In this code, the first `RAND()` selects the x-coordinate of a random point in the square, and the second `RAND()` selects the y-coordinate. Explain what value the code `(RAND()-0.5)^2+(RAND()-0.5)^2` will return for a random point.

ii What value does the code in `A1` return if the point is *inside* the circle?

iii What value does the code in `A1` return if the point is *outside* the circle?

c Monte fills the code down to the first 2000 cells of column `A`, and in cell `C1` enters `=4*AVERAGE(A:A)`. What value is the code `AVERAGE(A:A)` measuring, and what value should C1 be approaching?

d Use more cells in column `A` and use the `RECALCULATE` feature on your spreadsheet to investigate the accuracy of this method of determining π.

e This method could be used on your calculator. On calculators that provide the command `RAN#` to generate a random number in the interval $[0, 1)$, type the code `(Ran#-0.5)^2+(Ran#-0.5)^2`. Every time you enter this command it should use a new random point. If it is less than 0.25, count the point as in the circle. To make this procedure more accurate, do it as a class exercise and combine your results.

f Look up further details of the *Monte Carlo method* on the web.

ISBN 978-1-108-76630-2

16C Mean and variance of a distribution

The expected value (or mean) and the variance of a continuous probability distribution are obtained in almost the same ways as with discrete probability distributions. The main difference is that we replace the sum with its sigma notation $\sum$ by the integral with its integral notation $\int$. The Greek sigma and this early form of the letter S both correspond to the Latin letter S for 'sum'.

The expected value or mean of a continuous distribution

The *mean* or *expected value* of a discrete distribution is

$$\mathrm{E}(X) = \sum x\, p\,(x)\text{, summing over the whole distribution.}$$

The continuous analogue of addition is integration, so the continuous version is

$$\mathrm{E}(X) = \int_a^b xf(x)\,dx\text{, integrating over the whole interval } [a, b\,].$$

The variance and standard deviation of a continuous distribution

The *variance* of a discrete distribution has two equivalent forms:

$$\mathrm{Var}(X) = E\left((X-\mu)^2\right) = \sum (x-\mu)^2 p(x),$$
$$\mathrm{Var}(X) = E(X^2) - \mu^2 = \sum x^2 p(x) - \mu^2.$$

The continuous analogues of these two forms are

$$\mathrm{Var}(X) = E\left((X-\mu)^2\right) = \int_a^b (x-\mu)^2 f(x)\,dx,$$
$$\mathrm{Var}(X) = E(X^2) - \mu^2 = \int_a^b x^2 f(x)\,dx - \mu^2.$$

The equality of these two expressions is proven in Question 7 of Exercise 16C.

Example 8 16C

Apply all this to the chook at the start of Section 16B, where $f(x) = \frac{1}{18}x$.

SOLUTION

$$\begin{aligned}\mu &= \int_0^6 x \times \tfrac{1}{18}x\,dx \\ &= \tfrac{1}{54}\left[x^3\right]_0^6 \\ &= \tfrac{216}{54} - 0 \\ &= 4,\end{aligned}$$

so the chook's mean or expected distance from the centre is 4 metres.

ISBN 978-1-108-76630-2

$$\sigma^2 = \int_0^6 (x-4)^2 \times \tfrac{1}{18}x\,dx$$
$$= \tfrac{1}{18}\int_0^6 (x^3 - 8x^2 + 16x)\,dx$$
$$= \tfrac{1}{18}\left[\tfrac{1}{4}x^4 - \tfrac{8}{3}x^3 + 8x^2\right]_0^6$$
$$= \tfrac{1}{18}(324 - 576 + 288)$$
$$= 2,$$

OR

$$\sigma^2 = \int_0^6 x^2 \times \tfrac{1}{18}x\,dx - 4^2$$
$$= \tfrac{1}{72}\left[x^4\right]_0^6 - 16$$
$$= (18 - 0) - 16$$
$$= 2.$$

As with discrete distributions, the second form is usually easier for calculations.

The *standard deviation* is the square root of the variance, and has the same units as the values, so here $\sigma = \sqrt{2}$ metres.

6 MEAN OR EXPECTED VALUE, AND VARIANCE

Let $f(x)$ be a probability density function on a closed interval $[a, b]$.

- The *mean* or *expected value* $\mu = \mathrm{E}(X)$ is
$$\mathrm{E}(X) = \int_a^b x f(x)\,dx.$$
- The *variance* $\sigma^2 = \mathrm{Var}(X)$ is the expected value of the squared deviation from the mean,
$$\mathrm{Var}(X) = E\left((X-\mu)^2\right) = \int_a^b (x-\mu)^2 f(x)\,dx$$
- Alternatively, and usually easier in calculations, the variance is the expected value of the square, minus the square of the mean,
$$\mathrm{Var}(X) = E(X^2) - \mu^2 = \int_a^b x^2 f(x)\,dx - \mu^2.$$
- The standard deviation σ is the square root of the variance.

Example 9 16C

Find the mean and standard deviation of each PDF.

a $y = \frac{1}{8}$, for $0 \le x \le 8$

b $y = \frac{1}{50}x$, for $0 \le x \le 10$

SOLUTION

a
$$\mu = \int_0^8 \tfrac{1}{8}x\,dx$$
$$= \left[\tfrac{1}{16}x^2\right]_0^8$$
$$= 4 - 0$$
$$= 4,$$

$$\sigma^2 = \int_0^8 \tfrac{1}{8}x^2\,dx - 4^2$$
$$= \left[\tfrac{1}{24}x^3\right]_0^8 - 16$$
$$= \frac{512}{24} - 0 - 16$$
$$= \frac{16}{3},$$
$$\sigma = \tfrac{4}{3}\sqrt{3}.$$

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

b $y = \frac{1}{50}x$, for $0 \leq x \leq 10$

$$\mu = \int_0^{10} \frac{1}{50}x^2\,dx$$
$$= \left[\frac{1}{150}x^3\right]_0^{10}$$
$$= \frac{1000}{150} - 0$$
$$= 6\frac{2}{3},$$

$$\sigma^2 = \int_0^{10} \frac{1}{50}x^3\,dx - \left(6\frac{2}{3}\right)^2$$
$$= \left[\frac{1}{200}x^4\right]_0^{10} - \frac{400}{9}$$
$$= \frac{10000}{200} - 0 - \frac{400}{9}$$
$$= \frac{50}{9},$$
$$\sigma = \frac{5}{3}\sqrt{2}.$$

Exercise 16C

FOUNDATION

1 A function is defined by $f(x) = \frac{1}{10}$, where $0 \leq x \leq 10$.

a Show that $f(x)$ is a valid PDF (probability density function).

b Calculate the expected value using the formula $\mathrm{E}(X) = \int_a^b x f(x)\,dx$.

c Does your answer for the expected value agree with your understanding of expected value as an average value?

d Calculate the variance using the formula $\mathrm{Var}(X) = \int_a^b (x - \mu)^2 f(x)\,dx$, then find the standard deviation σ.

e Use the alternative formula for variance $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$ and confirm that your answer agrees with the previous result.

2 The previous question provides a mathematical model for selecting a random real number in the interval $[0, 10]$.

Use your calculator (or a spreadsheet) to generate a random number between 0 and 10 to as many decimal places as possible. Many calculators return a random number between 0 and 1, and you will need to multiply this answer by 10.

a Generate 20 such numbers, recording them in a table.

b Calculate the mean and standard deviation using your calculator.

c Do your results agree with the theoretical probabilities in the previous question?

d Our model includes the possibility of selecting a 10, but it is virtually certain that 10 will not be returned by the calculator's random number function. Does this affect the validity of our model and your results?

3 Define the function $f(x)$ by $f(x) = \frac{3}{2}x^2$, with domain $[-1, 1]$.

a Confirm that it is a valid PDF.

b Find the expected value $\mu = \mathrm{E}(X)$.

c Find the variance $\mathrm{Var}(X)$ and the standard deviation σ.

d Calculate $\int_{\mu-\sigma}^{\mu+\sigma} f(x)\,dx$ to determine what percentage of the population defined by this distribution lies within one standard deviation of the mean.

ISBN 978-1-108-76630-2

4 Repeat Question 3 for:

a $f(x) = 2x$, with domain $[0, 1]$

b $f(x) = |x|$, with domain $[-1, 1]$

c $f(x) = \frac{3}{64}x^2$, with domain $[0, 4]$ (final answer correct to three decimal places)

DEVELOPMENT

5 Consider the function defined by $f(x) = \dfrac{1}{c}$, for $0 \leq x \leq c$, where $c > 0$.

a Is this function a valid PDF?

b Calculate $\mathrm{E}(X)$. Is your answer as expected?

c Calculate $\mathrm{Var}(X)$.

d Compare your answer with the special case in Question 1.

e Use the results $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$ and $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$ to find the expected value for the translated uniform probability distribution with density function $g(x) = \dfrac{1}{c}$, for $h \leq x \leq h + c$.

f Find the expected value and variance of the uniform probability distribution defined on the interval $h \leq x \leq k$.

6 **a** Show that the function $f(x)$ is a valid PDF,

$$f(x) = \begin{cases} \frac{1}{8}, & \text{for } 0 \leq x < 2, \\ \frac{1}{4}, & \text{for } 2 \leq x \leq 5, \end{cases}$$

b Find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.

7 At the start of this chapter, we claimed that two expressions for the variance of a continuous distribution are equal,

$$\int_a^b (x - \mu)^2 f(x)\,dx = \int_a^b x^2 f(x)\,dx - \mu^2.$$

Prove this identity, starting with the LHS and expanding the integrand.

ENRICHMENT

Note: Except for Question 8, these questions involve improper integrals where the upper limit is ∞. Infinity is not a number, so you cannot substitute ∞. Instead, take the limit of the primitive as $x \to \infty$.

8 In Question 6 of Exercise 16A, we demonstrated that the area under a relative frequency polygon equals the area under a relative frequency histogram, and that both are equal to the total probability 1.

a Confirm that the relative frequency polygon in that question may be written piecewise as:

$$f(x) = \begin{cases} \frac{2}{10}x, & \text{for } 0 \leq x \leq 1, \\ \frac{1}{10}(3x - 1), & \text{for } 1 \leq x \leq 2, \\ \frac{1}{10}(-2x + 9), & \text{for } 2 \leq x \leq 3, \\ \frac{1}{10}(-3x + 12), & \text{for } 3 \leq x \leq 4. \end{cases}$$

b By integration, calculate $\mathrm{E}(X) = \displaystyle\int_0^4 x f(x)\,dx$ for the probability distribution using the PDF $f(x)$ defined above.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

c Compare the answer obtained by calculating the expected value for the discrete distribution using the table of values in Question 6 of Exercise 16A.

d What is your conclusion about the PDF as a generalisation of the frequency polygon?

9 Define $f(x) = \frac{3}{x^4}$, where $x \geq 1$.

a Show that $f(x)$ is a valid PDF.

b Evaluate $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.

c Calculate each probability.

i $P(X \leq 4)$ **ii** $P(X \geq 2)$ **iii** $P(2 \leq X \leq 5)$

d Find the cumulative distribution function $F(x) = P(X \leq x)$.

10 Consider the function $f(x) = e^{-x}$, where $x \geq 0$. In Question 16 of Exercise 16B, we showed that $f(x)$ is a valid PDF.

a Differentiate xe^{-x}, and hence integrate xe^{-x}.

b Evaluate $\mathrm{E}(X) = \int_0^\infty xe^{-x}\,dx$.

c Differentiate $x^2e^{-x} + 2xe^{-x} + 2e^{-x}$, and hence integrate x^2e^{-x}.

d Evaluate $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$ for the distribution with PDF $f(x)$.

11 Consider the function $f(x) = \frac{1}{\ln 2} \times \frac{1}{x(x+1)}$, where $x \geq 1$.

a Show that $f(x) = \frac{1}{\ln 2}\left(\frac{1}{x} - \frac{1}{x+1}\right)$.

b Hence show that $\int f(x)\,dx = \frac{1}{\ln 2} \times \ln\left(\frac{x}{x+1}\right)$, for $x \geq 1$.

c Hence show that $f(x)$ is a valid PDF in the domain $[1, \infty)$.

d Show that $\mathrm{E}(X)$ does not exist for this PDF.

12 a Prove that $\int_{-\infty}^{\infty} \frac{1}{1+x^2}\,dx = \pi$.

b Explain why the function $f(x) = \frac{1}{\pi(1+x^2)}$, defined over the whole real line $(-\infty, \infty)$, is a valid PDF.

c Use symmetry to evaluate the integral $\mathrm{E}(X)$.

