5.4 The chi-squared test

When two variables are independent, one does not affect the other. Here, you are finding out whether a person's gender influences their colour choice. You will learn more about mathematical independence in Chapter 8.

You may be interested in finding out whether or not certain sets of data are independent. Suppose you collect data on the favorite color of T-shirt for men and women. You may want to find out whether color and gender are independent or not. One way to do this is to perform a **chi-squared test (χ^2)** for independence.

To perform a chi-squared test (χ^2) there are four main steps.

Step 1: Write the **null** (H_0) and **alternative** (H_1) **hypotheses**.

H_0 states that the data sets are independent.
H_1 states that the data sets are not independent.

For example, the hypotheses for color of T-shirt and gender could be:
H_0: Color of T-shirt is independent of gender.
H_1: Color of T-shirt is not independent of gender.

Step 2: Calculate the chi-squared test statistic.

Firstly, you may need to put the data into a **contingency table**, which shows the frequencies of two variables. The elements in the table are the **observed** data. The elements should be frequencies (not percentages).

For the example above, the contingency table could be:

The main entries in this table form a 2×4 **matrix** (array of numbers) - do not include the row and column for the totals.

	Black	White	Red	Blue	Totals
Male	48	12	33	57	150
Female	35	46	42	27	150
Totals	83	58	75	84	300

If you are given the contingency table, you may need to extend it to include an extra row and column for the 'Totals'.

In examinations, the largest contingency table will be a 4×4.

From the observed data, you can calculate the **expected frequencies**. Since you are testing for independence, you can use the formula for the probability of independent events to calculate the expected values. So:

The expected number of men who like black T-shirts is

$$\frac{150}{300} \times \frac{83}{300} \times 300 = 41.5.$$

The expected number of men who like white T-shirts is

$$\frac{150}{300} \times \frac{58}{300} \times 300 = 29 \text{ and so on.}$$

The expected table of values would then look like this:

	Black	White	Red	Blue	Totals
Male	41.5	29	37.5	42	150
Female	41.5	29	37.5	42	150
Totals	83	58	75	84	300

Note:

- The expected values can **never** be less than 1.
- The expected values must be 5 or higher.
- If there are entries between 1 and 5, you can combine table rows or columns.

For calculations by hand, you need the expected frequencies to find the χ^2 value.

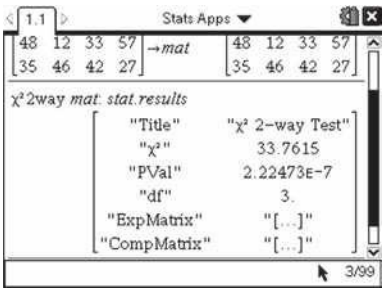
➔ To calculate the χ^2 value use the formula $\chi^2_{\text{calc}} = \sum \frac{(f_o - f_e)^2}{f_e}$, where f_o are the observed frequencies and f_e are the expected frequencies.

For our example,

$$\chi^2_{\text{calc}} = \frac{(48-41.5)^2}{41.5} + \frac{(12-29)^2}{29} + \frac{(33-37.5)^2}{37.5} + \frac{(57-42)^2}{42} + \frac{(35-41.5)^2}{41.5}$$
$$+ \frac{(46-29)^2}{29} + \frac{(42-37.5)^2}{37.5} + \frac{(27-42)^2}{42}$$
$$= 33.8$$

In examinations, you will only be expected to use your GDC to find the χ^2 value.

Using your GDC to find the χ^2 value, enter the contingency table as a matrix (array) and then use the matrix with the χ^2 2-way test.



Your GDC calculates the expected values for you but you must know how to find them by hand in case you are asked to show one or two calculations in an exam question. To see the matrix for the expected values, type 'stat.' and then select 'expmatrix' from the menu that pops up.

From the screenshot, you can see that $\chi^2_{\text{calc}} = 33.8$ (to 3 sf). This confirms our earlier hand calculation.

Step 3: Calculate the critical value.

First note the **level of significance**. This is given in examination questions but you have to decide which level to use in your project. The most common levels are 1%, 5% and 10%.

