

- (ii) Plot these data with the regression line.
- (iii) Plot the residuals against either the independent variable or the fitted values of y .
- (iv) Does the distribution of the residuals meet with the assumptions of regression?
- (v) What do you conclude about the relationship between engine size and fuel economy?

6 Analysing frequency data

This chapter covers tests for analysing frequency distributions of data measured on a nominal scale:

- Tests for association (or independence) between two variables measured on nominal scales
- Tests for goodness of fit between a measured and a theoretical distribution
- Consideration of alternative tests for association, goodness of fit and cases where there are more than two variables

So far we have considered the analysis of data that are measured at least on an ordinal scale (e.g. a ranking of tree leaf condition) or an interval/ratio scale (e.g. strontium-90 activity in becquerels per litre). We have only used nominal variables (i.e. data which are placed in categories) as a way of separating samples (e.g. males from females, urban from suburban, clean from polluted). We now need some way of analysing nominal scale data in their own right. Suppose that when we surveyed trees on polluted and clean sites (Chapter 4) we were unable to give the tree a rank score of condition and were merely able to say whether the tree was in poor condition or not. For each tree we would then have obtained a nominal measurement (good or poor) rather than a score from 1 to 6 (where good condition was 6). This would be the second piece of nominal data, because we also record site type: clean or polluted. How, then, would we analyse these data? We could compare the number of trees in good condition and the number of trees in poor condition on the two types of site using frequency analysis. There are two main types of frequency analysis: **tests for association** and tests for **goodness of fit**. The choice of test depends on whether we are examining the association of one measured frequency distribution with another (e.g. association of tree condition with site type), or whether we are testing one measured frequency distribution against a known theoretical distribution. The term 'theoretical distribution' sounds quite grand, but could simply be the expectation of equal frequencies of events (e.g. sex ratio: equal numbers of males and females).

Associations between frequency distributions

In a survey of public access agreements between local authorities and landowners, two variables were recorded: whether or not each authority had access agreements (in addition to public access via the public right-of-way network); and what type of local

Table 6.1 Data sheet for survey into public access agreements made between local authorities and landowners

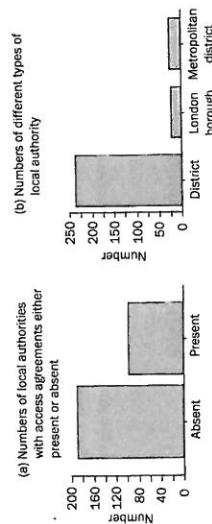
Row (data record number)	Access agreements present (P) or absent (A)	Type of authority: district (D), metropolitan district (M), or London borough (L)
1	A	M
2	P	D
3	A	M
4	A	D
5	P	L
etc.	etc.	etc.
↓	↓	↓
287	P	M

authority it was (i.e. whether it was a district, a metropolitan district or a London borough). An extract from the data sheet is shown in Table 6.1, where each row is an authority, and 287 authorities were surveyed all together (i.e. there were 287 rows).



In this example, there are two frequency distributions (see Figure 6.1): the distribution of the presence or absence of access agreements and the distribution of authority types. Suppose we wish to know whether there is any association between these two distributions, i.e. whether the different authority types differ in their tendency to have additional

access agreements with landowners. Another way of wording this is to say that we are examining whether the distributions are independent of each other (i.e. whether

Figure 6.1 Frequency distribution of types of local authority with access agreements either present or absent ($n = 287$)

having an access agreement is independent of authority type): for this reason, this test is also known as a **test for independence**. The data records (from Table 6.1) can be summarised in the contingency table shown in Worked Example 6.1a. Here the frequency counts for all possible combinations (presence and absence of access agreements for each local authority type) are laid out in a table similar to that used for cross-tabulation in Chapter 2 (Table 2.7).



WORKED EXAMPLE 6.1a Test for association between local authority types and presence or absence of access agreements: contingency table of observed values

	Agreements present	Agreements absent	Row totals
Districts	68	169	237
Metropolitan districts	20	7	27
London boroughs	10	13	23
Column totals	98	189	Grand total 287

The contingency table in Worked Example 6.1a is called a 3×2 contingency table (because there are three rows and two columns). Note that one distribution occupies the columns and the other the rows (it does not matter which way round these go). Since we are asking whether there is an association between the two distributions, the null hypothesis is that there is no association (i.e. that the distributions are independent of each other). We can then calculate what the expected values would be if authority type and the presence or absence of access agreements were distributed independently of each other. Or, worded another way, we calculate the frequencies expected if the distribution of access agreements was similar for each authority type. The first step is to work out the row and column totals (shown in Worked Example 6.1a), and then the expected frequencies are worked out by a simple formula which uses the row and column totals (see Box 6.1).

