Chapter 31

χ^2 hypothesis tests

Contents:

A The χ^2 goodness of fit test
B Estimating distribution parameters in a goodness of fit test
C Critical regions and critical values
D The χ^2 test for independence

OPENING PROBLEM

A multiple choice quiz has 10 questions. Each question has 4 choices, only 1 of which is correct.

The quiz is given to 150 people, and the number of correct answers for the participants is summarised in the table alongside.

A statistician analysing the data assumes that each person answers all questions by randomly guessing.

Number of correct answers	*Frequency*
0	7
1	34
2	52
3	36
4	12
5 or more	9

Things to think about:

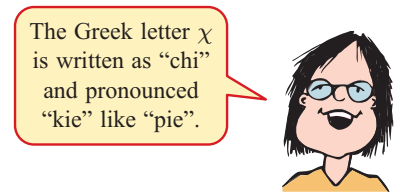
a Let X be the number of questions answered correctly by a person randomly guessing answers. What is the distribution of X?

b Do you think the statistician's model is appropriate? Explain your answer.

c How can we *measure* how appropriate the statistician's model is?

When we observe a variable in a population, we do not always know its distribution. If we *choose* a distribution to model the variable, we will want to know how well the distribution fits our observations.

In this Chapter we study **χ^2 (chi-squared)** tests to assess how appropriate a statistical model is.



A THE χ^2 GOODNESS OF FIT TEST

Suppose Rico rolls a die 60 times and obtains the rolls in the table. Since the relative frequencies or *proportions* of the outcomes are quite different, Rico claims that his die is *biased*.

Number	*Frequency*
1	20
2	10
3	5
4	8
5	7
6	10

Rico could use the binomial distribution to test a hypothesis about a *single* population proportion, as we did in **Chapter 30**.

For example, he could test the hypotheses H_0: $p = \frac{1}{6}$, H_1: $p \neq \frac{1}{6}$ where $p = \text{P}(\text{rolling a } 1)$.

However, these hypotheses do not take into account the probabilities of rolling the other numbers.

If X is the number rolled with Rico's die, then H_0 being true *might* mean that the die is fair.

x	1	2	3	4	5	6
$\text{P}(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

fair

However, there are also infinitely many biased distributions for X for which H_0 is true. An example is shown opposite.

x	1	2	3	4	5	6
$\text{P}(X = x)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{6}$

biased

To test whether his die is biased, Rico needs a test for *multiple* proportions, or a **probability distribution**. He needs the **χ^2 goodness of fit test**.

THE HYPOTHESES

For Rico's die, let p_1 be the probability of rolling a 1, p_2 be the probability of rolling a 2, and so on.

Our null hypothesis H_0 is that the die is fair. If this is the case, then each number is *equally likely* to occur on each roll. We therefore have H_0: $p_1 = \frac{1}{6}$, $p_2 = \frac{1}{6}$, ..., $p_6 = \frac{1}{6}$.

Our alternative hypothesis H_1 is that the die is *not* fair. If this is the case, then at least one outcome has a probability which is different from the others. We therefore have
H_1: at least one of $p_1, p_2, ..., p_6 \neq \frac{1}{6}$.

Consider a scenario with k categories. Let p_i be the population proportion of individuals in category i, where $p_1 + p_2 + ... + p_k = 1$.

The **hypotheses** in a χ^2 goodness of fit test have the form:

H_0: $p_1 = p_{01}$, $p_2 = p_{02}$, ..., and $p_k = p_{0k}$

H_1: at least one of $p_i \neq p_{0i}$

where p_{0i} is the population proportion of category i under the null hypothesis.

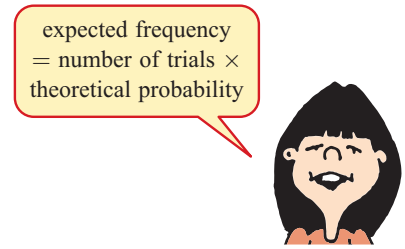
In Rico's experiment, $p_{01}, p_{02}, ..., p_{06}$ are all equal to $\frac{1}{6}$. However, in general the population proportions under H_0 do not have to all be the same.