GDC help on CD: *Alternative demonstrations for the TI-84 Plus and Casio FX-9860GII GDCs are on the CD.*

Now you need to calculate the number of **degrees of freedom**.

➔ To find the degrees of freedom for the chi-squared test for independence, use this formula based on the contingency table:

Degrees of freedom = (number of rows − 1) (number of columns − 1)

If the number of degrees of freedom is 1, you will be expected to use **Yates' continuity correction** to work out the chi-squared value. (In examinations the degrees of freedom will always be greater than 1.)

So, in our ongoing example, the number of degrees of freedom is
$(2 - 1) \times (4 - 1) = 3$

The level of significance and degrees of freedom can be used to find the critical value. However, in examinations, the **critical value** will always be given.

For our example, at the 1% level, the critical value is 11.345. At the 5% level, the critical value is 7.815. At the 10% level, the critical value is 6.251.

Step 4: Compare χ^2_{calc} against the critical value.

→ If χ^2_{calc} is **less than** the critical value then **do not reject** the null hypothesis.
If χ^2_{calc} is **more than** the critical value then **reject** the null hypothesis.

In our example, at the 5% level, 33.8 > 7.815. Therefore, we reject the null hypothesis that T-shirt color is independent of gender.

Using a GDC, you can compare the *p*-value against the significance level.

The *p*-value is the probability value. It is the probability of evidence against the null hypothesis.

→ If the *p*-value is **less** than the significance level then **reject** the null hypothesis.
If the *p*-value is **more** than the significance level then **do not reject** the null hypothesis.

Use the significance level as a decimal, so 1% = 0.01, 5% = 0.05 and 10% = 0.1.

So, for our example, *p*-value = 0.000 000 2 (see the GDC screenshot on page 234).

0.000 000 2 < 0.05, so we reject the null hypothesis.

→ **To perform a χ^2 test:**

1. Write the null (H_0) and alternative (H_1) hypotheses.
2. Calculate χ^2_{calc}:
   - **a** using your GDC (examinations)
   - **b** using the χ^2_{calc} formula (project work)
3. Determine:
   - **a** the *p*-value by using your GDC
   - **b** the critical value (given in examinations)
4. Compare:
   - **a** the *p*-value against the significance level
   - **b** χ^2_{calc} against the critical value

Investigation – shoe size and gender

Use the information that you collected at the beginning of this chapter to test if shoe size is independent of gender.

Example 13

One hundred people were interviewed outside a chocolate shop to find out which flavor of chocolate cream they preferred. The results are given in the table, classified by gender.

	Strawberry	Coffee	Orange	Vanilla	Totals
Male	23	18	8	8	57
Female	15	6	12	10	43
Totals	38	24	20	18	100

Perform a χ^2 test, at the 5% significance level, to determine whether the flavor of chocolate cream is independent of gender.

a State the null hypothesis and the alternative hypothesis.

b Show that the expected frequency for female and strawberry flavor is approximately 16.3.

c Write down the number of degrees of freedom.

d Write down the χ^2_{calc} value for this data.

The critical value is 7.815.

e Using the critical value or the *p*-value, comment on your result.

Answers

a H_0: Flavor of chocolate cream is independent of gender.
H_1: Flavor of chocolate cream is not independent of gender.

Write H_0 using 'independent of'.
Write H_1 using 'not independent of'.

b $\frac{43}{100} \times \frac{38}{100} \times 100 = 16.34$

So, the expected frequency for female and strawberry flavor is approximately 16.3.

From the contingency table:
Total for 'female' row = 43
Total for 'strawberry' column = 38
Total surveyed = 100

c Degrees of freedom = (2 − 1)(4 − 1) = 3

Degrees of freedom = (number of rows − 1) (number of columns − 1)
Here, there are 2 rows and 4 columns in the observed matrix of the contingency table.

d $\chi^2_{calc} = 6.88$

Using your GDC:
Enter the contingency table as a matrix. Use the matrix with χ^2 2-way test. Read off χ^2 value.
The p-value = 0.0758.

e 6.88 < 7.815; therefore, we do not reject the null hypothesis. There is enough evidence to conclude that flavor of chocolate cream is independent of gender.

