14 Testing for validity: Spearman's, hypothesis testing and χ^2 test for independence

In order to discover the characteristics of a population you might look carefully at a sample from it. But how do you use the data you obtain? How can you tell whether or not the data supports a hypothesis you might have about the population? And how can you be sure that your data is reliable and the test you chose valid? This chapter will discuss many different tests you could use and enable you to decide which one should be chosen to test your ideas about the population.

Concepts

- Relationships
- Validity

Microconcepts

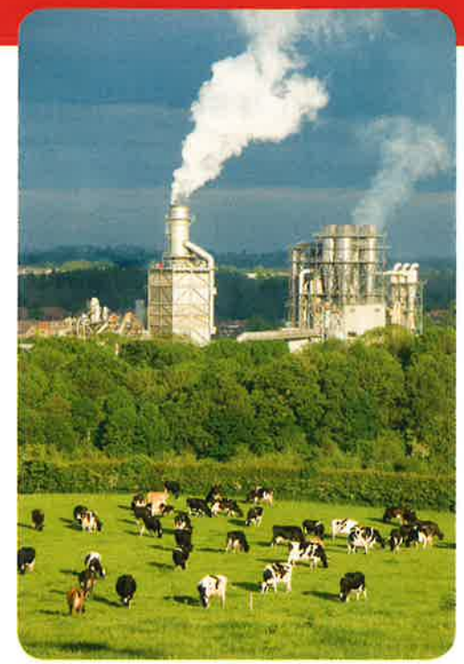
- Contingency tables
- Observed frequencies, expected frequencies
- Null hypothesis, alternative hypothesis
- Significance level
- Degrees of freedom
- Probability values (p-values)
- χ^2 test for independence, goodness of fit
- t-test
- Spearman's rank correlation

How can manufacturers determine whether a new product will be successful or not?

How can a visitor to a casino determine if dice are biased or not?

Are your preferences for food positively or negatively correlated with their nutritional value? Or neither?

How can a teacher (or the IB) tell whether two different versions of a test are equally difficult?



Scientists are concerned with the effect of air pollution on the growth of trees. They measured the heights in metres of 24 young trees of the same species in each of two different forest areas. The data are shown in the table.

It is claimed that the trees in area A are, on average, taller than those in area B.

How could you test this claim?

How should the scientists have ensured that their samples were not biased?

Which statistics can you calculate from the data?

How could you display the data to help test the claim?

Do you think there is enough evidence in these samples to make any claims about the tree heights in general?

Which tests can the scientists use to find out if air pollution has had an effect on the growth of trees in either forest?

How valid will the results of these tests be?

Which test will be the most reliable?

Will scientists be able to use the results of these tests to give feedback on the effect of air pollution on the growth of the trees?

Area A		Area B	
5.30	5.17	5.64	4.73
5.26	4.97	3.90	4.21
3.74	4.87	4.38	5.07
5.55	5.17	4.91	4.82
4.77	5.85	4.87	4.84
6.00	5.48	3.89	5.14
4.44	5.24	4.61	4.95
4.53	4.96	4.88	3.12
4.04	5.61	4.47	4.02
4.73	6.14	5.12	4.12
4.83	5.10	4.46	5.23
5.12	5.75	4.64	5.06

Developing inquiry skills

Write down any similar inquiry questions you might ask to decide whether a statement about other sets of data was true or not. It might be, for example, the heights of children, or the sizes of pebbles on a beach, or the fuel consumption of cars. What would you need to think about in each case?

Think about the questions in this opening problem and answer any you can. As you work through the chapter, you will gain mathematical knowledge and skills that will help you to answer them all.

Before you start

You should know how to:

1 Find the probability of independent events.
eg A fair dice and an unbiased coin are thrown.

Show that the probability of getting a 6 on the dice and a head on the coin are independent events.

$P(6) = \frac{1}{6}$, $P(\text{head}) = \frac{1}{2}$

$P(6 \cap \text{head}) = \frac{1}{36} = \frac{1}{6} \times \frac{1}{2}$

2 Find probabilities from a normal distribution.

Heights are normally distributed with a mean of 156 cm and standard deviation of 7 cm.

Find the probability that a person chosen at random has a height between 152 cm and 161 cm.

The answer is 0.479.

3 Find Pearson's correlation coefficient.
eg The data below shows the position and the number of goals scored for a football league.

Position	Goals
1	63
2	59
3	55
4	48
5	46
6	37
7	35
8	28
9	21
10	13

Find the value of Pearson's correlation coefficient and comment on your answer.
$r = -0.994$. It is strong and negative.

Skills check

Click here for help with this skills check

1 Numbers 1, 2, 3, 4, 5, 6 are written on cards.
S is the event of picking a square number.
E is the event of picking an even number.
Show that E and S are independent

2 The diameter of washers is normally distributed with mean = 35 mm and standard deviation = 3 mm. Find the probability that a washer chosen at random has a diameter less than 36 mm.

3 Find Pearson's correlation coefficient for the following data.

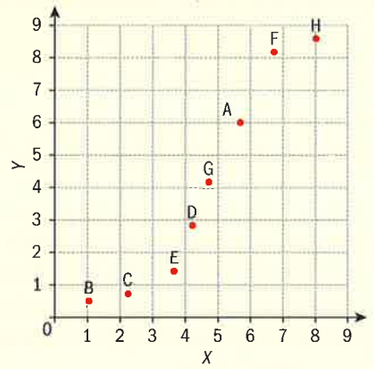
Weight (kg)	Height (cm)
36	128
38	131
39	134
41	138
39	140
42	142
41	145
42	146
54	149

14.1 Spearman's rank correlation coefficient

Investigation 1

Mould is grown in ten different petri dishes with different amounts of nutrients (X), and the area of the dish covered in mould after 48 hours (Y) is recorded. The results are given in the table and also shown on the graph.

X	Y
5.68	6.00
1.04	0.50
2.22	0.76
4.20	2.84
3.66	1.44
6.72	8.20
4.72	4.20
8.00	8.60



1 Calculate the Pearson's product moment correlation coefficient (PMCC) for this data and comment on your results.

Now give each data point a rank, which is the position of the point if the data were listed in order of size for each of the variables. For example, H would be ranked 1 for both X and Y. (It doesn't matter if we rank from largest to smallest, like this, or from smallest to largest; the result will be the same.)

2 **a** Complete the following table showing the ranks for each of the data points.

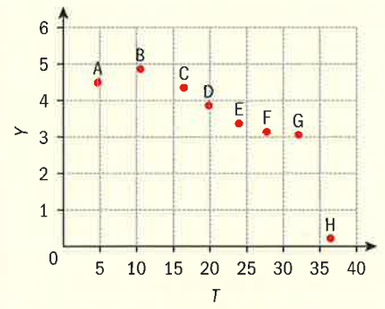
	A	B	C	D	E	F	G	H
X rank								1
Y rank								1

b Calculate the value of PMCC for these ranks.

c Comment on your result, relating it to the particular shape of the graph.

In another experiment the temperature (T) is varied and the area of the petri dish covered after 48 hours (Y) is recorded.

T	Y
4.95	4.50
10.49	4.86
16.40	4.36
19.80	3.86
23.90	3.38
27.70	3.14
32.30	3.06
36.40	0.22



Continued on next page

International-mindedness

In 1956, Australian statistician, Oliver Lancaster made the first convincing case for a link between exposure to sunlight and skin cancer using statistical tools including correlation and regression.

3 **a** Calculate the value of PMCC for this data and comment on your results.

b Complete the following table showing the ranks for each of the data points.

	A	B	C	D	E	F	G	H
T rank								1
Y rank								8

c Calculate the value of PMCC for the ranks and comment on your results.

d Discuss the features of the data that led to this value.

The PMCC of the rank values is called Spearman's rank correlation coefficient.

4 Factual What type of data is used for Spearman's?

5 Factual What type of data is used for Pearson's?

6 Conceptual What do correlation coefficients tell you about the relationship between two variables? When do you use which?

TOK

What is the difference between correlation and causation?

To what extent do these different processes affect the validity of the knowledge obtained?

The product moment correlation coefficient of the ranks of a set of data is called Spearman's rank correlation coefficient. The notation used in IB is r_s.

Spearman's correlation coefficient shows the extent to which one variable increases or decreases as the other variable increases. Such behaviour is described as "monotonic".

A value of 1 means the set of data is strictly increasing and a value of -1 means it is strictly decreasing.

A value of 0 means the data shows no **monotonic** behaviour.

Example 1

Find Spearman's rank correlation coefficient for the following sets of data.

a

x	23	34	17	23	29	45
y	12	10	14	11	11	8

b

x	1	2	3	4	5
y	6	7	8	8	16

a The ranks are

x	4.5	2	6	4.5	3	1
y	2	5	1	3.5	3.5	6

$r_s = -0.956$

1 When more than one piece of data have the same value the rank given to each is the average of the ranks. For example, the two values of x equalling 23 would have ranks 4 and 5; hence each is given a rank of $\frac{4+5}{2} = 4.5$.

The ranked data is put into a GDC and the PMCC obtained.

b The ranks are

x	5	4	3	2	1
y	5	4	2.5	2.5	1

$r_s = 0.975$

Often, when one of the variables increases at a fixed rate, for example measurements taken at one-minute intervals, the order of the ranks will be the reverse of the order of the data.

International-mindedness

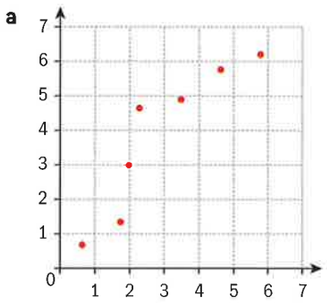
Karl Pearson (1857–1936) was an English lawyer and mathematician. His contributions to statistics include the product–moment correlation coefficient and the χ^2 test.

He founded the world's first university statistics department at the University College of London in 1911.

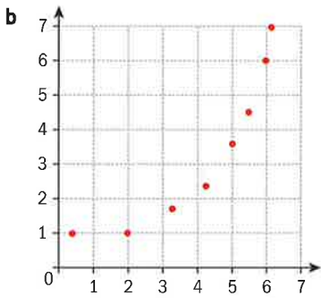
Exercise 14A

1 Write down the value of Spearman's correlation coefficient for each of the sets of data shown. Justify your answers.

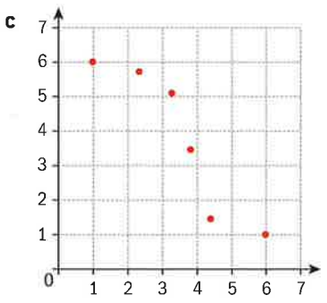
a



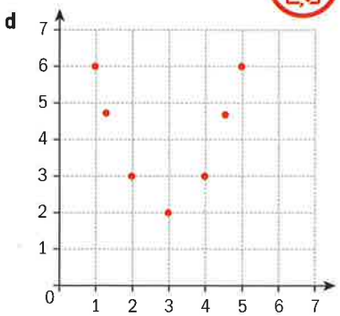
b



c



d



2 Find Spearman's rank correlation coefficient for the following data sets.

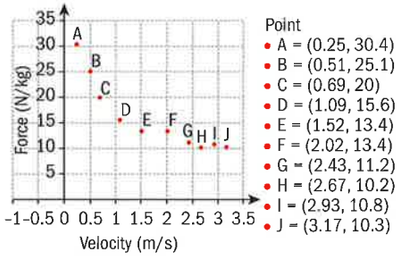
a

x	0	5	10	15	20	25	30
y	23	18	10	9	7	7	7

b

x	10	12	9	6	3	14	8
y	12	11	8	5	7	14	9

3 A sports scientist is testing the relationship between the speed of muscle movement and the force produced. In 10 tests the following data is collected.



a Explain why it might not be appropriate to use the PMCC in this case.

b Calculate Spearman's rank correlation coefficient (r_s) for this data.

c Interpret the value of r_s and comment on its validity.

4 A class took a mathematics test (marked out of 80) and an English test (marked out of 100) and the results are given in the table below.

Maths score	15	25	37	45	60	72	74	78	78	79	79
English score	44	47	42	49	52	44	54	59	69	78	89

a Calculate the PMCC for this data and comment on the result.

b Plot these points on a scatter diagram and comment on your result from part **a**.

c Calculate Spearman's rank correlation for this data and comment on your result.

d State, with a reason, which is the more useful measure of correlation.

5 In a blind tasting, customers are asked to rank ten different brands of coffee in terms of taste.

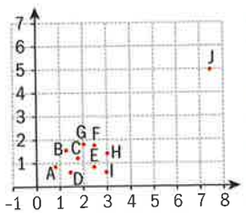
These rankings and the cost of the coffees in cents are given in the table below.

	A	B	C	D	E	F
Taste rank	1	2	3	4	5	6
Cost	450	360	390	320	350	300

a Explain why you cannot use PMCC in this case.

b Find Spearman's rank correlation coefficient for this data and comment on your answer.

6 Consider the following data set.



Point
- A = (0.82, 0.86)
- B = (1.28, 1.56)
- C = (1.78, 1.22)
- D = (1.46, 0.62)
- E = (2.46, 0.84)
- F = (2.48, 1.76)
- G = (2.02, 1.82)
- H = (3.02, 1.42)
- I = (2.98, 0.62)
- J = (7.46, 4.98)

a For this data, calculate the PMCC

i with the outlier J

ii without the outlier J.

b Calculate Spearman's rank correlation coefficient

i with the outlier J

ii without the outlier J.

c Comment on the results.

Reflect How does a scatter graph help you to interpret and compare two data sets?

When is it better to use Pearson's correlation coefficient and when is it better to use Spearman's?

The advantages of Spearman's over the PMCC are:

1 It can be used on data that is not linear.

2 It can be used on data which has been ranked even if the original data is unknown or cannot be quantified.

3 It is not greatly affected by outliers.

EXAM HINT

You may see the formula

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)},$$

where d_i represents the difference in ranks for the i th individual and n denotes the number of individuals.

This gives the correct value for r_s when there are no tied ranks. You will not be asked to use this formula in exams.

14.2 Hypothesis testing for the binomial probability, the Poisson mean and the product moment correlation coefficient

Until now you have been looking at samples of data and working out summary statistics related to the sample.

For example, if a scientist takes 20 plants from a field and measures their heights, he can use this data to find the mean or standard deviation of this particular sample of plants.

What does this tell us about the mean or standard deviation of all the plants in the field?

What might the accuracy of the prediction depend on?

Most of the work in statistics involves collecting a sample from a large population and from it estimating:

1 **parameters**; for example, the mean or correlation coefficient of a whole population

2 the **distribution** of the population; for example, whether or not the population is normally distributed.

Hypothesis testing

In statistics a **hypothesis** is a statement about unknown parameters or distributions.

The aim of a **statistical test** is to try and find data that supports your hypothesis.

Our initial hypothesis is called the **null hypothesis** and is written as H_0.

Every hypothesis test also has an **alternative hypothesis** that will be accepted if we reject H_0 and we write this as H_1.

For example, if we were interested in whether a population mean was 20 cm or more than 20 cm we could write

$H_0: \mu = 20$, $H_1: \mu > 20$

This is called a **one-tailed test** since you are only checking if the mean was more than 20 cm and not whether it could also be less than 20 cm.

If your alternative hypothesis is $H_0: \mu \neq 20$, then you have a **two-tailed test** as you are testing to see whether it is either greater than or less than 20.

Investigation 2

You need to test whether a coin is fair or biased in favour of either heads or tails. You decide to do an experiment in which you toss the coin ten times to see what happens.

1 In hypothesis testing it is important that you can test your null hypothesis mathematically. Suppose the probability of getting a head is p.

   a Explain why you would choose your null hypothesis to be $H_0: p = 0.5$ rather than $H_0: p \neq 0.5$

   b For your chosen null hypothesis, write down the alternative hypothesis.

2 **a** Without doing any calculations, write down a range of values for the number of heads to appear in the 10 tosses that will make you reject the null hypothesis that the coin is fair.

   b Also without doing any calculations, write down a range of values for the number of heads to appear in the 10 tosses for which you would not reject the null hypothesis but still might be suspicious that the coin is not fair. In this case, what might you do to try and be more certain about whether or not to reject the null hypothesis?

3 **a** Use the binomial distribution to work out the probability of the events listed in **2a** happening if the coin is fair.

   b Do you feel that this probability is small enough that you could reject the null hypothesis if one of the events listed occurred?

