Further Probability & Statistics

Chapter 8
Continuous random variables

In this chapter you will learn how to:

- use a probability density function that may be defined as a piecewise function
- use the general result $\mathrm{E}(\mathrm{g}(X)) = \int \mathrm{f}(x)\,\mathrm{g}(x)\,\mathrm{d}x$, where $\mathrm{f}(x)$ is the probability density function of the continuous random variable X, and $\mathrm{g}(X)$ is a function of X
- understand and use the relationship between the probability density function (PDF) and the cumulative distribution function (CDF), and use either to evaluate probabilities or percentiles
- use cumulative distribution functions of related variables in simple cases.

PREREQUISITE KNOWLEDGE

Where it comes from	What you should be able to do	Check your skills
AS & A Level Mathematics Probability & Statistics 1, Chapter 4	Calculate $E(X)$ and $Var(X)$.	1 <table><tr><td>x</td><td>1</td><td>2</td><td>3</td><td>4</td></tr><tr><td>$P(X = x)$</td><td>0.2</td><td>0.4</td><td>0.3</td><td>0.1</td></tr></table> Find $E(X)$ and $Var(X)$.
AS & A Level Mathematics Pure Mathematics 1, Chapter 9	Integrate and evaluate simple functions in a given interval.	2 Evaluate $\int_3^5 x^2 dx$.

What are continuous random variables?

A continuous random variable is a random variable that can take all values in an interval. It can be used to model quantities we measure, such as time or length.

A random variable could be a set of possible values from a random experiment. If the data can take any value within a given range then we say it is a continuous random variable. Suppose we measure the times people spend waiting for a bus at a bus stop. We know that a bus arrives every 13 minutes, but we do not know when the last bus arrived. Here, the waiting time is continuous and so we can model this situation with a continuous random variable. We can calculate mean waiting times, for example, the probability we will need to wait more than eight minutes.

In this chapter we shall study continuous random variables as well as their expectation and variance.

We shall use the **probability density function** (PDF) and **cumulative distribution function** (CDF) to calculate percentiles and probabilities. There are similarities with the work you did on discrete random variables and the normal distribution in AS & A Level Mathematics Probability & Statistics 1, Chapter 4 and Chapter 8. This work will help you to understand how the normal distribution is created.

We shall link related variables to find the PDF and CDF of functions of a variable.

8.1 The probability density function

A probability density function describes the probability of a continuous random variable in a similar way that a probability distribution table describes the probability of a discrete random variable.

The probability that a continuous random variable is equal to a particular value is always zero. This means we cannot use a table to describe the probability, so we use a function instead.

We need to know the conditions for a function on a given interval to represent a probability density function. Probability cannot be negative, and so a probability density function can *never* be negative, as shown in Key point 8.1.

KEY POINT 8.1

The function that defines the probability density function is always positive.

Condition 1: For $f(x)$ to represent a probability density function, $f(x) \geqslant 0$ for all values of x.

As the random variable is continuous, instead of adding values to evaluate probabilities over an interval, we must integrate the probability density function. Remember that integration can be used to evaluate the area bounded by a curve and the x-axis between particular limits. The area between the function and the x-axis defines the probability over an interval. The total area between the function and the x-axis must equal 1, as it represents the total probability of the probability density function, as shown in Key point 8.2.

KEY POINT 8.2

The area under the probability density function must equal 1.

This statement is equivalent to saying that the sum of all probabilities must equal 1.

Condition 2: For $f(x)$ to represent a probability density function, $\int f(x)\,dx = 1$ for all values of x.

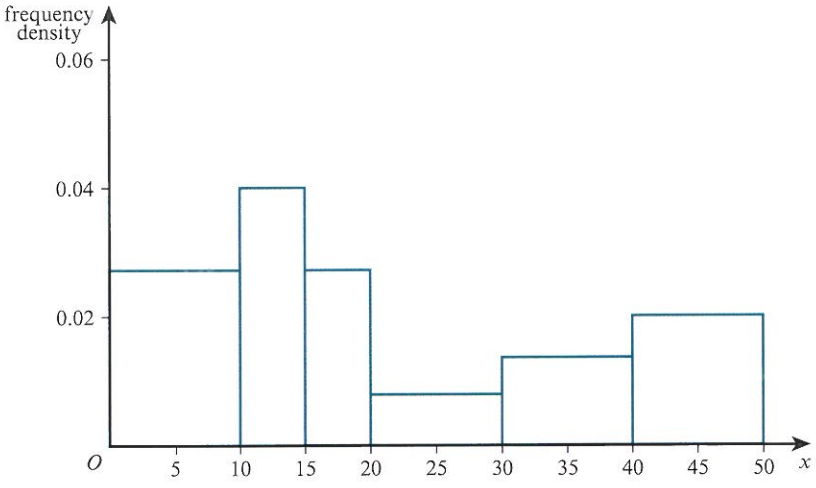
Conditions 1 and 2 must both be true for $f(x)$ to represent a probability density function.

Consider the following grouped continuous data.

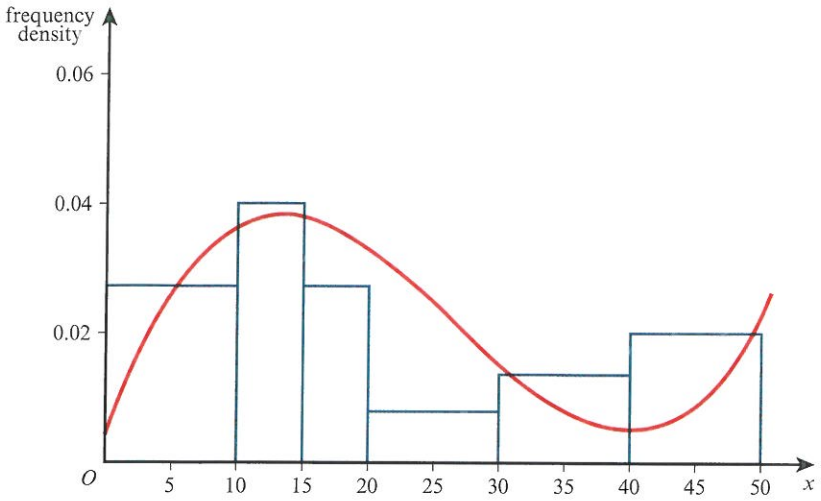
	Frequency	Relative frequency	Relative frequency density
$0 \leqslant x < 10$	40	$0.2\dot{6}$	$0.02\dot{6}$
$10 \leqslant x < 15$	30	0.2	0.04
$15 \leqslant x < 20$	20	$0.1\dot{3}$	$0.02\dot{6}$
$20 \leqslant x < 30$	10	$0.0\dot{6}$	$0.00\dot{6}$
$30 \leqslant x < 40$	20	$0.2\dot{6}$	$0.02\dot{6}$
$40 \leqslant x \leqslant 50$	30	0.2	0.02

If we were to draw a histogram of the continuous random variable, allowing the frequency to equal the area, we would have a histogram whose total area equals the frequency. If, instead, we considered the relative frequencies, then the area of the histogram would be 1. We know that relative frequency can represent probabilities.

The data are displayed in a histogram with a total area of 1.



In fact, this continuous data can be modelled using the following curve.

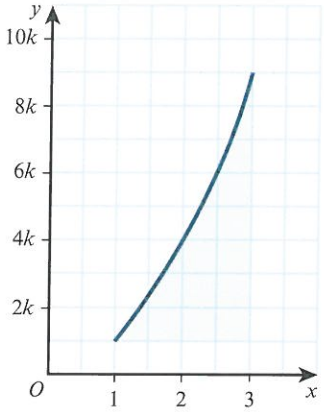


The area under this curve is also equal to 1.

WORKED EXAMPLE 8.1

Find the value of k for which $f(x) = kx^2$ could represent a probability density function over the interval $1 \leqslant x \leqslant 3$.

Answer



Check that the function is not negative. It is helpful to draw a sketch.

Alternatively, you may be required to show that a function is *always* positive for an interval.

$\int_1^3 kx^2 \, dx = 1$

Evaluate the integral with limits of 3 and 1, and equate to 1.

$\left[\frac{kx^3}{3}\right]_1^3 = 1$

Substitute in the limits.

$k\left(\frac{3^3}{3} - \frac{1^3}{3}\right) = 1$

Take out the common factor k.

$\frac{26k}{3} = 1$

Therefore, $k = \frac{3}{26}$.

Rearrange.

To find the probability between two values of the continuous random variable, integrate the PDF, $f(x)$, between those values, as shown in Key point 8.3. Note that, as $P(X = a) = 0$, $P(X < a)$ and $P(X \leqslant a)$ have the same value.

KEY POINT 8.3

The probability between two values of the continuous random variable is:

$$P(a < X < b) = \int_a^b f(x)\,dx = F(b) - F(a)$$

DID YOU KNOW?

In calculus, if we differentiate $f(x)$, we label it $f'(x)$ and call it the **derivative**.

If we integrate $f(x)$, we get $F(x)$ and call it the **primitive**.

We will learn more about this later but, in simple terms, when we integrate a probability density function $f(x)$ we find the cumulative distribution function $F(x)$. We can use cumulative distribution functions to calculate percentiles of distributions and probabilities. For example, $P(a \leqslant X \leqslant b) = F(b) - F(a)$.

When dealing with the normal distribution, we use $\Phi(z)$ to represent the cumulative distribution function. Φ is upper case phi in Greek. Its Latin equivalent is F. The probabilities are calculated in the same way from tables:

$$P(a \leqslant X \leqslant b) = \Phi(b) - \Phi(a)$$

REWIND

This is why, when dealing with discrete random variables in AS & A Level Probability & Statistics 1, Chapter 6, we used the following to represent the cumulative probability.

$F(x_0) = P(X \leqslant x_0)$

It is important to define fully the probability density function for all values of x. We must state for which values the PDF is valid, and for which values it is 0.

Consider the probability density function from Worked example 8.1: $f(x) = \frac{3x^2}{26}$ for $1 \leqslant x \leqslant 3$.

We should define this function for all values of x, so we write it as:

$$f(x) = \begin{cases} \frac{3x^2}{26} & 1 \leqslant x \leqslant 3 \\ 0 & \text{otherwise} \end{cases}$$

Now the function is defined for all values of x.

This notation can be used when dealing with probability density functions that are **piecewise functions**, as shown in Key point 8.4.

KEY POINT 8.4

Sometimes, probability density functions are represented by a combination of different functions, each corresponding to a part of the domain. Such probability density functions are called piecewise functions.

WORKED EXAMPLE 8.2

Consider the continuous random variable X, which has probability density function:

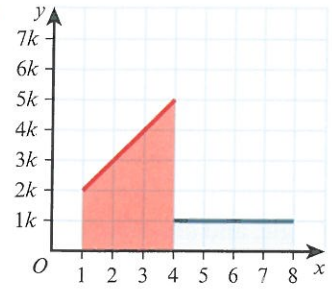
$$f(x) = \begin{cases} k(x+1) & 1 \leqslant x < 4 \\ k & 4 \leqslant x \leqslant 8 \\ 0 & \text{otherwise} \end{cases}$$

a Find the value of k.

b Calculate $P(2 \leqslant X < 6)$.

Answer

a



The function does *not* need to be piecewise continuous.

$$\int_1^4 k(x+1)\,dx + \int_4^8 k\,dx = 1$$

The total area must be 1.

$$k\left[\frac{x^2}{2} + x\right]_1^4 + k[x]_4^8 = 1$$

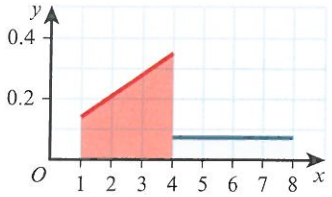
Take out k as a common factor to make the integration and algebra easier. Integrate and solve for k.

$$k\left(12 - \frac{3}{2}\right) + k(8-4) = 1$$

$$\frac{29k}{2} = 1$$

$$k = \frac{2}{29}$$

b



Ensure that the probabilities correspond to the domains of the PDF.

It is easy to make a numerical mistake, so show *all* of your working.

$$P(2 \leqslant X < 6) = P(2 \leqslant X < 4) + P(4 \leqslant X < 6)$$

$$\int_2^4 \frac{2}{29}(x+1)\,dx + \int_4^6 \frac{2}{29}\,dx$$

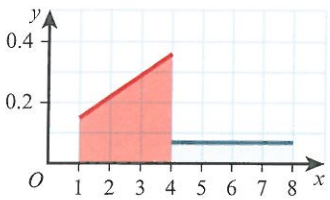
$$= \frac{2}{29}\left(\left[\frac{x^2}{2} + x\right]_2^4 + [x]_4^6\right)$$

$$= \frac{2}{29}((12-4) + (6-4))$$

$$P(2 \leqslant X < 6) = \frac{20}{29}$$

Alternatively:

Work out the area of the trapezium and the rectangle instead of using integration.



$$P(2 \leqslant X < 6) = P(2 \leqslant X < 4) + P(4 \leqslant X < 6)$$

For the trapezium:

Find the area of the trapezium.

$f(2) = \frac{6}{29}$, $f(4) = \frac{10}{29}$

$$P(2 \leqslant X < 4) = \frac{2\left(\frac{6}{29} + \frac{10}{29}\right)}{2} = \frac{16}{29}$$

For the rectangle:

Find the area of the rectangle, and add this to the area of the trapezium.

$P(4 \leqslant X < 6) = 2 \times \frac{2}{29} = \frac{4}{29}$

Total area $= \frac{16}{29} + \frac{4}{29} = \frac{20}{29}$ as before.

EXERCISE 8A

1 For each of the following, state whether or not it is a valid probability density function, giving a reason.

a $f(x) = \begin{cases} \frac{1}{10}(x+3) & 1 \leqslant x \leqslant 3 \\ 0 & \text{otherwise} \end{cases}$

b $f(x) = \begin{cases} -3x^2 + \frac{9}{2}x & 0 \leqslant x \leqslant 2 \\ 0 & \text{otherwise} \end{cases}$

c $f(x) = \begin{cases} x & 0 \leqslant x < 1 \\ 2 - x & 1 \leqslant x \leqslant 2 \\ 0 & \text{otherwise} \end{cases}$

d $f(x) = \begin{cases} x^2 & -1 \leqslant x \leqslant 2 \\ 0 & \text{otherwise} \end{cases}$

2 Sketch the following probability density functions.

a $f(x) = \begin{cases} \frac{1}{12}(x^2+3) & -1 \leqslant x \leqslant 2 \\ 0 & \text{otherwise} \end{cases}$

b $f(x) = \begin{cases} \frac{x}{4} & 0 \leqslant x < 2 \\ \frac{1}{2} & 2 \leqslant x \leqslant 3 \\ 0 & \text{otherwise} \end{cases}$

c $f(x) = \begin{cases} \frac{1}{20} & 0 \leqslant x < 5 \\ \frac{1}{96}(x-5) & 5 \leqslant x \leqslant 17 \\ 0 & \text{otherwise} \end{cases}$

3 Find the value of k for which $f(x) = \begin{cases} kx(3-x) & 0 \leqslant x \leqslant 3 \\ 0 & \text{otherwise} \end{cases}$ represents a probability density function.

4 Find the exact value of k for which $f(x) = \begin{cases} ke^{2(x-5)} & 3 \leqslant x \leqslant 4 \\ 0 & \text{otherwise} \end{cases}$ represents a probability density function.

5 For the given probability density function $f(x) = \begin{cases} k(x+1) & 5 \leqslant x \leqslant 9 \\ 0 & \text{otherwise} \end{cases}$, find:

a the value of k

b $P(X = 7)$

c $P(X < 8)$

d $P(X > 6)$

6 For the given probability density function $f(x) = \begin{cases} k(x^2 - 2x + 3) & 1 \leqslant x \leqslant 4 \\ 0 & \text{otherwise} \end{cases}$, find:

a the value of k

b $P(X < 2)$

c $P(1.5 \leqslant X < 3.5)$

7 Find the value of k for which $f(x) = \begin{cases} kx & 0 \leqslant x < 6 \\ \frac{k}{2}(9 - x) & 6 \leqslant x < 9 \\ 0 & \text{otherwise} \end{cases}$ represents a probability density function.

8 Find the value of k for which $f(x) = \begin{cases} \frac{k}{2}(x+1) & -1 \leqslant x < 3 \\ 2k & 3 \leqslant x < 4 \\ -\frac{2k}{3}(x+7) & 4 \leqslant x \leqslant 7 \\ 0 & \text{otherwise} \end{cases}$ represents a probability density function.

9 For the given probability density function $f(x) = \begin{cases} 4k & 5 \leqslant x < 7 \\ k(11 - x) & 7 \leqslant x < 11 \\ 0 & \text{otherwise} \end{cases}$, find:

a the value of k

b $P(X < 6)$

c $P(X > 8)$

d $P(6 < X \leqslant 10)$

10 For the given probability density function $f(x) = \begin{cases} k(6x - x^2) & 0 \leqslant x < 3 \\ 9k(5 - x) & 3 \leqslant x \leqslant 5 \\ 0 & \text{otherwise} \end{cases}$, find:

a the value of k

b $P(X < 3)$

c $P(X < 4.5)$

8.2 The cumulative distribution function

In this section, we shall see how to find a cumulative distribution function (CDF) from a probability density function (PDF), and vice versa.

We know from Section 8.1 that the area between the graph of a probability density function and the x-axis represents the probability. This area is found by integrating the PDF between suitable limits. If we integrate the PDF between the smallest value in the domain and use a variable as the upper limit, it will create a function that we can use to find the cumulative probability. We will not need to integrate the PDF every time. This is called the cumulative distribution function. We must define this for all values of x, as shown in Key point 8.5.

KEY POINT 8.5

Let the continuous random variable X have a probability density function $f(x)$. Then the cumulative distribution function is defined as:

$$F(x) = P(X \leqslant x) = \int_{-\infty}^{x} f(t)\,dt$$

Note that in Key point 8.5, since the limit is the variable x, we should not use x as the variable in the PDF. We simply choose a different letter here, known as a dummy variable.

Alternatively, instead of using limits, we could perform an indefinite integration. Then we would use the fact that the cumulative value at the right-hand end of the domain is 1 to find the constant of integration.

In Worked example 8.3, the probability density function consists of a single function. The first method shows the use of limits. The second method shows how we can use indefinite integration.

WORKED EXAMPLE 8.3

The continuous random variable X has probability density function $f(x) = \begin{cases} \frac{1}{12}x & 1 \leqslant x \leqslant 5 \\ 0 & \text{otherwise} \end{cases}$.

Find the cumulative distribution function.

Answer

Method 1: Using the limits

$F(x) = \int_{1}^{x} \frac{1}{12} t\,dt$ — Set up the integral. The lower limit is now 1. This is the smallest value in the domain.

$F(x) = \left[\frac{t^2}{24}\right]_{1}^{x}$ — Use t within the integral since x is within the limit.

$F(x) = \frac{x^2 - 1}{24}$ — Integrate and substitute in the limits.

It is worth checking that $F(5) = 1$ and $F(1) = 0$.

Method 2: Using indefinite integration

$F(x) = \int \frac{x}{12} dx$ — Since we are not using limits here, we can still use x as our variable.

$F(x) = \frac{x^2}{24} + c$

Using $F(1) = 0$ — We can use either $F(1) = 0$ or $F(5) = 1$ to find c:
$F(1) = 0$ since *no* probabilities have yet been added;
$F(5) = 1$ since *all* probabilities have been added.

$0 = \frac{1^2}{24} + c$

$c = -\frac{1}{24}$

$$F(x) = \frac{x^2}{24} - \frac{1}{24}$$

$$F(x) = \begin{cases} 0 & x < 1 \\ \dfrac{x^2 - 1}{24} & 1 \leqslant x \leqslant 5 \\ 1 & x > 5 \end{cases}$$

For both methods, define the cumulative distribution fully.

In Worked example 8.3, the probability density function has only one function. If the probability density function is a piecewise function, we need to find the cumulative distribution function for *each* piece of the function. We need to ensure that the cumulative probability from previous parts of the function is added. Worked example 8.4 shows two parts to the piecewise function.

Notice that in Worked example 8.4, the probability density function is not continuous. However, the cumulative distribution function *must* be continuous.

WORKED EXAMPLE 8.4

The continuous random variable X has probability density function $f(x) = \begin{cases} \frac{2}{29}(x+1) & 1 \leqslant x < 4 \\ \frac{2}{29} & 4 \leqslant x \leqslant 8 \\ 0 & \text{otherwise.} \end{cases}$

a Find the cumulative distribution function.

b Find $P(2 < X < 5)$.

c Find the value of a for which $P(X > a) = 0.1$.

Answer

a **Method 1**

Take each function in turn. A sketch graph is useful.

If $1 \leqslant x < 4$:

$$F(x) = \int_1^x \frac{2}{29}(t+1)\,dt$$

Find the cumulative distribution function for the first function.

$$= \left[\frac{2}{29}\left(\frac{t^2}{2} + t\right)\right]_1^x$$

$$= \frac{2}{29}\left(\frac{x^2}{2} + x\right) - \frac{2}{29}\left(\frac{1}{2} + 1\right)$$

$$F(x) = \frac{x^2}{29} + \frac{2x}{29} - \frac{3}{29}$$

And when $x = 4$, $F(4) = \frac{21}{29}$.

If $4 \leqslant x \leqslant 8$:

$$F(x) = F(4) + \int_4^x \frac{2}{29} dt$$

$$= \frac{21}{29} + \left[\frac{2t}{29}\right]_4^x$$

$$= \frac{21}{29} + \left(\frac{2x}{29} - \frac{8}{29}\right)$$

$$= \frac{13}{29} + \frac{2x}{29}$$

Since the cumulative distribution is continuous, the second domain starts at 4.

Check that $F(8) = 1$.

Method 2

If $1 \leqslant x < 4$:

Treat each part separately.

$$F(x) = \int \frac{2}{29}(x + 1)\, dx$$

$$= \frac{x^2}{29} + \frac{2x}{29} + c$$

Since $F(1) = 0$:

$$0 = \frac{1}{29} + \frac{2}{29} + c$$

Use the condition $F(1) = 0$.

Leading to $c = -\frac{3}{29}$.

Therefore, for this domain:

$$F(x) = \frac{x^2}{29} + \frac{2x}{29} - \frac{3}{29}$$

If $4 \leqslant x \leqslant 8$: $F(x) = \int \frac{2}{29} dx$

$$= \frac{2}{29}x + k$$

Here there is no need to add $F(4)$ since it becomes absorbed into the constant of integration.

Since $F(8) = 1$:

$$1 = \frac{16}{29} + k$$

There are two values we can use to calculate k. $F(4) = \frac{21}{29}$ and $F(8) = 1$. It is best to use $F(8) = 1$, since we may have calculated $F(4)$ incorrectly.

So, $k = \frac{13}{29}$

And therefore for $4 \leqslant x < 8$:

$$F(x) = \frac{2x}{29} + \frac{13}{29}$$

$$F(x) = \begin{cases} 0 & x < 1 \\ \frac{x^2}{29} + \frac{2x}{29} - \frac{3}{29} & 1 \leqslant x < 4 \\ \frac{2x}{29} + \frac{13}{29} & 4 \leqslant x \leqslant 8 \\ 1 & x > 8 \end{cases}$$

Define $F(x)$.

These can be factorised.

b Find $P(2 < X < 5)$.

Make sure that you use the correct part of $F(x)$ when evaluating the probability.

$$P(2 < X < 5) = P(X < 5) - P(X < 2)$$

$$= F(5) - F(2)$$

$$= \frac{2 \times 5 + 13}{29} - \frac{2^2 + 2 \times 2 - 3}{29}$$

$$= \frac{23}{29} - \frac{5}{29}$$

$$P(2 < X < 5) = \frac{18}{29}$$

c Find the value of a for which $P(X > a) = 0.1$.

Consider carefully in which domain the value of a will lie.

$F(a) = 0.9$

But since $F(4) = \frac{21}{29}$ and $F(8) = 1$ then

$F(4) < F(a) < F(8)$.

This implies $4 < a < 8$:

Check that your answer is in the correct domain.

$$F(a) = \frac{(2a + 13)}{29} = 0.9$$

$$a = 6.55$$

Percentiles

A **percentile** is a value that has a cumulative probability equal to a given probability.

For example, the 90th percentile, α, of a continuous random variable, X, is such that $P(X \leqslant \alpha) = 0.9$. Part **c** of Worked example 8.4 demonstrates how we find percentiles using the cumulative distribution function. a is the value below which 90% of the area lies. It is known as the 90th percentile. We need to be able to find a percentile, or show it correct to a given accuracy.

More generally, the nth percentile, α, of a continuous random variable, X, is $P(X \leqslant \alpha) = \frac{n}{100}$, as shown in Key point 8.6.

KEY POINT 8.6

The nth percentile, α, of a continuous random variable, X, is $\mathrm{P}(X \leqslant \alpha) = \frac{n}{100}$.

For the cumulative distribution function, $\mathrm{F}(\alpha) = \frac{n}{100}$, as shown in Key point 8.7.

KEY POINT 8.7

For the cumulative distribution function, $\mathrm{F}(\alpha) = \frac{n}{100}$ where α is the nth percentile.

The important percentiles are shown in Key point 8.8.

KEY POINT 8.8

The median	$\mathrm{F}(m) = 0.5$
The lower quartile	$\mathrm{F}(q_1) = 0.25$
The upper quartile	$\mathrm{F}(q_3) = 0.75$

WORKED EXAMPLE 8.5

Let X be a random variable with cumulative distribution function $\mathrm{F}(x) = \begin{cases} 0 & x < 0 \\ \dfrac{\mathrm{e}^x - 1}{\mathrm{e}^3 - 1} & 0 \leqslant x \leqslant 3 \\ 1 & x > 3. \end{cases}$

a Calculate the median.

b Calculate the lower and upper quartiles.

c Calculate the 40th percentile.

Answer

a $\mathrm{F}(m) = \dfrac{\mathrm{e}^m - 1}{\mathrm{e}^3 - 1} = \dfrac{1}{2}$ — Set the cumulative distribution function equal to $\frac{1}{2}$.

$\mathrm{e}^m = \dfrac{\mathrm{e}^3 - 1}{2} + 1$ — Rearrange and solve.

$m = \ln\left(\dfrac{\mathrm{e}^3 - 1}{2} + 1\right)$

$m = 2.355$

b $\mathrm{F}(q_1) = \dfrac{\mathrm{e}^{q_1} - 1}{\mathrm{e}^3 - 1} = \dfrac{1}{4}$ — Set the cumulative distribution function equal to $\frac{1}{4}$.

$\mathrm{e}^{q_1} = \dfrac{\mathrm{e}^3 - 1}{4} + 1$ — Rearrange and solve.

$$q_1 = \ln\left(\frac{e^3 - 1}{4} + 1\right)$$

$$q_1 = 1.753$$

$$F(q_3) = \frac{e^{q_3} - 1}{e^3 - 1} = \frac{3}{4}$$ Set the cumulative distribution function equal to $\frac{3}{4}$.

$$e^{q_3} = \frac{3(e^3 - 1)}{4} + 1$$ Rearrange and solve.

$$q_3 = \ln\left(\frac{3(e^3 - 1)}{4} + 1\right)$$

$$q_3 = 2.729$$

c $$F(\alpha) = \frac{e^{\alpha} - 1}{e^3 - 1} = 0.4$$ Set the cumulative distribution function equal to 0.4.

$$e^{\alpha} = 0.4(e^3 - 1) + 1$$ Rearrange and solve.

$$\alpha = \ln[0.4(e^3 - 1) + 1]$$

$$\alpha = 2.156$$

In Worked example 8.5, we were given the cumulative distribution function to start with. Sometimes we may need to find the CDF before calculating the median. Also, we may need to use some numerical methods to show that the value of a median, or in fact any percentile, is correct to a given level of accuracy.

WORKED EXAMPLE 8.6

Let X be a continuous random variable with probability density function $f(x) = \begin{cases} \frac{3}{10}\left(x^2 + \frac{1}{3}\right) & 0 \leqslant x \leqslant 2 \\ 0 & \text{otherwise.} \end{cases}$

Show that the median is 1.52, correct to 3 significant figures.

Answer

There are two methods we can use to solve this. First, we can find $F(x)$ and equate it to 0.5. Second, we could build this value into the limits for integration.

Method 1: Evaluating the integral directly

$$F(m) = P(X \leqslant m) = \int_0^m \frac{3}{10}\left(t^2 + \frac{1}{3}\right) dt = 0.5$$

$$\frac{3}{10}\left[\frac{t^3}{3} + \frac{t}{3}\right]_0^m = 0.5$$

$$\frac{1}{10}(m^3 + m) = 0.5$$

This leads to:

$g(m) = m^3 + m - 5 = 0$

This is a cubic that does not factorise.

You should not use a calculator to solve this as the question asks you to show that the median is 1.52 correct to 3 significant figures.

Since $m = 1.52$ (3 significant figures)

$m \in (1.515, 1.525)$.

$g(1.515) = 1.515^3 + 1.515 - 5 = -0.0077\ldots$.negative

$g(1.525) = 1.525^3 + 1.525 - 5 = 0.0715\ldots$.positive

Since $g(m)$ is continuous and there is a change in sign, m must be within the interval.

So $m = 1.52$ (3 significant figures).

This is covered in AS & A Level Pure Mathematics 3, Chapter 6.

Method 2

$$F(x) = \int_0^x \frac{3}{10}\left(t^2 + \frac{1}{3}\right) dt$$

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{10}(x^3 + x) & 0 \leqslant x \leqslant 2 \\ 1 & x > 2 \end{cases}$$

The advantage of finding the cumulative distribution function $F(x)$ rather than the probability directly is that it is possible to use this for other percentiles as well.

If $1.515 \leqslant m < 1.525$,

then $F(1.515) \leqslant 0.5 < F(1.525)$ and vice versa.

The principle here is the same. Instead of looking for a change in sign, we look for the values being either side of 0.5.

$F(1.515) = \frac{1}{10}(1.515^3 + 1.515) = 0.499\ldots < 0.5$

$F(1.525) = \frac{1}{10}(1.525^3 + 1.525) = 0.507\ldots > 0.5$

The $F(x)$ values are either side of 0.5.

Therefore $m = 1.52$ (3 significant figures).

Method 2 of Worked example 8.6 can be used to show any percentile to a given level of accuracy. We may not always be able to calculate the exact value.

We can find a cumulative distribution function from a probability density function by integrating. We may also need to find the PDF from a given CDF. This helps us calculate the mean or variance for a continuous random variable, as we cannot find this directly from the cumulative distribution function. We differentiate $F(x)$ to find $f(x)$, since differentiation is the inverse operation to integration. This is shown in Key point 8.9.

KEY POINT 8.9

Let the continuous random variable X have a cumulative distribution function $F(x)$. The probability density function is defined as:

$$f(x) = \frac{dF(x)}{dx}$$

If the function is piecewise, then, as in Worked example 8.7, make sure that all parts are differentiated.

WORKED EXAMPLE 8.7

A continuous random variable, X, has cumulative distribution function $F(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{108} & 0 \leqslant x < 6 \\ \frac{1}{54}\left(9x - \frac{x^2}{4} - 27\right) & 6 \leqslant x \leqslant 18 \\ 1 & \text{otherwise.} \end{cases}$

Find $f(x)$, the probability density function.

Answer

If $0 \leqslant x < 6$:

$\frac{dF(x)}{dx} = \frac{x}{54}$ ..... Differentiate $F(x)$ to find $f(x)$.

If $6 < x \leqslant 18$:

$\frac{dF(x)}{dx} = \frac{1}{54}\left(9 - \frac{x}{2}\right)$

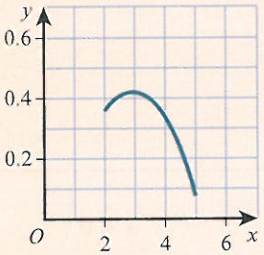
$f(x) = \begin{cases} \frac{x}{54} & 0 \leqslant x < 6 \\ \frac{1}{54}\left(9 - \frac{x}{2}\right) & 6 \leqslant x \leqslant 18 \\ 0 & \text{otherwise} \end{cases}$ ..... In the regions $x < 0$ and $x > 18$ we differentiate to 0. This is reflected in the 'otherwise' comment.

We can use the probability density function to find the mode of a function, as shown in Key point 8.10.

KEY POINT 8.10

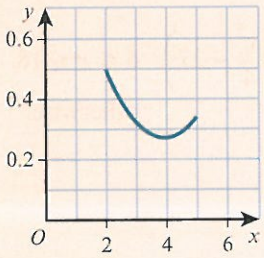
The mode is the highest point on a probability density function and so is either a stationary point or at the end points of the domain.

Given $f(x)$ defined as > 0 for $a \leqslant x \leqslant b$, then the mode is at $\frac{d}{dx}f(x) = 0\left[\frac{d^2}{dx^2}f(x) < 0\right]$ or a or b.



Here the mode is at the stationary point and is found at $\frac{df(x)}{dx} = 0$.

It is also a maximum $\left[\frac{d^2f(x)}{dx^2} < 0\right]$.



Here, we have a stationary point, but it is a minimum. We can see that the maximum value is at the start of the function.

The mode is a useful measure of central tendency, particularly if the data are highly skewed. We can also use the mode to discuss whether a dataset is positively or negatively skewed.

WORKED EXAMPLE 8.8

For each of the following probability density functions, find the mode.

a $f(x) = \begin{cases} \frac{1}{72}(8x - x^2) & 0 \leqslant x \leqslant 6 \\ 0 & \text{otherwise} \end{cases}$

b $f(x) = \begin{cases} \frac{1}{60}(2x + 3) & 2 \leqslant x \leqslant 7 \\ 0 & \text{otherwise} \end{cases}$

c $f(x) = \begin{cases} \frac{1}{48}(x^2 - 10x + 29) & 1 \leqslant x \leqslant 7 \\ 0 & \text{otherwise} \end{cases}$

Answer

a $f(x) = \begin{cases} \frac{1}{72}(8x - x^2) & 0 \leqslant x \leqslant 6 \\ 0 & \text{otherwise} \end{cases}$

We know that the stationary point is a maximum.

Stationary point:

$\frac{d}{dx}(f(x)) = \frac{1}{72}(8 - 2x) = 0$

Maximum where $x = 4$:

We need to work out the values of potential maxima.

$f(0) = 0$

$f(4) = \frac{2}{9}$

$f(6) = \frac{1}{6}$

The mode is therefore when $x = 4$.

Choose the greatest value.

b $f(x) = \begin{cases} \frac{1}{60}(2x + 3) & 2 \leqslant x \leqslant 7 \\ 0 & \text{otherwise} \end{cases}$

Since the function is linear, it has no stationary points.

$f(2) = \frac{7}{60}$

$f(7) = \frac{17}{60}$

The mode is therefore when $x = 7$.

Choose the greater value.

Alternatively, since the function is linear and always increasing, we can deduce the maximum value will be at $x = 7$.

c $f(x) = \begin{cases} \frac{1}{48}(x^2 - 10x + 29) & 1 \leqslant x \leqslant 7 \\ 0 & \text{otherwise} \end{cases}$

Find the stationary point.

Stationary point:

$\frac{d}{dx}(f(x)) = \frac{1}{48}(2x - 10) = 0$

Maximum where $x = 5$:

Work out the values of potential maxima.

$f(1) = \frac{5}{12}$

$f(5) = \frac{1}{12}$

$f(7) = \frac{1}{6}$

The mode is therefore when $x = 1$.

Choose the greatest value.

If the functions are more complicated, use a graph to help you, as shown in Worked example 8.9.

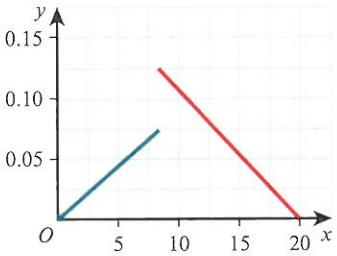
WORKED EXAMPLE 8.9

Given $f(x) = \begin{cases} \frac{1}{128}x & 0 \leqslant x < 8 \\ \frac{5}{24} - \frac{x}{96} & 8 \leqslant x \leqslant 20 \\ 0 & \text{otherwise} \end{cases}$, find the mode.

Answer

Mode = 8

It is easy to see where the mode is from the graph.



EXERCISE 8B

1 Find $F(x)$, the cumulative distribution function for:

$$f(x) = \begin{cases} \frac{2}{95}(5 - 3x) & -4 \leqslant x \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

2 Find $F(x)$, the cumulative distribution function for:

$$f(x) = \begin{cases} \frac{1}{9}(x^2 - 8x + 18) & 2 \leqslant x \leqslant 5 \\ 0 & \text{otherwise} \end{cases}$$

3 Find $F(x)$, the cumulative distribution function for:

$$f(x) = \begin{cases} \frac{1}{16} & 3 \leqslant x < 7 \\ \frac{1}{8} & 7 \leqslant x \leqslant 13 \\ 0 & \text{otherwise} \end{cases}$$

4 Find $F(x)$, the cumulative distribution function for:

$$f(x) = \begin{cases} \frac{4}{27}(x - 1) & 1 \leqslant x < 4 \\ \frac{4}{27}(11 - 2x) & 4 \leqslant x \leqslant \frac{11}{2} \\ 0 & \text{otherwise} \end{cases}$$

5 Find $F(x)$, the cumulative distribution function for:

$$f(x) = \begin{cases} \frac{1}{24} & 0 \leqslant x < 5 \\ \frac{1}{24}(x-4) & 5 \leqslant x < 7 \\ \frac{1}{8} & 7 \leqslant x < 12 \\ 0 & \text{otherwise} \end{cases}$$

6 For the given probability density function: $f(x) = \begin{cases} \frac{1}{28}(12-x) & 3 \leqslant x \leqslant 7 \\ 0 & \text{otherwise} \end{cases}$

a find $F(x)$

b calculate $F(5)$

c find $P(4 \leqslant x < 6)$

d find m such that $F(m) = 0.5$. Give your answer to 2 decimal places.

7 For the given probability density function: $f(x) = \begin{cases} \frac{3}{100}(x-2)(8-x) & 2 \leqslant x \leqslant 7 \\ 0 & \text{otherwise} \end{cases}$

a find $F(x)$, writing your answer in the form $\frac{-a}{100}(x-b)(x-c)^2$, where a, b, c are positive integers

b find $P(X \leqslant 4)$

c find $P(X > 5)$.

8 For the given probability density function: $f(x) = \begin{cases} \frac{1}{90}(13-x) & 4 \leqslant x < 7 \\ \frac{1}{270}(x+11) & 7 \leqslant x \leqslant 16 \\ 0 & \text{otherwise} \end{cases}$

a find $F(x)$

b find $P(X \leqslant 6)$

c find $P(X \leqslant 11)$

d find a such that $P(X \geqslant a) = 0.75$

e find m such that $F(m) = 0.5$, giving your answer to 3 significant figures.

9 For the given probability density function: $f(x) = \begin{cases} \frac{1}{99}(x^2 - 18x + 83) & 6 \leqslant x \leqslant 15 \\ 0 & \text{otherwise} \end{cases}$

a find $F(x)$

b find $P(X > 8)$

c show that the upper quartile is 14.3, to 1 decimal place.

PS **10** For the given probability density function: $f(x) = \begin{cases} \frac{12}{335}(x^3 + 4x^2 + 1) & -4 \leqslant x \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$

a find the mode

b show that the 40th percentile is -2.56, correct to 2 decimal places.

8.3 Calculating E(g(X)) for a continuous random variable

REWIND

In AS & A Level Mathematics Probability & Statistics 2, Chapter 2, we found both $E(aX + b)$ and $Var(aX + b)$ for discrete random variables. Also, in Chapter 4, we found $E(X)$ and $Var(X)$ for continuous random variables.

For discrete random variables we could simply recalculate the expectation and variance by redefining the variable, for example:

x	1	2	3
$P(X = x)$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{2}$

When $Y = 3X + 7$, we can write the probability distribution of Y as:

y	10	13	16
$P(Y = y)$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{2}$

We can calculate $E(3X + 7)$ and $Var(3X + 7)$ from this table.

For continuous random variables we cannot use this method. We need to be able to find $E(g(X))$ and $Var(g(X))$ from their probability density function, as shown in Key point 8.11.

KEY POINT 8.11

Let X be a continuous random variable with probability density function $f(x)$. Then:

$$E(X) = \int_{\forall x} xf(x)\,dx \text{ and } E(g(X)) = \int_{\forall x} g(x)f(x)\,dx$$

TIP

$\forall x$ means 'for all x'. This notation allows us to find the area when there are many domains for the continuous random variable.

There are similar integrals in AS & A Level Mathematics Probability & Statistics 2. You may have used these to calculate $E(X^2)$ to find $Var(X)$: $E(X^2) = \int_{\forall x} x^2 f(x)\,dx$ and $Var(X) = E(X^2) - [E(X)]^2$.

WORKED EXAMPLE 8.10

A continuous random variable, X, has probability density function $f(x) = \begin{cases} \frac{1}{5}\left(\frac{6x}{5} + \frac{1}{2}\right) & 0 \leqslant x \leqslant 2.5 \\ 0 & \text{otherwise.} \end{cases}$

a Find $E(X)$.

b Find $E(X(X + 1))$.

Answer

a $E(X) = \int_{\forall x} xf(x)\,dx$ — From the definition.

$$E(X) = \int_0^{2.5} x \times \frac{1}{5}\left(\frac{6x}{5} + \frac{1}{2}\right) dx$$

$$= \frac{1}{5}\int_0^{2.5} \frac{6x^2}{5} + \frac{x}{2} dx$$ Multiply and simplify.

$$= \frac{1}{5}\left[\frac{2x^3}{5} + \frac{x^2}{4}\right]_0^{2.5}$$

$$= \frac{1}{5}\left(\left(\frac{2(2.5)^3}{5} + \frac{2.5^2}{4}\right) - 0\right)$$ Substitute in limits.

$$= 1.5625$$ Always check to see if the answer makes sense. It must be between 0 and 2.5.

b $E(g(X)) = \int_{\forall x} g(x) f(x) dx$ From the definition.

$$E(X(X+1)) = \int_0^{2.5} x(x+1) \times \frac{1}{5}\left(\frac{6x}{5} + \frac{1}{2}\right) dx$$

$$= \frac{1}{5}\int_0^{2.5} \frac{6x^3}{5} + \frac{17x^2}{10} + \frac{x}{2} dx$$ Multiply out and collect like terms.

$$= \frac{1}{5}\left[\frac{3x^4}{10} + \frac{17x^3}{30} + \frac{x^2}{4}\right]_0^{2.5}$$

$$= \frac{1}{5}\left(\left(\frac{3(2.5)^4}{10} + \frac{17(2.5)^3}{30} + \frac{2.5^2}{4}\right) - 0\right)$$ Substitute in limits.

$E(X(X+1)) = 4.427083...$

$= 4.43$ (to 3 significant figures) Evaluate.

WORKED EXAMPLE 8.11

A continuous random variable, X, has probability density function $f(x) = \begin{cases} \frac{x}{12} & 0 \leqslant x < 3 \\ \frac{1}{8} & 3 \leqslant x < 8 \\ 0 & \text{otherwise} \end{cases}$.

Find $E\left(\frac{1}{X}\right)$.

Answer

$E(g(X)) = \int_{\forall x} g(x) f(x) dx$ From the definition.

$$E\left(\frac{1}{X}\right) = \int_0^3 \frac{1}{x}\left(\frac{x}{12}\right)dx + \int_3^8 \frac{1}{x}\left(\frac{1}{8}\right)dx$$

Since there are two domains, integrate over all values of x. Split the integral into sections to do this using the given domains.

$$= \int_0^3 \frac{1}{12}dx + \int_3^8 \frac{1}{8x}dx$$

$$= \left[\frac{x}{12}\right]_0^3 + \left[\frac{1}{8}\ln x\right]_3^8$$

Integrate.

$$= \left(\frac{1}{4} - 0\right) + \left(\frac{1}{8}\ln 8 - \frac{1}{8}\ln 3\right)$$

Evaluate.

$$= \frac{1}{8}\left(2 + \ln\left(\frac{8}{3}\right)\right)$$

Simplify.

$$= 0.372\,603\,6\ldots$$

Make sure you read the question carefully. You may be required to leave your answer in exact form.

$$= 0.373 \text{ (to 3 significant figures)}$$

DID YOU KNOW?

Another name for $E(X)$ is the first moment of the distribution of X and, using this, we can find a measure of centrality. It links to finding a centre of mass, using moments, in Mechanics.

$E(X^2)$ is called the second moment and links to measures of dispersion.

$E(X^n)$ is called the nth moment of the distribution of X. When $n = 3$, we can calculate some measures of skewness and, when $n = 4$, we are able to analyse kurtosis, a measure of how 'flat' a distribution is.

EXERCISE 8C

1 $f(x) = \begin{cases} \frac{x}{200} & 0 \leqslant x \leqslant 20 \\ 0 & \text{otherwise.} \end{cases}$

a Find $E(X)$. **b** Find $E(X^2)$. **c** Find $\text{Var}(X)$.

2 The continuous random variable X has probability density function given by $f(x) = \begin{cases} \frac{2}{21}(7 - x) & 2 \leqslant x \leqslant 5 \\ 0 & \text{otherwise.} \end{cases}$

a Find $E(X)$. **b** Find $E(X^2)$. **c** Find $\text{Var}(X)$.

3 The continuous random variable X has probability density function given by $f(x) = \begin{cases} \frac{1}{5} & 5 \leqslant x < 8 \\ \frac{1}{15} & 8 \leqslant x \leqslant 14 \\ 0 & \text{otherwise.} \end{cases}$

a Find $E(X)$. **b** Find $E(X^2)$. **c** Find $\text{Var}(X)$. **d** Find $\text{SD}(X)$.

4 The continuous random variable X has probability density function given by $f(x) = \begin{cases} \frac{3}{20} & 1 \leqslant x \leqslant 6 \\ \frac{1}{200}(16 - x) & 6 < x \leqslant 16 \\ 0 & \text{otherwise.} \end{cases}$

a Find $E(X)$. **b** Find $E(X^2)$. **c** Find $\text{Var}(X)$.

5 The continuous random variable X has probability density function given by $f(x) = \begin{cases} \frac{5}{64}\left(3 + \frac{1}{x^2}\right) & 1 \leqslant x \leqslant 5 \\ 0 & \text{otherwise.} \end{cases}$

Find $E(X)$, giving your answer in the form $a(b + \ln c)$.

PS 6 The continuous random variable X has probability density function given by $f(x) = \begin{cases} \frac{1}{56}(23 - x) & 7 \leqslant x \leqslant 11 \\ 0 & \text{otherwise.} \end{cases}$

Find $E(X(X - 1))$.