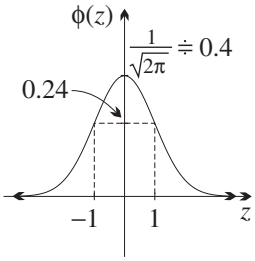
d Use the identity $\frac{x^2}{1+x^2} = 1 - \frac{1}{1+x^2}$ to show that $\mathrm{Var}(X)$ is not defined.

ISBN 978-1-108-76630-2

16D The standard normal distribution

Gauss and the early statisticians realised that one particular group of continuous distributions — the *Gaussian* or *normal distributions* — are particularly important. They occur in a wide variety of situations, and for reasons that we shall explain later, are involved in the study of every distribution, continuous or discrete.

Their graphs are generally referred to as *bell-shaped curves*, and every normal distribution can be obtained from every other normal distribution by shifting and stretching. We saw the shape of such a curve emerging when 20 coins were tossed at the end of Section 16A.



The graph of the *standard normal distribution* is sketched to the right. The equation of its probability density function is

$$\phi(z) = \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}} \text{ (the Greek letter } \phi \text{ is phi, corresponding to Latin f).}$$

It is standard practice to use Z rather than X for the standard normal random variable, and z rather than x for its values, so that the PDF is $\phi(z)$.

Sketching the curve

We already have all the tools for sketching this function from its equation. The first thing is to stop looking at the complicated denominator $\sqrt{2\pi}$, which is just a constant. Use your calculator to find that $\frac{1}{\sqrt{2\pi}} \fallingdotseq 0.4$, and start thinking of the formula as $\phi(z) \fallingdotseq \frac{2}{5}e^{-\frac{1}{2}z^2}$, which looks much more friendly.

- The y-intercept is $\phi(0) = \frac{1}{\sqrt{2\pi}} \fallingdotseq 0.4$, because $e^0 = 1$.
- When z is non-zero, the index $-\frac{1}{2}z^2$ is negative, so $e^{-\frac{1}{2}z^2} < e^0 = 1$. Hence the value at $z = 0$ is a global maximum, and the mode is therefore $z = 0$.
- The function is defined for all values of z, and is positive for all values of z.
- The function is even, with line symmetry in the y-axis, because replacing z by $-z$ leaves the equation unchanged.
- As $z \to \infty$, and as $z \to \infty$, the index $-\frac{1}{2}z^2$ quickly becomes a large negative number, so $e^{-\frac{1}{2}z^2}$ quickly becomes an extremely small positive number. Thus the z-axis is a horizontal asymptote in both directions.
- There are points of inflection at $z = -1$ and $z = 1$ (both have y-coordinate $\frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}} \fallingdotseq 0.24$). We have left the proof of this to the end of this section.

Why is $\phi(z)$ a probability density function?

We have now established that $\phi(z)$ has the graph shown above, but why is it a probability density function? Certainly we can see that it is always positive. But we also need to establish that

$$\int_{-\infty}^{\infty} \phi(z)\, dz = 1.$$

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

Unfortunately, this integral cannot be established using the techniques in this course. This fact is one of a small number of things that readers will have to accept for now, and perhaps prove in later years.

Making the total area under the curve have the value 1 is the reason why the denominator $\sqrt{2\pi}$ has been put there, and proving the result requires a proof that $\int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz = \sqrt{2\pi} \doteq 2.5$. The best that can be done is to confirm this result by trapezoidal rule approximations along the lines of Question 19 in Exercise 16D.

The mean and variance of the standard normal distribution

- The mean of the standard normal distribution is $\mu = 0$.
- Its variance is $\sigma^2 = 1$, and its standard deviation is therefore 1.

The fact that the mean is 0 is clear from the graph, because $y = \phi(z)$ is even, with line symmetry about the y-axis.

Establishing that the variance is 1, however, is a little more difficult because of some fancy integration, and has also been left to the end of this section.

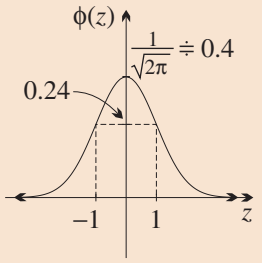
These results for the mean and standard deviation are closely tied to the turning point and inflections, so that μ and σ can be seen clearly on the graph.

- The mean $\mu = 0$ coincides with the maximum turning point at $z = 0$, which is the mode.
- The two inflections at $z = -1$ and $z = 1$ are each one standard deviation from the mean in opposite directions.

7 THE STANDARD NORMAL DISTRIBUTION

Let Z be the *standard normal random variable*.

- The probability density function of Z is $\phi(z) = \dfrac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}}$.
- The graph of the PDF is a *bell-shaped curve*, with global maximum at $z = 0$ (the mode) and points of inflection at $z = 1$ and $z = -1$.
- The mean is $\mu = 0$ and the standard deviation is $\sigma = 1$.
- The points of inflection are each one standard deviation from the mean.



In Section 16E we will be shifting and stretching the standard normal distribution. Whenever you see a curve that looks even vaguely normal, always look first at the turning point, then look at the two inflections and quickly estimate the standard deviation by eye.

Integrating to find probabilities

The probability that a standard normal random variable Z lies within one standard deviation of the mean is

$$P(-1 \leq Z \leq 1) = \int_{-1}^{1} \phi(z)\, dz.$$

Now we have a major inconvenience — the primitive of the function $\phi(z)$ cannot be written in terms of our usual range of exponential, trigonometric and algebraic functions. You have several options, all of which can be found online:

- Use a table of values for this integral — a short version is at the bottom of the next page.
- Use the Desmos graphing calculator or a statistics calculator that has the values of these integrals built in.

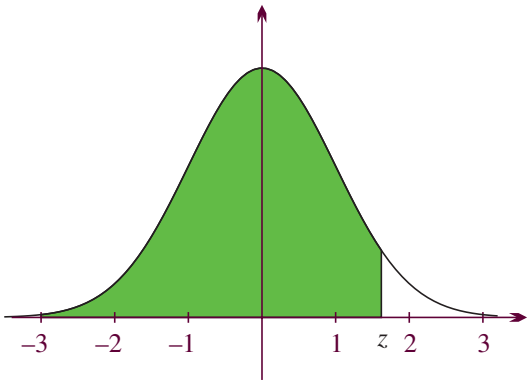
ISBN 978-1-108-76630-2

- Use a spreadsheet that has these integrals amongst its functions.
- Use specialised statistics software.

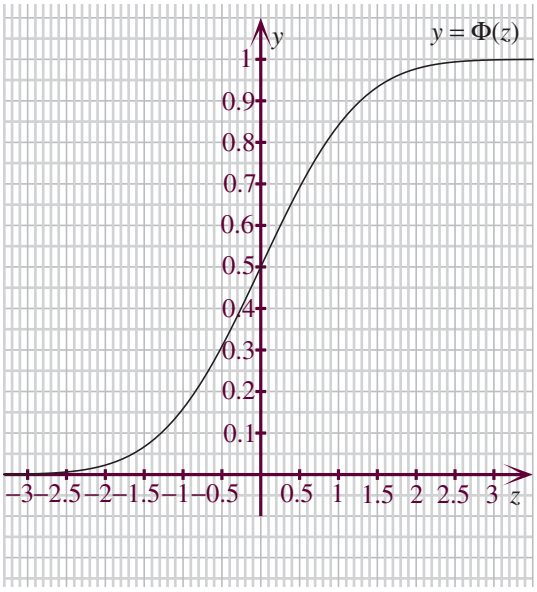
The cumulative distribution function

All these approaches will normally use the cumulative distribution function of the standard normal distribution. This CDF is usually denoted by $\Phi(z)$, using the uppercase version Φ of the Greek letter ϕ.

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



This is the graph of the standard normal probability density function $\phi(z)$.



This is the graph of the standard normal cumulative distribution function $\Phi(z)$.

The new curve $y = \Phi(z)$ has two horizontal asymptotes, $y = 0$ on the left, and $y = 1$ on the right. Because also $\phi(z)$ is even, $\Phi(z)$ has point symmetry in $(0, 0.5)$.

Here then are some further details about finding values of $\Phi(z)$.

- Below is a short table of values of $\Phi(z)$ for $0 \leq z < 4$, in steps of 0.1.
- Statistical calculators should have this function built in.
- In Excel, the function is `NORM.S.DIST`. The function has two arguments.
  - The first argument is the value of z (or the cell containing that value).
  - Set the second argument to true to obtain the value of the CDF $\Phi(z)$, and set it to `false` for the value of the PDF $\phi(z)$.
- See the interactive textbook for using Desmos.

The following short table of values of $\Phi(z)$ will be quite sufficient for most purposes in this chapter. Because of the even symmetry of the PDF $\phi(z)$, there is no need to give values of the CDF $\Phi(z)$ for negative values of z.

	first decimal place									
z	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0.	0.5000	0.5398	0.5793	0.6179	0.6554	0.6915	0.7257	0.7580	0.7881	0.8159
1.	0.8413	0.8643	0.8849	0.9032	0.9192	0.9332	0.9452	0.9554	0.9641	0.9713
2.	0.9772	0.9821	0.9861	0.9893	0.9918	0.9938	0.9953	0.9965	0.9974	0.9981
3.	0.9987	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

A more detailed table giving the values of z to three decimal places is to be found in the Appendix to Chapter 17.

ISBN 978-1-108-76630-2

Calculating probabilities of a normally distributed random variable

Calculating other probabilities for Z requires juggling integrals, preferably while looking at a graph of the PDF $y = \phi(z)$. Always keep two things in mind.

- The total area under the curve $y = \phi(z)$ is 1.
- The curve $y = \phi(z)$ is even — it has line symmetry in the y-axis.

The next worked example demonstrates all the methods required.

Example 10 10D

a Look up $P(Z \leq 2)$ and illustrate it as an area under $y = \phi(z)$.

b Illustrate each probability as an area under $y = \phi(z)$. Then calculate it using the value of $\Phi(2)$ found in part **a**. Keep looking back to the graph in part **a** while you juggle the intervals.

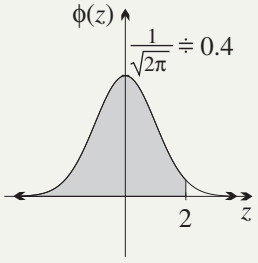
i $P(Z \geq 2)$ **ii** $P(Z \leq -2)$ **iii** $P(0 \leq Z \leq 2)$ **iv** $P(-2 \leq Z \leq 2)$

SOLUTION

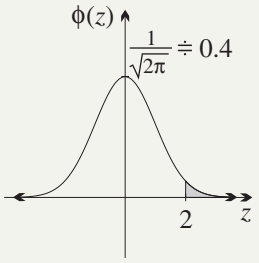
a From the table,

$$P(Z \leq 2) = \Phi(2) = 0.9772.$$

The area under $y = \phi(z)$ corresponding to $\Phi(2)$ is shaded in the diagram to the right.

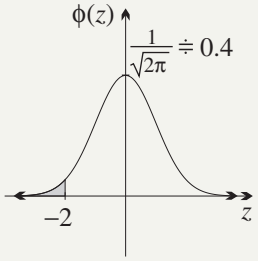


b i



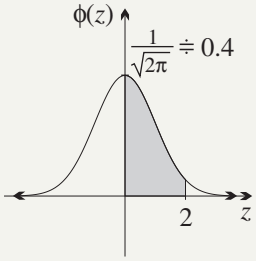
$P(Z \geq 2)$

ii



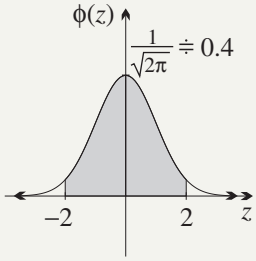
$P(Z \leq -2)$

iii



$P(0 \leq Z \leq 2)$

iv



$P(-2 \leq Z \leq 2)$

i $P(Z \geq 2) = 1 - P(Z \leq 2)$, because the total area is 1,
$\fallingdotseq 1 - 0.9772$
$\fallingdotseq 0.0228.$

The probability that $Z = 2$ exactly is zero,
so there is no need to distinguish between $\leq$ and $<$ or between $\geq$ and $>$.

ii $P(Z \leq -2) = P(Z \geq 2)$, because $\phi(z)$ is even,
$\fallingdotseq 0.0228$, from part **a**.

iii $P(0 \leq Z \leq 2)$
$= \Phi(2) - \Phi(0)$, using subtraction of areas,
$\fallingdotseq 0.9772 - 0.5$, because exactly half the scores are below the mean,
$\fallingdotseq 0.4772.$

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

iv $P(-2 \leq Z \leq 2)$
$= \Phi(2) - \Phi(-2)$
$\doteqdot 0.9772 - 0.0228$, by part **b**,
$\doteqdot 0.9544$,

OR

$P(-2 \leq Z \leq 2)$
$= 2 \times P(0 \leq Z \leq 2)$, by symmetry,
$\doteqdot 2 \times 0.4772$, by part **c**,
$\doteqdot 0.9544$.