THE TEST STATISTIC

In our study of probability we calculated the number of times we *expect* an event to occur given its theoretical probability.

For example, if Rico's die was fair then we would expect to see $60 \times \frac{1}{6} = 10$ of each number.

We are therefore interested in how the *observed* frequencies differ from their *expected* values.



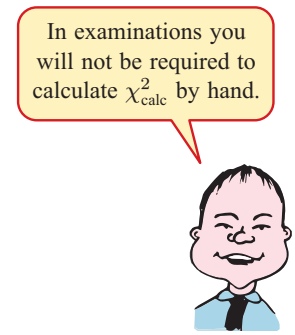
The **test statistic** for a χ^2 goodness of fit test is: $\chi^2_{\text{calc}} = \sum \dfrac{(f_{\text{obs}} - f_{\text{exp}})^2}{f_{\text{exp}}}$

where f_{obs} is an **observed** frequency

f_{exp} is an **expected** frequency.

For the die rolling example, we can calculate the test statistic χ^2_{calc} with the help of a table:

Number	f_{obs}	f_{exp}	$f_{\text{obs}} - f_{\text{exp}}$	$(f_{\text{obs}} - f_{\text{exp}})^2$	$\dfrac{(f_{\text{obs}} - f_{\text{exp}})^2}{f_{\text{exp}}}$
1	20	10	10	100	10
2	10	10	0	0	0
3	5	10	−5	25	2.5
4	8	10	−2	4	0.4
5	7	10	−3	9	0.9
6	10	10	0	0	0
				Total	13.8



So, $\chi^2_{\text{calc}} = 13.8$.

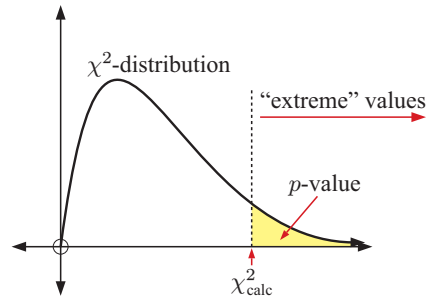
THE p-VALUE

In order to make a decision on whether or not to reject H_0 based on χ^2_{calc}, we need to calculate a ***p*-value** and compare it to the **significance level** α of the test.

For a t-test, we calculated the p-value as the probability of observing a value that is as "extreme" as the test statistic. Depending on the form of the alternative hypothesis H_1, this would correspond to the upper tail, lower tail, or both tails of the t-distribution.

In the goodness of fit test, if H_0 was not true then the difference between the observed frequencies and expected frequencies would be large. Hence the value of χ^2_{calc} would also be large. So, for the χ^2 goodness of fit test:

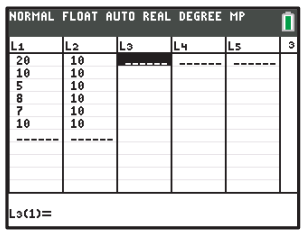
> ***p*-value** = probability of observing a value **greater than or equal to** χ^2_{calc}.

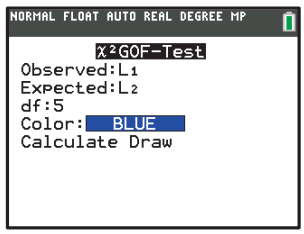


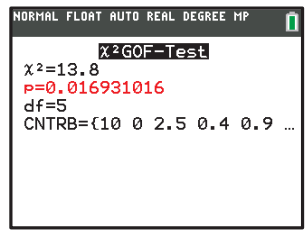
You can use your calculator to evaluate p-values.