The events, such as those in **2a**, for which the null hypothesis is rejected are called the **critical region**. Statisticians will not normally reject a null hypothesis unless the probability of an outcome occurring in the critical region is less than 0.05. This means that data would appear by chance in the critical region when H_0 is true less than 5% of the time. This is referred to as a 5% **significance level**.

4 **Conceptual** What is meant by a significance level of a test?

5 **a** Explain why your answer to **3a** is equal to the significance level of the test.

   b Find the significance level for a different choice of critical region.

   c Would you prefer this new one over the one chosen previously? Justify your answer.

   d Were your critical regions symmetrical in each case? Do they have to be?

TOK

If the result of a test is significant, what do we actually know?

6 **a** Can you ever be sure that a coin is biased? If so how, if not why not?

   b Suppose the result of your experiment fell just outside the critical region, does this mean that the null hypothesis is true?

7 **Conceptual** What information do we obtain from a hypothesis test?

If a statistic (a value) obtained from a sample falls in the critical region the null hypothesis is rejected.

Test for a binomial probability

Each statistic obtained from a sample has an associated p-value. The p-value is the probability of obtaining this value (or a more extreme one) if the null hypothesis is true.

If the p-value is less than the significance level the test is significant and the null hypothesis is rejected.

Example 2

A food scientist is trying to determine whether a new version of cheddar cheese is regarded as more tasty than the original type.

In order to do this he decides to carry out a test with 20 people in which they are given the two types of cheese without knowing which is the original and which is new, and he asks them to pick the one they prefer. His null hypothesis is that there is no preference, so each cheese is equally likely to be selected, and his alternative hypothesis is that the new cheese is preferred. He decides to perform the test with a 5% significance level.

Let X be the number of people in the test who prefer the new cheese.

1 If p is the proportion of people in the population who would prefer the new cheese, state the null and alternative hypotheses.

2 Find the critical region for this test.

3 State the lowest possible significance level of the test.

In the test, 18 out of the 20 people preferred the new cheese.

4 State the conclusion of the test.

5 **a** Find $P(X \geq 18)$ under the assumption that the null hypothesis is true.

   b How does this confirm your answer to question **3**?

1 $H_0: p = 0.5$ $H_1: p > 0.5$	The alternative hypothesis is often the thing you would like to be true; in this case, that the new cheese is preferred.

Continued on next page

2 Let X be the number of people in the sample who prefer the new cheese.

The critical region has to be to the right-hand side of the distribution as our alternative hypothesis is that the proportion is larger than 0.5.

Suppose the critical region is $X \geq r$

The boundary for the critical region of this test will be the smallest value of r for which the probability of obtaining at least this number is less than or equal to 0.05.

$P(X \geq r) \leq 0.05$

$r = 15$

$P(X \geq 14) = 0.0577$

$P(X \geq 15) = 0.0207$

The critical region is $X \geq 15$

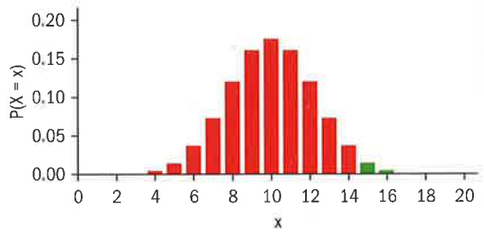
If your GDC only has a cumulative distribution function (cdf) find the acceptance region instead.

$P(X \leq s) > 0.95$

$P(X \leq 13) = 0.942$

$P(X \leq 14) = 0.979$

As the binomial distribution is discrete the critical region will therefore be $14 + 1 = 15$.



3 $P(X \geq 15) = 0.0207$

The probability is calculated using the function for the binomial distribution on the GDC.

The lowest possible significance level = 2.07%

The test could be given as a 2.07% test as the probability of being in the critical region if H_0 is true is equal to 0.0207. Normally though, a figure such as 5% or 1% is chosen.

4 As 18 lies in the critical region we reject the null hypothesis that the cheeses are preferred equally.

It can clearly be seen from the diagram above that 18 lies in the critical region and hence $P(X \geq 18)$ will be less than 0.05.

5 **a** $P(X \geq 18) = 0.000201$

b This probability is less than 0.05 and so 18 lies in the critical region.

This is the ***p*-value** for the test – and should not to be confused with the binomial probability p.

Finding the p-value is often the easiest way to test the significance of the test as the critical region does not need to be calculated.

The null hypothesis is rejected if either the test statistics falls in the critical region (it is beyond the critical value) or if the p-value is less than the significance level.

EXAM HINT

In examinations the test for a binomial probability will always be one-tailed.

Reflect How do you find a p-value for the binomial probability?

How do you find the critical region for the binomial probability?

Exercise 14B

1 Find the p-value for each of the following tests using the binomial distribution, where x is the number of successes. State whether or not H_0 should be rejected.

a $H_0: p = 0.3$, $H_1: p > 0.3$; $n = 12$, $x = 7$ and using a 5% significance level.

b $H_0: p = 0.4$, $H_1: p < 0.4$; $n = 20$, $x = 6$ and using a 10% significance level.

c $H_0: p = 0.7$, $H_1: p > 0.7$; $n = 10$, $x = 9$ and using a 10% significance level.

2 A medicine company claims that its treatment leads to a significant reduction in symptoms within one week for 60% of patients taking the treatment.

In order to test this claim a trial containing a sample of 30 people will take place.

a If p represents the proportion of people who benefit from the treatment, state null and alternative hypotheses for the test.

b Hence find the critical region for a test at the 5% significance level.

c When the test is performed, 14 people report a reduction in symptoms. State the conclusion of the test.

3 A show jumper claims the probability of their horse knocking down a fence (p) is only 0.1. A statistician wishes to test this claim by recording how many fences the horse knocks down in a single round of nine fences.

a Assuming a binomial distribution find the critical region for the test

$H_0: p = 0.1$, $H_1: p > 0.1$

b Give a reason why the binomial distribution might not be appropriate in this case.

4 A psychologist performs a test to see what proportion of 20 objects a subject can remember after viewing them for 30 seconds. The long-term proportion for the number of objects remembered is 0.3.

A student goes on a memory enhancement course and afterwards takes the same test and remembers 8 out of 20. Perform a hypothesis test to see whether the course improved the student's memory.

5 A bus company claims their buses are late less than 10% of the time. Jill notices that in a five-day period the bus is late two times.

Determine whether this is sufficient reason to reject the bus company's claim at the 5% significance level.

Reflect When is it appropriate to test for a binomial probability?

Test for the mean of a Poisson distribution

If a distribution is known to be Poisson, or if it has the characteristics of a Poisson distribution, a hypothesis test can be done to test for a particular value for the mean.

International-mindedness

The Poisson distribution is named after 19th century French mathematician Siméon Denis Poisson whose mentors were Lagrange and Laplace.

Example 3

The number of cars passing a school between 1 pm and 1.30 pm on a weekday can be modelled by a Poisson distribution with a mean of 32. A set of traffic lights is installed at one end of the road and it is hoped this will reduce the number of cars that use the road.

A teacher records the number of cars (X) that pass between 1 pm and 1.30 pm on five days during a school week.

a Find the critical region for a test at the 5% level.

b If the total number of cars is 140, state if there is there evidence at the 5% level that the number of cars has been reduced.

c Find the p-value for a test statistic of 140 cars and use it to verify your conclusion in part **b**.

a The total number of cars has a Po(160) distribution.

$H_0: \mu = 160$, $H_1: \mu < 160$

Let X be the number of cars that pass in the week.

Critical region is $X \leq 138$

The mean is calculated from $5 \times 32 = 160$.

Critical value is the smallest value (r) of X for which $P(X \leq r) < 0.05$.

X	Y_1
134	0.0197
135	0.0241
136	0.0292
137	0.0352
138	0.0421
139	0.0501
140	0.0592
141	0.0696
142	0.0813
143	0.0943
144	0.1089

X = 138

b $140 \geq 138$, so not significant. Therefore, there is insufficient evidence at the 5% level to reject H_0.

c $P(X \leq 140) = 0.0592$

$0.0592 > 0.05$, so not significant. Therefore, there is insufficient evidence at the 5% level to reject H_0.

Calculating the p-value is often the easiest way to do a test as the critical region does not need to be calculated.

Reflect How do you find a p-value for the mean of a Poisson distribution? How do you find the critical region for the mean of a Poisson distribution?

Exercise 14C

1 Let X have distribution Po(λ), and let x be the number of occurrences of the events in the given time interval. By finding the appropriate p-value carry out the following tests.

a $H_0: \lambda = 7.2$, $H_1: \lambda < 7.2$; $x = 5$, and using a 5% significance level

b $H_0: \lambda = 5.9$, $H_1: \lambda > 5.9$; $x = 10$ and using a 5% significance level

c $H_0: \lambda = 12.7$, $H_1: \lambda > 12.7$; $x = 23$ and using a 1% significance level

2 During the previous year the school bus has been late on average 1.8 times per week. A new route has been introduced for the new year and in the first two weeks the bus was late just twice. Determine whether this is sufficient evidence at the 5% significance level that the average has been reduced.

3 In the previous season a sports team conceded on average 2.2 goals per game. Over the summer they hired a specialist defence coach. The manager wants to use the first five games of the season to determine whether there is evidence of an improvement in their defence.

a Assuming a Poisson distribution, state appropriate null and alternative hypotheses for the test.

b Find the critical region for the test with a 5% significance level and state the actual significance level for the test.

c State one reason why the Poisson distribution might not be suitable in this case.

4 A hospital ward is being inspected to see whether or not it has more incidents of infections than the national average. The national average is 0.25 cases per week and the ward is monitored for 16 weeks.

a State a suitable test for whether or not the ward has a greater number of infections than the national average and state the critical region.

b State one condition that needs to be assumed for a Poisson model to be valid.

Reflect When is it appropriate to test for a Poisson mean?

Test for the product moment correlation coefficient

If you have a set of bivariate data, such as hours spent on homework by a student and his or her mid-term grades, then it would be possible to show these on a scatter diagram and find their correlation coefficient (r). This is unlikely to be the same value as the correlation coefficient for the whole population, which is normally written as ρ (pronounced "rho"), but for a large sample size it is likely to be quite similar.

If you need to calculate a line of best fit for your data you need to be sure that there is a linear correlation. Fortunately your GDC has an inbuilt function that allows you to do this by testing the null hypothesis

$H_0: \rho = 0$

against the alternatives $H_1: \rho < 0$, $H_1: \rho > 0$ or $H_1: \rho \neq 0$ depending on whether you are testing to see if there is a negative correlation, a positive correlation or either a negative or positive correlation.

In order to see whether the sample correlation is large enough so that the null hypothesis is rejected, a p-value is found. This is the probability of the sample correlation being at least as large as the one obtained if the two populations are in fact not correlated ($\rho = 0$). If the p-value is less than the significance level (normally 5% or 0.05) then there is strong evidence that the two variables have a linear correlation and as a consequence calculating the least squares regression line is appropriate.

Example 4

The number of times students are late for school and the distance they live from school is thought to be related. A sample of eight students is selected randomly and the data for the previous six weeks is checked. The following results were obtained.

Distance from school (km)	Number of times late	Distance from school (km)	Number of times late
5.2	5	2.3	1
1.4	2	2.8	3
6.7	0	7.0	2
8.8	6	0.5	0

Test at the 5% level whether there is a linear relationship between the two variables.

$H_0: \rho = 0$ $H_1: \rho \neq 0$	In a hypothesis test you must always clearly state both hypotheses. In this question you are not told to favour a positive or a negative correlation, so the alternative hypothesis is $H_1: \rho \neq 0$.
p-value = 0.190	There is no need to show any calculations, just give the p-value. All GDCs should be able to do this test directly.
0.190 > 0.05 hence no reason to reject the null hypothesis that there is not a linear relationship between the distance a student lives from school and the number of times they are late.	Notice that in the conclusion we are not saying that the null hypothesis is true, only that there is insufficient evidence to reject it. Notice the sample correlation coefficient is 0.516. Often this would be considered good evidence for a linear relationship but because the sample size is small it is not a significant result.

To do a hypothesis test for the population correlation coefficient (ρ) perform the following steps.

1. Write down the null hypothesis which is always $H_0: \rho = 0$.
2. Decide on the alternative hypothesis which will depend on whether you are looking for a positive correlation, a negative correlation or any linear relationship.
3. Find and write down the p-value.
4. Compare the p-value with the significance level of the test and write a conclusion.

If the p-value is less than the significance level, the result is **significant** and you "reject the null hypothesis that there is no correlation between the two variables".

If the p-value is greater than the significance level, the result is **not significant** and you would say "there is insufficient evidence to reject the null hypothesis that there is no correlation between the two variables".

You should not calculate a least squares regression line unless there is significant evidence of a linear relationship, though the level of significance you choose may vary depending on the context of the test.

Reflect How do you find a p-value when testing for PMCC = 0?

Exercise 14D

1 For each of the sets of data below:
   - i find the sample correlation coefficient
   - ii perform a test at the 10% significance level for the hypotheses $H_0: \rho = 0$ and $H_1: \rho \neq 0$
   - iii if appropriate, write down the least squares line of regression for y on x.

a

x	y
12	4
15	6
16	5

b

x	y	x	y
7	9	5	7
6	7	9	8
8	8	11	9

c

x	y	x	y	x	y
42	68	34	60	82	75
35	75	70	78	61	79
56	77	54	72	25	47
24	43	38	78	35	51

2 A business is trying to assess demand for its product when it is sold at different prices. It organizes a trial in which the prices in several shops are varied on different days and the number of sales is recorded. The results are shown below.

Price ($)	Sales	Price ($)	Sales
3.5	10	3	12
4	6	2	19
6	2	5	3
8	1	7	2

   - **a** Find the sample correlation coefficient.
   - **b** i Perform a test to see if there is a negative correlation between price and sales.
     - ii If the result of the test is significant, find the least squares regression line that would enable the business to estimate sales for a given price.
     - iii Use the line to find an estimate for the number of sales when the price is $5.50.
   - **c** Plot the original data on a graph and comment on the validity of the results obtained in parts **a** and **b**.

3 The total value of sales of ice cream from a shop and the maximum temperature were recorded over 10 days and the results are shown below.

Sales (€)	Temperature (°C)	Sales (€)	Temperature (°C)
156	23.2	191	27.3
175	25.6	182	26.6
178	28.4	187	24.6
201	31.3	162	22.3
207	30.2	158	18.5

a Test the hypothesis that there is a positive correlation between the total sales of ice cream at the shop and the maximum temperature.

b Find the equation of a regression line to estimate the sales of ice cream for a given maximum temperature.

c Hence find an estimate for the total sales when the temperature outside is 29 °C.

d Explain why the shop owner should not expect the regression line to give an accurate measure of sales for a maximum temperature of 35 °C. Justify your answer with reference to the context.

Reflect When is it appropriate to apply linear regression to bivariate data?

14.3 Testing for the mean of a normal distribution

There are many factors that affect the level of cholesterol in a person's bloodstream which makes it very difficult to know the probability distribution of cholesterol in a population as a whole.

This section explores how, for example, a research centre might test a new drug to see if it significantly reduces cholesterol, even without knowing the associated probability distribution.

The z-test

The normal distribution is the most common and important of all the distributions and has applications in many different disciplines. One of the main reasons for this is the **central limit theorem** which you met in chapter 13. An implication of the central limit theorem is that you do not need to know the distribution of the population you are sampling from in order to test for the population mean, so long as your sample size (n) is large enough. A figure of $n > 30$ is usually taken as sufficient, though this does depend on the distribution being sampled.

In the previous chapter the following result was demonstrated and is one of the most important results in Statistics.

If X has a mean of μ and a standard deviation of σ and if X is normally distributed, or the sample size (n) is large enough for the central limit theorem to apply, then $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$

International-mindedness

The standard version of the central limit theorem (CLT), was developed by the French mathematician Pierre-Simon Laplace in 1810 when he released and proved a generalization of his central limit theorem.

Investigation 3

A town council is trying to plan their budget for future years. To do this they need to know whether or not the average age of people in their town is older than the national average which is 40.0 years, with a standard deviation of 20.2 years.

They decide to test this belief by performing a hypothesis test and record the ages of a random sample of 100 people taken from the town records. They will then work out the sample average $\bar{x}$.