Example 11 10D

a Explain how to find $\Phi(0.7)$ from the table, and illustrate it.

b Use symmetry and the table of values of $\Phi(z)$ to find:

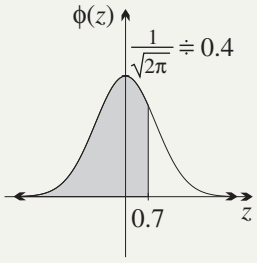
i $P(-2.5 \leq Z \leq -0.3)$ **ii** $P(-2.9 \leq Z \leq 0.6)$

SOLUTION

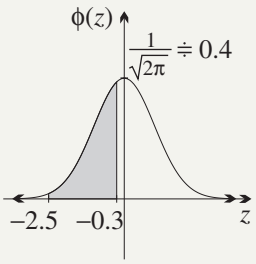
a To find $\Phi(0.7)$ from the table (as illustrated to the right):

- Look at the first row because 0.7 starts with '0.'.
- Then go to the column headed '.7'.

$P(Z \leq 0.7) = \Phi(0.7) \doteqdot 0.7580$.

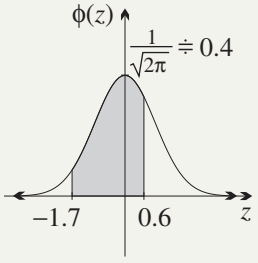


b i



$P(-2.5 \leq Z \leq -0.3)$
$= P(0.3 \leq Z \leq 2.5)$,
because $\phi(z)$ is even,
$= \Phi(2.5) - \Phi(0.3)$
$\doteqdot 0.9938 - 0.6179$
$\doteqdot 0.3759$.

ii



$P(-1.7 \leq Z \leq 0.6)$
$= P(-1.7 \leq Z \leq 0) + P(0 \leq Z \leq 0.6)$
$= P(0 \leq Z \leq 1.7) + P(0 \leq Z \leq 0.6)$
$= \Phi(1.7) - \Phi(0) + \Phi(0.6) - \Phi(0)$
$= \Phi(1.7) + \Phi(0.6) - 1$
$\doteqdot 0.6811$.

The empirical rule or the 68–95–99.7 rule

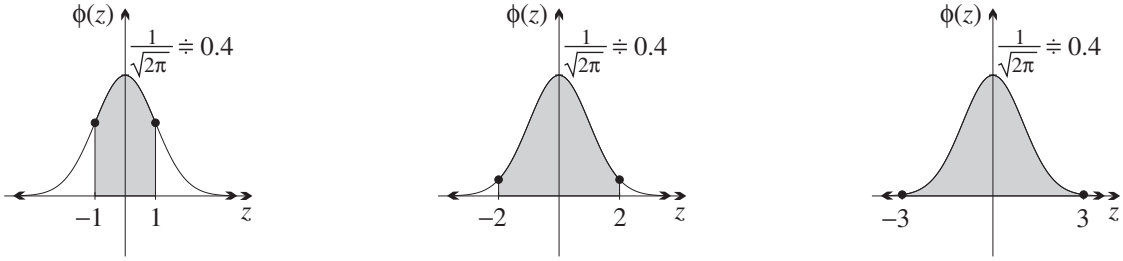
Sometimes statistics requires accurate results, and sometimes it uses very approximate methods. It turns out that in practical use, we constantly need to know the probabilities that a normally distributed variable is within 1, 2 or 3 standard deviations of the mean. That is intuitively straightforward, because

- Z has standard deviation $\sigma = 1$, and
- the two inflections make the region within one standard deviation of the mean stand out on the graph.

ISBN 978-1-108-76630-2

Here are those three results, derived from the table of values of $\Phi(z)$ and converted to the rounded percentages conventionally used in the empirical rule.

$$P(-1 \leq Z \leq 1) \doteq 0.6827 \doteq 68\%,$$
$$P(-2 \leq Z \leq 2) \doteq 0.9545 \doteq 95\%,$$
$$P(-3 \leq Z \leq 3) \doteq 0.9973 \doteq 99.7\%.$$



The percentages here are probabilities, but they are also predictions of what percentages of a normally distributed sample lie within 1, 2 or 3 standard deviations of the mean. These three results are so important that memorising them is part of learning to use the normal distribution, and together they are called the *empirical rule* or the *68–95–99.7 rule*.

8 THE EMPIRICAL RULE OR THE 68–95–99.7 RULE

In a normal distribution, the proportion of scores lying:

- within 1 standard deviation of the mean is 68%,
- within 2 standard deviations of the mean is 95%,
- within 3 standard deviations of the mean is 99.7%.

Example 12 10D

An experiment is run 1000 times. Its random variable is the standard normal variable Z. Answer these questions using the empirical rule only.

a How many scores greater than 2 would you expect?

b Find b if we would expect about 840 scores greater than b.

SOLUTION

a By the empirical rule, we expect $1000 \times 95\% = 950$ scores within $[-2, 2]$.
Because $\phi(z)$ is even, we expect $950 \div 2 = 475$ scores within $[0, 2]$.
Because 500 scores should be positive, we expect 25 scores greater than 2.

b We therefore expect 160 scores less than b, so in particular, b is negative.
Because $\phi(z)$ is even, we also expect 160 scores greater than $-b$.
Hence we expect $1000 - 160 - 160 = 680$ scores between $-b$ and b, so by the empirical rule, $b \doteq -2$.

Quartiles and percentiles

The graph of $y = \Phi(z)$ was drawn above on page 549. Approximation of the quartiles and percentiles can be found from this graph by drawing the appropriate horizontal lines. (We can also use interpolation on the table of values for $\Phi(z)$.)

ISBN 978-1-108-76630-2

Example 13 10D

a Find the 9th decile of the standard normal distribution.

b Find the third quartile Q_3 and the first quartile Q_1 of $\Phi(z)$.

c Using the IQR criterion, what proportion of the scores of a standard normal random variable would be expected to be outliers?

SOLUTION

a For the 9th decile, we need to solve $\Phi(z) \doteqdot 0.9$.
From the table, $\Phi(1.2) \doteqdot 0.8849$ and $\Phi(1.3) \doteqdot 0.9032$, with difference 0.0183 making the 9th decile about 1.28.
This agrees with the horizontal line with height 0.9 on the graph of $\Phi(z)$.

b For the upper quartile, we need to solve $\Phi(z) \doteqdot 0.75$.
From the table, $\Phi(0.6) \doteqdot 0.7257$ and $\Phi(0.7) \doteqdot 0.7580$, with difference 0.0323.
Thus $Q_3 \doteqdot 0.675$, and because $\phi(z)$ is even, $Q_1 \doteqdot -0.675$.
This agrees with the horizontal line with height 0.75 on the graph of $\Phi(z)$.

c From part **b**, the IQR is about 1.35, so $Q_3 + 1.5 \times \text{IQR} \doteqdot 2.70$.
The IQR criterion is that an outlier lies outside $-2.70 \leq z \leq 2.70$.

$$\begin{aligned} \text{Hence } P(Z \text{ is an outlier}) &= P(Z > 2.70) + P(Z < -2.70) \\ &= 2 \times P(Z > 2.70) \\ &= 2\big(1 - \Phi(2.70)\big) \\ &\doteqdot 0.007, \end{aligned}$$

so roughly 7 in 1000 scores would be expected to be outliers.

Note: Correct to five significant figures, $Q_3 \doteqdot 0.67449$. It is standard practice to round percentiles and quartiles of the normal to two decimal places, giving

$$Q_1 \doteqdot -0.67 \quad \text{and} \quad Q_3 \doteqdot 0.67 \quad \text{and} \quad \text{IQR} \doteqdot 1.35,$$

and using the interquartile range criterion, an outlier is a score outside the interval

$$-2.70 \leq z \leq 2.70.$$

In a normal distribution, the IQR criterion makes just under 1% of scores outliers.

The points of inflection

Showing that $y = \phi(z)$ has inflections at $z = -1$ and $z = 1$ requires the second derivative of $\phi(z)$, and is reasonably straightforward.

Differentiation of $y = e^{-\frac{1}{2}z^2}$ requires the chain rule,

$$\begin{aligned} \frac{dy}{dz} &= \frac{dy}{du} \times \frac{du}{dz} \\ &= e^{-\frac{1}{2}z^2} \times (-z) \\ &= -ze^{-\frac{1}{2}z^2}. \end{aligned}$$

Let $u = -\frac{1}{2}z^2$.
Then $y = e^u$.
Hence $\frac{du}{dz} = -z$,
and $\frac{dy}{du} = e^u$.

The function $\phi(z) = \dfrac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}}$ is a multiple of $e^{-\frac{1}{2}z^2}$,
so $\phi'(z) = -z\phi(z)$.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

Hence $\phi(z)$ has a stationary point at $z = 0$, which is a maximum turning point, because $\phi(z)$ is increasing for $z < 0$ and decreasing for $z > 0$.

z	-1	0	1
$\phi'(z)$	$\phi(-1)$	0	$-\phi(1)$
sign	$+$	0	$-$
	/	—	\

Notice in both the tables to the right that $\phi(z)$ is always positive.

For the second derivative, $\phi''(z) = \dfrac{d}{dz}\left(\phi'(z)\right)$

$$= \frac{d}{dz}\left(-z\phi(z)\right),$$

and applying the product rule with $u = -z$ and $v = \phi(z)$,

$$\begin{aligned}\phi''(z) &= -\phi(z) - z\phi'(z)\\ &= -\phi(z) + z^2\phi(z)\\ &= \phi(z)(z^2 - 1).\end{aligned}$$

So there are points of inflection at $z = 1$ and at $z = -1$.

z	-2	-1	0	1	2
$\phi''(z)$	$3\phi(-2)$	0	$-\phi(0)$	0	$3\phi(2)$
sign	$+$	0	$-$	0	$+$
	⌣	·	⌢	·	⌣

The mean and standard deviation

The integrals involved in the calculation of mean and standard deviation require some rather sophisticated techniques. The mean is given by the integral

$$\mathrm{E}(Z) = \int_{-\infty}^{\infty} z\phi(z)\,dz.$$

The integrand $z\phi(z)$ is an odd function, because it is the product of an odd function z and an even function $\phi(z)$. Hence the integral is zero.

This argument assumes that the integral converges. To avoid this assumption,

$$\begin{aligned}\mathrm{E}(Z) &= \int_{-\infty}^{\infty} z\phi(z)\,dz\\ &= \left[-\phi(z)\right]_{-\infty}^{\infty} \quad \text{because we showed above that } \phi'(z) = -z\phi(z)\\ &= 0 - 0 \quad \text{because } \phi(z) \to 0 \text{ as } z \to \infty \text{ and as } z \to -\infty.\end{aligned}$$

Because the mean is zero, the variance is given by the integral

$$\mathrm{Var}(Z) = \int_{-\infty}^{\infty} z^2\phi(z)\,dz.$$

We showed above while finding the second derivative of $\phi(z)$ that

$$\phi''(z) = \phi(z)(z^2 - 1),$$

and rearranging, $z^2\phi(z) = \phi''(z) + \phi(z)$.

Hence

$$\begin{aligned}\mathrm{Var}(Z) &= \int_{-\infty}^{\infty} z^2\phi(z)dz\\ &= \int_{-\infty}^{\infty} \phi''(z)dz + \int_{-\infty}^{\infty} \phi(z)dz.\\ &= \left[\phi'(z)\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \phi(z)dz.\\ &= 0 + 1\\ &= 1.\end{aligned}$$

The first integral above is zero, because the integrand $\phi'(z) = -z\phi(z)$ is odd, as we saw before. The second integral above is 1 because $\phi(z)$ is a probability density function.

ISBN 978-1-108-76630-2

Exercise 16D

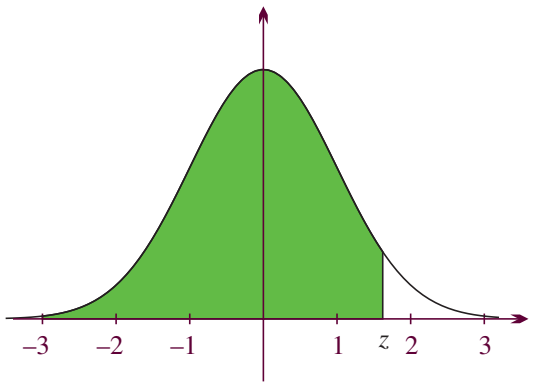
FOUNDATION

The purpose of this exercise is to build familiarity with the symmetry of the standard normal curve. It is not intended that any technology be used for the values of the standard normal distribution in the early questions, because it is important to maximise interaction with the curve and its shape.

