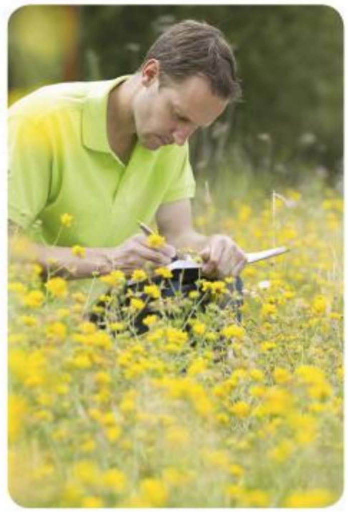
8.2 χ^2 test for independence



So far 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, she can use this data to find the mean or standard deviation of this particular sample of the 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 as a discipline involves collecting a sample and, from it, estimating:

- **parameters**; for example, the mean or correlation coefficient of a whole population, when all you have are these values for a sample of the population
- the **distribution** of the population; for example, whether or not the population is normally distributed (see Chapter 7).

Hypothesis testing

In statistics a **hypothesis** is a statement about the unknown parameters or features of a data set.

The aim of a **statistical test** is to try to find out whether the data supports your hypothesis. If it does, this is an example of **statistical inference**: you have inferred something from the statistics of the sample you are considering.

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

Every hypothesis has an **alternative hypothesis** that will be accepted if H_0 is rejected; we write this as H_1.

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

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

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 10 times to see what happens.

Continued on next page

Statistics and probability

1 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. If this is the case, how might you be able to reduce your suspicions?

2 a Use the binomial distribution to work out the probability of the events listed in **1a** 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?

3 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 $\mathrm{H}_0: p = 0.5$ rather than $\mathrm{H}_0: p \neq 0.5$.

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

4 The events, such as those in **1a**, 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 even when H_0 is true less than 5% of the time. This is referred to as a 5% **significance level**.

a Write down the significance level of your test.

b Find the significance level for a different choice of critical region. Would you prefer this one over the one chosen previously? Justify your answer.

c Comment on the symmetry of your critical regions and discuss whether they have to be symmetrical.

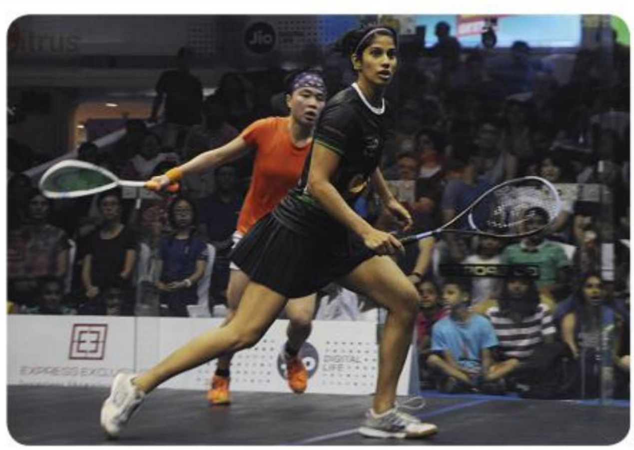
5 a Deduce whether you can ever be sure that the coin is biased; explain your answer.

b Suppose the result of your experiment fell just outside the critical region. Explain whether or not this means that the null hypothesis is true.

6 **Conceptual** What is meant by the significance level of a hypothesis test?

7 **Conceptual** What is the purpose of hypothesis testing?

Saanvi is a member of a sports club. She has noticed that more males play squash than females, and is interested to find out whether 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 shown in the table.



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 wants to know whether or not the choice of game is independent of gender.

A χ^2 **test for independence** shows whether two data sets are independent of each other or not. It can be performed at various significance levels.

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

Her null hypothesis, H_0, is:

H_0 : Preferred racket game is independent of gender.

Her alternative hypothesis is:

H_1 : Preferred racket game is not independent of gender.

EXAM HINT

In the examination it will only be tested at the 1%, 5% or 10% significance level.

Investigation 3

Complete the following table. This is the table of **observed frequencies**, f_o, and is called a contingency table.