1 Explain why the ages of people in a large population are unlikely to follow a normal distribution.

2 Explain why the distribution of $\bar{X}$ can be assumed to be normal.

3 State null and alternative hypotheses for the council's test.

Because $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$ we would expect $\bar{X}$ on average to be quite close to μ.

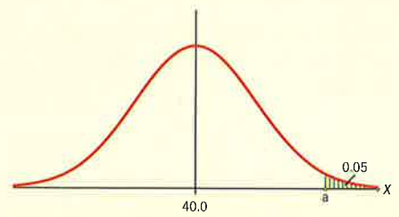
4 Explain why you would expect $\bar{X}$ to be particularly close to μ for larger sample sizes.

As a consequence of its distribution we would choose a test in which $\mathrm{H}_0: \mu = 40.0$, which would be accepted if $\bar{X}$ is close to 40.0 and rejected if it is far above it.

As before the test will have a critical region. If the value of $\bar{X}$ falls within this region then H_0 is rejected.

The town council decide to perform their test at the 5% significance level.

Let a be the critical value on the boundary of the critical region, so that $\mathrm{P}(\bar{X} > a) = 0.05$.



5 **a** Write down the distribution of $\bar{X}$ under the null hypothesis.

b Hence show that the critical region for the test is $\mathrm{P}(\bar{X} > 43.32)$.

The value of the sample mean calculated by the town council was $\bar{x} = 42.3$.

6 State the conclusion of the town council's test.

7 Check your conclusion by calculating $\mathrm{P}(\bar{X} > 42.3)$ when H_0 is true.

Explain why this is the p-value for the test.

8 **Factual** How do you find the critical value of a one-tailed test?

9 **Conceptual** **a** What conditions are necessary to be able to use the normal distribution to test for a population mean?

b If the conditions in part **a** apply, which distribution is used to find the critical region for a test for a population mean?

TOK

When is the normal distribution a valid model?

The z-test uses the sample mean to test for the population mean. It can be used whenever the population standard deviation is known and when either the population is normally distributed, or the sample size is large enough for the CLT to apply.

Example 5

A machine fills bags of flour with a labelled weight of 1 kg. To make sure the bags are being filled correctly a sample of 40 is taken and their weights measured. The sample mean is found to be 995 g. From past experience it is known that the standard deviation of the bags filled by the machine is 20 g.

a Use the p-value to test whether there is sufficient evidence at the 5% level that the machine is filling the bags to less than the correct weight.

b Find the critical region for the test.

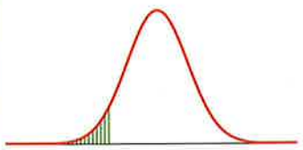
a H_0: $\mu = 1000$, H_1: $\mu < 1000$

p-value will be $P(\bar{X} < 995) = 0.0569$

μ is standard IB notation for the population mean, so does not need to be defined on each occasion it is used.

The question does not tell us that the distribution of weights is normally distributed but, because the sample size is greater than 30, we can assume it is by the central limit theorem. This means we can use the z-test.

The p-value is the probability of being further from than the test statistic.



$0.0569 > 0.05$, not significant so insufficient evidence to reject H_0 that the bags are being filled to the correct average weight.

b $\bar{X} \sim N\left(1000, \frac{20^2}{40}\right)$

Let a be the boundary of the critical region (the critical value).

$P(\bar{X} < a) = 0.05$

EXAM HINT

Your GDC will have a function (z-test) that will enable you to obtain the p-value directly.

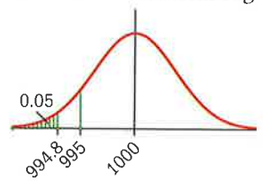
$a = \text{invnorm}\left(0.05, 1000, \frac{20}{\sqrt{40}}\right) = 994.8$

Critical region is $\bar{X} < 994.8$

The notation used here is for illustration only as it is a particular calculator notation and will vary on different calculators.

Note the distribution used for $\bar{X}$ is the one given above.

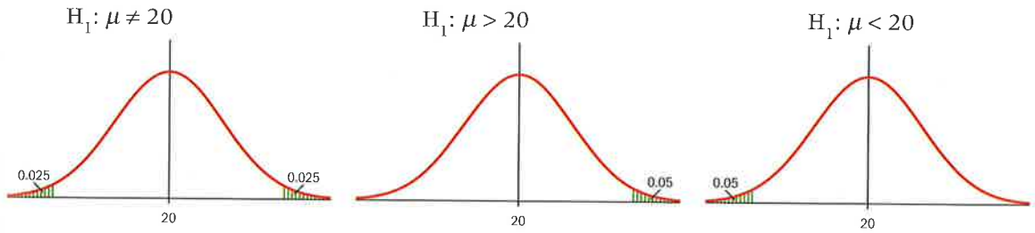
You can confirm your answer to part **a** as 995 does not lie in the critical region.



Two-tailed tests

When performing a two-tailed test (for example, one in which the alternative hypothesis is H_1: $\mu \neq 20$ rather than H_1: $\mu > 20$ or H_1: $\mu < 20$), the probability that the test statistic falls in the critical region is split between the two sides.

For example, for a 5% test we have the following critical regions:



Example 6

The times taken by an athlete to run a circuit near his home can be modelled by a normal distribution with a mean of 15.4 minutes and a standard deviation of 0.62 minutes. The athlete's work takes him away from home for six months and on his return he is interested to see whether his average times have changed. He records his times over the first five days after his return and obtains the following times in minutes:

15.4, 15.5, 14.9, 15.2, 15.1

a Use the p-value to perform a test at the 5% significance level to see if his average time to complete the circuit has changed.

b Find the critical region for the test.

a H_0: $\mu = 15.4$, H_1: $\mu \neq 15.4$

p-value = 0.516

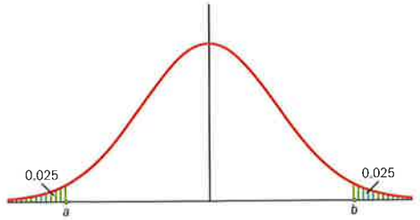
$0.516 > 0.05$, not significant so no reason to reject H_0 that his average time is still 15.4 minutes.

Note that this is a two-tailed test.

As the test is performed on the GDC there is no need to show any method. The hypotheses, p-value and conclusion are all that are required.

Continued on next page

b Let the two critical values be a and b.



The 5% is split between the two critical regions.

$P(\overline{X} < a) = 0.025$

$a = \text{invnorm}\left(0.025, 15.4, \frac{0.62}{\sqrt{5}}\right) = 14.86$

$P(\overline{X} > b) = 0.025$

$P(\overline{X} < b) = 0.975$

$b = \text{invnorm}\left(0.975, 15.4, \frac{0.62}{\sqrt{5}}\right) = 15.94$

Some GDCs will be able to find the value of b without converting it to $\overline{X} < b$.

b could also be found using the value of a as the two values will be symmetrical about the mean value, 15.4.

Critical region is $\overline{X} < 14.86$, $\overline{X} > 15.94$

Reflect What are two-tailed tests?

Exercise 14E

1 The data below is taken from a normal population. Test to see whether it could have come from one with the given mean.

a $H_0: \mu = 124$, $H_1: \mu \neq 124$; $\sigma = 10$. Use a 5% significance level.

122, 134, 138, 128, 129

b $H_0: \mu = 12.2$, $H_1: \mu \neq 12.2$; $\sigma = 1.3$. Use a 5% significance level.

11.0, 9.4, 11.9, 12.1, 10.3

c $H_0: \mu = 0.043$, $H_1: \mu \neq 0.043$; $\sigma = 0.012$. Use a 10% significance level.

0.044, 0.051, 0.040, 0.054, 0.048

2 A pot of paint should cover 24 m^2 of wall. In a test, 32 pots of paint were tested and it was found that they covered on average 23.3 m^2. It is known that the standard deviation of the area covered is 1.8 m^2. The testers wish to determine whether this is sufficient evidence at the 5% significance level that the average coverage of the paint is less than 24 m^2.

a Explain why it is possible to use a z-test in this case.

b State the null and alternative hypotheses.

c Determine the p-value for the test and the conclusion from the test.

d Find the critical region for this test and use it to verify your result in part **c**.

3 It is known that the time between the arrivals of two buses is normally distributed with a standard deviation of 2.1 minutes. The bus company claims that the mean time between buses is 8.3 minutes.

John believes the average time is in fact longer than this so he decides to record some of the times between buses and this data is shown in the table below.

Time (minutes) between buses	8.0	8.7	9.2	8.4	8.5

a State the null and alternative hypotheses for John's test.

b Find the mean of John's sample ($\bar{x}$).

Let X be the times between buses.

c Write down the distribution of $\overline{X}$ under the null hypothesis.

d Hence find $P(\overline{X} > \bar{x})$ assuming the null hypothesis is true.

e Write down the conclusion of John's test.

f Verify your answer to part **d** by performing the test using the inbuilt function on your GDC.

g Find the critical region for the test.

The *t*-test

You will have noticed in the previous questions that though you were testing for an unknown population mean, you were assuming that you knew the population variance.

In most cases this is unrealistic, so the tester needs to estimate the population variance from the sample.

The value used is not the variance of the sample (s_n^2) but the **unbiased estimator of the population variance** (s_{n-1}^2). An unbiased estimator is one that will on average tend towards the value of the parameter being estimated; in this case σ^2.

The two values are linked by the equation $s_{n-1}^2 = \frac{n}{n-1} s_n^2$ which is given in the formula books.

Calculators are likely to use different symbols for s_{n-1} and s_n. Make sure you know which is which.

In examinations if using the inbuilt testing functions you need to be aware whether you have to enter s_{n-1} or s_n. Depending on what is given in the question you may need to use the formula above to convert between the two.

Unfortunately, using the data to estimate the population variance adds an extra degree of uncertainty. The level of uncertainty will depend on the sample size. If the sample is large then not much uncertainty will be introduced as the estimate of the variance will be close to the actual value. If the sample is small there will be more uncertainty.

The extra degree of uncertainty means the distribution of $\overline{X}$ can no longer be regarded as normally distributed but instead follows a t-distribution.

The particular t-distribution you need to use depends on the sample size. For a sample of size n we use the $T(n-1)$ distribution where $n-1$ is referred to as the degrees of freedom (the number of independent observations) and is often written as ν. The term comes from the fact that you know the value of s_{n-1} so if you only knew $n-1$ of the numbers you could work out the final one.

The t-distribution approaches the normal distribution for large values of n.

HINT

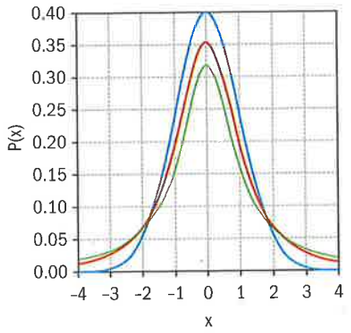
The use of the t-test still requires either the background population to have a normal distribution or the sample size to be large enough for the central limit theorem to apply.

International-mindedness

W Gosset was employed by Guinness in order to improve the taste and quality of their beer. In order to monitor the quality of hops which were used in the brewing of Guinness, he invented the t-test. His pen name was Student. Hence, it is sometimes referred to as Student's t-test.

The graph on the right shows the t-distribution with one degree of freedom (green), two degrees of freedom (red) and the N(0,1) distribution.

The t-test is performed using the GDC in a very similar way to the z-test.



Example 7

a In order to test the hypotheses H_0: $\mu = 8.2$, H_1: $\mu < 8.2$ a sample of 14 is taken and the mean of the sample is found to be 8.15 and the standard deviation 0.07.

Test at the 5% significance level whether the sample is from the population given or one with a smaller mean.

b The sample below is thought to have come from a normal population with a mean of 34.5. Test this belief at a 5% significance level.

34.3	30.2	29.7	34.4	33.6	35.7	34.0	33.9	35.1	34.5

a $s_{n-1} = \sqrt{\frac{n}{n-1}}s_n = \sqrt{\frac{14}{13}} \times 0.07 = 0.0726$

p-value = 0.0115 < 0.05, significant so reject H_0: $\mu = 8.2$

If your GDC requires you to enter s_{n-1} you need to convert using this formula.

Use the t-test (one sample) on your GDC. If it asks for the sample size, input 14. If it asks for the degrees of freedom (ν) input $14 - 1 = 13$.

b H_0: $\mu = 34.5$, H_1: $\mu \neq 34.5$

p-value = 0.161 > 0.05, not significant so insufficient evidence to reject H_0

Because the answers are obtained directly from the GDC there is no need to show your method, just the hypotheses, the p-value and the conclusion.

Reflect When should the t-test be used instead of the z-test?

How do you calculate the degrees of freedom for a t-test?

How do you find the unbiased estimator for the variance if given the sample variance?

Exercise 14F

1 Let $\bar{x}$ be the mean of a sample of size n taken from a population with a normal distribution. Carry out the following tests.

a H_0: $\mu = 6.7$, H_1: $\mu < 6.7$; $\bar{x} = 6.4$, $s_{n-1} = 0.6$, $n = 8$ using a 5% significance level

b H_0: $\mu = 124$, H_1: $\mu \neq 124$; $\bar{x} = 121$, $s_n = 10$, $n = 9$ using a 5% significance level

c H_0: $\mu = 0.85$, H_1: $\mu > 0.85$; $\bar{x} = 0.864$, $s_n = 0.012$, $n = 6$ using a 5% significance level

2 The data below is taken from a normal population. Test to see whether it could have come from a population with the given mean.

a H_0: $\mu = 123$, H_1: $\mu \neq 123$. Use a 5% significance level

122, 134, 138, 128, 129

b H_0: $\mu = 12.2$, H_1: $\mu \neq 12.2$. Use a 5% significance level

11.0, 9.4, 11.9, 12.1, 10.3

c H_0: $\mu = 0.042$, H_1: $\mu \neq 0.042$. Use a 10% significance level

0.044, 0.051, 0.040, 0.054, 0.048

3 The average historic temperature in July is 28.2 °C. In a test the temperature is measured each day in two consecutive years (62 days). The average temperature recorded is 28.5 °C with a standard deviation of 3.2 °C. The researchers wish to determine whether this is sufficient evidence at the 5% significance level to state that the temperature is above the historical average.

a Explain why it is possible to use a t-test in this case.

b State the null and alternative hypotheses.

c Find the p-value for the test and state the conclusion from the test.

4 The time taken to travel between two towns by bus can be modelled as a normal distribution. The bus company claims that the mean journey time between the towns is 83 minutes.

Jackie believes the average time is in fact longer than this, so she decides to record some of her journeys on the bus and the data is shown below.

Journey time (minutes) between towns	82	87	92	84	85

a State the null and alternative hypotheses for Jackie's test.

b For this sample find

i the mean **ii** the standard deviation

iii the unbiased estimator of the population standard deviation.

c Find the p-value and write down the conclusion of the test.

Confidence intervals

Often, rather than testing data obtained from a sample against a single value for the population mean, it is more convenient to have a range of values within which the mean of the population is likely to lie.

These intervals always have a confidence level attached to them. For example, a 95% confidence interval means that on 95% of all occasions such a sample was selected the population mean would fall within the calculated boundaries.

When the population from which the sample is taken can be regarded as being normally distributed, confidence intervals for the population

TOK

In the absence of knowing the value of a parameter, will an unbiased estimator always be better than a biased one?

means can easily be worked out using the GDC. As with hypothesis testing, the normal distribution is used when the population variance is known and the $T(n-1)$ distribution is used when it has to be estimated from the data.

The GDC will always give a confidence interval that is centred on the sample mean.

International-mindedness

Tolerances in design and manufacturing are based on confidence intervals. For example, engineering tolerance is the permissible limit on dimensions such as an axle for a car might be 2 mm ± 1 mm.

Example 8

The sample below is taken from a population which can be modelled by a normal distribution. Find a 95% confidence interval for the population mean.

34.3	29.5	38.1	27.5	29.2	37.0

The 95% confidence interval is (27.91, 37.29).	This can be obtained directly from the calculator using the t-distribution. The interval can be stated in words or given using interval notation as has been done here.

Reflect What is a confidence interval?