The summary below is repeated as an appendix at the end of this chapter.

A brief summary of the standard normal probability distribution

The graph to the right is the *standard normal probability density function* $y = \phi(z)$.



The shaded area represents the value of the corresponding *cumulative distribution function*

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

The table below gives some values of the probabilities $\Phi(z) = P(Z \leq z)$. For example,

$$P(Z \leq 1.6) = \Phi(1.6) = \int_{-\infty}^{1.6} \phi(z)\,dz \doteqdot 0.9452.$$

	first decimal place									
z	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0.	0.5000	0.5398	0.5793	0.6179	0.6554	0.6915	0.7257	0.7580	0.7881	0.8159
1.	0.8413	0.8643	0.8849	0.9032	0.9192	0.9332	0.9452	0.9554	0.9641	0.9713
2.	0.9772	0.9821	0.9861	0.9893	0.9918	0.9938	0.9953	0.9965	0.9974	0.9981
3.	0.9987	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

A more detailed table giving the values of z to three decimal places is to be found in the Appendix to Chapter 17.

For many purposes, all that is required is the *empirical rule*, or *68–95–99.7 rule*,

$P(-1 \leq Z \leq 1) \doteqdot 68\%$
$P(-2 \leq Z \leq 2) \doteqdot 95\%$
$P(-3 \leq Z \leq 3) \doteqdot 99.7\%$

1 Use the table above to look up the following probabilities for the standard normal distribution. Record your answers correct to four decimal places.

a $P(Z \leq 0)$ **b** $P(Z \leq 1)$ **c** $P(Z \leq 2)$ **d** $P(Z \leq 1.5)$
e $P(Z < 0.4)$ **f** $P(Z \leq 2.3)$ **g** $P(Z < 1.2)$ **h** $P(Z \leq 5)$

2 Explain from the graph above why $P(Z > a) = 1 - P(Z \leq a)$. Then use the standard normal table with this complementary result to find:

a $P(Z > 0)$ **b** $P(Z > 1)$ **c** $P(Z > 2)$ **d** $P(Z \geq 2.4)$
e $P(Z > 1.3)$ **f** $P(Z > 0.7)$ **g** $P(Z \geq 1.6)$ **h** $P(Z > 8)$

3 **a** Use the symmetry of the standard normal graph above to explain why if $a > 0$, then $P(Z < -a) = 1 - P(Z \leq a)$. (You need not memorise this result).

b Use this result to find:

i $P(Z < -1.2)$ **ii** $P(Z \leq -2.3)$ **iii** $P(Z < -0.2)$
iv $P(Z < -3.2)$ **v** $P(Z < -5)$ **vi** $P(Z \leq -0.7)$
vii $P(Z < -1.6)$ **viii** $P(Z \leq -1.4)$ **ix** $P(Z < -0)$

ISBN 978-1-108-76630-2

4 **a** Use symmetry to explain why $P(Z \leq 0) = 0.5$.

b Hence use symmetry and the standard normal table to find:

i $P(0 < Z \leq 1.3)$ **ii** $P(0 < Z \leq 2.4)$ **iii** $P(0 < Z \leq 0.7)$

iv $P(-2.4 \leq Z < 0)$ **v** $P(-1.1 \leq Z < 0)$ **vi** $P(-0.7 \leq Z < 0)$

vii $P(0 < Z \leq 1.6)$ **viii** $P(-1.3 \leq Z \leq 0)$ **ix** $P(0 < Z \leq 5)$

c Find:

i $P(-1.3 \leq Z < 1.3)$ **ii** $P(-2.4 < Z \leq 2.4)$ **iii** $P(-0.8 < Z < 0.8)$

iv $P(-2.9 < Z \leq 2.9)$ **v** $P(-0.4 \leq Z < 0.4)$ **vi** $P(-1.5 < Z \leq 1.5)$

5 Match these eight probabilities into four pairs with equal values.

a $P(Z \leq 2)$ **b** $P(Z \leq -1)$ **c** $P(Z \leq 1.2)$ **d** $P(Z = 4)$

e $P(Z < 2)$ **f** $P(Z = 2.3)$ **g** $P(Z \geq 1)$ **h** $P(Z > -1.2)$

6 Repeat the previous question for these eight values:

a $P(Z \leq 5)$ **b** $P(Z > -1.7)$ **c** $P(Z < 5)$ **d** $P(Z \geq 2)$

e $P(Z = 3)$ **f** $P(Z \leq -2)$ **g** $P(Z \leq 1.7)$ **h** $P(Z = 1.2)$

DEVELOPMENT

7 **a** Explain why $P(a \leq Z \leq b) = P(Z \leq b) - P(Z < a)$.

b Use this result to find:

i $P(1.2 \leq Z < 1.5)$ **ii** $P(0.2 \leq Z < 2.3)$ **iii** $P(0.6 \leq Z < 1.7)$

iv $P(-2 \leq Z < -1.2)$ **v** $P(-4 \leq Z < -0.2)$ **vi** $P(-2.7 \leq Z < -1)$

c Similarly find:

i $P(-1.5 \leq Z < 2.2)$ **ii** $P(-0.9 \leq Z < 1.2)$ **iii** $P(-2.9 \leq Z < 1.3)$

8 Use just the two values $\Phi(1.2) = 0.8849$ and $\Phi(1.8) = 0.9641$, and the symmetry of $\phi(z)$, and your knowledge of its properties as a PDF, to find:

a $P(Z \leq 0)$ **b** $P(Z = 4)$ **c** $P(Z > 1.8)$ **d** $P(Z \leq 1.2)$

e $P(Z \geq 1.2)$ **f** $P(0 \leq Z \leq 1.2)$ **g** $P(Z \leq -1.8)$ **h** $P(Z \geq -1.2)$

i $P(1.2 \leq Z \leq 1.8)$ **j** $P(-1.8 \leq Z \leq 1.2)$

9 Use the standard normal table to find:

a $P(Z \leq 1.3)$ **b** $P(Z = 2.4)$ **c** $P(Z > 0.4)$ **d** $P(Z \leq 1.7)$

e $P(Z \geq -1.3)$ **f** $P(0 \leq Z \leq 1.5)$ **g** $P(Z \leq -0.8)$ **h** $P(Z \geq 0.2)$

i $P(1.1 \leq Z \leq 1.5)$ **j** $P(-1.3 \leq Z \leq 2.2)$

10 Use the standard normal table to find these probabilities. Recall from the probability chapter of the Year 11 book that 'and' and 'or' correspond to intersection and union.

a $P(Z \leq 1.2 \text{ or } Z \geq 1.8)$ **b** $P(Z \leq 1.8 \text{ and } Z \geq 1.2)$

c $P(Z \leq 0.2 \text{ or } Z \geq 1.6)$ **d** $P(Z \leq 2.4 \text{ and } Z \geq 1.7)$

11 Repeat any of the previous questions using a calculator or other technology in place of the standard normal tables.

ISBN 978-1-108-76630-2

12 Use the empirical rule (also called the 68–95–99.7 rule) to find:

a $P(Z \leq 0)$ **b** $P(Z \leq 1)$ **c** $P(Z \leq 2)$

d $P(Z < -1)$ **e** $P(0 \leq Z \leq 3)$ **f** $P(0 \leq Z < 1)$

g $P(-2 \leq Z \leq 0)$ **h** $P(-3 < Z \leq -2)$ **i** $P(-1 \leq Z \leq 1)$

j $P(-3 < Z \leq 1)$ **k** $P(-2 \leq Z < 1)$ **l** $P(-2 \leq Z \leq 7)$

13 Use the empirical rule to find the value of b in each case.

a $P(-b \leq Z \leq b) = 0.68$ **b** $P(0 \leq Z \leq b) = 0.475$

c $P(Z \geq b) = 84\%$ **d** $P(-2b \leq Z \leq b) = 0.815$

e $P(-3b \leq Z \leq 3b) = 0.997$ **f** $P(Z^2 \leq b) = 0.95$

14 Use the standard normal table in reverse to find the value of a, given that:

a $P(Z < a) = 0.7257$ **b** $P(Z \leq a) = 0.9893$

c $P(Z < -a) = 0.1151$ **d** $P(Z < a) = 0.2119$

e $P(-a \leq Z < a) = 0.7286$ **f** $P(-a < Z \leq a) = 0.9906$

15 A professional bowler discovers that when he bowls at a central target, his results form a standard normal distribution, where Z is the distance in centimetres from the target to where the ball hits on each bowl.

a Use the empirical rule to find the probability that his result lies:

i within 1 centimetre of the central target,

ii further to the left than 2 centimetres to the right of the target,

iii more than 3 centimetres from the target.

b In how many centimetres either side of the target do 50% of the bowls strike? You will need to use the standard normal table in reverse for this question.

16 Give a mathematical explanation, and also a practical explanation and example, for the result $P(Z = a) = 0$ for any a.

17 Consider the standard normal curve, $\phi(z) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$.

a Test your knowledge of this curve:

i What is the domain?

ii Is it odd, even or neither?

iii Write down the equation of any axis of symmetry.

iv What is the area under the curve and above the horizontal axis?

v What are the z-coordinates of the points of inflection?

vi What are the coordinates of the maximum turning point?

vii What are the z-intercepts?

b Test your knowledge of the associated standard normal distribution:

i What is its mean? **ii** What is its mode?

iii What is its median? **iv** What is its standard deviation?

c Without looking, write down its probability density function.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

18 [Graphing the standard normal distribution]
The purpose of this question is to use our calculus and curve-sketching skills to draw a graph of $y = f(x)$, where

$$f(x) = e^{-\frac{1}{2}x^2},$$

and then use this graph to sketch the standard normal density function $y = \phi(x)$.

a Show that $f(x)$ is an even function.

b Show that $f'(x) = -xe^{-\frac{1}{2}x^2}$ and $f''(x) = (x^2 - 1)e^{-\frac{1}{2}x^2}$.

c Show that there is a unique stationary point. Find its coordinates and determine its nature.

d Show that there are two points of inflection, and that they occur one standard deviation either side of the mean. Find their coordinates.

e Explain what happens to $f(x)$ as $x \to \infty$ and $x \to -\infty$.

f Graph $y = f(x)$.

g Now use stretching to draw the graph of $y = \phi(x)$.

19 **a** **i** Use the trapezoidal rule with five function values (that is, four intervals) to estimate the integral $\int_0^1 \phi(z)dz$.

ii Double this value to estimate the probability that a value will lie within one standard deviation of the mean on the standard normal curve.

iii Why do you know that this will be an underestimate of the true result?

iv Is this in good agreement with the empirical rule and the standard normal table?

b Use the trapezoidal rule with five function values to determine the probability that:

i a value will lie within two standard deviations of the mean,

ii a value will lie within three standard deviations of the mean.

c Use a spreadsheet to increase your number of intervals to say 10, 20, 50, and 100, and observe the convergence.

ENRICHMENT

20 In this question, you may assume the result $\int_{-\infty}^{\infty} \phi(z)\,dz = 1$, where $\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$ is the PDF of the standard normal distribution.

a Write down the integral for $\mathrm{E}(Z)$ and use the symmetry of the integrand to explain why $\mathrm{E}(Z) = 0$.

b Differentiate $ze^{-\frac{1}{2}z^2}$ and hence integrate $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{1}{2}z^2} dz$.

c Evaluate $\mathrm{Var}(Z)$.

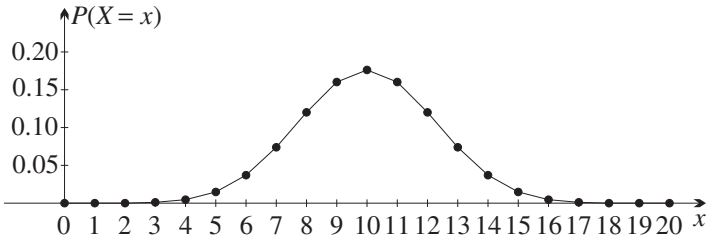
Calculators

Once you are fluent with the diagrams for the standard normal and the associated calculations in this exercise, you should check on your calculator to see whether it has the function $\Phi(z)$. If it does, then as suggested in Question 11, practise until you can use the calculator confidently. The same questions will be quite adequate.