Sport	Tennis	Badminton	Squash	Totals
Male	10			
Female				
Totals				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 are independent, find the expected number of males who like tennis best.

These are called the **expected frequencies**, f_e.

Continued on next page

TOK

The use of *p*-values and null hypothesis testing is "surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students."

— *William Rozeboom*

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

5 Complete the table of expected frequencies.

Sport	Tennis	Badminton	Squash	Totals
Male				
Female				
Totals				

Note that the expected frequencies must be greater than 5. If there are expected frequencies less than 5 then you will need to combine rows or columns.
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	Totals
Male				33
Female				27
Totals	19	20	21	60

EXAM HINT
You will not need to do this in exams, but you may need to in your internal assessments.

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 **Conceptual** What does the number of degrees of freedom represent?

The formula for the degrees of freedom is

$$v = (\text{rows} - 1)(\text{columns} - 1)$$

In examinations, v will always be greater than 1.

Investigation 3 (continued)

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?

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

12 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.

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

14 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**.

EXAM HINT

In examinations you will use your GDC to find this value. Go to Matrix, enter the number of rows and the number of columns (not including the totals, if included in the question). Enter the data and store the matrix. Next go to Stat tests, chi-squared 2-way test, enter the letter that you stored the matrix under and press OK. You will see the chi-squared value, the p-value and the number of degrees of freedom.

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.

Exercise 8B

1 Misty was interested to find out whether preference for car colour was dependent on gender. She asked 80 of her friends and the results are shown in the table.

Colour of car	White	Black	Red	Blue	Totals
Male	6	14	10	8	38
Female	12	8	9	13	42
Totals	18	22	19	21	80

a Show that the expected number of males who prefer black cars is 10.45.

b Show that the expected number of females who prefer white cars is 9.45.

c Find the χ^2 value.

2 Max was watching people coming out of a take-away coffee shop. He was noting down the gender and whether they had bought a small, medium or large coffee. The results were:

Coffee size	Small	Medium	Large	Totals
Female	22	16	18	56
Male	12	14	28	54
Totals	34	30	46	110

a Show that the expected number of males who buy a small coffee is 16.7.

b Show that the expected number of females who buy a large coffee is 23.4.

c Find the χ^2 value.

3 Brenda works in a pet shop. She watches the rabbits, guinea pigs and hamsters to see if they eat the lettuce or carrots first. Her results are:

Pet	Rabbits	Guinea pigs	Hamsters
Lettuce	16	16	28
Carrots	34	18	18

a Show that the expected number of rabbits who eat carrots first is 26.9.

b Show that the expected number of hamsters who eat lettuce first is 21.2.

c Find the χ^2 value.

4 Ziyue asked her year group how they travelled to school. Her results are shown in the table.

Transport	Car	Bus	Bicycle	Walk
Male	12	12	28	8
Female	21	13	15	11

a Show that the expected number of males who come by bicycle is 21.5.

b Show that the expected number of females who come by car is 16.5.

c Use your GDC to find the χ^2 value.

Investigation 3 (continued)

15 Use the entries in the tables from earlier in the investigation for the observed and expected frequencies to find the χ^2 test statistic. Complete the following:

f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
10	10.45	−0.45	0.2025	0.1937...
			$\sum \frac{(f_o - f_e)^2}{f_e}$	

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.

16 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. With a larger table, and hence more degrees of freedom, 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 two degrees of freedom at the 5% significance level is $\chi^2_{5\%} = 5.991$.

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

18 Use your GDC with the observed values given above and verify your previously obtained value for the test statistic.

Your GDC also gives you a p-value. This is the probability value. If the p-value is smaller than the level of significance then you do not accept the null hypothesis.

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

20 **Conceptual** How would you use the result of the chi squared test to determine whether two variable are independent or not?

21 **Conceptual** What is the purpose of the χ^2 test?