Exercise 14G

1 Find the 95% confidence interval for the population mean given the samples below. The population can be assumed to be normally distributed.

a $\bar{x} = 12.4,\ s_{n-1} = 3.2,\ n = 12$

b $\bar{x} = 62.3,\ s^2_{n-1} = 4.2,\ n = 120$

c $\bar{x} = 6.3,\ s^2_n = 5.2,\ n = 10$

d $\bar{x} = 2.3,\ s_n = 0.2,\ n = 7$

e $\bar{x} = 4.6,\ \sigma = 0.2,\ n = 9$ (where σ is the population standard deviation)

2 Find the 99% confidence interval for the population mean based on the following samples.

a 12.4, 13.6, 10.9, 12.5, 11.9

b −2.3, −4.5, 0.2, 5.1, −0.9

3 A sample has a mean $\bar{x} = 21.2$ and $s_{n-1} = 1.4$.

a Find a 99% confidence interval for the population mean when

i $n = 10$ **ii** $n = 20$

iii $n = 50$ **iv** $n = 100$

b Comment on your answers to part **a**.

4 **a** Find the 95% confidence interval for the population mean based on the sample below.

12.2, 14.4, 11.6, 15.1, 13.7

b State one assumption that you made in your calculation.

c Comment on a claim that the population mean is 15.3.

Reflect What is the effect of sample size on the width of the confidence interval?

Two-sample tests

One of the main uses for the t-test is in comparing two different samples and asking if they could have come from an identical population.

The assumption is always that the distributions are the same and the standard deviations are equal, so the test is just whether or not the populations they come from have the same mean. Before carrying out the test, you have to consider whether $\bar{X}$ can be regarded as normally distributed.

Example 9

Mr Arthur gives his two chemistry groups the same test. He wants to find out if there is any difference between the achievement levels of the two groups.

The results are:

Group 1	54	62	67	43	85	69	73	81	47	92	55	59	68	72
Group 2	73	67	58	46	91	48	82	81	67	74	57	66		

a Write down the null and alternative hypotheses.

b Perform a t-test at the 5% significance level.

c Write down the conclusion to the test.

Let the two population means be μ_1 and μ_2	Notice that the two groups do not need to be exactly the same size.
a $H_0: \mu_1 = \mu_2$ There is no difference between the grades in Group 1 and the grades in Group 2. $H_1: \mu_1 \neq \mu_2$ There is a difference between the grades in Group 1 and the grades in Group 2.	This will be a **two-tailed test** as you want to know if Group 1 is better or worse than Group 2.
b p-value = 0.816	All GDCs will have a two-sample t-test option. If you are given the choice between pooled and not pooled, select pooled. This assumes the variances of the populations the samples are taken from are equal and is the assumption that will be used in examinations.
c $0.816 > 0.05$, not significant so no reason to reject the null hypothesis that there is no significant difference between the two groups.	

Reflect What is the difference between pooled and non-pooled tests and which should you use?

Exercise 14H

1 Petra noticed that one of her apple trees grew in the shade and the other did not. She wanted to find out if apples from the tree in the shade weighed less than those in the sun. She picked nine apples from each tree and weighed them in grams.

Tree in shade	75	82	93	77	85	78	91	83	92
Tree not in shade	74	81	95	79	95	82	93	88	90

Perform a t-test at the 10% significance level to test whether the apples from the tree in the shade weigh less than those in the sun.

2 A pharmaceutical company claims to have invented a new pill to aid weight loss. They claim that people taking these pills will lose more weight than people not taking them. A total of twenty people are weighed and tested. Ten people are given the new pills and the other ten are given a placebo. After two months the people are weighed again and any weight loss, in kg, is noted in the table below.

New remedy	1.2	2.4	1.6	3.5	3.2	4.6	2.5	0.8	1.2	3.9
Placebo	0.6	0	1.0	1.3	2.1	0.7	1.9	2.4	0.3	1.0

Perform a t-test at the 1% significance level to see if those taking the pills are losing more weight on average.

Paired samples

A special type of two-sample test occurs when the two samples are paired in an obvious way. This might be the score of the same people in two different tests, or the time taken by a group of athletes to run two separate courses.

The null hypothesis is the same as in the previous test, namely the two populations have the same mean, but it is rephrased to say the difference in the two means is equal to zero and the test is done on the differences rather than on the two samples separately.

TOK

Can we claim that one product is better on average than another if there is a large overlap between the confidence intervals of the two means?

How can statistical data influence our decision making in this case?

Example 10

Five candidates attended a revision course hoping to improve their chemistry grades. They were tested before the course started and again at the end of the course. The results were as follows.

Candidate	1	2	3	4	5
Score before course	64	43	29	56	61
Score after course	72	60	33	55	62

Determine at the 5% level whether the course improved the candidates' performance in their chemistry tests.

The differences between the scores after the course and before the course are:	It is possible to do the subtraction for each candidate in turn, but most GDCs allow you to subtract two lists. If the data set is larger, then this would be the best method to use.
8 | 17 | 4 | −1 | 1 Let μ_D be the population mean for the difference between the two scores. H_0: $\mu_D = 0$, H_1: $\mu_D > 0$ p-value = 0.0713 > 0.05, not significant so no reason to reject H_0 that the grades have not improved.	The test is performed in the usual way using the differences and a one-sample t-test.

Reflect How is a paired-sample t-test different from a two-sample t-test?
What do you need to do before testing a paired sample?

Exercise 14I

1 An oil company claims to have developed a fuel that will increase the distance travelled for every litre of fuel.

Ten scooters are filled with one litre of normal fuel and driven to see how far they can go. They are then filled with one litre of the new fuel and driven over the same track until the second litre is used up. The distances travelled, in km, are shown in the table below.

Original fuel	36	38	44	42	45	39	48	51	48	43
New fuel	43	39	51	49	53	48	52	46	53	49

Test at the 1% significance level whether the new fuel has increased the distance travelled.

2 A financial company claims to be able to increase any investment by 5% in six months.

To test this claim a reporter invested \$100 six times over a period of a few weeks and checked the values of each when the six months were ended. The results are shown in the table below.

Value at start	100	100	100	100	100	100
Value at end	105	110	106	108	103	99

a State the null and alternative hypotheses.

b Carry out the test to see whether or not the company's claim can be rejected at the 5% level.

3 It is felt that a new drug has a positive effect in reducing cholesterol. A sample of ten patients was taken and the level of cholesterol measured at the start of the treatment and again six weeks later.

The results are shown in the table below.

Total cholesterol before (mmol/l)	8.5	9.2	8.7	9.6	7.9	8.8	8.9	9.5	10.0	8.4
Total cholesterol after (mmol/l)	7.9	7.5	8.5	8.1	6.2	6.8	7.5	8.8	9.4	7.6

a Use an appropriate test, with a 5% significance level, to see if the drug has had a positive effect on reducing the level of cholesterol.

It is now given that during the treatment the patients were also encouraged to eat more healthily. A large control group did not take the drug but were also encouraged to eat more healthily. It was found that their cholesterol on average dropped by 0.7 mmol/l.

b Given this new information, state an appropriate test to see if the drug has a beneficial effect and carry out this test.

c **i** State whether you would advise the drug manufacturers to do further tests.

ii If you were to do another test, comment on how the test could be improved.

14.4 χ^2 test for independence

A **chi-squared (χ^2) test for independence** can be performed to find out if two data sets are independent of each other or not. It can be performed at various significance levels. In the examination it will only be tested at the 1%, 5% or 10% significance level.

Saanvi is a member of a sports club. She has noticed that more males play squash than females, and is interested to find out if there is any relationship between gender and favourite racket game. She sent around a survey to the other members in the club to find out which game they prefer: tennis, badminton or squash. The results of the survey are:

Male	Male	Male	Female	Female	Female
Tennis	Badminton	Badminton	Badminton	Badminton	Tennis
Tennis	Squash	Squash	Badminton	Squash	Badminton
Squash	Squash	Squash	Tennis	Badminton	Squash
Squash	Squash	Badminton	Tennis	Tennis	Badminton
Squash	Tennis	Squash	Squash	Badminton	Badminton
Squash	Tennis	Tennis	Tennis	Badminton	Tennis
Squash	Tennis	Tennis	Badminton	Squash	Squash
Badminton	Tennis	Tennis	Badminton	Tennis	Badminton
Tennis	Squash	Badminton	Tennis	Tennis	Badminton
Badminton	Squash	Squash			
Squash	Badminton	Squash			

Saanvi decides to perform a χ^2 test for independence at the 5% significance level to find out if the preferred game is independent of gender or not. She will need hypotheses for this test.

Her **null hypothesis** is

H_0: Preferred racket game is independent of gender.

And her **alternative hypothesis** is

H_1: Preferred racket game is not independent of gender.

Investigation 4

Complete the following table for Saanvi's data. This is the table of **observed frequencies**, f_o, and is called a contingency table.

Sport	Tennis	Badminton	Squash	Total	
Male	10				
Female					
Total				60	

1 Calculate the probability that a person chosen at random is male.

2 Calculate the probability that a person chosen at random likes tennis best.

3 If these two probabilities are independent, find the probability that a person chosen at random is male and likes tennis best.

4 There are 60 people in total. If the events were independent, find the expected number of males who like tennis best.

Under the null hypothesis that the two attributes are independent of each other, the column and row totals can be used to calculate expected frequencies (f_e) for each of the cells.

5 Complete the table of expected frequencies.

Sport	Tennis	Badminton	Squash	Total
Male				
Female				
Total				

In the table of expected frequencies, the **totals** of the rows and columns are fixed to match the numbers of males and females and players of each sport in the sample. In this example:

Sport	Tennis	Badminton	Squash	Total
Male				33
Female				27
Total	19	20	21	60

Continued on next page

International-mindedness

German mathematician Friedrich Robert Helmert, a colleague of Gauss, wrote of the χ^2 test in 1876 in German texts, but they were not translated into English. In 1900 English statistician Karl Pearson wrote of his own version.

6 Find the smallest number of entries that you need to calculate by multiplying probabilities before you can fill in the rest of the table from the numbers already there.

7 If your table had three rows and three columns, find the smallest number of entries that you would need to calculate by multiplying probabilities.

8 If your table had three rows and four columns, find the smallest number of entries that you would need to calculate by multiplying probabilities.

9 Find the smallest number of entries if the table had n rows and m columns.

This number is called the **degrees of freedom**, often written as ν. This is because you only have a "free" choice for the numbers that go into that many cells. After that, the remaining numbers are fixed by the need to keep the totals the same.

10 Factual What does the number of degrees of freedom represent?

11 Conceptual What do the "expected values" tell us?

EXAM HINT
In examinations, ν will always be greater than 1.

The formula for the degrees of freedom (ν) is:
$\nu = (\text{rows} - 1)(\text{columns} - 1)$

HINT
Expected values must be greater than 5. If there are expected values less than 5 then you will need to combine rows or columns. See Example 11.

Investigation 5

To decide whether two variables are likely to be independent it is necessary to compare the observed values with those expected. If the observed values are a long way from the expected values then you can deduce that the two variables are unlikely to be independent and reject the null hypothesis. But how do you measure how far away they are, and, if you have a measure, how do you decide when the difference is large enough to reject the null hypothesis?

1 Looking back at the results, which categories are furthest from the expected values? Which are closest?

2 Find the sum of the differences between the observed and expected values in the tables above and comment on how suitable this would be as a measure of how far apart they are.

3 Comment on an advantage of squaring the differences before adding them.

4 Comment on a disadvantage of using this sum as a measure of the distance between the observed and expected values.

In order to make sure that differences are in proportion, it would be better to divide each difference squared by the expected value (as long as the expected value is not too small).

This calculation will give you the χ^2 **test statistic**.

The χ^2 test statistic is

$$\chi^2_{\text{calc}} = \sum \frac{(f_o - f_e)^2}{f_e}$$

where f_o are the observed values and f_e are the expected values.

If this number is larger than a **critical value** then reject the null hypothesis. If it is smaller than the critical value then accept the null hypothesis.

Your GDC is likely to give you the p-value. This is often the easiest method to perform the test.

As before, the null hypothesis is rejected if the p-value is less than the significance level for the test.

Investigation 5 (continued)

The null and alternative hypotheses for a χ^2 test for independence are:

H_0: the two variables are independent of each other

H_1: the two variables are not independent of each other.

5 Use the entries in the tables above for the observed and expected frequencies to find the χ^2 test statistic. The calculation begins:

$$\chi^2_{calc} = \frac{(10-10.45)^2}{10.45} + \ldots$$

For a 5% significance level the critical value is chosen so that the probability of the test statistic being greater than this value if the two variables are independent is 0.05.

6 Will the critical value be larger or smaller for a 1% significance level than for a 5% significance level?

The size of the critical value also depends on the number of degrees of freedom, as more numbers are being added to create the test statistic. In examinations you will always be given the critical value if you need to use it.

The critical value for 2 degrees of freedom at the 5% significance level is $\chi^2_{5\%} = 5.991$.

7 Use this value and your test statistic to decide whether or not to accept the null hypothesis.

8 Use the inbuilt function on your GDC for the observed values given above and verify your previous answer for the test statistic.

Also make sure you know how to find the expected values on your GDC as you will need to check them each time to make sure they are all greater than 5.

Your GDC also gives you a p-value. As before this is the probability of obtaining the particular test statistic calculated or a higher value. If the p - value is smaller than the level of significance then you do not accept the null hypothesis.

EXAM HINT
In examinations you will use your GDC to find the value of your test statistic.

Continued on next page

9 Use the p-value that you found on your GDC and the significance level of 5% to reach a conclusion Saanvi's test.

10 **Factual** What are the null and alternative hypotheses for a χ^2 test for independence?

11 **Conceptual** How can you use the result of a χ^2 test to determine if there is a relationship between two variables?

Example 11

A randomly selected group of 80 people were asked what their favourite genre of music was: pop, classical, folk or jazz. The results are in the table below.

	Pop	Classical	Folk	Jazz	Total
Male	18	9	3	8	38
Female	22	6	7	7	42
Total	40	15	10	15	80

A χ^2 test was carried out at the 10% significance level.

a Write down the null and alternative hypotheses.

b Show that the expected value for a female liking pop is 21.

c Find the full table of expected values.

d Combine two columns so that all expected values are greater than 5 and write down the new observed and expected tables.

e Write down the degrees of freedom for the new table.

f Use your GDC to find the χ^2 test statistic and the p-value for this test.

g Determine whether the null hypothesis is accepted or not.

a H_0: Favourite genre of music is independent of gender.

H_1: Favourite genre of music is not independent of gender.

Even if not told to do so, you must always state the null and alternative hypotheses when doing a test.

b $\frac{42}{80} \times \frac{40}{80} \times 80 = 21$

Because the command term is "show that" you cannot just get this value from your GDC.

c

	Pop	Classical	Folk	Jazz
Male	19	7.125	4.75	7.125
Female	21	7.875	5.25	7.875

You should always check the expected values to make sure all of them are greater than 5.

d

Expected	Pop	Classical	Folk and Jazz
Male	19	7.125	11.875
Female	21	7.875	13.125

Observed	Pop	Classical	Folk and Jazz
Male	18	9	11
Female	22	6	14

The Folk column could have been joined with either of the other two adjacent columns.

e $(3-1) \times (2-1) = 2$

Remember that you subtract one from the number of rows and columns, then multiply.

f $\chi^2 = 1.1629...$ and $p = 0.559...$

$0.559 > 0.10$ and so the result is not significant and there is no reason to reject the null hypothesis that favourite genre of music is independent of gender.

Reflect When do you need to combine columns/rows?

Exercise 14J

1 Sixty people were asked what their favourite flavour of chocolate was (milk, dark, white).

The results are shown in the table below.

	Milk	Dark	White	Total
Male	10	17	5	32
Female	8	6	14	28
Total	18	23	19	60

Perform a χ^2 test at the 1% significance level to see if favourite flavour of chocolate and gender are independent.

The critical value for this test is $\chi^2_{1\%} = 9.210$.

2 Nandan wanted to know whether or not the number of hours on social media had an influence on average grades (GPA). He collected the following information:

	Low GPA	Average GPA	High GPA	Total
0–9 hours	4	23	58	85
10–19 hours	23	45	32	100
> 20 hours	43	33	9	85
Total	70	101	99	270

He decided to perform a χ^2 test at the 10% significance level, to find out if there is a connection between GPA and number of hours on social media.

a State the null hypothesis and the alternative hypothesis.

b Show that the expected frequency for 0–9 hours and a high GPA is 31.2.

c Show that the number of degrees of freedom is 4.

d Write down the χ^2 test statistic and the p-value for this data.