Note that, in analysing for associations between frequency distributions, most computer programs will accept the data either in a pair of columns, each containing the codes for a single distribution as in Table 6.1, or as a contingency table as in Worked Example 6.1a (for details of data entry for analysis by computer, see Tables B.4 and B.5, Appendix B). Note also that some programs require numerical codes rather than letters.

Worked Example 6.1b shows the calculation of each expected value. Note that, as a check on your arithmetic, the expected values should have the same row and column totals as the observed values, taking into account any rounding errors. We can see that

Box 6.1 Formula for the expected values for testing associations between frequency distributions

The expected values for each cell of a contingency table are calculated as:

$$\text{Expected value (E)} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

This formula can be arrived at from basic probability theory (see Box 2.4). From the data in Worked Example 6.1a, there are 287 authorities in total. Of these, 237 are districts. If the null hypothesis is true, an authority randomly drawn from the 287 would have a $237/287$ chance of being a district. Similarly with access agreements, 98 have access agreements, so there is a $98/287$ chance of randomly drawing an authority with such agreements. If authority type and having access agreements are independent events, then the two probabilities are multiplied to find the probability of being a district and having access agreements:

$$\frac{237}{287} \times \frac{98}{287} = 0.282$$

This is a probability (and thus it lies between 0 and 1); to turn this into an expected number of districts with access agreements, we multiply the probability by the grand total (287):

$$\frac{237}{287} \times \frac{98}{287} \times 287 = 80.927$$

Note that this calculation simplifies to the general formula by cancelling 287 on the top and bottom of the equation:

$$\frac{237 \times 98}{287} = 80.927$$

there are differences between the expected and the observed values in Worked Example 6.1b, for example there are fewer districts with agreements than we would expect if authority type and presence of agreements were independent, and more metropolitan districts with agreements than expected. To identify whether these differences are significant, we need to calculate a test statistic to find whether this difference between the observed and expected frequencies is bigger than we would expect by chance. The statistic, χ^2 (not to be confused with the lower case x^2 used in earlier chapters), is first calculated for each cell of the contingency table and then summed to obtain the total χ^2 value (see Box 6.2). The distribution of this test statistic approximates to a statistical distribution known as chi-square (pronounced 'ky-square', where 'ky' rhymes with 'my', and given the symbol χ^2). In fact, this test is commonly known as a **chi-square test**. Many text books give the test statistic the symbol χ^2 instead of X^2 , but strictly speaking this is not correct because X^2 only approximates to χ^2 (see Sokal and Rohlf, 1995).

For our access agreement data, the total χ^2 value is 23.168 (see Worked Example 6.1c). To see if this is significant, we compare the total χ^2 value to a table of chi-square values (see the shaded area of Table 6.2) at the appropriate degrees of

WORKED EXAMPLE 6.1B Test for association between local authority types and presence or absence of access agreements: table of expected values (E). The observed values (O) are taken from Worked Example 6.1a. The formula for calculating E is (row total $\times$ column total)/grand total (see Box 6.1)

	Agreements present		Agreements absent		Row totals	
	O	Calculating E	O	Calculating E	E	
Districts	68	$(237 \times 98)/287 = 80.927$	169	$(237 \times 189)/287 = 156.073$	237	
Metropolitan districts	20	$(27 \times 98)/287 = 9.220$	7	$(27 \times 189)/287 = 17.780$	27	
London boroughs	10	$(23 \times 98)/287 = 7.854$	13	$(23 \times 189)/287 = 15.146$	23	
Column totals	98		189		Grand total 287	

freedom, given, since we have three rows and two columns, by $df = (3 - 1)(2 - 1) = 2$ (see Box 6.2). At 2 degrees of freedom, our calculated value (23.168) is higher than the table values of chi-square for either $P = 0.05$ (5.99) or $P = 0.01$ (9.21), telling us that we have a highly significant association ($P < 0.01$) between authority type and presence or absence of access agreements.