Using the given critical value, check:
χ^2_{calc} < critical value → do not reject, or
χ^2_{calc} > critical value → reject.

Or, using the p-value, check:
p-value < significance level → reject, or
p-value > significance level → do not reject.
Significance level = 5% = 0.05. So, 0.0758 > 0.05 and we do not reject the null hypothesis.

Example 14

Members of a club are required to register for one of three games: billiards, snooker or darts.
The number of club members of each gender choosing each game in a particular year is shown in the table.

	Billiards	Snooker	Darts
Male	39	16	8
Female	21	14	17

Perform a χ^2 test, at the 10% significance level, to determine if the chosen game is independent of gender.

a State the null hypothesis and the alternative hypothesis.

b Show that the expected frequency for female and billiards is approximately 27.1.

c Write down the number of degrees of freedom.

d Write down the χ^2_{calc} value for this data.

The critical value is 4.605.

e Using the critical value or the *p*-value, comment on your result.

Answers

a H_0: The choice of game is independent of gender.
H_1: The choice of game is not independent of gender.

b $\left(\frac{52}{115}\right)\left(\frac{60}{115}\right)(115) = 27.130$

≈ 27.1

So, the expected frequency for female and billiards is approximately 27.1.

c Degrees of freedom = $(2 - 1)(3 - 1) = 2$

d $\chi^2_{\text{calc}} = 7.79$

e $7.79 > 4.605$; therefore, we reject the null hypothesis. There is enough evidence against H_0 to conclude that the choice of game is not independent of gender.

Expected value table from the GDC:

	Billiards	Snooker	Darts
Male	32.9	16.4	13.7
Female	27.1	13.6	11.3

The p-value = 0.0203
Or, using the p-value,
0.0203 < 0.10. Therefore, we reject the null hypothesis.

Exercise 5H

EXAM-STYLE QUESTIONS

1 300 people were interviewed and asked which genre of books they mostly read. The results are given below in a table of observed frequencies, classified by age.

		Genre			
		Fiction	**Non-fiction**	**Science fiction**	**Totals**
Age	**0–25 years**	23	16	41	80
	26–50 years	54	38	38	130
	51+ years	29	43	18	90
	Totals	106	97	97	300

Perform a χ^2 test, at the 5% significance level, to determine whether genre of book is independent of age.

a State the null hypothesis and the alternative hypothesis.

b Show that the expected frequency for science fiction and the 26–50 age group is 42.

c Write down the number of degrees of freedom.

d Write down the χ^2_{calc} value for this data.

The critical value is 9.488.

e Using the critical value or the p-value, comment on your result.

2 Tyne was interested in finding out whether natural hair color was related to eye color. He surveyed all the students at his school. His observed data is given in the table below.

		Hair color			
		Black	**Brown**	**Blonde**	**Totals**
Eye color	**Brown/Black**	35	43	12	90
	Blue	8	27	48	83
	Green	9	20	25	54
	Totals	52	90	85	227

Perform a chi-squared test, at the 10% significance level, to determine if hair color and eye color are independent.

a State the null hypothesis and the alternative hypothesis.

b Find the expected frequency of a person having blonde hair and brown eyes.

c Write down the number of degrees of freedom.

d Write down the chi-squared value for this data.

The critical value is 7.779.

e Using the critical value or the p-value, comment on your result.

EXAM-STYLE QUESTIONS

3 Three different flavors of dog food were tested on different breeds of dog to find out if there was any connection between favorite flavor and breed. The results are given in the table.

	Beef	Chicken	Fish	Totals
Poodle	13	11	8	32
Boxer	15	10	10	35
Terrier	16	12	9	37
Great Dane	17	11	8	36
Totals	61	44	35	140

A χ^2 test, at the 5% significance level, is performed to investigate the results.

a State the null hypothesis and the alternative hypothesis.

b Show that the expected frequency of a Boxer's favorite food being chicken is 11.

c Show that the number of degrees of freedom is 6.

d Write down the χ^2_{calc} value for this data.