7 The continuous random variable X has probability density function given by $f(x) = \begin{cases} 1 - \frac{x}{4} & 1 \leqslant x \leqslant 3 \\ 0 & \text{otherwise.} \end{cases}$

Find the exact value of $E(e^X)$.

8 The continuous random variable X has probability density function given by

$f(x) = \begin{cases} -\frac{3}{16}(x^2 - 10x + 22) & 4 \leqslant x \leqslant 6 \\ 0 & \text{otherwise.} \end{cases}$

Find $E\left(\frac{1}{X^2}\right)$. Give your answer in the form $a \ln b + c$.

9 The continuous random variable X has probability density function given by $f(x) = \begin{cases} \frac{4}{15} & 0 \leqslant x \leqslant 3 \\ \frac{1}{10} & 3 < x \leqslant 5 \\ 0 & \text{otherwise.} \end{cases}$

Find $E((X - 2)^2)$.

10 The continuous random variable X has probability density function given by $f(x) = \begin{cases} \frac{1}{8}(5 - x) & 1 \leqslant x < 3 \\ \frac{1}{32}(x - 3) & 3 \leqslant x \leqslant 7 \\ 0 & \text{otherwise.} \end{cases}$

Find $E\left(\frac{1}{X}\right)$.

PS 11 For the given cumulative distribution function: $F(x) = \begin{cases} 0 & x < 1 \\ \frac{1}{36}(x^3 + 5x - 6) & 1 \leqslant x \leqslant 3 \\ 1 & x > 3 \end{cases}$

a show that the median is 2.32, correct to 3 significant figures

b find the mode

c find $E(X)$

d use your answers to parts **a**, **b** and **c** to comment on the skewness of the distribution.

8.4 Finding the probability density function and cumulative distribution function of $Y = g(X)$

In Section 8.3 we saw how to find $E(Y)$, where $Y = g(X)$. We now need to calculate the probability density function and the cumulative distribution function for the continuous random variable $Y = g(X)$. This will allow us to calculate percentiles and probabilities for these functions. With a discrete random variable, we can simply recalculate the probability distribution and cumulative distribution. We will work with a discrete random variable with the following probability distribution. This will help us to develop some ideas that we can use later with continuous random variables.

x	1	2	3
$P(X = x)$	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{3}{8}$

And cumulative distribution:

x	1	2	3
$F(x)$	$\frac{1}{2}$	$\frac{5}{8}$	1

Consider $Y = X^2$. The probability distribution is:

y	1	4	9
$P(Y = y)$	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{3}{8}$

Or, equivalently, using X:

x	1	2	3
$P(X = \sqrt{y})$	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{3}{8}$

And cumulative distribution $G(y)$:

y	1	4	9
$G(y)$ $P(Y \leqslant y)$	$\frac{1}{2}$	$\frac{5}{8}$	1

Or, equivalently, using X:

x	1	2	3
$F(\sqrt{y})$ $P(X \leqslant \sqrt{y})$	$\frac{1}{2}$	$\frac{5}{8}$	1

If $Y = h(X)$, then $G(y) = P(X \leqslant h^{-1}(y))$.

Consider $Y = -X$. The probability distribution is:

y	−3	−2	−1
$P(Y = y)$	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{1}{2}$

And cumulative distribution $G(y)$:

y	-3	-2	-1
$G(y) = P(Y \leqslant y)$	$\frac{3}{8}$	$\frac{1}{2}$	1

Or equivalently, using X:

x	1	2	3
$F(-y)$ $P(X \geqslant y) = 1 - P(X \leqslant -y)$	$\frac{1}{2}$	$\frac{5}{8}$	1

Here, we can see that if $Y = h(X)$, then $G(y) = 1 - P(X \leqslant h^{-1}(y))$.

Consider $Y = \frac{1}{X}$. The probability distribution is:

y	$\frac{1}{3}$	$\frac{1}{2}$	1
$P(Y = y)$	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{1}{2}$

And cumulative distribution $G(y)$:

y	$\frac{1}{3}$	$\frac{1}{2}$	1
$G(y) = P(Y \leqslant y)$	$\frac{3}{8}$	$\frac{1}{2}$	1

Or, equivalently, using X:

x	1	2	3
$F\left(\frac{1}{y}\right)$ $P\left(X \geqslant \frac{1}{y}\right) = 1 - P\left(X \leqslant \frac{1}{y}\right)$	$\frac{1}{2}$	$\frac{5}{8}$	1

Here, we can see also that if $Y = h(X)$, then $G(y) = 1 - P(X \leqslant h^{-1}(y))$ for a discrete random variable.

For continuous variables we cannot do this. Instead, we use a similar idea with the cumulative distribution function of $Y = g(X)$, as shown in Key point 8.12.

KEY POINT 8.12

For a continuous random variable, X, with cumulative distribution function $F(x)$ and a function $Y = h(X)$, we find the cumulative distribution function $G(y)$ by:

$$G(y) = P(Y \leqslant y) = \begin{cases} P(X \leqslant h^{-1}(y)) = F(h^{-1}(y)) \\ \text{or} \\ P(X \geqslant h^{-1}(y)) = 1 - F(h^{-1}(y)) \end{cases}$$

WORKED EXAMPLE 8.12

Consider the continuous random variable X with cumulative distribution function $\mathrm{F}(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{16} & 0 \leqslant x \leqslant 4 \\ 1 & x > 4. \end{cases}$

Find the cumulative distribution function of $Y = X^2$.

Answer

$$\mathrm{F}(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{16} & 0 \leqslant x \leqslant 4 \\ 1 & x > 4 \end{cases}$$

$Y = X^2$

Note that as X increases, so does Y.

This becomes important later.

Consider $\mathrm{G}(y) = \mathrm{P}(Y \leqslant y)$, which is also $\mathrm{P}(X \leqslant \sqrt{y}) = \mathrm{F}(\sqrt{y})$

Apply this now to the CDF.

$$\mathrm{F}(\sqrt{y}) = \begin{cases} 0 & \sqrt{y} < 0 \\ \frac{(\sqrt{y})^2}{16} & 0 \leqslant \sqrt{y} \leqslant 4 \\ 1 & \sqrt{y} > 4 \end{cases}$$

This is the same as $\mathrm{F}(x)$ with the function applied.

$$\mathrm{G}(y) = \begin{cases} 0 & y < 0 \\ \frac{y}{16} & 0 \leqslant y \leqslant 16 \\ 1 & y > 16 \end{cases}$$

Write in terms of y.

This is the full description of the cumulative distribution function for X^2.

In Worked example 8.12 we saw how to find the CDF of a function of X from the CDF of the continuous random variable X.

What if we start with the probability density function of X and need to find the PDF of $Y = \mathrm{g}(X)$?

We cannot do this directly, but there is a way, using the material covered so far:

$\mathrm{f}(x)$ the PDF of $X \rightarrow \mathrm{F}(x)$ the CDF of $X \rightarrow \mathrm{G}(y)$ the CDF of $Y \rightarrow \mathrm{g}(y)$ the PDF of Y

WORKED EXAMPLE 8.13

A continuous random variable, X, has probability density function $\mathrm{f}(x) = \begin{cases} \frac{8}{3x^3} & 1 \leqslant x \leqslant 2 \\ 0 & \text{otherwise} \end{cases}$. Find the probability density function of $Y = \frac{X^2}{4}$.

Answer

Find $F(x)$ first:

Always calculate the cumulative distribution function first.

$$F(x) = \int_1^x \frac{8}{3t^3}\,dt$$

$$= \left[-\frac{4}{3t^2}\right]_1^x$$

$$= -\frac{4}{3x^2} - \left(-\frac{4}{3}\right)$$

$$= \frac{4}{3}\left(1 - \frac{1}{x^2}\right)$$

Therefore, $F(x) = \begin{cases} 0 & x < 1 \\ \frac{4}{3}\left(1 - \frac{1}{x^2}\right) & 1 \leqslant x \leqslant 2 \\ 1 & x > 2 \end{cases}$

Now consider the function $Y = \frac{X^2}{4}$.

Make X the subject.

$$X = 2\sqrt{Y}$$

$$G(y) = P(Y \leqslant y) = P(X \leqslant 2\sqrt{y})$$

Apply this to the cumulative function.

$$G(y) = F(2\sqrt{y})$$

$$F(2\sqrt{y}) = \frac{4}{3}\left(1 - \frac{1}{(2\sqrt{y})^2}\right)$$

$$G(y) = \frac{4}{3} - \frac{1}{3y}$$

The domain of $F(x)$ is $1 \leqslant x \leqslant 2$.

Now consider the domain.

And of $F(2\sqrt{y})$: $1 \leqslant 2\sqrt{y} \leqslant 2$

Apply the function, rearranging to make y the subject.

$$1 \leqslant 4y \leqslant 4$$

$$\frac{1}{4} \leqslant y \leqslant 1$$

Therefore, $G(y) = \begin{cases} 0 & y < \frac{1}{4} \\ \frac{4}{3} - \frac{1}{3y} & \frac{1}{4} \leqslant y \leqslant 1 \\ 1 & y > 1 \end{cases}$

It is useful to check:

$$G\left(\frac{1}{4}\right) = \frac{4}{3} - \frac{1}{3\left(\frac{1}{4}\right)} = 0$$

$$G(1) = \frac{4}{3} - \frac{1}{3(1)} = 1$$

$g(y) = \frac{dG(y)}{dy}$ — Differentiate.

$g(y) = \frac{1}{3y^2}$

And $g(y) = \begin{cases} \frac{1}{3y^2} & \frac{1}{4} \leqslant y \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$ — As required. Remember to define fully the probability density function.

When we deal with reciprocal or negative functions, we need to be very careful about how we define the cumulative distribution function. For example, as in Worked example 8.14, if the function is $Y = \frac{1}{X}$, ensure that you define the correct domain.

WORKED EXAMPLE 8.14

Let the continuous random variable X have probability density function

$$f(x) = \begin{cases} \frac{8}{3x^3} & 1 \leqslant x \leqslant 2 \\ 0 & \text{otherwise.} \end{cases}$$

Find the probability density function of $Y = \frac{1}{X}$.

Answer

$F(x) = \int \frac{8}{3x^3} dx$ — Find $F(x)$.

$= -\frac{4}{3x^2} + c$ — Find the constant of integration instead of using limits.

$-\frac{4}{3(2)^2} + c = 1$ — Use $F(2) = 1$.

Therefore, $c = \frac{4}{3}$.

And $F(x) = \frac{4}{3}\left(1 - \frac{1}{x^2}\right)$.

Now consider the function $Y = \frac{1}{X}$. — Since we are taking the reciprocal, change the inequality. Think about why the inequality switches over in this example.

$G(y) = P(Y \leqslant y) = P\left(X \geqslant \frac{1}{y}\right) = 1 - P\left(X \leqslant \frac{1}{y}\right)$

$G(y) = 1 - F\left(\frac{1}{y}\right)$

$= 1 - \frac{4}{3}(1 - y^2)$

$= \frac{1}{3}(4y^2 - 1)$

The domain for $F(x)$ is $1 \leqslant x \leqslant 2$.

The same care is required for the domain.

$1 \leqslant \frac{1}{y} \leqslant 2$

The domain for $G(y)$ is therefore

Check the domain for the reciprocal function.

$\frac{1}{2} \leqslant y \leqslant 1$.

Therefore, $G(y) = \begin{cases} 0 & y < \frac{1}{2} \\ \frac{1}{3}(4y^2 - 1) & \frac{1}{2} \leqslant y \leqslant 1 \\ 1 & y > 1 \end{cases}$

Check:

$G\left(\frac{1}{2}\right) = \frac{1}{3}\left(4\left(\frac{1}{2}\right)^2 - 1\right) = 0$

$G(1) = \frac{1}{3}(4(1)^2 - 1) = 1$

$g(y) = \frac{dG(y)}{dy}$

Differentiate to find $g(y)$.

$g(y) = \frac{8y}{3}$

$g(y) = \begin{cases} \frac{8y}{3} & \frac{1}{2} \leqslant y \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$

Define fully $g(y)$.

TIP

Always take your time when finding the CDF of a function of X, and notice when you need to change the inequality.

EXERCISE 8D

1 $F(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{400} & 0 \leqslant x \leqslant 20 \\ 1 & x > 20 \end{cases}$. Find the cumulative distribution function of $A = X^2$.

2 $F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{10}(x^3 + x) & 0 \leqslant x \leqslant 2 \\ 1 & x > 2 \end{cases}$. Find the cumulative distribution function of $A = X^3$.

3 $F(x) = \begin{cases} 0 & x < -4 \\ \frac{1}{95}(-3x^2 + 10x + 88) & -4 \leqslant x \leqslant 1 \\ 1 & \text{otherwise} \end{cases}$. Find the cumulative distribution function $A = 3X - 22$.

4 The continuous random variable X has cumulative distribution function given by

$F(x) = \begin{cases} 0 & x < 1 \\ \frac{1}{3}(x^2 - 1) & 1 \leqslant x \leqslant 2 \\ 1 & x > 2 \end{cases}$. Find the cumulative distribution function of:

a $A = X^2$ b $B = \sqrt{X}$

5 The continuous random variable X has cumulative distribution function given by

$$F(x) = \begin{cases} & 0x < 0 \\ \frac{1}{300}(x^2 + 20x) & 0 \leqslant x \leqslant 10 \\ & 1x > 10 \end{cases}$$

Find the cumulative distribution function of $Y = 100X^2$.

6 The continuous random variable X has probability density function given by $f(x) = \begin{cases} \frac{1}{8}(4 - x) & 0 \leqslant x \leqslant 4 \\ 0 & \text{otherwise.} \end{cases}$

a Find $F(x)$.

b Find the cumulative distribution function of $Y = 3X - 2$.

c Find the probability density function of Y.

7 The continuous random variable X has probability density function given by $f(x) = \begin{cases} \frac{2}{x^2} & 1 \leqslant x \leqslant 2 \\ 0 & \text{otherwise.} \end{cases}$

a Find $F(x)$.

b Find the cumulative distribution of $Y = \frac{X^2}{4}$.

c Find the probability density function of Y.

8 The continuous random variable X has cumulative distribution function given by

$$F(x) = \begin{cases} 0 & x < 1 \\ -\frac{25}{24}\left(\frac{1}{x^2} - 1\right) & 1 \leqslant x \leqslant 5 \\ 1 & \text{otherwise.} \end{cases}$$

a Find the cumulative distribution function of $Y = \frac{1}{X}$.

b Find the probability density function of Y.

9 The continuous random variable X has probability density function given by $f(x) = \begin{cases} \frac{2}{25}(5 - x) & 0 \leqslant x \leqslant 5 \\ 0 & \text{otherwise.} \end{cases}$

a Find $F(x)$.

b Find the cumulative distribution of $Y = 5 - 2X$.

c Find $P(Y < 2)$.

d Find $P(-2 < Y < 2)$.

e Find the probability density function of Y.

M **10** A circular ink blot has radius r, described by the probability distribution $\mathrm{f}(r) = \begin{cases} \frac{2}{25}(6-r) & 1 \leqslant r \leqslant 6 \\ 0 & \text{otherwise.} \end{cases}$

a Find $\mathrm{F}(r)$, the cumulative density function.

b Find $\mathrm{G}(A)$, the cumulative distribution function for the area of the ink blot.

c Find the probability density function for the area of the ink blot.

WORKED PAST PAPER QUESTION

The continuous random variable X has probability density function f given by

$$\mathrm{f}(x) = \begin{cases} \frac{1}{6}x & 2 \leqslant x \leqslant 4 \\ 0 & \text{otherwise} \end{cases}$$

i Find the cumulative distribution function of X.

The continuous random variable Y is defined by $Y = X^3$. Find

ii the probability density function of Y

iii the value of k for which $\mathrm{P}(Y \geqslant k) = \frac{7}{12}$

Cambridge International AS & A Level Further Mathematics 9231 Paper 21 Q7 November 2016

Answer

i $\mathrm{f}(x) = \begin{cases} \frac{1}{6}x & 2 \leqslant x \leqslant 4 \\ 0 & \text{otherwise} \end{cases}$

For $2 \leqslant x \leqslant 4$: Integrate to find $\mathrm{F}(x)$.

$\mathrm{F}(x) = \int_2^x \frac{1}{6}t \, \mathrm{d}t = \left[\frac{t^2}{12}\right]_2^x$ Use a dummy variable.

$= \frac{x^2}{12} - \frac{2^2}{12}$ Evaluate.

$\mathrm{F}(x) = \frac{x^2}{12} - \frac{1}{3}$

$\mathrm{F}(x) = \begin{cases} 0 & x < 2 \\ \frac{x^2}{12} - \frac{1}{3} & 2 \leqslant x \leqslant 4 \\ 1 & x > 4 \end{cases}$ Define fully.

ii $G(y) = P(Y < y) = P(X^3 < y)$

State G(y) in terms of X.

$$= P\left(X < y^{\frac{1}{3}}\right) = F\left(y^{\frac{1}{3}}\right)$$

$$G(y) = \frac{y^{\frac{2}{3}}}{12} - \frac{1}{3}$$

Define G(y).

$$g(y) = \frac{1}{18}y^{-\frac{1}{3}}$$

Differentiate to get g(y).

$$2 \leqslant y^{\frac{1}{3}} \leqslant 4$$

$$8 \leqslant y \leqslant 64$$

$$g(y) = \begin{cases} \frac{1}{18}y^{-\frac{1}{3}} & 8 \leqslant y \leqslant 64 \\ 0 & \text{otherwise} \end{cases}$$

Define g(y) fully.

iii $P(Y \geqslant k) = 1 - P(Y \leqslant k)$

Consider the cumulative probability.

$$= 1 - G(k)$$

$$1 - \frac{k^{\frac{2}{3}}}{12} + \frac{1}{3} = \frac{7}{12}$$

$$\frac{16}{12} - \frac{k^{\frac{2}{3}}}{12} = \frac{7}{12}$$

$$k^{\frac{2}{3}} = 9$$

$$k = 9^{\frac{3}{2}}$$

Solve for k.

$$k = 27$$

Checklist of learning and understanding

Probability density functions:

- For $f(x)$ to represent a probability density function, $f(x) \geqslant 0$ for all values of x.
- $\int_{\forall x} f(x)\,dx = 1$

Cumulative distribution functions:

- $F(x) = \int_{-\infty}^{x} f(t)\,dt$
- $\frac{dF(x)}{dx} = f(x)$

Expectation of functions of X:

- $E[g(X)] = \int_{\forall x} g(x) f(x)\,dx$

Finding the cumulative distribution function of a function of X:

- $G(y) = P(Y \leqslant y) = \begin{cases} P(X \leqslant h^{-1}(y)) = F(h^{-1}(y)) \\ \text{or} \\ P(X \geqslant h^{-1}(y)) = 1 - F(h^{-1}(y)) \end{cases}$

END-OF-CHAPTER REVIEW EXERCISE 8

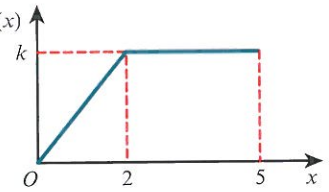
1 The time, T seconds, between successive cars passing a particular checkpoint on a wide road has probability density function f given by $\mathrm{f}(t) = \begin{cases} \frac{1}{100}e^{-0.01t} & t \geqslant 0 \\ 0 & \text{otherwise.} \end{cases}$

i State the expected value of T.

ii Find the median value of T.

iii Sally wishes to cross the road at this checkpoint and she needs 20 seconds to complete the crossing. She decides to start out immediately after a car passes. Find the probability that she will complete the crossing before the next car passes.

Cambridge International AS & A Level Further Mathematics 9231 Paper 21 Q7 November 2014

2 

The continuous random variable X takes values in the interval $0 \leqslant x \leqslant 5$ only. For $0 \leqslant x \leqslant 5$ the graph of its probability density function f consists of two straight line segments, as shown in the diagram.

a Find k and show that f is given by $\mathrm{f}(x) = \begin{cases} \frac{1}{8}x & 0 \leqslant x \leqslant 2 \\ \frac{1}{4} & 2 \leqslant x \leqslant 5 \\ 0 & \text{otherwise.} \end{cases}$

b The random variable Y is given by $Y = X^2$.

i Find the probability density function of Y.

ii Show that $\mathrm{E}(Y) = 10.25$.

iii Show that the median of Y is the square of the median of X.

Cambridge International AS & A Level Further Mathematics 9231 Paper 23 Q11 November 2012

3 The lifetime, in years, of an electrical component is the random variable T, with probability density function f given by

$$\mathrm{f}(t) = \begin{cases} A\mathrm{e}^{-\lambda t} & t \geqslant 0, \\ 0 & \text{otherwise,} \end{cases}$$

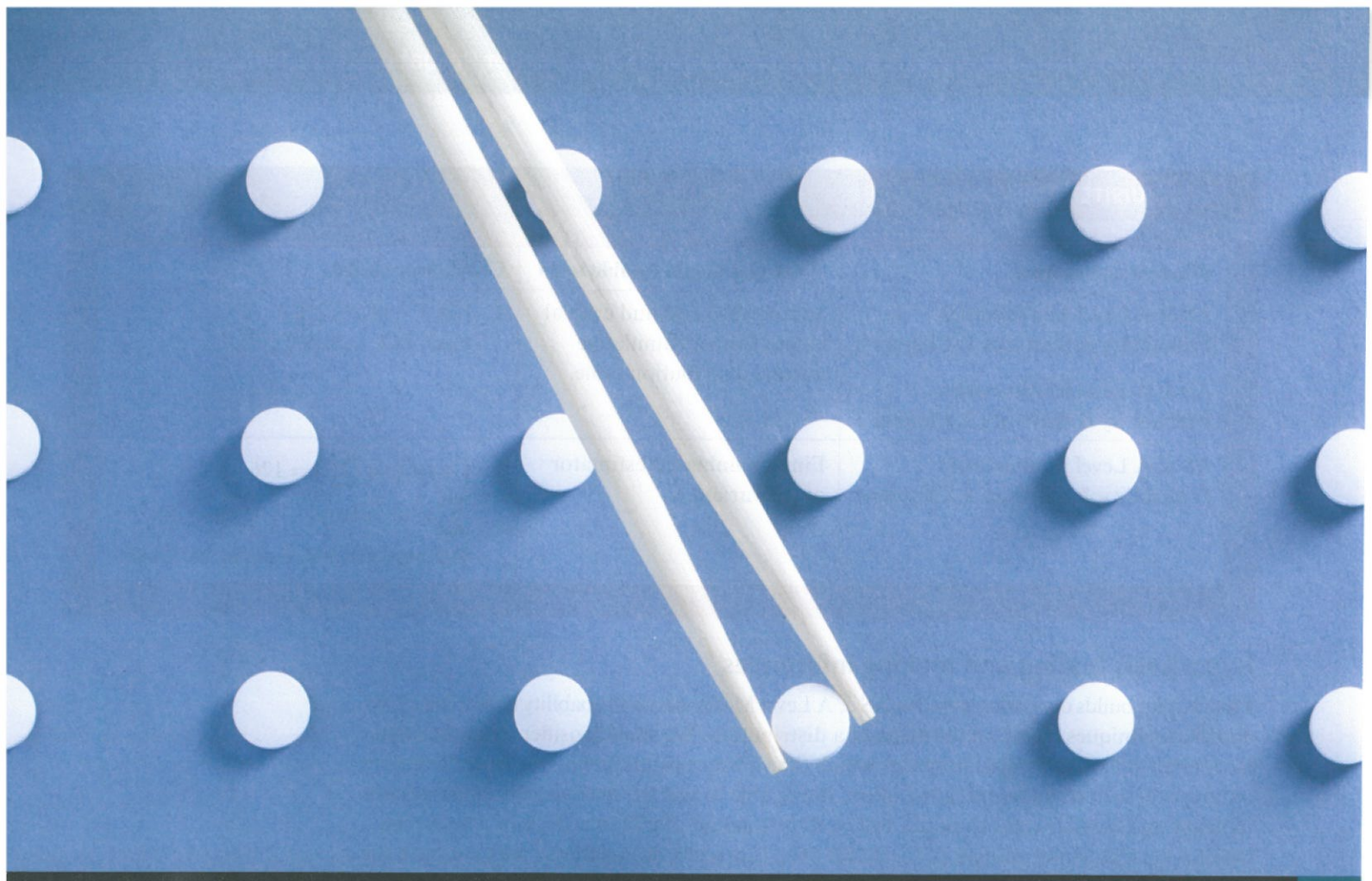
where A and λ are positive constants.

i Show that $A = \lambda$.

It is known that out of 100 randomly chosen components, 16 failed within the first year.

ii Find an estimate for the value of λ, and hence find an estimate for the median value of T.

Cambridge International AS & A Level Further Mathematics 9231 Paper 22 Q8 November 2013



Chapter 9
Inferential statistics

In this chapter you will learn how to:

- formulate and carry out a hypothesis test concerning the mean for a small sample, using the *t*-test
- calculate a pooled estimate of a population variance from two samples
- formulate and carry out a hypothesis test concerning the difference in means, using:
  - a two-sample *t*-test
  - a paired sample *t*-test
  - a test using the normal distribution
- determine a confidence interval for a population mean based on a small sample, using the *t*-distribution
- determine a confidence interval for the difference in population means.

PREREQUISITE KNOWLEDGE

Where it comes from	What you should be able to do	Check your skills
AS & A Level Mathematics Probability & Statistics 1, Chapter 8 AS & A Level Mathematics Probability & Statistics 2, Chapter 3	Standardise and find critical values from a cumulative normal distribution table.	1 Let $X \sim N(24, 2^2)$. Find $P(X \leqslant 26.34)$.
AS & A Level Mathematics Probability & Statistics 2, Chapter 6	Find an unbiased estimator of the variance.	2 Given that $\sum x = 126$, $\sum x^2 = 514$ and $n = 37$, find the unbiased estimator of the variance.

Hypothesis testing and making inferences

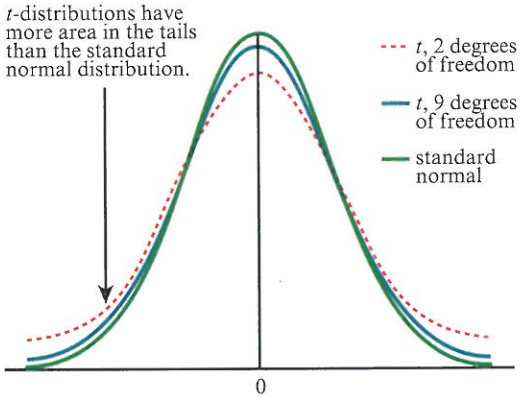
This chapter builds on work covered in AS & A Level Mathematics Probability & Statistics 2, to develop techniques based on the mean of a distribution. We shall consider situations that have small samples or populations for which the variance is unknown. We shall then carry out hypothesis tests concerning the mean. Being able to test the mean is very important in industry and medicine, for example, to test if the amount of effective drug in a headache tablet is correct. If not enough of the active drug is present, the medicine may not have the desired effect. We can also use this type of test to see whether the machine that is making the tablets is putting enough of the effective drug into the tablets.

9.1 *t*-distribution

When we collect a sample from a normal distribution to carry out a hypothesis test concerning the mean, we need to make some assumptions in order to use the normal distribution. We assume that the population variance is known, or that the sample size is sufficiently large that we can use s^2, the unbiased estimator of the variance, instead of the population variance, σ^2.

If the sample size is small, and we do not know what the variance is, then it is no longer appropriate to use the unbiased estimator. The *t*-distribution was developed so that the unbiased estimator could be used. It is a better model in this situation.

The diagram shows that the *t*-distribution is different from the normal distribution because the density in the tails is greater than the density in the tails of the standard normal distribution. This means that the *t*-distribution has greater probability in the tails and less probability in the centre compared to the standard normal distribution. The *t*-distribution is a family of distributions with $(n - 1)$ **degrees of freedom**. The number of degrees of freedom refers to the number of independent observations in a set of data. The number of degrees of freedom of the *t*-distribution relate to the sample size, *n*, and as the sample size increases, the *t*-distribution looks more like the normal distribution.



Hence, the distribution of the *t* statistic from samples of size 8 would be described by a *t*-distribution having 8 – 1 or 7 degrees of freedom. Similarly, a *t*-distribution having 15 degrees of freedom would be used

with a sample of size 16. As the degrees of freedom tend to infinity, the distribution becomes more like the normal distribution.

KEY POINT 9.1

Given n data points, the unbiased estimator of the variance, s^2, is calculated using:

$$s^2 = \frac{1}{n-1}\sum (x - \bar{x})^2$$

or

$$s^2 = \frac{1}{n-1}\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right) = \frac{1}{n-1}\left(\sum x^2 - n\bar{x}^2\right)$$

This can be written as $s^2 = \frac{S_{xx}}{n-1}$.

TIP

The unbiased estimator of the variance is also sometimes written as $\hat{\sigma}^2$. Generally, if θ is a parameter, then $\hat{\theta}$ is an estimator. In this book, we will use s^2.

Key point 9.1 gives a convenient way of calculating the unbiased estimator of the variance, s^2.

WORKED EXAMPLE 9.1

The wingspans of six Monarch butterflies are measured (in cm) and recorded as:
8.8, 9.6, 9.2, 9.1, 9.9, 8.7
Calculate the unbiased estimator for the variance of these data.

Answer

Method 1

$\sum x = 55.3$

First calculate the estimator of the mean, $\bar{x}$, using $\bar{x} = \frac{\sum x}{n}$.

$\bar{x} = \frac{55.3}{6} = 9.2166\ldots$

$x - \bar{x}$	$(x - \bar{x})^2$
−0.41667	0.173611
0.383333	0.146944
−0.01667	0.000278
−0.11667	0.013611
0.683333	0.466944
−0.51667	0.266944

Use the most accurate values when working to avoid rounding errors.

$\sum (x - \bar{x})^2 = 1.068333$

So $s^2 = \frac{1.068333}{6-1} = 0.214$ (to 3 significant figures)

Use $s^2 = \frac{1}{n-1}\sum (x - \bar{x})^2$

Method 2: Using summative data

This method helps with analysis of variance (ANOVA) which is studied at degree level.

$\sum x = 55.3$

$\sum x^2 = 510.75$

$$S_{xx} = \sum x^2 - \frac{(\Sigma x)^2}{n}$$
$$= 510.75 - \frac{55.3^2}{6}$$
$$= 1.068333$$

So $s^2 = \frac{1.068333}{6-1} = 0.214$ (to 3 significant figures)

We now have the tools to perform a hypothesis test concerning the population mean. This is where a small sample is taken from an underlying normal distribution with unknown variance. In other words, we assume the population is normally distributed.

There are several important steps when performing a hypothesis test for significance. First, we define the null hypothesis H_0. When dealing with a parametric test, H_0 is the assumed value of the parameter. It is this assumption that we are testing. Then we propose an alternative hypothesis H_1. There could be many different versions of this but, for the purposes of this course, it depends on whether we are looking at a one-tailed or a two-tailed test. Consequently, we refer only to H_0 and state whether we reject or do not reject it. Does rejecting H_0 mean that H_1 is accepted?

For an underlying normal distribution with unknown variance, a small sample of size n is taken. To carry out a hypothesis test on the mean, the t-distribution with $(n-1)$ degrees of freedom is used to find the critical value. In fact, $\overline{X} \sim t_{n-1}$. We state the null and alternative hypotheses as shown in Key point 9.2.

KEY POINT 9.2

For an underlying population that is normally distributed, with unknown variance, the null and alternative hypotheses will be:

H_0: $\mu = k$, where k is the assumed value of the mean that we are testing

H_1: $\begin{cases} \mu < k \\ \mu > k \\ \mu \neq k \end{cases}$ depending on whether we are performing a one-tailed or two-tailed test

The test statistic $= \frac{\overline{x} - \mu}{s/\sqrt{n}}$, where $\overline{x}$ is the sample mean and $s = \sqrt{\frac{S_{xx}}{n-1}}$ is the unbiased estimator of the standard deviation.

The critical value for a significance level $100(1-\alpha)\%$ is:

One-tailed	$t_{\alpha, n-1}$
Two-tailed	$t_{\frac{\alpha}{2}, n-1}$

WORKED EXAMPLE 9.2

Monarch butterflies are bred at a butterfly farm. Monarch butterflies should grow to have a mean wingspan of 9.4 cm. The breeders are concerned that their butterflies are not growing as well as they could be, so a sample of six Monarch butterflies is taken and their wingspans measured (in cm). These are recorded as: 8.8, 9.6, 9.2, 9.1, 9.9, 8.7. Assuming that the wingspans are normally distributed and using a 10% significance level, investigate whether the wingspans of the butterflies are less than 9.4 cm.

Answer

Important facts:
- assume underlying normal distribution
- variance is unknown
- sample size is small.

This suggests that we need a *t*-test.

We are testing whether the population mean is **below** 9.4.
This is a one-tailed test.
The significance level is 10%.

Make sure that you state whether you have a one- or two-tailed test.

$H_0: \mu = 9.4$, $H_1: \mu < 9.4$
Test statistics:

Define the null and alternative hypotheses. We must use parameters in the hypothesis test. This is the assumed value for μ when calculating the test statistic.

$\bar{x} = 9.22$ (3 significant figures) as found in Worked example 9.1

$s = 0.462$ (3 significant figures) as found in Worked example 9.1

Note $s^2 = 0.214$ was found earlier, so $s = 0.462$

$$\text{Test statistic} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$
$$= \frac{9.22 - 9.4}{0.462/\sqrt{6}}$$
$$= -0.9715$$

Use $\frac{\bar{x} - \mu}{s/\sqrt{n}}$ and make the critical value positive or negative accordingly.

The critical value for this test is $t_{0.9,5} = -1.476$.

We look at the *t*-distribution table. Since the test is one-tailed, at the 10% significance level we need the 90th percentile with 5 degrees of freedom.

Our test statistic is negative, so we need to consider the negative value for *t* too.

p	0.75	0.90	0.95
$v = 1$	1.000	3.078	6.314
2	0.816	1.886	2.920
3	0.765	1.638	2.353
4	0.741	1.533	2.132
5	0.727	1.476	2.015
6	0.718	1.440	1.943
7	0.711	1.415	1.895
8	0.706	1.397	1.860

Now we need to decide whether to reject H_0 or not.

Since $-0.9715 > -1.476$ this means the test statistic is not in the critical region, so we do not reject H_0.

Consider H_0, the assumption we made. Note we do not state we accept H_1 but instead state we do not reject H_0.

There is insufficient evidence to suggest that the mean wingspan of the Monarch butterflies is less than 9.4 cm.

Write a conclusion in context.

Worked example 9.3 demonstrates a two-tailed test.

WORKED EXAMPLE 9.3

A random sample of 12 workers from a mobile phone assembly line is selected from a large number of workers. A manager asks each of these workers to assemble a phone at their normal working speed. The times taken, in minutes, to complete these tasks are recorded below.

43.2, 41.6, 49.3, 48.2, 44.2, 40.6, 39.7, 43.4, 44.9, 45.1, 46.2, 43.2

Assuming that this sample comes from an underlying normal population, investigate the claim that the population mean is 45 minutes. Use a 5% significance level.

Answer

Important facts:
- underlying normal distribution assumed
- variance is unknown
- sample size is small.

This suggests that we need a t-test.

Test whether the population mean is different from 45.
This is a two-tailed test.
The significance level is 5%.

Decide whether this is a one- or two-tailed test.

$H_0: \mu = 45$
$H_1: \mu \neq 45$

Define the null and alternative hypotheses. Use parameters in the hypothesis test. This is the assumed value for μ when calculating the test statistic.

Test statistic:

$\Sigma x = 529.6$
$\Sigma x^2 = 23462.88$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

$$S_{xx} = 23462.88 - \frac{529.6^2}{12} = 89.8666\ldots$$

$$s = \sqrt{\frac{89.8666\ldots}{11}} = 2.858268\ldots$$

Use $s^2 = \frac{S_{xx}}{n-1}$, where $S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{529.6}{12} = 44.133\ldots$$

$s = 2.86$ (to 3 significant figures)

$\bar{x} = 44.1$ (to 3 significant figures)

$$\text{Test statistic} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{44.1 - 45}{2.86/\sqrt{12}} = -1.05$$

Always use $\frac{\bar{x} - \mu}{s/\sqrt{n}}$ and make the critical value positive or negative accordingly.

The critical value for this test is $t_{0.975,\,11} = -2.201$.

p	0.75	0.90	0.95	0.975
$\nu = 1$	1.000	3.078	6.314	12.71
2	0.186	1.886	2.920	4.303
3	0.765	1.638	2.353	3.182
4	0.741	1.533	2.132	2.776
5	0.727	1.476	2.015	2.571
6	0.718	1.440	1.943	2.447
7	0.711	1.415	1.895	2.365
8	0.706	1.397	1.860	2.306
9	0.703	1.383	1.833	2.262
10	0.700	1.372	1.812	2.228
11	0.697	1.363	1.796	2.201
12	0.695	1.356	1.782	2.179

Since this is a two-tailed test at the 5% significance level, look at the 97.5th percentile with 11 degrees of freedom, in the t-distribution table.

See if the test statistic is in the critical region. In this case, this will be when the test statistic is less than the critical value, if negative, or greater than the critical value, if positive. Since the test statistic is negative, we will need to compare it with the negative critical value.

Since $-1.05 > -2.201$, the test statistic is not in the critical region and hence we do not reject H_0.

It is best to refer to H_0.

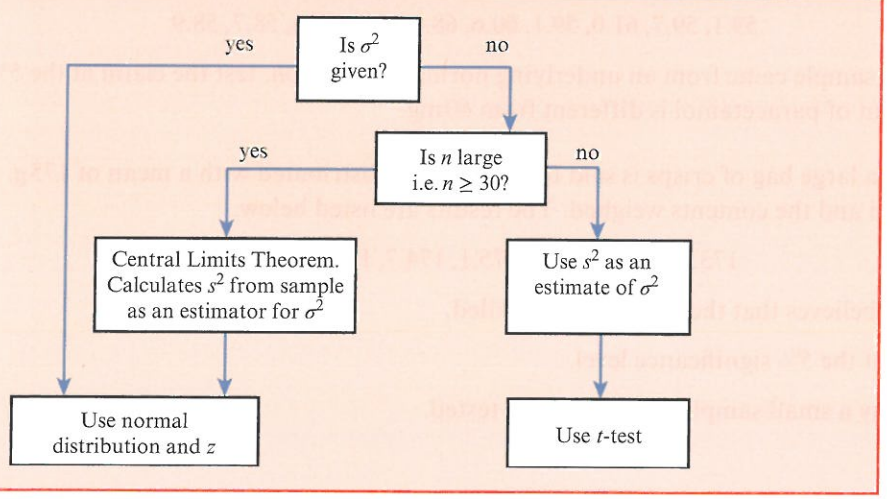
There is insufficient evidence to claim that the population mean is not 45 minutes.

We should always write full conclusions in context. Note we should not state there is sufficient evidence the population mean is 45 minutes, but instead state there is insufficient evidence the population mean is not 45 minutes.

It is very important to know which test to perform when testing the mean. In AS & A Level Mathematics Probability & Statistics 2, Chapter 5, you carried out a hypothesis test concerning the mean. In that case, the underlying distribution was known to be normal, or the sample size was large enough to use the central limit theorem. You were also given the population variance. In this chapter, you have carried out hypothesis tests for small samples or where the population variance was unknown. It is very important that you choose the most appropriate test to carry out.

The flowchart shown in Key point 9.3 can help you to decide which is the most appropriate test to use.

KEY POINT 9.3



FAST FORWARD

If the underlying distribution is not normal, and the sample size is small, we can use non-parametric tests, as shown in Chapter 11.

EXERCISE 9A

1 For the given data, find unbiased estimates for the mean and variance.

a 12, 16, 17, 19, 13, 14, 11, 16, 19, 21, 14, 15

b 143, 154, 156, 145, 144, 132, 135, 148, 171, 124

2 In each case, state the magnitude of the test statistic for the given value of n and stated significance level.

a $n = 11$, one-tailed 5%

b $n = 21$, one-tailed 2.5%

c $n = 15$, two-tailed 5%

d $n = 25$, two-tailed 1%

e $n = 8$, one-tailed 10%

f $n = 18$, two-tailed 10%

3 State the null and alternative hypotheses for the following tests.

a The population mean differs from 41.

b The population mean is greater than 7.3.

c The population mean has decreased from 54.2.

d The population mean has not changed from 6.5.

4 For the given test statistic, sample sizes and significance levels, state whether you would reject or not reject the null hypothesis.

a Test statistic = 1.96, $n = 10$, 5% one-tailed

b Test statistic = –2.764, $n = 8$, 1% one-tailed

c Test statistic = 1.451, $n = 15$, 10% two-tailed

d Test statistic = –2.341, $n = 11$, 5% two-tailed

5 In a given week, 12 babies are born in hospital. Assume that this sample came from an underlying normal population. The length of each baby is routinely measured and is listed below (in cm):

49, 50, 45, 51, 47, 49, 48, 54, 53, 55, 45, 50

a Find unbiased estimators for the mean and variance.

The average length of babies is thought to be 50.5 cm. There is a concern that this is an overestimate.

b Test this claim at the 5% significance level based on this sample.

6 A drugs manufacturer claims that the amount of paracetamol in tablets is 60 mg. A sample of ten tablets is taken and the amount of paracetamol in each is recorded:

59.1, 59.7, 61.0, 59.1, 60.6, 68.9, 60.2, 58.6, 58.7, 58.9

Assuming that this sample came from an underlying normal population, test the claim at the 5% significance level that the amount of paracetamol is different from 60 mg.

7 The weight, X g, of a large bag of crisps is said to be normally distributed with a mean of 175 g. A sample of eight bags is opened and the contents weighed. The results are listed below.

173.2, 171.5, 176.3, 175.1, 174.7, 174.2, 176, 174.5

A consumer group believes that the bags are underfilled.

a Test this claim at the 5% significance level.

b Suggest why only a small sample of packets was tested.

8 At a petrol station, the manager thinks that one of the pumps is not working properly and is giving out more petrol than it should. She decides to test this claim by filling up ten buckets with 5 litres, according to the pump. The results, in cm^3, are given below (1 litre $= 1000\ cm^3$).

It is assumed that the amounts given are from a normal distribution.

5001, 5002, 5009, 4996, 4997, 5001, 5003, 5006, 5013, 5013

Using this sample, test at the 5% significance level whether or not the petrol pump is giving out too much petrol.

9.2 Hypothesis tests concerning the difference in means

We may test whether two populations have equal population means. In Section 9.1 we learned that the type of test we need to perform depends on a number of factors. These factors include: whether or not we know the population variances; whether the underlying distribution is normal, or the means approximate to a normal distribution by the application of the central limit theorem; and the size of the sample.

To start with, we will assume three things:

- the underlying distributions are normal
- the populations are independent
- the population variance of the two populations is the same (but may be unknown).

TIP

The second assumption can be tested using an F-test, but this is beyond the scope of this course.

From AS & A Level Mathematics Probability & Statistics 2, Chapter 4:

If $X \sim N(\mu, \sigma^2)$, then $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

If $X \sim N(\mu_x, \sigma_x^2)$ is independent of $Y \sim N(\mu_y, \sigma_y^2)$, then $X - Y \sim N(\mu_x - \mu_y, \sigma_x^2 + \sigma^2{}_y)$ and

$\overline{X} - \overline{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$.

We standardise to find a z-value, as shown in Key point 9.4. This acts as the test statistic for the difference in means.

KEY POINT 9.4

z-value is given by $Z = \dfrac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \sim N(0, 1)$.

WORKED EXAMPLE 9.4

A group of 50 children and 70 adults participate in a maths activity. The mean time taken for the children to complete the activity is 45.3 seconds, with a standard deviation of 3.2 seconds. For the adults, the mean time is 46.1 seconds with a standard deviation of 2.8 seconds. Assuming the completion times are normally distributed with equal variances, test at the 5% significance level whether or not the children are faster at completing the activity.

Answer

First consider the conditions and assumptions:
- n is large for both populations
- both have an underlying mean
- the variances are equal but unknown.

Estimate σ^2 with s_x^2 and s_y^2.

Let $C \sim N(45.3, 3.2^2)$ represent the population of children.

Let $A \sim N(46.1, 2.8^2)$ represent the population of adults.

If the children are faster than the adults, then $\mu_A > \mu_C$.

Rewrite this as $\mu_A - \mu_C > 0$.

$H_0: \mu_A - \mu_C = 0$
$H_1: \mu_A - \mu_C > 0$

You could have this the other way round, but a positive value for H_1 will reduce errors in interpretation.

Test statistic $= \dfrac{(\overline{A} - \overline{C}) - (\mu_A - \mu_C)}{\sqrt{\dfrac{s_A^2}{n_A} + \dfrac{s_C^2}{n_C}}}$

Estimate σ^2 with s_x^2 and s_y^2 under the assumption of $H_0: \mu_A - \mu_C = 0$.

$$= \frac{(46.1 - 45.3) - 0}{\sqrt{\dfrac{2.8^2}{70} + \dfrac{3.2^2}{50}}} = 1.4213\ldots$$

This is a one-tailed test at the 5% significance level so the critical value is 1.96.
Since $1.4213 < 1.96$, the test statistic is not in the critical region and we should not reject H_0.

Always discuss H_0, the assumption you made.

There is insufficient evidence to suggest that the children perform the maths activity faster than the adults.

Always write full contextualised conclusions.

FAST FORWARD

In Worked example 9.4, we are given data that is normally distributed. We are given large samples and can therefore use s^2 in place of σ^2. In other situations we must use different techniques, as shown in Section 9.3.

Sometimes we need to carry out two-sample tests, such as comparing the mean of two distributions of unknown but equal variances. If the sample sizes are too small to allow us to use s_x^2 and s_y^2 as estimators, we need to pool these variances (combine them).

Consider two samples of size n_x and n_y. From underlying normal distributions the pooled estimate of the population variance is:

$$s_p^2 = \frac{\sum (x - \overline{x})^2 + \sum (y - \overline{y})^2}{n_x + n_y - 2}$$

However, if you are given the unbiased estimators of the variance for each sample,

$$s_x^2 = \frac{\sum (x - \overline{x})^2}{n_x - 1} \text{ and } s_y^2 = \frac{\sum (y - \overline{y})^2}{n_y - 1},$$

then $\sum (x - \overline{x})^2 + \sum (y - \overline{y})^2 = (n_x - 1)s_x^2 + (n_y - 1)s_y^2$.

Then the pooled estimate of the population variance, as shown in Key point 9.5, is $s_p^2 = \dfrac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$.

But how small should n be? Generally, we would need $n < 15$.

KEY POINT 9.5

For two samples of size n_x and n_y the pooled estimate of the population variance is:

$$s_p^2 = \frac{\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2}{n_x + n_y - 2}$$

If you know the unbiased estimators of the variance for each sample, $s_x^2 = \dfrac{\sum (x - \bar{x})^2}{n_x - 1}$ and $s_y^2 = \dfrac{\sum (y - \bar{y})^2}{n_y - 1}$, then $\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2 = (n_x - 1)s_x^2 + (n_y - 1)s_y^2$.