Having demonstrated that there is a significant association between authority type and the presence of additional access agreements, we now return to the data to interpret these results. This can be done by looking at the individual cell χ^2 components (see

Box 6.2 Formula for the test statistic χ^2

The components of the test statistic are first calculated for each cell of the contingency table:

$$\text{Cell } \chi^2 = \frac{(O - E)^2}{E}$$

where: O is the observed (measured) value and E the expected (calculated) value.

The individual cell χ^2 values are then summed to give the total χ^2 :

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

The total χ^2 value is then compared with a table of chi-square values (χ^2 : an extract from which is shown in Table 6.2) at $df = (r - 1)(c - 1)$, where r is the number of rows and c the number of columns: see Box 6.3 for an explanation of the logic behind this calculation of the degrees of freedom.



WORKED EXAMPLE 6.1C Test for association between local authority types and presence or absence of access agreements: calculating the test statistic (χ^2). Observed and expected values are taken from Worked Examples 6.1a and 6.1b, respectively. The formula for the cell χ^2 values is $(O - E)^2/E$ (see Box 6.2).

	Agreements present		Agreements absent	
	Calculating cell χ^2 values	Cell χ^2	Calculating cell χ^2 values	Cell χ^2
Districts	$(68 - 80.926)^2/80.926 =$	2.0649	$(169 - 156.073)^2/156.073 =$	1.0707
Metropolitan districts	$(20 - 9.220)^2/9.220 =$	12.6058	$(7 - 17.780)^2/17.780 =$	6.5363
London boroughs	$(10 - 7.854)^2/7.854 =$	0.5866	$(13 - 15.146)^2/15.146 =$	0.3042

$$\text{Total } \chi^2 = 2.0649 + 12.6058 + 0.5866 + 1.0707 + 6.5363 + 0.3042 = 23.168$$

This value is then compared with values in a chi-square table (see the shaded area of Table 6.2) at degrees of freedom calculated as:

$$df = (r - 1)(c - 1) = (3 - 1)(2 - 1) = 2$$

where r is the number of rows and c the number of columns.

Worked Example 6.1c. The biggest contributor to the total χ^2 value stems from metropolitan districts (i.e. those covering towns and cities) having more than the expected number of additional access agreements. Interpretation of significant associations may depend on some knowledge of the material under study. In this example, towns and cities may have less access via the existing right-of-way network than do more rural areas,

Table 6.2 Selected values of chi-square (χ^2). Reject the null hypothesis if the calculated value of χ^2 is greater than the table χ^2 value. Shading indicates the critical values for the example referred to in the text. A more comprehensive table of χ^2 values is given in Table D.8 (Appendix D)

df	χ^2	
	P = 0.05	P = 0.01
1	3.841	6.635
2	5.991	9.210
3	7.815	11.345
4	9.488	13.277
5	11.070	15.086

Box 6.3 Logic underlying the calculation of degrees of freedom for a test of association

When calculating the degrees of freedom for a test of association between frequency distributions, we do not use the number of observations, but rather the number of cells. In a 2×3 contingency table, there are $2 \times 3 = 6$ cells with a frequency in each. However, as for other statistical tests, the assumption of contingency table analysis is that each sample is independent. We used row and column totals to calculate our expected values. For this size of table, we can see that with knowledge of the row and column totals, only two of the cells can be randomly drawn; the remaining cells are not independent, but can be calculated from the information that we already have. The degrees of freedom are therefore 2. This logic is extended for any size of table to give the formula:

$$df = (r - 1)(c - 1)$$

where r is the number of rows and c is the number of columns.

	2 x 3 contingency table	
	Present	Absent
District	68	?
Metropolitan	20	?
London	?	?
Total	98	189

The special case of 2 x 2 contingency tables

When the degrees of freedom are 1 (e.g. in a 2×2 contingency table), the calculated χ^2 value is not accurate (it is higher than it should be) and there is an increased chance of rejecting the null hypothesis when in fact it is true (i.e. concluding that there is an association when there is not: a type I error – see Chapter 3). An adjustment, called Yates' correction for continuity, can be made to the calculated value of χ^2 for 2×2 contingency tables. This is not a perfect solution, since it tends to over-compensate and reduce the chances of obtaining a significant result to a greater extent than we would wish (i.e. giving an increased chance of a type II error). One way round this is to calculate the χ^2 value both with and without the correction and compare the significance values obtained. If both agree, then you can be confident about the result. However, if the corrected value is not significant, but the uncorrected analysis gives a significant result, then the results are ambiguous and the true probability lies around 0.05. Under these circumstances, you should be cautious about your interpretation, and consider repeating the investigation using a larger sample size.