It is possible that the calculator also has the inverse function, denoted by $\Phi^{-1}(z)$ or something similar. This would allow you, for example, to find the value of z for which $\Phi(z) = 0.3$, so that there would be no need to use interpolation.

ISBN 978-1-108-76630-2

16E General normal distributions



In Section 16A we graphed the probabilities of obtaining x heads when 20 coins are thrown, and joined the 21 points to form a polygon. The polygon suggests very much that we should be approximating it with a bell-shaped normal curve. But the curve suggested by the graph is certainly not the standard normal curve, for two reasons:

- The mean is not zero.
- A glance at the inflections shows that the standard deviation is not 1.

This section extends the normal distribution to bell-shaped curves in general.

Shifting and stretching the standard normal distribution

We can estimate the means and the standard deviation from the graph and from the experiment.

- We know that the polygon above is symmetric about $x = 10$, because, for example, the probabilities of obtaining 7 heads from 20 throws, and 13 heads from 20 throws, are equal. Thus the mean is exactly 10.
- We can roughly estimate the standard deviation by looking at where the points of inflection would be if the points were joined up by a curve. The steepest intervals are the interval from $x = 7$ to $x = 8$, and the interval from $x = 12$ to $x = 13$. Let us estimate the points of inflection to be at $x = 7.5$ and $x = 12.5$. That would give a standard deviation of about 2.5.

Some further theory in the Extension 1 course (the *binomial distribution*) tells us that the true standard deviation is $\sigma = \sqrt{5}$, which is approximately 2.236. We now need to stretch and then shift the standard normal distribution to get a curve that may help understand the graph above. That is, we need to produce a normal distribution with $\mu = 10$ and $\sigma = \sqrt{5} \fallingdotseq 2.236$.

For the rest of this chapter, we will drop the approximately equals sign $\fallingdotseq$ because nearly all our numbers are estimates or approximations.

Stretching to accommodate the standard deviation

First, stretch the standard normal distribution horizontally by a factor of σ. This is done by replacing x by $\frac{x}{\sigma}$, as discussed in Section 3H. The standard normal is $y = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, so the result is

$$y = \phi\left(\frac{x}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \quad \left(\text{always look at } \frac{1}{\sqrt{2\pi}} \text{ and think } \tfrac{2}{5} \text{ or } 0.4\right).$$

When this stretching is done, the inflections at $x = 1$ and $x = -1$ become inflections at $x = \sigma$ and $x = -\sigma$. This is because stretching transforms concave-up pieces of curve to concave-up pieces, and concave-down pieces of curve to concave-down pieces.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

This function, however, is not a probability density function, because the stretching has increased the area under the curve by a factor of σ, so that the area is now σ and not 1. To correct this, we have to stretch vertically by a factor of $\frac{1}{\sigma}$, giving what is once again a probability density function,

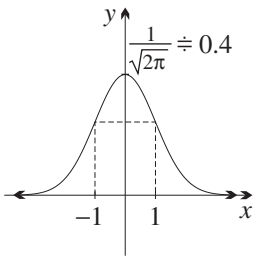
$$y = \frac{1}{\sigma}\phi\left(\frac{x}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{x^2}{2\sigma^2}}.$$

Shifting to accommodate the mean

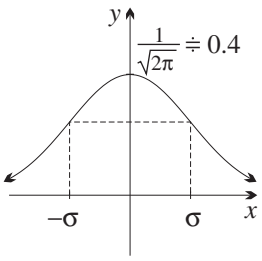
Once the standard deviation has been sorted out, shift the curve μ to the right to make the mean μ instead of 0. This is done by replacing x by $x - \mu$, giving

$$y = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

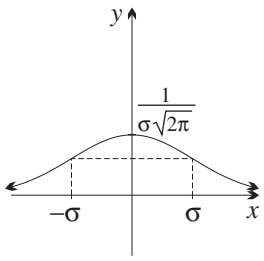
This time the area does not need to be adjusted, and the inflections continue to be one standard deviation from the mean, that is, at $x = \mu - \sigma$ and $x = \mu + \sigma$.



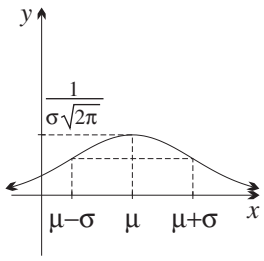
$y = \phi(x)$



$y = \phi\left(\frac{x}{\sigma}\right)$



$y = \frac{1}{\sigma}\phi\left(\frac{x}{\sigma}\right)$



$y = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)$

Summarising the transformation

Let $f(x)$ be the new stretched and shifted probability density function. Taking account of the horizontal and the vertical stretches, and then the horizontal shift, we can write $f(x)$ in terms of $\phi(x)$,

$$f(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right).$$

The last diagram above shows this transformed normal distribution. The continuous probability distribution described by the new function $f(x)$ is called the *normal distribution with mean μ and standard deviation σ*.

These are the important things to notice about the curves above.

- The first, third and fourth graphs are all normal distribution functions. In particular, all have area 1 under the curve.
- The third and fourth graphs both have standard deviation σ — look at the two points of inflection.
- The fourth graph has mean μ — look at the symmetry about $x = \mu$.
- The fourth graph has standard deviation σ — look at the two points of inflection σ to the right of the mean μ, and σ to the left of μ.

The four successive sketches above were drawn using the numerical values $\mu = 3$ and $\sigma = 2$. You can see in the fourth graph that if you take $\mu = 3$, then the inflections are at $x = 3 - 2 = 1$ and at $x = 3 + 2 = 5$.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

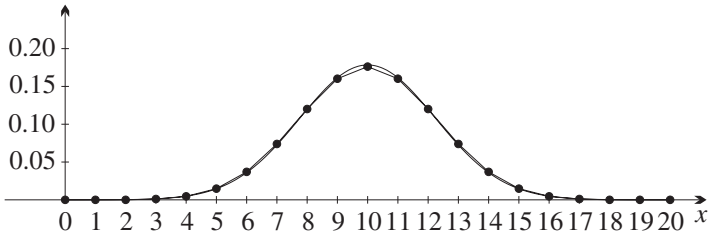
9 THE GENERAL NORMAL DISTRIBUTION

Let $f(x)$ be the probability density function describing a normal distribution with mean μ and standard deviation σ.

- The PDF $f(x)$ is obtained from the standard normal PDF by:
  - stretching horizontally with factor σ and vertically with factor $\frac{1}{\sigma}$,
  - then shifting right by μ units.
- The transformed PDF is therefore $f(x) = \frac{1}{\sigma}\phi\left(\frac{x - \mu}{\sigma}\right)$.
- Thus the transformed PDF has equation $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- The points of inflection of $f(x)$ are each one standard deviation from the mean, that is, at $x = \mu - \sigma$ and at $x = \mu + \sigma$.
- In any normal distribution, the mean, the median and the mode coincide.

Comparison with the 20 coin tosses

Graphed below is the polygon of the 20 coin tosses, together with the normal PDF with $\mu = 10$ and $\sigma = \sqrt{5}$. The vertical scale is the same for both graphs, but there is no name on the vertical axis. This is because for the discrete distribution the name is $P(X = x)$, and for the continuous distribution the name is $f(x)$ or y.



The fit is a very good approximation, but it is not exact. It looks very much as if the fit would get better with more and more coin tosses. The example clearly shows how useful the normal is in approximating complicated probability distributions — in this case a discrete distribution. Historically, this coin-tossing experiment was the first use of the normal to approximate another distribution.

Working with the general normal distribution — *z*-scores

We have seen that every normal distribution is obtained from the standard normal distribution by transformations. In order to work with any normal distribution, we need to convert back to the standard normal. The key to this is *z-scores*.

In any distribution, normal or not, the z-score of a score x is the number of standard deviations above the mean. This is easily calculated by the formula

$$z\text{-score} = \frac{x - \mu}{\sigma}.$$

We need to be able to convert from values of x to z-scores, and back from z-scores to values of x.

The two equations are

$$z = \frac{x - \mu}{\sigma} \quad \text{and} \quad x = \mu + \sigma z.$$

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

For example, check conversions both ways for this table of scores and corresponding z-scores for a distribution with mean $\mu = 10$ and standard deviation $\sigma = 2$.

x	4	5	6	7	8	9	10	11	12	13	14	15	16
z-score	-3	-2.5	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	2.5	3

10 THE z-SCORES OF A RANDOM VARIABLE

Suppose that X is a random variable, normal or not, with mean μ and standard deviation σ.

- The *z-score* of a score x is the number of standard deviations that x lies above the mean. If the z-score is negative, then x lies below the mean.
- Thus the conversions between z-scores and values of x are given by

$$z = \frac{x - \mu}{\sigma} \quad \text{and} \quad x = \mu + \sigma z.$$

- If the distribution is normal, the z-scores allow the values and features of the standard normal distribution to be applied.
- For sample data (not population data), use $\bar{x}$ for the mean and s for the standard deviation.

Example 14 16E

A dataset has mean $\bar{x} = 12$ and standard deviation $s = 3.60$. Answer these questions correct to two decimal places.

a What scores would be 1, 2 and 3 standard deviations from the mean?

b How many standard deviations from the mean are scores of 24, 11 and 7.7?

SOLUTION

a We can do part **a** either using the formula for conversion from z scores to x-values, or working verbally with 'the number of standard deviations from the mean'.

For $z = 1$,
$x = \bar{x} + sz$
$= 12 + 3.60$
$= 15.60$,
and for $z = -1$,
$x = 12 - 3.60$
$= 8.40$.

For $z = 2$,
$x = \bar{x} + 2sz$
$= 12 + 7.20$
$= 19.20$,
and for $z = -1$,
$x = 12 - 7.20$
$= 4.80$.

For $z = 3$,
$x = \bar{x} + 3sz$
$= 12 + 10.80$
$= 22.80$,
and for $z = -1$,
$x = 12 - 10.80$
$= 1.20$.

OR

The scores one SD from $\bar{x}$ are $\bar{x} + s = 12 + 3.60 = 15.60$, and $\bar{x} - s = 12 - 3.60 = 8.40$.

The scores two SDs from $\bar{x}$ are $\bar{x} + 2s = 12 + 7.20 = 19.20$, and $\bar{x} - 2s = 12 - 7.20 = 4.80$.

The scores three SDs from $\bar{x}$ are $\bar{x} + 3s = 12 + 10.80 = 22.80$, and $\bar{x} - 3s = 12 - 10.80 = 1.20$.

ISBN 978-1-108-76630-2

b For $x = 24$, $z = \dfrac{x - \bar{x}}{s}$
$= \dfrac{24 - 12}{3.6}$
$= 3.33,$
3.33 SDs above the mean.

For $x = 11$, $z = \dfrac{x - \bar{x}}{s}$
$= \dfrac{11 - 12}{3.6}$
$= -0.28,$
0.28 SDs below the mean.

For $x = 7.7$, $z = \dfrac{x - \bar{x}}{s}$
$= \dfrac{7.7 - 12}{3.6}$
$= -1.19,$
1.19 SDs below the mean.

Example 15 16E

A normally distributed random variable X has mean 100 and standard deviation 20.

a Write down the two conversion formulae between z-scores and values of x.

b Find: **i** $P(X \leq 110)$ **ii** $P(X \geq 90)$

c Find (nearest whole number) the value of a such that $P(X \leq a) = 0.98$.

SOLUTION

a $z = \dfrac{x - 100}{20}$ and $x = 100 + 20z$.

b i $P(X \leq 110)$
$= P(Z \leq 0.5)$
$= 0.69$

ii $P(X \geq 90) = P(Z \geq -0.5)$
$= P(Z \leq 0.5)$ ($\phi(x)$ is even)
$= 0.69$

c From the table, $\Phi(2.0) = 0.9772$ and $\Phi(2.1) = 0.9821$,
so by interpolation, $\Phi(2.06) = 0.98$.
Converting back to x-values, $a = 100 + 20 \times 2.06 = 141$.

Quartiles, the empirical rule, and the IQR criterion for outliers

If a distribution is normal, we can use z-scores to apply results already calculated about the standard normal distribution. Suppose then that we have a normally distributed random variable X with mean μ and standard deviation σ.

The empirical rule, or 68–95–99.7 rule:

When the experiment is run a large number of times, these are the expectations.

