This chapter looks at two techniques for doing this. We begin with the simple case of two categorical variables and discover the chi-square statistic (which we're not really discovering because we've unwittingly come across it countless times before). We then extend this model to look at relationships between several categorical variables.

18.4. Theory of analysing categorical data ①

We will begin by looking at the simplest situation that you could encounter; that is, analysing two categorical variables. If we want to look at the relationship between two categorical variables then we can't use the mean or any similar statistic because we don't have any variables that have been measured continuously. Trying to calculate the mean of a categorical variable is completely meaningless because the numeric values you attach to different categories are arbitrary, and the mean of those numeric values will depend on how many members each category has. Therefore, when we've measured only categorical variables, we analyse frequencies. That is, we analyse the number of things that fall into each combination of categories. If we take an example, a researcher was interested in whether animals could be trained to line-dance. He took 200 cats and tried to train them to line-dance by giving them either food or affection as a reward for dance-like behaviour. At the end of the week he counted how many animals could line-dance and how many could not. There are two categorical variables here: **Training** (the animal was trained using either food or affection, not both) and **Dance** (the animal either learnt to line-dance or it did not). By combining categories, we end up with four different categories. All we then need to do is to count how many cats fall into each category. We can tabulate these frequencies as in Table 18.1 (which shows the data for this example), and this is known as a **contingency table**.

Table 18.1 Contingency table showing how many cats will line-dance after being trained with different rewards

		Training		
		Food as reward	***Affection as reward***	***Total***
Could they dance?	Yes	28	48	76
	No	10	114	124
	Total	38	162	200

18.4.1. Pearson's chi-square test ①

If we want to see whether there's a relationship between two categorical variables (i.e., does the number of cats that line-dance relate to the type of training used?) we can use the Pearson's **chi-square test** (Fisher, 1922; Pearson, 1900). This is an extremely elegant statistic based on the simple idea of comparing the frequencies you observe in certain categories to the frequencies you might expect to get in those categories by chance. All the way back in Chapters 2, 7 and 10 we saw that if we fit a model to any set of data we can evaluate that model using a very simple equation (or some variant of it):

$$\text{Deviation} = \sum(\text{observed} - \text{model})^2$$

This equation was the basis of our sums of squares in regression and ANOVA. Now, when we have categorical data we can use the same equation. There is a slight variation in that we divide by the model scores as well, which is actually much the same process as dividing the sum of squares by the degrees of freedom in ANOVA. So, basically, what we're doing is standardizing the deviation for each observation. If we add all of these standardized deviations together the resulting statistic is Pearson's chi-square (χ^2) given by:

$$\chi^2 = \sum \frac{\left(\text{observed}_{ij} - \text{model}_{ij}\right)^2}{\text{model}_{ij}} \tag{18.1}$$

in which i represents the rows in the contingency table and j represents the columns. The observed data are, obviously, the frequencies in Table 18.1, but we need to work out what the model is. In ANOVA the model we use is group means, but as I've mentioned we can't work with means when we have only categorical variables so we work with frequencies instead. Therefore, we use 'expected frequencies'. One way to estimate the expected frequencies would be to say 'well, we've got 200 cats in total, and four categories, so the expected value is simply 200/4 = 50'. This would be fine if, for example, we had the same number of cats that had affection as a reward and food as a reward; however, we didn't: 38 got food and 162 got affection as a reward. Likewise there are not equal numbers that could and couldn't dance. To take account of this, we calculate expected frequencies for each of the cells in the table (in this case there are four cells) and we use the column and row totals for a particular cell to calculate the expected value:

$$Model_{ij} = E_{ij} = \frac{\text{row total}_i \quad \text{column total}_j}{n}$$

where n is simply the total number of observations (in this case 200). We can calculate these expected frequencies for the four cells within our table (row total and column total are abbreviated to RT and CT respectively):

$$\text{Model}_{\text{Food, Yes}} = \frac{\text{RT}_{\text{Yes}} \quad \text{CT}_{\text{Food}}}{n} = \frac{76 \quad 38}{200} = 14.44$$

$$\text{Model}_{\text{Food, No}} = \frac{\text{RT}_{\text{No}} \quad \text{CT}_{\text{Food}}}{n} = \frac{124 \quad 38}{200} = 23.56$$

$$\text{Model}_{\text{Affection, Yes}} = \frac{\text{RT}_{\text{Yes}} \quad \text{CT}_{\text{Affection}}}{n} = \frac{76 \quad 162}{200} = 61.56$$

$$\text{Model}_{\text{Affection, No}} = \frac{\text{RT}_{\text{No}} \quad \text{CT}_{\text{Affection}}}{n} = \frac{124 \quad 162}{200} = 100.44$$