In Chapter 4, the condition of trees on two types of site (polluted and clean) was recorded on a scale of 1 to 6 (where 1 was poor condition). Suppose the researcher had instead classified the trees as 'poor' (for those with a score of 1, 2 or 3) or 'good' (for those with a score of 4, 5 or 6). These data can be entered into a 2×2 contingency table (see Worked Example 6.2). To apply the correction factor, simply subtract 0.5 from the absolute value of the $O - E$ component of the original formula before squaring it (see Box 6.4).

Box 6.4 Formula for Yates' correction for continuity

The formula for situations with 1 degree of freedom, using Yates' correction, is:

$$\chi^2 = \sum \frac{(|O - E| - 0.5)^2}{E}$$

where O are the observed values and E are the expected values. $|O - E|$ instructs you to take the absolute value of $O - E$. Remember that the absolute value ignores any negative sign and so is always positive.

**WORKED EXAMPLE 6.2 Test for association between tree condition and type of site**

First we complete a 2×2 contingency table by filling in the observed values. Expected values (in brackets) are calculated as (row total $\times$ column total) / grand total (see Box 6.2):

	Poor	Good	Row totals
Polluted	5 (3)	5 (7)	10
Clean	1 (3)	9 (7)	10
Column totals	6	14	Grand total: 20

Now we can calculate the test statistic, incorporating Yates' correction for continuity, as $\Sigma[(|O - E| - 0.5)^2/E]$ (see Box 6.4):

	Poor	Good
Polluted	$(5 - 3 - 0.5)^2 / 3 = 0.75$	$(5 - 7 - 0.5)^2 / 7 = 0.321$
Clean	$(1 - 3 - 0.5)^2 / 3 = 0.75$	$(9 - 7 - 0.5)^2 / 7 = 0.321$
Total	$\chi^2 = 0.75 + 0.75 + 0.321 + 0.321 = 2.143$	

When we look up the value calculated in Worked Example 6.2 in the chi-square table (Table D.8, Appendix D) at $df = 1$, we find that our value of 2.14 is less than the critical value (3.84), and there is therefore no significant association between the condition of trees and whether or not the site is polluted. If we had not applied the correction factor, our calculated value would have been 3.81 which, although still not significant, is close to the critical value. Since both the corrected and uncorrected values give a non-significant result, we can be confident that there is:

no significant association between the condition of the trees and the polluted nature of the site (corrected $\chi^2 = 2.14$, $df = 1$, $P > 0.05$).

It may seem surprising that these data are not significant using a test for associations between frequency distributions when there was a highly significant effect using a Mann-Whitney U test on the condition scores in Chapter 4 (see Worked Example 4.2). This example serves to illustrate a general point about downgrading data. By converting our scores to nominal categories (i.e. from an original range of scores from 1 to 6 to either 'good' or 'poor'), we have discarded much information from the original data, and thus lost sensitivity in detecting possible differences between the two types of site. In this example, had the researcher felt confident enough to give tree condition a rank score, valuable information would have been saved. In addition, the sample sizes here are quite low. Tests of association are more powerful when the sample size is large and more reliable when the mean expected frequency is large (see the section on the assumptions of chi-square frequency analysis later in this chapter).

Goodness of fit against theoretical distributions

When we wish to compare observed frequencies of one distribution with an external hypothesis, the frequency analysis is known as a goodness of fit test. Imagine we were interested in ascertaining whether the ethnic mix of visitors to a nature reserve was representative of that of the local community. In a survey of 500 visitors, 349 were classified as coming from ethnic group A, 117 as coming from ethnic group B and 34 as coming from ethnic group C (see the first two columns in Worked Example 6.3). These observed frequencies can be compared with expected frequencies based on known percentages of the different ethnic groups from census data: 64% (or 0.64 as a proportion) of the local population are of ethnic group A, 30% (0.3) are B and 6% (0.06) are C. Since this census information has not come from our study, it is the external hypothesis which is used to calculate the expected frequency distribution. The null hypothesis would be that there is no significant difference between the ethnic make-up of the visitors to the reserve and the make-up of the local community. The first step is to use the proportions from the census to generate the expected frequency of the ethnic groups in a sample of 500 (the fourth column in Worked Example 6.3). The expected values are calculated by multiplying the proportion expected from the census data by 500. Note that the sum of the expected values should also be 500.