- If the χ^2 test statistic is less than the critical value then you accept the null hypothesis. If it is greater than the critical value then you do not accept the null hypothesis.

$$\chi^2_{\text{calc}} < \text{critical value} \Rightarrow \text{accept } H_0$$
$$\chi^2_{\text{calc}} > \text{critical value} \Rightarrow \text{reject } H_0$$

- If the p-value is greater than the significance level (0.01, 0.05 or 0.10) then you accept the null hypothesis. If it is less than the significance level then you do not accept the null hypothesis.

$$p\text{-value} > 0.05 \Rightarrow \text{accept } H_0$$
$$p\text{-value} < 0.05 \Rightarrow \text{reject } H_0$$

TOK

How can a mathematical model give us knowledge even if it does not yield accurate predictions?

You can use **either** the test statistic and the critical value **or** the p-value and the significance level to reach a conclusion. You may only be given one of these values (the critical value or the significance level), so you should know how to use both.

Example 2

Eighty people were asked for their favourite genre of music: pop, classical, folk or jazz. The results are in the following table.

Genre	Pop	Classical	Folk	Jazz	Totals
Male	18	9	4	7	38
Female	22	6	7	7	42
Totals	40	15	11	14	80

A χ^2 test was carried out at the 1% significance level. The critical value for this test is 11.345.

a Write down the null and alternative hypotheses.

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

c Write down the number of degrees of freedom.

Continued on next page

d Find the χ^2 test statistic and the p-value.

e State whether the null hypothesis is accepted or not, giving a reason for your answer.

a H_0 : Favourite music genre is independent of gender H_1 : Favourite music genre is not independent of gender **b** E(female liking pop) = P(female) × P(likes pop) × total $= \frac{42}{80} \times \frac{40}{80} \times 80 = 21$ **c** $\nu = (2-1)\times(4-1) = 3$ **d** $\chi^2 = 1.622...$ and $p = 0.654...$ **e** $0.654 > 0.01$ or $1.622 < 11.345$ and so you accept the null hypothesis: favourite music genre is independent of gender.	Remember that you do not count the Totals row or column.

Example 3

American bulldogs are classified by height, h, as Pocket, Standard or XL. Pockets have $h < 42$ cm high, Standards have $42 \le h < 50$ and XLs have $50 \le h < 58$. At a dog show, Marius measures and weighs 50 dogs. He is interested to find out whether class of dog is independent of weight and decides to perform a χ^2 test at the 5% significance level. The results are shown in the table.

Height	Weight	Height	Weight	Height	Weight
36	30	42	38	50	39
37	33	42	39	51	41
37	36	43	36	51	42
38	31	43	44	52	45
38	38	44	42	52	45
39	32	44	48	52	51
39	39	45	46	53	53
39	42	46	49	54	55
40	41	46	38	54	48
40	43	46	42	54	56
40	38	47	46	55	58
41	38	47	50	55	51
41	44	47	52	56	54
41	46	48	49	56	53
41	45	48	48	56	55
41	47	48	42	57	58
		49	53	57	59

a Find the mean weight of the 50 dogs.

b Complete the following contingency table:

Class	Pocket	Standard	XL
< mean			
≥ mean			

c Write down the null and alternative hypotheses.

d Write down the number of degrees of freedom.

e Show that the expected number of XL dogs that weigh less than the mean is 8.16.

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

g Comment on your answer.

a Mean weight = 44.96

b

Class	Pocket	Standard	XL
< mean	13	8	3
≥ mean	3	9	14

c H_0 : Class of dog is independent of weight.
H_1 : Class of dog is not independent of weight.

d $v = 2$

e $\frac{17}{50} \times \frac{24}{50} = 8.16$

f $\chi^2 = 13.4$ and p-value $= 0.001\,25$

g $0.001\,25 < 0.05$, therefore do not accept the null hypothesis.

Exercise 8C

1 Pippa sends out a questionnaire to 50 of her classmates asking what their favourite sport is. She wants to conduct a χ^2 test at the 10% significance level to find out whether favourite sport is independent of gender.

The results are as follows:

Males: cycling, cycling, basketball, football, football, football, basketball, basketball, basketball, basketball, football, football, football, cycling, cycling, basketball, basketball, basketball, basketball, basketball, cycling, cycling, cycling.