So the pooled estimate of the population variance is $s_p^2 = \dfrac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$.

Let's consider how Key point 9.5 affects the test statistic:

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\dfrac{\sigma_x^2}{n_x} + \dfrac{\sigma_y^2}{n_y}}}$$

If we make the assumption that the variances are equal and that the pooled estimator can be used, $\sigma_x^2 = \sigma^2{}_y = s_p^2$, we have:

$$\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y} = s_p^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)$$

Since n is small, we know that this will be modelled as a t-distribution, as shown in Key point 9.6.

KEY POINT 9.6

t-distribution:

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{s_p^2\left(\dfrac{1}{n_x} + \dfrac{1}{n_y}\right)}} \sim t_{n_x + n_y - 2}$$

WORKED EXAMPLE 9.5

A shopkeeper believes that playing music in his shop encourages customers to spend more money. To test this belief, he records how much money is collected for a ten-day period while music is playing and then for an eight-day period without music. The sales, in thousands of dollars, are summarised as follows.

With music	$\sum x = 960.1$	$\sum x^2 = 92274.44$
Without music	$\sum y = 748.2$	$\sum y^2 = 70041.16$

Assuming these data are randomly sampled from normal distributions with the same variance, test the shopkeeper's claim, using a 5% significance level.

Answer

First consider the assumptions and conditions:
- n is small for both populations
- both have an underlying mean
- the variances are equal but unknown.

Since n is small and the variances are unknown, we must use a t-test and hence the pooled estimator.

Let X represent the sales with music and Y represent the sales without music.

Define the variables to be used.

If the sales with music are greater than without, then $\mu_x > \mu_y$.

Rewrite this as $\mu_x - \mu_y > 0$.

$H_0: \mu_x - \mu_y = 0$

$H_1: \mu_x - \mu_y > 0$

The test statistic is:

$$\frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{s_p^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$$

If the sample variances are given, then $s_p^2 = \dfrac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$ is used.

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n_x} = 92274.44 - \frac{960.1^2}{10} = 95.239$$

$$S_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n_y} = 70041.16 - \frac{748.2^2}{8} = 65.755$$

$$s_p^2 = \frac{S_{xx} + S_{yy}}{n_x + n_y - 2} = \frac{95.239 + 65.755}{10 + 8 - 2} = 10.062125$$

Calculate s_p^2.

We can use the form:

$$s_p^2 = \frac{S_{xx} + S_{yy}}{n_x + n_y - 2}$$

$$\overline{x} = \frac{\Sigma x}{n_x} = \frac{960.1}{10} = 96.01$$

$$\overline{y} = \frac{\Sigma y}{n_y} = \frac{748.2}{8} = 93.525$$

Calculate $\overline{x}$ and $\overline{y}$.

Test statistic:

$$\frac{(96.01 - 93.525) - 0}{\sqrt{10.062125\left(\frac{1}{10} + \frac{1}{8}\right)}} = 1.65154$$

The critical value is $t_{0.95,\,16} = 1.746$.

We have $10 + 8 - 2 = 16$ degrees of freedom here.

p	0.75	0.90	0.95
10	0.700	1.372	1.812
11	0.697	1.363	1.796
12	0.695	1.356	1.782
13	0.694	1.350	1.771
14	0.692	1.345	1.761
15	0.691	1.341	1.753
16	0.690	1.337	1.746
17	0.689	1.333	1.740

Since $1.65154 < 1.746$, the test statistic is not in the critical region so we do not reject H_0.

Always refer to H_0, the assumption you made.

There is insufficient evidence to suggest that playing music increases the sales in the shop.

Write a full contextualised conclusion.

EXERCISE 9B

1 Given the sample variance and sample size of each set of data, find the pooled estimate of variance.

a $s_x^2 = 13.2 \quad n_x = 15$

$s_y^2 = 11.9 \quad n_y = 13$

b $s_x^2 = 161.2 \quad n_x = 21$

$s_y^2 = 158.7 \quad n_y = 24$

c $s_x^2 = 32.1 \quad n_x = 60$

$s_y^2 = 48.6 \quad n_y = 40$

2 For the following pairs of data sets, find an estimate for the pooled variance.

a

X	27	19	15	19	21
	18	17	16	20	28

Y	32	31	27	26
	29	30	28	14

b

X	23	25	26	19	22	21
	28	25	26	19	23	

Y	14	19	21	20	17
	16	18	15	21	

3 For each question, state the magnitude of the test statistic for the given values of n_x and n_y and stated significance level.

a $n_x = 8, n_y = 6$, one-tailed 5%

b $n_x = 14, n_y = 10$, one-tailed 2.5%

c $n_x = 8, n_y = 7$, two-tailed 5%

d $n_x = 20, n_y = 12$, two-tailed 1%

e $n_x = 11, n_y = 14$, one-tailed 10%

f $n_x = 17, n_y = 12$, two-tailed 10%

4 For each of the following, state the null and alternate hypotheses.

a The difference in population means is not 0.

b The population mean for X is greater than the population mean for Y.

c The population mean for X is five units greater than the population mean for Y.

d The difference in the population means is not six units.

M 5 Two examiners are marking an examination paper, and it is believed that examiner A is more strict than examiner B. The results from several papers are added together for each examiner, and presented in the following table.

	Sample size	Sum of marks
Examiner A	16	689
Examiner B	12	636

Test the claim at the 5% significance level, assuming that the marks are normally distributed with a standard deviation of 15.

M 6 Takahē birds are native to New Zealand and are very rare. The male birds and female birds look very similar. The only way of differentiating males from females is to measure their weights. It is known that the female bird is slightly smaller than the male, and so weighing them could be a way of identifying the gender of an adult Takahē bird.

The weights of ten male and eight female Takahē birds are measured, and the summative statistics are presented in the following table.

	Sample size	Sum	Sum of squares
Male	10	28.1	79.4
Female	8	21.5	58.3

a Find s_p^2, the pooled estimator of the population variance.

b Test, at the 5% significance level, whether male Takahē birds are heavier than female Takahē birds assuming the weights are normally distributed.

M 7 A company that makes computers must transport them from its warehouse to the delivery centre, with one lorry delivery per day. In three weeks' time, the usual route will have roadworks stopping the traffic for six weeks. The local council says that the alternative route will add not more than ten minutes to the route. The manager of the company does not think that this is true and so, for the next 14 days, he asks eight of the company's lorry drivers to travel the new route, and six to travel the old route.

Old	34	45	36	47	42	43		
New	47	51	47	50	53	51	50	45

a Find s_p^2, the pooled estimate of the population variance.

b Test, at the 5% significance level, whether the manager is justified in his complaint assuming the times are normally distributed.

M 8 Samples are taken from two different types of honey and the viscosity (i.e. how 'runny' the honey is) is measured.

Honey	Mean	Standard deviation	Sample size
A	114.44	0.62	4
B	114.93	0.94	6

Assuming normal distributions, test at the 5% significance level whether there is a difference in the viscosity of the two types of honey.

9.3 Paired *t*-tests

In Section 9.2, we looked at whether or not two samples with the same variances have the same mean. We looked at the difference in means, sometimes referred to as an 'unpaired test', since there is no mechanism to 'pair' the data values.

If we need to measure the effect of a variable on a set of data, then we measure twice: before the change in variable, and after. This repeated measures design allows us to pair the points in the two datasets. In this situation, a paired *t*-test would be the most appropriate test to perform. Instead of measuring the difference in the means, we measure the mean of the differences, as shown in Key point 9.7.

TIP

In Social Sciences, the unpaired test is also known as an independent samples design.

KEY POINT 9.7

For a paired *t*-test, with n pairs of data, and H_0: $\mu_d = k$ (typically $\mu_d = 0$)

the test statistic $= \dfrac{\bar{d} - k}{\dfrac{s_d}{\sqrt{n}}}$

where $d_i = x_i - y_i$, and $\bar{d}$ and s_d are the sample mean and standard deviation of $D \sim N\left(\mu_d, \dfrac{s_d^2}{n}\right)$.

We test using a *t*-distribution with $(n - 1)$ degrees of freedom.

The only assumption for this test is that the difference is approximately normally distributed. As a consequence, if your dataset has outliers, this test is not appropriate. Worked example 9.6 works through a paired *t*-test.

WORKED EXAMPLE 9.6

A diagnostic test is taken by ten students before a revision session, and then again after completing the revision session. Their scores are presented in the following table.

Student	A	B	C	D	E	F	G	H	I	J
Before	45	54	49	51	53	64	71	55	78	43
After	49	56	56	54	61	70	72	60	82	51

Using a paired *t*-test, and assuming the differences in scores are normally distributed, test at the 5% significance level whether the revision was effective.

Answer

	Before	After	Difference
A	45	49	4
B	54	56	2
C	49	56	7
D	51	54	3
E	53	61	8
F	64	70	6
G	71	72	1
H	55	60	5
I	78	82	4
J	43	51	8

First, calculate all of the differences so we can calculate s_d and $\bar{d}$.

$\Sigma d = 48$

$\Sigma d^2 = 284$

$\bar{d} = \frac{\Sigma d}{n} = \frac{48}{10} = 4.8$

$S_{dd} = \Sigma d^2 - \frac{(\Sigma d)^2}{n}$ — Find the summative values.

$S_{dd} = 284 - \frac{48^2}{10} = 53.6$

$s_d^2 = \frac{S_{dd}}{n-1} = \frac{53.6}{9} = 5.96$ (to 3 significant figures)

$\bar{d} = 4.8$ — Now find the unbiased estimates.

$s_d^2 = 5.96$ (to 3 significant figures)

Carrying out the test, let $\bar{D} \approx N\left(\mu_d, \frac{s_d^2}{n}\right)$. — We can now perform our hypothesis test. We require an approximately normal distribution.

$H_0: \mu_d = 0$

$H_1: \mu_d > 0$ — If the revision has been effective, then we would anticipate the difference to increase, and so the test is one-tailed test.

$$\text{Test statistic} = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{4.8}{\sqrt{\frac{5.96}{10}}} = 6.22$$

The critical value is $t_{0.95,9} = 1.833$. — The critical value at the 5% significance level must be found.

Since the test statistic is greater than the critical value, the test statistic is in the critical region. So we reject H_0 as there is sufficient evidence to suggest that the difference in scores has increased and so the revision has been effective. — A well-formed conclusion is required. We reject the null hypothesis and state there is sufficient evidence to suggest the revision has been effective.

In Worked example 9.6 we are measuring whether there is a difference in the mean. Worked example 9.7 considers whether the difference in the mean is greater than, or less than, a certain value.

WORKED EXAMPLE 9.7

This uses the same scenario as in Worked example 9.6.

Student	A	B	C	D	E	F	G	H	I	J
Before	45	54	49	51	53	64	71	55	78	43
After	49	56	56	54	61	70	72	60	82	51

Using a paired t-test, and assuming the differences in scores are normally distributed, test at the 5% significance level whether students have increased their scores by four marks or more.

Answer

$\overline{d} = 4.8$
$s_d^2 = 5.96$ (to 3 significant figures)

Use the same estimators as before.

Let $\overline{D} \approx N\left(\mu_d, \frac{s_d^2}{n}\right)$.

$H_0: \mu_d = 4$

$H_1: \mu_d > 4$

Test statistic $= \frac{\overline{d} - 4}{\frac{s_d}{\sqrt{n}}} = \frac{0.8}{\sqrt{\frac{5.96}{10}}} = 1.04$

Carry out the test.

The critical value is $t_{0.95, 9} = 1.833$.

We need the critical value at the 5% significance level.

Since the test statistic is less than the critical value, the test statistic is not in the critical region. We do not reject H_0. There is insufficient evidence to suggest that the difference in scores has increased by four or more.

A well-formed conclusion is required.

EXERCISE 9C

1 For each of the following pairs of data, find the sample mean of the difference, $\overline{d}$, and the unbiased estimator of the variance of the distance, s_d^2.

a

X	124	139	128	119	119	112	113	128	113
Y	127	117	121	126	119	125	118	118	127

b

X	34.3	30.8	32.8	27.5	26.3	27.8	35.1	31.1
Y	27.3	28.5	30.5	29	28	33	31.6	28

X	28.5	31.5	30.7	29.5
Y	32.8	28.1	30.9	30

c

X	75	84	80	66	78	97	68	86
Y	81	81	86	90	87	76	88	89

X	86	73	97	70	72
Y	84	92	85	90	91

2 For the given null hypotheses, find the test statistic of the following summative data.

a $H_0: \mu_d = 0$
$\mu_d = 0.344, s_d^2 = 121, n = 11$

b $H_0: \mu_d = 0$
$\mu_d = -0.688, s_d^2 = 11.62, n = 10$

c $H_0: \mu_d = -5$
$\mu_d = -5.82, s_d^2 = 182.3, n = 12$

3 For the data in question **2a–c**, state the magnitude of critical value, given the following alternate hypothesis and significance level.

a $H_1: \mu_d \neq 0$, significance level 5%

b $H_1: \mu_d > 0$, significance level 5%

c $H_1: \mu_d < -5$, significance level 2.5%

M 4 A biologist investigates the effect of a new food on Takahē male birds. Eight birds are weighed (in kg). They are then fed the new food for 14 days and weighed again.

Let us assume that the weight gains are normally distributed.

Initial weight (kg)	2.67	2.93	3.12	3.21	2.64	2.73	2.86	2.91
Weight after 14 days (kg)	2.71	3.01	3.19	3.24	2.6	2.78	2.84	2.97

Test, at the 2.5% significance level, to investigate whether there has been a significant increase in the weight of the Takahē male birds.

M 5 A diet programme aims at trying to help people lose at least 2 kg in weight within five weeks of starting the programme. A sample of eight participants are asked to volunteer to take part in the experiment and their weight at the beginning of the programme and after five weeks is measured and recorded. The following estimators were calculated: $\bar{d} = 2.225$; $s_d^2 = 0.931589$. Test, at the 5% significance level, the claim that participants will lose at least 2 kg of weight within the first five weeks of the programme, assuming the weight losses are normally distributed.

M 6 Police trainees are given a test to assess how good their memory is. After seeing ten car plates for 15 seconds each, they must write down as many as they can remember. The trainees then attend a memory improvement course. After this week-long course, they are retested. The results of the tests for eight police trainees are presented in the following table.

Number correct before course	6	5	6	5	7	5	4	6
Number correct after course	6	8	6	7	9	8	9	6

Test, at the 5% significance level, whether the course has made a difference to the trainees' scores, assuming the differences in scores are normally distributed.

M 7 A company sends its employees to a psychologist to try to improve their sales productivity. The following table shows the sales figures, in thousands of dollars, of six employees before and after seeing the psychologist.

	A	B	C	D	E	F
Before	10	8	15	38	60	90
After	14	9	16	42	80	83

Test, at the 5% significance level, whether the visits to the psychologist have improved sales productivity, assuming the increases in sales are normally distributed.

9.4 Confidence intervals for the mean of a small sample

Confidence intervals are another useful tool in statistical inference.

In a hypothesis test, we are interested in finding the critical region for the test. A confidence interval can be thought of as the **acceptance region** instead.

Confidence intervals are commonly misinterpreted so it is important to know the following concepts:

- A confidence interval is created from a sample taken.
- If another sample is taken, then a different confidence interval will be found.
- The population mean, although unknown, is fixed and so we assess whether the confidence interval contains the population mean.

If $\bar{x}$ is the mean of a random sample of size n from a normal distribution with population mean μ, and unbiased estimator of the variance s^2, a $100(\alpha - 1)\%$ confidence interval for μ is as shown in Key point 9.8.

KEY POINT 9.8

A $100(\alpha - 1)\%$ confidence interval for μ is given by:

$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1}\frac{s}{\sqrt{n}}$$

Consider that $\bar{X} \sim N(\mu, \sigma^2)$ and that $\bar{x}$, the sample mean and s^2 the unbiased estimator of the variance are calculated with a small sample size.

Let us consider performing a hypothesis test at the 90% significance level. The critical values for this test would be $\pm t_{0.95, n-1}$.

We could now unstandardise the critical values to calculate the limits for μ which would lead us to not reject the null hypothesis.

$$\pm t_{0.95, n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$\pm t_{0.95, n-1}\frac{s}{\sqrt{n}} = \bar{x} - \mu$$

$$\mu = \bar{x} \pm t_{0.95, n-1}\frac{s}{\sqrt{n}}$$

And so

$$\bar{x} - t_{0.95, n-1}\frac{s}{\sqrt{n}} \leqslant \mu \leqslant \bar{x} + t_{0.95, n-1}\frac{s}{\sqrt{n}}$$

is the acceptance region for this hypothesis test. We can also call this the 90% confidence interval for μ.

It is written as:

$$\left(\bar{x} - t_{0.95, n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{0.95, n-1}\frac{s}{\sqrt{n}}\right)$$

More generally,

Let us consider performing a hypothesis test at the $100(\alpha - 1)\%$ significance level. The critical values for this test would be $\pm t_{\frac{\alpha}{2}, n-1}$.

We could now unstandardise the critical values to calculate the limits for μ which would lead us to not reject the null hypothesis.

$$\pm t_{\frac{\alpha}{2}, n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$\pm t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}} = \bar{x} - \mu$$

$$\mu = \bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$$

And so

$$\bar{x} - t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}} \leqslant \mu \leqslant \bar{x} + t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$$

is the acceptance region for this hypothesis test. We can also call this the $100(\alpha - 1)\%$ confidence interval for μ.

It is written as:

$$\left(\bar{x} - t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}} \right)$$

WORKED EXAMPLE 9.8

A random sample of people queueing for a train ticket are asked how long they have been waiting in the queue before buying their ticket. Their replies, in minutes, are 12, 17, 21, 9, 14, 19.

a Assuming a normal distribution, calculate a 90% confidence interval for the mean stated waiting time.
b Comment on the train company's claim that the mean waiting time is ten minutes.

Answer

a Since we have a small sample, and the population variance is unknown, we must consider a t-distribution.

$\Sigma x = 92$
$\Sigma x^2 = 1512$

Calculate the unbiased estimators of the mean and variance.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{92}{6} = 15.333$$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 1512 - \frac{92^2}{6} = 101.333\ldots$$

$$s^2 = \frac{S_{xx}}{n-1} = \frac{101.33\ldots}{5} = 20.267$$

We are creating a 90% confidence interval, so:

$100(\alpha - 1)\% = 90\%$

Therefore, $\frac{\alpha}{2} = 0.95$

The confidence interval required is:

$\bar{x} \pm t_{0.95,\,5} \frac{s_x}{\sqrt{n}}$

The interval can be stated each time – it does not need to be derived.

From tables, $t_{0.95,\,5} = 2.015$.

Use the correct p-value when reading from the t-tables.

And so the confidence interval is:

$15.33 \pm 2.015 \frac{\sqrt{20.267}}{\sqrt{6}}$

And the 90% confidence interval is (11.627, 19.033).

Write the confidence interval like this.

b The confidence interval (CI) does not contain the claimed value of ten minutes. In fact, the confidence interval is wholly above the claimed value.

This means that the train company is underestimating the waiting time in the queue.

EXERCISE 9D

1 For the following confidence intervals, find the value from the t-distribution that must be used.

a 90% confidence interval, $n = 6$

b 90% confidence interval, $n = 8$

c 95% confidence interval, $n = 12$

d 80% confidence interval, $n = 7$

2 For the given sample sizes and values of s_x^2, find the value of the standard error $\left(\frac{s}{\sqrt{n}}\right)$.

a $n = 8 \quad s_x^2 = 9$

b $n = 6 \quad s_x^2 = 12$

c $n = 9 \quad s_x^2 = 22$

d $n = 5 \quad s_x^2 = 4.2$

3 Given that the data comes from an underlying normal distribution, and that $\bar{x} = 13.2$, $s_x^2 = 18$, find the confidence interval for the sample size stated.

a 90% confidence interval, $n = 8$

b 90% confidence interval, $n = 6$

c 90% confidence interval, $n = 9$

d 95% confidence interval, $n = 7$

e 99% confidence interval, $n = 9$

f 80% confidence interval, $n = 10$

4 For the following dataset, find a 95% confidence interval.

12, 15, 16, 18, 17, 15

Write each end of the interval to 2 decimal places.

M 5 A car rental company claims that, on average, its class C-type car will use 7.1 litres of fuel per 100 km when travelling at 60 km/h. Seven class C cars are tested on a test track and their fuel use over 100 km is measured.

Fuel usage	7.236	7.113	7.098	7.198	7.143	7.151	7.132

a Assuming a normal distribution, find a 95% confidence interval for the mean amount of fuel used.

b Comment on the claim by the car rental company that its class C cars use 7.1 litres of fuel per 100 km.

M 6 While on holiday, Yushan likes to stay in youth hostels. The company that owns the hostels claims that the average price of a night's stay is \$43. Yushan spends one night each in six different hostels. The prices that she pays are:

\$46, \$46, \$48, \$42, \$40, \$38

Calculate a 90% confidence interval for these data, assuming a normal distribution for the prices paid.

7 The contents of jars of beans may be assumed to be normally distributed. The contents, in grams, of a random sample of nine jars are as follows.

460, 449, 458, 455, 461, 456, 459, 457, 453

a Calculate a 95% confidence interval for these data.

b The jar has 'Contains 454 g' written on the label. Comment on this claim based on your calculated confidence interval.

8 The waiting time for a particular train that runs daily is measured over 36 days. The average waiting time is found to be 37.2 minutes, with a standard deviation of 3.2 minutes. Find a 99% confidence interval for the waiting times for the train, assuming the waiting times are normally distributed.

9.5 Confidence intervals for the difference in means

In Section 9.2, we considered setting up a hypothesis test for the difference in means with the following assumptions:

- the underlying distributions are normal
- the populations are independent
- the population variance of the two populations is the same (but may be unknown)
- n is large.

For this, we modelled the difference on the means as a normal distribution as:

$$\overline{X} - \overline{Y} \sim \mathrm{N}\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$$

Upon standardising, this creates a z-value of:

$$z = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \sim \mathrm{N}(0, 1)$$

The confidence interval can be derived from this test statistic. Consider using the value of z for a known percentage value. In Section 9.4, we used the t-statistic, but this time we will use the z-statistic. Depending on the assumptions and types of distribution, it is possible to create confidence intervals for many mean calculations. For example, an estimated confidence interval for the mean of a Poisson distribution can be calculated by approximating it to a normal distribution. Let $\overline{x}$ be the mean of a random sample of size

n_x from a normal distribution with population mean μ_x and unbiased estimator of the variance s_x^2, and let $\bar{y}$ be the mean of a random sample of size n_y from a normal distribution with population mean μ_y and unbiased estimator of the variance s_y^2, with the conditions:

- X and Y are independent populations
- the population variance of the two populations is the same (but may be unknown)
- n is large.

Then a $100(\alpha - 1)\%$ confidence interval for the difference in means can be calculated as shown in Key point 9.9.

KEY POINT 9.9

A $100(\alpha - 1)\%$ confidence interval for the difference in means is:

$$\bar{x} - \bar{y} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

We also need to consider small samples, for which $n < 30$. Here, we pool the variances to get the best estimate, and then model the difference as a t-distribution.

When we pool the variances like this, the hypothesis yields the following test statistic where s_p^2 is the pooled estimate of the population variance.

$$t_{\frac{\alpha}{2}, n_x+n_y-2} = \frac{(X - Y) - (\mu_x - \mu_y)}{\sqrt{s_p^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$$

A $100(\alpha - 1)\%$ confidence interval for the difference in means for small samples can be calculated as shown in Key point 9.10.

KEY POINT 9.10

The $100(\alpha - 1)\%$ confidence interval for the difference in means for small samples is:

$$(\bar{x} - \bar{y}) \pm t_{\frac{\alpha}{2}, n_x+n_y-2} \times s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$$

where $s_p^2 = \dfrac{\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2}{n_x + n_y - 2}$.

We can calculate s_p^2 using $\dfrac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$ if we know the unbiased estimators s_x^2 and s_y^2.

It is very important to use the correct test based on the sample size.

WORKED EXAMPLE 9.9

A group of 60 men and 70 women participate in a maths activity. The mean time taken for the men to complete the activity is 45.3 seconds, with a standard deviation of 3.2 seconds. For the women, the mean time is 46.1 seconds with a standard deviation of 2.8 seconds. Assuming the completion times are normally distributed with equal variances, find a 90% confidence interval for the difference in the means.

Answer

$\bar{x} - \bar{y} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$

Since n_x and n_y are large in this case, we can use a normal distribution.

Let X represent the completion times for women and Y the completion times for men.

$\bar{x} = 46.1,\ s_x = 2.8,\ n_x = 70$

$\bar{y} = 45.3,\ s_y = 3.2,\ n_y = 60$

$46.1 - 45.3 \pm 1.6449\sqrt{\frac{2.8^2}{70} + \frac{3.2^2}{60}}$

Since we are looking for a 90% confidence interval, we consider $z_{0.95} = 1.6449$.

$= 0.8 \pm 0.874\,534\ldots$

$(-0.0745, 1.6745)$

Given one confidence interval, it is possible to calculate a different confidence interval for the same sample.

WORKED EXAMPLE 9.10

Given that a 90% confidence interval is $(-0.0745, 1.6745)$, calculate a 99% confidence interval.

Answer

We currently know the following.

From the tables: $z_{0.95} = 1.6449$.

$\bar{x} - 1.6449 \times \frac{s}{\sqrt{n}} = -0.0745$

and $\bar{x} + 1.6449 \times \frac{s}{\sqrt{n}} = 1.6745$

$2\bar{x} = 1.6$

$\bar{x} = 0.8$

Solve simultaneously for $\bar{x}$ and $\frac{s}{\sqrt{n}}$.

$\frac{3.2898s}{\sqrt{n}} = 1.749$

Therefore:

$\frac{2.576s}{\sqrt{n}} = 1.749 \times \frac{2.576}{3.2898} = 1.370$

For a 99% CI, we consider $z_{0.995} = 2.576$.

And the 99% confidence interval becomes:

$(0.8 - 1.370, 0.8 + 1.370)$

$(-0.570, 2.170)$

In Worked example 9.11, we shall find the pooled estimate of the population variance and construct the confidence interval, as shown in Key point 9.11.

KEY POINT 9.11

A $100(\alpha - 1)\%$ confidence interval for the difference in means is:

$$(\bar{x} - \bar{y}) \pm t_{\frac{\alpha}{2}, n_x + n_y - 2} \times s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$$

WORKED EXAMPLE 9.11

A shopkeeper believes that playing music in his shop encourages customers to spend more money in his shop. To test this, he records how much money was collected for a ten-day period while music was playing and then an eight-day period when it wasn't. The sales, in thousands of dollars, are summarised as follows.

With music	$\sum x = 960.1$	$\sum x^2 = 92274.44$
Without music	$\sum y = 748.2$	$\sum y^2 = 70041.16$

Assuming these data are randomly sampled from normal distributions with the same variance, find the 90% confidence interval for the difference in means.

Answer

$\bar{x} = 96.01$
$\bar{y} = 93.525$
$s_p^2 = 10.062125$

We saw this in Worked example 9.5, and so we can state the statistics calculated.

The confidence interval can be calculated by:

$$(\bar{x} - \bar{y}) \pm t_{\frac{\alpha}{2}, n_x + n_y - 2} \times s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$$

$$= (96.01 - 93.525) \pm 1.746 \times \sqrt{10.062125} \sqrt{\frac{1}{10} + \frac{1}{8}}$$

We are using $t_{0.95, 16} = 1.746$.

$$= 2.485 \pm 1.746 \times 1.5047$$

So the 90% confidence interval is:
$(-0.142, 5.112)$

WORKED EXAMPLE 9.12

A chemist has developed a fuel additive and claims that it reduces the fuel consumption of cars. Eight randomly selected cars were each filled with 20 litres of fuel and driven around a race circuit. Each car was tested twice, once with the additive and once without it. The distances in miles that each car travelled before running out of fuel are given in the table below.

Car	1	2	3	4	5	6	7	8
Distance without additive	163	172	195	170	183	185	161	176
Distance with additive	168	185	187	172	180	189	172	175

Assuming a normal distribution, find a 90% confidence interval for the difference in the distances travelled.

Answer

Since this is a matched pairs design, we are repeating the experiment on each car then we consider the difference between each pair of data points.

With additive	Without additive	Difference
168	163	5
185	172	13
187	195	−8
172	170	2
180	183	−3
189	185	4
172	161	11
175	176	−1

First, find the differences.

$\sum d = 23, \sum d^2 = 409$

Calculate $\overline{d}$ and s_d^2.

$\overline{d} = \frac{23}{8} = 2.875$

$S_{dd} = 409 - \frac{23^2}{8} = \frac{2743}{8} = 342.875$

$s_d = \sqrt{\frac{342.875}{7}} = 6.999$

We use: $s_d^2 = \frac{S_{dd}}{7}$

$\overline{d} \pm t_{0.95,7} \times \frac{s_d}{\sqrt{8}}$

Calculate the confidence interval.

$2.875 \pm 1.895 \times \frac{6.999}{\sqrt{8}}$

$t_{0.95,7} = 1.895$ from tables.

$2.875 - 4.68855 = -1.814$

$2.875 + 4.68855 = 7.564$

$(-1.814, 7.564)$

EXERCISE 9E

1 For the following confidence intervals, find the value that must be used from the z-distribution.

a 90% confidence interval, $n_x = 40, n_y = 60$

b 95% confidence interval, $n_x = 30, n_y = 40$

c 99% confidence interval, $n_x = 40, n_y = 35$

d 80% confidence interval, $n_x = 80, n_y = 100$

2 For the following confidence intervals, find the value that must be used from the t-distribution.

a 90% confidence interval, $n_x = 8, n_y = 6$

b 95% confidence interval, $n_x = 14, n_y = 9$

c 99% confidence interval, $n_x = 15, n_y = 15$

d 80% confidence interval, $n_x = 8, n_y = 12$

3 By first calculating either $\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}$ or $s_p\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}$, as appropriate, find the stated confidence interval for $\mu_x - \mu_y$. You may assume that the data has come from an underlying normal distribution.

a

$s_x^2 = 13.2$	$n_x = 40$	$\bar{x} = 8.31$
$s_y^2 = 14.6$	$n_y = 50$	$\bar{y} = 7.92$

90% confidence interval

b

$s_x^2 = 11.356$	$n_x = 12$	$\bar{x} = 36.08$
$s_y^2 = 11.643$	$n_y = 8$	$\bar{y} = 36.75$

90% confidence interval

c

$s_x^2 = 433.9$	$n_x = 8$	$\bar{x} = 127.25$
$s_y^2 = 292.9$	$n_y = 7$	$\bar{y} = 124.42$

95% confidence interval

4 A 95% confidence interval using z values is (9.642, 14.558).

a Calculate a 90% confidence interval.

b Calculate a 99% confidence interval.

M 5 Two newly discovered trees, X and Y, are thought to belong to the same species. Leaf measurements are made on each tree and the estimators tabulated.

Tree	Number of leaves sampled	Mean length (cm)	Variance (cm^2)
X	10	14.3	0.50
Y	12	15.1	1.52

a Calculate a pooled estimate for the population variance.

b Find the 90% confidence interval for the difference in the means of the leaf lengths.

M 6 A psychologist wishes to investigate the effect of sleep deprivation on reaction times. Eight students volunteer to take a test, which measures their reaction time, and then re-take the test after being awake for 36 hours. Their reaction times, in milliseconds, are recorded.

	A	B	C	D	E	F	G	H
Before	19.3	11.1	10.3	12.4	13.6	13.2	14.6	15.2
After	20.5	13.5	14.2	12.9	15.3	15.2	16.2	15.9

a Calculate the 99% confidence interval for the difference in reaction times.

b Interpret your confidence interval regarding the effect of sleep deprivation on reaction times.

M 7 In a large school, a sample of 50 boys and 60 girls complete a 100 m race. The estimators of the data are given.

	Sample size	Mean time taken (s)	Variance (s^2)
Boys	50	13.7	2.56
Girls	60	14.9	3.43

Find the 95% confidence interval for the difference in times between the boys and girls to complete the 100 m race.

M 8 An economist believes that a typical basket of weekly food, bought by a family of four, costs more in Eastville than in Weston. Seven stores are randomly selected in each of these two towns and the cost of the basket recorded (in $).

Eastville	13.21	13.97	13.76	13.11	13.25	13.98	13.03
Weston	12.93	13.13	12.98	13.01	12.99	13.21	13.01

Calculate the 95% confidence interval for the difference in means, assuming the costs are normally distributed.

WORKED PAST PAPER QUESTION

A company decides that its employees should follow an exercise programme for 30 minutes each day, with the aim that they lose weight and increase productivity. The weights, in kg, of a random sample of 8 employees at the start of the programme and after following the programme for 6 weeks are shown in the table.

Employee	A	B	C	D	E	F	G	H
Weight before (kg)	98.6	87.3	90.4	85.2	100.5	92.4	89.9	91.3
Weight after (kg)	93.5	85.2	88.2	84.6	95.4	89.3	86.0	87.6

a Assuming that loss in weight is normally distributed, find a 95% confidence interval for the mean loss in weight of the company's employees.

b Test at the 5% significance level whether, after the exercise programme, there is a reduction of more than 2.5 kg in the population mean weight.

Cambridge International AS & A Level Further Mathematics 9231 Paper 23 Q8 June 2011

Answer

a

Difference	5.1	2.1	2.2	0.6	5.1	3.1	3.9	3.7

First, calculate the differences in weight.

$\sum d = 25.8, \ \sum d^2 = 100.14$

$\bar{d} = \frac{25.8}{8} = 3.225$

$s_d^2 = \frac{1}{8-1}\left(100.14 - \frac{25.8^2}{8}\right) = 1.555^2$

95% confidence interval

$\bar{d} \pm t_{0.975,7} \times \frac{s}{\sqrt{n}}$

$= 3.225 \pm 2.365 \times \frac{1.555}{\sqrt{8}}$

Use the unbiased estimators.

$[1.92, 4.53]$

Since the sample size is small, we must consider the t-distribution.

b $H_0: \mu_b - \mu_a = 2.5, \ H_1: \mu_b - \mu_a > 2.5$

State the hypotheses. When setting the parameters, label them sensibly, for example, μ_b for before and μ_a for after.

$\bar{d} = 3.225$

$s_d^2 = 1.555^2$

Calculate $\bar{d}$ and s_d^2.

$$t = \frac{\bar{d} - (\mu_b - \mu_a)}{\frac{s}{\sqrt{n}}} = \frac{3.225 - 2.5}{\frac{1.555}{\sqrt{8}}} = 1.32$$

Calculate the test statistic.

$t_{0.95,7} = 1.89$

We are assuming an underlying normal distribution.
We have a small sample size with unknown variance.
We need to use the t-distribution for our critical value.

Since $1.32 < 1.89$, we do not reject H_0. There is insufficient evidence that the difference in the means is greater than 2.5.

Checklist of learning and understanding

Hypothesis test for the mean, with a small sample:

- The test statistic is $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim t_{n-1}$.
- s can be used in place of σ when the population variance is unknown.

Pooled estimate of a population variance from two samples:

- The pooled estimate of the population variance, s_p^2 can be found from

$$s_p^2 = \frac{\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2}{n_x + n_y - 2}.$$

Difference in means: two-sample t-test:

- Assume:
  - underlying distributions are normal
  - populations are independent
  - population variance of the two populations is the same (but may be unknown).
- The test statistic is $T = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{s_p^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \sim t_{n_x + n_y - 2}$.

Difference in means: paired sample t-test:

- Assume:
  - differences are normally distributed
  - population variance of the two populations is the same (but may be unknown)
  - data are matched pairs (repeated measures design).
- The test statistic is $\frac{\bar{d} - k}{\frac{s_d}{\sqrt{n}}}$.

Difference in means: normal distribution:

- Assume:
  - underlying distributions are normal
  - large sample sizes
  - populations are independent
  - population variance of the two populations is the same (but may be unknown).

- The test statistic is $Z = \dfrac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\dfrac{\sigma_x^2}{n_x} + \dfrac{\sigma_y^2}{n_y}}} \sim N(0, 1)$.

Confidence interval for a mean from a small sample:

- If $\overline{x}$ is the mean of a random sample of size n from a normal distribution with population mean μ, a $100(\alpha - 1)\%$ confidence interval for μ is given by $\overline{x} \pm t_{\frac{\alpha}{2}, n-1}\dfrac{s}{\sqrt{n}}$.

Confidence interval for the difference in population means:

- A $100\ (\alpha - 1)\%$ confidence interval for the difference in means for small samples is given as:

  $(\overline{x} - \overline{y}) \pm t_{\frac{\alpha}{2}, n_x + n_y - 2} \times s_p \sqrt{\dfrac{1}{n_x} + \dfrac{1}{n_y}}$

- A $100(\alpha - 1)\%$ confidence interval for the difference in means (for large n) is given as:

  $(\overline{x} - \overline{y}) \pm z_{\frac{\alpha}{2}} \sqrt{\dfrac{s_x^2}{n_x} + \dfrac{s_y^2}{n_y}}$

- A $100(\alpha - 1)\%$ confidence interval for the difference in means for matched pairs is given as:

  $\overline{d} \pm t\left(\frac{\alpha}{2}, n-1\right) \times \dfrac{s_d}{\sqrt{n}}$

END-OF-CHAPTER REVIEW EXERCISE 9

1 a A gardener P claims that a new type of fruit tree produces a higher annual mass of fruit than the type that he has previously grown. The old type of tree produced 5.2 kg of fruit per tree, on average. A random sample of 10 trees of the new type is chosen. The masses, x kg, of fruit produced are summarised as follows.

$$\sum x = 61.0$$

$$\sum x^2 = 384.0$$

Test, at the 5% significance level, whether gardener P's claim is justified, assuming a normal distribution.

b Another gardener Q has his own type of fruit tree. The masses, y kg, of fruit produced by a random sample of 10 trees grown by gardener Q are summarised as follows.

$$\sum y = 70.0$$

$$\sum y^2 = 500.6$$

Test, at the 5% significance level, whether the mean mass of fruit produced by gardener Q's trees is greater than the mean mass of fruit produced by gardener P's trees. You may assume that both distributions are normal and you should state any additional assumption.

Cambridge International AS & A Level Further Mathematics 9231 Paper 21 Q9 June 2013

2 A random sample of 10 observations of a normally distributed random variable X gave the following summarised data, where $\bar{x}$ denotes the sample mean.

$$\sum x = 70.4, \sum (x - \bar{x})^2 = 8.48$$

Test, at the 10% significance level, whether the population mean of X is less than 7.5.

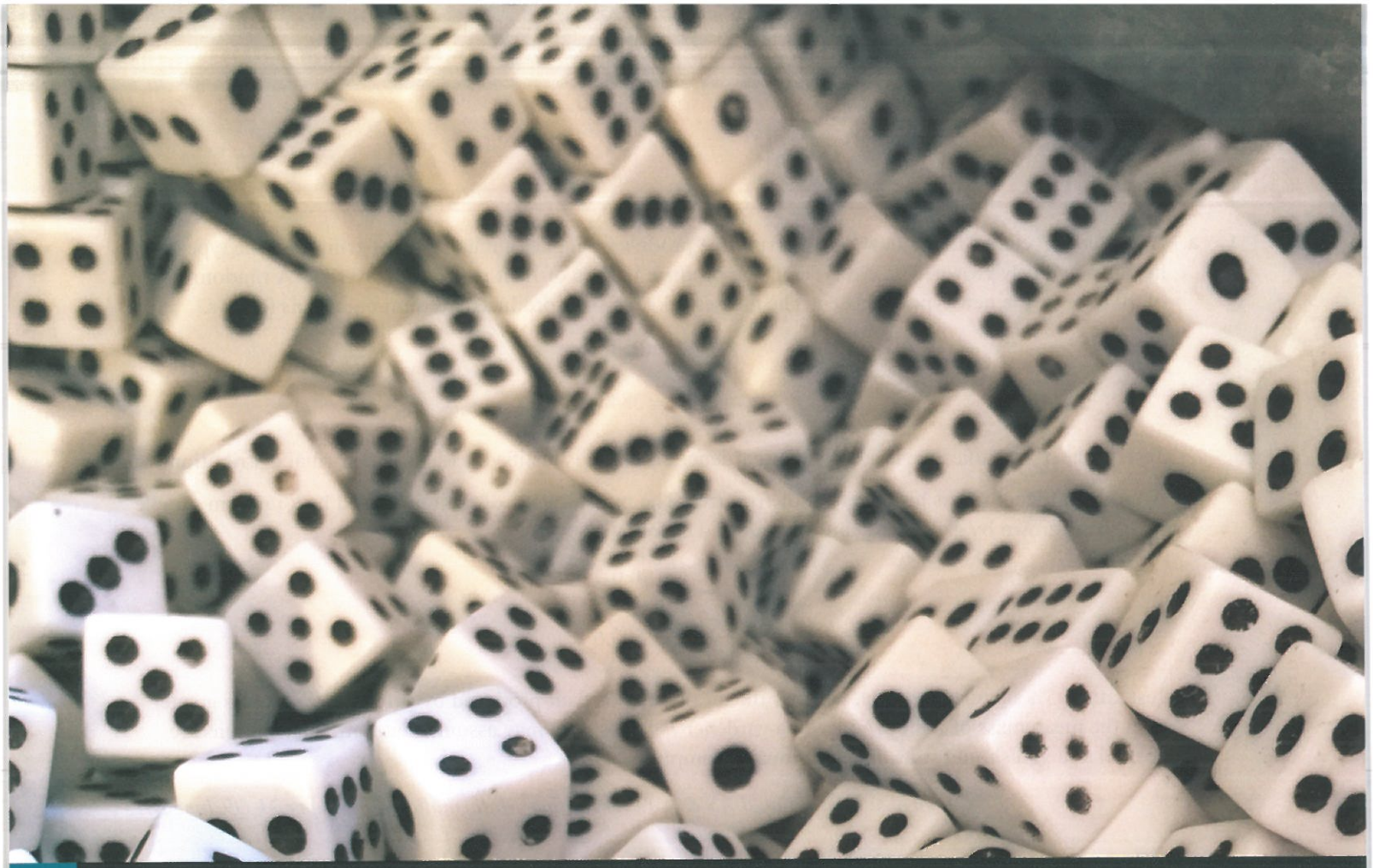
Cambridge International AS & A Level Further Mathematics 9231 Paper 22 Q7 November 2013

3 A random sample of 50 observations of a random variable X and a random sample of 60 observations of a random variable Y are taken. The results for the sample means, $\bar{x}$ and $\bar{y}$, and the unbiased estimates for the population variances, s_x^2 and s_y^2, respectively, are as follows.

$$\bar{x} = 25.4 \quad \bar{y} = 23.6 \quad s_x^2 = 23.2 \quad s_y^2 = 27.8$$

A test at the $\alpha\%$ significance level, of the null hypothesis that the population means of X and Y are equal, against the alternative hypothesis that they are not equal, is carried out. Given that the null hypothesis is not rejected, find the set of possible values of α.

Cambridge International AS & A Level Further Mathematics 9231 Paper 21 Q6 November 2014



Chapter 10
Chi-squared tests

In this chapter you will learn how to:

- fit a theoretical distribution, as prescribed by a given hypothesis, to given data
- use a χ^2-test, with the appropriate number of degrees of freedom, to carry out the corresponding goodness of fit analysis
- use a χ^2-test, with the appropriate number of degrees of freedom, for independence in a contingency table.

PREREQUISITE KNOWLEDGE

Where it comes from	What you should be able to do	Check your skills
AS & A Level Mathematics Probability & Statistics 1, Chapters 6 & 7 AS & A Level Mathematics Probability & Statistics 2, Chapter 2	Calculate probabilities from discrete random variables such as binomial, Poisson and geometric.	1 Let $X \sim \text{Bin}(10, 0.2)$. Find $P(X \leqslant 3)$. 2 Let $Y \sim \text{Po}(2.2)$. Find $P(Y > 3)$. 3 Let $G \sim \text{Geo}(0.3)$. Find $P(G \leqslant 4)$.
AS & A Level Mathematics Probability & Statistics 2, Chapter 5	Calculate probabilities from the normal distribution.	4 Let $X \sim \text{N}(42, 6)$. Find $P(X \geqslant 40.2)$.
Chapter 8	Calculate probabilities from any continuous random variable.	5 Let X represent a continuous random variable with a probability density function f given by: $f(x) = \begin{cases} \frac{3}{80}(x^2 + 3) & -2 \leqslant x \leqslant 3 \\ 0 & \text{otherwise} \end{cases}$ Find $P(-1 \leqslant X < 2)$.

Testing statistical models

In this chapter, we shall test whether the data provided fit a particular distribution. Sometimes the parameters will be given, but sometimes they will not.

We shall also test how closely categorical data are associated. For example, we could test if there is an association between hair colour and eye colour, or between colour blindness and gender.

These statistical tests are vital in social sciences: Psychology and Sociology, in particular. Most of the data collected in these subjects are categorical. This allows researchers to analyse any associations between these types of data.

10.1 Forming hypotheses

We shall test how well the observed data from an experiment fits the expected values from a distribution. For example, consider rolling a die. We may wish to test if the values on the die have an equal chance of being selected, that is the die is not biased. This means that we are looking to see if the data fit a **discrete uniform distribution**. A discrete uniform distribution occurs when each outcome is equally likely so it has the same probability.

To perform a hypothesis test, we need to define a null hypothesis as in Chapter 9.

For the die, we could set up these hypotheses:

H_0: There is no difference between the observed data and the expected values.

H_1: There is a difference between the observed data and the expected values.

This is vague, but it shows the fundamental premise behind the test. We need to ensure that our hypotheses are related to the situation we are testing, as shown in Key point 10.1.

We could have written these hypotheses:

H_0: A discrete uniform distribution is a good-fit model.

H_1: A discrete uniform distribution is not a good-fit model.

It is very important that the hypotheses are well written and that our conclusions refer to the initial problem.

KEY POINT 10.1

When defining your null and alternative hypotheses, make sure you refer to the distribution that you are using to model your observed data.

In your conclusion, make sure you address the initial problem. You will see this in the examples regarding rolling die.

Let us continue with the idea of rolling a die. In an experiment, we roll a die 180 times and collect the following data.

Number, n	1	2	3	4	5	6
Observed frequency	29	31	34	39	23	24

We will use N to denote $\sum O_i$, the sum of the observed frequencies.

To test whether the die is biased or not, we need to consider how well this data set fits a discrete uniform distribution.