The critical value is $\chi^2_{10\%} = 7.779$.

e Comment on your result.

3 Hubert wanted to find out if the number of people walking their dog was related to the time of day. He kept a record during 120 days and the results are in the table below.

	Morning	Afternoon	Evening	Total
0–5 people	8	6	18	32
6–10 people	13	8	23	44
> 10 people	21	7	16	44
Total	42	21	57	120

Test, at the 5% significance level, if there is a connection between time of day and number of people walking their dog.

The critical value is $\chi^2_{5\%} = 9.488$.

4 Samantha wanted to find out if there was a connection between the type of degree that a person had and their annual salary in dollars. She interviewed 120 professionals and her observed results are shown in the table below.

	BA	MA	PhD	Total
< $ 60 000	9	6	3	18
$ 60 000– $ 120 000	11	17	10	38
> $ 120 000	7	13	24	44
Total	27	36	37	100

Test, at the 1% significance level, if there is a connection between type of degree and salary.

a State the null hypothesis and the alternative hypothesis.

b State which cell has an expected value less than 5.

c Combine two **rows** so that all expected values are greater than 5.

d Find the p-value for this data.

e Comment on your result.

Developing inquiry skills

Let the height of the trees in the opening problem be h. Divide the heights into small, medium and large where small is $h \leq 4.5$ m, medium is $4.5 < h \leq 5$ and large $h > 5.0$.

Use these categories to form a contingency table for the heights in areas A and B and test at the 5% significance level whether the height of a tree is independent of the area in which it is growing.

Does the conclusion of the test support the hypothesis that the trees from area A, on average, have a greater height than those from area B. Justify your answer.

TOK

What does it mean if a data set passes one test but fails another?

14.5 χ^2 goodness-of-fit test

The uniform and normal distributions

Investigation 6

Jiang wonders whether the die he was given is fair. He rolls it 300 times. His results are shown in the table.

Number	Frequency
1	35
2	52
3	47
4	71
5	62
6	33

1 Write down the probability of throwing a 1 with a fair die.

2 If you throw a fair die 300 times, how many times would you expect to throw a 1?

3 Write down the expected frequencies for throwing a fair die 300 times.

Since all the expected frequencies are the same, this is known as a **uniform distribution**.

4 **Factual** Is the formula for the χ^2 test suitable to test whether Jiang's results fit this uniform distribution?

The null hypothesis is H_0: Jiang's die is fair.

5 Write down the alternative hypothesis.

6 Given that the critical value at the 5% significance level for this test is 11.07, use the formula for the χ^2 test, $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$, to find out if Jiang's results could be taken from a uniform distribution.

Normally you would solve this using your GDC, which may ask you to enter the degrees of freedom.

7 **Factual** What is the number of degrees of freedom in a χ^2 goodness-of-fit test? (Consider in how many cells you have free choices when completing the expected values table.)

8 Write down the number of degrees of freedom for this test.

9 Using your GDC verify the value for the test statistic found above and write down the associated p-value.

10 What is your conclusion from this test?

11 **Conceptual** What is the purpose of the χ^2 goodness-of-fit test?

These types of test are called "goodness-of-fit" tests as they are measuring how closely the observed data fits with the expected data for a particular distribution. The test for independence using contingency tables is an example of a goodness-of-fit test, but you can test for the goodness-of-fit with any distribution.

In a χ^2 goodness-of-fit test, the degrees of freedom $\nu = (n-1)$.

Example 12

Marius works in a fish shop. One week he measures 250 fish before selling them. His results are in the table below.

Length of fish, x cm	Frequency
$x<12$	5
$12\leq x<15$	22
$15\leq x<18$	71
$18\leq x<21$	88
$21\leq x<24$	52
$24\leq x<27$	10
$27\leq x$	2

Marius is told that the lengths of the fish should be modelled by a normal distribution with a mean of 19 cm and standard deviation of 3 cm, so he decides to perform a χ^2 goodness-of-fit test at the 5% significance level to find out if the fish that he measured could have come from a population with this distribution.

a Write down his null and alternative hypotheses.

b Find the probability that a fish is less than 12 cm long.

c Hence find the expected number of fish whose length is less than 12 cm.

The table below shows the expected values of 250 normally distributed fish with mean of 19 cm and standard deviation of 3 cm for the given intervals.

Length of fish, x cm	Probability	Expected frequency
$x<12$		
$12\leq x<15$	0.0814	20.3
$15\leq x<18$		
$18\leq x<21$		
$21\leq x<24$	0.2047	51.2
$24\leq x<27$	0.04396	10.99
$27\leq x$	0.00383	0.958

d Find the missing values.

e Perform the χ^2 goodness-of-fit test, writing down the degrees of freedom used.

(The critical value for this test is 9.488.)

a Let the length of the fish be X.

H_0: X is normally distributed with a mean of 19 cm and a standard deviation of 3 cm.

H_1: X is not normally distributed with a mean of 19 cm and a standard deviation of 3 cm.

b $P(X < 12) = 0.00981...$

c 2.45

Use the normal cdf function on your GDC to find the probability and then multiply your answer by 250 to get the expected number.

$250 \times 0.00981... = 2.45$

d

$18\leq x<21$	0.3781	94.5
$21\leq x<24$	0.2047	51.2

These values are calculated by first finding the probability and then multiplying by 250.

e Two of the expected frequencies are less than 5 so the rows need to combined.

Length of fish, x cm	Observed frequency	Expected frequency
$x<15$	27	22.75
$15\leq x<18$	71	69.6
$18\leq x<21$	88	94.5
$21\leq x<24$	52	51.2
$24\leq x$	12	11.9

There are 4 degrees of freedom.

$\chi^2 = 1.28$ and the p-value = 0.864

Either: $1.28 < 9.488$,

or $0.864 > 0.05$

Hence not significant so no reason to reject the null hypothesis.

Unlike the test for independence you will need to enter the expected values as well as the observed values.

Either critical values or p-values can be used, though in examinations the critical value will not always be given so the p-value must be used.

Reflect How do you perform a goodness-of-fit test for a uniform or normal distribution?

Exercise 14K

1 Terri buys 10 packets of sweets and counts how many of each colour, yellow, orange, red, purple and green, there are. In total she has 600 sweets.

According to the packaging, the colours should be evenly distributed with 20% of each colour in a bag.

The results for Terri's 10 bags are:

Colour	Frequency
Yellow	104
Orange	132
Red	98
Purple	129
Green	137

a Find the expected frequencies.

b Write down the degrees of freedom.

c Perform a goodness-of-fit test at the 5% significance level to find out if Terri's data fits a uniform distribution. Remember to write down the null and alternative hypotheses and to state your conclusion.

The critical value for this test is 9.488.

2 The last digit on 500 winning lottery tickets is recorded in the table below.

Last number	0	1	2	3	4	5	6	7	8	9
Frequency	44	53	49	61	47	52	39	58	42	45

a Each digit should be equally likely to occur. Write down the table of expected values.

b Perform a goodness-of-fit test at the 10% significance level to find out if the data fits a uniform distribution.

3 The grades for an economics exam for 300 university students are as follows.

Grade, x%	Frequency
$x < 50$	8
$50 \leq x < 60$	72
$60 \leq x < 70$	143
$70 \leq x < 80$	71
$80 \leq x$	6

The guidelines for the exam state that grades should be normally distributed with mean of 65% and standard deviation of 7.5%.

a Complete the expected frequency table for this distribution.

Expected frequency	6.8			68.92	6.8

b Perform a goodness-of-fit test at the 10% significance level to find out if the data could have been taken from this normal distribution.

4 The heights of elephants are normally distributed with mean of 250 cm and standard deviation of 11 cm.

250 elephants are measured and the results shown in the table below.

Height, h cm	$h < 235$	$235 \leq h < 245$	$245 \leq h < 255$	$255 \leq h < 265$	$265 \leq h$
Frequency	10	69	88	63	20

a Complete the expected frequency table.

Expected frequency	21.6				

b Perform a goodness-of-fit test at the 5% significance level to find out if the observed data indicates that the population of elephants from which it was taken follows the given normal distribution.

5 The scores for IQ tests are normally distributed with mean of 100 and standard deviation of 10.

Cinzia gives an IQ test to all 200 IBDP students in the school. Her results are in the table below.

Score, x	Frequency
$60 \leq x < 90$	18
$90 \leq x < 100$	39
$100 \leq x < 110$	78
$110 \leq x < 120$	46
$120 \leq x < 130$	10
$130 \leq x < 140$	9

Cinzia wants to test if these results are taken from a population with the same distribution and performs a χ^2 goodness-of-fit test at the 1% significance level.

a Write down the null and alternative hypotheses.

b Explain why you should use the intervals $x < 90$ and $x \geq 130$ when working out the expected values for the lower and upper intervals.

c Hence calculate the expected values for each interval for an N(100,10) distribution.

d Combine rows to ensure all expected values are greater than 5 and give the new expected and observed values.

e Write down the number of degrees of freedom.

The critical value is 6.251.

f Find the χ^2 test statistic and the p-value and state the conclusion for the test.

g **i** Find the mean and standard deviation of the observed data.

ii From a consideration of the answers above and the distribution of the observed data, explain why the χ^2 test has not shown that the data is not normally distributed.

iii If you wished to test to see if the data was normally distributed, state how you might adapt the test.

The Poisson and binomial distributions

Example 13

Flaws in a length of material are thought to be modelled by a Poisson distribution with a mean of two flaws per metre.

Fifty 1 m lengths of material are inspected and the number of flaws in each are recorded in the table below.

Number of flaws	0	1	2	3	≥ 4
Frequency	5	10	18	11	6

a If $X \sim \text{Po}(2)$ find $P(X = 0), P(X = 1), P(X = 2), P(X = 3)$ and $P(X \geq 4)$.

b Hence find the expected values if the number of flaws follows a Poisson distribution with a mean of two flaws per metre.

c Write down the null and alternative hypotheses and the degrees of freedom for the test.

d Find the p-value.

e State the conclusion for this test.

a

Number of flaws	0	1	2	3	≥ 4
Probability	0.135	0.271	0.271	0.180	0.143

Most calculators have a function that will allow all these values to be read directly from a table.

To find $P(X \geq 4)$ you can use either the cumulative distribution function on the calculator or subtract the other values from 1.

Continued on next page

b

Number of flaws	0	1	2	3	≥ 4
Expected value	6.77	13.5	13.5	9.02	7.15

These results are obtained from multiplying the probabilities by the number of 1 m lengths.

c H_0: The number of flaws in the material follows a Poisson distribution with a mean of two flaws per metre.

H_1: The number of flaws in the material does not follow a Poisson distribution with a mean of two flaws per metre.

Degrees of freedom = 5 − 1 = 4

As all the expected values are greater than 5 the degrees of freedom is just the number of cells minus one.

d p-value = 0.479

e 0.479 > 0.05

This result is not significant so no reason to reject H_0 that the number of flaws follows a Poisson distribution.

Example 14

In a trial three coins are tossed.

a Find the probability of obtaining: 0 heads, exactly 1 head, exactly 2 heads, exactly 3 heads.

Hagar tosses three coins 200 times and makes a note of the number of heads each time.

Her results are as follows.

Number of heads	Frequency
0	28
1	67
2	83
3	22

She is interested to find out if her coins are fair and so performs a χ^2 goodness-of-fit test at the 5% significance level on her data.

b Use the probabilities for B(3, 0.5) and the fact that Hagar tossed the coins 200 times, to find the expected values for the number of heads.

c Write down the null and alternative hypotheses and the degrees of freedom for the test.

The critical value is 7.815.

d Find the χ^2 value and the p-value.

e State the conclusion for this test.

a

Number of heads	Probability
0	0.125
1	0.375
2	0.375
3	0.125

Most calculators have a function that will allow all these values to be read directly from a table.

b

Number of heads	Expected value
0	25
1	75
2	75
3	25

Multiply the list of probabilities obtained in part **a** by 200.

In binomial questions, take care not to confuse the number of times the experiment is repeated (200) with the number of times the action is repeated in each binomial trial (3).

An equivalent null hypothesis would be H_0: All the coins are fair.

c H_0: The number of heads has a B(3, 0.5) distribution.

H_1: The number of heads does not have a B(3, 0.5) distribution.

The number of degrees of freedom is 3.

Remember the number of degrees of freedom equals the number of cells minus one.

d $\chi^2 = 2.43$, p-value = 0.489. Either: 2.43 < 7.815, or 0.489 > 0.05

e The result is not significant, so there is no reason to reject the null hypothesis.

The conclusion could also be that not all the coins are fair.

Reflect How do you perform a goodness-of-fit test for a binomial or Poisson distribution?

TOK

In practical terms, is saying that a result is significant the same as saying that it is true?

How does language influence our perception?

Exercise 14L

1 Esmerelda rolls two dice 250 times. She records the number of sixes that she rolls.

Number of sixes	0	1	2
Frequency	135	105	10

a i Find the probabilities of scoring 0, 1 or 2 sixes when rolling two fair dice.

ii Hence, calculate the missing entries in the table of expected frequencies when rolling two fair dice.

Number of sixes	0	1	2
Expected frequencies			6.94

b Perform a goodness-of-fit test at the 5% significance level to find out if the data fits a $B\left(2, \frac{1}{6}\right)$ distribution. Remember to write down the null and alternative hypotheses.

The critical value for this test is 5.991.

2 It is thought that the number of goals scored in a sports match can be modelled by a Poisson distribution. To test this theory the number of goals in 50 games was collected and the data is shown in the table below.

Goals	0	1	2	3	4	5	≥6
Frequency	7	10	11	10	6	4	2

It is known that the average number of goals per game over a season is 2.4.

a Taking the mean as 2.4 and assuming a Poisson distribution, calculate the missing entries in the table below.

Goals	0	1	2	3	4	5	≥6
Probability							
Expected frequency							

b Use a hypothesis test at the 10% significance level to test whether the number of goals scored in a game follows a Poisson distribution with a mean of 2.4.

3 Advait sows three seeds in 50 different pots. The packet claims that the probability that a seed will germinate is 0.75 and the germination of a seed is independent of the other seeds in the packet.

The number of seeds that germinated in each pot is shown below.

Number of seeds germinating	0	1	2	3
Frequency	5	10	15	20

a Using the binomial distribution B(3, 0.75), find the expected probabilities of 0, 1, 2 or 3 seeds germinating.

Advait wishes to test the manufacturer's claim by using the χ^2 goodness-of-fit test and a 5% significance level.

b Write down the null and alternative hypotheses.

c Find the table of expected frequencies.

d Write down the degrees of freedom.

e Find the p-value for this test and hence whether the number of seeds germinating is consistent with a B(3, 0.75) distribution.

4 The number of people joining a queue at an airport during the hour before a flight departs is thought to follow a Poisson distribution with a mean of 4.2 in every five-minute interval.

To test this hypothesis, data is collected over a period of five days and the number of people arriving each five minutes is shown in the table below.

Number of people	0	1	2	3	4	5	6	7	≥8
Frequency	5	7	6	7	8	6	4	8	9

a Perform a χ^2 goodness-of-fit test at the 5% significance level to test the following hypotheses.

H_0: The number of people arriving in the queue each five minutes follows a Po(4.2) distribution.

H_1: The number of people arriving in the queue each five minutes does not follow a Po(4.2) distribution.

Show your table of expected values and clearly indicate any columns that have been combined.

b A test can be significant due to the data not following the particular distribution given, or through the wrong choice of parameters.

i Find the mean and the variance of the original sample taken.

ii Hence give two reasons why the significant result was probably due to the distribution not being Poisson rather than the parameter not being incorrect.

5 In an experiment a student has to guess the symbol on a card held up by a researcher. There are four possible symbols to choose from and each student will be tested five times.

a If a student is guessing randomly, write down the distribution of X, the number of cards they guess correctly.

b Hence find the probabilities that X is equal to 0, 1, 2, 3, 4 and 5.

A group of 500 students sit the test and the results are:

Number correct, x	0	1	2	3	4	5
Frequency	104	193	139	49	13	2

c Perform a goodness-of-fit test at the 5% significance level to find out if the data supports the hypothesis that the students are guessing randomly.

Developing inquiry skills

	Section A	Section B
Small $h \leq 4.5$	3	9
Medium $4.5 < h \leq 5.0$	7	9
Large $h > 5.0$	14	6

From previous research it is known that the heights of this species of tree follow a normal distribution with a mean of 4.9 m and a standard deviation of 0.5 m.