Females: football, football, football, basketball, basketball, cycling, basketball, basketball, basketball, cycling, cycling, cycling, football, football, football, cycling, basketball, basketball, cycling, football, football, football, football, cycling, cycling, cycling, basketball.

a Set up a contingency table to display the results.

b Write down the null and alternative hypotheses.

c Write down the number of degrees of freedom.

d Find the χ^2 value and the p-value.

e Check that expected values are greater than 5.

The critical value is 4.605.

f Write down the conclusion of the test.

g Comment on whether the p-value supports this conclusion.

2 A survey was conducted to find out which type of bread males and females prefer. Eighty people were interviewed outside a baker's shop and the results are shown below.

Bread	White	Brown	Corn	Multi-grain	Totals
Male	14	10	7	8	39
Female	17	6	6	12	41
Totals	31	16	13	20	80

Using the χ^2 test at the 5% significance level, determine whether the favourite type of bread is independent of gender.

a State the null hypothesis and the alternative hypothesis.

b Show that the expected frequency for female and white bread is approximately 17.9.

c Write down the number of degrees of freedom.

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

The critical value is 7.815.

e Comment on your result.

3 Three hundred people of different ages were interviewed and asked which genre of film they mostly watched (thriller, comedy or horror). The results are shown below.

Film type	Thriller	Comedy	Horror	Totals
0–20 years	13	26	41	80
20–50 years	54	48	28	130
51+ years	39	43	8	90
Totals	106	117	77	300

Using the χ^2 test at the 10% significance level, determine whether the genre of film watched is independent of age.

a State the null hypothesis and the alternative hypothesis.

b Show that the expected frequency for preferring horror films between the ages of 20 to 50 years is 33.4.

c Write down the number of degrees of freedom.

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

e Comment on your result.

4 Three different flavours of dog food were tested on different breeds of dogs to see whether there was any connection between favourite flavour and breed. The results are shown in the table below.

Flavour	Beef	Chicken	Lamb	Totals
Boxer	14	6	8	28
Labrador	17	11	10	38
Poodle	13	8	14	35
Collie	6	5	8	19
Totals	50	30	40	120

Perform a χ^2 test at the 5% significance level to test whether favourite flavour is independent of breed of dog.

a State the null hypothesis and the alternative hypothesis.

b Write down the table of expected frequencies.

c Combine the results for collie and poodle so that all expected values are greater than five and write down the new table of observed values.

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

The critical value is 9.488.

e Comment on your result for a 5% significance level.

5 Sixty people were asked what their favourite flavour of chocolate was (milk, pure, white). The results are shown in the table below.

Flavour	Milk	Pure	White	Totals
Male	10	17	5	32
Female	8	6	14	28
Totals	18	23	19	60

A χ^2 test at the 1% significance level is set up.

a State the null hypothesis and the alternative hypothesis.

b Write down the number of degrees of freedom.

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

The critical value is 9.210.

d Comment on your result.

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

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

He decided to perform a χ^2 test at the 10% significance level to find out whether there is a connection between GPA and number of hours spent 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 Write down the number of degrees of freedom.

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

The critical value is 7.779.

e Comment on your result.

7 Hubert wanted to find out whether the number of people walking their dog was related to the time of day. He kept a record covering 120 days and the results are shown in the following table.

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

Test, at the 5% significance level, whether there is a connection between time of day and number of people walking their dog. The critical value for this test is 9.488.

8 Carle has a part-time job working at a corner shop. He decides to see whether there is a connection between the temperature and the number of bottles of water sold. His observations are in the table below.

Temperature	$<21°C$	21°C–30°C	$>30°C$	Totals
<30 bottles	14	23	12	49
30–60 bottles	10	31	17	58
>60 bottles	8	26	9	43
Totals	32	80	38	150

Test, at the 1% significance level, whether there is a connection between temperature and the number of bottles of water sold.

a State the null hypothesis and the alternative hypothesis.

b Write down the number of degrees of freedom.

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

The critical value is 13.277.

d Comment on your result.