GRAPHICS CALCULATOR INSTRUCTIONS

For the die rolling example, we obtain:

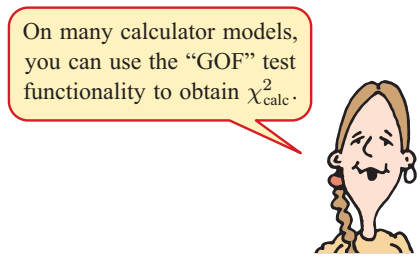






So, the p-value ≈ 0.0169. For a test of H_0 on a significance level of $\alpha = 0.05$, this is enough evidence to reject H_0 in favour of H_1.

We therefore conclude that Rico's die is biased.



DEGREES OF FREEDOM

In finding the p-value with your calculator, you should have noticed that you also need a value for "df". Like in the t-test, this stands for "**degrees of freedom**".

In a χ^2 goodness of fit test, the value of df is the number of values that are *free to vary*.

For example, consider the population proportions of 3 categories p_1, p_2, and p_3. Since they are proportions, they must add to 1. We can therefore write any one of the proportions in terms of the other two. For example, $p_1 = 1 - p_2 - p_3$.

So, only two of the proportions are *free to vary*.

For a χ^2 goodness of fit test, **df = number of categories − 1**.

SUMMARY OF THE χ^2 GOODNESS OF FIT TEST

Step 1: State the **null hypothesis** H_0 and the **alternative hypothesis** H_1.

Step 2: State the **significance level** α.

Step 3: Calculate the value of the **test statistic**: $\chi^2_{calc} = \sum \frac{(f_{obs} - f_{exp})^2}{f_{exp}}$.

Step 4: Use technology to calculate the ***p*-value**, using df = number of categories − 1.

Step 5: Reject H_0 if p-value $\leqslant \alpha$, otherwise accept H_0.

Step 6: Interpret your decision in the context of the problem. Write your conclusion in a sentence.

Notice the similarity between the steps of the goodness of fit test and all of the other hypothesis tests we have seen so far.

Example 1 **Self Tutor**

The table alongside shows the grades received by university students taking a second year Computer Science course.

Grade	*Frequency*
High Distinction	16
Distinction	21
Credit	21
Pass	59
Fail	34
Total	151

In the following semester, a new coordinator is appointed for the course. The new coordinator is concerned by the high number of High Distinctions and Distinctions awarded, and wants to determine if the course should be adjusted.

It is considered usual if 5% of students receive a High Distinction, 10% receive a Distinction, 15% receive a Credit, 40% receive a Pass, and the remaining 30% receive a Fail.

Conduct a χ^2 goodness of fit test to determine whether the course should be adjusted with a 5% level of significance.

Step 1: Let p_1, p_2, p_3, p_4, and p_5 be the population proportions of students who receive a High Distinction, Distinction, Credit, Pass, and Fail respectively.

The hypotheses that should be tested are:

H_0: $p_1 = 0.05$, $p_2 = 0.1$, $p_3 = 0.15$, $p_4 = 0.4$, $p_5 = 0.3$

H_1: at least one of $p_1 \neq 0.05$, $p_2 \neq 0.1$, ..., or $p_5 \neq 0.3$.

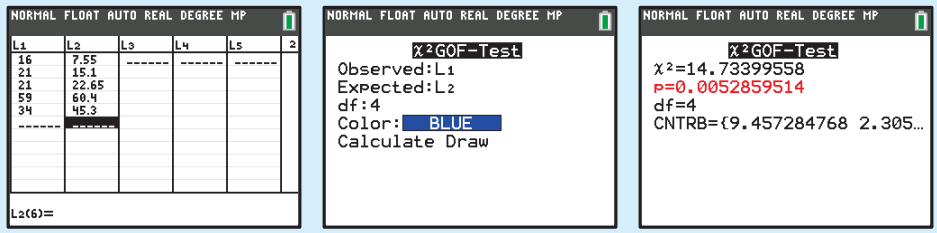
Step 2: The significance level is $\alpha = 0.05$.