The discrete uniform distribution will have the following probability distribution.

$$\mathrm{P}(X = x) = \begin{cases} \frac{1}{6} & x = 1, 2, 3, 4, 5, 6 \\ 0 & \text{otherwise} \end{cases}$$

The expected frequencies from this distribution (and, in fact, any discrete distribution we choose to use) can be calculated using the formula $N \times \mathrm{P}(X = x)$.

Number, n	1	2	3	4	5	6
Expected frequency	30	30	30	30	30	30

The last cell is coloured red for a reason. Since we know the total is 180, and we know the sum of the other expected values is 150, the last value is predetermined. It is not calculated from the probability distribution. This seems trivial at the moment, but will become very important when calculating the test statistic and modelling its distribution.

The first five expected values (these can be *any* five) are free, independent variables. The final expected value is predetermined and is not independent of the others. It is calculated from the others rather than from the probability distribution.

This is called a **constraint** and reduces the **free variables** in the system by one. We call these degrees of freedom.

We will need to refine how we find degrees of freedom, but we can make use of the formula shown in Key point 10.2.

KEY POINT 10.2

For a goodness-of-fit test, the number of degrees of freedom, ν (the Greek letter nu), in the system is found using:

ν = number of expected values − 1 − number of parameters estimated

In the case of rolling a die, we have six expected values. We subtract 1 because of the constraint on the system and we have no parameters to estimate. (This will be developed further through the following examples.)

Let's look back at our die:

Number, n	1	2	3	4	5	6
Observed frequency	29	31	34	39	23	24
Expected frequency	30	30	30	30	30	30

We need to calculate a test statistic to be able to carry out the test.

We want to calculate the *difference* relative to the expected value (E_i) and the corresponding observed value (O_i). We want this test statistic to be equal to 0 if the observed and expected frequencies are the same.

The construction is very similar to finding the sample variance.

We first calculate $\frac{(O_i - E_i)^2}{E_i}$ for each pair of values and then we add these up:

$$X^2 = \sum \left(\frac{(O_i - E_i)^2}{E_i} \right)$$

This is the test statistic that we will use.

Which distribution can we use to model the test statistic? If we meet the following necessary conditions then $X^2 \sim \chi^2(\nu)$, where ν is the number of degrees of freedom required:

- each O_i represents a frequency
- all E_i are greater than five
- the classes all form a sample space; that is, each observation taken fits uniquely into a single category.

This is shown in Key point 10.3. Note that χ^2 is a **chi-squared** (from the Greek letter chi; pronounced 'kye-squared') statistic.

KEY POINT 10.3

Given the condition that all $E_i \geqslant 5$, then $\sum \left(\frac{(O_i - E_i)^2}{E_i} \right) \sim \chi^2(\nu)$ is a good approximation.

DID YOU KNOW?

The $\chi^2(\nu)$ distribution is the sum of ν independent values from squared standardised normal distributions:

$$\chi^2(\nu) = \sum_1^{\nu} Z_i^2, \text{ where } Z_i \sim N(0, 1^2)$$

In the case of the die discussed previously, we have five independent variables (the final E was a constraint on the system), and so we can model the X^2 statistic, the sum of the squared normal distributions, as $Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2 = \chi^2(5)$.

The condition that $E_i \geqslant 5$ is linked to the similar condition that the frequency of the bars of a histogram must also by greater than five. It is useful to research this yourself. It has something to do with the effect of outlying data on the test statistic. As the sample size grows, this condition can be relaxed as the approximations to the squared normal distributions improve. However, in this course we must use the condition that each $E_i \geqslant 5$.

WORKED EXAMPLE 10.1

Let us return to the experiment with the die.

An experiment is carried out to test whether a die is biased or not. A die is rolled 180 times and the following observations are tabulated.

Number, n	1	2	3	4	5	6
Observed frequency	29	31	34	39	23	24

Test, at the 5% significance level, whether the die is biased.

Answer

If the die is biased, then we cannot fit the data to a discrete uniform distribution. If the die is unbiased, then a discrete uniform distribution would be a good fit.

Define the hypotheses.

H_0: A discrete uniform distribution is a good fit.
H_1: A discrete uniform distribution is not a good fit.

The assumed probability distribution is:

$$P(X = x) = \begin{cases} \frac{1}{6} & x = 1, 2, 3, 4, 5, 6 \\ 0 & \text{otherwise} \end{cases}$$

So we can calculate the expected frequencies and, hence, the value of each $\frac{(O_i - E_i)^2}{E_i}$.

Expected frequencies are calculated using $N \times P(X = x)$.

Number	Observed	Expected	$\frac{(O_i - E_i)^2}{E_i}$
1	29	30	0.0333
2	31	30	0.0333
3	34	30	0.5333
4	39	30	2.7
5	23	30	1.6333
6	24	30	1.2

$$\chi^2 = \sum \left(\frac{(O_i - E_i)^2}{E_i} \right) = 6.1333$$

$\nu = 6 - 1 - 0$
$\nu = 5$

Calculate the degrees of freedom. Since we have not estimated any parameters, we have five degrees of freedom.

The critical value will be:

Find the critical value from the χ^2-distribution table.

Even though the hypothesis is two-tailed, we consider only the upper tail for goodness of fit since we are measuring how much the test statistic *exceeds* a specific value.

p	0.95
$v = 1$	3.841
2	5.991
3	7.815
4	9.488
5	11.07
6	12.59

$\chi^2_5(0.95) = 11.07$

Since $6.1333 < 11.07$, there is insufficient evidence to reject H_0.

Always refer to H_0.

There is insufficient evidence to state that a discrete uniform distribution is not a good fit.

Refer back to the question when writing the conclusion.

There is insufficient evidence to suggest that the die is biased.

TIP

There is another way of calculating this test statistic.

$$\sum\left(\frac{(O_i - E_i)^2}{E_i}\right)$$

$$= \sum\left(\frac{O_i^2 - 2O_iE_i + E_i^2}{E_i}\right)$$ Multiply out.

$$= \sum\left(\frac{O_i^2}{E_i}\right) - 2\Sigma(O_i) + \Sigma(E_i)$$ Simplify.

$$= \sum\left(\frac{O_i^2}{E_i}\right) - 2N + N$$ Now we have established that $\Sigma O_i = N$.

$$\chi^2 = \sum\left(\frac{O_i^2}{E_i}\right) - N$$ And, in fact, $\Sigma E_i = N$.

This is a convenient form to use, but some information that may be useful for further analysis may be lost. We will mention this when we consider contingency tables in Section 10.4.

EXERCISE 10A

1 Find the values of:

a $\chi^2_8(0.9)$ b $\chi^2_{11}(0.95)$ c $\chi^2_4(0.99)$ d $\chi^2_{21}(0.995)$

2 For the following distributions and sample sizes, write the table of expected values.

a $P(X = x) = \begin{cases} \frac{x^2}{30} & x = 1, 2, 3, 4 \\ 0 & \text{otherwise} \end{cases}$

$n = 150$

b Data values a, b, c, d are in the ratio $4:4:2:2$; $n = 240$

c $P(X = x) = \begin{cases} \frac{1}{7} & x = 2, 3, 4, 5, 6, 7, 8 \\ 0 & \text{otherwise} \end{cases}$

$n = 280$

3 For the following sets of observed (O) and expected (E) data, calculate the value of χ^2, the test statistic.

a

O	32	59	12	14	41	32
E	40	50	20	15	40	25

b

O	8	23	39	42	42	41	36	22	11
E	11	24	35	40	45	40	35	24	11

c

O	35	61	70	55	39
E	30	60	90	50	30

4 The following dataset shows values for what is thought to have come from the probability distribution.

$P(X = x) = \begin{cases} \frac{5 - x}{10} & x = 1, 2, 3, 4 \\ 0 & \text{otherwise} \end{cases}$

n	1	2	3	4
Observed frequency	35	29	12	4

A test at the 5% significance level will be carried out.

a Find the expected values.

b State the hypotheses.

c Find the value of the test statistic.

d State the degrees of freedom.

e Write down the critical value.

f Conclude the hypothesis test.

M 5 The population of a country is known to have blood groups O, A, B and AB in the ratio $5:3:2:1$. 220 people are randomly selected from the population of a neighbouring country. Their blood group is assessed and the results tabulated.

Blood type	O	A	B	AB
Frequency	87	73	49	11

Test at the 5% significance level whether or not the neighbouring country's population has the same proportions of blood groups.

M 6 A company is preparing invoices to send to their customers. Before they are sent, they are checked and the daily number of mistakes found over a two-week period are recorded.

	Week 1					Week 2				
	M	Tu	W	Th	F	M	Tu	W	Th	F
Errors	29	17	13	15	22	26	16	25	18	19

Test, at the 5% level of significance, whether a uniform distribution is a good-fit model.

10.2 Goodness of fit for discrete distributions

Testing a binomial distribution as a model

We can now apply these general principles to known parametric distributions. With the binomial distribution, we may be asked to see whether the dataset fits a binomial distribution with a given probability of success (p), or we may be asked whether it fits a binomial distribution without the parameter stated. In this case, we would need to estimate the parameter. This will affect the degrees of freedom, as stated in the formula:

ν = number of expected values − 1 − number of estimated parameters

The parameter is p. This can be estimated by knowing that $E(X) = np$. So the estimate of p is:

$$\hat{p} = \frac{\overline{x}}{n}, \text{ where } \overline{x} = \frac{\Sigma r_i \times O_i}{N}, \text{ where } r_i = 0, 1, 2\ldots, n, \text{ the mean from the observed dataset.}$$

TIP

Be careful not to confuse n (the number of trials in the binomial distribution) with N (the number of times the experiment is repeated).

WORKED EXAMPLE 10.2

The data in the table are thought to be binomially distributed.

x	0	1	2	3	4	5	6	7
Frequency	10	34	63	48	29	10	4	2

Test, at the 5% significance level, the claim that the data are binomially distributed.

Answer

$$\overline{x} = \frac{\Sigma(r_i \times O_i)}{N} = \frac{508}{200} = 2.54$$

Since we are not given a parameter, we must estimate this first.

Therefore, $\hat{p} = \frac{2.54}{7} = 0.363$ (3 significant figures)

H_0: A binomial distribution is a good fit.
H_1: A binomial distribution is not a good fit.

Define the hypotheses. Notice there is no reference to the value of the population proportion, as this is unknown.

We calculate the expected values by $200 \times {}^nC_r(0.363)^r(0.637)^{n-r}$.

Expected values are calculated using $E_i = N \times P(X = r)$ and we have $X \sim \text{Bin}(7, 0.363)$.

x	Observed	Expected
0	10	8.525
1	34	33.985
2	63	58.064
3	48	55.113
4	29	31.387
5	10	10.725
6	4	2.036
7	2	0.166

TIP

All values should be written to 3 decimal places. However, more accurate values should be used in calculations.

x	Observed	Expected	$\frac{(O_i - E_i)^2}{E_i}$
0	10	8.525	0.255
1	34	33.985	0.000
2	63	58.064	0.420
3	48	55.113	0.918
4	29	31.387	0.182
5–7	16	12.927	0.731

As some $E_i < 5$ we need to combine the cells for when $x = 5, 6$ and 7.

After combining we have six expected values.

$$\chi^2 = \sum\left(\frac{(O_i - E_i)^2}{E_i}\right) = 2.506$$

$\nu = 6 - 1 - 1 = 4$

Consider the number of degrees of freedom: we need the extra −1 because we are now required to estimate a parameter.

The critical value is $\chi^2_4(0.95) = 9.488$. Since $2.506 < 9.488$, there is insufficient evidence to reject H_0. There is insufficient evidence to state that a binomial distribution is not a good fit.

Note, throughout the question, no reference has been made to the population proportion. This is because it is unknown.

Testing a Poisson distribution as a model

With the Poisson distribution, we need to see whether the dataset fits a Poisson distribution with a given rate (the parameter here is λ). Alternatively, we may need to estimate the parameter.

The parameter is λ. This can be estimated using:

$$\hat{\lambda} = \frac{\Sigma r_i \times O_i}{N}$$

WORKED EXAMPLE 10.3

The data in the table are thought to be modelled as a Poisson distribution with a mean of 2.5.

Number, n	0	1	2	3	4	5	6	7–
Frequency	8	34	42	28	26	5	5	2

Test the claim, at the 5% level of significance, that the data can be modelled as a Poisson distribution with mean 2.5.

Answer

In this case, the parameter is given and so it does not need to be estimated.

H_0: The Poisson (2.5) model is a good fit.

H_1: The Poisson (2.5) model is not a good fit.

Define the hypotheses.

Here we can refer to the population parameter in the hypotheses, as it is a known value.

We calculate the expected values here by:

$150 \times e^{-2.5} \times \frac{2.5^r}{r!}$

For the Poisson distribution,
$P(X = r) = \frac{e^{-\lambda}\lambda^r}{r!}$.
The final E value is calculated as 150 – (sum of the others).

x	Observed	Expected
0	8	12.313
1	34	30.782
2	42	38.477
3	28	32.064
4	26	20.040
5	5	10.020
6	5	4.175
7–	2	2.129

Since $E < 5$ for some values, we need to combine the final two categories.

x	Observed	Expected	$\frac{(O_i - E_i)^2}{E_i}$
0	8	12.313	1.511
1	34	30.782	0.336
2	42	38.477	0.323
3	28	32.064	0.515
4	26	20.040	1.772
5	5	10.020	2.515
6–	7	6.304	0.077

$\chi^2 = \sum\left(\frac{(O_i - E_i)^2}{E_i}\right) = 7.049$

Calculate the test statistic.

$\nu = 7 - 1 = 6$

We now consider the degrees of freedom.

No parameters were estimated.

The critical value is $\chi^2_6(0.95) = 12.59$.

Since $7.049 < 12.59$, there is insufficient evidence to reject H_0.

There is insufficient evidence to state that a Poisson (2.5) model is not a good fit.

EXERCISE 10B

1 It is believed that some data, $N = 80$, can be modelled by $X \sim \text{Bin}(6, 0.3)$. The table of expected values is:

x_i	0	1	2	3	46
E_i	9.412	24.202	r	s	5.638

a Find r.

b Find s.

Write your answers to 3 decimal places.

2 It is believed that some data, $N = 100$, can be modelled by $X \sim \text{Po}(2.9)$. The table of expected values is:

x_i	0	1	2	3	4	5	6–
E_i	5.502	15.957	r	22.367	s	9.405	7.417

a Find r.

b Find s.

Write your answers to 3 decimal places.

3 It is believed that the following observed data follow a binomial distribution.

x_i	0	1	2	3	4	5	6	7
O_i	3	10	15	16	9	3	2	2

Let $\hat{p}$ be the unbiased estimator for the proportion, p.

a Show that $\hat{p} = 0.393$ to 3 significant figures.

b Using the value found in **a**, find the values of r and s in the following table of expected data.

x_i	0–1	2	3	4	5–7
E_i	10.078	r	17.304	s	5.378

c Explain why it was necessary to combine some of the columns together.

4 For the data below, calculate the test statistic and state how many degrees of freedom are required, assuming no parameters need to be estimated.

O	20	31	15	12	10	2
E	24	35	14	7	6	4

For each following question, clearly state:

- your hypotheses
- the value of your test statistic
- the degrees of freedom required
- the critical value
- your conclusion.

M 5 150 students take a multiple-choice test consisting of six questions. The numbers of correct answers are tabulated.

x	0	1	2	3	4	5	6
Observed	2	10	30	36	44	22	6

Test, at the 5% significance level, whether a binomial distribution is a good model for the data.

M 6 The number of accidents on a road per day is recorded for 80 days, giving the following results.

No. accidents	0	1	2	3	4	5
Frequency	13	21	23	11	7	5

It is thought that the dataset models a Poisson distribution with a rate of 2.5 accidents per day. Test this claim at the 5% significance level.

M 7 The owner of a small ski hostel records the demand for rooms during high season.

Rooms required	0	1	2	3	4
No. nights	13	27	41	13	6

a Show that the mean demand for rooms per night is 1.72.

A test is to be carried out at the 1% significance level to show that a Poisson distribution is a good model. The expected values are:

Rooms required	0	1	2	3	4
Expected values	17.91	p	26.49	q	9.62

b Find the value of p, and hence q.

c Carry out the hypothesis test.

10.3 Goodness of fit for continuous distributions

Testing a normal distribution as a model

With the normal distribution, as with all continuous distributions, we need to note carefully how the data are grouped so that we calculate the expected values correctly. Worked examples 10.4 and 10.5 highlight this point. We also need to be aware that two parameters may now need to be estimated: μ and σ^2.

If required, μ can be estimated by $\bar{x}$, and σ^2 can be estimated by s^2. This will affect the number of degrees of freedom, as shown in Key point 10.4.

KEY POINT 10.4

When fitting a normal distribution, the number of degrees of freedom are:

$\nu = n - 1$ if no parameters are estimated

$\nu = n - 1 - 1$ if one parameter is estimated

$\nu = n - 1 - 2$ if two parameters are estimated.

WORKED EXAMPLE 10.4

During observations on the weights of 150 newborn babies, the following data are observed and recorded to 1 decimal place.

Weight (kg)	2.0–2.4	2.5–2.9	3.0–3.4	3.5–3.9	4.0–4.4	4.5–4.9
Frequency	8	14	54	43	22	9

It is believed that the data follow a normal distribution, with variance 0.4. Test this belief at the 10% significance level.

Answer

Here, we are given the population variance, but we need to estimate the mean. We also need to set out the data in a way that will enable us to calculate the expected values.

To estimate the mean, use the midpoints of the groups in the usual calculation.

$\sum fx = 522$

$n = 150$

$\hat{\mu} = \frac{522}{150} = 3.48$

H_0: A normal distribution with variance 0.4 is a good fit.

H_1: A normal distribution with variance 0.4 is not a good fit.

Define the hypotheses.

a	b	Probability of interval	Observed value	Expected value	$\frac{(O_i - E_i^2)}{E_i}$
	2.45	0.0517	8	7.755	0.008
2.45	2.95	0.1493	14	22.397	3.148
2.95	3.45	0.2801	54	42.010	3.422
3.45	3.95	0.2902	43	43.532	0.007
3.95	4.45	0.1661	22	24.922	0.343
4.45		0.0626	9	9.383	0.016

Show the information correctly so you can calculate the expected values. It is very important that the expected values add up correctly to the total of observed data. For this to happen, we must consider the entire probability distribution and ensure that the probabilities add up to 1.

We must use the probability at the lower tail and the upper tail to include all values. This means that the lower group should be $X < 2.45$ rather than $1.95 \leqslant X < 2.45$. Similarly, for the upper tail, we should consider $X \geqslant 4.45$ rather than $4.45 \leqslant X < 4.95$.

Standardise $X \sim N(3.48, 0.4)$ to calculate the probabilities.

For example: $P(X < 2.45)$

$$= P\left(Z < \frac{2.45 - 3.48}{\sqrt{0.4}}\right) = 0.0517$$

$$\chi^2 = \sum \left(\frac{(O_i - E_i)^2}{E_i}\right) = 6.942$$

$\nu = 6 - 1 - 1 = 4$

We have estimated one parameter, $\hat{\mu} = 3.48$, and so we have $\nu = n - 1 - 1$ degrees of freedom.

The critical value is $\chi^2_4(0.9) = 7.779$.

Find the critical value.

Since $6.942 < 7.779$, there is insufficient evidence to reject H_0. There is insufficient evidence to state that a normal distribution with variance 0.4 is not a good model.

Worked example 10.5 introduces a continuous uniform distribution and fits it to the data. The grouped data are described in a different way. It is very important that you are aware of how the data are presented, as this can affect the accuracy of the estimators required.

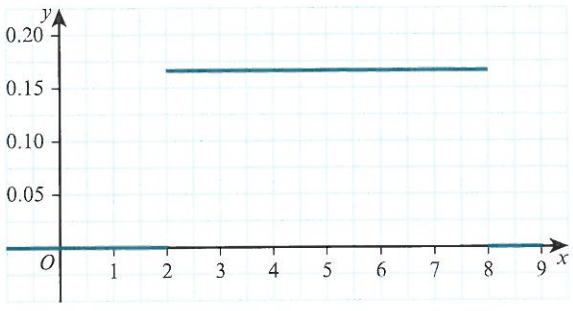
A continuous uniform distribution is sometimes referred to as a *rectangular* distribution due to the shape of its probability density function. The distribution over an interval has the same probability density for all values.

Let $X \sim U[a, b]$.

Then $f(x) = \begin{cases} \frac{1}{b-a} & a \leqslant x \leqslant b \\ 0 & \text{otherwise.} \end{cases}$

For example, consider $X \sim U[2, 8]$:

$f(x) = \begin{cases} \frac{1}{6} & 2 \leqslant x \leqslant 8 \\ 0 & \text{otherwise} \end{cases}$



WORKED EXAMPLE 10.5

The waiting times for a bus are observed over 120 days and the results are noted.

Time, t (min)	0–	10–	20–	30–	40–	50–60
Frequency	14	25	27	25	15	14

The departure time of the previous bus is unknown in each case. It is believed that the waiting times are uniformly distributed over one hour. Test this claim at the 10% level of significance.

Answer

We are modelling this as a uniform distribution on the interval $[0, 60]$ and so we can define our hypotheses on this basis.

Define the hypotheses.

H_0: $U[0, 60]$ is a good model.

H_1: $U[0, 60]$ is not a good model.

t (min)	Observed	Expected	$\frac{(O_i - E_i^2)}{E_i}$
$0 \leqslant t < 10$	14	20	1.8
$10 \leqslant t < 20$	25	20	1.25
$20 \leqslant t < 30$	27	20	2.45
$30 \leqslant t < 40$	25	20	1.25
$40 \leqslant t < 50$	15	20	1.25
$50 \leqslant t < 60$	14	20	1.8

The probability function for $U[a, b]$ is

$$P(X = x) = \begin{cases} \frac{1}{b-a} & a \leqslant x \leqslant b \\ 0 & \text{otherwise.} \end{cases}$$

$$\chi^2 = \sum \left(\frac{(O_i - E_i)^2}{E_i} \right) = 9.8$$

$\nu = 6 - 1 = 5$

Calculate the number of degrees of freedom.

$\chi^2_5(0.9) = 9.236$

Find the critical value.

Since $9.8 > 9.236$, there is sufficient evidence to reject H_0.

There is sufficient evidence to suggest that $U[0, 60]$ does not fit the data.

EXERCISE 10C

1 For each of the following distributions, write the table of expected values.

a $P(X = x) = \begin{cases} \frac{1}{5} & 2 \leqslant x \leqslant 7 \\ 0 & \text{otherwise} \end{cases}$

x	O_i
$2 \leqslant x < 3$	18
$3 \leqslant x < 4$	10
$4 \leqslant x < 4.5$	10
$4.5 \leqslant x < 5$	11
$5 \leqslant x \leqslant 7$	31

b $P(X = x) = \begin{cases} \frac{2}{9} & 6.5 \leqslant x \leqslant 11 \\ 0 & \text{otherwise} \end{cases}$

x	O_i
$6.5 \leqslant x < 7.5$	20
$7.5 \leqslant x < 8$	8
$8 \leqslant x < 8.5$	11
$8.5 \leqslant x < 9$	6
$9 \leqslant x < 10$	21
$10 \leqslant x < 11$	15

c $P(X = x) = \begin{cases} \frac{x^2 - 1}{228} & 3 \leqslant x \leqslant 9 \\ 0 & \text{otherwise} \end{cases}$

x	O_i
$3 \leqslant x < 6$	127
$6 \leqslant x < 7$	86
$7 \leqslant x < 8$	119
$8 \leqslant x < 8.5$	79
$8.5 \leqslant x \leqslant 9$	89

2 For each distribution in question **1a–c**, calculate the value of the test statistic, χ^2.

a

x	O_i
$2 \leqslant x < 3$	18
$3 \leqslant x < 4$	10
$4 \leqslant x < 4.5$	10
$4.5 \leqslant x < 5$	11
$5 \leqslant x \leqslant 7$	31

b

x	O_i
$6.5 \leqslant x < 7.5$	20
$7.5 \leqslant x < 8$	8
$8 \leqslant x < 8.5$	11
$8.5 \leqslant x < 9$	6
$9 \leqslant x < 10$	21
$10 \leqslant x < 11$	15

c

x	O_i
$3 \leqslant x < 6$	127
$6 \leqslant x < 7$	86
$7 \leqslant x < 8$	119
$8 \leqslant x < 8.5$	79
$8.5 \leqslant x \leqslant 9$	89

3 For each of the following sets of information, find:

i the number of degrees of freedom

ii the critical value.

a Number of E_i cells after combining = 9. The data are believed to fit a normal distribution with mean 6. Test at 5% significance.

b Number of E_i cells after combining = 7. The data are believed to fit a normal distribution.

Test at 10% significance.

c Number of E_i cells after combining = 11. The data are believed to fit a normal distribution with mean 6 and variance 4.

Test at 2.5% significance.

d Number of E_i cells after combining = 15. The data are believed to fit a normal distribution with variance 5.

Test at 5% significance.

4 Let $X \sim N(35, 4^2)$. Find:

a $P(X < 30.5)$ b $P(30.5 \leqslant X < 40.5)$ c $P(40.5 \leqslant X < 46.5)$ d $P(X \geqslant 46.5)$

For each following question, clearly state:

- your hypotheses
- the value of the test statistic
- the number of degrees of freedom required
- the critical value
- your conclusion.

M 5 A machine is designed to cut metal into strips of length 25 m, to the nearest metre. The lengths of 100 cut pieces are grouped and recorded.

Length cut (m)	Frequency
$24.5 \leqslant l < 24.75$	21
$24.75 \leqslant l < 25$	29
$25 \leqslant l < 25.25$	27
$25.25 \leqslant l < 25.5$	23

It is believed that the machine is equally likely to cut the metal to any length between 24.5 and 25.5 m. Test this claim at the 5% level of significance.

M 6 A company makes climbing rope, which is cut to lengths of 50 m with a standard deviation of 1.5 m. A sample of 150 pieces of rope is measured and the results are recorded.

Length (m)	Frequency
$l < 48$	0
$48 \leqslant l < 49$	2
$49 \leqslant l < 50$	22
$50 \leqslant l < 51$	30
$51 \leqslant l < 52$	33
$52 \leqslant l < 53$	30
$53 \leqslant l < 54$	25
$54 \leqslant l < 55$	8

Test, at the 5% significance level, whether the lengths of pieces of rope can be modelled as a normal distribution according to the parameters suggested.

M 7 At the end of a statistics course, 110 students sit an examination. The marks are grouped into classes, as shown in the following table.

Marks (X)	Number of students
0–29	22
30–34	8
35–39	8
40–49	16
50–59	20
60–69	16
70–90	20

$\sum x = 5305$, $\sum x^2 = 392247.5$

It is believed that the mean for the population is 47.5.

a Show that the unbiased estimator for the standard deviation of these data is 35.38.

b Test, at the 0.5% significance level, whether the data fit a normal distribution with mean 47.5.

8 It is believed that the following data fit the model $f(x) = \begin{cases} \frac{1}{40} e^{-\left(\frac{x}{4}\right)} & 0 < x \\ 0 & \text{otherwise.} \end{cases}$

x	0–	20–	40–	60–	90–	120–
Frequency	40	20	15	14	8	3

Test this claim at the 5% significance level.

EXPLORE 10.1

Sometimes, statistics can be manipulated to ensure that the outcome of a test supports a person's claim.

Consider a tree nursery. It has recently reduced its spending on the amount and quality of the fertiliser used to help the trees to grow. The expected heights of the trees after six months should be:

Height (cm)	0–10	11–20	21–30	31–40	41–50	51–60	61–70	71–80	81–90
Expected frequencies	2	2	5	4	9	11	2	3	2

The nursery manager has been told that the fertiliser is not good enough and the trees are not growing as they should be. For non-scientific reasons, the nursery manager wishes to show that this is not true. He carries out a 2.5% significance level χ^2-test to show that the observed data and expected data have the same distribution. The observed heights of the trees after six months were:

Height (cm)	0–10	11–20	21–30	31–40	41–50	51–60	61–70	71–80	81–90
Observed frequencies	4	4	5	11	6	4	3	2	1

How can the nursery manager perform a χ^2-test to show that the observed and expected data come from the same distribution and, hence, the trees are growing as they should? Think about how to group categories.

10.4 Testing association through contingency tables

Another application of the χ^2-distribution is to look for an association between two criteria. We describe this as testing whether two criteria are *independent*. This is a particularly powerful test as it allows us to deal with data in categories, such as the association between eye colour and hair colour. Again, we need to be very specific and careful with the language that we use. As long as the two criteria we want to test can be split into distinct categories, we can create a **contingency table** and then use χ^2 to test for an association between the criteria.

It is important to include row totals, column totals and the grand total when using contingency tables.

The total of each row is called R_i.

The total of each column is called C_j.

		Hair colour			Row totals
		Brown	Blonde	Red	
Eye colour	Brown	63	31	6	$100 = R_1$
	Blue	26	20	14	$60 = R_2$
	Green	11	19	10	$40 = R_3$
Column totals		$100 = C_1$	$70 = C_2$	$30 = C_3$	$200 = T$

This is a contingency table. Notice that each data point will be placed uniquely into one of the nine categories, thus satisfying one of the conditions for using a χ^2-distribution.

This contingency table shows the *observed* data.

We can see that the eye colours are in the ratio 5 : 3 : 2 and the hair colours are in the ratio 10 : 7 : 3. However, no individual row or column matches these ratios.

If there is no association between the criteria, the ratios between the totals should also be reflected in each row and in each column. We use the ratios to find each expected value, as shown in Key point 10.5.

KEY POINT 10.5

To calculate each expected value in a contingency table:

$$E_{ij} = \frac{R_i \times C_j}{T}$$

This will guarantee that the ratios between the totals are reflected in each row and each column, and will give us a table that assumes there is no association between eye colour and hair colour. We can now set up our hypotheses.

WORKED EXAMPLE 10.6

Consider the previous contingency table. Find the expected values.

Answer

		Hair colour			Row totals
		Brown	Blonde	Red	
Eye colour	Brown	$\frac{100 \times 100}{200}$	$\frac{100 \times 70}{200}$	$\frac{100 \times 30}{200}$	100
	Blue	$\frac{60 \times 100}{200}$	$\frac{60 \times 70}{200}$	$\frac{60 \times 30}{200}$	60
	Green	$\frac{40 \times 100}{200}$	$\frac{40 \times 70}{200}$	$\frac{40 \times 30}{200}$	40
Column totals		100	70	30	200

We use the formula $E_{ij} = \frac{R_i \times C_j}{T}$ for each cell.

		Hair colour			Row totals
		Brown	Blonde	Red	
Eye colour	Brown	50	35	15	100
	Blue	30	21	9	60
	Green	20	14	6	40
Column totals		100	70	30	200

Notice that in the table of expected values, each row is now in the ratio 10 : 7 : 3 and each column is in the ratio 5 : 3 : 2, as required.

To perform a χ^2-test, we also need to consider how many degrees of freedom there are in the system.

Consider the previous table (expected values):

		Hair colour			Row totals
		Brown	Blonde	Red	
Eye colour	Brown	50			100
	Blue			9	60
	Green	20	14		40
Column totals		100	70	30	200

Since we are given the totals, we have sufficient information (and the necessary information) to be able to generate the whole table. The four data values shown in red are the only free independent variables that exist. We can calculate the remaining five values from these four. Note that these are not the only four values that could be used. Some combinations of four values will not work as they will not allow us to calculate the rest of the values. The minimum number of expected values that must be calculated independently is four in this case. We can calculate the number of degrees of freedom as (number of rows − 1)(number of columns − 1) = (3 − 1)(3 − 1) = 4 as shown in Key point 10.6.

KEY POINT 10.6

The number of degrees of freedom of an $r \times c$ contingency table is:

$$\nu = (r - 1)(c - 1)$$

It is now possible to perform a hypothesis test to see whether there is an association between two criteria (provided the criteria are independent).

WORKED EXAMPLE 10.7

Given the data below, conduct a hypothesis test, at the 5% significance level, to see whether there is an association between eye colour and hair colour.

Observed values		Hair colour			Row totals
		Brown	Blonde	Red	
Eye colour	Brown	63	31	6	100
	Blue	26	20	14	60
	Green	11	19	10	40
Column totals		100	70	30	200

Answer

H_0: There is no association between eye colour and hair colour.

H_1: There is an association between eye colour and hair colour.

First define the hypotheses.

We could have stated the hypotheses as:

H_0: Eye colour and hair colour are independent.

H_1: Eye colour and hair colour are not independent.

Expected values		Hair colour			Row totals
		Brown	Blonde	Red	
Eye colour	Brown	50	35	15	100
	Blue	30	21	9	60
	Green	20	14	6	40
Column totals		100	70	30	200

We can now calculate the expected values.

Observed	Expected	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
63	50	3.38
26	30	0.533
11	20	4.05
31	35	0.457
20	21	0.048
19	14	1.786
6	15	5.4
14	9	2.778
10	6	2.667

We calculate $\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ for each pair. In this case, this is called the contribution.

$\chi^2 = \sum\left(\frac{(O_{ij} - E_{ij})^2}{E_{ij}}\right) = 21.098$

Calculate the test statistic.

Degrees of freedom:
$\nu = (r-1)(c-1) = 4$

Calculate the number of degrees of freedom.

$\chi^2_4(0.95) = 9.488$

Find the critical value.

Since $21.098 > 9.488$, there is sufficient evidence to reject H_0.

There is sufficient evidence to suggest an association between eye colour and hair colour.

Write your conclusion to the hypothesis test.

A different way of saying this would be:
There is sufficient evidence to suggest that eye colour and hair colour are not independent.

In Worked example 10.7, each individual $\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ was referred to as a *contribution*.

This is an appropriate name since each value contributes to the test statistic. In fact, the relative size of this can give us more information when we reject the null hypothesis. It shows the major contributing factor that causes us to reject the null hypothesis.

We could use the form $\sum\left(\frac{O_{ij}^2}{E_{ij}}\right) - N$, but then we would not be able to deduce this information.

WORKED EXAMPLE 10.8

A research student collects information regarding the age of adults and the amount of debt that they have accumulated. The information collected is presented in the following table.

		Amount of debt	
		⩽ $7500	> $7500
Age (years)	⩽ 35	45	68
	> 35	15	32

Test, at the 5% level of significance, to decide whether there is an association between age and amount of debt.

Answer

H_0: There is no association between age and amount of debt.

H_1: There is an association between age and amount of debt.

Define the hypotheses.

Observed		Amount of debt		
		⩽ $7500	> $7500	
Age (years)	⩽ 35	45	68	113
	> 35	15	32	47
		60	100	160

Calculate the totals for each row and column.

Expected		Amount of debt	
		⩽ $7500	> $7500
Age (years)	⩽ 35	42.375	70.625
	> 35	17.625	29.375

Calculate the expected values using $E_{ij} = \frac{R_i \times C_j}{T}$.

There is no $E_{ij} < 5$, so we do not need to consider combining categories.

Observed	Expected	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
45	42.375	0.163
15	17.625	0.391
68	70.625	0.098
32	29.375	0.235

$\chi^2 = \sum\left(\frac{(O_{ij} - E_{ij})^2}{E_{ij}}\right) = 0.886$

Calculate the test statistic.

Degrees of freedom:

Calculate the number of degrees of freedom.

$\nu = (r - 1)(c - 1)$

$\nu = (2 - 1)(2 - 1)$

$\chi^2_1(0.95) = 3.841$

Find the critical value.

Since $0.886 < 3.841$, there is insufficient evidence to reject H_0. There is insufficient evidence to state there is an association between age and amount of debt.

E In Worked example 10.8, the number of categories is very small and so does not yield good results. Using a 2×2 contingency table in this case means that the χ^2-distribution does not approximate the test statistic particularly well. In this case it would have been better to use Yates' correction, where each contribution is calculated as $\frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}$.

For AS & A Level Further Mathematics, we are not required to use this, but it is important for further study.

WORKED EXAMPLE 10.9

In a school, the IGCSE results of 380 students are compared to see if there is an association between the grade gained in Mathematics and the grade gained in English. The results are shown in the table.

		Mathematics grade				
		A	B	C	D	E
English grade	A	33	23	9	4	1
	B	23	44	24	8	1
	C	14	30	28	11	2
	D	7	17	25	17	4
	E	1	6	19	22	7

a Calculate a table of expected values.

b Which columns would you combine and why?

c Which rows might you consider combining? State the advantages and disadvantages of combining these rows.

d Combining both rows and columns as suggested, perform a test, at the 1% significance level, to see whether there is an association between grades achieved in English and in Maths.

Answer

a

Observed values		Mathematics grade					
		A	B	C	D	E	
English grade	A	33	23	9	4	1	70
	B	23	44	24	8	1	100
	C	14	30	28	11	2	85
	D	7	17	25	17	4	70
	E	1	6	19	22	7	55
		78	120	105	62	15	380

Calculate the totals for each row and column.

Expected values		Mathematics grade				
		A	B	C	D	E
English grade	A	14.368	22.105	19.342	11.421	2.763
	B	20.526	31.579	27.632	16.316	3.947
	C	17.447	26.842	23.487	13.868	3.355
	D	14.368	22.105	19.342	11.421	2.763
	E	11.289	17.368	15.197	8.974	2.171

Calculate each expected value using $E_{ij} = \frac{R_i \times C_j}{T}$.

b Combining the last two columns will ensure that all $E_{ij} \geqslant 5$.

We need to combine columns if any $E_{ij} < 5$.

c We could also combine the D and E rows. An advantage of this is that we are able to compare grade groupings in the same way: A, B, C, D/E is the same for each subject. The disadvantage of this is that we have lost information and also the critical value is reduced.

d H_0: There is no association between grades achieved in Maths and in English.

H_1: There is an association between grades achieved in Maths and in English.

Define the hypotheses.

Observed data table with totals for each row and column:

Combine the necessary columns/rows.

		Mathematics grade				
		A	B	C	D/E	
English grade	A	33	23	9	5	70
	B	23	44	24	9	100
	C	14	30	28	13	85
	D/E	8	23	44	50	125
		78	120	105	77	380

Expected value table:

		Mathematics grade			
		A	B	C	D/E
English grade	A	14.37	22.11	19.34	14.18
	B	20.53	31.58	27.63	20.26
	C	17.45	26.84	23.49	17.22
	D/E	25.66	39.47	34.54	25.33

O_{ij}	E_{ij}	$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
33	14.37	24.16
23	20.53	0.30
14	17.45	0.68
8	25.66	12.15
23	22.11	0.04
44	31.58	4.89
30	26.84	0.37
23	39.47	6.88
9	19.34	5.53
24	27.63	0.48
28	23.49	0.87
44	34.54	2.59
5	14.18	5.95
9	20.26	6.26
13	17.22	1.04
50	25.33	24.03
	Sum	96.20

Calculate the test statistic.

$\chi^2 = 96.20$

$\nu = (4-1)(4-1) = 9$

Calculate the number of degrees of freedom, using $\nu = (r-1)(c-1)$.

$\chi^2_9(0.99) = 21.67$

Since $96.20 > 21.67$ there is sufficient evidence to reject H_0.

Always add a conclusion, referring to the question.

There is an association between the grades achieved in Maths and in English.

EXERCISE 10D

1 a For the following table of observed data, calculate the expected values E_{11}, E_{31} and E_{33}.

20	30	10
30	50	20
10	20	10

b For the following table of observed data, calculate the expected values E_{12}, E_{24} and E_{13}.

14	20	24	22
8	8	13	11

Write your expected values to 2 decimal places where required.

c For the following table of observed data, calculate the table of expected data

8	12	7
12	11	8
16	23	9
18	12	4

Write your values to 2 decimal places, where required.

2 For the following observed data, write the table of expected data and calculate the test statistic.

a

10	30
20	40

b

36	30
58	76

c

36	30
24	30
58	76

d

16	4	10
14	4	20
18	5	23

3 For each given table of expected data, state how many degrees of freedom would be required.

a

13.44	14.84	13.72
16.64	18.37	16.99
17.92	19.79	18.29

b

10	20	30
21.67	43.33	65
28.33	56.67	85
40	80	120

c

10.52	21.03	31.55	1.90
21.68	43.37	65.05	3.90
28.48	56.96	85.44	5.12
39.32	78.64	117.96	7.08

4 Two categories X and Y are thought to be associated. The table of observed data is:

	Y_1	Y_2
X_1	23	36
X_2	42	136

a Give the table of expected values, to 2 decimal places.

b Calculate the test statistic.

c Test, at the 5% significance level, whether there is an association between category X and category Y. Clearly state your hypotheses.

For each following question, clearly state:

- your hypotheses
- the value of the test statistic
- the degrees of freedom required
- the critical value
- your conclusion.

M 5 A bank manager obtains information on 150 randomly selected loans made by the bank in the previous year. The loans are classified as either good or toxic. The manager also looks at the age groups of the people provided with the loan.

		Age group (years)			Total
		18–25	Over 25–35	Over 35	
Loan type	Good	41	32	27	100
	Toxic	23	17	10	50
	Total	64	49	37	150

Carry out a test, at the 10% significance level, to see if the loan type is independent of age group.

M 6 Last year, 500 students in England entered a poetry competition. Eighty of the entries were published in a book. Each student was required to state which region of England they lived in: north, south, east or west. 140 students indicated they were from the north, 120 from the south and 90 from the east. 15% of students from the north had their poems published, 10% from the south were published and 20% from the east were published.

a Complete the following contingency table.

	N	S	E	W	Total
Selected	21				
Rejected					
Total	140				500

b Test, at the 5% significance level, whether there is an association between being published and the region in which the students lived.

M 7 Residents of three towns, A, B and C, are surveyed on how good their mobile phone reception is while at home, choosing from good, satisfactory or poor. A random sample of responses are gathered from each town and tabulated.

	Good	Satisfactory	Poor
A	32	42	16
B	27	24	24
C	16	9	10

Test, at the 5% significance level, whether there is an association between the town and the quality of mobile phone reception.

WORKED PAST PAPER QUESTION

Random samples of employees are taken from two companies, A and B. Each employee is asked which of three types of coffee (cappuccino, latte, and ground) they prefer. The results are shown in the following table.

	Cappuccino	Latte	Ground
Company A	60	52	32
Company B	35	40	31

a Test, at the 5% significance level, whether coffee preferences of employees are independent of their company.

Larger random samples, consisting of N times as many employees from each company, are taken. In each company, the proportions of employees preferring the three types of coffee remain unchanged.

b Find the least possible value of N that would lead to the conclusion, at the 1% significance level, that coffee preferences of employees are not independent of their company.

Cambridge International AS & A Level Further Mathematics 9231 Paper 21 Q10 June 2012

Answer

a Calculate the totals for each row and column.

	Cappuccino	Latte	Ground
Company A	$\frac{144 \times 95}{250}$	$\frac{144 \times 92}{250}$	$\frac{144 \times 63}{250}$
Company B	$\frac{106 \times 95}{250}$	$\frac{106 \times 92}{250}$	$\frac{106 \times 63}{250}$

First we need to find the expected values.
Remember:
$E_{ij} = \frac{R_i \times C_j}{T}$

	Cappuccino	Latte	Ground
Company A	54.72	52.992	36.288
Company B	40.28	39.008	26.712

H_0: There is no association between the company and coffee preference
H_1: There is an association between the company and coffee preference

Define the hypotheses.

Calculate the test statistic.

O	E	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
60	54.72	0.509 474
52	52.992	0.018 57
32	36.288	0.506 695
35	40.28	0.692 115
40	39.008	0.025 227
31	26.712	0.688 34

$\chi^2 = 2.44$

Because of rounding, 2.45 is also acceptable here.

$\nu = (2 - 1)(3 - 1) = 2$

Calculate number of the degrees of freedom.

$\chi^2_2(0.95) = 5.991$

Compare the critical value with the test statistic.

Since $2.44 < 5.991$, we do not reject H_0. There is insufficient evidence to suggest there is an association between the company and the coffee type.

b Since the proportions of data are the same, just N times bigger, the test statistic $X^2 = N \times 2.44$.

At the 1% significance level, with two degrees of freedom, the critical value is $\chi^2_2(0.99)$.

$\chi^2_2(0.99) = 9.21$

To reject the test, we require $N \times 2.44 > 9.21$.

$N > 3.77$

Hence, $N_{\min} = 4$.

Checklist of learning and understanding

Goodness of fit:

- When fitting data to a distribution, we combine cells to ensure that $E_i \geqslant 5$.

  Then the test statistic is calculated by $\chi^2 = \sum \left(\frac{(O_i - E_i)^2}{E_i} \right) \sim \chi^2(\nu)$ and will have

  ν = number of expected values − 1 − number of estimated parameters.

Known distributions:

Distribution	Degrees of freedom
Binomial	$\nu = n - 1$ if p not estimated $\nu = n - 2$ if p estimated
Poisson	$\nu = n - 1$ if λ not estimated $\nu = n - 2$ if λ estimated
Normal	$\nu = n - 1$ if μ and σ^2 not estimated $\nu = n - 2$ if μ or σ^2 estimated $\nu = n - 3$ if μ and σ^2 estimated

Contingency tables:

- These are used to look for an association between two criteria or independence.
- Each expected value can be found from $E_{ij} = \frac{R_i \times C_j}{T}$, where R_i is the ith row total and C_j is the jth column total.
- Rows or columns can be combined to ensure that each $E_{ij} \geqslant 5$.
- The test statistic is then calculated by $\chi^2 = \sum \left(\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right)$.
- The number of degrees of freedom is $\nu = (r - 1)(c - 1)$ where r is the number of rows and c is the number of columns in the table.

END-OF-CHAPTER REVIEW EXERCISE 10

1 A family was asked to record the number of letters delivered to their house on each of 200 randomly chosen weekdays. The results are summarised below:

Number of letters	0	1	2	3	4	5	$\geqslant 6$
Number of days	57	60	53	25	4	1	0

a It is suggested that the number of letters delivered each weekday has a Poisson distribution. By finding the mean and variance for this sample, comment on the appropriateness of this suggestion.

The following table includes some of the expected values, correct to 3 decimal places, using a Poisson distribution with mean equal to the sample mean for the above data.

Number of letters	0	1	2	3	4	5	$\geqslant 6$
Expected number of days	53.964	70.693	p	q	6.622	1.735	0.463

b i Show that $p = 46.304$, correct to 3 decimal places, and find q.

ii Carry out a goodness of fit test at the 10% significance level.

Cambridge International AS & A Level Further Mathematics 9231 Paper 23 Q10 June 2011

2 A random sample of 200 is taken from the adult population of a town and classified by age group and preferred type of car. The results are given in the following table.

	Hatchback	Estate	Convertible
Under 25 years	32	11	17
Between 25 and 50 years	45	24	6
Over 50 years	31	16	18

Test, at the 5% significance level, whether preferred type of car is independent of age group.

Cambridge International AS & A Level Further Mathematics 9231 Paper 23 Q8 June 2014

3 A random sample of 80 observations of the continuous random variable X was taken and the values are summarised in the following table.