- 68% lie within one standard deviation of the mean,
  — that is, 68% lie within the interval $\mu - \sigma \leq x \leq \mu + \sigma$.
- 95% lie within two standard deviations of the mean,
  — that is, 95% lie within the interval $\mu - 2\sigma \leq x \leq \mu + 2\sigma$.
- 99.7% lie within three standard deviations of the mean,
  — that is, 99.7% lie within the interval $\mu - 3\sigma \leq x \leq \mu + 3\sigma$.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

The first and third quartile:

- We saw in Section 16C that the third quartile of the standard normal is $z = 0.67$.
  This is 0.67 standard deviations above the mean,
  Hence the third quartile of the transformed distribution is $Q_3 = \mu + 0.67\sigma$.
  Alternatively, using the formula, $x = \mu + z\sigma = \mu + 0.67\sigma$.
- We saw in Section 16C that the first quartile of the standard normal is $z = -0.67$.
  This is 0.67 standard deviations below the mean,
  Hence the first quartile of the transformed distribution is $Q_1 = \mu - 0.67\sigma$.
  Alternatively, using the formula, $x = \mu + z\sigma = \mu - 0.67\sigma$.
- The standard normal has interquartile range 1.35,
  so for the transformed distribution, IQR $= 1.35\sigma$.

The IQR criterion for outliers:

- We showed below worked Example 13 that the IQR criterion characterises as outliers any scores lying outside the interval $-2.70 \leq x \leq 2.70$.
  Hence for the transformed distribution, we characterise as outliers any scores lying outside the interval $\mu - 2.70\sigma \leq x \leq \mu + 2.70\sigma$.

Example 16 16E

A dataset with 1000 scores is known to be a sample from a normally distributed variable X with mean $\mu = -32.6$ and standard deviation $\sigma = 5.7$.

a Describe what the empirical rule predicts about the data.

b Using z-scores, and finding $\Phi(x)$ from a table or using technology, predict roughly how many scores will:
i lie in $[-30, \infty)$, **ii** lie in $(-\infty, -40]$, **iii** lie in $[-40, -30]$.

c Using the IQR criterion $\mu - 2.70\sigma \leq x \leq \mu + 2.70\sigma$ for scores that are not outliers, roughly how many outliers would you expect?

SOLUTION

a About 680 scores will lie within one SD from the mean, that is, in $[-38.3, -26.9]$.
About 950 scores will lie within two SDs from the mean, that is, in $[-44.0, -21.2]$.
About 997 scores will lie within three SDs from the mean, that is, in $[-49.7, -15.5]$.

b i For -30, $z = \dfrac{x - \mu}{\sigma}$

$= \dfrac{-30 + 32.6}{5.7}$

$= 0.456$,

so $P(X \geq -30) = P(Z \geq 0.456)$

$= 0.324$,

predicting roughly 324 such scores.

ii For -40, $z = \dfrac{x - \mu}{\sigma}$

$= \dfrac{-40 + 32.6}{5.7}$

$= -1.298$,

so $P(X \leq -40) = P(Z \leq -1.298)$

$= P(Z \geq 1.298)$

$= 1 - P(Z \leq 1.298)$

$= 1 - 0.903$

$= 0.097$,

predicting roughly 97 such scores.

ISBN 978-1-108-76630-2

iii $P(-40 \leq X \leq -30)$
$= 1 - (P(X \leq -40) + P(X \geq -30))$ (no need for z-scores)
$= 1 - (0.324 + 0.097)$ (using parts **i** and **ii**)
$= 0.579$, predicting roughly 579 such scores.

c This does not need to be recalculated. Worked Example 13**c** showed that roughly 7 in 1000 scores are outliers by the IQR criterion for the standard normal, and because the calculation depended only on z-scores, this is valid for any normally distributed variable.

Exercise 16E

FOUNDATION

Note: There is a brief summary of the normal distribution, including a graph, a table and the empirical rule, in the Appendix at the end of this chapter.

1 In each part, calculate the z-scores corresponding to the given value of x, and state how many standard deviations each value of x lies above or below the mean.

a $\mu = 4, \sigma = 1, x = 5$
b $\mu = 13, \sigma = 3, x = 7$
c $\mu = 0.5, \sigma = 0.25, x = 0.75$
d $\mu = 1, \sigma = 3, x = -5$
e $\mu = 114, \sigma = 1.2, x = 120$
f $\mu = 2.35, \sigma = 0.05, x = 2.20$

2 a Use the formula $z = \dfrac{x - \mu}{\sigma}$ to find the z-score when:

i $\mu = 50, \sigma = 4$ and $x = 60$,
ii $\mu = 450, \sigma = 25$ and $x = 375$,
iii $\mu = 3.19, \sigma = 0.12$ and $x = 3.85$,
iv $\mu = 23, \sigma = 8$ and $x = 25$.

b Which of the results in part **a** are:

i furthest from the mean,
ii above the mean,
iii below the mean,
iv within 2 standard deviations from the mean,
v not within the middle 68% of the data?

3 Use z-scores to convert these probability statements for the normal random variable X with mean 4 and standard deviation 2 into probability statements on the standard normal random variable Z. For example, $P(X \leq 7) = P(Z \leq 1.5)$.

a $P(X \leq 5)$
b $P(X > 4.5)$
c $P(X \leq 2)$
d $P(X \geq 1)$
e $P(0 \leq X \leq 3)$
f $P(0.5 \leq X \leq 4.5)$

4 A certain quantity is normally distributed with mean 5 and standard deviation 2. Convert the following probabilities to probabilities involving the standard normal distribution, and then use the empirical rule to find them.

a $P(X \geq 5)$
b $P(3 \leq X \leq 7)$
c $P(X \leq 9)$
d $P(X \geq 1)$
e $P(-1 \leq X \leq 7)$
f $P(1 \leq X \leq 3)$

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

5 Use the empirical rule to find the following probabilities for a normally distributed random variable with the given parameters.

a $P(10 \leq X \leq 18)$, given mean $\mu = 12$ and standard deviation $\sigma = 2$.

b $P(X \geq 42)$, given mean $\mu = 37$ and standard deviation $\sigma = 5$.

c $P(X \geq 4.5)$, given mean $\mu = 4$ and standard deviation $\sigma = 0.25$.

6 Find each probability for a normally distributed random variable X with the given parameters. You will need to use the table of values for the standard normal distribution, or a statistics calculator, or other technology such as a spreadsheet, or online resources.

a $P(3 \leq X \leq 7)$, given mean $\mu = 5$ and standard deviation $\sigma = 0.8$.

b $P(X \geq 20)$, where $\mu = 4$ and $\sigma = 10$.

c $P(X \leq 8)$, where $\mu = 12$ and $\sigma = 5$.

d $P(X \geq -39)$, where $\mu = 0$ and $\sigma = 30$.

e $P(X < 36)$, where $\mu = 20$ and $\sigma = 10$.

f $P(3 < X \leq 5)$, where $\mu = 8$ and $\sigma = 2$.

7 Explain what it means for a score x if the corresponding z-score is:

a positive, **b** negative, **c** zero.

DEVELOPMENT

8 A distribution is known to be normal with mean $\mu = 73$ and $\sigma = 8$. A researcher records the following data values from this distribution:

69, 80, 95, 50, 43, 90, 52, 98, 45

a Write down the data values that lie within one standard deviation of the mean.

b Write down the data values that lie within three standard deviations of the mean.

c Write down the data values that lie more than two standard deviations below the mean.

d Write down the data values that are more than two and a half standard deviations above the mean.

e The researcher believes that these data values were obtained randomly. Do they seem to fit the expected distribution for a normal random variable? Construct a stem-and-leaf plot for the data and comment on the shape of the data.

9 The results of an English examination and a mathematics examination are approximately normally distributed with these parameters:

English:	$\mu = 65\%$	$\sigma = 10$
Mathematics:	$\mu = 62\%$	$\sigma = 15$

a For each student below, determine the z-scores for the two results and state which is more impressive:

i Student A's result in English (90%), or their result in mathematics (92%),

ii Student B's result in English (57%), or their result in mathematics (53%),

iii Student C's result in English (80%), or their result in mathematics (77%).

b What is the probability that a mathematics student obtains over 95%?

c What is the probability that a student's English mark is greater than the mean of the mathematics marks?

ISBN 978-1-108-76630-2

10 The results of an experiment are known to be normal, with mean 50 and standard deviation 10. The experiment is run 600 times.

a Describe what the empirical rule predicts about the data.

b Using z-scores, and using a table of the standard normal, or a calculator, or statistics software, predict roughly how many scores will:

i lie in $[-\infty, 55]$, **ii** lie in $[35, 50]$, **iii** lie in $[38, 62]$.

c Using the IQR criterion $\mu - 2.70\sigma \leq x \leq \mu + 2.70\sigma$ for scores that are not outliers, roughly how many outliers would you expect?

ENRICHMENT

11 At a certain school, Biology has four assessments. The mean and standard deviation for these assessments are recorded in the table below.

Assessment	Mean	SD
1	60	10
2	65	8
3	75	4
4	63	12

a Jack obtained 50, 53 and 67 for the first three assessments, but was absent for the fourth assessment due to a fall.

i Find the z-score for each of Jack's results.

ii Use these z-scores to find Jack's average deviation from the mean.

iii Hence estimate a mark for Jack in the fourth assessment.

iv What are the advantages of this method over simply giving Jack the average for his scores in the first three assessments?

v Are there any disadvantages to the method?

b Jill obtains 64, 70 and 79 for the first three assessments, but due to a tumble could not attend the final assessment. Use the same method to estimate Jill's missing result.

Calculators

Once you are confident with z-scores and their use with probability calculations, you may want to check whether your calculator can handle z-scores, and if so, practise until you can do things quickly. Be careful, however, because automating these transformations gets in the way of understanding 'the number of standard deviations from the mean'.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

16F Applications of the normal distribution

A great number of common situations follow a normal distribution, or follow it approximately enough for practical purposes. The questions in Exercise 16F are self-explanatory, given the previous theory, and one worked example should be sufficient introduction.

Example 17 16F

The Happytime Chocolate Company manufactures a 100 g chocolate–nougat bar. As with any manufacturing process, these chocolate bars do not all have precisely the same weight, and these bars are known to be normally distributed with standard deviation 2 g. To reduce the number of complaints, the company has adjusted its machinery so that the mean is 102 g (such an adjustment does not affect the standard deviation).

a Using the empirical rule where possible, and tables or technology otherwise, find the percentage of chocolate bars:
- **i** of weight less than the stated weight of 100 g,
- **ii** of weight greater than 105 g.

b What would the mean weight need to be for there to be less than 1 chocolate bar in 1000 under 100 g?

SOLUTION

a i A weight of 100 g is 1 standard deviation below the mean of 102 g.

By the empirical rule. $P(-1 \leq Z \leq 1) = 68\%$,

and using the complement, $P(Z < -1 \text{ or } Z > 1) = 32\%$,

so by the even symmetry, $P(Z < -1) = 32\% \div 2$,

$= 16\%$.

ii A weight of 105 g is 1.5 standard deviations above the mean of 102 g.

From the table. $P(Z \leq 1.5) = 93\%$,

so using complements, $P(Z > 1.5) = 7\%$.

b Reading the table backwards, $P(Z \leq 3.1) = 0.999$,

so by symmetry and complements, $P(Z \leq -3.1) = 0.001$.

Hence we need to make the mean 3.1σ above 100 g,

meaning that we set the controls so that $\mu = 100 + 3.1 \times 2 = 106.2$ g.

Exercise 16F FOUNDATION

The first four questions of this exercise should be completed using the empirical rule (or the 68–95–97.7 rule) rather than the standard normal probability table or technology.

There is a brief summary of the normal distribution, including a graph, a table and the empirical rule, in the Appendix at the end of this chapter.

1 The results of a school's English examination are found to be normally distributed with mean 70 and standard deviation 10.
- **a** What percentage of the pupils score over 50?
- **b** What percentage of the pupils score under 80?

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

2 The results in an examination are approximately normally distributed with mean 68 and standard deviation 9. In a cohort of 2000, how many students will be expected to score:

a more than 95, **b** less than 50, **c** between 59 and 86?

3 A machine produces screws that are an average of 2 cm long, with a standard deviation of 0.1 cm. The screw lengths are approximately normally distributed.

a What is the probability that a screw will be undersized, if this is taken to mean more than 2 standard deviations below the mean?

b In a batch of 2400, use z-scores to find how many screws are longer than 2.3 cm?

4 Apples of a certain variety are to be sold in packages in a supermarket. Their diameters are normally distributed with mean 68 mm and standard deviation 2 mm. Apples are discarded if their diameter is more than 72 mm or less than 64 mm. What percentage are discarded?