Given that we now have these model values, all we need to do is take each value in each cell of our data table, subtract from it the corresponding model value, square the result, and then divide by the corresponding model value. Once we've done this for each cell in the table, we just add them up!

$$\begin{aligned}\chi^2 &= \frac{(28-14.44)^2}{14.44} + \frac{(10-23.56)^2}{23.56} + \frac{(48-61.56)^2}{61.56} + \frac{(114-100.44)^2}{100.44} \\ &= \frac{(13.56)^2}{14.44} + \frac{(-13.56)^2}{23.56} + \frac{(-13.568)^2}{61.56} + \frac{(13.56)^2}{100.44} \\ &= 12.73 + 7.80 + 2.99 + 1.83 \\ &= 25.35\end{aligned}$$

This statistic can then be checked against a distribution with known properties. All we need to know is the degrees of freedom and these are calculated as $(r - 1)(c - 1)$ in which r is the number of rows and c is the number of columns. Another way to think of it is the number of levels of each variable minus one multiplied. In this case we get $df = (2 - 1)(2 - 1) = 1$. If you were doing the test by hand, you would find a critical value for the chi-square distribution with $df = 1$ and if the observed value was bigger than this critical value you would say that there was a significant relationship between the two variables. These critical values are produced in the Appendix, and for $df = 1$ the critical values are 3.84 ($p = .05$) and 6.63 ($p = .01$), and so because the observed chi-square is bigger than these values it is significant at $p < .01$. However, if you use **R**, it will simply produce an estimate of the precise probability of obtaining a chi-square statistic at least as big as (in this case) 25.35 if there were no association in the population between the variables.

18.4.2. Fisher's exact test ①

There is one problem with the chi-square test, which is that the sampling distribution of the test statistic has an *approximate* chi-square distribution. The larger the sample is, the better this approximation becomes, and in large samples the approximation is good enough to not worry about the fact that it is an approximation. However, in small samples the approximation is not good enough, making significance tests of the chi-square distribution inaccurate. This is why you often read that to use the chi-square test the expected frequencies in each cell must be greater than 5 (see section 18.5). When the expected frequencies are greater than 5, the sampling distribution is probably close enough to a perfect chi-square distribution for us not to worry. However, when the expected frequencies are too low, it probably means that the sample size is too small and that the sampling distribution of the test statistic is too deviant from a chi-square distribution to be of any use.

Fisher came up with a method for computing the exact probability of the chi-square statistic that is accurate when sample sizes are small. This method is called **Fisher's exact test** (Fisher, 1922) even though it's not so much a test as a way of computing the exact probability of the chi-square statistic. This procedure is normally used on 2×2 contingency tables (i.e., two variables each with two options) and with small samples. However, it can be used on larger contingency tables and with large samples, but on larger contingency tables it becomes computationally intensive and you might find **R** taking a long time to give you an answer. In large samples there is really no point because it was designed to overcome the problem of small samples, so you don't need to use it when samples are large.

18.4.3. The likelihood ratio ②

An alternative to Pearson's chi-square is the likelihood ratio statistic, which is based on maximum-likelihood theory. The general idea behind this theory is that you collect some data and create a model for which the probability of obtaining the observed set of data is maximized, then you compare this model to the probability of obtaining those data under the null hypothesis. The resulting statistic is, therefore, based on comparing observed frequencies with those predicted by the model:

$$L\chi^2 = 2\sum \text{observed}_{ij} \ln\left(\frac{\text{observed}_{ij}}{\text{model}_{ij}}\right) \tag{18.2}$$

in which i and j are the rows and columns of the contingency table and ln is the natural logarithm (this is the standard mathematical function that we came across in Chapter 8, and you can find it on your calculator, usually labelled as ln or $\log_e$). Using the same model and observed values as in the previous section, this would give us:

$$\begin{aligned} L\chi^2 &= 2\left[28\times\ln\left(\frac{28}{14.44}\right)+10\times\ln\left(\frac{10}{23.56}\right)+48\times\ln\left(\frac{48}{61.56}\right)+114\times\ln\left(\frac{114}{100.44}\right)\right] \\ &= 2\left[28\times 0.662+10\times -0.857+48\times -0.249+114\times 0.127\right] \\ &= 2\left[18.54-8.57-11.94+14.44\right] \\ &= 24.94 \end{aligned}$$