Interval	$2 \leqslant x < 3$	$3 \leqslant x < 4$	$4 \leqslant x < 5$	$5 \leqslant x < 6$
Observed frequency	36	29	9	6

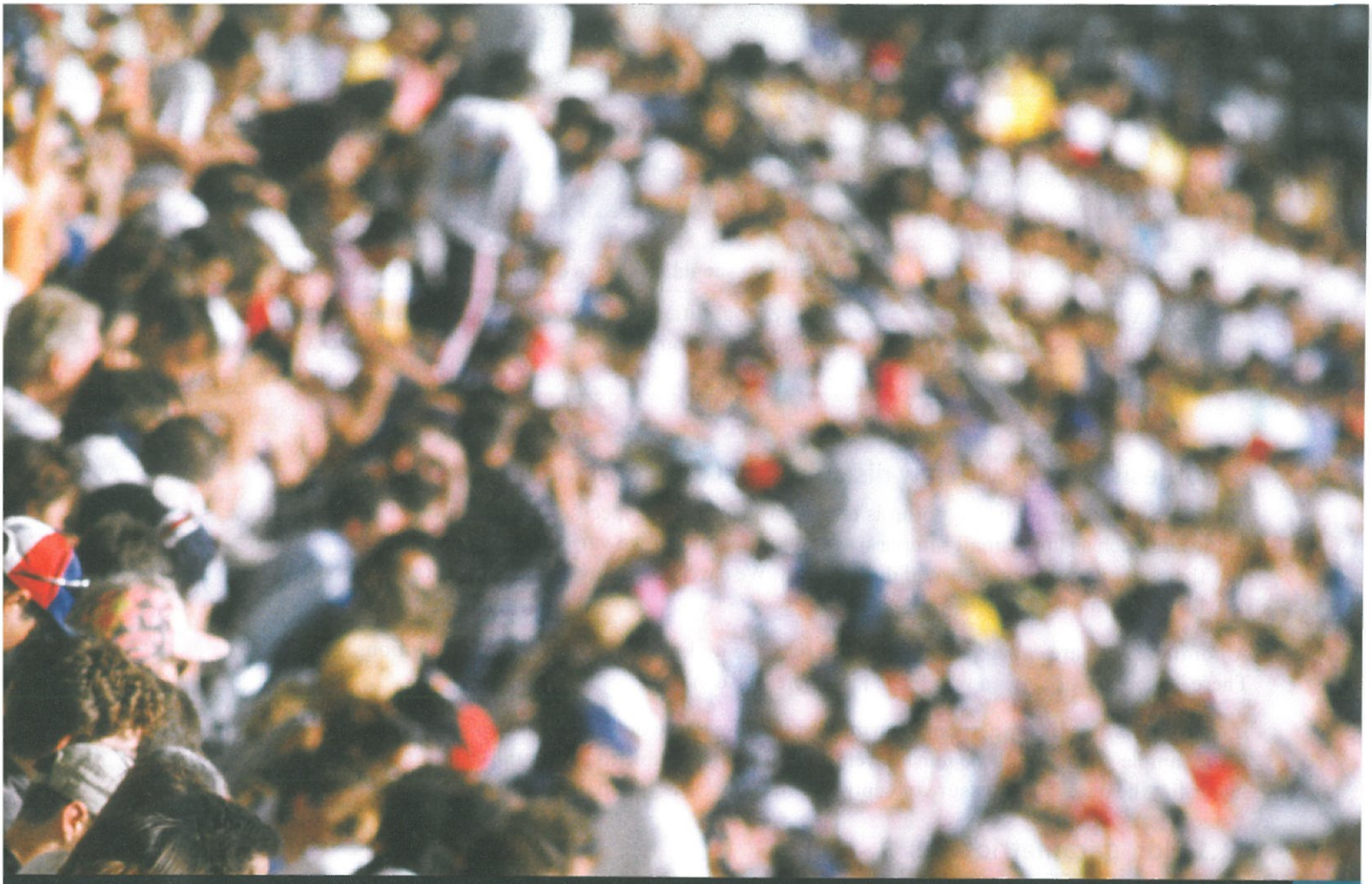
It is required to test the goodness of fit of the distribution having probability density function f given by

$$f(x) = \begin{cases} \dfrac{3}{x^2} & 2 \leqslant x < 6, \\ 0 & \text{otherwise.} \end{cases}$$

a Show that the expected frequency for the interval $2 \leqslant x < 3$ is 40 and calculate the remaining expected frequencies.

b Carry out a goodness of fit test, at the 10% significance level.

Cambridge International AS & A Level Further Mathematics 9231 Paper 22 Q7 June 2013



Chapter 11
Non-parametric tests

In this chapter you will learn how to:

- understand the idea of a non-parametric test and when it might be useful
- understand the basis of the sign test, the Wilcoxon signed-rank test and the Wilcoxon rank-sum test
- use a single-sample sign test and a single-sample Wilcoxon signed-rank test to test a hypothesis concerning a population median
- use a paired-sample sign test, a Wilcoxon matched-pairs signed-rank test and a Wilcoxon rank-sum test, as appropriate, to test for identity of populations.

PREREQUISITE KNOWLEDGE

Where it comes from	What you should be able to do	Check your skills
AS & A Level Mathematics Probability & Statistics 1, Chapter 7	Find probabilities from the binomial distribution.	1 Given $X \sim \text{Bin}(10, 0.5)$, find $P(X \leqslant 2 \cup X \geqslant 9)$.

Hypothesis testing with few distributional assumptions

In Chapter 9 we carried out hypothesis tests for the mean. We specified a number of conditions such as large/small sample size and known/unknown variance and we assumed that there was a underlying normal distribution. We carried out a test using the normal distribution or a t-test.

All of these tests are called parametric tests because we know the underlying distribution. There is a population mean which is a parameter we can test.

If we do not know the underlying distribution then we carry out a **non-parametric test**. If the sample size is small, we must develop new tests to be able to gather information regarding the data. The measure of centrality in these cases is usually the median.

11.1 Non-parametric tests

Sometimes when we wish to perform a hypothesis test, we are not able to assume the type of distribution from which the sample data came. In fact, it may not have a distribution. This causes some problems as we cannot assume a population parameter, for example, the mean or variance. In this case, we say that the data is non-parametric. A variety of non-parametric tests have been developed to cater for this. Each test is based on certain assumptions so, depending what we can assume, we can choose the correct test.

This chapter is split into two main parts, but the ideas are interlinked. In the first part we focus on single-sample statistics. In the second part we focus on two-sample statistics. Most of the ideas in the second part will be introduced in the first part. The assumptions for each test are listed in the following table.

Type of test	Test	Assumptions
Single sample	Sign test	• The underlying data are continuous • The data are independent
	Wilcoxon signed-rank test	• The underlying data are symmetric • The underlying data are continuous • The data are independent
Two sample	Paired sign test	• The data are in matched pairs • The differences between matched pairs are continuous • The data are independent
	Wilcoxon matched-pairs signed-rank test	• The data are in matched pairs • The differences between matched pairs are symmetric • The differences between matched pairs are continuous • The data are independent
	Wilcoxon rank-sum test	• The two samples are independent • The underlying data are symmetric • The underlying data are continuous

11.2 Single-sample sign test

We can use the single-sample sign test when we wish to see whether data differ from a stated value for the median. It is important to understand that the median is *not* a parameter, as we do not know the underlying distribution. This test is based on the assumptions listed in the previous table.

To perform the single-sample sign test, mark the values that are greater than the stated median with a + sign, and mark those that are less than the stated median with a − sign. If the data are well distributed about the median, we would expect an equal number of + and − signs. So there should be a probability of 0.5 that any data point is above the median and a probability of 0.5 that it is below the median.

Some people think that the single-sample sign test is quite a crude method, but it does give some useful information. The single-sample sign test is a special case of the binomial test, when n is the number of data points and the probability of 'success' is 0.5, as shown in Key point 11.1.

KEY POINT 11.1

Given n data points, a single-sample sign test is created using $X \sim \text{Bin}(n, 0.5)$. The test statistic can be the number of + signs, that is the number of data points greater than the median. We can calculate the probability that X is above this test statistic, below this test statistic, or either in the case of a two-tailed test.

This can be expressed as $P(X \leqslant ts \mid X \sim \text{Bin}(n, 0.5))$ or $P(X \geqslant ts \mid X \sim \text{Bin}(n, 0.5))$ where ts stands for test statistic.

WORKED EXAMPLE 11.1

It is believed that the following dataset comes from a population with median 135.

150	130	125	140	170
140	190	180	175	165
160	130	140	140	145

Perform a single-sample sign test, at the 5% significance level, to test this claim.

Answer

H_0: The population median is 135.

H_1: The population median is not 135.

First, state the hypotheses.

Notice that this is a two-tailed test.

Value	Sign	Value	Sign
150	+	140	+
140	+	140	+
160	+	175	+
130	−	140	+
190	+	170	+
130	−	165	+
125	−	145	+
180	+		

Consider which values are above or below the stated median.

Here, the test statistic is 12, as there are 12 values above the stated median.

Consider $X \sim \text{Bin}(15, 0.5)$:

Since 12 is greater than $\frac{n}{2}$, which is 7.5, we need consider only the top tail.

$$P(X \geqslant 12) = {}^{15}C_{12}(0.5)^{15} + {}^{15}C_{13}(0.5)^{15} + {}^{15}C_{14}(0.5)^{15} + {}^{15}C_{15}(0.5)^{15}$$

$$P(X \geqslant 12) = 0.017\,578$$

Since $0.017\,578 < 0.025$, the test statistic of 12 is in the critical region and, therefore, we reject H_0.

Since we are looking at a two-tailed test, we consider 2.5% as our critical value.

There is sufficient evidence to suggest the population median is not 135.

E In a situation where we have zero instead of + or −, the data point is discounted. This is not required in this course.

It is possible to approximate the sign test to a normal distribution for large n ($n > 10$ is considered large here), as shown in Key point 11.2.

KEY POINT 11.2

Let $S = min$(number of + signs, number of − signs) then $\text{E}(S) = \frac{n}{2}$, $\text{Var}(S) = \frac{n}{4}$.

For large n (> 10), $T \sim \text{N}\left(\frac{n}{2}, \frac{n}{4}\right)$, we can use the normal approximation of the binomial with $p = 0.5$. We must also make sure that we use a continuity correction. As we are approximating a discrete distribution with a continuous distribution, our z-value is:

$$z = \frac{S^+ - \mu + 0.5}{\sigma}$$

For example, if the test statistic is $S^+ = 5$, and we have approximated to $T \sim \text{N}(15, 7.5)$, we calculate the z-value as $z = \frac{5.5 - 15}{\sqrt{7.5}}$ since any value from 5 up to 5.5 rounds down to 5, and we are looking in this case at $\text{P}(X < 15)$.

EXERCISE 11A

1 For the following dataset, calculate S^+ based on each of the stated null hypotheses.

133	132	155	145	157
140	152	167	148	158
163	126	179	126	149
182	138	172	178	166

a H_0: population median is 150.
b H_0: population median is 141.
c H_0: population median is 156.
d H_0: population median is 165.

2 For each of the following, find $P(X \leqslant S^+ | X \sim \text{Bin}(n, 0.5))$.

a $S^+ = 3, n = 12$ **b** $S^+ = 2, n = 9$ **c** $S^+ = 4, n = 15$ **d** $S^+ = 5, n = 20$

3 For each of the following, find $P(X \geqslant S^+ | X \sim \text{Bin}(n, 0.5))$.

a $S^+ = 11, n = 14$ **b** $S^+ = 10, n = 12$ **c** $S^+ = 13, n = 15$ **d** $S^+ = 17, n = 25$

4 For the given hypotheses, sample size and value of S^+, state whether the null hypothesis should be rejected or not rejected at the 5% significance level. Do not use a normal approximation.

a H_0: the population median is 38.
H_1: the population median is less than 38.
$S^+ = 4, n = 15$

b H_0: the population median is 16.2.
H_1: the population median is not 16.2.
$S^+ = 5, n = 20$

c H_0: the population median is 154.
H_1: the population median is greater than 154.
$S^+ = 11, n = 15$

5 For each of the tests in question **4a–c**, state $E(S)$ and $\text{Var}(S)$ and hence the value of the test statistic, when approximating to the normal distribution.

M **6** Acid rain is a problem for many areas. Rain is described as being acidic if its pH level is below 5.2. In a city, a biologist takes 12 water samples and measures the pH.

4.73	5.06	4.87	4.88
5.04	5.16	5.07	5.09
5.11	5.24	5.25	5.25

Test, at the 5% significance level, whether the rain in this city is acidic.

M **7** During 2017, the weekly amount of pocket money given to a random sample of 15-year-old children living in Chicago is:

$9.50, $10.50, $12.50, $7.50, $20.00, $13.00, $14.00, $16.00, $17.50, $10.00

During 2016, the median amount of pocket money given to 15-year-olds in Chicago was $11.00. Carry out a sign test to investigate the claim that the median weekly amount of pocket money given to 15-year-old children in Chicago has increased since 2016. Test this claim at the 10% level of significance.

M 8 A website advertises used cars for sale. During September 2018, 12 cars of similar age and of the same model are for sale. The asking prices for the cars ($AUS) are:

5999, 8900, 7000, 6499, 7500, 7999, 8450, 6500, 7250, 8150, 4999, 5600

a Investigate, at the 10% significance level, whether the median asking price for such cars is $7675.

b Still using the 10% level of significance, above what value would the median need to be to make the test significant?

M 9 The tax office claims that it takes 60 minutes to fill out their tax form. A researcher believes that it takes longer than this. A random sample of 20 people are selected, and the recorded times taken to complete the form are listed below.

55	62	63	68	70
71	58	62	64	69
69	72	59	62	66
68	69	72	69	63

Use a suitable approximation to test the claim, at the 5% significance level, that the time taken to complete the form is more than 60 minutes.

11.3 Single-sample Wilcoxon signed-rank test

If we know that our underlying data are symmetric, we can refine the sign test by performing a Wilcoxon signed-rank test. This test factors in the magnitude of the rank, as well as whether it is above or below the median. Another condition placed on the use of this test is that the underlying data are continuous.

To perform the single-sample Wilcoxon signed-rank test, we rank the differences in the data points from the stated population median. The test statistic is the smaller value of the sums of the negative ranks and the sums of the positive ranks.

As shown in Key point 11.3, a Wilcoxon signed-rank test can be performed under the conditions that:

- the underlying data are symmetric
- the underlying data are continuous
- the data are independent.

KEY POINT 11.3

Where P is the sum of the ranks corresponding to the positive differences from the stated median and N is the sum of ranks corresponding to the negative differences from the stated median:

$$T = \min(P, N)$$

T is the test statistic for the Wilcoxon signed-rank test. Critical values can be found in the statistical tables. If the test statistic is below the critical value, we reject H_0.

Even though the data are continuous, we are measuring the sums of ranks, and so the distribution of T is *discrete*. Also, it is worth noting that P can fall between the values 0 and $\frac{n(n+1)}{2}$. As all of the data points lie about the stated population median, their ranks will be between 1 and n, and the sum of these integers is $\frac{n(n+1)}{2}$.

The closer our test statistic is to 0, the more extreme the data; that is, the more likely data are to be above or below the stated population median. This is why we need our test statistic to be below the critical value from the tables.

WORKED EXAMPLE 11.2

The weights (in kg) of ten randomly selected Spanish mackerel are recorded:

1.6, 1.1, 2.1, 2.4, 2.2, 2.9, 2.6, 2.3, 2.7, 1.9

Test, at the 5% significance level, whether the median weight is greater than 1.8 kg.

Answer

H_0: The population median weight of Spanish mackerel is 1.8 kg.

H_1: The population median weight of Spanish mackerel is greater than 1.8 kg.

Define the hypotheses.

Weight, W_i	W_i – Median	P	N
1.6	−0.2		2
1.1	−0.7		7
2.1	0.3	3	
2.4	0.6	6	
2.2	0.4	4	
2.9	1.1	10	
2.6	0.8	8	
2.3	0.5	5	
2.7	0.9	9	
1.9	0.1	1	
	Sums:	46	9

To perform this test, we first need to rank the magnitude of differences of each data point from the stated population median. Ignoring signs, start with the smallest difference and give this rank 1, the next smallest difference is given rank 2 and so on.

We can check P and N here using the fact that:

$$P + N = \frac{n(n+1)}{2}$$

So the test statistic here is $T = \min(P, N) = 9$.

We look up the critical value in the statistical tables:

	Level of significance			
One-tailed	0.05	0.025	0.01	0.005
Two-tailed	0.1	0.05	0.02	0.01
$n = 6$	2	0		
7	3	2	0	
8	5	3	1	0
9	8	5	3	1
10	(10)	8	5	3
11	13	10	7	5

We are conducting a one-tailed test here at the 5% significance level.

The critical value here is 10.

Since $9 < 10$, (test statistic < critical value), there is sufficient evidence to reject H_0.

Be careful, as we require test statistic < critical value to reject H_0 here. This is different from the other tests performed in Chapter 9. We are testing whether the test statistic is significantly smaller than would happen by chance.

There is sufficient evidence to suggest that the population median is not 1.8 kg.

Write a conclusion in context.

E If the ranks are tied, for example, two values both have rank 3, they occupy the 3rd and 4th placings and so we allocate them a tied rank of 3.5. However, this is beyond the scope of this course.

It is possible to approximate the Wilcoxon signed-rank test to a normal distribution for large n, as shown in Key point 11.4.

KEY POINT 11.4

Given the statistic $T = \min(P, N)$, then:

$$E(T) = \frac{n(n+1)}{4}$$

$$\text{Var}(T) = \frac{n(n+1)(2n+1)}{24}$$

And for large n:

$$T \sim N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$$

We use a continuity correction since we are approximating a discrete distribution with a continuous distribution. Our z-value is:

$$z = \frac{T - \mu + 0.5}{\sigma}$$

WORKED EXAMPLE 11.3

In a clinical trial, the survival times, in weeks, for 19 patients with non-Hodgkin's lymphoma are recorded.

37	54	73	89	94	110	112	123	129	132
148	151	173	189	201	204	213	276	281	

Test, at the 5% significance level, whether the median differs from 150.

Answer

H_0: The population median is 150.

State the hypotheses.

H_1: The population median is different from 150.

W_i	W_i – Med	$\lvert W_i$ – Med$\rvert$	P	N
37	−113	113		17
54	−96	96		16
73	−77	77		15
89	−61	61		13
94	−56	56		12
110	−40	40		9
112	−38	38		7
123	−27	27		6
129	−21	21		4
132	−18	18		3
148	−2	2		2
151	1	1	1	
173	23	23	5	
189	39	39	8	
201	51	51	10	
204	54	54	11	
213	63	63	14	
276	126	126	18	
281	131	131	19	
		Sum	86	104

Set up the table of ranks for the data.

Ignoring signs, start with the smallest difference and give this rank 1, the next smallest difference is given rank 2 and so on.

We can check

$$P + N = 86 + 104 = 190 = \frac{19(19+1)}{2} = \frac{n(n+1)}{2}$$

$T = \min(P, N) = 86$

$$\mathrm{E}(T) = \frac{n(n+1)}{4} = \frac{19 \times 20}{4} = 95$$

$$\mathrm{Var}(T) = \frac{n(n+1)(2n+1)}{24} = \frac{19 \times 20 \times 39}{24} = 617.5$$

Calculate $\mathrm{E}(T)$ and $\mathrm{Var}(T)$ so we can approximate to the normal.

$$z = \frac{86.5 - 95}{\sqrt{617.5}}$$
$$= -0.342$$

Use this statistic from $T \sim N(95, 617.5)$ and standardise it using $z = \frac{T - \mu + 0.5}{\sigma}$.

$P(Z \leqslant -0.342) = 0.3662$

Since this is negative, but two-tailed, we consider only the bottom tail.

Since $0.3662 > 0.025$, we do not reject H_0.

Since this is greater than 2.5%, it is not in the critical region.

We could instead have compared -0.342 with the critical value for the two-tailed test, -1.96.

Since $-0.342 > -1.96$, we do not reject H_0.

There is insufficient evidence to suggest that the population median differs from 150.

E When using tied ranks (which is beyond this course) the calculation for the variance overestimates the variance. To compensate, we count the number of ranks that are tied, t, and reduce the variance by $\frac{t^3 - t}{48}$. So:

$$\text{Var}(T) = \frac{n(n+1)(2n+1)}{24} - \frac{t^3 - t}{48}$$

EXERCISE 11B

1 Assuming that a Wilcoxon signed-rank test is appropriate for the data, calculate T (the test statistic) based on the null hypotheses stated.

67	81	94
71	88	97
72	90	102
75	91	104
77	92	105

a H_0: population median is 85.

b H_0: population median is 100.

c H_0: population median is 70.

2 For each sample size and significance level given, state the critical value for a Wilcoxon signed-rank test.

a $n = 8$, 5% significance, one-tailed

b $n = 15$, 1% significance, one-tailed

c $n = 18$, 2% significance, two-tailed

d $n = 9$, 10% significance, two-tailed

3 State the assumptions required for the use of the Wilcoxon signed-rank test.

4 For each of the following tests, find:

- $E(T)$ and $\text{Var}(T)$
- the test statistic when approximating to the normal distribution.

Also state whether you would reject or not reject the null hypothesis.

a H_0: the population median is 142.1.
H_1: the population median is less than 142.1.
5% significance level
$T = 175$, $n = 30$

b H_0: the population median is 40.6.
H_1: the population median is not 40.6.
10% significance level
$T = 59$, $n = 20$

c H_0: the population median is 16.3.
H_1: the population median is greater than 16.3.
2.5% significance level
$T = 260$, $n = 40$

M 5 A psychology student carries out a test on short-term memory. She shows 20 commonly used words to ten 18-year-old males. As soon as the 20 words have been shown, the psychologist asks the participants to write down as many words as they can remember in five minutes. The sample of 18-year-old males can be seen as representative of the population of 18-year-old males.

The number of words correctly remembered by the participants are:

15, 7, 12, 14, 11, 10, 4, 13, 9, 2

The median number of words remembered by 65-year-old males in this test is four. Carry out a Wilcoxon signed-rank test, at the 5% level of significance, to investigate whether the median number of words remembered by the 18-year-old males is greater than that for the 65-year-old males.

M 6 Trials are carried out on a new tablet to help ease joint pain for people with chronic arthritis. A randomly selected sample of eight patients who have been suffering from arthritis are given the new tablets. Each participant measures the time it takes for the pain to stop after taking a new tablet as soon as they wake in the morning. The times, in minutes, are:

34, 44, 25, 30, 8, 27, 41, 31

The average waiting time for the old type of tablet is 43 minutes after awakening.

a Carry out a Wilcoxon signed-rank test, at the 5% significance level, to investigate whether the new tablets offer faster pain relief.

b Give a reason why the Wilcoxon signed-rank test might be preferred to a sign test.

M 7 The student council of a large school believes that the average time that the A Level students spend on individual study has increased because students are more aware of the need to achieve high grades. In 2018, the average time per week of the school term that students spent on individual study was 11.2 hours.

A random sample of ten students are asked to record the amount of time on individual study for three weeks during October 2016. The average times, in hours, per week are then calculated:

12, 13.2, 14.1, 10.8, 9.6, 11.3, 17.6, 14.3, 12.1, 19.2

Test, at the 5% level of significance, whether the average amount of time spent on individual study has increased from 2015.

M 8 Managers at a busy international airport are studying the times taken by arriving passengers to pass immigration, collect their luggage, then pass through customs. It is known that in the past this was 50 minutes. Some changes have been made to the queuing system in the hope of reducing this time. A random sample of 45 arriving passengers is taken and the rank sums calculated as $P = 55$, $N = 410$. Using a suitable approximation, test, with a 1% significance level, whether the median waiting time has reduced.

11.4 Paired-sample sign test

We can extend the idea of the sign test to work with paired-sample data by looking for a positive or negative difference. Nevertheless, the principles behind the sign test remain the same.

WORKED EXAMPLE 11.4

Data are collected on the time, in seconds, it takes nine children to tie up their left shoelace and their right shoelace.

Child	Left (s)	Right (s)
A	42	45
B	38	36
C	51	52
D	42	39
E	31	35
F	48	49
G	61	62
H	38	39
I	44	45

Test, at the 10% level of significance, whether there is a difference in the time it takes for the children to tie each shoelace.

Answer

H_0: There is no difference in the time taken to tie their left and right shoelaces.

Define the hypotheses.

H_1: There is a difference in the time taken to tie their left and right shoelaces.

Child	Left (s)	Right (s)	
A	42	45	–
B	38	36	+
C	51	52	–
D	42	39	+
E	31	35	–
F	48	49	–
G	61	62	–
H	38	39	–
I	44	45	–

Set $L_i - R_i$ as the difference.

The test statistic is 2.

Let the number of + signs be the test statistic.

$P(X \leqslant 2) = {}^9C_0(0.5)^9 + {}^9C_1(0.5)^9 + {}^9C_2(0.5)^9$

$P(X \leqslant 2) = 0.089\,844$

Use: $X \sim \text{Bin}(9, 0.5)$

The test is two-tailed, but we need to consider only the lower tail.

Since $0.089\,844 > 0.05$, the test statistic of 2 is not in the critical region. Therefore, there is insufficient evidence to reject H_0. There is insufficient evidence to say there is a difference in the times taken for children to tie their left and right shoelaces.

The probability will be 5%, as the test is two-tailed.

EXERCISE 11C

1 For the following paired datasets, a paired sign test will be performed. Calculate S^+.

a

11	7	10	7	13	6	6	9	11
12	13	7	8	8	5	6	8	12

b

16	22	12	23	19	15
17	22	14	22	18	17

c

163	162	166	157	158	153
160	162	163	158	167	156

2 a For the following dataset, state the value of n to be used in the paired sign test. Give a reason for your answer.

A	160	158
B	159	159
C	167	158
D	166	163
E	162	163
F	163	166
G	166	165
H	159	164
I	166	166
J	166	164
K	161	161
L	159	162
M	158	157
N	161	166

b Find the value of S^+ for the dataset in part **a**.

3 In each case, state whether you would reject or not reject H_0 at the stated significance level.

H_0: the population medians are equal.

H_1: the population medians are not equal.

a $n = 9$, $S^+ = 3$, 10% significance level

b $n = 6$, $S^+ = 1$, 5% significance level

c $n = 8$, $S^+ = 1$, 10% significance level

4 For the dataset in question **2a**:

a state whether a normal approximation would be appropriate, giving a reason for your answer

b find $E(S^+)$ and $Var(S^+)$

c by assuming that a normal approximation would be appropriate, find the test statistic that would be used in performing a hypothesis test.

M 5 A new drug to help ease bronchitis, a lung infection, is developed and needs to be inhaled using an aerosol. Two types of aerosol (X and Y) have been developed and the company wishes to test whether there is a difference in the average effectiveness of the aerosols. Ten patients participate in the trial, in which the patient breathes in before using the aerosol and afterwards. The percentage increase in air intake is measured and recorded.

	A	B	C	D	E	F	G	H	I	J
X	27	22	11	41	19	57	48	41	35	37
Y	23	18	7	18	24	63	31	23	21	28

Test, at the 5% significance level, whether or not there is any difference in the average effectiveness of the aerosols.

M 6 An eye hospital treats a large number of patients who have one eye normal, and the other eye suffers from a thinning of the cornea. Seven such patients are randomly selected and the thickness of their cornea on their good eye and poor eye is measured, in micrometres.

	A	B	C	D	E	F	G
Good eye	512	502	516	484	476	390	498
Poor eye	503	505	493	480	477	355	491

Using a sign test, at the 10% significance level, investigate whether there is any difference in the thickness of the cornea between the two eyes.

M 7 At a research centre, a trial is conducted to see if a new fertiliser gives a better yield of potatoes than the usual fertiliser. Ten plots of land are available for the trial. Each plot is spilt into two equal halves: one half is treated with the new fertiliser, the other half with the usual fertiliser. The following table gives the yield, in kg, per half plot.

	Plot									
	1	2	3	4	5	6	7	8	9	10
Usual	20.2	22.0	17.8	20.6	26.8	20.9	21.2	16.5	20.8	12.9
New	20.8	24.3	17.0	21.2	27.7	19.4	22.6	17.3	20.9	13.1

Using a sign test, at the 5% significance level, investigate whether the new fertiliser produces an increased average yield.

11.5 Wilcoxon matched-pairs signed-rank test

The Wilcoxon matched-pairs signed-rank test is used when we have matched pairs of data, as for the sign test, but when we can assume that the differences in the pairs of the data are symmetric, as shown in Key point 11.5. The process is the same as for the single-sample Wilcoxon signed-rank test. Worked example 11.5 demonstrates how to use the Wilcoxon matched-pairs signed-rank test.

KEY POINT 11.5

When we have matched pairs of data of unknown distributions, but the differences between them are thought to be symmetric, it is appropriate to use a Wilcoxon matched-pairs signed-rank test. We test to see whether the paired-difference median is 0.

WORKED EXAMPLE 11.5

An investigation is carried out into the effectiveness of two types of post-operative pain relief drug: Drug 1 and Drug 2. Seven adults agree to take Drug 1 on one day, and Drug 2 on the second. The time, in hours, of pain relief is recorded.

	Drug 1	Drug 2
A	4.1	3.9
B	3.2	3.3
C	5.3	5.0
D	5.1	4.6
E	4.2	4.6
F	3.8	3.2
G	3.6	4.3

Test, using the matched-pairs Wilcoxon signed-rank test, at the 5% significance level, whether Drug 2 gives longer pain relief than Drug 1.

Answer

H_0: The times are the same before and after.

H_1: The times afterwards have increased.

Define the hypotheses.

This is a one-tailed test.

Before	After	Difference	P	N
4.1	3.9	0.2	2	
3.2	3.3	−0.1		1
5.3	5	0.3	3	
5.1	4.6	0.5	5	
4.2	4.6	−0.4		4
3.8	3.2	0.6	6	
3.6	4.3	−0.7		7
		Sum	16	12

First calculate the test statistic.

Calculate differences and rank them, keeping track of positive and negative differences.

$T = \min(P, N) = 12$

And so the test statistic = 12.

Find the critical value in the statistical tables:

We are carrying out a 5% one-tailed test.

	Level of significance			
One-tailed	0.05	0.025	0.01	0.005
Two-tailed	0.1	0.05	0.02	0.01
$n = 6$	2	0		
7	(3)	2	0	
8	5	3	1	0
9	8	5	3	1
10	10	8	5	3
11	13	10	7	5

Since $12 > 3$ (test statistic > critical value), there is insufficient evidence to reject H_0.

There is insufficient evidence to suggest that Drug 2 gives longer pain relief.

EXERCISE 11D

1 For the following pairs of data, calculate the value of $T = \min(P, N)$.

a

A	B
6.2	6.7
7.6	7.3
5.7	5.3
6.2	6
8.4	7.8
7.2	7.3

b

A	B
15.4	13.8
13.1	13.4
14	14.1
13.8	14.1
15.4	15
14	14.7
15.2	15
15.4	14.4
15.1	13.6

c

A	B
12.1	13
13.6	13
14.2	13.5
15.3	14.5
14.7	14.5
14.9	13

2 For the hypotheses, sample size and test statistic given, state whether you would reject or not reject the null hypothesis.

a H_0: the population medians are equal.
H_1: the population medians are not equal.
$n = 8$, $T = 4$, significance level 5%

b H_0: the scores are the same.
H_1: the scores have decreased.
$n = 14$, $T = 20$, significance level 5%

c H_0: the weights are the same.
H_1: the weights have increased.
$n = 9$, $T = 6$, significance level 2%

3 For a sample size of $n = 18$, a Wilcoxon signed-rank test is to be carried out. Where $T = \min(P, N)$:

a Find $E(T)$.

b Find $\text{Var}(T)$.

c Given that the value of T is 59 and that a normal approximation is appropriate, calculate the test statistic to be used.

d For the given hypotheses, state whether at 5% you would reject or not reject the null hypothesis.

H_0: the population medians are equal.

H_1: the population medians are not equal.

M 4 Percentage marks are obtained for a random sample of eight A Level students for their AS & A Level examinations in Statistics and in Mechanics. The following table shows their results.

	Statistics	Mechanics
A	53	57
B	64	62
C	72	79
D	61	58
E	72	67

	Statistics	Mechanics
F	58	52
G	59	60
H	71	79

a Test, at the 10% significance level, whether there is a difference in the students' scores in Mechanics and Statistics.

b State any assumptions you have made.

M 5 The manufacturer of a brand of smartphone wishes to know what its customers think about the performance of the phone before and after introducing a new processor. The manufacturer selects, at random, ten customers. Each customer is given a phone without the new processor and a phone with the new processor. Each customer is then asked to rate 'on a scale of 1–20' the performance of each smartphone (20 being the best).

	A	B	C	D	E	F	G	H	I	J
Original processor	15	17	13	17	14	12	13	7	7	10
New processor	12	15	18	18	10	19	19	15	17	19

Test, at the 5% significance level, whether customers think that the phone with the new processor is better.

M 6 A particular type of tree can develop a virus that creates black spots on the leaves. The more spots, the more the virus has infected the tree. Two different virus treatments have been developed: treatment X and treatment Y. A sample of eight leaves is chosen. Each leaf is divided into two equal sections without removing it from the tree. On one half, treatment X is used and on the other, treatment Y. The number of black spots on each half is given in the table.

	A	B	C	D	E	F	G	H
X	41	37	34	12	7	19	23	16
Y	24	21	30	11	5	14	20	9

Test, at the 5% significance level, whether there is a difference between the two treatments.

M 7 It is believed that identical twins have similar IQ levels. Thirty pairs of identical twins participate in an IQ test and their results are recorded. There are no tied or zero ranks. The sum of positive ranks is 272, and the sum of negative ranks is 193.

a Find $\mathrm{E}(T)$.

b Find $\mathrm{Var}(T)$.

c Using a suitable approximation, test, at the 5% significance level, whether there is a difference between the IQ scores of a set of identical twins.

11.6 Wilcoxon rank-sum test

The Wilcoxon matched-pairs signed-rank test requires data to be paired, and groups of data must be of equal size. What if we have two independent groups of different sizes and we want to test for a difference in their medians? To do this, we perform the Wilcoxon rank-sum test to see whether the data are from the same distribution. This test is very similar in design to the independent t-test.

To perform the Wilcoxon rank-sum test, we rank the data first, as if it were from one population. We then sum the ranks for each group separately. These sums are the test statistics.

The calculation of the Wilcoxon rank-sum test statistic is quite tricky. We have two samples, one of size m and the other of size n, and we let $m \leqslant n$. Let R_m be the sum of the ranks from the group of size m. We have not defined which way we should rank the data.

Let R_m be the ranking given, then $m(n+m+1) - R_m$ will create the rank sum of the smaller sample when ranked the opposite way round, explained as follows.

There are $n+m$ data points in total. If a data point is the ath value when ranked one way and the bth value when ranked the other, then the sum of these two ranks will be $n+m+1$.

We have m data points in the smaller sample, so the sum of all of these sums of ranks is $m(n+m+1)$. To find the rank sum when they are ranked the other way round, we use $m(n+m+1) - R_m$.

To avoid having to think too carefully about the ranking order, we define the test statistic as $W = \min(R_m, m(n+m+1) - R_m)$, as shown in Key point 11.6.

We can find the critical values from the data tables given.

KEY POINT 11.6

If two samples have sizes m and n, where $m \leqslant n$, R_m is the sum of the ranks of the items in the sample of size m, the test statistic is:

$$W = \min(R_m, m(n+m+1) - R_m)$$

WORKED EXAMPLE 11.6

Researchers are investigating the effect of vitamin B12 on the size of the brain. A sample of males aged between 25 and 40 years is selected. Nine of them are known to have low B12 levels and seven are known to have high B12 levels. After a brain scan, the ratio of brain volume to skull capacity is recorded.

Low B12 levels	0.795	0.798	0.802	0.805	0.806	0.807	0.808	0.81	0.812
High B12 levels	0.786	0.789	0.792	0.796	0.799	0.8	0.803		

Carry out a Wilcoxon rank-sum test, at the 5% significance level, to see whether the level of vitamin B12 affects the size of the brain.

Answer

H_0: level of B12 has no effect on brain size.

H_1: level of B12 has an effect on brain size.

We can also state H_0 as the samples are from the same population.

		Low B12	High B12
0.812	1	1	
0.810	2	2	
0.808	3	3	
0.807	4	4	
0.806	5	5	
0.805	6	6	
0.803	7		7
0.802	8	8	
0.800	9		9
0.799	10		10
0.798	11	11	
0.796	12		12
0.795	13	13	
0.792	14		14
0.789	15		15
0.786	16		16
	Sum	53	83

First, rank the whole dataset. Note which group each value comes from. Then add up the ranks for each category.

Use the value of sums of the group with the smaller sample size.

The following table shows what we would get if we ranked them the other way round.

		Low B12	High B12
0.812	16	16	
0.810	15	15	
0.808	14	14	
0.807	13	13	
0.806	12	12	
0.805	11	11	
0.803	10		10
0.802	9	9	
0.800	8		8
0.799	7		7
0.798	6	6	
0.796	5		5
0.795	4	4	
0.792	3		3
0.789	2		2
0.786	1		1
	Sum	100	36

Calculate the test statistic.

$R_m = 83$ since this is the rank sum from the smaller-sized sample.

$m(n + m + 1) - R_m = 7(9 + 7 + 1) - 83 = 36$

The test statistic is the minimum of 83 and 36, which is $W = 36$.

Find the critical value.

	Level of significance		
One-tailed	0.05	0.025	0.01
Two-tailed	0.1	0.05	0.02
n	$m = 7$		
7	39	36	34
8	41	38	35
9	43	(40)	37
10	45	42	39

Since $36 < 40$, there is sufficient evidence to reject H_0.

There is evidence to suggest that level of vitamin B12 affects brain size.

For large n and m ($n \geqslant 10, m \geqslant 10$) it is possible to approximate W as a normal distribution, $W = \min(R_m, m(n+m+1) - R_m)$ as shown in Key point 11.7.

KEY POINT 11.7

For large n and m ($n \geqslant 10, m \geqslant 10$) it is possible to approximate W as a normal distribution:

$$E(W) = \frac{m(n+m+1)}{2}$$

$$\text{Var}(W) = \frac{mn(n+m+1)}{12}$$

We must also make sure that we use a continuity correction. Since we are approximating a discrete distribution with a continuous distribution, our z-value is $z = \frac{W - \mu + 0.5}{\sigma}$.

DID YOU KNOW?

There is another test that is statistically equivalent to the Wilcoxon rank-sum test. This is the Mann–Whitney U-test. The only difference is the measure of the test statistic, and hence the table of critical values, but it is equivalent. You may come across Mann–Whitney U-test in the Social Sciences.

WORKED EXAMPLE 11.7

A company is investigating a new production technique to improve the quality of camera lenses for a phone. Samples of the lenses are given to a camera expert who is asked to rank the lenses, with rank 1 being the highest quality. The expert does not know which production technique has been used.

Lens	A	B	C	D	E	F	G	H	I	J	K	L
Method	old	new	new	old	old	new	old	new	old	old	old	new
Rank	12	1	2	9	10	5	21	6	20	22	23	17

Lens	M	N	O	P	Q	R	S	T	U	V	W	X
Method	new	new	old	old	old	new	old	new	old	new	new	old
Rank	14	13	3	4	19	11	24	16	18	8	7	15

Using a suitable approximation as shown in Key point 11.7, test, at the 5% significance level, whether there is a difference in the quality of production techniques.

Answer

H_0: There is no difference in the quality of the two samples.

Define the hypotheses.

H_1: There is a difference in the quality of the two samples.

$m = 11$ (new)

$n = 13$ (old)

Since we are approximating, we need to know only the sizes of the two samples and the rank sum.

$$E(R_m) = \frac{m(n+m+1)}{2} = \frac{11(25)}{2} = 137.5$$

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

$$\text{Var}(R_m) = \frac{mn(n+m+1)}{12} = \frac{11 \times 13(25)}{12} = 297\frac{11}{12}$$

$R_m = 1 + 2 + 5 + 6 + 17 + 14 + 13 + 11 + 16 + 8 + 7$

The test statistic is the minimum of R_m and $m(n + m + 1) - R_m$.

$R_m = 100$

$m(n + m + 1) - R_m = 175$

And so $W = \min(R_m, m(n + m + 1) - R_m) = 100.$

$$z = \frac{100.5 - 137.5}{\sqrt{\left(297\frac{11}{12}\right)}} = -2.144$$

Find the z-test statistic, remembering to make the continuity correction.

$P(Z \leqslant -2.144) = 1 - 0.984 = 0.0160$

Since we have a two-tailed test, compare the probability with the critical value for 2.5%.

Since $0.0160 < 0.025$, the test statistic is in the critical region and so we have sufficient evidence to reject H_0.

We could have compared -2.144 with the critical value for the two-tailed test, -1.96.

Since $-2.144 < -1.96$, the test statistic is in the critical region and so we have sufficient evidence to reject H_0.

There is a difference in quality between samples of camera lenses made by different production techniques.

EXPLORE 11.1

We are given a table of critical values for all of the tests that we carry out. Sometimes it is not clear where these values come from. In this activity, we shall find some of the critical values of the Wilcoxon rank-sum test.

Let us consider the situation where we have sample sizes $m = 4$ and $n = 6$. Here we have ten ranks, four of which must be assigned to the sample of size four.

The first case is the rankings $\{1, 2, 3, 4\}$, with the rank sum of 10.

The second case is a rank sum of 11, created with the rankings $\{1, 2, 3, 5\}$.

1 Find the sets of ranks that give a rank sum of:

a 12 **b** 13 [three sets]

c 14 [five sets] **d** 15 [six sets]

2 How many possible sets of four ranks are there from ten?

3 Copy and complete the following table.

P(rank sum = 10)	$\frac{1}{210}$	0.004762
P(rank sum $\leqslant$ 11)	$\frac{1+1}{210}$	0.009524
P(rank sum $\leqslant$ 12)		
P(rank sum $\leqslant$ 13)		
P(rank sum $\leqslant$ 14)		
P(rank sum $\leqslant$ 15)		

The critical value, c, of a 5% one-tailed test is the greatest value of c that satisfies $P(X \leqslant c) \leqslant 0.05$.

From the work shown, we can see that this is $c = 13$.

And so we conclude that for $m = 4, n = 6$ the critical value for a one-tailed test at the 5% significance level is 13.

This is reflected in the critical value table.

	Level of significance					
One-tailed	0.05	0.025	0.01	0.05	0.025	0.01
Two-tailed	0.1	0.05	0.02	0.1	0.05	0.02
n	$m = 3$			$m = 4$		
3	6	–	–			
4	6	–	–	11	10	–
5	7	6	–	12	11	10
6	8	7	–	13	12	11

We can now also confirm that the critical values for tests at the 2.5% and 1% significance levels are correct.

EXPLORE 11.2

Use the internet to research the Kruskal–Wallis test for non-parametric data. We can use the Kruskal–Wallis test to compare three or more samples.

EXERCISE 11E

1 For each of the following datasets, calculate the values of:

i R_m

ii $m(n + m + 1) - R_m$

iii W

a

X	Y
5.7	4.8
4.4	3.9
4.9	4.1
4.6	4.2
5.2	4.7
4.5	4.3
	5.3

b

X	Y
17.3	19.8
19.7	18.1
19.4	19.2
18.3	18.6
18.7	18.4
18.9	18.0
19.1	18.5
	17.6
	17.8
	17.7

c

X	Y
21.5	22.0
21.6	22.9
21.8	22.6
22.4	22.5
23.0	22.8
	22.1
	23.1
	22.7

d

X	Y
114.2	115.8
116.1	115.1
115.4	116.0
116.8	117.1
116.2	117.2

2 For the corresponding dataset in question **1a–d**, state the critical value for the following significance levels.

a Significance level 5% one-tailed

b Significance level 10% two-tailed

c Significance level 5% two-tailed

d Significance level 1% one-tailed

3 For the following dataset sizes, find $\mathrm{E}(R_m)$ and $\mathrm{Var}(R_m)$.

a $n = 12, m = 9$ b $n = 13, m = 12$ c $n = 9, m = 8$ d $n = 15, m = 13$

4 a For the following dataset sizes and W values, find the test statistic if a normal approximation is used.
$n = 18, m = 12$; $W = 146$

b These hypotheses are given:
H_0: Population medians are equal.
H_1: Population medians are not equal.
Testing at 10%, state whether you would reject or not reject the null hypothesis.

M 5 It is believed that the calorie content of a chicken sausage is different to that of a vegetarian sausage. The calorie contents of eight vegetarian sausages and seven chicken sausages are measured and tabulated.

Vegetarian	186	181	176	135	184	190	111	132
Chicken	129	137	102	106	94	142	149	

Using a Wilcoxon rank-sum test, investigate this claim with a 5% significance level.

M 6 Plants of the same species grow on opposite sides of a river. It is thought that plants on the north side grow taller than those on the south side. Eight randomly selected plants are chosen from each side of the river, removed and their lengths measured to the nearest cm. By using a Wilcoxon rank-sum test, investigate this claim with a 5% significance level.

North side	23	21	20	28	27	19	24	25
South side	18	15	22	14	16	22	17	26

M 7 A supermarket manager wishes to investigate the quality of plums. Plums can be delivered from source to the supermarket either chilled (below 5 °C) or at a constant temperature of 12 °C. The amount of time (in days) it takes for the plums to become soft and therefore overripe is measured and ranked. The highest rank is the quickest to become overripe.

Stored at 12 °C	3	8	9	4	11	10
Chilled	2	7	1	5	6	

Carry out an appropriate non-parametric test, at the 5% level of significance, to investigate whether the chilled plums take longer to become overripe.

M 8 Mr Sum wishes to investigate whether a student's test score depends on whether the test is taken in the morning or in the afternoon. He selects a random sample of 35 students of similar ability, and randomly assigns some of them to take the test in the morning and the rest to take the test in the afternoon. The students taking the test in the morning are kept away from the students taking the test in the afternoon. The ordered scores are given in the following table (M for morning sitting, A for afternoon sitting).

M	31		M	64
M	32		A	65
M	38		A	66
M	39		M	67
A	41		A	68
A	43		A	69
M	44		M	70
A	47		A	72
M	48		A	73
A	49		A	75
M	51		A	76
A	56		A	78
M	57		M	81
A	58		A	82
M	59		A	85
M	60		A	86
M	62		A	88
A	63			

Using a suitable approximation, test, at a 2% level of significance, whether exam performance is affected by the session in which a student takes the examination.

WORKED EXAM-STYLE QUESTION

The following table shows the systolic blood pressure (mm Hg) of a random sample of eight students before and after a six-week training period.