5 The IQ (*Intelligence Quotient*) test is designed to give a qualitative measure of a person's intelligence. In Australia, IQ is approximately normally distributed with mean 98 and standard deviation 15.

a According to one definition, a genius is defined to be someone with an IQ over 140. What percentage of the Australian population would this be?

b In a population of 25 million, how many geniuses would you expect?

DEVELOPMENT

6 A very famous and early experiment into cholesterol levels, called the Framingham study, found that the average cholesterol level in the population of adult males who did not go on to develop heart disease was 219 mg/mL, with standard deviation 41 mg/mL. Assuming that doctors call a reading of above 240 mg/mL *high*, what percentage of this population could be said to have high cholesterol?

7 In Australian adult males, height is found to be normally distributed with mean 176 cm and standard deviation 7.5 cm. A doorway is designed so that 90% of this population can enter without ducking.

a Read the supplied standard normal distribution table backwards to find the z-score such that $P(Z < z) = 90\%$, correct to 2 decimal places.

b Hence find the minimum height of the doorway.

c In the Dinaric alps, the mean and standard deviation of the heights of adult males are respectively 185 and 7.5 centimetres. A customer orders a special design for the doorway so that 95% of adult males can enter without ducking.

i Explain why a reasonable estimate from the table such that $P(Z < z) = 0.95$ is 1.65.

ii Find the minimum design height of the door.

8 A company has a machine designed to fill cereal boxes. It dispenses cereal according to a normal distribution with mean 500 g and standard deviation 2 g. To ensure that boxes are above the advertised weight at least 95% of the time, what weight should be recorded as the weight on each box?

9 The length of gestation (pregnancy) in human females is approximately normally distributed with mean 266 days and standard deviation 16 days.

a Nine months is about $0.75 \times 365 \div 274$ days. What percentage of females give birth before 274 days?

b If 266 days is considered 'on time', what percentage of females give birth more than:

i 1 week early, **ii** one week late?

ISBN 978-1-108-76630-2

10 A certain study indicates that the pulse rate of an adult male aged 20–39 is about 71 with standard deviation about 9. The data are approximately normally distributed.

- **a** What percentage of this population would be expected to have *bradycardia*, which is defined to be a slow pulse rate below 60 beats/minute?
- **b** *Tachycardia* is defined to be a pulse rate greater than 100 beats/minute. What percentage of the population might be expected to fall in this category?
- **c** Repeat part **a**–**b** for females aged 20–39, whose mean is about 76 and standard deviation is about 9.5.

ENRICHMENT

11 The apples in Question 4 must also fit within regulation weight guidelines. Suppose that the weights are normally distributed, with 97.7% of the apples weighing more than 100 g and 69.1% weighing less than 115 g. Find the mean and standard deviation of the weights of the apples. Use the supplied normal distribution table.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

16G Investigations using the normal distribution

These questions are intended to be investigations using technology — statistical calculators, spreadsheets, statistics software, or online resources. Many of the investigations can be broadened or extended into projects. The exercise is long, and it is certainly not intended that all questions be attempted.

Many of the investigations use *sampling of the mean*. This concept is the reason why the normal distribution plays such a central role in all statistics. The underlying theorem is the *central limit theorem*.

The following Challenge paragraphs explain a particular case of the theorem very briefly. It would perhaps be better read after the idea has been encountered in one or more investigations.

Sampling of the mean

Suppose that we have a random variable X. The distribution may be discrete, or continuous and normal, or continuous and not normal. Suppose that this distribution has mean μ and standard deviation σ, neither of which we know.

We want to find the mean of this distribution, so we do the obvious thing — we *sample* the variable X. That is, we run n independent trials of the experiment and take the average of these results as our estimate for the mean μ. What this procedure has actually done is generate a new random variable Y. The procedure described is:

- Take n independent samples of the random variable X, thus generating n values of X.
- Find the mean of the n samples, and assign this mean to a random variable Y.

The central limit theorem says that in most situations, this new random variable Y:

- has the same mean μ as X (which is obvious),
- has variance $\frac{\sigma^2}{n}$, that is, its standard deviation is $\frac{\sigma}{\sqrt{n}}$ (nearly obvious),
- *tends towards a normal distribution as the number n of samples increases.*

The significance of this theorem is that sampling any random variable to find its mean generates approximately a normal distribution as more and more samples are taken. Thus the normal distribution is involved in the study of every distribution, continuous or discrete.

The normal approximation of the 'toss 20 coins' polygon in Section 16E is historically one of the first examples of the theorem. It will be studied in some detail in Sections 17C–17D of the final chapter on binomial distributions.

Exercise 16G

INVESTIGATION

Some of the questions below involve the use of a spreadsheet such as Excel, LibreOffice Calc or GoogleDocs. The instructions below are directly relevant for a recent version of Excel on Windows, but may be adapted depending on available software. More serious investigations could use a general programming language such as Python, or a statistical programming language such as R.

1 [Sampling of the mean] The data from many common experiments, with any distribution, can be displayed as normally distributed data using the following technique called *sampling from the mean*. This is the reason why the normal distribution is so important.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

A student generated three real-valued random numbers between 0 and 15. The mean of these three numbers was recorded and the original three numbers discarded. This was repeated 1000 times, and the results were recorded as grouped data in the table below.

class	0–1	1–2	2–3	3–4	4–5	5–6	6–7	7–8
class centre x	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5
frequency f	2	5	24	46	88	116	138	158
relative frequency f_r								

class	8–9	9–10	10–11	11–12	12–13	13–14	14–15
class centre x	8.5	9.5	10.5	11.5	12.5	13.5	14.5
frequency f	144	113	78	47	27	10	4
relative frequency f_r							

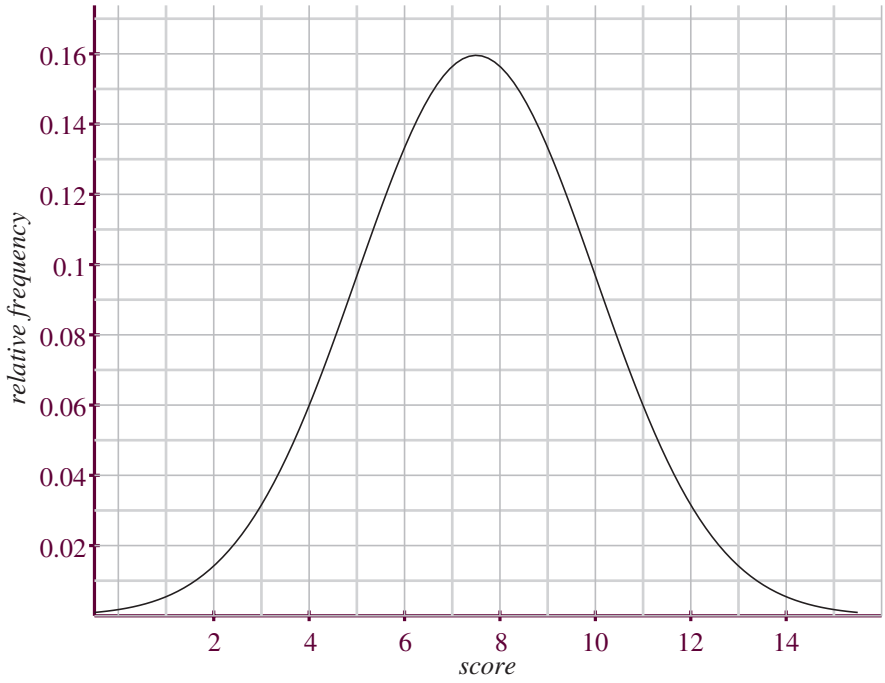
a Use your calculator or a spreadsheet to evaluate the mean and standard deviation of this data, using the given class centres and frequencies. Round your answers correct to 1 decimal place.

b Complete the table by filling in the relative frequencies. Now take the class centres x as the values, and take the relative frequencies as estimates of the probabilities $P(x)$, and evaluate the mean and standard deviation using the formulae

$$\mathrm{E}(X) = \sum xP(x), \quad \text{and}$$
$$\mathrm{Var}(X) = E(X^2) - \mathrm{E}(X)^2, \quad \text{where } \mathrm{E}(X^2) = \sum x^2P(x).$$

Check that your answers agree with the previous results.

c



In the graph above, the normal probability density curve with the mean and standard deviation you just found has been plotted. Photocopy the graph and construct a relative frequency histogram for your grouped data on the same diagram. Add the relative frequency polygon by joining the top centres of the histogram rectangles. What do you notice?

ISBN 978-1-108-76630-2

d How do you think that you might improve the match between the normal curve and the relative frequency polygon?

e Using integration, calculate the mean $\mathrm{E}(Y)$ and variance $\mathrm{Var}(Y)$ for a uniform continuous random variable with $P(Y) = \frac{1}{15}$ on the interval $0 \leq Y \leq 15$. Confirm that the mean and variance obtained in part **a** above are related by the formulae:

$$\mathrm{E}(X) = \mathrm{E}(Y) \qquad \text{and} \qquad \mathrm{Var}(X) = \frac{\mathrm{Var}(Y)}{n},$$

where $n = 3$ (because three random numbers were averaged at each stage).

2 In this question we will replicate the experiment discussed in the previous question using a spreadsheet. In this experiment we will generate 3 real-valued random numbers between 0 and 12 and calculate their mean. This step will be carried out 100 times, giving a set of 100 means that form a distribution called *sampling of the mean*.

a A fragment of a spreadsheet is shown below. In each cell `A2:C2` we have entered `=RAND()*12`. In cell `E2` we have entered a formula to calculate the mean of the three numbers. Enter all this in your spreadsheet.

	A	B	C	D	E	F
1					Average	
2	1.234	4.578	6.914		`=AVERAGE(A2:C2)`	

b Fill down row 2 one hundred times, finishing on row 101.

c Type the formula `=AVERAGE(E2:E101)` in cell `I2` and `=STDEV.P(E2:E101)` in cell `J2` to calculate the mean and standard deviation of this sample of means.

d We will now group the data in intervals of 1 unit. In cells `K3:V3` enter the starting value of each class, that is 0, 1, . . . , 11. Also record a final 12 in cell `W3` to record the end of the data. Cells `K4:V4` will calculate the class centre and cells `K5:V5` will calculate the frequency for each class. Finally cells `K6:V6` will divide the frequency by 100 (the number of times we ran the experiment) to calculate the relative frequency. The formulae are shown in the spreadsheet fragment below for cells `K4:K6`. You should fill the formulae from cells `K4:K6` across to cells `V4:V6`.

	⋯	I	J	K	L	M	N	O
1	⋯	Mean	SD					
2	⋯	5.901	1.897					
3	⋯			0	1	2	3	4
4	⋯			`=(L3+K3)/2`				
5	⋯			`=countif($E:$E, "<"&L3)−countif($E:$E,"<"&K3)`				
6	⋯			`=K5/100`				

(When using the code `$E:$E`, make sure that there are no other entries in column `E`.)

e Construct a histogram using the data from cells `K6:V6`. If your program allows, you should label the horizontal axis using your class centres and ensure that there are no gaps between the rectangles (Excel: Click on bars and select `FORMAT DATA SERIES`), because there should not be gaps between the rectangles of a histogram.

ISBN 978-1-108-76630-2

f If your data do not generate a distribution that looks normal, you might like to recalculate with a fresh set of random numbers. In Excel this option is available under the 'Formulas' tab, option `Calculate Now` (shortcut `F9`).

g If random numbers are generated from the interval $0 \leq x \leq c$, then theory claims that

$$\mu \doteqdot \frac{c}{2} \qquad \text{and} \qquad \sigma^2 \doteqdot \frac{c^2}{12n}.$$

Test these two formulae agree with your results obtained in part **c**.

h If your distribution is normal, then approximately 68% of the data should be within 1 standard deviation of the mean. Theory predicts that the mean was 6 and the standard deviation was 2. Check that:

i approximately 68 of the 100 numbers fell in the interval $[4, 8]$,

ii approximately 95 of the 100 numbers fell in the interval $[2, 10]$.

i A more correct experimental approach to improve the normality of the distribution would be to run the experiment with more trials. Adapt the spreadsheet for 1000 trials. Remember to adjust cells `K6:V6` for the new experiment. Test your improved experiment by repeating part **h**.

3 In the experiment in Questions 1 and 2 we took means of a continuous distribution, but you can also take the mean of discrete data and approximate it by a normal distribution.

a In Question 2, replace cells `A2:C101` by `RANDBETWEEN(1,6)+RANDBETWEEN(1,6)`, simulating the result of throwing two dice and recording their sum. Note that this is not a uniform distribution.

b This could also be done as a practical experiment using a pair of dice.