The test statistic is again χ^2 and is calculated using the same formula as before (Box 6.2). The calculation is shown in stages in the final three columns of Worked Example 6.3. For goodness of fit tests, the degrees of freedom are the number of categories minus 1. In this case there are 3 categories, so the degrees of freedom are 2. We look up our value of χ^2 in a chi-square table at the appropriate degrees of freedom (see

Table D.8, Appendix D). Since our value (10.42) is higher than the table value at both $P = 0.05$ (5.99) and $P = 0.01$ (9.21), we can conclude that:

The frequency of the different ethnic groups at the reserve is not representative of the local population ($\chi^2 = 10.42$, $df = 2$, $P < 0.01$).



WORKED EXAMPLE 6.3 *Goodness of fit test between ethnicity of visitors to a nature reserve and the ethnic mix of the local community. The formula for calculating the expected values is shown in Box 6.1*

Ethnic group	Observed values (O)	Calculating the expected values (E)	E	O - E	(O - E) ²	Cell χ^2 values (O - E) ² / E
A	349	500×0.64	= 320	29	841	2.628
B	117	500×0.30	= 150	-33	1089	7.260
C	34	500×0.06	= 30	4	16	0.533
Total	500		500			$\chi^2 = 10.42$

Looking at the final column of Worked Example 6.3, we can see that ethnic group B makes the greatest contribution to the overall χ^2 value. When we compare the observed and expected values for this group, this suggests that people belonging to ethnic group B are underrepresented at the reserve.

Note that, in analysing for goodness of fit between an expected and observed frequency distribution, many computer programs will accept the data in a pair of columns, one for the expected and one for the observed distribution (for details of data entry for analysis by computer, see Tables B.4 and B.5, Appendix B).

This test can be applied irrespective of the number of categories, e.g. the frequency of ethnic groups A, B, C, D, E could be tested against an expected distribution (with four degrees of freedom: number of categories minus 1). However, the test may be unreliable if the expected frequencies fall below 2 (see the assumptions of chi-square frequency analysis listed later in this chapter). Goodness of fit tests with only two categories (i.e. the degrees of freedom are 1) are sometimes considered to be something of a special case (see below).

The special case of two-category goodness of fit tests

If we have two categories within which to place our data records, the degrees of freedom are 1, and there is a falsely enhanced chance of obtaining a significant result (type I error). Here, we can use a correction for continuity (as we did for 2×2

contingency tables). The formula for calculating χ^2 is exactly the same as that for 2×2 contingency tables (see Box 6.4). Again, the correction overcompensates (gives an increased chance of a type II error) and so the same guidelines apply: the results can be taken with confidence unless the corrected analysis turns an initially significant result into a non-significant one. Under these circumstances it is worth repeating the study with a larger sample size.

A useful application of this two-category test is to test for equality in a sex ratio, with the null hypothesis (external hypothesis) being that there are 50% males and 50% females. Where the expected values are equal for each category, as in this case, the goodness of fit test is sometimes known as a test for **homogeneity** (i.e. a test for evenness). In a survey of invertebrate diversity using pitfall traps, an ecologist catches 91 spiders, only 33 of which are female. It seems that the trapping technique may be biased towards catching males, but is it significant? Of the



91 spiders, if the sex ratio was 1 : 1, we would have expected half (45.5) to be female. The calculation of χ^2 (by Yates' correction formula given in Box 6.4) is given in Worked Example 6.4. Consulting the chi-square table at $df = 1$ (see Table D.8, Appendix D), we find that our calculated value of 6.33 lies between the table values for $P = 0.01$ ($\chi^2 = 6.63$) and $P = 0.05$ ($\chi^2 = 3.84$), therefore we reject the null hypothesis of sex ratio unity and conclude that the sex ratio is significantly biased towards males ($P < 0.05$). Note that if we had not used the correction factor, χ^2 would have been 6.868 which is greater than the table value at $P = 0.01$ (6.63). Thus, the use of the correction alters the confidence we have in our result, but does not effect the main conclusion that:

the sex ratio of spiders caught in pitfall traps is significantly biased towards males (corrected $\chi^2 = 6.33$, $df = 1$, $P < 0.05$).