Student	1	2	3	4	5	6	7	8
Before training	130	170	125	170	130	130	145	160
After training	120	163	120	135	143	136	144	120

a Stating clearly your hypotheses, test, using the Wilcoxon signed-rank test, whether or not there is evidence that the training has reduced blood pressure. Use a 5% level of significance.

At a later date a random sample of 30 students undertake a six-week training period. Analysis of their results using the Wilcoxon signed-rank test gives $T = 132$.

b Stating clearly your hypotheses and using a 5% level of significance, test whether or not there is evidence that the training has reduced blood pressure.

Answer

a H_0: Population median blood pressure is unchanged.

Define the hypotheses.

H_1: Population median blood pressure has decreased.

	1	2	3	4	5	6	7	8
Before–after	10	7	5	35	−13	−6	1	40
Rank	5	4	2	7	6	3	1	8
Signed rank	5	4	2	7	−6	−3	1	8

Calculate the sums of positive ranks, P, the sums of negative ranks, N, and the test statistic, T.

$P = 27, N = 9$

$T = \min(P, N) = 9$

Critical value is 5.

The test is one-tailed.

Since $9 > 5$, do not reject H_0.

Compare the test statistic with the critical value.

There is no evidence to suggest that the median blood pressure has decreased.

Conclude in context.

b H_0: Population median blood pressure is unchanged.

Define the hypotheses.

H_1: Population median blood pressure has decreased.

$$E(T) = \frac{n(n+1)}{4} = 232.5$$

$$\text{Var}(X) = \frac{n(n+1)(2n+1)}{24} = 2363.75$$

Calculate $E(T)$ and $\text{Var}(T)$. We can approximate to the normal distribution as n is large.

$$\text{Test statistic} = z = \frac{T - \mu + 0.5}{\sigma}$$

$$= \frac{132 - 232.5 + 0.5}{\sqrt{2363.75}}$$

$$= -2.06$$

Calculate the test statistic, remembering continuity corrections.

Critical value $= -1.645$

Find the critical value.

Since $-2.06 < -1.645$, reject H_0.

Compare the test statistic with the critical value.

The population median blood pressure has decreased.

Conclude in context.

Checklist of learning and understanding

Single-sample sign test:

- Given n data points, a sign test is created using $X \sim \text{Bin}(n, 0.5)$. The test statistic can be the number of + signs, that is the number of data points greater than the median. We can calculate the probability that X is above this test statistic, below this test statistic, or either in the case of a two-tailed test.

Wilcoxon signed-rank test:

- A Wilcoxon signed-rank test can be performed when:
  - the underlying data are symmetric
  - the underlying data are continuous.
- Where:
  - P is the sum of the ranks corresponding to the positive differences from the stated median
  - N is the sum of the ranks corresponding to the negative differences from the stated median
  - $T = \min(P, N)$ is the test statistic.
- Given the statistic $T = \min(P, N)$, then $\text{E}(T) = \dfrac{n(n+1)}{4}$, $\text{Var}(T) = \dfrac{n(n+1)(2n+1)}{24}$.
  For large n: $T \sim \text{N}\left(\dfrac{n(n+1)}{4}, \dfrac{n(n+1)(2n+1)}{24}\right)$ allowing for an approximate z-test to be done using $z = \dfrac{T - \mu + 0.5}{\sigma}$.

Wilcoxon matched-pairs signed-rank test:

- A Wilcoxon matched-pairs signed-rank test can be performed when:
  - the difference between matched-pairs is symmetric
  - the difference between matched-pairs is continuous.
- Where:
  - P is the sum of the ranks corresponding to the positive differences between the matched pairs
  - N is the sum of ranks corresponding to the negative differences between the matched pairs
  - $T = \min(P, N)$ is the test statistic.
- Given the statistic $T = \min(P, N)$, then $\text{E}(T) = \dfrac{n(n+1)}{4}$ and $\text{Var}(T) = \dfrac{n(n+1)(2n+1)}{24}$.
  For large n, $T \sim \text{N}\left(\dfrac{n(n+1)}{4}, \dfrac{n(n+1)(2n+1)}{24}\right)$, allowing for an approximate z-test with $z = \dfrac{T - \mu + 0.5}{\sigma}$.

Wilcoxon rank-sum test:

- A Wilcoxon rank-sum test can be performed when the two samples are independent, where:
  - the two samples have sizes m and n, where $m \leqslant n$
  - R_m is the sum of the ranks of the items in the sample of size m
  - the test statistic is $W = \min(R_m, m(n+m+1) - R_m)$.
- Given the test statistic W, then $\mathrm{E}(W) = \dfrac{m(n+m+1)}{2}$ and $\mathrm{Var}(W) = \dfrac{mn(n+m+1)}{12}$.

  For large n and m ($n \geqslant 10, m \geqslant 10$) it is possible to approximate W as a normal distribution:
  - $W \sim \mathrm{N}\left(\dfrac{m(n+m+1)}{2}, \dfrac{mn(n+m+1)}{12}\right)$, allowing for an approximate z-test with $z = \dfrac{W - \mu + 0.5}{\sigma}$.

END-OF-CHAPTER REVIEW EXERCISE 11

M 1 The dining room in a school is some distance away from the building that has all of the classrooms in it. The school believes that students take longer walking back from the dining room after lunch than they do walking there. The school records the time taken (in seconds) by ten randomly chosen students to walk to the dining room and ten students to walk back to the main school. The recorded times are presented in the following tables.

To the dining hall	62	58	69	84	45
	96	116	89	51	75

From the dining hall	67	85	68	100	49
	121	139	87	54	88

a i Using these data, state which non-parametric test would be most appropriate. Give a reason for your choice and any assumptions you need to make.

ii Carry out a test of the school's belief at the 5% significance level.

b Later on, the school discovers that the person who collected the data has used the same ten students and has recorded them in the same order.

i Which test is now most appropriate to use? Give a reason for your answer.

ii Using this new information, carry out a test of the school's belief at the 5% significance level.

M 2 The blood cholesterol levels of 30 males and 20 females are measured. These data are shown in the following table.

Males	621	550	104	303	384
	1080	1061	771	206	1203
	810	259	610	770	382
	829	385	479	1301	551
	92	723	105	478	417
	1081	383	205	207	258
Females	208	482	94	194	370
	973	683	72	50	162
	149	215	304	127	233
	189	529	191	974	710

Using a suitable approximation, test, at the 5% significance level, whether the blood cholesterol levels of females and males differ. Is the assumption that the dataset is symmetric justified?

M 3 An investigation is conducted into the pollution levels in a major city. The number of 2.5 mm particles can be measured using the Air Quality Index (AQI). The World Health Organization recommends that an AQI of 50 or below will not have a significant effect on health. The AQI is measured for 14 consecutive days. The data are shown in the following table.

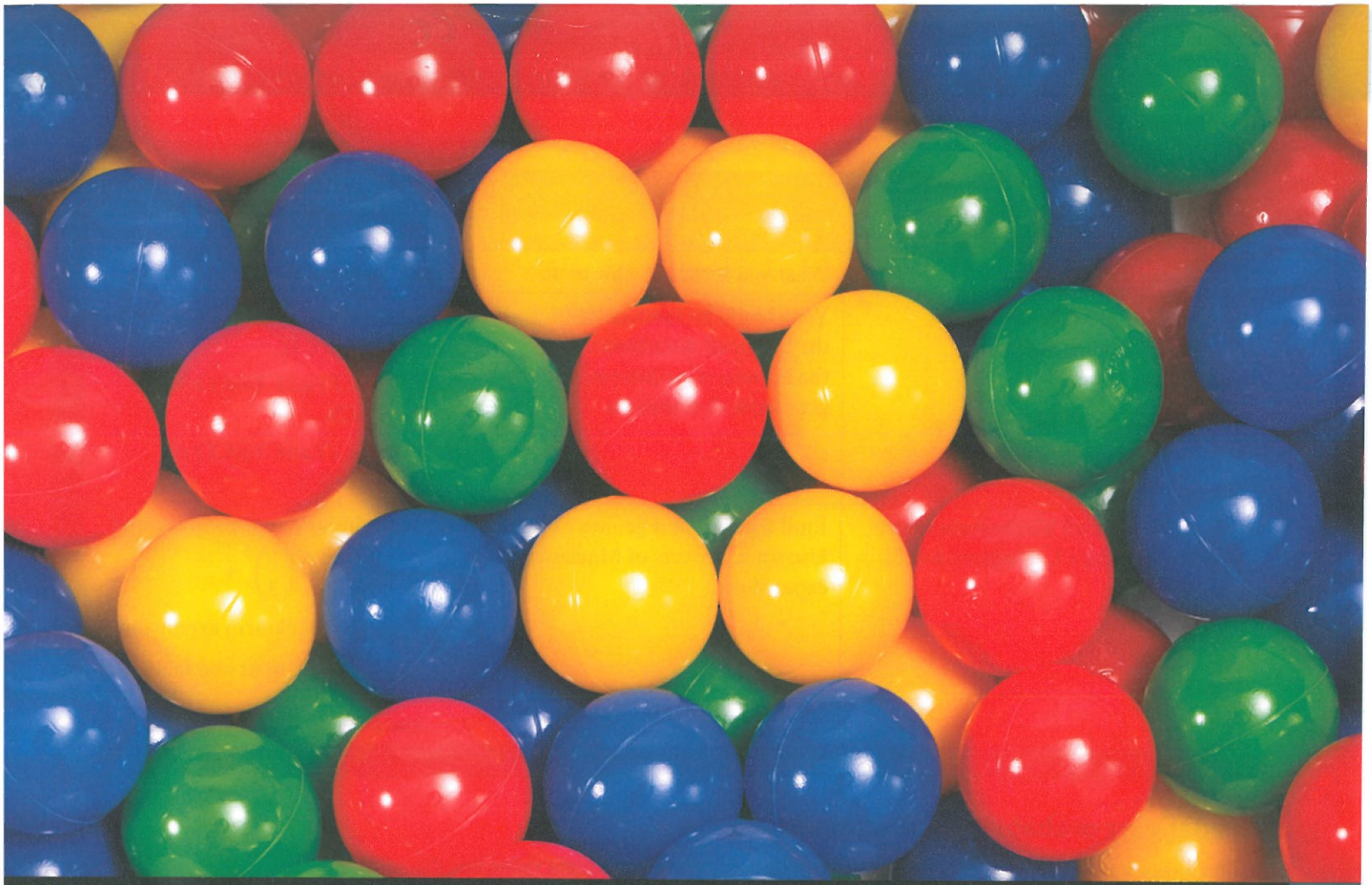
45	132	87	103	67	46	90
78	54	79	81	44	65	82

a Explain why the following tests cannot be carried out:

i a *t*-test

ii a single-sample Wilcoxon signed-rank test.

b Carry out an appropriate test, at the 5% significance level, to establish whether there is evidence that the AQI is above 50 in the city.



Chapter 12
Probability generating functions

In this chapter you will learn how to:

- understand the concept of a probability generating function (PGF)
- construct and use the PGF for given distributions, including:
  - discrete uniform distribution
  - binomial distribution
  - geometric distribution
  - Poisson distribution
- use formulae for the mean ($E(X)$) and variance ($Var(X)$) of a discrete random variable in terms of its PGF, and use these formulae to calculate the mean and variance of a given probability distribution
- use the result that the PGF of the sum of independent random variables is the product of the PGFs of those random variables (the convolution theorem)
- find the probability generating function of a linear transformation of random variables
- generalise to three or more random variables.

PREREQUISITE KNOWLEDGE

Where it comes from	What you should be able to do	Check your skills
AS & A Level Mathematics Probability & Statistics 1, Chapters 7 & 8 AS & A Level Mathematics Probability & Statistics 2, Chapters 2 & 4	You should be familiar with the binomial distribution, the Poisson distribution, and the geometric distribution.	1 Find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ of $X \sim \mathrm{Bin}(8, 0.4)$. 2 Find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ of $X \sim \mathrm{Po}(2)$. 3 Find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ of $X \sim \mathrm{Geo}(0.3)$.
AS & A Level Mathematics Pure Mathematics 1, Chapters 6 & 21	Find the sum of a geometric series. Use some aspects of Maclaurin expansions.	4 Find the sum to the nth term of $\frac{1}{3} + \left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^3 + \ldots$ 5 Find the Maclaurin expansion of $\frac{5}{(3-2t)^2}$ up to and including the t^3 term.

Redefining probability distributions

The discrete uniform distribution is a distribution where each discrete value has the same probability of occurring. For instance, when rolling a fair die, the probability of each outcome is $\frac{1}{6}$.

x	1	2	3	4	5	6
$\mathrm{P}(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Generally:

$$\mathrm{P}(X = x) = \begin{cases} \frac{1}{n} & x = x_1, x_2, \ldots, x_n \\ 0 & \text{otherwise} \end{cases}$$

In this chapter, we shall study a different way of describing a probability distribution. We focus on finding the probability generating functions (PGFs) of discrete probability distributions. This gives an elegant and efficient way of finding expected values and variances. A PGF gives a concise form for a probability distribution and allows much greater analysis. By recognising the expansions of functions, PGFs enable us to describe the structure of infinite discrete distributions, such as the Poisson distribution and geometric distribution. We can therefore use PGFs to find the probabilities when the value of the discrete random variable is very large indeed.

12.1 The probability generating function

Let X represent a discrete random variable, with values x_i:

x	x_1	x_2	x_3	...	x_n
$\mathrm{P}(X = x)$	$\mathrm{P}(X = x_1)$	$\mathrm{P}(X = x_2)$	$\mathrm{P}(X = x_3)$	...	$\mathrm{P}(X = x_n)$

We can create a function $G_X(t)$ using this table:

$$G_X(t) = P(X = x_1)t^{x_1} + P(X = x_2)t^{x_2} + P(X = x_3)t^{x_3} + \ldots + P(X = x_n)t^{x_n}$$

This function is called the **probability generating function** (PGF). It can be written as a single summation:

$$G_X(t) = \sum_x t^{x_i} P(X = x_i)$$

This is called the closed form of the probability generating function. Notice that the expression for $G_X(t)$ is the same as that for the expectation function, $E(t^X)$, and so $G_X(t) = \sum_x t^{x_i} P(X = x_i) = E(t^X)$, as shown in Key point 12.1.

KEY POINT 12.1

$$G_X(t) = \sum_x t^{x_i} P(X = x_i) = E(t^X)$$

The variable t is called a dummy variable in this case, and has no significance itself, but t does have an important role in finding the expectation of X and higher moments of expectation.

WORKED EXAMPLE 12.1

Consider the following probability distribution.

x	0	1	2	3	4	5	6
$P(X = x)$	0.1	0.2	0.3	0.15	0.1	0.1	0.05

Write down the PGF for the random variable X.

Answer

Apply the general form for the PGF.

$G_X(t) = \sum_x t^{x_i} P(X = x_i)$

$G_X(t) = 0.1t^0 + 0.2t^1 + 0.3t^2 + 0.15t^3 + 0.1t^4 + 0.1t^5 + 0.05t^6$

Sometimes it is useful to do this in table form first.

In Worked example 12.1, the values of the random variable occur in a sequence. This does not have to be the case, as Worked example 12.2 shows.

WORKED EXAMPLE 12.2

Consider the following probability distribution.

x	2	4	5	10
$P(X = x)$	0.1	0.2	0.3	0.4

Write down the PGF for the random variable X.

Answer

Apply the general form for the PGF.

$G_X(t) = \sum_x t^{x_i} P(X = x_i)$

$G_X(t) = 0.1t^2 + 0.2t^4 + 0.3t^5 + 0.4t^{10}$

Notice that there does not need to be any specific pattern in the values that the random variable can take.

At a trivial level, you can think of a PGF as a different way of presenting the information given by a probability distribution table. As you will discover throughout this chapter, PGFs allow us to calculate much more.

You may have noticed in Worked examples 12.1 and 12.2 that the probabilities are just the coefficients of each of the terms in t. The sum of these probabilities is 1.

We can see this by evaluating the probability generating function when $t = 1$:

$$G_X(1) = \sum_x 1^{x_i} P(X = x_i) = \sum_x P(X = x_i) = 1$$
$$G_X(1) = 1$$

Something that is a little harder to spot is that if we differentiate the probability generating function with respect to t, we will multiply each term by the value x_i:

$$G'_X(1) = \sum_x x_i(1)^{x_i - 1} P(X = x_i)$$

And then evaluating at $t = 1$ gives:

$$G'_X(1) = \sum_x x_i(1)^{x_i - 1} P(X = x_i) = E(X)$$

So $G'_X(1) = E(X)$, as shown in Key point 12.2.

KEY POINT 12.2

$$G'_X(1) = E(X)$$

You may notice that if all of the values of x are non-negative integer values, then $G_X(t)$ forms a polynomial in t. This may be finite or infinite, depending on the context.

REWIND

From your work on series in AS & A Level Pure Mathematics 1 Chapter 6, you may have observed that if there is a pattern in the PGF, it may be possible to express it as a function rather than a summation. This can lead to an efficient way of finding $E(X)$.

WORKED EXAMPLE 12.3

Let X be a discrete random variable, as shown in the probability distribution given by:

x	1	2	3	4	5
$P(X = x)$	0.2	0.2	0.2	0.2	0.2

Find the probability generating function for X.

Answer

The PGF is: Use the definition of $G_X(t)$.

$$G_X(t) = \sum_x t^{x_i} P(X = x_i)$$
$$G_X(t) = 0.2t^1 + 0.2t^2 + 0.2t^3 + 0.2t^4 + 0.2t^5$$
$$= 0.2t(1 + t + t^2 + t^3 + t^4)$$ Factorise.

FAST FORWARD

In the work on Maclaurin expansions in Chapter 21 Section 21.4, you should see that if there is a pattern in the PGF, it may be possible to express it as a function rather than a summation. This is an efficient way of finding $E(X)$ and higher moments of expectation.

$(1 + t + t^2 + t^3 + t^4) = S_5 = \dfrac{1 - t^5}{1 - t}$

$G_X(t) = \dfrac{0.2t(1 - t^5)}{1 - t}$

Notice that the distribution from the table is a uniform distribution.

Notice that this expression is the sum of the first five terms of a geometric series with first term 1 and common ratio t.

So we can use the formula $S_n = \dfrac{a(1 - r^n)}{1 - r}$ from AS & A Level Pure Mathematics 1.

Discrete uniform distribution

Let X be a discrete random variable with $P(X = x_i) = \begin{cases} \frac{1}{n} & i = 1, \ldots n \\ 0 & \text{otherwise.} \end{cases}$

Then $G_X(t) = \dfrac{t(1 - t^n)}{n(1 - t)}$, as shown in Key point 12.3.

KEY POINT 12.3

For a uniform distribution:

$$G_X(t) = \frac{t(1 - t^n)}{n(1 - t)}$$

WORKED EXAMPLE 12.4

Let $X \sim \text{Bin}(5, 0.2)$. Find the probability generating function for X.

Answer

As a reminder, the probability distribution would be:

x	$P(X = x)$
0	0.8^5
1	$5 \times 0.8^4 \times 0.2$
2	$10 \times 0.8^3 \times 0.2^2$
3	$10 \times 0.8^2 \times 0.2^3$
4	$5 \times 0.8^1 \times 0.2^4$
5	0.2^5

Use the binomial formula from AS & A Level Probability & Statistics 1:

$P(X = x) = \binom{n}{x} p^x q^{n-x}$

The PGF is:

$G_X(t) = \sum_x t^{x_i} P(X = x_i)$

$G_X(t) = 0.8^5 + 5 \times 0.8^4 \times 0.2t + 10 \times 0.8^3 \times 0.2^2 t^2 + 10 \times 0.8^2 \times 0.2^3 t^3 + 5 \times 0.8^1 \times 0.2^4 t^4 + 0.2^5 t^5$

$= (0.8 + 0.2t)^5$

This is a binomial expansion and can be factorised.

Binomial distribution

Let $X \sim \text{Bin}(n, p)$. Then $G_X(t) = (q + pt)^n$, as shown in Key point 12.4.

KEY POINT 12.4

For the binomial distribution:

$$G_X(t) = (q + pt)^n$$

WORKED EXAMPLE 12.5

Let $X \sim \text{Geo}\left(\frac{1}{5}\right)$. Find the probability generating function for X.

Answer

As a reminder, the probability distribution would be:

Using the geometric formula from AS & A Level Probability & Statistics 1:

$$P(X = x) = \left(\frac{4}{5}\right)^{x-1}\left(\frac{1}{5}\right)$$

x	$P(X = x)$
1	$\frac{1}{5}$
2	$\left(\frac{4}{5}\right)\left(\frac{1}{5}\right)$
3	$\left(\frac{4}{5}\right)^2\left(\frac{1}{5}\right)$
...	...

The PGF is:

$$G_X(t) = \sum_x t^{x_i} P(X = x_i)$$

$$G_X(t) = \frac{1}{5}t + \left(\frac{4}{5}\right)\left(\frac{1}{5}\right)t^2 + \left(\frac{4}{5}\right)^2\left(\frac{1}{5}\right)t^3 + \ldots$$

$$= \frac{t}{5}\left(1 + \left(\frac{4t}{5}\right) + \left(\frac{4t}{5}\right)^2 + \ldots\right)$$

This is the sum of a geometric series to infinity. So we can use the formula $S_\infty = \frac{a}{1-r}$ from AS & A Level Pure Mathematics 1.

$$= \frac{\frac{t}{5}}{\left(1 - \frac{4t}{5}\right)}$$

This gives us a generalised form, but we need to simplify it.

$$= \frac{t}{5 - 4t}$$

Geometric distribution

Let $X \sim \text{Geo}(p)$. Then $G_X(t) = \dfrac{pt}{1-qt}$, as shown in Key point 12.5.

KEY POINT 12.5

For a geometric distribution:

$$G_X(t) = \frac{pt}{1-qt}$$

Poisson distribution

Let $X \sim \text{Po}(\lambda)$. Then $G_X(t) = e^{\lambda(t-1)}$, as shown in Key point 12.6.

KEY POINT 12.6

For a Poisson distribution:

$$G_X(t) = e^{\lambda(t-1)}$$

PROOF 12.1

The following table shows the probability distribution table for $X \sim \text{Po}(\lambda)$, using $P(X = x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$.

x	$P(X = x)$
0	$e^{-\lambda}$
1	$\dfrac{\lambda e^{-\lambda}}{1!}$
2	$\dfrac{\lambda^2 e^{-\lambda}}{2!}$
3	$\dfrac{\lambda^3 e^{-\lambda}}{3!}$
...	...

And so the PGF is:

$$G_X(t) = e^{-\lambda} + \frac{\lambda e^{-\lambda}}{1!}t + \frac{\lambda^2 e^{-\lambda}}{2!}t^2 + \frac{\lambda^3 e^{-\lambda}}{3!}t^3 + \ldots$$

$$= e^{-\lambda}\left(1 + \lambda t + \frac{(\lambda t)^2}{2!} + \frac{(\lambda t)^3}{3!} + \ldots\right)$$

From Chapter 21, we have the following Maclaurin expansion.

$$e^x \approx 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$$

Therefore, the PGF becomes:

$$G_X(t) = e^{-\lambda}(e^{\lambda t})$$
$$= e^{\lambda(t-1)}$$

as required.

EXERCISE 12A

1 For each of the following distributions, write down the probability generating function, $G_X(t)$.

a $X \sim \text{Bin}(20, 0.3)$ b $X \sim \text{Bin}(10, 0.25)$ c $X \sim \text{Bin}(50, 0.04)$

2 For each of the following distributions, write down the probability generating function, $G_X(t)$.

a $X \sim \text{Po}(4)$ b $X \sim \text{Po}(2.3)$ c $X \sim \text{Po}(12)$

3 For each of the following distributions, write down the probability generating function, $G_X(t)$.

a $X \sim \text{Geo}(0.1)$ b $X \sim \text{Geo}(0.7)$ c $X \sim \text{Geo}(0.4)$

4 Find the probability generating function for:

x	1	2	3	4	5	6	7
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$

5 Find the probability generating function for:

x	2	4	6	8	10
$P(X = x)$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$

Write your answer in the form $at^b(1 + t^2)^c$.

6 Find the probability distribution for the following probability generating function.

$$G_X(t) = \frac{t^3}{6}(2 + t^4 + 2t^8)$$

7 Find the probability distribution for the following probability generating function.

$$G_X(t) = \left(\frac{4}{5} + \frac{t}{5}\right)^7$$

8 A distribution has a probability generating function of:

$$G_X(t) = \frac{1}{3}\left(\frac{2 + t}{2 - t}\right)$$

a Find the probabilities for when $x = 0, 1, 2, 3$.

b Find a general formula for $P(X = k), k \geqslant 1$.

P 9 Independent trials, each with the probability of 'success' p, are carried out. The random variable X counts the number of trials up to and including that on which the first success is obtained. Write down an expression for $P(X = x)$ for $x = 1, 2, \ldots$. Show that the probability generating function of X is $G_X(t) = pt(1 - qt)^{-1}$.

10 Consider the probability generating function $G_X(t) = \dfrac{k}{(5 - 2t)^2}$. Find the value of k.

12.2 Mean (E(X)) and variance (Var(X)) using the probability generating function

In this section, we shall find the mean and variance using the PGF of a discrete random variable. We shall discover why expressing the probabilities in a functional way using a polynomial is a very powerful tool.

Here are a few important results that we shall use:

$$\mathrm{E}(X) = \sum_{\forall x} x\,\mathrm{P}(X = x)$$

$$\mathrm{E}(X^2) = \sum_{\forall x} x^2\,\mathrm{P}(X = x)$$

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$$

These lead to the following results. First, as shown in Section 12.1, Key point 12.2:

$$\mathrm{E}(X) = \mathrm{G}_X'(1)$$

Differentiating $\mathrm{G}_X(t)$ twice:

$$\mathrm{G}_X(t) = \sum_x t^x \mathrm{P}(X = x)$$

$$\mathrm{G}_X'(t) = \sum_x x t^{x-1} \mathrm{P}(X = x)$$

$$\mathrm{G}_X''(t) = \sum_x x(x-1) t^{x-2} \mathrm{P}(X = x)$$

Evaluating at $t = 1$:

$$\mathrm{G}_X''(1) = \sum_x x(x-1)\mathrm{P}(X = x)$$

$$\mathrm{G}_X''(1) = \sum_x (x^2 - x)\mathrm{P}(X = x)$$

$$\mathrm{G}_X''(1) = \sum_x x^2 \mathrm{P}(X = x) - \sum_x x\,\mathrm{P}(X = x)$$

$$\mathrm{G}_X''(1) = \mathrm{E}(X^2) - \mathrm{E}(X)$$

$$\mathrm{E}(X^2) = \mathrm{G}_X''(1) + \mathrm{G}_X'(1)$$

Therefore:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$$

$$\mathrm{Var}(X) = \mathrm{G}_X''(1) + \mathrm{G}_X'(1) - [\mathrm{G}_X'(1)]^2$$

as shown in Key point 12.7.

KEY POINT 12.7

$$\mathrm{Var}(X) = \mathrm{G}_X''(1) + \mathrm{G}_X'(1) - [\mathrm{G}_X'(1)]^2$$

WORKED EXAMPLE 12.6

A bag contains five red balls and three green balls. The balls are taken out one at a time, the colour is noted, and then it is replaced. Let X be the number of times a ball is removed until a green ball is chosen.

a State the PGF of X.

b Calculate the mean and variance of X.

Answer

a In this question, we can see that the outcomes follow a geometric distribution:

$$X \sim \text{Geo}\left(\frac{3}{8}\right)$$

Always define a random variable before using it. Here, $p = \frac{3}{8}$ and $q = \frac{5}{8}$.

Its PGF is:

$$G_X(t) = \frac{pt}{1 - qt}$$

$$= \frac{\left(\frac{3}{8}\right)t}{1 - \left(\frac{5}{8}\right)t}$$

$$= \frac{3t}{8 - 5t}$$

Simplify.

We will need to differentiate this twice in part **b**.

b $G'_X(t) = \frac{3(8 - 5t) - 3t(-5)}{(8 - 5t)^2}$

The quotient rule is needed here.

$$= \frac{24}{(8 - 5t)^2}$$

$$= 24(8 - 5t)^{-2}$$

$G''_X(t) = 24 \times -5 \times -2(8 - 5t)^{-3}$

Find the second derivative.

Use the chain rule.

$$= \frac{240}{(8 - 5t)^3}$$

$G'_X(1) = \frac{24}{9} = 2\frac{2}{3}$

Evaluate at $t = 1$.

$G''_X(1) = \frac{240}{27} = 8\frac{8}{9}$

$E(X) = G'_X(1) = 2\frac{2}{3}$

Use the standard results.

$\text{Var}(X) = G''_X(1) + G'_X(1) - [G'_X(1)]^2$

$$= \frac{80}{9} + \frac{8}{3} - \frac{64}{9}$$

$$= \frac{40}{9} = 4\frac{4}{9}$$

Generally, the previous results offer only a different way of calculating the mean and variance. Worked examples 12.7 and 12.8 demonstrate some properties of distributions using the probability generating functions.

WORKED EXAMPLE 12.7

Prove that for $X \sim \text{Po}(\lambda)$:

a $\text{E}(X) = \lambda$ **b** $\text{Var}(X) = \lambda$

Answer

a $G_X(t) = e^{\lambda(t-1)}$ — Consider the PGF of the Poisson distribution.

$G'_X(t) = \lambda e^{\lambda(t-1)}$ — Differentiate.

$G'_X(1) = \lambda e^{\lambda(1-1)}$ — Let $t = 1$.

$G'_X(1) = \lambda$

$\text{E}(X) = \lambda$ — As required.

b $G'_X(t) = \lambda e^{\lambda(t-1)}$ — Consider $G'_X(t)$.

$G''_X(t) = \lambda^2 e^{\lambda(t-1)}$ — Differentiate again.

$G''_X(1) = \lambda^2$ — Let $t = 1$.

$\text{Var}(X) = G''_X(1) + G'_X(1) - [G'_X(1)]^2$ — Use the given identity.

$\text{Var}(X) = \lambda^2 + \lambda - \lambda^2$

$\text{Var}(X) = \lambda$ — As required.

We shall use the properties of the mean and variance to calculate unknown probabilities using the PGF, as in Worked example 12.8.

WORKED EXAMPLE 12.8

A discrete random variable has the following probability distribution.

x	0	1	2
$\text{P}(X = x)$	a	b	c

The mean is $\frac{2}{3}$ and the variance is $\frac{5}{9}$. Find a, b and c.

Answer

First, express this using the PGF: — $G_X(t) = \sum_x t^{x_i} \text{P}(X = x_i)$

$G_X(t) = a + bt + ct^2$

$G'_X(t) = b + 2ct$

$G''_X(t) = 2c$ — Differentiate twice to find the expectation and variance.

$G_X(1) = a + b + c$

$G'_X(1) = b + 2c$ — Evaluate at $t = 1$.

$G''_X(1) = 2c$

$1 = a + b + c$ **(1)**

$\frac{2}{3} = b + 2c$ **(2)**

$\frac{5}{9} = 2c + \frac{2}{3} - \frac{4}{9}$

$\frac{1}{3} = 2c$ **(3)**

From **(3)**: $c = \frac{1}{6}$

From **(2)**: $b = \frac{1}{3}$

From **(1)**: $a = \frac{1}{2}$

$G_X(1) = 1$

$E(X) = G'_X(1)$

$\text{Var}(X) = G''_X(1) + G'_X(1) - [G'_X(1)]^2$

We can use the derivatives of $G_X(t)$ to help us find $P(X = r)$ of probability generating functions. We use the idea that the expansion of a function of this type coincides with its Maclaurin expansion.

Consider $G_X(t) = \frac{t}{2 - t}$.

This can be expanded binomially to give $\frac{t}{2} + \frac{t^2}{4} + \frac{t^3}{8} + \frac{t^4}{16} + \ldots$.

The probabilities for each r can be seen here clearly, but what if the expansion is not a known one?

We can consider the Maclaurin expansion of any function as:

$$G_X(t) = G_X(0) + G'_X(0)t + \frac{G''_X(0)t^2}{2!} + \frac{G'''_X(0)t^3}{3!} + \ldots$$

And so the probabilities could also be calculated using:

$$P(X = r) = \frac{G^{(r)}_X(0)}{r!}$$

where $G^{(r)}_X(t)$ is the rth derivative of $G_X(t)$.

EXERCISE 12B

1 For the following probability generating function:

$$G_X(t) = \frac{t^2}{10}(1 + 2t + 3t^2 + 2t^4 + t^5)$$

a Find $G'_X(1)$. **b** Find $G''_X(1)$. **c** Find $E(X)$. **d** Find $\text{Var}(X)$.

2 For the following probability generating function:

$$G_X(t) = \frac{t}{16}(1 + 4t + 6t^2 + 4t^3 + t^4)$$

a Find $G'_X(1)$. **b** Find $G''_X(1)$. **c** Find $E(X)$. **d** Find $\text{Var}(X)$.

3 Find $E(X)$ and $\text{Var}(X)$ of the following probability generating function.

$$G_X(t) = \frac{t^4}{10}(3 + 5t + t^3 + t^5)$$

4 Find $E(X)$ and $Var(X)$ of the following probability generating function.

$$G_X(t) = \frac{9}{(5-2t)^2}$$

5 In a game, the probability that player A wins on her rth go can be described as a discrete random variable X, with probability function $P(X = r) = \frac{1}{2^r}$, for $r = 1, 2, 3, \ldots$.

a Find the probability generating function.

b Find $E(X)$ and $Var(X)$.

6 Find $E(X)$ and $Var(X)$ of the following probability generating function,

$$G_X(t) = \frac{p}{1-qt}$$

where $q = 1 - p$.

7 Find $E(X)$ and $Var(X)$ of the following probability generating function.

$$G_X(t) = \frac{t+2}{(2-t^2)(4-t)}$$

M

8 A regular octagonal spinner is made up of eight isosceles triangles. Two of the triangles have a score of 2, three triangles have a score of 'a' and three triangles have a score of 'b', where $a < b$. The expectation and variance of the spinner are given by:

$$E(X) = 4.25$$

$$Var(X) = \frac{75}{16}$$

a Find an expression for $G_X(t)$, $G'_X(t)$ and $G''_X(t)$.

b Hence, find the values of a and b.

E

All the examples covered so far have involved discrete random variables where each value of the random variable has a probability associated with it. This means we can set up the probability generating function easily and calculate the mean and variance from this.

This is not possible to do if the random variable is continuous. To calculate the mean and variance for a continuous random variable we use the **moment generating function**. This technique also works for a discrete random variable.

$M_X(\theta) = \sum_x e^{\theta x} P(X = x)$	For a discrete random variable, X.
$M_X(\theta) = \int e^{\theta x} f(x)\,dx$	For a continuous random variable, X, with a probability density function $f(x)$.

The following results come from the moment generating function.

$$E(X) = M'(0)$$
$$E(X^r) = M^r(0)$$
$$Var(X) = M''(0) - [M'(0)]^2$$

This is beyond the Further Mathematics course.

12.3 The sum of independent random variables

In AS & A Level Probability & Statistics 2, Chapter 3, you learned that, if there are two independent variables of the same distribution (for example the normal), $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$, we can create the distribution of $X + Y$. This will be $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. The same is true for the Poisson distribution. We need to have some method for finding the PGF of $X + Y$ for any pairing of independent random variables X and Y. We will consider discrete random variables in this case.

Consider the following two distributions as an example. In Section 12.4, we will extend this idea to three or more random variables.

Let X have the following probability distribution.

x	0	1	2
$P(X = x)$	p_0	p_1	p_2

Let Y have the following probability distribution.

y	0	1	2
$P(Y = y)$	q_0	q_1	q_2

Then the distribution of $X + Y$ is:

$x + y$	$P(X + Y = x + y)$
0	$P(X = 0 \cap Y = 0)$
1	$P(X = 0 \cap Y = 1) + P(X = 1 \cap Y = 0)$
2	$P(X = 0 \cap Y = 2) + P(X = 1 \cap Y = 1) + P(X = 2 \cap Y = 0)$
3	$P(X = 1 \cap Y = 2) + P(X = 2 \cap Y = 1)$
4	$P(X = 2 \cap Y = 2)$

If X and Y are independent, then $P(X = x_i \cap Y = y_j) = P(X = x_i) \times P(Y = y_j) = p_i q_j$.

The previous table now becomes:

$x + y$	$P(X + Y = x + y)$
0	$p_0 q_0$
1	$p_1 q_0 + p_0 q_1$
2	$p_2 q_0 + p_1 q_1 + p_0 q_2$
3	$p_2 q_1 + p_1 q_2$
4	$p_2 q_2$

Now consider the PGF of $X + Y$:

$$G_{X+Y}(t) = p_0 q_0 + (p_1 q_0 + p_0 q_1)t + (p_2 q_0 + p_1 q_1 + p_0 q_2)t^2 + (p_2 q_1 + p_1 q_2)t^3 + p_2 q_2 t^4$$

This can be rewritten as:

$$G_{X+Y}(t) = (p_0 + p_1 t + p_2 t^2)(q_0 + q_1 t + q_2 t^2)$$

And we notice that these are the PGFs of X and of Y:

$$G_{X+Y}(t) = G_X(t) \times G_Y(t)$$

This is called the **convolution theorem**.

The convolution theorem

Let X and Y be two independent discrete random variables with PGFs $G_X(t)$ and $G_Y(t)$. The probability generating function of $X + Y$ is given as $G_{X+Y}(t) = G_X(t) \times G_Y(t)$, as shown in Key point 12.8.

KEY POINT 12.8

$$G_{X+Y}(t) = G_X(t) \times G_Y(t)$$

We will use this result directly to find the probability generating function of the sum of two independent random variables.

WORKED EXAMPLE 12.9

The discrete random variables X and Y have the following probability distributions.

x	1	2	3
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$

y	2	4	6
$P(Y = y)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Assuming that X and Y are independent:

a find the PGF of $X + Y$

b write down the probability distribution of $X + Y$

c show that $E(X + Y) = E(X) + E(Y)$ and $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.

Answer

a $G_X(t) = \frac{1}{4}t + \frac{1}{4}t^2 + \frac{1}{2}t^3$

and $G_Y(t) = \frac{1}{3}t^2 + \frac{1}{3}t^4 + \frac{1}{3}t^6$

First, consider the PGFs of X and Y.

$G_{X+Y}(t) = G_X(t) \times G_Y(t)$

Since X and Y are assumed to be independent, we can use the convolution theorem.

$$G_{X+Y}(t) = \left(\frac{1}{4}t + \frac{1}{4}t^2 + \frac{1}{2}t^3\right) \times \left(\frac{1}{3}t^2 + \frac{1}{3}t^4 + \frac{1}{3}t^6\right)$$

$$= \frac{t^3}{12}(1 + t + 2t^2) \times (1 + t^2 + t^4)$$

Take out $\frac{1}{4}t \times \frac{1}{3}t^2$ as a common factor.

$$= \frac{t^3}{12}(1 + t^2 + t^4 + t + t^3 + t^5 + 2t^2 + 2t^4 + 2t^6)$$

Multiply.

$$= \frac{t^3}{12}(1 + t + 3t^2 + t^3 + 3t^4 + t^5 + 2t^6)$$

Simplify.

b

$x+y$	$P(X+Y=x+y)$
3	$\frac{1}{12}$
4	$\frac{1}{12}$
5	$\frac{1}{4}$
6	$\frac{1}{12}$
7	$\frac{1}{4}$
8	$\frac{1}{12}$
9	$\frac{1}{6}$

c $G_{X+Y}(t) = \frac{1}{12}(t^3 + t^4 + 3t^5 + t^6 + 3t^7 + t^8 + 2t^9)$

Remember the powers of t relate to the values of the distribution, and the coefficients relate to their respective probabilities.

$G'_{X+Y}(t) = \frac{1}{12}(3t^2 + 4t^3 + 15t^4 + 6t^5 + 21t^6 + 8t^7 + 18t^8)$

Find both derivatives first.

$G''_{X+Y}(t) = \frac{1}{12}(6t + 12t^2 + 60t^3 + 30t^4 + 126t^5 + 56t^6 + 144t^7)$

$G'_{X+Y}(1) = \frac{1}{12}(3 + 4 + 15 + 6 + 21 + 8 + 18)$

Evaluate the derivatives at $t = 1$.

$= \frac{75}{12} = 6\frac{1}{4}$

$G''_{X+Y}(1) = \frac{1}{12}(6 + 12 + 60 + 30 + 126 + 56 + 144)$

$= \frac{434}{12} = 36\frac{1}{6}$

$E(X+Y) = G'_{X+Y}(1) = 6\frac{1}{4}$

$E(X+Y) = G'_X(1)$

$\text{Var}(X+Y) = \frac{434}{12} + \frac{75}{12} - \left[\frac{75}{12}\right]^2 = \frac{161}{48} = 3\frac{17}{48}$

$\text{Var}(X) = G''_X(1) + G'_X(1) - [G'_X(1)]^2$

Similarly:

$G_X(t) = \frac{1}{4}t + \frac{1}{4}t^2 + \frac{1}{2}t^3$

$G'_X(t) = \frac{1}{4} + \frac{2}{4}t + \frac{3}{2}t^2$ $\quad G'_X(1) = \frac{9}{4}$

$G''_X(t) = \frac{1}{2} + 3t$ $\quad G''_X(1) = \frac{7}{2}$

$E(X) = \frac{9}{4}$

$\text{Var}(X) = \frac{7}{2} + \frac{9}{4} - \left(\frac{9}{4}\right)^2 = \frac{11}{16}$

And:

$$G_Y(t) = \frac{1}{3}t^2 + \frac{1}{3}t^4 + \frac{1}{3}t^6$$

$$G'_Y(t) = \frac{2}{3}t + \frac{4}{3}t^3 + 2t^5 \qquad G'_Y(1) = 4$$

$$G''_Y(t) = \frac{2}{3} + 4t^2 + 10t^4 \qquad G''_Y(1) = \frac{44}{3}$$

$$E(Y) = 4$$

$$\text{Var}(Y) = \frac{44}{3} + 4 - 4^2 = \frac{8}{3}$$

$$E(X) + E(Y) = \frac{9}{4} + 4 = 6\frac{1}{4} = E(X + Y)$$

$$\text{Var}(X) + \text{Var}(Y) = \frac{11}{16} + \frac{8}{3} = \frac{161}{48} = 3\frac{17}{48} = \text{Var}(X + Y)$$

Therefore, $E(X + Y) = E(X) + E(Y)$ and $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.

Consider $E(Y)$ and $\text{Var}(Y)$ in the same way as for $E(X)$ and $\text{Var}(X)$.

Compare the two sets of results.

The PGF of a function of a random variable

Consider a discrete random variable that is a function of another variable, as seen in Chapter 8.

$Y = aX + b$, where X has PGF $G_X(t)$.

We can find the PGF of Y by considering the alternative definition of the PGF:

$$G_X(t) = E(t^X)$$

Consider $Y = aX + b$, then:

$$\begin{aligned} G_Y(t) &= E(t^Y) = E(t^{aX+b}) \\ &= E(t^{aX}t^b) = t^b E(t^{aX}) \\ &= t^b E[(t^a)^X] \\ &= t^b G_X(t^a) \end{aligned}$$

Therefore, $G_{aX+b}(t) = t^b G_X(t^a)$, as shown in Key point 12.9.

KEY POINT 12.9

$$G_{aX+b}(t) = t^b G_X(t^a)$$

From this definition, we can now formally reproduce some of the results shown in Chapter 11.

Let us find the expectation and variance of $Y = aX + b$ using the PGF of X.

$G_{aX+b}(t) = t^b G_X(t^a)$	
$G'_{aX+b}(t) = bt^{b-1}G_X(t^a) + at^{a-1}t^b G'_X(t^a)$	Differentiate using the product rule and the chain rule.
$G'_{aX+b}(1) = b1^{b-1}G_X(1^a) + a1^{a-1}1^b G'_X(1^a)$ $G'_{aX+b}(1) = b \times 1 + aG'_X(1)$ $E(aX+b) = aE(X) + b$	 $G_X(1) = 1$ $G'_X(1) = E(X)$
$G''_{aX+b}(t) = b(b-1)t^{b-2}G_X(t^a) + abt^{a-1}t^{b-1}G'_X(t^a)$ $+ a(a+b-1)t^{a+b-2}G'_X(t^a)$ $+ a^2t^{a-1}t^{a-1}t^b G''_X(t^a)$ $G''_{aX+b}(1) = b(b-1) + abG'_X(1) + a(a+b-1)G'_X(1) + a^2G''_X(1)$	Differentiate again. Let $t = 1$.
$\text{Var}(aX+b) = G''_{aX+b}(1) + G'_{aX+b}(1) - [G'_{aX+b}(1)]^2$ $\text{Var}(aX+b) = b(b-1) + abE(X) + a(a+b-1)E(X)$ $+ a^2G''_X(1) + [aE(X)+b] - [aE(X)+b]^2$ $= b^2 - b + abE(X) + a^2E(X) + abE(X) - aE(X) + a^2G''_X(1)$ $+ aE(X) + b - a^2[E(X)]^2 - 2abE(X) - b^2$ $= a^2[E(X) + G''_X(1) - [E(X)]^2]$ $\text{Var}(aX+b) = a^2\text{Var}(X)$	$\text{Var}(X) = G''_X(1) + E(X) - [E(X)]^2$

WORKED EXAMPLE 12.10

A discrete random variable, X, has the probability distribution:

x	1	2	3	4	5
$P(X=x)$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{3}{9}$	$\frac{2}{9}$	$\frac{1}{9}$

a Find $G_X(t)$, the PGF of X.

b Given that $Y = 4 - 7X$, find $G_Y(t)$, the PGF of Y.

Answer

a $G_X(t) = \frac{1}{9}t + \frac{2}{9}t^2 + \frac{3}{9}t^3 + \frac{2}{9}t^4 + \frac{1}{9}t^5$ — Use the definition for $G_X(t)$.

$= \frac{t}{9}(1 + 2t + 3t^2 + 2t^3 + t^4)$ — Factorise.

b $G_Y(t) = G_{-7X+4}(t)$ — Use the definition $G_{aX+b}(t) = t^b G_X(t^a)$.