Use the data above to test the heights of the trees from each of the areas separately and see if the observed values are consistent with both samples being taken from this distribution.

Is there sufficient evidence to say these two samples were not taken from the given normal population?

What do your results suggest about the likelihood of the trees from area A having a greater height than those from area B?

Estimating parameters

Often in a χ^2 test you are not comparing the data with a fixed distribution but just want to test whether it is normal, binomial or Poisson. In these cases it makes no sense to choose an arbitrary mean, standard deviation or probability (in the case of binomial), but instead use values estimated from the observed data.

Estimating parameters from the data does affect the degrees of freedom used. For example, if you use the same mean as the observed data for your expected values then when all but two cells are filled in, the other two are fixed as there will be only two numbers which will ensure both the totals and the means match up between observed and expected.

> To obtain the number of degrees of freedom, take the number of cells minus one, and then subtract one for each of the parameters estimated.

For example if you have 10 cells and you have estimated the mean from the data, the number of degrees of freedom will be $10 - 1 - 1 = 8$.

Example 15

The lengths of fish caught in a lake are thought to be normally distributed. To test this belief 200 fish were caught and measured and the results are shown in the table below.

Length (x cm)	$0<x<10$	$10<x<15$	$15<x<20$	$20<x<25$	$25<x<30$	$30<x<40$
Number of fish	45	55	38	27	25	10

Using estimates of the mean and standard deviation of the population taken from the sample data, test the hypothesis at the 5% level that the lengths of the fish are normally distributed.

From the sample $\bar{x} = 16.1$ cm, $s_{n-1} = 8.46$ cm H_0: The fish in the lake have an $N(16.1, 8.46^2)$ distribution. H_1: The fish in the lake do not have an $N(16.1, 8.46^2)$ distribution. Expected values are: 41.3, 42.5, 45.9, 35.2, 19.3, 9.6 p-value = 0.0332 $0.0332 < 0.05$, the result is not significant so there is in sufficient evidence to reject the null hypothesis that the lengths of fish in the lake follow a normal distribution.	These values are obtained from the GDC using the mid-interval values. Make sure you use the unbiased estimator, and not the standard deviation of the sample. Expected values are calculated by multiplying the probabilities by 200. Make sure that you use $x < 10$ and $x > 30$ for the upper and lower intervals when calculating expected values.

Reflect How are the degrees of freedom calculated if extra parameters are estimated?

The binomial distribution

For the binomial distribution $B(n, p)$ you will need to estimate the value of p from the observed data. To do this, use the fact that the expected value for a binomial is np and hence p can be estimated using $\bar{x} = np$. Remember n is the number of trials within each of the experiments, not the number of times the experiment is repeated.

Example 16

An archer fires five arrows at a target, aiming for the "bullseye" in the centre. She feels that she has an equal chance of hitting the bullseye with each shot, that each shot is independent of the ones that have gone before and so the binomial distribution is a good model to use.

To test this belief she looks back over her records and notes the number of times she has hit the bullseye in the last 150 sets of five arrows fired. These results are recorded in the table below.

Number of bullseyes	0	1	2	3	4	5
Frequency	5	22	28	45	40	10

Perform a χ^2 goodness-of-fit test to test the following hypotheses.

H_0: The number of bullseyes follows a binomial distribution.

H_1: The number of bullseyes does not follow a binomial distribution.

Let the number of bullseyes be $X \sim B(5, p)$ For the observed data $\bar{x} = 2.82$ $\Rightarrow p = \frac{2.82}{5} = 0.564$ Expected values are 2.36, 15.3, 39.5, 51.2, 33.1, 8.6 Combine the first two columns to get 17.7, 39.5, 51.2, 33.1, 8.6 Degrees of freedom = $5 - 1 - 1 = 3$ p-value = $0.0137 < 0.05$ The result is significant at the 5% significance level so we reject the null hypothesis that the data follows a binomial distribution,	The formula used is $p = \frac{\bar{x}}{n}$. These are obtained from the GDC. As the first column has expected value less than 5, it needs to be combined with the second column. p was estimated from the observed data so the degrees of freedom is the number of cells minus 2.

Exercise 14M

1 It is claimed that the lifespan of light bulbs is normally distributed with a mean lifespan of 1200 hours.

400 light bulbs are tested and the results shown in the table.

Lifespan, h hours	Frequency
$900 \leq h < 1000$	24
$1000 \leq h < 1100$	52
$1100 \leq h < 1200$	92
$1200 \leq h < 1300$	164
$1300 \leq h < 1400$	42
$1400 \leq h < 1500$	26

a Use the data given to estimate the standard deviation of the light bulbs.

b Copy and complete the expected frequency table assuming a normal distribution with a mean of 1200 and the standard deviation calculated in part **a**.

Expected frequency	19.7					19.7

c Write down the degrees of freedom.

d Perform a goodness-of-fit test at the 5% significance level to find out if the data fits a normal distribution.

2 The number of boys in 100 families with three children is shown below.

Number of boys	0	1	2	3
Frequency	16	29	32	17

A statistician wishes to check whether the probability of a boy being born into one of these families is always the same and the gender of each child is independent of all the others, so he decides to test for a binomial distribution.

a Find the mean of the data and hence the probability (p) that a child in the sample is a boy.

b Perform a goodness-of-fit test at the 1% significance level to find out if the data fits a binomial distribution with the probability of a boy equal to p.

3 The number of fish per day caught in a lake by each angler is thought to follow a Poisson distribution. To test this belief the number of fish caught in an hour by 80 anglers is recorded and the results are shown in the table below.

Number of fish	0	1	2	3	4	5	≥6
Frequency	7	10	15	21	14	9	4

a Given that the maximum number of fish caught be any angler was six find the mean of the sample.

b Using this mean as an approximation for the population mean perform a χ^2 goodness-of-fit test to see if the observed values are consistent with the data coming from a Poisson distribution.

4 The weights of grade 9 children are thought to be normally distributed.

The district nurse weighs 200 grade 9 students and her results are in the table below.

Weight, w kg	Frequency
$40 \leq w < 45$	12
$45 \leq w < 50$	59
$50 \leq w < 55$	52
$55 \leq w < 60$	68
$60 \leq w < 65$	9

a Find the mean and standard deviation of the sample.

b Using the calculated values for the mean and standard deviation, perform a χ^2 goodness-of-fit test at the 5% significance level to test whether the data is taken from a normal distribution.

c From a consideration of the observed data, conjecture why the results might not follow a normal distribution.

5 A scratch card has ten covered discs which are scratched off to reveal a prize. The company says the prizes are distributed randomly and independently among the cards. Abi is suspicious of this because she has heard of people who have won lots of prizes, whereas she has won very few. Scratch cards are normally sold in batches of five.

Abi contacts all her friends and family and asks them to record how many prizes are won in each batch of five. The results of the survey are shown below. No one won more than five prizes.

Number of prizes	0	1	2	3	4	5
Frequency	10	13	18	12	6	1

Abi decides to conduct a χ^2 goodness-of-fit test to see if the prizes are randomly distributed. Initially assume that the probability of winning a prize from a single disc is p.

a Write down the distribution for the number of prizes won in a batch of five cards if the prizes are distributed randomly.

b Use the data collected by Abi to find an estimate for p, the probability of winning a prize.

c Use a χ^2 goodness-of-fit test with this value of p to see whether the prizes are likely to be distributed randomly and independently.

6 The number of phone calls a company receives each five minutes between 1 pm and 2 pm is thought to follow a Poisson distribution with a mean of 2.0.

Over a one-week period the number of calls received every 5 minutes are recorded. The results are shown in the table.

Number of calls / 5 minutes	0	1	2	3	4	≥5
Frequency	4	8	21	12	10	5

a Perform a χ^2 goodness-of-fit test to see if the data supports the hypothesis that the number of calls is distributed as Po(2.0).

A statistician commented that a significant result did not show that the distribution of phone calls did not follow a Poisson distribution, only that it did not follow one with a mean of 2.

b Use a more appropriate value for the mean to test whether or not the phone calls coming into the business could follow a Poisson distribution.

You may assume that there were three intervals in which there were five calls to the business, one interval in which there was six calls, and one interval in which there were seven calls.

14.6 Choice, validity and interpretation of tests

The process of conducting a statistical test is far more than just analysing the data using one of the techniques covered in the previous sections. In this section we will look at the considerations you need to make in choosing which test to use, the checks to make sure your test is both valid and reliable and the limits on the interpretations you can make from your data.

TOK

Do you think that people from very different backgrounds are able to follow mathematical arguments, as they possess deductive ability?

Collection of data

In the natural sciences most data collected will have a numerical value. It is important to be aware of any possible errors in the collection methods. The errors might be **systematic** in which the errors will have a non-zero mean and often follow some kind of pattern, or **random** due to natural variation or other unknown factors, which might be expected to have a zero mean.

An example of systematic error might be a weighing machine that always adds on a fixed amount or a fixed percentage.

Though a systematic error might give a good correlation, any line of regression will be affected by it so it will give misleading results.

Selecting the sample

It is important that the sample chosen matches the needs of the test. In Chapter 2 you met simple random, convenience, systematic, quota and stratified sampling methods.

It is important that the right method is used for the question you are trying to answer.

For example, if you want to do a survey about the level of satisfaction with the provision of sport in a school you need to decide in advance whether you want all years represented (and if so in what proportions – quota or stratified).

Collecting data using questionnaires

A multiple-choice or short-answer survey will provide data that is easy to analyse but it needs to be used carefully for a variety of reasons.

- Answers may be restrictive, so not enough information is obtained.
- People might not answer honestly (particularly if the survey is not anonymous).
- The question might be interpreted in different ways.
- The questionnaire needs to be complete, as extra questions cannot be asked later.

Data mining

Data mining occurs when lots of pairs of variables are considered to see if any significant results can be found. If enough variables are compared it is very likely that such results can be found, as significant results will happen by chance 5% of the time for a 5% significance level. There might be a great temptation to publish these results, even though they have little meaning.

Example 17

A questionnaire is compiled asking a group of people a series of questions about five unrelated qualities. The answers to each set of questions are then compared with the answers to each of the other sets using a χ^2 test for independence, to see if there is significant evidence that they are not independent.

a State the number of tests the compiler of the questionnaire will have to perform if they are to compare every possible pair.

b Given that the tests are all at the 5% significance level, find the probability that at least one of them will prove to be significant, even if the qualities referred to in the questions are independent of each other.

a 10	This can be calculated by listing or reasoning that each of the five attributes needs to be paired with four others which makes 20 combinations. Because these count each pairing twice the answer needs to be divided by 2.
b P(at least one significant pairing) $= 1 - P(\text{none})$ $= 1 - 0.95^{10} = 0.401$	The probability of the 10 tests resulting in at least one significant result is about 40% so without further evidence nothing meaningful can be said about having found one.

Data mining is a useful technique to highlight possible connections between variables but if there is no external evidence in support of a connection then the test must be repeated to see if similar results are obtained. If this is not done then the test is not valid.

Reflect What is a systematic error?

What are the advantages and disadvantages of using a questionnaire to collect data?

Why is important to collect appropriate data for a test?

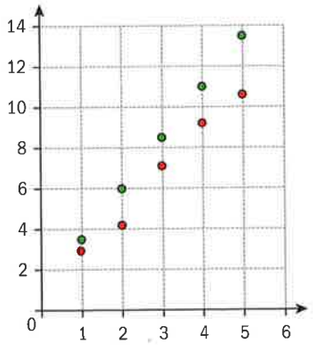
Exercise 14N

1 The two sets of data below have been taken from a distribution in which $y = 2x + 1$. One has been subject to a systematic error and the errors for the other follow a normal distribution with a mean of 0.

x	1	2	3	4	5
y	2.9	4.2	7.1	9.2	10.6

x	1	2	3	4	5
y	3.2	5.4	7.6	9.8	12.0

Both sets of data are shown in the graph below.



a State which data set shows a systematic error and which shows a random error.

b Comment on the results you would obtain if you found a correlation coefficient for each.

c Comment on the usefulness of any line of regression.

2 One hundred people were assessed in ten different attributes. The results for the comparison of two of these attributes are shown in the table below.

	0–10	11–20	21–30	Total
Positive	30	12	28	70
Negative	14	11	5	30
Total	44	23	33	100

a Perform the χ^2 test at the 5% significance level to see if there is evidence for the two attributes not being independent.

The assessor then carries out the test for all different pairings of the 10 attributes.

b Show that there are 45 different pairings for him to test.

c Assuming all the attributes are in fact independent, determine the probability that if all the pairs were tested at least one pair would yield a significant result at the 5% level.

d State what else you would need to know before producing a final conclusion about the result in part **a.**

3 Suppose two schools both teach the IB diploma programme. School A has approximately 60% girls and school B has approximately 40%. Each school has a very different style of teaching to the other.

Both schools agree to a sample of their students taking a standardized test at the start of the course, which will then be compared with their final results to see how much they have improved.

Given it is assumed that girls show more improvement between the test and the final result than boys, state how you would select your sample if you were interested in:

a i finding out which school is better in terms of how much their students improve

ii finding out which of the two teaching methods are better

iii testing whether in fact girls do perform better than boys.

b State a test you could use for each of the cases above, and comment on any assumptions being made.

4 To collect information on how many adults (defined as anyone who is no longer at school) in a particular area live with their parents, two methods of data collection were suggested.

A Use the results from the most recent census (data collected six years ago). In the census all households were visited and it was recorded if any children of the householders were also living there.

B Have pollsters visit 1000 homes chosen randomly from a database to find out the necessary information.

a State one reason why using the census might be a good way to collect this information, and one reason why it might not.

It is decided to use the pollsters to visit 1000 homes.

Assume that whether or not adult children live at home in a particular house is independent of all other houses. Assume also all households answer honestly.

b If the actual proportion of households with adult children living at home is 0.15 find the probability that the survey will obtain a figure for the proportion between

i 0.14 and 0.16 **ii** 0.1 and 0.2.

Comment on your results.

c Write down a possible question or questions that might be used by the pollsters to gain the required information.

5 A statistician wishes to test for differences between males and females so has them complete a survey answering questions in eight different categories. The answers to each of the questions consist of five boxes in which box 5 indicates "strongly agree" and box 1 indicates "strongly disagree".

The statistician carries out a t-test to compare the average of the answers for each of the eight categories with each of the other categories to see if there is a significant difference in the answers for males and for females. He finds that indeed one of the categories does show a significant difference and he publishes his findings.

As an external assessor you are asked to report on the validity of the test.

a State what you would want to know about the sample chosen.

b Calculate a probability to decide if his result is meaningful or not.

Choosing a valid test

A valid test is one that measures the quality it claims to be measuring.

For example, a chemistry test on the periodic table will provide data about how well the students have learned the periodic table.

A test such as this, which covers all the content being tested, is said to have **content** validity.

The assumption of content validity needs to be carefully considered. For example, if you are asking in a survey about how happy a person is, can you be sure that their response is a measure of their happiness or just a measure of how happy people say they are?

A test on quadratic equations certainly has content validity for testing how good a student is at doing quadratic equations.

In addition if previous experience with a test has shown that a high score in the test is a good indicator of success in more advanced algebra then it is said to also have **criterion** validity as a test for the later success. If it relates particularly to future events then it is also referred to as **predictive** validity. An example of a test with predictive validity might be a total points score in the IB diploma and future success at university level.

Reliability of tests

A test or other means of collecting data is regarded as **reliable** if it produces similar results on each occasion it is carried out in similar circumstances. Its key attribute is repeatability.

A reliable test is not necessary a valid one. For example, a test on quadratic equations could be very reliable because students will obtain similar scores each time they take it, but it would be not valid as a test for assessing their ability in a visual arts course.

All valid tests though have to be reliable themselves and use data from a reliable source, so it is important to test for reliability if not sure.

Test–retest

To test that your means of collecting data is reliable the same test or questionnaire can be given to the same group after a period of time. If it is reliable there should be a strong correlation between the results on the two occasions the test is taken.

Of course there might be intervening factors between the two tests and hence even for the most reliable test there is unlikely to be a perfect correlation between the two, but a high value should be a good indication of reliability.

Parallel forms

A large number of questions is split into two parts, each part containing a range of questions designed to measure the required attributes and of as similar standard as possible.