Step 3:

Grade	f_{obs}	f_{exp}	$\frac{(f_{obs} - f_{exp})^2}{f_{exp}}$
High Distinction	16	$151 \times 0.05 = 7.55$	≈ 9.4573
Distinction	21	$151 \times 0.1 = 15.1$	≈ 2.3053
Credit	21	$151 \times 0.15 = 22.65$	≈ 0.1202
Pass	59	$151 \times 0.4 = 60.4$	≈ 0.0325
Fail	34	$151 \times 0.3 = 45.3$	≈ 2.8188
		Total	≈ 14.734

So, $\chi^2_{calc} \approx 14.7$.

Step 4: $\text{df} = 5 - 1 = 4$



Using technology, p-value $\approx 0.005\,29$.

Step 5: Since p-value $< 0.05 = \alpha$, we have enough evidence to reject H_0 in favour of H_1 on a 5% level of significance.

Step 6: Since we have accepted H_1, we conclude that the course should be adjusted.

LIMITATIONS

In order for χ^2 to be distributed appropriately, the sample size n must be sufficiently large. Generally, n is sufficiently large if none of the expected frequencies is less than 5.

In cases where there are expected frequencies less than 5, we can **combine** similar categories to get more reliable results.

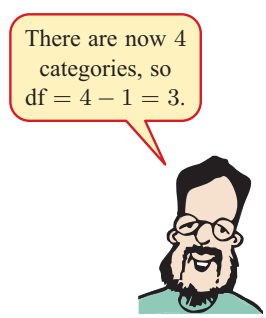
For example, when investigating the number of serious sports injuries treated by a hospital, the following expected frequencies are calculated. Since the expected frequencies of two age groups are less than 5, the data for some categories is combined.

Original table

Age group	*Expected frequency*
10 - 15	14.9
15 - 20	14.1
20 - 30	15.6
30 - 40	4.4
40 - 50	0.9
50+	0.2

Combined

Age group	*Expected frequency*
10 - 15	14.9
15 - 20	14.1
20 - 30	15.6
30+	$4.4 + 0.9 + 0.2 = 5.5$



EXERCISE 31A

1 When a coin was tossed 96 times, 54 heads were observed. A χ^2 goodness of fit test is to be conducted to determine if the coin is biased with a 5% level of significance.

a Write down the hypotheses that should be considered.

b How many tails were observed?

c Calculate the *expected* frequencies of heads and tails assuming the coin is fair.

d Use a table to help you calculate χ^2_{calc}.

e State the number of degrees of freedom.

f Use technology to find the p-value.

g Is there enough evidence to conclude that the coin is biased?

2 In the last election, 54% of voters voted for party A, 30% of voters voted for party B, and the rest voted for party C.

A polling agency conducted a survey asking 300 voters which party they are going to vote for in the upcoming election. The results are shown alongside.

Party	A	B	C
Voters	141	105	54

Conduct a χ^2 goodness of fit test with a 1% level of significance to determine whether there has been a change in the proportions of voters supporting each party since the last election.

3 Brian owns a chocolate café. He wants to start offering ice cream in addition to his chocolate menu items. He initially makes the same amount of chocolate, strawberry, vanilla, honeycomb, and choc-chip ice cream, assuming that the flavours will be equally liked.

The number of sales of each ice cream flavour are shown in the table alongside.

Conduct a χ^2 goodness of fit test with a 10% level of significance to determine whether Brian should change the amount of each ice cream flavour that he makes.

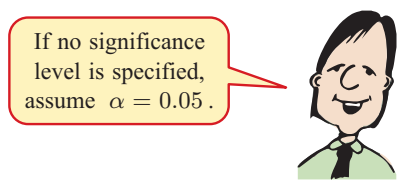
Flavour	*Sales*
chocolate	54
strawberry	48
vanilla	35
honeycomb	28
choc-chip	40
Total	205

4 Of the people living in London in 2001, 71.2% identified as White, 12.1% as Asian/Asian British, 10.9% as Black/Black British, 3.2% as mixed ethnicity, and 2.6% as other ethnicities.