$G_{-7X+4}(t) = t^4 G_X(t^{-7})$

$G_{-7X+4}(t) = t^4 \times \frac{t^{-7}}{9}(1 + 2(t^{-7}) + 3(t^{-7})^2 + 2(t^{-7})^3 + (t^{-7})^4)$

$G_Y(t) = \frac{1}{9t^3}\left(1 + \frac{2}{t^7} + \frac{3}{t^{14}} + \frac{2}{t^{21}} + \frac{1}{t^{28}}\right)$

EXERCISE 12C

1 A discrete random variable X has probability distribution:

x	1	2
$P(X = x)$	0.4	0.6

A discrete random variable Y has probability distribution:

y	0	1	3	5
$P(Y = y)$	0.25	0.25	0.25	0.25

Given that X and Y are independent, find:

a $G_X(t)$ **b** $G_Y(t)$ **c** $G_{X+Y}(t)$

2 A discrete random variable X has probability distribution:

x	−2	−1	0	1	2
$P(X = x)$	0.1	0.2	0.4	0.2	0.1

A discrete random variable Y has probability distribution:

y	2	3	4	5	6
$P(Y = y)$	0.2	0.2	0.2	0.2	0.2

Given that X and Y are independent, find:

a $G_X(t)$ **b** $G_Y(t)$ **c** $G_{X+Y}(t)$

3 A discrete random variable, X, has the probability distribution:

x	1	3	5
$P(X = x)$	0.2	0.5	0.3

A discrete random variable, Y, has the probability distribution:

y	0	2	4
$P(Y = y)$	0.3	0.4	0.3

a Find $G_X(t)$. **b** Find $G_Y(t)$.

Hence, given that X and Y are independent:

c find $G_{X+Y}(t)$ **d** write down the probability distribution of $X + Y$.

4 Let $X \sim \text{Geo}(0.4)$ and $Y \sim \text{Geo}(0.6)$, where X and Y are independent.

a Write down an expression for $G_{X+Y}(t)$. **b** Express $\dfrac{5}{(5-2t)(5-3t)}$ in partial fractions.

c Hence, find $P(X + Y = k)$ for $k = 2, 3, 4$.

5 Let $X \sim \text{Bin}(3, 0.2)$ and $Y \sim \text{Po}(2)$, where X and Y are independent.

a Find the probability generating function for $X + Y$.

b Find the value of $E(X + Y)$.

c Find the value of $G''_{X+Y}(t)$. Write your answer in the form:

$$\frac{2e^{2t}}{125e^2}(t+4)(at^2+bt+c)$$

d Find the value of $\text{Var}(X + Y)$.

P 6 Let $X \sim \text{Po}(a)$ and $Y \sim \text{Po}(b)$. Prove that if X and Y are independent:

a $\text{E}(X + Y) = a + b$

b $\text{Var}(X + Y) = a + b$

7 Let $\text{G}_X(t) = \dfrac{3t}{5 - 2t}$.

a For $Y = 5X$, find $\text{G}_Y(t)$.

b For $Y = X + 7$, find $\text{G}_Y(t)$.

8 Let $\text{G}_X(t) = \text{e}^{4(t-1)}$.

a For $Y = X - 1$, find $\text{G}_Y(t)$.

b For $Y = 3 - 2X$, find $\text{G}_Y(t)$.

P 9 A discrete random variable, X, has the probability generating function:

$$\text{G}_X(t) = \frac{3}{10 - 7t}$$

a Find the probability generating function for $Y = 3X + 4$.

b Show that $\text{E}(Y) = 11$.

c Find $\text{P}(Y = k)$ for $k = 4, 7, 10$.

12.4 Three or more random variables

In the previous sections, we considered only two random variables and only a linear combination of one random variable. If the variables are independent, we can generalise this to deal with more than two random variables, and also linear combinations of these variables, as shown in Key points 12.10 and 12.11.

KEY POINT 12.10

For independent random variables, X_i, with corresponding PGF $\text{G}_{X_i}(t)$, then:

$$\text{G}_{X_1 + \ldots + X_n}(t) = \text{G}_{X_1}(t) \times \ldots \times \text{G}_{X_n}(t)$$

If n independent discrete random variables all have the same PGF, then this formula reduces to:

$$\text{G}_{X_1 + \ldots + X_n}(t) = [\text{G}_X(t)]^n$$

This is different from:

$$\text{G}_{nX}(t) = \text{G}_X(t^n)$$

KEY POINT 12.11

For independent random variables, X_i, with corresponding PGF $\text{G}_{X_i}(t)$, then:

$$\text{G}_{aX_1 + bX_2} = \text{G}_{X_1}(t^a) \times \text{G}_{X_2}(t^b)$$

WORKED EXAMPLE 12.11

Find the probability generating function for the total number of 8s when a fair eight-sided die is rolled six times.

Answer

When one die is rolled, X has the following probability distribution, where X = number of 8s:

Find the PGF.

x	0	1
$P(X = x)$	$\frac{7}{8}$	$\frac{1}{8}$

And so its PGF is $G_X(t) = \frac{7}{8} + \frac{1}{8}t$.

This is called a Bernoulli distribution.

Since we are rolling the die six times, we define $Y = X_1 + X_2 + \ldots + X_6$.

And so the PGF of Y is $G_Y(t) = [G_X(t)]^6$.

These are all independent, so we can use the property $G_{aX_1 + bX_2} = G_{X_1}(t^a) \times G_{X_2}(t^b)$.

$$G_Y(t) = \left(\frac{7}{8} + \frac{1}{8}t\right)^6$$

Notice that this is the same as the PGF of $Y \sim \text{Bin}\left(6, \frac{1}{8}\right)$.

EXERCISE 12D

1 Consider the discrete random variable X, with probability generating function:

$$G_X(t) = \frac{t^2}{5}(4 + t)$$

X_1 and X_2 are two independent observations. Given that $Y = X_1 + X_2$, find:

a $G_Y(t)$ **b** the probability distribution of Y.

2 Consider the discrete random variable X, with probability distribution:

x	−1	0	1
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Three independent observations of X are taken and the values are added to form Y. Three independent observations of X are taken and the values of each X_i added to form Y where $Y = X_1 + X_2 + X_3$.

a Find $G_X(t)$. **b** Find $G_y(t)$. **c** Find $P(Y = 1)$.

3 Two independent random variables X and Y have corresponding probability generating functions:

$$G_X(t) = e^{3(t-1)}$$

and

$$G_Y(t) = e^{5(t-1)}$$

It is given that $A = Y - X$.

a A student made the following attempt to find $G_A(t)$:

Step 1: $G_{X+Y}(t) = G_X(t) \times G_Y(t)$

Step 2: Therefore $G_{Y-X}(t) = G_Y(t) \div G_X(t)$

Step 3: $G_{Y-X}(t) = e^{5(t-1)} \div e^{3(t-1)}$

Step 4: $G_{Y-X}(t) = e^{2(t-1)}$

Which step contains the error in the student's working?

b Find the correct probability generating function of A, $G_A(t)$.

4 Consider the discrete random variable X, with probability distribution:

X	1	2	3
$P(X = x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

and Y, with probability distribution:

y	4	5	6
$P(Y = y)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

It is given that $A = 2Y - 3X$.

a Find $G_X(t)$. **b** Find $G_Y(t)$. **c** Find $G_A(t)$.

5 Consider the independent discrete random variables X, Y and Z, where:

$G_X(t) = (0.7 + 0.3t)^3$

$G_Y(t) = (0.3 + 0.7t)^4$

$G_Z(t) = \frac{3t}{10 - 7t}$

Find the probability generating function for the following.

a $A = 2X + Y + 1$ **b** $B = X - Y - 3Z + 4$ **c** $C = 3X + 2Y + Z - 1$

6 An observation from $X \sim \text{Geo}(0.3)$ is taken five times: $X_1, \ldots X_5$.

a Find the probability generating function of $Y = (X_1 + \ldots + X_5)$.

b Find $P(Y \leqslant 8)$ to 4 significant figures.

7 Let X be a discrete random variable with probability distribution:

x	1	2	3
$P(X = x)$	0.2	0.3	0.5

Four independent observations are made. Y is the sum of these observations.

a Find the probability density function of Y.

b Find $E(Y)$. **c** Find $\text{Var}(Y)$.

8 Let $X_1 \sim \text{Geo}(0.1)$, $X_2 \sim \text{Geo}(0.2)$ and $X_3 \sim \text{Geo}(0.3)$.

Let $Y = 3X_1 + 2X_2 + X_3$.

a Find $G_Y(t)$. **b** Find $P(Y = 7)$.

EXPLORE 12.1

Consider two die, A and B. Each face on each die has an equal probability of occurring.

Die A has faces 1, 3, 4, 5, 6, 8.

Die B has faces 1, 2, 2, 3, 3, 4.

1 Find the probability generating function of each die and, hence, the probability generating function for Z, the sum of the scores on both die.

2 What do you notice about the probability distribution for the sum of scores?

WORKED EXAM-STYLE QUESTION

Jamil and Yao are revising for a Maths exam by randomly selecting questions from a large question bank. They play a game by taking it in turns to answer questions. The first person to get a question right wins the game. Jamil answers a question correctly $\frac{1}{4}$ of the time. Yao answers a question correctly $\frac{1}{5}$ of the time. Assume that all question attempts are independent. Let X be the total number of questions attempted. Jamil will start the game.

a Find the probability generating function for X.

b Find $E(X)$.

c Find $Var(X)$.

Answer

a $X = 1$ (Jamil starts the game and wins)

$P(X = 1) = \frac{1}{4}$

$X = 2$ (Jamil is incorrect, then Yao is correct).

$P(X = 2) = \frac{3}{4} \times \frac{1}{5} = \frac{3}{20}$

$X = 3$ (Jamil incorrect, Yao incorrect, Jamil correct).

$P(X = 3) = \left(\frac{3}{4} \times \frac{4}{5}\right) \times \frac{1}{4} = \left(\frac{3}{5}\right) \times \frac{1}{4}$

$P(X = 4) = \frac{3}{4} \times \frac{4}{5} \times \frac{3}{4} \times \frac{1}{5} = \left(\frac{3}{5}\right) \times \frac{3}{20}$

$P(X = 5) = \frac{3}{4} \times \frac{4}{5} \times \frac{3}{4} \times \frac{4}{5} \times \frac{1}{4} = \left(\frac{3}{5}\right)^2 \times \frac{1}{4}$

$P(X = 6) = \frac{3}{4} \times \frac{4}{5} \times \frac{3}{4} \times \frac{4}{5} \times \frac{3}{4} \times \frac{1}{5} = \left(\frac{3}{5}\right)^2 \times \frac{3}{20}$

And so on.

It is worth finding a few probabilities to spot any patterns in them. There may be a standard PGF that we can apply.

$$G_X(t) = \frac{1}{4}t + \left(\frac{3}{5}\right)\left(\frac{1}{4}\right)t^3 + \left(\frac{3}{5}\right)^2\left(\frac{1}{4}\right)t^5 + \ldots$$
$$+ \left(\frac{3}{20}\right)t^2 + \left(\frac{3}{5}\right)\left(\frac{3}{20}\right)t^4 + \left(\frac{3}{5}\right)^2\left(\frac{3}{20}\right)t^6 + \ldots$$
$$= \left(\frac{t}{4}\right)\left(\frac{1}{1-\frac{3}{5}t^2}\right) + \left(\frac{3t^2}{20}\right)\left(\frac{1}{1-\frac{3}{5}t^2}\right)$$
$$= \frac{5t + 3t^2}{4(5 - 3t^2)}$$

We can look for a pattern in the odd and even values for X. When X is odd, Jamil wins. When X is even, Yao wins.

Each of these patterns is a sum to infinity of geometric series, so we can use the formula $S_\infty = \frac{a}{1-r}$ from AS & A Level Pure Mathematics 1.

Write in closed form and simplify.

b $G_X(t) = \frac{5t + 3t^2}{4(5 - 3t^2)}$

$$G_X'(t) = \frac{4(5 - 3t^2)(5 + 6t) - (5t + 3t^2)(4)(-6t)}{4^2(5 - 3t^2)^2}$$
$$= \frac{100 + 120t - 60t^2 - 72t^3 + 120t^2 + 72t^3}{4^2(5 - 3t^2)^2}$$
$$= \frac{5(3t^2 + 6t + 5)}{4(5 - 3t^2)^2}$$

Differentiate.

$$G_X'(1) = \frac{35}{8}$$

$$E(X) = \frac{35}{8}$$

Set $t = 1$ to find $E(X)$.

c $G_X''(t) = \frac{4(5 - 3t^2)^2(5)(6t + 6) - 5(3t^2 + 6t + 5)(4)(2)(-6t)(5 - 3t^2)}{4^2(5 - 3t^2)^4}$

$$= \frac{(5 - 3t^2)(5)(3t + 3) - 5(3t^2 + 6t + 5)(-6t)}{2(5 - 3t^2)^3}$$
$$= \frac{75t + 75 - 45t^3 - 45t^2 + 90t^3 + 180t^2 + 150t}{2(5 - 3t^2)^3}$$
$$= \frac{15(3t^3 + 9t^2 + 15t + 5)}{2(5 - 3t^2)^3}$$

$$G_X''(1) = 30$$

$$\text{Var}(X) = G_X''(1) + G_X'(1) - [G_X'(1)]^2$$

$$\text{Var}(X) = 30 + \frac{35}{8} - \left(\frac{35}{8}\right)^2$$
$$= \frac{975}{64}$$

We need $G_X''(1)$ to find $\text{Var}(X)$:
$\text{Var}(X) = G_X''(1) + G_X'(1) - [G_X'(1)]^2$

Checklist of learning and understanding

For a discrete probability distribution:

- The probability generating function is $G_X(t) = E(t^X) = \sum_x t^x P(X = x)$.

Standard probability generating functions:

Probability distribution	$P(X = r)$	$G_X(t)$
$\text{Bin}(n, p)$	${}^nC_r q^{n-r} p^r$	$(q + pt)^n$
$\text{Po}(\lambda)$	$\frac{e^{-\lambda}\lambda^r}{r!}$	$e^{\lambda(t-1)}$
$\text{Geo}(p)$	$q^{r-1}p$	$\frac{pt}{1-qt}$

$E(X)$ and $\text{Var}(X)$:

- $G_X(1) = 1$
- $E(X) = G'_X(1)$
- $\text{Var}(X) = G''_X(1) + G'_X(1) - [G'_X(1)]^2$
- $\text{Var}(X) = G''_X(1) + E(X) - [E(X)^2]$

For two independent random variables:

- $G_{X+Y}(t) = G_X(t) \times G_Y(t)$

To generalise to three or more random variables:

- $G_{X_1+\ldots+X_n}(t) = G_{X_1}(t) \times \ldots \times G_{X_n}(t)$

The probability generating function of a linear transformation:

- $G_{aX+b}(t) = t^b G_X(t^a)$

END-OF-CHAPTER REVIEW EXERCISE 12

1 The discrete random variable X is the number of times we throw a pair of fair die to get a sum of eight. Find the probability generating function, as well as the expected number of throws and the variance of X.

2 A discrete random variable, X, has the probability distribution function:

$$P(X = x) = \frac{k}{e^x} \quad x = 0, 1, 2, 3, \ldots$$

a Find the value of k and the probability generating function.

b Find $E(X)$.

c Find $\text{Var}(X)$.

P 3 A game consists of rolling two die, one die with six sides, the other die with eight sides, and adding the scores together. Both die are fair, with the first numbered 1 to 6 and the second numbered 1 to 8. Show that the probability generating function of Z, where Z is the sum of the scores on the two die, is:

$$\frac{t^2}{48} \times \frac{(1 - t^6)(1 - t^8)}{1 - t^2}$$

P 4 The variable Y can take only the values $1, 2, 3, \ldots$ and is such that $P(Y = r) = kP(X = r)$, where $X \sim \text{Po}(\lambda)$. Show that the probability generating function of Y is given by:

$$G_Y(t) = \frac{e^{\lambda t} - 1}{e^{\lambda} - 1}$$

CROSS-TOPIC REVIEW EXERCISE 2

1 A random sample of 40 observations of a random variable X, and a random sample of 25 observations of a random variable Y, are taken. The sample means and unbiased estimates are:

$$\bar{x} = 13.6 \qquad \bar{y} = 11.2 \qquad s_x = 6.3 \qquad s_y = 7.1$$

A test is performed at a significance level of $\alpha\%$ using the following hypotheses:

$$H_0: \mu_x - \mu_y = 0$$
$$H_1: \mu_x - \mu_y > 0$$

Given that H_0 is not rejected, find the possible values of α.

2 A school is considering buying a large number of safety devices to install in the classrooms. The devices are designed to activate sprinklers if the temperature in the room exceeds 65 °C. A sample of the devices are tested by slowly increasing the temperature and noting the temperatures, in °C, at which the sprinklers are activated. The results are as follows.

57.8 62.9 63.7 71.2 60.6 69.5 64.7 65.7

Use a t-test and the 10% significance level to examine whether the mean temperature at which the sprinklers are activated is 65 °C. Assume that the sample is random and mean temperatures are normally distributed.

3 A random variable X has a probability distribution, $f(x)$, given by:

$$f(x) = \begin{cases} ke^{-3x} & 0 \leqslant x \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

a Show that $k = \dfrac{3e^3}{e^3 - 1}$.

b Find the median value of X.

4 A random sample of five metal rods produced by a machine is taken. Each rod is tested for hardness. The results, in suitable units, are as follows.

524 526 520 523 530

i Assuming a normal distribution, calculate a 95% confidence interval for the population mean.

Some adjustments are made to the machine. Assume that a normal distribution is still appropriate and that the population variance remains unchanged. A second random sample, this time of ten metal rods, is now taken. The results for hardness are as follows.

525 520 522 524 518 520 519 525 527 516

ii Stating suitable hypotheses, test at the 10% significance level whether there is any difference between the population means before and after the adjustments.

Cambridge International AS & A Level Further Mathematics 9231 Paper 22 Q9 November 2011

5 A biased die has probability p of showing a 6. The random variable X counts the number of trials up to and including the trial in which the first 6 is obtained. The random variable Y counts the number of trials up to and including the trial in which the nth 6 is obtained.

a Write down an expression for $P(X = x)$ for $X = 1, 2, \ldots$. Show that the probability generating function of X is:

$$G_X(t) = pt(1 - qt)^{-1}$$

where $q = 1 - p$. Hence show that the mean and variance of X are, respectively:

$$E(X) = \frac{1}{p}$$

$$\text{Var}(X) = \frac{q}{p^2}$$

b Given that the trials for X are independent and that $Y = X_1 + X_2 + \cdots + X_n$, find:

i the probability generating function for Y

ii $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$.

6 The owner of three driving schools, A, B and C, wished to assess whether there was an association between passing the driving test and the school attended. He selected a random sample of learner drivers from each of his schools and recorded the numbers of passes and failures at each school. The results that he obtained are shown in the table below.

	Driving school attended		
	A	B	C
Passes	23	15	17
Failures	27	25	43

Using a χ^2-test and a 5% level of significance, test whether there is an association between passing or failing the driving test and the driving school attended.

Cambridge International AS & A Level Further Mathematics 9231 Paper 2 Q8 November 2010

7 Trees of the same species grow on opposite sides of a river valley. The heights of eight randomly selected trees from each side of the river valley are measured and the results, in metres, are given in the table.

East side (m)	13.6	20.1	7.8	18.2	8.5	7.7	16.0	13.4
West side (m)	5.9	7.5	9.6	11.0	4.1	7.1	10.1	12.2

Carry out a Wilcoxon rank-sum test, at the 5% level of significance, to investigate whether there is any difference in the average heights of the trees from the two sides of the river valley.

Interpret your conclusion in context.

8 The continuous random variable X has probability density function, f, given by:

$$\mathrm{f}(x) = \begin{cases} \frac{1}{15}((x-3)^2 + 2) & 0 \leqslant x \leqslant 3 \\ 0 & \text{otherwise} \end{cases}$$

The random variable Y is defined by $Y = X^2$.

a Show that the cumulative distribution function $\mathrm{G}(y)$ is given by:

$$\mathrm{G}(y) = \begin{cases} 0 & y < 0 \\ \frac{1}{45}\left(y^{\frac{3}{2}} - 9y + 33y^{\frac{1}{2}}\right) & 0 \leqslant y \leqslant 9 \\ 1 & y > 9 \end{cases}$$

b Show that the median is 0.75, correct to 2 decimal places.

c Find $\mathrm{E}(Y)$.

9 A charity launched a campaign to raise awareness of low pay amongst full-time workers in company canteens. The charity now believes the median weekly wage in this job sector has increased.

After the campaign, a random sample of ten canteen workers was asked how much they earned in the previous week. The results, in NZ$, were as follows.

156.45 145.50 151.30 150.70 156.10 151.15 144.40 146.60 163.75 157.60

Before the campaign the median weekly wage for full-time workers in the company canteens was NZ$ 147.50.

Carry out a Wilcoxon signed-rank test to determine whether there has been an increase in the median wage after the campaign. Use the 5% level of significance.

10 It has been found that 60% of the computer chips produced in a factory are faulty. As part of quality control, 100 samples of 4 chips are selected at random, and each chip is tested. The number of faulty chips in each sample is recorded, with the results given in the following table.

Number of faulty chips	0	1	2	3	4
Number of samples	2	12	27	49	10

The expected values for a binomial distribution with parameters $n = 4$ and $p = 0.6$ are given in the following table.

Number of faulty chips	0	1	2	3	4
Number of samples	2.56	15.36	34.56	34.56	12.96

i Show how the expected value 34.56 corresponding to 2 faulty chips is obtained.

ii Carry out a goodness of fit test at the 5% significance level, and state what can be deduced from the outcome of the test.

Cambridge International AS & A Level Further Mathematics 9231 Paper 2 Q9 November 2009

11 The Tax Bureau claims that people typically take 140 minutes to fill in a tax form. A researcher believes that this claim is incorrect and that, generally, it takes people longer to complete the form. She recorded the time (in minutes) it took ten people to complete the form. The results are given below.

151 138 132 149 145 152 141 148 162 146

Carry out a sign test, at the 5% significance level, to investigate whether the average time to complete the form is greater than 140 minutes.

FURTHER PROBABILITY & STATISTICS PRACTICE EXAM-STYLE PAPER

1 An experiment is carried out to investigate memory in two situations: aural and visual. A random sample of 12 students are shown 20 objects and then each student is asked to recall as many of the objects as possible. Subsequently a list of 20 different objects is read out to them and again each student is asked to recall as many as possible. The results are given below.

Student	A	B	C	D	E	F	G	H	J	K	L	M
Visual	14	15	10	9	8	13	17	12	14	7	15	6
Aural	12	13	9	10	7	14	16	10	11	8	14	5

Use a sign test to determine whether or not there is evidence that students can recall objects with greater accuracy when they are presented visually rather than aurally. Use a 5% significance level. **[8]**

2 A university librarian selects a random sample of seven students from the physics department and records the number of books each student borrows in a particular month. The results are as follows.

$$11, 13, 6, 8, 10, 17, 5$$

During the same month the librarian selects a random sample of eight students from the geography department and the number of books each of these students borrows that month is as follows.

$$19, 9, 20, 15, 16, 12, 14, 18$$

a Using the Wilcoxon rank-sum test at the 5% level of significance, test whether or not there is any difference between the median number of books borrowed by students from these two departments. **[7]**

b Explain briefly how your test would be modified if 50 students had been randomly selected from each department. **[3]**

3 Four torpedoes are fired independently from a ship at a target. Each one has a $\frac{1}{3}$ probability of hitting the target. The random variable X represents the number of hits and has probability generating function:

$$G_X(t) = \frac{1}{81}(2 + t)^4$$

a Find the mean and the variance of X. **[2]**

A second ship fires at the same target and the random variable Y, representing its number of hits, has probability generating function:

$$G_Y(t) = \frac{1}{243}(2 + t)^5.$$

Given that X and Y are independent:

b find the probability generating function of $Z = X + Y$ **[2]**

c calculate the mean and the variance of Z. **[6]**

4 The manager of a hotel collected data on the usage of the facilities at the hotel by its guests. A random sample from her records is summarised below.

Facility	Male	Female
Spa	40	68
Swimming pool	26	33
Business centre	52	31

Making your method clear, test whether or not there is any evidence of an association between gender and use of the hotel's facilities. Use a 5% significance level. **[11]**

5 A continuous random variable has probability density function f given by

$$f(x) = \begin{cases} \frac{1}{30}(x+4) & 0 \leqslant x < 2 \\ \frac{7}{96} & 2 \leqslant x < 6 \\ \frac{1}{60}(15-x) & 6 \leqslant x \leqslant 9 \\ 0 & \text{otherwise.} \end{cases}$$

a Find $E(X)$. [3]

b Find the cumulative distribution function. [5]

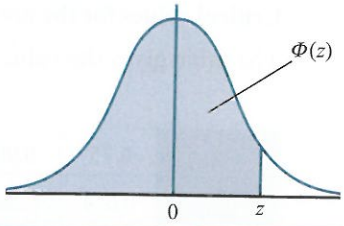
c Find the median. [3]

THE STANDARD NORMAL DISTRIBUTION FUNCTION

If Z is normally distributed with mean 0 and variance 1, the table gives the value of $\Phi(z)$ for each value of z, where

$$\Phi(z) = P(Z \leqslant z).$$

Use $\Phi(-z) = 1 - \Phi(z)$ for negative values of z.



z	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
											ADD								
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359	4	8	12	16	20	24	28	32	36
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	4	8	12	16	20	24	28	32	36
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	4	8	12	15	19	23	27	31	35
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517	4	7	11	15	19	22	26	30	34
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879	4	7	11	14	18	22	25	29	32
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224	3	7	10	14	17	20	24	27	31
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	3	7	10	13	16	19	23	26	29
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852	3	6	9	12	15	18	21	24	27
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133	3	5	8	11	14	16	19	22	25
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389	3	5	8	10	13	15	18	20	23
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	2	5	7	9	12	14	16	19	21
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830	2	4	6	8	10	12	14	16	18
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015	2	4	6	7	9	11	13	15	17
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177	2	3	5	6	8	10	11	13	14
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319	1	3	4	6	7	8	10	11	13
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441	1	2	4	5	6	7	8	10	11
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545	1	2	3	4	5	6	7	8	9
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633	1	2	3	4	4	5	6	7	8
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706	1	1	2	3	4	4	5	6	6
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767	1	1	2	2	3	4	4	5	5
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817	0	1	1	2	2	3	3	4	4
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857	0	1	1	2	2	2	3	3	4
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890	0	1	1	1	2	2	2	3	3
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916	0	1	1	1	1	2	2	2	2
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936	0	0	1	1	1	1	1	2	2
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952	0	0	0	1	1	1	1	1	1
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964	0	0	0	0	1	1	1	1	1
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974	0	0	0	0	0	1	1	1	1
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981	0	0	0	0	0	0	0	1	1
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986	0	0	0	0	0	0	0	0	0

Critical values for the normal distribution

The table gives the value of z such that $P(Z \leqslant z) = p$, where $Z \sim N(0, 1)$.

p	0.75	0.90	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995
z	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

5 Proof

6 Proof

7 Proof

8 Proof

End-of-chapter review exercise 7

1 Proof

2 For all cases 1, $-1 + (1 + n)!$, Proof

3 Proof

Cross-topic review exercise 1

1 Proof, $\frac{1}{3}n(7n^2 - 1)$

2 **a** $S_1 = 5, S_3 = 122$

b $y^3 - 25y^2 + 10y - 1 = 0$

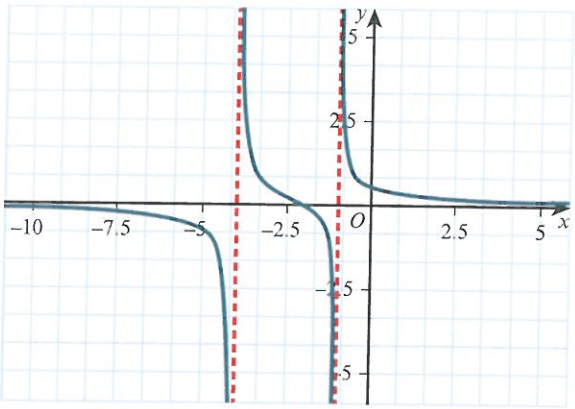
c 14878

3 **a** $x = -4, x = -1, y = 0$

b $\frac{dy}{dx} = \frac{-x^2 - 4x - 6}{(x^2 + 5x + 4)^2}$

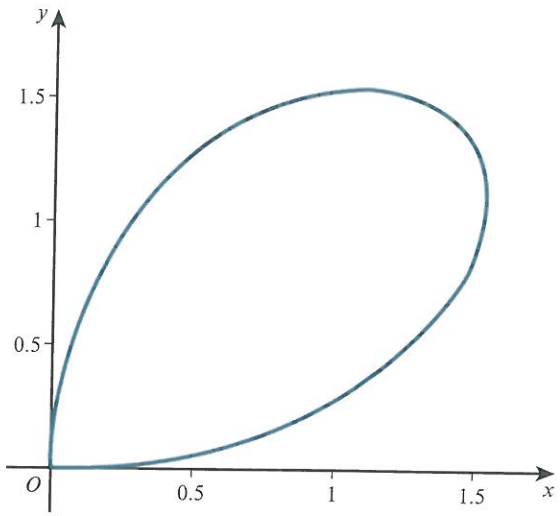
c 0

d Intercepts are $(-2, 0), \left(0, \frac{1}{2}\right)$



4 $\mathbf{A} = \begin{pmatrix} 1 & -2 & 3 \\ -1 & 4 & 1 \\ 2 & -2 & 9 \end{pmatrix}$ means the planes all meet at a single point

5 **a**



b $\frac{8\sqrt{3}}{9}$ **c** $\frac{\pi}{2}$

6 **a** 90° **b** $\frac{\sqrt{6}}{2}$

7 Proof

8 **a** 4 **b** Proof

9 $\frac{1}{4r-2} - \frac{1}{4r+2}; \frac{1}{2} - \frac{1}{4n+2}; \frac{1}{2}$

10 **i** $y = 5 + \frac{4}{x-2} - \frac{4}{x+3}$ **ii** $y = 1.8$

iii $x = 2, x = -3, y = 5$ **iv** $1.8 < k < 5$

11 $B = \begin{pmatrix} -\frac{1}{4} & -\frac{5}{2} & \frac{5}{2} \\ \frac{1}{2} & \frac{1}{2} & -1 \\ -\frac{1}{4} & \frac{3}{2} & -1 \end{pmatrix}$

12 Proof

8 Continuous random variables

Prerequisite knowledge

1 $E(X) = 2.3, \text{Var}(X) = 0.81$

2 $\frac{98}{3}$

Exercise 8A

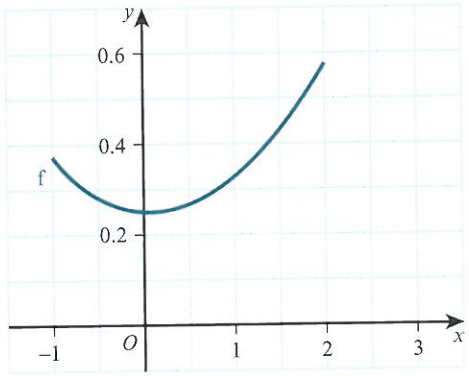
1 **a** Yes, because $f(x) \geqslant 0$ and $\int f(x)dx = 1$

b No, $\int f(x)dx = 1$ but $f(x) < 0$ for some x

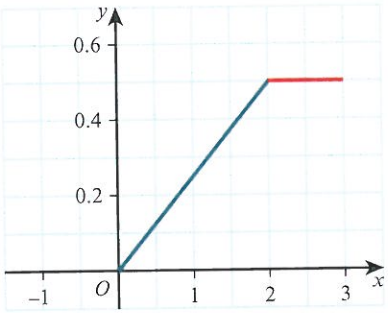
c Yes because $f(x) \geqslant 0$ and $\int f(x)dx = 1$

d No since $\int f(x)dx \neq 1$

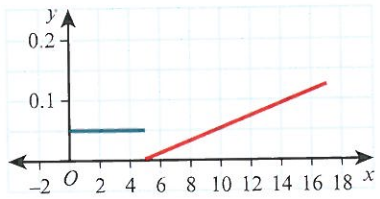
2 **a**



b



c



3 $k = \frac{2}{9}$

4 $k = \frac{2e^4}{e^2 - 1}$

5 **a** $k = \frac{1}{32}$ **b** 0 **c** $\frac{45}{64}$ **d** $\frac{51}{64}$

6 **a** $k = \frac{1}{15}$ **b** $\frac{7}{45}$ **c** $\frac{11}{18}$

7 $k = \frac{4}{81}$

8 $k = -\frac{1}{19}$

9 **a** $k = \frac{1}{16}$ **b** $\frac{1}{4}$ **c** $\frac{9}{32}$ **d** $\frac{23}{32}$

10 **a** $\frac{1}{36}$ **b** $\frac{1}{2}$ **c** $\frac{31}{32}$

Exercise 8B

1 $$F(x) = \begin{cases} 0 & x < -4 \\ \frac{1}{95}(-3x^2 + 10x + 88) & -4 \leqslant x \leqslant 1 \\ 1 & \text{otherwise} \end{cases}$$

2 $$F(x) = \begin{cases} 0 & x < 2 \\ \frac{1}{27}(-68 + 54x - 12x^2 + x^3) & 2 \leqslant x \leqslant 5 \\ 1 & x > 5 \end{cases}$$

3 $$F(x) = \begin{cases} 0 & x < 3 \\ \frac{1}{16}(x - 3) & 3 \leqslant x \leqslant 7 \\ \frac{1}{8}(x - 5) & 7 \leqslant x \leqslant 13 \\ 1 & x > 13 \end{cases}$$

4 $$F(x) = \begin{cases} 0 & x < 1 \\ \frac{2}{27}(x^2 - 2x + 1) & 1 \leqslant x < 4 \\ \frac{2}{27}(-2x^2 + 22x - 47) & 4 \leqslant x \leqslant \frac{11}{2} \\ 1 & x > \frac{11}{2} \end{cases}$$

5 $$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{24} & 0 \leqslant x < 5 \\ \frac{1}{48}(x^2 - 8x + 25) & 5 \leqslant x < 7 \\ \frac{1}{8}(x - 4) & 7 \leqslant x < 12 \\ 1 & x > 12 \end{cases}$$

6 **a** $$F(x) = \begin{cases} 0 & x < 3 \\ \frac{1}{56}(-x^2 + 24x - 63) & 3 \leqslant x \leqslant 7 \\ 1 & x > 7 \end{cases}$$

b $\frac{4}{7}$ **c** $\frac{1}{2}$ **d** 4.72

7 **a** $$F(x) = \begin{cases} 0 & x < 2 \\ -\frac{1}{100}(x - 11)(x - 2)^2 & 2 \leqslant x \leqslant 7 \\ 1 & x > 7 \end{cases}$$

b $\frac{7}{25}$ **c** $\frac{23}{50}$

8 **a** $F(x) = \begin{cases} 0 & x < 4 \\ \frac{1}{180}(-x^2 + 26x - 88) & 4 \leqslant x < 7 \\ \frac{1}{540}(x^2 + 22x - 68) & 7 \leqslant x \leqslant 16 \\ 1 & x > 16 \end{cases}$

b $\frac{8}{45}$ **c** $\frac{59}{108}$ **d** 7 **e** $m = 10.4$

9 **a** $F(x) = \begin{cases} 0 & x < 6 \\ \frac{1}{99}\left(\frac{x^3}{3} - 9x^2 + 83x - 246\right) & 6 \leqslant x \leqslant 15 \\ 1 & x > 15 \end{cases}$

b 0.872 **c** Proof

10 **a** $-\frac{8}{3}$ **b** Proof

Exercise 8C

1 **a** $\frac{40}{3}$ **b** 200 **c** $\frac{200}{9}$

2 **a** $\frac{23}{7} = 3.286$ **b** 11.5 **c** $\frac{69}{98} = 0.704$

3 **a** 8.3 **b** 75.4 **c** 6.51 **d** 2.55

4 **a** $4.96 = \left(\frac{119}{24}\right)$ **b** $33.9 = \left(\frac{407}{12}\right)$

c 9.33 using exact values

5 $\frac{5}{64}(36 + \ln 5)$

6 $\frac{502}{7} = 71.7$

7 $\frac{e(e^2 - 2)}{2}$

8 $\frac{15}{8}\ln\left(\frac{3}{2}\right) - \frac{23}{32}$

9 $\frac{5}{3}$

10 0.482

11 **a** $f(x) = \begin{cases} \frac{1}{36}(3x^2 + 5) & 1 \leqslant x \leqslant 3 \\ 0 & \text{otherwise} \end{cases}$

Median, $F(2.315) < 0.5$, $F(2.325) > 0.5$,
Therefore median = 2.32 correct to 3 significant figures

b Mode = 3

c $E(X) = 2.22$

d Since mean < median < mode the skew is negative.

Exercise 8D

1 $G(a) = \begin{cases} 0 & a < 0 \\ \frac{a}{400} & 0 \leqslant a \leqslant 400 \\ 1 & a > 400 \end{cases}$

2 $G(a) = \begin{cases} 0 & a < 0 \\ \frac{1}{10}(a + \sqrt[3]{a}) & 0 \leqslant a \leqslant 8 \\ 1 & a > 8 \end{cases}$

3 $G(a) = \begin{cases} 0 & a < -34 \\ \frac{-1}{285}a(34 + a) & -34 \leqslant a \leqslant -19 \\ 1 & a > -19 \end{cases}$

4 **a** $G(a) = \begin{cases} 0 & a < 1 \\ \frac{1}{3}(a - 1) & 1 \leqslant a \leqslant 4 \\ 1 & a > 4 \end{cases}$

b $G(b) = \begin{cases} 0 & b < 1 \\ \frac{1}{3}(b^4 - 1) & 1 \leqslant b \leqslant \sqrt{2} \\ 1 & b > \sqrt{2} \end{cases}$

5 $G(y) = \begin{cases} 0 & y < 0 \\ \frac{1}{300}(y + 20\sqrt{y}) & 0 \leqslant y \leqslant 100 \\ 1 & y > 100 \end{cases}$

6 **a** $F(x) = \begin{cases} 0 & x < 0 \\ -\frac{1}{16}(x^2 - 8x) & 0 \leqslant x \leqslant 4 \\ 1 & x > 4 \end{cases}$

b $G(y) = \begin{cases} 0 & y < -0 \\ -\frac{1}{144}(y^2 - 20y - 44) & -0 < y \leqslant 10 \\ 1 & y > 10 \end{cases}$

c $g(y) = \begin{cases} \frac{1}{72}(10 - y) & -0 \leqslant y \leqslant 10 \\ 0 & \text{otherwise} \end{cases}$

7 **a** $F(x) = \begin{cases} 0 & x < 1 \\ \frac{2(x - 1)}{x} & 1 \leqslant x \leqslant 2 \\ 1 & x > 2 \end{cases}$

b $G(y) = \begin{cases} 0 & y < \frac{1}{4} \\ \frac{2\sqrt{y}-1}{\sqrt{y}} & \frac{1}{4} \leqslant y \leqslant 1 \\ 1 & y > 1 \end{cases}$

c $g(y) = \begin{cases} \frac{1}{2y^{\frac{3}{2}}} & \frac{1}{4} \leqslant y \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$

8 **a** $G(y) = \begin{cases} 0 & y < \frac{1}{5} \\ 1 + \frac{25}{24}(y^2 - 1) & \frac{1}{5} \leqslant y \leqslant 1 \\ 1 & y > 1 \end{cases}$

b $g(y) = \begin{cases} \frac{50y}{24} & \frac{1}{5} \leqslant y \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$

9 **a** $F(x) = \begin{cases} 0 & x < 0 \\ -\frac{1}{25}(x^2 - 10x) & 0 \leqslant x \leqslant 5 \\ 1 & x > 5 \end{cases}$

b $G(y) = \begin{cases} 0 & y < -5 \\ \frac{1}{100}(y+5)^2 & -5 \leqslant y \leqslant 5 \\ 1 & y > 5 \end{cases}$

c $\frac{49}{100}$ **d** $\frac{2}{5}$

e $g(y) = \begin{cases} \frac{1}{50}(y+5) & -5 \leqslant y \leqslant 5 \\ 0 & \text{otherwise} \end{cases}$

10 **a** $F(r) = \begin{cases} 0 & r < 1 \\ \frac{1}{25}(-r^2 + 12r - 11) & 1 \leqslant r \leqslant 6 \\ 1 & r > 6 \end{cases}$

b

$$G(A) = \begin{cases} 0 & A < \pi \\ -\left(\frac{1}{25\pi}\right)(A - 12\sqrt{A\pi} + 11\pi) & \pi \leqslant A \leqslant 36\pi \\ 1 & A > 36\pi \end{cases}$$

c $g(A) = \begin{cases} \frac{1}{25\pi}\left(\frac{6\sqrt{\pi}}{\sqrt{A}} - 1\right) & \pi \leqslant A \leqslant 36\pi \\ 0 & \text{otherwise} \end{cases}$

End-of-chapter review exercise 8

1 **i** $E(T) = 100$ **ii** 69.3 **iii** 0.819

2 **a** $k = \frac{1}{4}$

b **i** $g(y) = \begin{cases} \frac{1}{16} & 0 \leqslant y \leqslant 4 \\ \frac{1}{8\sqrt{y}} & 4 \leqslant y \leqslant 25 \\ 0 & \text{otherwise} \end{cases}$

ii Proof

iii $m_x = 3$, $m_y = 9$, hence $m_y = m_x^2$

3 **i** Proof

ii $\lambda = 0.174$, median $= \lambda^{-1} \ln 2 = 3.98$

9 Inferential statistics

Prerequisite knowledge

1 0.879

2 2.359

Exercise 9A

1 **a** $\bar{x} = 15.583$, $s^2 = 9.174$

b $\bar{x} = 145.2$, $s^2 = 177.96$

2 **a** $t_{0.95,\,10} = 1.812$ **b** $t_{0.975,\,20} = 2.086$

c $t_{0.975,\,14} = 2.145$ **d** $t_{0.995,\,24} = 2.797$

e $t_{0.9,\,7} = 1.415$ **f** $t_{0.95,\,17} = 1.740$

3 **a** $H_0: \mu = 41$, $H_1: \mu \neq 41$

b $H_0: \mu = 7.3$, $H_1: \mu > 7.3$

c $H_0: \mu = 54.2$, $H_1: \mu < 54.2$

d $H_0: \mu = 6.5$, $H_1: \mu \neq 6.5$

4 **a** $t_{0.95,\,9} = 1.833$; reject

b $t_{0.99,\,7} = 2.998$; do not reject

c $t_{0.95,\,14} = 1.761$; do not reject

d $t_{0.975,\,10} = 2.228$; reject

5 **a** $\bar{x} = 49.7$, $s^2 = 10.42$

b H_0: $\mu = 50.5$, H_1: $\mu < 50.5$, $t = -0.894$, $t_{0.95,\,11} = -1.796$. Do not reject H_0: The mean length is 50.5 cm, based on this sample.

6 H_0: $\mu = 60$, H_1: $\mu \neq 60$, $t = 0.494$, $t_{0.975,\,9} = 2.262$, Reject H_0: The amount of paracetamol is different from 60 mg.

7 **a** H_0: $\mu = 175$, H_1: $\mu < 175$, $\bar{x} = 174$, $s^2 = 2.377$, $t = -1.032$, $t_{0.95,\,7} = -1.895$. Do not reject H_0: The crisp packets weigh 175 g.

b The packets that are tested cannot be sold afterwards, so as few as possible should be opened (this is called destruction testing).

8 H_0: $\mu = 5000$, H_1: $\mu > 5000$, $\bar{x} = 5004.1$, $s^2 = 36.32$, $t = 2.151$, $t_{0.95,\,9} = 1.83$

Reject H_0: The pump is giving more than 5 litres of petrol.

Exercise 9B

1 **a** $s_p^2 = 12.6$ **b** $s_p^2 = 159.86$

c $s_p^2 = 38.67$

2 **a** $s_p^2 = 24.68$ **b** $s_p^2 = 7.746$

3 **a** $t_{0.95,\,12} = 1.782$ **b** $t_{0.975,\,22} = 2.074$

c $t_{0.975,\,13} = 2.160$ **d** $t_{0.995,\,30} = 2.750$

e $t_{0.9,\,23} = 1.319$ **f** $t_{0.95,\,27} = 1.703$

4 **a** H_0: $\mu_x - \mu_y = 0$, H_1: $\mu_x - \mu_y \neq 0$

b H_0: $\mu_x - \mu_y = 0$, H_1: $\mu_x - \mu_y > 0$

c H_0: $\mu_x - \mu_y = 5$, H_1: $\mu_x - \mu_y > 5$

d H_0: $\mu_x - \mu_y = 6$, H_1: $\mu_x - \mu_y \neq 6$

5 NB: we are given $\sigma = 15$, H_0: $\mu_A - \mu_B = 0$, H_1: $\mu_A - \mu_B < 0$. (If more severe, A should be giving fewer marks on average.) Test statistic $= -1.7348$, $z = -1.96$. Do not reject H_0: Examiner A is not more severe than examiner B.

6 **a** $s_p^2 = 0.0599$ (4 decimal places)

b H_0: $\mu_M - \mu_F = 0$, H_1: $\mu_M - \mu_F > 0$, $t = 1.056$, $t_{0.95,\,16} = 1.746$

Do not reject H_0: Male Takahē birds are not significantly heavier than female Takahē birds.

7 **a** $s_p^2 = 15.03$

b H_0: $\mu_N - \mu_O = 10$, H_1: $\mu_N - \mu_O > 10$, $t = -0.915$, $t_{12,\,0.95} = -1.782$

Do not reject H_0: The council is correct in stating that the new route will not add more than ten minutes to the journey.

8 H_0: $\mu_B - \mu_A = 0$, H_1: $\mu_B - \mu_A \neq 0$, $t = 0.91$, $t_{0.975,\,8} = 2.306$

Do not reject H_0: There is no difference in the viscosities of the two types of honey.

Exercise 9C

1 **a** $\bar{d} = -0.333$, $s_d^2 = 135$

b $\bar{d} = 0.683$, $s_d^2 = 12.578$

c $\bar{d} = -5.75$, $s_d^2 = 190.93$

2 **a** 0.104 **b** −0.638 **c** −0.210

3 **a** $t_{0.975,\,10} = 2.228$ **b** $t_{0.95,9} = 1.833$

c $t_{0.975,\,11} = 2.201$

4 $\bar{d} = 0.0338$, $s_d^2 = 0.00183$, H_0: $\mu_d = 0$, H_1: $\mu_d > 0$, $t = 2.233$, $t_{0.975,\,7} = 2.365$

Reject H_0: The weight after the 14 days has increased significantly.

5 H_0: $\mu_d = 2$, H_1: $\mu_d > 2$, $\bar{d} = 2.225$, $s_d = 0.931589$, $t = 0.659$, $t_{0.95,\,7} = 1.895$

Do not reject H_0: The claim that participants will lose at least 2 kg in the first five weeks is not true.

6 H_0: $\mu_d = 0$, H_1: $\mu_d \neq 0$, $t = 2.934$, $t_{0.975,\,7} = 2.365$

Reject H_0: The course did not make a difference to the memory test scores.