The critical value is 12.59.

e Using the critical value or the *p*-value, comment on your result.

4 Eighty people were asked to identify their favorite film genre. The results are given in the table below, classified by gender.

	Adventure	Crime	Romantic	Sci-fi	Totals
Male	15	12	2	12	41
Female	7	9	18	5	39
Totals	22	21	20	17	80

A χ^2 test, at the 1% significance level, is performed to decide whether film genre is independent of gender.

a State the null hypothesis and the alternative hypothesis.

b Show that the expected frequency of a female's favorite film genre being crime is 10.2.

c Write down the number of degrees of freedom.

d Write down the chi-squared value for this data.

The critical value is 11.345.

e Using the critical value or the *p*-value, comment on your result.

EXAM-STYLE QUESTIONS

5 Kyu Jin was interested in finding out whether or not the number of hours spent playing computer games per week had an influence on school grades. He collected the following information.

	Low grades	Average grades	High grades	Totals
0–9 hours	6	33	57	96
10–19 hours	11	35	22	68
> 20 hours	23	22	11	56
Totals	40	90	90	220

Perform a chi-squared test, at the 5% significance level, to decide whether the grade is independent of the number of hours spent playing computer games.

a State the null hypothesis and the alternative hypothesis.
b Show that the expected frequency of a high grade and 0–9 hours of playing computer games is 39.3.
c Show that the number of degrees of freedom is 4.
d Write down the χ^2_{calc} value for this data.

The critical value is 9.488.

e Using the critical value or the p-value, comment on your result.

6 The local authority conducted a survey in schools in Rotterdam to determine whether the employment grade in the school was independent of gender. The results of the survey are given in the table.

	Directors	Management	Teachers	Totals
Male	26	148	448	622
Female	6	51	1051	1108
Totals	32	199	1499	1730

Perform a χ^2 test, at the 10% significance level, to determine whether the employment grade is independent of gender.

a State the null hypothesis and the alternative hypothesis.
b Write down the table of expected frequencies.
c Write down the number of degrees of freedom.
d Write down the chi-squared value for this data.

The critical value is 4.605.

e Using the critical value or the p-value, comment on your result.

EXAM-STYLE QUESTIONS

7 Ayako had a part-time job working at a sushi restaurant. She calculated the average amount of sushi sold per week to be 2000. She decided to find out if there was a relationship between the day of the week and the amount of sushi sold. Her observations are given in the table.

	< 1700	1700–2300	> 2300	Totals
Monday–Wednesday	38	55	52	145
Thursday–Friday	39	65	55	159
Saturday–Sunday	43	60	63	166
Totals	120	180	170	470

Perform a χ^2 test, at the 5% significance level, to determine whether the amount of sushi sold is independent of the day of the week.

a State the null hypothesis and the alternative hypothesis.

b Show that the expected frequency of selling over 2300 sushi on Monday–Wednesday is 52.4.

c Write down the number of degrees of freedom.

d Write down the χ^2_{calc} value for this data.

The critical value is 9.488.

e Using the critical value or the *p*-value, comment on your result.

8 Haruna wanted to investigate the connection between the weight of dogs and the weight of their puppies. Her observed results are given in the table.

		Puppy			
		Heavy	**Medium**	**Light**	**Totals**
Dog	**Heavy**	23	16	11	50
	Medium	10	20	16	46
	Light	8	15	22	45
	Totals	41	51	49	141

Perform a χ^2 test, at the 1% significance level, to determine whether a puppy's weight is independent of its parent's weight.

a State the null hypothesis and the alternative hypothesis.

b Show that the expected frequency of a medium dog having a heavy puppy is 13.4.

c Write down the number of degrees of freedom.

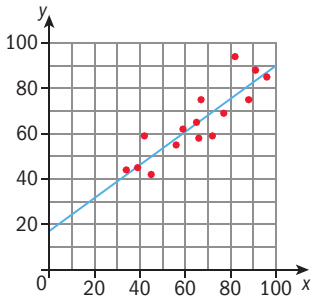
d Write down the χ^2_{calc} value for this data.

The critical value is 13.277.

e Using the critical value or the *p*-value, comment on your result.