The table alongside shows the number of people recorded for each ethnic group in the 2011 UK Census.

Conduct a χ^2 goodness of fit test to determine whether there was a significant change in London's demographics between 2001 and 2011.

Ethnic group	*Frequency*
White	4 887 435
Asian/Asian British	1 511 546
Black/Black British	1 088 640
Mixed	405 279
Other	281 041
Total	8 173 941



5 In Australia, the NAPLAN tests are used to gauge the literacy and numeracy skills of students. The students are allocated into "bands" based on the score they obtain. The results for the Year 9 students at a particular school, and the national percentages for each band, are shown below.

Band	*School frequency*	*National percentage*
10	5	7.9%
9	9	16.7%
8	55	29.8%
7	53	29.7%
6	23	13.5%
5 and below	5	2.4%
Total	150	100%

a Use the national percentages to calculate the expected frequency for each band.

b Conduct a χ^2 goodness of fit test with a 1% significance level to determine whether there is a substantial difference between the school's results and the rest of the nation.

c Explain why you may wish to combine the "Band 6" and "Band 5 and below" categories.

d Combine the appropriate categories and repeat **b**. Comment on your results.

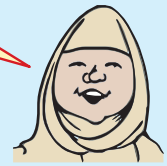
Example 2 Self Tutor

The data alongside shows the number of children born to 150 Indian women in a 5-year period in the 19th century. Test at a 5% level of significance, whether the data is binomial with parameters $n = 5$ and $p = 0.5$.

Number of children	*Number of women*
0	4
1	19
2	41
3	52
4	26
5	8

Step 1: The hypotheses are:
H_0: the data is from B(5, 0.5)
H_1: the data is not from B(5, 0.5).

We do not need to specify each individual probability in the hypothesis because they are implied by the distribution we are considering.



Step 2: The significance level is $\alpha = 0.05$.

Step 3: We calculate the probabilities for each value of x given $X \sim$ B(5, 0.5):

x	0	1	2	3	4	5
$P(X = x)$	0.031 25	0.156 25	0.3125	0.3125	0.156 25	0.031 25

This gives us the following expected frequency table:

Number of children (x)	f_{obs}	f_{exp}
0	4	$150 \times 0.031\,25 = 4.6875$
1	19	$150 \times 0.156\,25 = 23.4375$
2	41	$150 \times 0.3125 = 46.875$
3	52	$150 \times 0.3125 = 46.875$
4	26	$150 \times 0.156\,25 = 23.4375$
5	8	$150 \times 0.031\,25 = 4.6875$

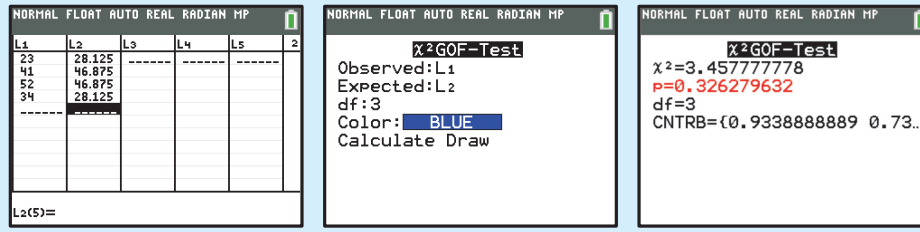
< 5

There are expected frequencies less than 5, so we combine "categories" appropriately:

Number of children (x)	f_{obs}	f_{exp}	$\frac{(f_{obs} - f_{exp})^2}{f_{exp}}$
$\leqslant 1$	23	28.125	≈ 0.9339
2	41	46.875	≈ 0.7363
3	52	46.875	≈ 0.5603
$\geqslant 4$	34	28.125	≈ 1.2272
		Total	≈ 3.4578

So, $\chi^2_{calc} \approx 3.46$.