One test is then given to a group of students and, very shortly afterwards, the second one is also given. If the test is reliable then there should be a strong correlation between the scores achieved on each of the tests.

The advantages and disadvantages of each method include the following.

- For parallel testing you have to create a large number of questions of equal difficulty to measure the same quality.
- Proving two parallel tests are equivalent is difficult.
- Test–retest can be affected by intervening factors.

TOK

Given that a set of data may be approximately fitted by a range of curves, where would we seek for knowledge of which equation is the "true" model?

Exercise 140

1 **a** State the necessary requirements on a population for the following tests to be valid.

i a test for the binomial probability, p

ii a test for the mean of a Poisson distribution

iii a t-test.

b If you are not sure they do satisfy the necessary requirements, state the test you could perform to see if the data might come from the required distribution.

2 In a model for the spread of an infection in a small hospital, the average number of people a person will infect per day while he or she is infected is r. The hospital authorities collects data for the total number of people infected each day and compares the additional number infected with the number of people already infected and does this over a period of two weeks.

a The doctor wishes to test for the value of r to see if it is higher than what is normally expected, which might be a sign of bad practice in the hospital. To do this she assumes the number of infections can be modelled by a Poisson distribution with mean r. Comment on this assumption. Justify your answer.

b If it is a reasonable assumption, state a suitable statistical test, and comment on the reliabilty of any results.

3 A survey is done in a school to see if the weekly allowance received by a student is independent of the distance they live from school.

a State two tests which are used to test for independence.

b State the more appropriate in this case.

c Explain how you might set up the test having collected the data.

4 A school is experimenting with a survey to see how students rate the food served in the canteen. Before using the data they decide to test the survey for reliability and opt to do this by the test–retest method, giving the students the same questionnaire to answer four weeks after they first filled it in.

a Explain why a test–retest measure of reliability might be better than a parallel forms measure.

The average score given by eight students in each of the tests is recorded below.

Student	A	B	C	D	E	F	G	H
Test 1	4.2	6.4	5.4	4.0	4.8	5.0	4.8	5.6
Test 2	5.0	6.4	5.8	4.6	5.6	5.6	5.4	5.4

For a sample of size of eight a correlation coefficient above 0.8 will indicate good reliability.

b Find the correlation coefficient and state whether or not the survey is reliable.

c Find the median and the quartiles for the two sets of data and hence draw two box plots.

The canteen says that it has improved over the four weeks between the two tests.

d Comment on whether your answer to part **c** supports this. Justify your answer.

e Carry out a suitable test to decide whether or not this is true at the 5% significance level. Use your answer to part **c** to justify any assumptions you are making.

5 A statistician in a company thinks he has spotted a correlation between an employee's height and their salary. The managing director thinks this is nonsense and asks him to do a test which will hopefully demonstrate the two events are independent.

The statistician feels he can assume both variables are normally distributed and so decides to see whether there is evidence that the correlation between the two factors is positive. He then collects data from ten employees chosen randomly from the workforce and records their heights and salaries.

Height (cm)	145	178	167	155	182	186	191	150	162	177
Salary ($000s)	21	42	27	25	55	37	52	34	36	40

a Carry out an appropriate test and state the conclusion.

The statistician's form began with the following.

Employee no.	Male/female	Height	Salary
12 465	F	145	21
23 412	M	178	42
13 457	F	167	27

b Comment on the validity or otherwise of the test carried out.

c It is felt that a more significant factor affecting salary might be whether an employee is male or female. Assuming more data can be collected, state a test you might do to test this hypothesis.

Type I and type II errors and the importance of the significance level

Different significance levels can be used depending on how certain you want to be about the test results.

Suppose a vital component in a power station is tested to see if it is at a dangerous threshold or not. In this test a significant result would mean it is assumed to be safe when it is in fact dangerous.

Suppose testing takes place once a day.

A 5% significance level would mean that 5% of the time, roughly once every three weeks, a dangerous component would be kept in position. Though the consequences of keeping it in position need to be taken into consideration, this seems unreasonably high for a potentially dangerous situation, and so a lower significance level should be chosen.

A **type I** error is rejecting H_0 when H_0 is true; we can write its probability as $P(C|H_0 \text{ true})$ where C stands for the test statistic being in the critical region.

For the normal or t-distributions this is the same as the significance level, and for discrete distributions it is equal to the probability of rejecting H_0, which might not be exactly the same as the quoted significance level.

A **type II** error is accepting H_0 when it is not true; we can write its probability as $P(C'|H_1 \text{ true})$. The actual value of the probability depends on the particular value of the parameter.

Unfortunately, for a given sample size, reducing the chance of a type I error increases the chance of a type II error and vice versa.

It is therefore very important to consider the balance between the two. In some circumstances, such as the ones mentioned above, it is more important to avoid a type I error than to avoid a type II error. In the case of the power station a type II error would mean replacing a component before it needed to be replaced.

International-mindedness

In physics, during the search for the Higgs boson, the calculated p-value was 5.5×10^{-7}. It was felt necessary to have such a small value as the consequences were so large for our understanding of particle physics.

TOK

When is it more important not to make a type I error and when is it more important not to make a type II error?

Investigation 7

A laboratory test is being performed on a new drug. Previous drugs achieved a rating of 0.856 with a standard deviation of 0.01. The higher the rating, the more effective the drug. It is intended that the new drug will be tested on 40 samples to see if it has a higher rating. The test is initially conducted at a 5% significant level.

Assuming the standard deviation is the same as for the previous drugs:

1 state the null and alternative hypotheses for the test

2 find the critical region for the test and comment on the meaning of your answer in the given context

3 state the probability of a type I error for this test.

4 The probability of a type II error is $P(\bar{X} < a)$ if H_1 is true. State the value of a.

5 Why is it not possible to work out the probability of a type II error without further information?

Given that the population mean for the rating of the new drug is 0.858:

6 find the probability of a type II error and comment on your result.

7 If the significance level of the test was reduced to 1%

- **a** find the new critical region
- **b** hence find the probability of a type II error
- **c** comment on the effect of reducing the probability of a type I error.

8 On the same axes sketch the distribution of $\bar{X} \sim N\left(\mu, \frac{0.01}{\sqrt{40}}\right)$ for $\mu = 0.856$ and $\mu = 0.858$

- **a** when $\mu = 0.856$ and shade the area that represents the probability of a type I error
- **b** when $\mu = 0.858$ and shade the area that represents the probability of a type II error.

9 **a** Investigate the effect of changing the size of the sample on the probability of a type II error for a fixed significance level.

b Explain this effect by consideration of the standard deviation of $\bar{X}$.

10 **Factual** Explain what is meant by a type I and type II error and why is it important to considering both.

11 Due to extra funding the new drug is now tested on 500 samples and the sample mean is found to be 0.8568.

- **a** Find the p-value and state the conclusion of the test at the 5% level.
- **b** Medical testers might refer to a result as being 'statistically significant but not clinically significant'. Explain what this might mean for the result obtained in part **a**.

12 **Conceptual** Why is it important to consider type I and type II errors?

The probability of a type I error is the probability of rejecting H_0 when H_0 is true. For the normal distribution this will be equal to the significance level; for a discrete distribution this will be the probability of the statistic falling in the critical region.

In order to find the probability of a type II error, first find the critical region under the null hypothesis.

The probability of a type II error is the probability of the statistic not being in the critical region. This is calculated using a value for the parameter chosen from the alternative hypothesis.

Example 18

A machine produces components needed for a software company. The probability of a fault occurring in the production of a single component has to be less than 0.02. A sample of size 50 is taken from the output and tested to see if any were faulty. A test was performed with the hypotheses H_0: $p = 0.02$ and H_1: $p > 0.02$ at a 5% significance level.

a State a suitable model for the number of faults in the sample; include any additional assumptions you are making.

b Find

i the critical region for the test

ii the probability of a type I error.

c Earlier testing indicates that the probability of a fault is 0.04. If this is the case find the probability of a type II error.

a Let X be the number of faults in the sample. $X \sim B(50, 0.02)$ For the binomial model to be suitable we need to assume that the faults occur independently of each other.	
b i $P(X \geq a) < 0.05$ $P(X \leq a - 1) > 0.95$ $a - 1 = 3 \Rightarrow a = 4$ The critical region is $X \geq 4$	The critical value is the smallest value of a for which this inequality is satisfied. If your GDC will do right tail probabilities the answer can be obtained directly from the first line. If your GDC can only do cumulative probabilities (left tail) then find the acceptance region instead, which is $P(X \leq a - 1)$.
ii $P(X \geq 4) = 0.0178$ which is equal to the probability of a type I error.	This can be calculated as $1 - P(X \leq 3)$.
c $P(X \leq 3 \mid p = 0.04) = 0.861$ which is equal to the probability of a type II error.	A type II error is the probability of being in the acceptance region (not in the critical region) when H_0 is not true.

Example 19

In order to satisfy quality control the mean number of flaws in aluminium sheets must be less than or equal to 0.6 flaws per metre length. A length of 7 m is inspected.

Assuming the number of flaws follows a Poisson distribution:

a state the distribution of the number of flaws (X) in the length sampled, assuming an average of 0.6 flaws per metre

b state the hypotheses for the test

c find the critical region for the test at the 5% significance level

d find the probability of

i a type I error

ii a type II error, given the mean is in fact 0.72 flaws per metre.

a $X \sim Po(4.2)$	The mean is equal to $0.6 \times 7 = 4.2$.
b H_0: $\mu = 4.2$, H_1: $\mu > 4.2$	
c $P(X \geq a) \leq 0.05$ $P(X \leq a - 1) \geq 0.95 \Rightarrow a - 1 = 8$ $a = 9$ Critical region $X \geq 9$	Some GDCs will be able to produce the critical region without the need to use the cumulative distribution function to find the acceptance region.
d i $P(X \geq 9 \mid \mu = 4.2) = 0.028$	The probability of a type I error is the probability of being in the critical region when H_0 is true.
ii $0.72 \times 7 = 5.04$ $P(X \leq 8 \mid \mu = 5.04) = 0.929$	The probability of a type II error is the probability of falling outside the critical region under the assumptions of H_1.

Exercise 14P

1 For the samples below taken from a normal distribution, find the probability of a type II error, for a test at the 5% significance level.

a H_0: $\mu = 7$, H_1: $\mu > 7$, $n = 30$, $\sigma = 0.4$, true value of $\mu = 7.1$

b H_0: $\mu = 12.1$, H_1: $\mu < 12.1$, $n = 20$, $\sigma = 0.2$, true value of $\mu = 12.0$

2 For the samples below taken from a binomial distribution find the probability of

i a type I error **ii** a type II error

for a test at the 5% significance level.

a H_0: $p = 0.6$, H_1: $p > 0.6$, $n = 30$, true value of $p = 0.68$

b H_0: $p = 0.45$, H_1: $p < 0.45$, $n = 40$, true value of $p = 0.43$

3 A survey is taken to count the number of cars that pass a point each hour. For the hypothesis tests below n is the number of hours used in the survey. Assuming the number of cars can be modelled by a Poisson distribution, find the probability of

i a type I error **ii** a type II error

for a test at the 5% significance level.

a H_0: $\mu = 4.6$, H_1: $\mu < 4.6$, $n = 30$, true value of $\mu = 4.5$

b H_0: $\mu = 2.1$, H_1: $\mu > 2.1$, $n = 20$, true value of $\mu = 2.8$

4 A sample of size 15 is taken from a normal population with $\sigma = 5.3$.

The following hypotheses are tested at the 5% level.

$H_0: \mu = 51$, $H_1: \mu \neq 51$

a Find the critical regions for the test.

b Hence find the probability of a type II error if $\mu = 51.5$.

5 Two types of radioactive substances emit particles at different rates. A emits on average 50 particles per second and B emits on average 54 particles per second. The emission rates of both particles can be considered as being normally distributed with a known variance of 36.

Sophie believes she is testing substance A and records the emission rate in ten different experiments.

a Given $H_0: \mu = 50$ and $H_1: \mu > 50$ find the critical region for her test at the 5% level of significance.

b Find the probability of a type II error if the substance is in fact substance B.

c State one way Sophie could reduce the probability of a type II error.

6 It is thought that in a crowded city with a large population the proportion of people who have a car is 0.3. To test this belief it is decided to take a sample of 50 people and record how many have a car. A 5% significance level is chosen.

a State the distribution of the sample under H_0.

b Find the critical region for rejecting the null hypothesis in favour of the alternative hypothesis that a larger proportion than 0.3 have a car.

c For the above test calculate the probability of a type II error when the population proportion is in fact:

i 0.4 **ii** 0.5

7 The number of people entering the casualty department at a hospital during a weekday evening, 6–10 pm, can be regarded as a Poisson distribution with a mean of 8.2 people per hour.

It is hoped that the introduction of a hospital phone line, which patients can call with concerns, will reduce this number and so a phone line is trialled for a period of several weeks.

During the trial the number of patients arriving over 10 randomly chosen hours during a weekday evening is recorded.

a Explain why it was important to select the 10 hours for the survey randomly, rather than just record the numbers over three consecutive evenings.

b Write down possible null and alternative hypotheses, based on the total number of people who come into the hospital during the 10 hours chosen, to test whether the number of patients arriving has been reduced.

If the mean number of patients arriving per hour in the sample is less than or equal to 7.5 per hour the management will assume the rate of arrivals has fallen and will continue the development of the phone line.

c Describe what would be a type I error for this test and calculate its probability.

d If the rate has gone down to 7.8 patients per hour, find the probability of a type II error and describe what this would mean in the given context.

8 A machine designed to put jam into doughnuts is set to deliver an average of 1.20 cm^3 of jam per doughnut. The machine is checked regularly to ensure that the mean does not deviate from this amount. At each of the checks a sample of 20 is taken and the amount of jam dispensed is measured. Assume that the amount of jam follows a normal distribution with a standard deviation of 0.1 cm^3, and the test is performed at the 5% significance level.

a State the null and alternative hypotheses.

b Write down the probability of a type I error.

Given that the actual amount of jam delivered is 1.17 cm^3:

c find the probability of a type II error.

Investigation 8

The actual meaning of a p-value is often misunderstood. It gives the probability of the data observed (or more extreme data) occurring if the null hypothesis is true, but what the testers often really want to know is the probability that the null hypothesis is true given the data, and that is not immediately accessible.

1 A test is performed with hypotheses $H_0: \mu = 5$ and $H_1: \mu \neq 5$. The p-value was equal to 0.02. The person performing the test claimed the result meant that the probability the mean is 5 is 0.98.

Using H_0 to signify the event the null hypothesis is true, and D to signify the event the data (or more extreme data) is obtained, write down:

a an expression for the probability the researcher thinks he is giving.

b the probability he is actually giving.

2 **a** In addition to the above notation let H_1 signify the alternative hypothesis is true. Draw a tree diagram to show the four possible situations for the test when the results D are obtained. The first two branches should be the events H_0 is true and H_1 is true.

b On your tree diagram indicate the p-value.

c **i** Hence explain why

$$P(H_0 \mid D) = \frac{P(H_0) \times P(D \mid H_0)}{P(H_0) \times P(D \mid H_0) + P(H_0) \times P(D' \mid H_0)}$$

ii Find a similar formula linking $P(H_1|D)$ with $P(D|H_1)$.

3 It is felt that some athletic training camps have begun to use an illegal treatment to improve performance. In standardised tests of athletes before and after undergoing this treatment performance normally increases by 15%. Other indicators (investigative journalism, whistle-blowers etc) lead the athletics organisation to believe that 10% of all camps are using this method.

From previous data is known that the improvement in performance of athletes at a particular camp has been normally distributed with a mean of 10% and a standard deviation of 7.5%.

In a random check to see if the camp is now using the illegal treatment five athletes from the camp have their performance measured before and after the training, and a test is done with the following hypotheses,

H_0: The mean improvement (μ) is 10%, $H_1: \mu > 10\%$.

The test was carried out and it was found that $\bar{x} = 15.52$.

a Assuming the standard deviation has not changed, verify the p-value for this result is 0.0499.

ii Explain why this does not mean the probability the camp was using the illegal treatment is approximately 0.95.

ii Without doing further calculations would you intuitively rate a value of 15.52% improvement as strong evidence for the athletes in the camp undergoing the illegal treatment?

b Find the probability of obtaining this result (or a more extreme one) if the athletes had been undergoing the illegal treatment ($P(D|H_1)$).