WORKED EXAMPLE 6.4 *Test for homogeneity of sex ratio in spiders caught using pitfall trapping*

	Observed O	Expected E	O - E	(O - E - 0.5) ²	Cell χ^2 values (O - E - 0.5) ² / E
Males	58	45.5	12.5	144	3.165
Females	33	45.5	-12.5	144	3.165
					$\chi^2 = 6.33$

Goodness of fit against model distributions

Expected frequencies can be derived from a model distribution and compared to observed frequencies using a goodness of fit test. There are many possible

distributions to which data might conform. In Chapter 2 we found that measurement data often conform to the normal distribution, and that an assumption of parametric tests (including the t test) is that data are normally distributed. A few statistics packages use goodness of fit tests to test the data against this assumption. However, it has been suggested that some alternative methods are better – see, for example, Zar (1999) for details.

Other theoretical distributions correspond to the way in which data are dispersed in time or space. For example, counts of plants, animals, people, buses, etc., may show them to be evenly (regularly) spaced, randomly spaced or clumped. If counts of relatively rare items are distributed randomly, then they may follow a distribution called the Poisson distribution (where the variance equals the mean). Regularly spaced items may conform to the binomial distribution (where the variance is small in comparison to the mean) and clumped items to the negative binomial distribution (where the variance is greater than the mean). Perfect distributions of these may be calculated and compared with the observed data using goodness of fit tests. See texts such as Zar (1999) or Sokal and Rohlf (1995) for details.

Assumptions of chi-square frequency analysis

In tests of association or goodness of fit the following assumptions should be satisfied:

- 1 The observed values should be counts (not proportions or percentages – these should be calculated back to their original frequencies before analysis).
- 2 The grand total must represent the number of independent individuals in the study.
- 3 Additionally, calculations of χ^2 (as opposed to some other tests of association or goodness of fit – see the next section) are unreliable when the expected frequencies are low. In practice, when testing at a critical value of $P = 0.05$, the mean of the expected values (the grand total divided by the total number of cells in the table) should be at least 2 for goodness of fit tests (although if there are only two categories, each expected value should be at least 5) and at least 6 for tests of association (see Zar, 1999). If expected values are lower than this, then categories can be combined where appropriate, to increase the numbers of observations (and hence the expected values) in a particular category. If the critical value needs to be at P less than 0.05 (e.g. $P = 0.01$), then higher expected values are required (see Zar, 1999).

Fisher's exact test and G tests

There are alternatives to chi-square tests for frequency analysis, the best known of which are Fisher's exact test and the G test. Fisher's exact test is used only for 2×2 contingency tables where the research design has effectively fixed the values of both

the row and column totals. It is not often appropriate for environmental and geographical data.

G tests can be used as an alternative to chi-square tests for most purposes, indeed some authors recommend G tests in preference to chi-square tests, although this has yet to become established in either textbooks or journal articles. G tests may perform slightly better than chi-square tests, especially if some of the expected frequencies are low (see point 3 under the assumptions of chi-square frequency analysis above). See texts such as Sokal and Rohlf (1995) for further details.

Testing for associations between more than two distributions

It may be that the ecologist who collected the data for tree condition on polluted and clean sites in Worked Example 6.2 decided to repeat the survey on additional polluted and clean sites. It could then be tempting to combine the new data with the existing data into one contingency table. However, this would assume that the distributions of tree condition and site type were drawn from the same underlying frequency distributions on both occasions. It is only permissible to combine the data for the two (or more) occasions if the two (or more) tests are first run individually and then tested for consistency (see Zar, 1999, for details). If they are shown to be consistent, the data can then be combined and analysed using either a chi-square or a G test.

If a third variable in addition to site type and tree condition had been measured (e.g. different species of tree), then the contingency table could be viewed as being three-dimensional. This type of research design can be analysed using a G test (see Sokal and Rohlf, 1995, for details).