Step 4: df $= 4 - 1 = 3$



Using technology, p-value ≈ 0.326.

Step 5: Since p-value $> 0.05 = \alpha$, we do not have enough evidence to reject H_0. We therefore accept H_0.

Step 6: Since we have accepted H_0, we conclude that the data is from the binomial distribution B(5, 0.5).

6 Answer the **Opening Problem** on page **830**. Use a 10% level of significance in your hypothesis test.

7 It is claimed that the data in the table comes from a Poisson distribution with mean 5.
Test this claim using a goodness of fit test at a 1% level of significance.

Score	$\leqslant 3$	4	5	6	$\geqslant 7$	*Total*
Frequency	6	9	10	7	4	36

ACTIVITY MENDEL'S DATA

Gregor Johann Mendel (1822 - 1884) is often credited as being "the father of modern genetics".

One of his most important contributions was the use of recessive and dominant genes to describe inherited traits.

During his research, he conducted experiments involving over 28 000 plants, the majority of which were pea plants. The table below shows the results of one such experiment.

Gregor Johann Mendel

Type of pea	*Frequency*
Yellow round seeds	315
Green round seeds	108
Yellow wrinkled seeds	101
Green wrinkled seeds	32
Total	556

According to Mendel's model, he expected the ratio

yellow round : green round : yellow wrinkled : green wrinkled $= 9 : 3 : 3 : 1$

What to do:

1 Calculate the expected frequency of each type of pea.

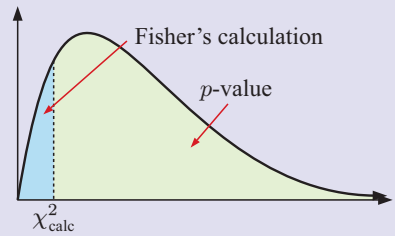
2 In the context of a χ^2 goodness of fit test, show that $\chi^2_{\text{calc}} \approx 0.470$ and p-value ≈ 0.925.

3 What can we conclude based on these results?

THEORY OF KNOWLEDGE

In 1936, **Sir Ronald Aylmer Fisher** analysed Mendel's experiments described in the previous Activity.

He considered the probability that Mendel's results were *consistent* with his expectations, and found that the probability of observing a test statistic *less* than $\chi^2_{\text{calc}} \approx 0.470$ was ≈ 0.075.



In fact, Fisher observed results like this for all of Mendel's experiments. By combining all of Mendel's data, Fisher found that the probability of getting data as good as Mendel's was about 4 in 100 000. Fisher concluded that Mendel had manipulated the data to obtain the results he desired.

1 Does Fisher's finding invalidate the importance of Mendel's contribution to biology and genetics?

Today, the manipulation of data and "data mining" is a major problem in research. Many academic publications only report findings which are "statistically significant", as these findings are more likely to yield more interesting results and lead to further research.

2 Will all claims in academic publications on statistical data necessarily be true? You might want to consider your findings in **Investigation 1** in **Chapter 30**.

3 Discuss the role that statistical interpretation plays in research ethics.

The MMR vaccine controversy was caused by a fraudulent paper published in 1998 which claimed a causal relationship between the MMR vaccine and autism in children. The paper has since been retracted after thorough investigation. However, because of the widespread misconceptions that it has caused, it has been cited as "perhaps the most damaging medical hoax of the last 100 years".

4 Research the details of the MMR vaccine controversy. In particular, consider how the authors collected and used the data cited in the original paper.

5 What other things should a statistician be mindful of when analysing data?

6 Who do you think was responsible for the damage caused by the MMR controversy? Was it the authors of the paper, the media, or the general public?

B ESTIMATING DISTRIBUTION PARAMETERS IN A GOODNESS OF FIT TEST

When the parameters of the distribution of interest are unknown, we must estimate them from the data.

For example, suppose in **Example 2** that we want to test whether the data is binomial with $n = 5$, but without specifying the probability of success p in advance.