7 H_0: $\mu_d = 0$, H_1: $\mu_d > 0$, $t = 1.057$, $t_{0.95,\,5} = 2.015$

Do not reject H_0: The visits of the psychologist have not had a positive impact on sales productivity.

Exercise 9D

1 **a** $t_{0.95,\,5} = 2.015$ **b** $t_{0.95,\,7} = 1.895$

c $t_{0.975,\,11} = 2.201$ **d** $t_{0.9,\,6} = 1.440$

2 **a** 1.061 **b** 1.414

c 1.563 **d** 0.917

3 a (10.358, 16.042) b (9.710, 16.690)
c (10.570, 15.830) d (9.276, 17.124)
e (8.455, 17.945) f (11.344, 15.056)

4 (13.32, 17.68)

5 a $\bar{x} = 7.153$, $s^2 = 0.002347$, $t_{6, 0.975} = 2.447$, [7.108, 7.198]
b Since the confidence interval is totally above the value of 7.1, the claim by the car hire company is wrong.

6 [40.10, 46.57]

7 a [453.6, 459.3]
b The confidence interval contains 454 (just!) and so the claim is justified.

8 NB: since n is large we can use the Central Limit Theorem and we can use s^2 as an unbiased estimator for σ^2. CI: we use $z_{0.995}$, $37.2 \pm 2.576 \times \frac{3.2}{\sqrt{36}}$, [35.83, 38.57]

Exercise 9E

1 a 1.645 b 1.960
c 2.576 d 1.282

2 a $t_{0.95, 12} = 1.782$ b $t_{0.975, 21} = 2.080$
c $t_{0.995, 28} = 2.763$ d $t_{0.9, 18} = 1.330$

3 a $\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} = 0.789$, confidence interval (−0.907, 1.687)
b $s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}} = 1.546$, confidence interval (−3.350, 2.010)
c $s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}} = 9.939$, confidence interval (−18.643, 24.303)

4 a (10.037, 14.163) b (8.87, 15.33)

5 a $s_p^2 = 1.061$
b $t_{0.9, 20} = 1.725$, [−1.56, −0.04]

6 a [0.419, 3.081]
b Since the confidence interval is wholly above the value of 0, we can conclude that sleep deprivation does affect reaction time.

7 Since both n are large, we can use the z-statistic to do this. $z_{0.975} = 1.96$, [0.555, 1.845]

8 NB: we need to use pooled variance here. $s_p^2 = 0.0909$, [−0.590, 0.111]

End-of-chapter review exercise 9

1 a $\bar{x} = 6.1$, $s_x^2 = 1.15^2 = 1.322$, H_0: $\mu = 5.2$, H_1: $\mu > 5.2$, $t = 2.47$, $t_{0.95, 9} = 1.83$
Reject H_0: The mean is greater than 5.2.
b H_0: $\mu_P - \mu_Q = 0$, H_1: $\mu_P - \mu_Q < 0$. Assume distributions have equal variances.
$\bar{y} = 7.0$, $s_y^2 = 1.085^2$, $s^2 = 1.118^2$, $t = 1.8$, $t_{18, 0.95} = 1.73$
Reject H_0: Therefore mean of Q is greater.

2 H_0: $\mu = 7.5$, H_1: $\mu < 7.5$, $\bar{x} = 7.04$, $s^2 = 0.9707^2$, $t = 1.499$, $t_{0.9, 9} = 1.383$
Reject H_0: The mean is less than 7.5.

3 $s^2 = 0.9273$; $z = 1.869$; $\alpha \leq 6.2$

10 Chi-squared tests

Prerequisite knowledge

1 0.879

2 0.181

3 0.7599

4 0.618

5 0.45

Exercise 10A

1 a 13.36 b 19.68 c 13.28 d 41.40

2 a

x_i	1	2	3	4
E_i	5	20	45	80

b

	a	*b*	*c*	*d*
E_i	80	80	40	40

c

x_i	2	3	4	5	6	7	8
E_i	40	40	40	40	40	40	40

3 a $\chi^2 = 8.472$
b $\chi^2 = 1.837$
c $\chi^2 = 8.494$

4 **a**

n	1	2	3	4
Expected frequency	32	24	16	8

b H_0: the model is a good fit
H_1: the model is not a good fit

c 4.3229 **d** $\nu = 4 - 1 = 3$

e $\chi^2_3(0.95) = 7.815$

f Since $4.3229 < 7.815$, there is insufficient evidence to reject H_0: Therefore the model is a good fit.

5

n	O	A	B	AB
Expected frequency	100	60	40	20

H_0: The blood groups are in the proportion 5:3:2:1
H_1: The blood groups are not in the proportion 5:3:2:1
$\chi^2 = 10.5817$, $\nu = 4 - 1 = 3$, $\chi^2_3(0.95) = 7.815$
Since $10.5817 > 7.815$, there is sufficient evidence to reject H_0: Therefore the blood groups in the neighbouring country are not in the proportion 5:3:2:1.

6 H_0: A uniform distribution is a good fit
H_1: A uniform distribution is not a good fit
$\chi^2 = 12.5$, $\nu = 10 - 1 = 9$, $\chi^2_9(0.95) = 16.92$
Do not reject H_0: A uniform distribution is a good fit.

Exercise 10B

1 **a** 25.931

b 14.817

(Or, if s is found first, $s = 14.818$ and r will then be 25.930.)

2 **a** 23.137 **b** 16.215

3 **a** Proof

b $r = 16.036$, $s = 11.204$

c It is necessary that each $E_i \geqslant 5$. Combining the columns in this way guarantees that.

4 $\chi^2 = 5.167$; degrees of freedom = 5. The last two columns need to be combined.

5 H_0: Binomial is a good fit
H_1: Binomial is not a good fit
$\chi^2 = 2.944$, $\chi^2_3(0.95) = 7.815$
Do not reject H_0: Binomial is a good fit.

6 H_0: Poisson $\lambda = 2.5$ is a good fit
H_1: Poisson $\lambda = 2.5$ is not a good fit
$\chi^2 = 2.944$, $\chi^2_5(0.95) = 11.070$
Reject H_0: Poisson with $\lambda = 2.5$ is not a good fit.

7 **a** Proof

b $p = 30.80$, $q = 15.19$

c H_0: Poisson is a good fit
H_1: Poisson is not a good fit
$\chi^2 = 11.442$, $\chi^2_3(0.99) = 11.34$
Reject H_0: A Poisson distribution is not a good model.

Exercise 10C

1 **a**

E_i	16	16	8	8	32

b

E_i	18	9	9	9	18	18

c

E_i
131.58
90.643
121.35
73.556
82.871

The last group has been adjusted to ensure that $\Sigma E_i = 500$.

2 **a** 4.156 **b** 2.778

c 1.299

3 **a** **i** $\nu = 9 - 1 - 1 = 7$

ii $\chi^2_7(0.95) = 14.07$

b **i** $\nu = 7 - 1 - 2 = 4$

ii $\chi^2_4(0.9) = 7.779$

c **i** $\nu = 11 - 1 = 10$

ii $\chi^2_{10}(0.975) = 20.48$

d **i** $\nu = 15 - 1 - 1 = 13$

ii $\chi^2_{13}(0.95) = 22.36$

4 **a** 0.1303 **b** 0.7851

c 0.0825 **d** 0.00202

5 H_0: Machine is equally likely to cut between 24.5 and 25.5 m

H_1: Machine is not equally likely to cut between 24.5 and 25.5 m

This could also be: H_0: L can be modelled as continuous uniform on [24.5, 25.5]

H_1: L cannot be modelled as continuous uniform on [24.5, 25.5]

$\chi^2 = 1.6$, $\nu = 4 - 1 = 3$, $\chi^2_3(0.95) = 7.815$

Do not reject H_0: Machine is equally likely to cut between 24.5 and 25.5 m.

6 H_0: $L \sim N(50, 1.5^2)$ is a good model

H_1: $L \sim N(50, 1.5^2)$ is not a good model

$\chi^2 = 223.861$, $\nu = 6 - 1 = 5$, $\chi^2_5(0.95) = 11.070$

Reject H_0: $L \sim N(50, 1.5^2)$ is not a good model for these data.

7 **a** Proof

b H_0: a normal distribution with $\mu = 47.5$ is a good model

H_1: a normal distribution with $\mu = 47.5$ is not a good model

$\chi^2 = 17.20$, $\nu = 7 - 1 - 1 = 5$,
$\chi^2_5(0.995) = 16.75$

Reject H_0: A normal distribution with $\mu = 47.5$ is not a good model.

8 H_0: the model is a good fit

H_1: the model is not a good fit

$\chi^2 = 1.097$, $\nu = 5 - 1 = 4$, $\chi^2_4(0.95) = 9.488$

Do not reject H_0: The model is a good fit.

Exercise 10D

1 **a** $E_{11} = 18$, $E_{31} = 12$, $E_{33} = 8$

b $E_{12} = 18.67$, $E_{24} = 11$, $E_{13} = 24.67$

c

10.41	11.19	5.4
11.96	12.84	6.2
18.51	19.89	9.6
13.12	14.08	6.8

The shaded cells have been adjusted to ensure that rows and column add correctly.

2 **a**

12	28
18	42

$\chi^2 = 0.794$

b

31.02	34.98
62.98	71.02

$\chi^2 = 2.251$

c

30.66	35.34
25.09	28.91
62.25	71.75

$\chi^2 = 2.366$

d

28.63	7.75	31.61
19.37	5.25	21.39

The first two rows have been combined to ensure $E_{ij} \geqslant 5$.

$\chi^2 = 0.386$

3 **a** 4 **b** 6

c 6. Either the first two rows or the last two columns must be combined

4 **a**

16.18	42.82
48.82	129.18

b $\chi^2 = 5.271$

c H_0: there is no association between X and Y

H_1: there is an association between X and Y

$\nu = 1$; $\chi^2_1(0.95) = 3.481$

Reject H_0: There is an association between X and Y.

5 H_0: Age group and loan type are independent

H_1: Age group and loan type are not independent

Expected values:

Loan type	Age group			Totals
	18–25	25–35	Over 35	
Good	42.67	32.67	24.67	100
Toxic	21.3	16.3	12.3	50
Totals	64	49	37	150

$\chi^2 = 0.898$, $\nu = (2 - 1)(3 - 1) = 2$, $\chi^2_2(0.9) = 4.605$

Do not reject H_0: Age group and loan type are independent.

6 a

	N	S	E	W	Total
Selected	21	12	18	29	80
Rejected	119	108	72	121	420
Total	140	120	90	150	500

b H_0: There is no association between region and being accepted

H_1: There is an association between region and being accepted

Expected table:

22.4	19.2	14.4	24	80
117.6	100.8	75.6	126	420
140	120	90	150	500

$\chi^2 = 5.630$, $\nu = (4-1)(2-1) = 3$,
$\chi^2_3(0.95) = 7.815$

Do not reject H_0: There is no association between region and being accepted.

7 H_0: There is no association between town and quality of mobile phone reception

H_1: There is an association between town and quality of mobile phone reception

$\chi^2 = 8.210$, $\nu = (3-1)(3-1) = 4$,
$\chi^2_4(0.95) = 9.488$

Do not reject H_0: There is no association between town and quality of mobile phone reception.

End-of-chapter review exercise 10

1 a Mean = 1.31, variance = 1.21

Since these are roughly the same, a Poisson distribution seems appropriate.

b i $q = 20.219$

ii H_0: Poisson is a good fit, H_1: Possion is not a good fit

Need to combine last three categories.

$\chi^2 = 5.542$, $\nu = 5 - 1 - 1$, $\chi^2_3(0.9) = 6.251$

Do not reject H_0: The Poisson is a good model.

2 H_0: Car type is independent of age group
H_1: Car type is not independent of age group

Or H_0: There is no association between car type and age group

H_1: There is an association between car type and age group

χ^2 test statistic = 12.6 (1 decimal place), $\nu = 4$, $\chi^2_4(0.95) = 9.488$

Since $12.6 > 9.488$ we reject H_0: There is sufficient evidence to suggest that car type and age group are not independent (there is an association).

3 a

$2 \leqslant x < 3$	40
$3 \leqslant x < 4$	20
$4 \leqslant x < 5$	12
$5 \leqslant x < 6$	8

b H_0: f(x) is a good fit, H_1: f(x) is not a good fit

$\chi^2 = 5.7$, $\nu = 4 - 1$, $\chi^2_3(0.9) = 6.251$

Since $5.7 < 6.251$, do not reject H_0: f(x) is a good fit

11 Non-parametric tests

Prerequisite knowledge

1 0.0654 (3 sf)

Exercise 11A

1 a 11 b 14 c 9 d 6

2 a 0.0730 b 0.0898
c 0.0592 d 0.0207

3 a 0.0287 b 0.0193
c 0.0037 d 0.0539

4 a Do not reject H_0. b Reject H_0.
c Reject H_0.

5 a E(S) = 7.5, Var(S) = 3.75, $z = -1.549$

b E(S) = 10, Var(S) = 5, $z = -2.012$

c E(S) = 7.5, Var(S) = 3.75, $z = 2.066$

6 H_0: The population median is 5.2
H_1: the population median is below 5.2

Test statistic = 3, $P(X \leqslant 3) = 0.0730$

Since $0.0730 > 0.05$, do not reject H_0: There is insufficient evidence to suggest that the pH is below 5.2.

7 H_0: Population median is 11, H_1: Population median is greater than 11.

Test stat = 6, $P(X \geqslant 6 | X \sim \text{Bin}(10, 0.5)) = 0.377$

Do not reject H_0: Population median weekly pocket money has not increased from 2016 to 2017.

8 a H_0: The population median is 7675, H_1: The population median is not 7675.

Test statistic = 4,
$P(X \leqslant 4 | X \sim \text{Bin}(12, 0.5)) = 0.194$

Do not reject H_0: The population median is 7675.

b Median would need to be greater than 8150 (so only 2 values created a positive)

9 H_0: the population median is 60, H_1: the population median is greater than 60.

$$S^+ = 16, z = \frac{16.5 - \left(\frac{20}{2}\right)}{\sqrt{\frac{20}{4}}} = 2.907, z_{0.95} = 1.6449$$

Reject H_0: The population median time is greater than 60 minutes.

Exercise 11B

1 a $P = 70, N = 50, T = 50$

b $P = 8, N = 112, T = 8$

c $P = 117, N = 3, T = 3$

2 a 5 b 19

c 32 d 8

3 The underlying data are symmetric; the underlying data are continuous; the data are independent.

4 a $E(T) = 232.5$, $\text{Var}(T) = 2363.75$
$z = -1.172$
Do not reject H_0.

b $E(T) = 105$, $\text{Var}(T) = 717.5$
$z = -1.699$
Reject H_0.

c $E(T) = 410$, $\text{Var}(T) = 5535$
$z = -2.009$
Reject H_0.

5 H_0: The population median is 4, H_1: The population median is greater than 4.

$T = 4$, Critical value = 10

Reject H_0: The population median is greater than 4.

6 a H_0: The population median is 43, H_1: The population median is below 43.

$T = \min(1, 35) = 1$, Critical value = 8

Reject H_0: The population median is below 43.

b Wilcoxon signed-rank is preferred because the magnitudes of the differences are taken into account.

7 H_0: The population median is 11.2, H_1: The population median is greater than 11.2.
$T = \min(7, 48) = 7$, Critical value = 10

Reject H_0: The population median is 11.2.

8 H_0: The population median is 50, H_1: The population median is less than 50.

$T = 55$, $E(T) = 517.5$, $\text{Var}(T) = 7848.75$

$$z = \frac{55.5 - 517.5}{\sqrt{7848.75}}$$

$z = -5.215$

$z_{0.99} = -2.326$

Reject H_0: The population median is less than 50.

Exercise 11C

1 a 4 b 2 c 2

2 a $n = 11$
Since data points B, I, K, have a zero difference they must be ignored.

b 6

3 a $P(X \leqslant 3 | X \sim \text{Bin}(9, 0.5)) = 0.254$
Do not reject H_0.

b $P(X \leqslant 1 | X \sim \text{Bin}(6, 0.5)) = 0.109$
Do not reject H_0.

c $P(X \leqslant 1 | X \sim \text{Bin}(8, 0.5)) = 0.0352$
Reject H_0.

4 a The value of n in this case is 11, which is greater than 10 and so an approximation would be valid.

b $E(S^+) = 5.5$, $\text{Var}(S^+) = 2.75$

c $z = 0.603$

5 H_0: There is no difference in the population medians

H_1: There is a difference in the population medians

Test statistic = 2 or 8 (choose 2 to use)

$P(X \leqslant 2 | X \sim \text{Bin}(10, 0.5)) = 0.0547$

Do not reject H_0: There is no difference in the average effectiveness of the aerosols.

6 H_0: The thickness of the cornea is the same in each eye

H_1: The thickness of the cornea is different in each eye

Test statistic = 5,

$P(X \geqslant 5 | X \sim \text{Bin}(7, 0.5)) = 0.227$

Since $0.227 > 0.05$, do not reject H_0: There is no difference in the thickness of the cornea.

7 H_0: The population median yields are the same

H_1: The population median yield for the new fertiliser is greater

Test statistic = 8,

$P(X \geqslant 8 | X \sim \text{Bin}(10, 0.5)) = 0.0547$

Do not reject H_0: The population median yields are the same.

Exercise 11D

1 **a** $P = 15, N = 6; T = 6$

b $P = 14, N = 31; T = 14$

c $P = 15, N = 5; T = 5$

2 **a** Critical value = 3; do not reject H_0.

b Critical value = 25; reject H_0.

c Critical value = 3; do not reject H_0.

3 **a** 85.5

b 527.25

c −1.132

d Critical value = −1.96; do not reject H_0.

4 **a** H_0: the scores are the same, H_1: the scores are different.

$T = \min(16, 20) = 16$, Critical value = 5

Do not reject H_0: There is no difference between the scores of the eight students.

b The data are symmetric.

5 H_0: Phones with old and new processors are the same

H_1: Phones with the new processor are better

$T = \min(9, 46) = 9$, Critical value = 10

Reject H_0: The phones with the new processors are rated higher by customers.

6 H_0: Treatments have no difference (in the population median number of spots)

H_1: There is a difference (in the population median number of spots)

$T = \min(36, 0) = 0$, Critical value = 3

Reject H_0: There is a difference (in the population median number of spots).

7 **a** $E(T) = 232.5$ **b** $\text{Var}(T) = 2363.75$

c H_0: The pairs of identical twins have the same IQ

H_1: The pairs of identical twins have a difference in their IQ

$T = \min(272, 193) = 193$, $ts = -0.81245$, $z_{0.975} = -1.96$

Do not reject H_0: Identical twins have the same IQ.

Exercise 11E

1 **a** **i** 52 (32)

ii 32 (52)

iii 32

b **i** 75 (51)

ii 51 (75)

iii 51

c **i** $R_m = 24$ (46)

ii 46 (24)

iii 24

d **i** 25 or 30

ii 30 or 25

iii 25

2 **a** 29 **b** 45 **c** 21 **d** 16

3 **a** $E(R_m) = 99$, $\text{Var}(R_m) = 198$

b $E(R_m) = 156$, $\text{Var}(R_m) = 338$

c $E(R_m) = 72$, $\text{Var}(R_m) = 108$

d $E(R_m) = 188.5$, $\text{Var}(R_m) = 471.25$

4 **a** $E(R_m) = 186$, $\text{Var}(R_m) = 558$, $z = -1.672$

b Critical value $= -1.645$; reject H_0.

5 H_0: The calorie content of the chicken sausage is the same as the vegetarian one

H_1: The calorie content of the chicken sausage is different to the vegetarian one

$R_m = 74$, $m(n + m + 1) - R_m = 38$, $W = 38$, Critical value $= 38$

Reject H_0: The calorie content of the chicken sausage is different to the vegetarian one.

6 H_0: There is no difference in the population median length on each side of the river

H_1: There is a difference in the population median length on each side of the river

$R_m = 48$, $m(n + m + 1) - R_m = 88$, $W = 48$, Critical value $= 51$

Reject H_0: There is a difference in the lengths of the plants on the two sides of the river.

7 H_0: There is no difference in the time taken to become overripe

H_1: Chilled transport takes longer to become overripe

$R_m = 21$, $m(n + m + 1) - R_m = 39$, $W = 21$, Critical value $= 20$

Do not reject H_0: There is no difference in the time taken to become overripe.

8 H_0: There is no difference in the scores between the two sessions

H_1: There is a difference in the scores between the two sessions

$E(R_m) = 270$, $\text{Var}(R_m) = 900$, $ts = -2.483$, $P(Z \leqslant -2.483) = 0.0065$

Since we have a two-tailed test, $0.0065 < 0.01$ and so we reject H_0: There is a difference in performance between the morning session and the afternoon session.

End-of-chapter review exercise 11

1 **a** **i** Since we are comparing two sets of data, we must use a two-sample test. Since the data is not matched, we must carry out a Wilcoxon rank-sum test. To do this we must assume that the samples are independent. (The data are already continuous.)

ii H_0: there is no difference in the population medians of the two samples

H_1: The population median for returning from the dining hall is greater than the population median going to the dining hall

One-tail 5% Wilcoxon rank-sum test: Test statistic $= 94$, Critical value $= 82$

Since $94 > 82$, do not reject H_0: There is no difference in the population medians of the two samples.

b **i** The data is now matched pairs and so Wilcoxon matched-pairs signed-rank test would be appropriate.

ii H_0: The median difference of the paired values is 0

H_1: The median difference of the paired values is greater than 0

$T = 3$, Critical value for $n = 10$ is 10

Since $3 < 10$ we reject H_0: The median difference is greater than 0 and so the time taken to walk back is greater than the time to walk to the dining room.

2 H_0: There is no difference in the population medians of men's and women's cholesterol levels

H_1: There is a difference in the population medians of men's and women's cholesterol levels

$R_m = 633$, $m(n + m + 1) - R_m = 387$, $W = 387$, $E(W) = 510$, $\text{Var}(W) = 2550$

Test statistic $= -2.246$, Critical value $= -1.96$ (two-tailed 5%)

Since $-1.96 > -2.246$ reject H_0: There is a difference in the population medians of men's and women's cholesterol levels.

When drawing a stem and leaf diagram, we can see that the data are symmetric and so the assumption is justified.

3 **a** **i** The data cannot be assumed to be normal.

ii The data cannot be assumed to be from a symmetric distribution.

b H_0: The population median is 50, H_1: The population median is greater than 50

Test statistic: 11(+) or 3(−), $P(X \geqslant 11) = 0.0287$

Since $0.0287 < 0.05$, do not reject H_0: The population median is 50

12 Probability generating functions

Prerequisite knowledge

1 $E(X) = 3.2$, $Var(X) = 1.92$

2 $E(X) = 2$, $Var(X) = 2$

3 $E(X) = \frac{10}{3}$, $Var(X) = 7.778$

4 $\frac{1 - \left(\frac{1}{3}\right)^n}{2}$

5 $\frac{5}{9} + \frac{20t}{27} + \frac{20t^2}{27} + \frac{160t^3}{243}$

Exercise 12A

1 a $(0.7 + 0.3t)^{20}$

b $(0.75 + 0.25t)^{10}$

c $(0.96 + 0.04t)^{50}$

2 a $e^{4(t-1)}$ b $e^{2.3(t-1)}$ c $e^{12(t-1)}$

3 a $\frac{t}{10 - 9t}$ b $\frac{7t}{10 - 3t}$ c $\frac{2t}{5 - 3t}$

4 $G_X(t) = \frac{t}{16}(4 + 2t + 4t^2 + t^3 + t^4 + 2t^5 + 2t^6)$

5 $G_X(t) = \frac{1}{16}(t^2 + 4t^4 + 6t^6 + 4t^8 + t^{10})$

$G_X(t) = \frac{t^2}{16}(1 + t^2)^4$

6

x	3	7	11
$P(X = x)$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{3}$

7 $X \sim \text{Bin}(7, 0.2)$

8 a $P(X = 0) = \frac{1}{3}$, $P(X = 1) = \frac{1}{3}$, $P(X = 2) = \frac{1}{6}$, $P(X = 3) = \frac{1}{12}$

b $P(X = k) = \frac{1}{3 \times 2^{k-1}}$

9 $P(X = x) = q^{x-1}p$, Proof

10 $k = 9$

Exercise 12B

1 a 3.9 b 15.2 c 3.9 d 3.89

2 a 3 b 7 c 3 d 1

3 $E(X) = 5.3$, $Var(X) = 2.21$

4 $E(X) = \frac{4}{3}$, $Var(X) = \frac{20}{9}$

5 a $G_X(t) = \frac{t}{2 - t}$

b $E(X) = 2$, $Var(X) = 2$

6 $E(X) = \frac{q}{p}$, $Var(X) = \frac{q}{p^2}$

7 $E(X) = \frac{8}{3}$, $Var(X) = \frac{26}{3}$

8 a $G_X(t) = \frac{1}{8}(2t^2 + 3t^a + 3t^b)$

$G'_X(t) = \frac{1}{8}(4t + 3at^{a-1} + 3bt^{b-1})$

$G''_X(t) = \frac{1}{8}(4 + 3a(a - 1)t^{a-2} + 3b(b - 1)t^{b-2})$

b $a = 3, b = 7$

Exercise 12C

1 a $\frac{t}{5}(2 + 3t)$

b $\frac{1}{4}(1 + t + t^3 + t^5)$

c $\frac{t}{20}(2 + 3t)(1 + t + t^3 + t^5)$

2 a $\frac{1}{10t^2}(1 + 2t + 4t^2 + 2t^3 + t^4)$

b $\frac{t^2(1 - t^5)}{5(1 - t)}$

c $\frac{(1 - t^5)}{50(1 - t)}(1 + 2t + 4t^2 + 2t^3 + t^4)$

3 a $G_X(t) = \frac{t}{10}(2 + 5t^2 + 3t^4)$

b $G_Y(t) = \frac{1}{10}(3 + 4t^2 + 3t^4)$

c $G_{X+Y}(t) = \frac{t}{100}(2 + 5t^2 + 3t^4)(3 + 4t^2 + 3t^4)$

d

$x + y$	1	3	5	7	9
P	0.06	0.23	0.35	0.27	0.09

4 a $G_{X+Y} = \frac{6t^2}{(5 - 2t)(5 - 3t)}$

b $\frac{2}{2t-5} - \frac{3}{3t-5}$

c

k	$P(X+Y=k)$
2	$\frac{6}{25}$
3	$\frac{6}{25}$
4	$\frac{114}{625}$

5 **a** $G_{X+Y}(t) = e^{2(t-1)}(0.8+0.2t)^3$

b $E(X) = 2.6$

c $G''_X(t) = \frac{2e^{2t}}{125e^2}(t+4)(2t^2+22t+59)$

d $\text{Var}(X+Y) = 2.48$

6 **a** Proof **b** Proof

7 **a** $\frac{3t^5}{5-2t^5}$

b $\frac{3t^8}{5-2t}$

8 **a** $\frac{e^{4(t-1)}}{t}$

b $t^3e^{4(t^{-2}-1)}$

9 **a** $G_Y(t) = \frac{3t^4}{10-7t^3}$

b Proof

c

y	$P(Y=k)$
1	0.3
4	0.21
7	0.147

Exercise 12D

1 **a** $\frac{t^4}{25}(4+t)^2$

b

y	4	5	6
$P(Y=y)$	$\frac{16}{25}$	$\frac{8}{25}$	$\frac{1}{25}$

2 **a** $\frac{1}{4}\left(\frac{1}{t}+2+t\right)$

b $\frac{1}{64}\left(\frac{1}{t}+2+t\right)^3$

c $\frac{15}{64}$

3 **a** 2

b $e^{\frac{(t-1)(5t-3)}{t}}$

4 **a** $\frac{t}{3}(1+t+t^2)$

b $\frac{t^4}{4}(2+t+t^2)$

c $\frac{t^5}{12}(t^4+2t^2+t+2+t^{-1}+2t^{-3}+t^{-4}+2t^{-6})$

5 **a** $G_A(t) = t(0.7+0.3t^2)^3(0.3+0.7t)^4$

b $G_B(t) = \frac{3(0.7+0.3t)^7}{10t^3-7}$

c $G_C(t) = \frac{3(0.7+0.3t^3)^3(0.3+0.7t^2)^4}{10-7t}$

6 **a** $G_Y(t) = \frac{243t^5}{(10-7t)^5}$ **b** 0.05797

7 **a** $G_Y(t) = \frac{t^4}{10\,000}(2+3t+5t^2)^4$

b $E(Y) = \frac{46}{5}$

c $\text{Var}(Y) = 2.44$

8 **a** $G_Y(t) = \frac{3t^6}{(10-9t^3)(5-4t^2)(10-7t)}$

b $\frac{21}{5000}$

End-of-chapter review exercise 12

1 $G_X(t) = \frac{5t}{36-31t}$, $E(X) = 7.2$, $\text{Var}(X) = 44.64$

2 **a** $k = 1$, $G_X(t) = \frac{et}{e-t}$

b $E(X) = \frac{e^2}{(e-t)^2}$

c $\text{Var}(X) = \frac{e^2(t^2-2t-2e-2et)}{(e-t)^4}$

3 Proof

4 Proof

Cross-topic review exercise 2

1 $s^2 = 3.00865$

$z = 1.384$

$\alpha \leqslant 8.32\%$

2 $H_0: \mu = 65$, $H_1: \mu \neq 65$

Test statistic $= -0.314$

Critical value $= t_{0.95,7} = -1.895$

There is insufficient evidence to suggest that the sprinklers are not activated at $65°C$.

3 **a** Proof

b 0.614

4 **i** [520.0, 529.2] or $524.6 \pm 4.6[1]$

ii $H_0: \mu_b - \mu_a = 0$, $H_1: \mu_b - \mu_a \neq 0$

$s^2 = 12.711$

Test statistic $= 1.52$

Critical value $= 1.64$

There is insufficient evidence to suggest a difference between the two means.

5 **a** $P(X = x) = p(1-p)^{x-1}$

b **i** $G_Y(t) = p^n t^n (1-qt)^{-n}$

ii $E(Y) = \frac{n}{p}$, $Var(Y) = \frac{nq}{p^2}$

6 H_0: No association between test results and school

H_1: An association between test results and school

Test statistic $= 3.68$

Critical value $= \chi^2_2(0.95) = 5.99$

Do not reject H_0: This is no association between test results and school.

7 H_0: There is no difference between the population medians

H_1: There is a difference between the population medians

$R_m = 48$ or 88

$W = 48$

Critical value $= 49$

Reject H_0: There is a difference in the population medians, therefore there is a difference in the average height of trees on the two sides of the river.

8 **a** Proof

b Proof

c 1.74

9 H_0: The population median is 147.50

H_1: The population median is greater than 147.50

$P = 49$, $N = 6$

$T = \min(49, 6) = 6$

Critical value $= 10$

Reject H_0: There is sufficient evidence to suggest the median is greater than 147.50.

10 **i** $100 \times {}^4C_2 \times 0.4^2 \times 0.6^2 = 34.56$

ii H_0: A binomial B(4, 0.6) is a good model.

H_1: A binomial B(4, 0.6) is not a good model.

$\chi^2 = 9.22$

$\chi^2_3(0.95) = 7.815$

Reject H_0: Probability of faulty chips is not 0.6.

11 H_0: Population median time taken $= 140$

H_1: Population median time taken > 140

Test statistic: $S^+ = 8\,(S^- = 2)$

$P(S^+ \geqslant 8) = 0.054688$

Do not reject H_0: the time taken to fill in the forms is 140 minutes.

13 Projectiles

Prerequisite knowledge

1 magnitude of deceleration $= 0.36\,m\,s^{-2}$

2 $t = 5.74\,s$ and $v = 37.4\,m\,s^{-1}$

Exercise 13A

1 **a** $45.62\,m\,s^{-1}$

b 88.68 m

c 27.17° or 62.83°

2 $42.75\,m\,s^{-1}$

3 $14.59\,ms^{-1} < u < 16.31\,ms^{-1}$

4 36.9°

5 $10\sqrt{2} < u < 20$

6 $7.75\,m\,s^{-1}$

7 **a** $\sqrt{60}\,m\,s^{-1}$ **b** 2.75 m

8 $6\sqrt{15}\,m$

9 2.62 s

10 **a** 2.32 s **b** $25\,m\,s^{-1}$

Further Probability & Statistics practice exam-style paper

1 H_0: Recall for visual and aural is the same
H_1: Recall from visual is better than aural.
Test statistic: $S^+ = 9 (S^- = 3)$
$P(S^+ \geqslant 9) = 0.072998$
Do not reject H_0:
Objects presented visually are not recalled more accurately than objects presented aurally.

2 **a** H_0: There is no difference between the population medians of the two departments
H_1: There is a difference between the population medians of the two departments
$m = 7, n = 8$
$R_m = 75$
$m(n + m + 1) - R_m = 37$
$W = 37$
Critical value = 38
Sufficient evidence to reject H_0: There is a difference in the population medians of the two departments.

b A normal approximation can be used with
$E(T) = 2525$
$\text{Var}(T) = 21042$
$T \approx N(2525, 21042)$

3 **a** $E(X) = \frac{4}{3}, \text{Var}(X) = \frac{8}{9}$

b $G_Z(t) = \frac{1}{19683}(2 + t)^9$

c $E(Z) = 3, \text{Var}(Z) = 2$

4 H_0: No association between gender and facilities used
H_1: An association between gender and facilities used

$$\sum \left(\frac{(O - E)^2}{E} \right) = 12.7$$

$\chi^2_2(0.95) = 5.991$
Reject H_0:
There is an association between gender and the facilities used at the hotel.

5 **a** 4.297

b $$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{60} y(8 + y) & 0 \leqslant x < 2 \\ \frac{1}{96}(7x + 18) & 2 \leqslant x < 6 \\ \frac{1}{120}(30x - x^2 - 69) & x \geqslant 6 \\ 0 & \text{otherwise} \end{cases}$$

c Median $= \frac{30}{7}$

Further Mechanics practice exam-style paper

1 Speed $18.8\,\text{m s}^{-1}$ and direction angle 22.9° below the horizontal.

2 49.3°

3 **a** $v = 3u - \frac{4}{3m}t^3$ **b** $t = (9mu)^{\frac{1}{3}}$

4 **a** Proof **b** $\sqrt{3ga} < u < \sqrt{\frac{15}{2}ga}$

5 **a** $\ddot{x} = -\frac{50}{3}x, \omega = \sqrt{\frac{50}{3}}$

b 0.612 s

6 **a** $v_Q = \frac{9}{7}u, v_P = -\frac{3}{14}u$, opposite directions

b $\frac{135}{28}mu^2\,\text{J}$ **c** $\frac{45}{7}mu\,\text{N}$

Further Pure Mathematics 2 practice exam-style paper

1 $y = \sec x \ln \sin x + c \sec x$

2 $\frac{\pi}{12}$

3 **a** $\left(\begin{array}{ccc|c} 1 & -2 & 1 & 0 \\ 0 & 3 & 2 & 2 \\ 0 & 0 & a - 4 & b - 2 \end{array}\right)$

b $a \neq 4, b \in R$

c $a = 4, b = 2$

d $a = 4, b \neq 2$

4 13

5 Proof, $\ln(3 + \sqrt{8})$, $\ln(2 + \sqrt{3})$

Glossary

χ^2-test: a test that can be used to look for the association between two sets of categorical data, or to perform a goodness of fit test

ρ, the Greek letter rho: density (pronounced 'row' (spelt rho))

A

Acceptance region: the values of the test statistic for which we do not reject the null hypothesis

Angle of depression: the angle formed by the line of sight and the horizontal plane for an object below the horizontal

Angle of elevation: the angle formed by the line of sight and the horizontal plane for an object above the horizontal

Angular speed: the velocity of a body rotating about a fixed point, measured as the rate of change of the angle turned per unit of time

Arithmetic sequence: a sequence in which each successive term is obtained by adding the same constant value

Asymptote: a line that a curve tends towards

Augmented matrix: a matrix that is formed by combining the columns of two matrices

Auxiliary equation: an algebraic equation of degree n that is based upon an nth degree differential equation

B

Barycentre: the point between two objects, such as planets, where the objects are perfectly balanced with each other

Boundaries: limiting or bounding lines

Boundary conditions: a set of conditions that limit the possible solutions of differential equations

Breaking equilibrium: when the net force on an object is no longer zero

C

Cardioids: a type of polar curve that is heart shaped

Catenary: a naturally occuring shape observed in cases such as telephone cables and rope bridges, modelled by the hyperbolic cosine function

Cayley–Hamilton theorem: states that every square matrix satisfies its own characteristic equation

Centre of mass: the point at which the entire mass of a body may be considered

Characteristic equation: the polynomial of degree n that relates to the eigenvalues of a square matrix

Chi-squared: a family of distributions which are used to test association and goodness of fit

Circular functions: sine, cosine and tangent are called circular functions as they are derived from the unit circle

Circular motion: motion that occurs about a fixed point where the distance is constant

Coalesce: when two objects join together upon collision; occurs when there is zero elasticity between the objects

Coefficient of restitution: the measure of how elastic the collision is between two objects

Column: a vertical collection of terms, such as in a matrix

Common perpendicular: when two or more lines or planes are such that a vector is at right angles to both of them

Comparison test: the use of a similar sum to compare against the original; this comparative sum is known to converge or diverge

Complementary function (CF): the general solution of the auxiliary equation of a linear differential equation

Composite: made up of several different parts or elements

Compressed: reduced in size due to external forces

Confidence interval: an interval for which there is a given probability that the population mean lies within that interval

Conic section: a special curve created by cutting through a right circular cone with a plane

Conical pendulum: a pendulum that performs horizontal circles about a centre that is vertically below where the string is attached

Conservation of energy: the total energy of an isolated system remains constant throughout the motion

Constraint: a rule or condition

Contingency table: a two-way table to display categorical data used in a chi-squared test

Convergent: a series is convergent if the sequence of its partial sums approaches a limit

Convolution theorem: in statistics, a theorem that allows evaluation of the probability generating function of the sum of two independent discrete random variables

Cross product: two vectors, $\mathbf{u}$ and $\mathbf{v}$, are crossed to form a vector that is perpendicular to both $\mathbf{u}$ and $\mathbf{v}$

Cube roots of unity: the three roots of the cubic equation $z^3 - 1 = 0$

Cubic equation: a polynomial with a leading term of power 3

Cumulative distribution function: a function that relates probability to the area under the graph for a probability density function that defines a continuous random variable

Cusp: a point on a curve at which two branches meet such that their tangents are equal

D

Deformation: the altering of the shape of an object

Degrees of freedom: the number of independent pieces of information that contribute to the estimate of a parameter

Denominator: the bottom portion of a fraction

Derivative: a function or value obtained from differentiating the original function

Determinant: a value obtained from the elements of a square matrix, usually used to represent the scaling factor from a transformation

Diagonalisable: a square matrix is known to be diagonalisable if it is similar to a diagonal matrix

Differential equation: an equation that contains the original function and at least the first derivative; the order of the differential equation is determined by the highest derivative in the equation

Differentiation: the process of finding the gradient function

Directly proportional: a relationship between two variables such that they increase in the same ratio

Discontinuity: a point on a curve in which $f(a)$ does not exist; a gap exists in the curve

Discrete uniform distribution: a distribution in which the random variable takes specific values, and each value has an equal probability

Discriminant: a function obtained from the coefficients of a polynomial, allowing the deduction of the number of roots of the polynomial in question

Displacement: the position of an object relative to its starting point, measured as a vector

E

Eigenvalue: a value obtained from solving the characteristic equation of a square matrix

Eigenvector: a vector that maps to a factor of itself when a matrix is applied to it, the direction being unchanged

Elastic: a material that has the ability to stretch beyond its natural length when a force is applied to it

Elastic potential energy (EPE): the energy stored in an elastic body that has been stretched or compressed

Element: a value in a matrix

Ellipse: a curve surrounding two focal points, where the sum of the distances of a point on the curve to these two focal points is always constant

Energy: the measure of mechanical energy stored in a system, comprising kinetic energy, potential energy and elastic potential energy

Enlargement: a transformation that increases or decreases the area or volume of an existing shape, a stretch along all coordinate axes

Equilibrium: a state in which the resultant forces on an object are zero

Extension: the extra length created when an elastic object is stretched beyond its natural length

F

Free variable: a variable that does not correspond to a pivot column in a row reduced matrix

Frustum: a right circular cone with a smaller right circular cone cut off by slicing the cone to give a larger and smaller circular face

G

General solution (GS): a solution to a differential equation with undetermined constants

H

Homogeneous differential equation: a differential equation that includes terms in only one unknown function, e.g. y, and its derivatives, e.g. $\frac{dy}{dx}, \frac{d^2y}{dx^2}$. It is possible to arrange the terms to give zero on one side of the equation

Hooke's law: a law that relates the extension of a string or spring, or the compression of a spring, to the force applied

Hyperbola: a curve surrounding two focal points, where the difference of the distances of a point on the curve to these two focal points is always constant

Hyperbolic function: hyperbolic functions are derived from the unit hyperbola

Hyperbolic identities: relationships between hyperbolic functions similar to their trigonometric equivalents

I

Implicit: a function or expression that is not expressed directly in terms of independent variables

Impulse: a force applied over a given time interval

Induction: a method of proof in which a base case is shown to be true, then successive steps are shown also to be true, completing the proof

Inertia: the resistance of any physical object to change its current state of motion

Inextensible: a spring or string that cannot be stretched beyond its natural length

Inhomogeneous differential equation: a differential equation that can include terms in two different functions. Moving all the terms in one function, e.g. $\frac{dy}{dx}, \frac{d^2y}{dx^2}$, to one side of the equation leaves a function, e.g. $f(x)$, on the other side instead of zero

Initial conditions: values that are defined or stated when the modelling of an observation is set in motion

Intersection: the point at which two or more objects, or functions, meet

Invariant: a point, or a set of points, that never change their value

Inverse matrix: a square matrix that can be multiplied by the original matrix to produce an identity matrix

Iteration: a repeat of a mathematical procedure applied to the result of a previous iteration

L

Lamina: a 2-dimensional surface with both mass and density

Limiting friction: a maximum value of static friction for which motion is impeded

Line of intersection: a line that is common to two or more planes in 3-dimensional space

Linear motion: motion that occurs in a straight line; it can be described with one spatial dimension

Logarithmic form: meaning that the answer should be written in exact form, using logarithms, usually ln

M

Matrix (plural: matrices): a rectangular array that consists of elements that are numbers or expressions, arranged in rows and columns

Model: an equation or system of equations that are used to closely resemble an observed phenomenon

Modulus of elasticity: a value that measures the resistance of an object to being stretched or compressed

Moment: a turning effect produced by a force acting at a distance on an object

Momentum: the quantity of motion of a moving body, measured as a product of its mass and velocity

N

Natural length: the original length of a spring or string before any forces act upon it

Newton's equations of motion: the set of equations that govern motion where the acceleration is constant

Newton's experimental law: a law that relates the velocities of two objects before and after collision

Non-parametric test: a test that does not require knowledge of the underlying distribution

Non-singular matrix: a matrix that has a non-zero determinant and an inverse

***n*th roots of unity:** the n solutions of the complex equation $z^n = 1$

O

Oblique: a type of asymptote that is neither horizontal nor vertical

Oblique collision: a collision in which the line of centres of the two objects is not parallel to both the objects' direction of motion

Order (of a matrix): the order of a matrix is the size of the matrix defined by the number of rows (m) and columns (n) and written $m \times n$

Osborne's rule: a rule that changes trigonometrical identities to hyperbolic identities

P

Parabola: a plane curve formed by intersecting a right circular cone with a plane that is parallel to the generator of the cone

Parabolic trajectory: a trajectory modelled by motion in a 2-dimensional plane for which the acceleration is constant

Particle: a small point mass used to represent a larger object

Particular integral (PI): a function used to convert inhomogeneous equations to homogeneous equations

Particular solution: a solution generated by given initial or boundary conditions

Percentile: a measure indicating the value below which a given percentage of observations in a group of observations fall

Perfectly elastic: a collision between two particles in which no kinetic energy is lost

Piecewise function: a function which is defined by several sub-functions, each applying to a certain interval of the domain

Polar coordinates: a 2- or 3-dimensional system for which the distance from the origin and the angle turned through are ordinates

Polynomial: a function consisting of many terms of a variable, with each term having a different non-negative integer power

Position vector: a vector that measures displacement from a given origin

Primitive: the inverse of a derivative, an indefinite integral

Probability density function: a function that describes the relative likelihood for the random variable to take on a given value

Probability generating function: a function that describes the probability of the discrete variable having a value, but in the form of a polynomial

Projectile: a particle or object that, once thrown, continues to move under its own inertia and the force of gravity

Q

Quadratic: a polynomial with a leading term of power 2

Quartic: a polynomial with a leading term of power 4

R

Rational function: an algebraic fraction in which the numerator and the denominator are polynomials

Reduced row echelon form: a matrix that has only the leading diagonal of elements that are non-zero

Reduction: a way of simplifiying an integral through integration by parts and a recurrence relation

Reflection: a transformation in which all points in the image are equidistant from a mirror line with their original positions

Rigid body: a body that remains in equilbrium in all directions

Root: a solution of an equation

Rotation: a transformation in which a plane figure rotates about a fixed point

Row: a horizontal collection of terms, such as in a matrix

Row echelon form: a matrix that has a lower triangle of zeros and a leading diagonal of non-zero elements

S

Scalar equation of a plane: the standard definition of a plane, written in the form $\mathbf{r} \cdot \mathbf{n} = \mathbf{a} \cdot \mathbf{n}$

Scalar product: the result of projecting the length of one vector parallel to the direction of another vector

Sequence: a set of mathematically ordered values or terms

Series: the sum of the terms of a sequence

Shearing: each point in a shape is displaced by an amount that is proportional to its distance from a fixed parallel invariant line

Singular matrix: a matrix that has a zero determinant, and as a consequence it cannot be inverted

Stretch: a type of transformation in which curve, or shape has either its x or y values changed by a scale factor

T

Top-heavy fraction: a fraction where the numerator's algebraic expression is the same degree or higher than that of the denominator's expression

Turning point: a point on a curve at which the gradient is equal to zero and where the gradient is of a different sign on either side of the turning point

U

Uniform: identical or consistent throughout

Unit hyperbola: a curve with the equation $x^2 - y^2 = 1$

Unit vector: a vector of magnitude 1

V

Vector determinant: *see* Determinant

Vector equation of a line: the vector representation of a line, written in the form $\mathbf{r} = \mathbf{a} + \mathbf{b}t$

Vector equation of a plane: the vector representation of a plane, written in the form $\mathbf{r} = \mathbf{a} + \mathbf{b}s + \mathbf{c}t$

Vector product: the crossing of two vectors to create a common perpendicular vector, also known as the cross product

Z

Zero matrix: a matrix that has all its elements as zeros

Index