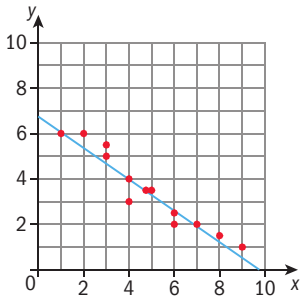
Extension material on CD:
Worksheet 5 - Useful statistical techniques for the project

3 a, c moderate, positive, linear correlation



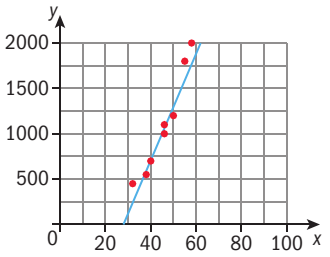
b 65.3 and 65.1 d 54%

4 a, c moderately strong, negative, linear correlation



b 4.83 and 3.5 d 4.5

5 a, c strong, positive, linear correlation



b 45.6 and 1100

d 1500

Exercise 5F

1 0.931 very strong and positive

2 a 0.880

b strong, positive

3 –0.891 strong and negative

4 0.936 very strong and positive

5 0.990 very strong and positive

6 0.200 very weak and positive

7 0.985 very strong and positive

8 0.580 moderate and positive

Exercise 5G

1 a 0.994 very strong and positive

b $y = 1.47x + 116$

c 1586 rupees

2 a 0.974

b $y = 0.483x + 15.6$

c 19.5 cm

3 a mean of $x = 68.6$ and standard deviation of $x = 6.55$

mean of $y = 137.7$ and standard deviation of $y = 5.97$

b –0.860

c strong and negative

d $y = -0.784x + 191.5$

e 137 s

4 a 0.792

b $y = 0.193x + 1.22$ c 4

5 a $y = 0.0127x + 0.688$

b 1.58 AUD

6 a $y = 0.751x + 11.6$ b 49

7 a $y = 1.04x - 2.53$ b 60

8 a $y = 0.279x + 2.20$

b 13.4 hours

Exercise 5H

1 a H_0: Genre of books is independent of age

H_1: Genre of books is not independent of age

b $130 \times \frac{97}{300} = 42.0$

c 4 d 26.9

e 26.9 > 9.488 so reject null hypothesis

2 a H_0: Hair color is independent of eye color

H_1: Hair color is not independent of eye color

b $85 \times \frac{90}{227} = 33.7$

c 4 d 44.3

e 44.3 > 7.779 so reject the null hypothesis

3 a H_0: Favorite flavor is independent of breed

H_1: Favorite flavor is not independent of breed

b $35 \times \frac{44}{140} = 11$

c $(3 - 1)(4 - 1) = 6$

d 0.675

e 0.675 < 12.59 so do not reject the null hypothesis

4 a H_0: Film genre is independent of gender

H_1: Film genre is not independent of gender

b $39 \times \frac{21}{80} = 10.2$

c 3 d 19.1

e 19.1 > 11.345 so reject the null hypothesis

5 a H_0: Grade is independent of number of hours spent playing computer games

H_1: Grade is not independent of number of hours spent playing computer games

b $90 \times \frac{96}{220} = 39.27 \approx 39.3$

c $(3 - 1)(3 - 1) = 4$ d 42.1

e 42.1 > 9.488 so reject the null hypothesis

6 a H_0: Employment grade is independent of gender

H_1: Employment grade is not independent of gender

b

11.5	71.5	539
20.5	127.5	960

c 2 d 180

e 180 > 4.605 so reject the null hypothesis

7 a H_0: Amount of sushi is independent of day of the week

H_1: Amount of sushi is not independent of day of the week

b $170 \times \frac{145}{470} = 52.4$

c 4 d 0.840

e 0.840 < 9.488 so do not reject the null hypothesis.

8 a H_0: Puppy's weight is independent of its parent's weight

H_1: Puppy's weight is not independent of its parent's weight

b $46 \times \frac{41}{141} = 13.38 \approx 13.4$

c 4

d 13.7

e 13.7 > 13.277 so reject the null hypothesis