You may assume the standard deviation is unchanged.

c **i** Use your values from **3a** and **3b**, plus the information given in the initial stem to add all the probabilities to the branches of a copy of the tree diagram from question **2a**.

ii Hence find the probability of obtaining this test result (or a more extreme one), using all the information available ($P(D)$).

iii Hence use the formula from c ii to find the probability the camp had been using the illegal treatment.

iv What conclusion might you draw from the test?

4 Explain why the p-value is not the same as the probability the null hypothesis is true.

5 **Conceptual** What does a p-value actually tell you? What is a common misconception?

Exercise 14Q

1 Bruno claims he has extra sensory perception (ESP). A test is done in which three cards have one of five possible shapes drawn on them and Bruno has to guess the shape on the cards. If he possesses some form of ESP it is expected he will have a better than average chance of guessing the cards.

When the test is performed Bruno guesses the shapes on all three cards correctly.

a Explain why the p-value for this test is equal to 0.008.

b Bruno claims that this means the probability he has ESP is 0.992.

Explain why he is wrong.

c Explain what else you would need to consider if you were to try to find the probability of the existence of ESP.

Let H_0 be the event the results were obtained by chance and H_1 be the event Bruno has ESP. Let D be the event of obtaining the result or a more extreme one.

From other data the researcher believes that the probability ESP exists is 0.01.

d Draw a tree diagram to show the four possibilities from the test, with the first two branches indicating having ESP (H_1) and not having ESP (H_0). You may assume that if a person has ESP they will guess all three cards correctly.

e Hence find the probability that

i Bruno would guess all three cards correctly.

ii Bruno has ESP (note: this is $P(H_1|D)$).

f **i** Comment on which probability used was the least certain.

ii A different researcher believes the probability of ESP existing is 0.001.

If this was the case how would your answer to part **e ii** change?

2 It is known that a particular infection in a hospital will occur by chance in 1.0% of the patients admitted. It is known that poor practices in a hospital will lead to an increased likelihood of this infection occurring.

It is decided to look at 100 patients from a particular hospital and if three or more are found to have contracted the infection then the hospital will be required to review its practices.

The hypotheses used are H_0: $p = 0.01$, H_1: $p > 0.01$

a **i** Find the significance level of this test.

ii Does this mean the probability the hospital has bad practices is over 0.92?

It is know from previous research that 10% of hospitals have bad practices that lead to an increase infection rate. In these hospitals the probability of three or more infections is equal to 1.

A particular hospital is tested and it has three cases of the infection. No other information is available about this hospital.

b **i** Find the probability that the hospital has bad practices ($p > 0.01$).

ii What conclusion would you draw from this information?

3 Investigators have seized a package of drugs. They are keen to know whether this came from inside their country or from outside the country. A chemical in the drugs can be used to test the country of origin. On this test those from inside the country return a value which has a mean of 5.2 and those from outside the country a mean of 4.6. In each case the standard deviation can be assumed to be 1.2.

The drugs are sent away to be tested and 16 tests are run on the sample sent.

A one-tailed test of the sample mean $\bar{x}$ is set up with H_0: $\mu = 5.2$ and H_1: $\mu < 5.2$ and at a 5% significance level.

a Find the critical region for this test.

b Find the probability of a type II error if the drugs are in fact from outside the country.

It is known that generally 90% of the drugs in the area of the seizure are from inside the country and 10% from outside the country.

c Find the probability that $\overline{X}$ will fall in the critical region.

d If having done the tests the sample mean is found to lie in the critical region find the probability the drugs came from inside the country.

Chapter summary

- The product moment correlation coefficient of the ranks of a set of data is called Spearman's rank correlation coefficient. The notation used in IB is r_s.
- Spearman's correlation coefficient shows the extent to which one variable increases or decreases as the other variable increases.
- A value of 1 means the set of data is strictly increasing and a value of -1 means it is strictly decreasing.
- A value of 0 means the data shows no **monotonic** behaviour.
- The null hypothesis is rejected if either the test statistic falls in the critical region (it is beyond the critical value) or the p-value is less than the significance level.
- If a statistic is such that the null hypothesis is rejected we say the result is significant.
- The hypothesis test for the population correlation coefficient (ρ) will have H_0: $\rho = 0$. The p-value can be obtained from the GDC.
- When testing for a population mean, use the z-test if the population standard deviation is known and the t-test if not.
- When testing for the differences between two means use the pooled t-test.
- When the two groups are paired find the difference between each pair and test H_0: $\mu_D = 0$.
- Calculators are likely to use different symbols for s_{n-1} and s_n. Make sure you know which is which.
- In examinations if using the inbuilt testing functions you need to be aware whether you have to enter s_{n-1} or s_n. Depending on what is given in the question you may need to use the formula above to convert between the two.
- When testing for a binomial probability p which is not given in the question, p can be estimated from $\frac{\bar{x}}{n}$.
- A χ^2 **test for independence** can be performed to find out if two data sets are independent of each other or not. The GDC will produce a table of expected frequencies and a p-value. If any expected frequencies are less than 5 then adjacent rows or columns need to be merged.
- In a χ^2 goodness-of-fit test, the degrees of freedom $\nu = (n-1)$.
- To obtain the number of degrees of freedom, take the number of cells minus one, and then subtract one for each of the parameters estimated.
- A probability of a type I error is the probability of rejecting H_0 when H_0 is true. For the normal distribution this will be equal to the significance level, for a discrete distribution this will be the probability of the statistic falling in the critical region.
- In order to find the probability of a type II error, first find the critical region under the null hypothesis.
- The probability of a type II error is the probability of the statistic not being in the critical region. This is calculated using a value for the parameter chosen from the alternative hypothesis.

Developing inquiry skills

The initial claim was: the mean height of the trees from area A is smaller than the mean height of the trees from area B.

Which would be the best test to use to address this claim?

What conditions are necessary and have you tested to see if these conditions have been met?

Carry out the test and state the conclusion.

Chapter review

Click here for a mixed review exercise

1 Prabu took a note of the heights of 12 of her classmates and timed how many seconds it took them to run the 100-metre dash. Her data is in the table below.

Height (cm)	Time (seconds)
151	17.5
153	18
153	16.5
154	16
155	15.4
159	13.2
162	14
164	13.7
164	13.2
168	12.5
175	12
181	12

a Calculate Spearman's rank correlation coefficient (r_s) for this data.

b Interpret the value of r_s and comment on its validity.

2 The colour of eggs laid by three different types of hens was recorded.

	Leghorn	Brahma	Sussex
White eggs	5	23	14
Brown eggs	25	7	16

Phoebe was interested to find out if the colour of the eggs was independent of the type of hen. She decided to perform a χ^2 test at the 5% significance level on her data.

a Write down the null and alternative hypotheses.

b Show that the expected value of a Leghorn laying a white egg is 14.

c Write down the degrees of freedom.

d Find the χ^2 test statistic and the p-value.

e The critical value is 5.991; find the conclusion for this test, justifying your answer both in terms of the critical value and the p-value.

3 Marilu tosses two unbiased coins 60 times. The number of tails that she gets is given in the table below.

Number of tails	0	1	2
Frequency	12	34	14

a Show that the expected frequency for tossing 0 tails is 15.

b Find the table of expected frequencies.

c Write down the degrees of freedom.

d Perform a goodness-of-fit test at the 5% significance level to find out if the data fits a binomial distribution.

The critical value for this test is 5.991.

e State the conclusion for the test, justifying your answer.

4 Mrs Nelson gave her two grade 12 classes the same history test. She wanted to find out if they were, on average, of an equal standard.

The results of the test are:

Class 1	79	63	42	88	95	57	73	61	82	76	51	48
Class 2	65	78	85	49	59	91	68	74	82	56		

a Write down the null and alternative hypotheses.

b State whether this is a one-tailed test or a two-tailed test.

c Perform a t-test at the 5% significance level, stating your conclusion.

5 a Find a 95% confidence limit for the population mean if the following random sample is taken from the population which is assumed to have a normal distribution.

1.3, 1.5, 1.7, 1.4, 1.5, 1.6, 1.9

b Find the 90% confidence interval for the population mean if a sample of size 10 has a mean of 22.1 cm and a standard deviation of 0.8 cm.

6 For many years the number of tornadoes in a particular area during August has followed a Poisson distribution with a zmean of 7.5.

It is thought that climate change might be making the occurrence of more likely. The records for the two most recent years are looked at and the number of tornadoes is found to be 19.

Test at the 5% level the hypothesis that the number of tornadoes is now greater than 7.5.

7 The success rate of a medication is claimed to be 82%.

To test this claim the medication is given to a sample of 42 patients and the number of patients who benefit from the medication, X, is noted.

a Write down a suitable distribution for X and justify your answer.

b State suitable null and alternative hypotheses for the test.

c State the critical region for the test.

d State the probability of a type I error.

8 Consider the following one-tailed test. H_0: $\mu = 30$ against H_1: $\mu > 30$.

The population is normally distributed with a known variance of 9 and a test is performed with a sample of size 4 with the following significance levels:

i 5%

ii 1%

a Write down the probability of a type I error in each case.

b Find the critical region in each case and hence the probability of a type II error given that $\mu = 32$.

c Comment on the effect on the probability of a type II error if:

i the sample size is increased

ii the significance level is reduced

iii the true value of μ increases, for example if $\mu = 34$.

9 A research team asks some employees about various aspects of the company they work for.

Following the survey the company takes action on some of the issues raised and later the company repeats the survey.

a Explain what is meant by the test–retest test for reliability.

A sample of the results is shown in the table below where the average level of satisfaction for each employee is given as a score out of 10.

	A	B	C	D	E	F	G	H
First test	5.4	6.6	4.3	8.1	5.5	3.2	4.6	7.1
Second test	5.6	7.0	4.7	8.1	5.0	3.8	4.3	7.7

b Find the product moment correlation coefficient for the sample and comment on the reliability of the test.

c Perform an appropriate test to see if the sample supports the belief that the levels of satisfaction have improved. State any assumptions you are making.

Exam-style questions

10 **P1:** A conman uses a coin to play a game. This might be an ordinary fair coin, with one side a Head and the other a Tail; or it might be a double-headed coin. The police are suspicious and formulate the following hypotheses.

H_0 : It is a fair coin

H_1 : It is a double headed coin.

A test is to be performed by tossing the coin four times. The following decision rules are made:

If there is at least one Tail, we will conclude that it is a fair coin.

If all four tosses produce a Head, we will conclude that it is a double-headed coin.

a Find the probability of making a Type I error. (3 marks)

b Find the probability of making a Type II error. (3 marks)

11 **P2:** In the country of Sodor there are three Television channels: Alphaview; Betaview; and Peppaview.

A sample of children were asked what their favourite TV channel was. The results are recorded in the table below.

	Alpha	Beta	Peppa
Up to 5 years old	10	10	40
Between 6 and 10 years old	15	25	30
Between 11 and 15 years old	15	35	20

a Perform a suitable test to decide, at the 1% significance level, whether the children's age and preference of TV channel are independent or not.

In your answer, you should include the test used, the test hypotheses, the degrees of freedom, the p-value and the conclusion (with a reason) of the test. (7 marks)

b Give the expected values in a table similar to the one above, and comment on their size in the context of the validity of the test. (4 marks)

12 **P2:** The heights of eight Welsh policemen were measured in centimetres as 170, 174, 176, 175, 171, 165, 179, 180

The heights of 10 Scottish policemen were measured in centimetres as 173, 176, 179, 180, 181, 178, 175, 177, 182, 177

a Find the sample mean height of **i** the Welsh policemen **ii** the Scottish policemen. (2 marks)

You can assume that both sets of values have a common unknown variance.

b Carry out a test at the 5% level to determine if the population mean of Welsh policemen is smaller than that of Scottish policemen. In your answer, you should include the test used (with a reason), the test hypotheses, the p-value and the conclusion of the test (with a reason). (7 marks)

c State if your conclusion would have been any different if working at the 1% significance level. (2 marks)

13 **P1:** The number of "likes" that Narcissus receives on his web site each hour is known to satisfy the Poisson distribution. Narcissus claims that on average he receives 20 "likes" every hour. We suspect that he is exaggerating. We monitor his web site over a period of 6 hours and find that he received 100 "likes" during this time.

Test Nar.issus's claim against our suspicion at the 5% significance level. In your answer, you should include whether this is a one-tailed or two-tailed test, the test hypotheses, calculation of the p-value, and the conclusion (with a reason) of the test. (7 marks)

14 **P2:** **a** State the central limit theorem. (4 marks)

Let the random variable X be the number of hours that a random student spends using their mobile phone each day. It is known that X has a population variance of $\sigma^2 = 1$ hour.

b A sample of 100 students were asked how many hours they spent using their phone in a day and the sample mean was $\bar{x} = 5$ hours. (3 marks)

Calculate the 95% confidence interval for the population mean, μ.

15 **P1:** It is claimed that the tread on the front tyres of rally cars wears more quickly than the tread on the rear tyres. To test this claim, the wear, in millimetres, on the front tyres and on the back tyres of 10 rally cars is measured at the end of a rally event.

The paired data is given in the table below.

Car	1	2	3	4	5	6	7	8	9	10
Front wear	4.1	4.0	4.2	3.9	4.8	4.5	4.6	5.0	4.7	4.7
Rear wear	3.9	4.0	3.8	4.0	4.2	4.5	4.4	4.9	4.8	4.4

Test the above claim at the 5% significance level. In your answer, you should include the test used (with a reason), the test hypotheses, the p-value and the conclusion of the test (with a reason). (8 marks)

16 **P1:** Let the random variable X be the height of a female, in centimetres, and let the random variable Y be the number of pets that she owns. A random sample of size 6 was taken and the paired data is shown in the table below.

Height	170	165	160	155	157	162
Number of pets	0	4	3	0	1	2

Use this data to test at the 5% significance level if there is a linear relationship between these two variables. (8 marks)

17 **P1:** The manufacturers of toys that come in surprise eggs claim that the colour of the toys;

Blue, Pink, Purple, and Green; appear in the ratios 3:4:2:1 respectively.

A sample of 100 toys was taken and the number of each colour is shown in the table below.

Colour	Blue	Pink	Purple	Green
Number	32	37	23	8

Test the manufacturer's claim at the 10% significance level. (9 marks)

Rank my maths!

Approaches to learning: Collaboration, Communication

Exploration criteria: Personal engagement (C), Reflection (D), Use of mathematics (E)

IB topic: Spearman's rank, hypothesis test

The task

In this task you will be designing an experiment that will compare the rankings given by different students in your class. You will then use your results to determine how similar they are and what agreement there is and then to test the significance of that relationship.

The experiment

Select one student in your group to be the experimenter. The other students in your group will be the subject of the experiment.

The experimenter is going to determine whether there is any similarity between the tastes of the other students in their group by asking them to rank a set of selections from the least to the most favourite.

Would you expect the rankings of the students in your group to be the same, similar or completely different?

What would help you to predict who might have similar tastes (where the strongest correlation would be)?

Under what circumstances might the rankings be similar and under what circumstances might they be different?

In your group, discuss:

What could you use the Spearman's rank method to compare?

What other ideas can you think of that this could be used to test?

The experimenter should prepare their own set of selections for the other members of the group to compare in the area they have chosen to investigate.

They should make a prediction about how strong they think the rank correlation will be between the other students with regards to the experiment they are doing.

The experimenter will give their set of selections to the group.

The other students will rank the set of selections from favourite (1) to least favourite (n), where n is the number of selections.

What does the experimenter need to be aware of when providing instructions?

The experimenter should record the rankings in a table.

The students who are doing the ranking should not collaborate or communicate with each other.

Why is this important?

The results

Now find the Spearman rank correlation between the students.

Do the students display strong correlations?

Are the results what were expected.

Discuss as a group why the original hypothesis may have been accurate or inaccurate.

Write a conclusion for the experiment based on the results you have found so far.

The statistical test

What spearman's rank value would mathematically be considered to be high?

It is possible to test the significance of a relationship between the two rankings. The test is similar to the one conducted for the product moment correlation coefficient.

Conduct a hypothesis test for the Spearman's rank value(s) you have found.

What conclusions can you draw from this? How confident are you in your results?

Extension

What other experiments or data collection exercises can you think of that will be suitable for a statistics-based exploration that will result in a hypothesis test like the one in this task?

Use the examples and questions in this chapter to give you some inspiration and then design your experiment and how you will conduct it.