4 [Normal approximation to the binomial] If we throw 10 coins, what is the probability of obtaining exactly 5 heads? This is called a *binomial probability*, because each coin produces one of only two possible outcomes, 0 and 1 (the prefix 'bi-' means 'two'). It is another discrete probability distribution that may be modelled by a continuous normal probability distribution.

a Throw 10 coins and record the number of heads (if you are short of cash, you could throw 10 dice and record the number of dice showing an even number).

b Repeat this 100 times, recording your results in a frequency table as you go:

Heads	0	1	2	3	4	5	6	7	8	9	10
Tally											
Freq											

c Calculate the mean and standard deviation of your distribution.

d Draw a histogram of your experiment and add the frequency polygon. Does it look normal?

e Assuming a mean of 5 and a standard deviation of about 1.5, one standard deviation either side of the mean should represents an outcome of 4, 5 or 6 heads. (Why?) Do you find 68% of the numbers fall within one standard deviation of the mean?

f To improve your results, repeat the experiment more times. If this is done as a class exercise in groups, groups could collate their data into one frequency table.

g [Technology] A spreadsheet is a good tool to record your data and construct a histogram.

ISBN 978-1-108-76630-2

h [Technology] You may wish to try simulating the whole experiment in a spreadsheet. Here is a fragment of a spreadsheet to generate 10 coin flips and record the number of heads. The formula in cell `A1` is duplicated in cells `B1:J1` and returns a 1 for a head and a zero for a tail. The formula in cell `K1` counts the number of heads.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	`=randbetween(0,1)`				1	0	0	1	1	0	`=sum(A1:J1)`		

5 [For further investigation of Question 4]

a Theory indicates that if I count the number of heads on n throws and if p is the probability of a head, then

$$\bar{x} = \frac{n}{2} \qquad \text{and} \qquad s^2 = np(1 - p).$$

Check this for the experiment above.

b Try varying n.

c Try varying p — throw 15 dice and count the number of dice showing more than 4 $\left(\text{so that } p = \frac{1}{3}\right)$. Or, on your spreadsheet use the code `=if(randbetween(1,6)>4,1,0)`.

d Statisticians have a rule of thumb — for a fairly good approximation to a normal distribution, np and $n(1 - p)$ should be at least 5. Test this out.

6 [For further investigation] In Question 1 and 2 we generated a new set of data by averaging the means of three random numbers. If real numbers are generated from the interval $[0, c]$, the *central limit theorem* predicts that the mean and standard deviation of these data will be

$$\mu \doteqdot \frac{c}{2} \qquad \text{and} \qquad \sigma^2 \doteqdot \frac{c^2}{12n},$$

and that the approximation to normal should be increasingly good as $n \to \infty$. Investigate what happens as you use larger and larger values of n. Does your distribution look increasingly normal?

Note that if you wish to use an interval length not equal to 1 when grouping your data, then you will need to graph relative frequency per unit of width on the vertical axis.

7 It is reasonable to suppose that height follows a bell-shaped distribution, because most of a fairly homogeneous population cluster around the mean height and rapidly tails off further from the mean.

Collect the height of students in your year group. This could be done in classes and results shared, or results could be entered in an online survey.

a Using the techniques of this section, group the results, then graph the histogram and frequency polygon. You will need to choose your interval width so that enough students lie in the central classes to generate a good histogram.

b Does the curve look normal?

c Calculate the mean and standard deviation.

d Assuming that the results are approximately normal, test whether the expected number of students lie within one and two standard deviations of the mean.

e Can you improve the normality of your results by restricting your population to a certain age group or ethnic group?

ISBN 978-1-108-76630-2

8 Using DESMOS, or other graph sketching program, we can investigate the normal distribution curve with mean μ and standard deviation σ,

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Your curve sketching program may recognise entering *pi* for 3.14159265. . . and *e* for 2.718281828. . ., but many programs will not recognise *sigma* or *mu* and you will need to replace them with *s* and *m* (or other pronumerals):

$$y = \frac{1}{s\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2s^2}}.$$

a Create sliders for m and s, if possible restricting $-6 \leq s \leq 6$ and $-3 \leq m \leq 3$.

b Adjust the vertical scale (say $-0.1 \leq y \leq 1.1$) so that the graph with its 'bell shape' is clearly displayed.

c Set $s = 1$ and verify that adjusting m shifts the central mean and median of the symmetric curve.

d With $s = 1$ and $m = 0$, determine the highest point of the curve. Using the equation above, what are the exact coordinates of this maximum point on the curve?

e What is the effect of adjusting s to the 'fatness' and height of the curve?

f What are the heights of the curve when $s = 1$, $s = 2$ and $s = 4$? Comment.

g In DESMOS, define the function

$$f(x) = \frac{1}{s\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2s^2}} \qquad \text{and then enter} \qquad \int_{-s}^{s} f(x)\,dx$$

Check the claim that 68% of the data lie within one standard deviation of the mean.

h Adjust the limits of the integral and check that 95.4% of the data lie within two standard deviations of the mean, and 99.7% of the data lie within three standard deviations of the mean.

i Graph-sketching programs are not always good at handling ∞. Use the table for the standard normal curve and check the value obtained for $P(X \leq 1)$ obtained by DESMOS with

$$\int_{-3}^{1} f(x)\,dx.$$

How small a value is needed for the lower limit to get a value accurate to 9 decimal places, assuming that the answer should be $P(X \leq 1) \doteqdot 0.841344746$?

9 [Investigation] The *Galton board* is a machine designed to generate a bell-shaped curve by means of a set of steel balls falling through a triangular array of pegs. It is possible to buy Galton boards, or there are animations of this device on the web. The original Galton board was designed by Sir Francis Galton (1822–1911).

a Test this device or view animations of the device in action. How good a normal approximation does it produce?

b How does it work?

ISBN 978-1-108-76630-2

Chapter 16 Review

Review activity

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

Chapter 16 Multiple-choice quiz

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

Review

Chapter review exercise

Note: There is a brief summary of the normal distribution, including a graph, a table and the empirical rule, in the Appendix at the end of this chapter.

1 A simple experiment measures the length of time in hours that a certain drug is retained in a patient's system. The following preliminary data were recorded:

0.9	1.4	2.1	2.3	2.6	2.2	2.4	2.7	3.6	3.7
4.1	4.3	4.4	4.4	4.7	5.1	5.2	6.1	6.3	7.1

a Complete the following table for these data.

x	0–1	1–2	2–3	3–4	4–5	5–6	6–7	7–8	Sum
cc	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	—
Tally									
f									—
cf									—
f_r									—
cf_r									

b Draw a relative frequency histogram and polygon for the dataset.

c Draw a cumulative relative frequency histogram and polygon (ogive) for the dataset.

d By adding appropriate horizontal lines to your graph, find:

i the median Q_2,

ii the quartiles Q_1 and Q_3,

iii the ninth decile,

iv the eighty-fifth percentile.

e The dataset appears (almost) bimodal, with many data points falling in two specific intervals. Advise the medical researcher how to proceed next.

2 State whether each of these sentences is true or false.

a The ogive is joined to the top centre of each rectangle of the cumulative frequency histogram.

b The area under the frequency polygon is 1.

c The probability density function of a continuous probability distribution is the analogue of the relative frequency polygon of a discrete distribution.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

d A probability density function $f(x)$ defined on the interval $a \leq x \leq b$ satisfies the two conditions $f(x) \geq 0$, for all x in the interval, and $\int_a^b f(x)\,dx = 1$.

e Every normal distribution is related to the standard normal distribution by stretches and a horizontal shift.

f Approximately 99% of all data lie within three standard deviations of the mean.

3 Let $f(x) = \frac{1}{20}$, where $-10 \leq x \leq 10$.

a Show that $f(x)$ is a probability density function.

b What special name is given to this type of distribution, where the density function takes the same value across its domain?

c Calculate its expected value.

d Calculate its variance and standard deviation.

4 Let $f(x) = \frac{3}{16}(4 - x^2)$, $0 \leq x \leq 2$.

a Show that $f(x)$ is a probability density function (PDF).

b Find its cumulative density function (CDF).

c The CDF is graphed to the right. Use this graph to estimate:

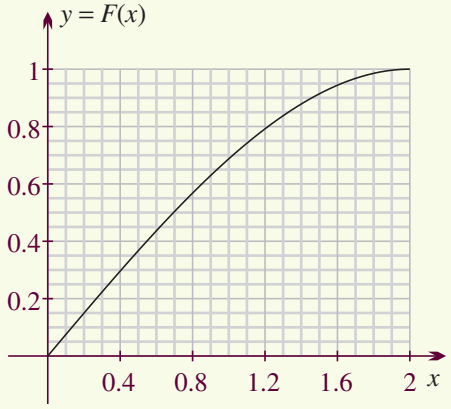
i the three quartiles Q_1, Q_2 and Q_3,

ii the sixth decile,

iii $P(X \leq 1.2)$,

iv $P(X \geq 0.3)$,

v $P(0.2 \leq X \leq 0.4)$.



5 Use your standard normal table and a knowledge of the symmetry of the curve to find:

a $P(Z < 0)$ **b** $P(Z < 1.3)$ **c** $P(-1.8 < Z < 1.8)$

d $P(Z > 0.5)$ **e** $P(Z < -0.2)$ **f** $P(-0.1 < Z < 1.2)$

6 Find the given probability for the normal distribution with given mean and standard deviation. Use the empirical rule (the 68–95–99.7 rule) to estimate:

a $P(X \leq 16)$ if $\mu = 10, \sigma = 3$ **b** $P(X \geq 3.5)$ if $\mu = 5, \sigma = 1.5$

c $P(1.85 \leq X \leq 2.3)$ if $\mu = 2, \sigma = 0.15$ **d** $P(13.65 \leq X \leq 14.1)$ if $\mu = 15, \sigma = 0.45$

7 Repeat the previous question, but this time use your standard normal distribution tables.

a $P(X \leq 22.5)$ if $\mu = 20, \sigma = 5$ **b** $P(X \geq 62)$ if $\mu = 50, \sigma = 10$

c $P(3.96 \leq X \leq 4.3)$ if $\mu = 4, \sigma = 0.2$ **d** $P(6.79 \leq X \leq 8.09)$ if $\mu = 5.75, \sigma = 1.3$

8 A washing machine manufacturer has tested the design of its machines and found them to have an expected life of 6 years 4 months with a standard deviation of 15 months.

a If a family buys one of their machines, what is the probability that it will last more than eight years?

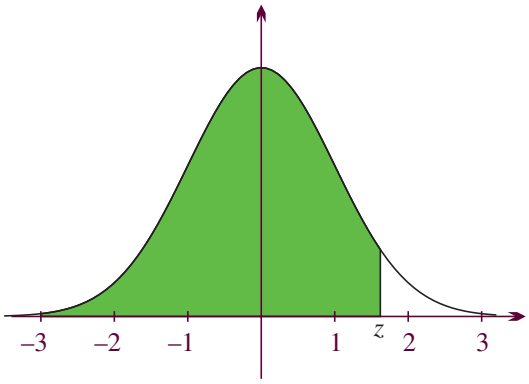
b The manufacturer is deciding on whether to launch a promotion and advertise a five-year warranty on its machines. How many machines could they expect to come to the end of their life within the five-year period?

ISBN 978-1-108-76630-2

Appendix: The standard normal distribution

A brief summary of the standard normal probability distribution

The graph to the right is the *standard normal probability density function* $y = \phi(z)$.



The shaded area represents the value of the corresponding *cumulative distribution function*

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

The table below gives some values of the probabilities $\phi(z) = P(Z \leq z)$. For example,

$$P(Z \leq 1.6) = \Phi(1.6) = \int_{-\infty}^{1.6} \phi(z)\, dz \doteq 0.9452.$$

	first decimal place									
z	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0.	0.5000	0.5398	0.5793	0.6179	0.6554	0.6915	0.7257	0.7580	0.7881	0.8159
1.	0.8413	0.8643	0.8849	0.9032	0.9192	0.9332	0.9452	0.9554	0.9641	0.9713
2.	0.9772	0.9821	0.9861	0.9893	0.9918	0.9938	0.9953	0.9965	0.9974	0.9981
3.	0.9987	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

A more detailed table giving the values of z to three decimal places is to be found in the Appendix to Chapter 17.

For many purposes, all that is required is the *empirical rule*, or *68–95–99.7 rule*,

$$P(-1 \leq Z \leq 1) \doteq 68\%$$
$$P(-2 \leq Z \leq 2) \doteq 95\%$$
$$P(-3 \leq Z \leq 3) \doteq 99.7\%$$

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2