Summary

By now you should be able to:

- recognise when frequency data are appropriate for contingency table analysis (for association between two variables measured on nominal scales), calculate the expected values and work out the degrees of freedom;
- calculate the test statistic (χ^2) for association between frequency distributions from the observed and expected values and assess its significance;
- recognise when data are appropriate for a goodness of fit test (i.e. to test frequencies against an expected external distribution), calculate expected frequencies from probabilities and work out the degrees of freedom;
- calculate χ^2 for goodness of fit from the observed and expected values and assess its significance;
- interpret the direction of any significant χ^2 (whether for association or goodness of fit) by looking at the individual χ^2 components.
- be aware that other techniques are appropriate for the association between more than two frequency distributions.

Questions

- 6.1 As part of a study into the effects of woodland management on bird distribution, 36 woods were surveyed for the presence (P) or absence (A) of long-tailed tits and classified as one of three woodland types: pre-thicket (PT), thicket (T) or high forest (HF). The following data were obtained:

Wood number	Wood type	Presence or absence of long-tailed tits	Wood number	Wood type	Presence or absence of long-tailed tits
1	PT	P	19	HF	A
2	HF	P	20	T	P
3	PT	A	21	PT	P
4	PT	P	22	PT	P
5	HF	A	23	HF	A
6	T	P	24	T	P
7	HF	A	25	PT	A
8	T	P	26	PT	P
9	HF	P	27	HF	P
10	T	P	28	T	P
11	PT	A	29	PT	A
12	T	A	30	PT	P
13	T	P	31	T	P
14	HF	P	32	HF	A
15	PT	A	33	T	P
16	T	P	34	HF	A
17	HF	A	35	PT	P
18	HF	P	36	HF	A

- (i) Arrange the data in a 3×2 contingency table (including row, column and grand totals).
(ii) Calculate the expected frequencies, place them in a table and check that the row, column and grand totals agree with those for the table of observed values.
(iii) Calculate the individual χ^2 components and place them in a table.
(iv) Calculate the total χ^2 , df and P values.
(v) What do you conclude regarding the distribution of long-tailed tits in the different types of woodland?
- 6.2 Trees were surveyed for the presence of lichen. Each lichen found was recorded as either occurring on the north, east, south or west quadrant of the tree: 36 lichens were found on north-facing quadrants, 50 on east-facing quadrants, 29 on south-facing quadrants and 45 on west-facing quadrants.
- (i) For lichens occurring on the north, east, south and west sides of trees, use the external hypothesis that lichen orientation is random, and display in a table the observed and expected number of lichens and the component χ^2 values in a goodness of fit test.
(ii) Calculate the total χ^2 , df and P values to complete the goodness of fit test.
(iii) What can you conclude about the distribution of lichens on trees?

7 Differences between more than two samples

To investigate potential differences between more than two samples we need an extension of the tests described in Chapter 4. The tests are classified not only on their requirement for the data to be normally distributed, but also according to whether we are looking at more than one independent variable at a time. This chapter covers:

- One-way analysis of variance for normally distributed data
- Kruskal-Wallis one-way analysis of variance using ranks for data measured at least on an ordinal scale
- Two-way analysis of variance for normally distributed data
- Two-way analysis of variance using ranks for data measured at least on an ordinal scale
- Consideration of alternative models of analysis of variance for more complicated experimental and survey designs

In Chapter 4 we compared the means (or medians) of two samples using a t test (for normally distributed data) or a Mann-Whitney U test (for ordinal or non-normal data). However, many experiments or surveys involve the comparison of three or more samples. For example, in a survey of household energy bills, we might place houses in three categories depending on the type of insulation they have – loft insulation, double-glazing or minimal insulation. To investigate the effect of insulation type on household energy bill, it might seem logical to perform three t tests (comparing 'loft' with 'double-glazing', 'loft' with 'minimal' and 'double-glazing' with 'minimal'). This would certainly be possible. However, if we had more samples, the computation would become very long-winded and time-consuming (for seven samples, for example, we would need 21 separate tests). There is a more serious objection to conducting multiple t tests: that is, when we

compute several tests we increase our chances of obtaining a significant result by chance alone. Remember that the critical probability value is 0.05. This is the value at which the null hypothesis is rejected because there is only a 5% probability that the result occurred by chance. We can turn this around and say that on 5% of occasions the null hypothesis is rejected when in fact it is true (this is a type I error: see Chapter 3). It now becomes clear that if we do 20 tests, at