As with Pearson's chi-square, this statistic has a chi-square distribution with the same degrees of freedom (in this case 1). As such, it is tested in the same way: we could look up the critical value of chi-square for the number of degrees of freedom that we have. As before, the value we have here will be significant because it is bigger than the critical values of 3.84 ($p = .05$) and 6.63 ($p = .01$). For large samples this statistic will be roughly the same as Pearson's chi-square, but is preferred when samples are small.

18.4.4. Yates's correction ②

When you have a 2 × 2 contingency table (i.e., two categorical variables each with two categories) then Pearson's chi-square tends to produce significance values that are too small (in other words, it tends to make a Type I error). Therefore, Yates suggested a correction to the Pearson formula (usually referred to as **Yates's continuity correction**). The basic idea is that when you calculate the deviation from the model (the $\text{observed}_{ij} - \text{model}_{ij}$ in equation (18.1)) you subtract 0.5 from the absolute value of this deviation before you square it. In plain English, this means you calculate the deviation, ignore whether it is positive or negative, subtract 0.5 from the value and then square it. Pearson's equation then becomes:

$$\chi^2 = \sum \frac{(|\text{observed}_{ij} - \text{model}_{ij}| - 0.5)^2}{\text{model}_{ij}}$$

For the data in our example this just translates into :

$$\begin{aligned} \chi^2 &= \frac{(13.56-0.5)^2}{14.44}+\frac{(13.56-0.5)^2}{23.56}+\frac{(13.56-0.5)^2}{61.56}+\frac{(13.56-0.5)^2}{100.44} \\ &= 11.81+7.24+2.77+1.70 \\ &= 23.52 \end{aligned}$$

The key thing to note is that it lowers the value of the chi-square statistic and, therefore, makes it less significant. Although this seems like a nice solution to the problem there is a fair bit of evidence that this overcorrects and produces chi-square values that are too small! Howell (2006) provides an excellent discussion of the problem with Yates's correction for continuity, if you're interested; all I will say is that, although it's worth knowing about, it's probably best ignored.

18.5. Assumptions of the chi-square test ①

It should be obvious that the chi-square test does not rely on assumptions such as having continuous normally distributed data like most of the other tests in this book (categorical data cannot be normally distributed because they aren't continuous). However, the chi-square test still has two important assumptions:

- Pretty much all of the tests we have encountered in this book have made an assumption about the independence of data and the chi-square test is no exception. For the chi-square test to be meaningful it is imperative that each person, item or entity contributes to only one cell of the contingency table. Therefore, you cannot use a chi-square test on a repeated-measures design (e.g., if we had trained some cats with food to see if they would dance and then trained the same cats with affection to see if they would dance, we couldn't analyse the resulting data with Pearson's chi-square test).
- The expected frequencies should be greater than 5. Although it is acceptable in larger contingency tables to have up to 20% of expected frequencies below 5, the result is a loss of statistical power (so the test may fail to detect a genuine effect). Even in larger contingency tables no expected frequencies should be below 1. Howell (2006) gives a nice explanation of why violating this assumption creates problems. If you find yourself in this situation consider using Fisher's exact test (section 18.4.2).

Finally, although it's not an assumption, it seems fitting to mention in a section in which a gloomy and foreboding tone is being used that proportionately small differences in cell frequencies can result in statistically significant associations between variables if the sample is large enough (although it might need to be very large indeed). Therefore, we must look at row and column percentages to interpret any effects we get. These percentages will reflect the patterns of data far better than the frequencies themselves (because these frequencies will be dependent on the sample sizes in different categories.

18.6. Doing the chi-square test using R ①

There are two ways in which categorical data can be entered: enter the raw scores, or enter weighted cases. We'll look at both in turn.

18.6.1. Entering data: raw scores ①

If we input the raw scores, it means that every row of the data editor represents each entity about which we have data (in this example, each row represents a cat). So, you would create two codings (**Training** and **Dance**). Training would contain two values – one to indicate food was a reward, and one to indicate affection was a reward. Dance would contain Yes, or No, depending on whether the cat danced. There were 200 cats in all and so there are 200 rows of data. This is how the data are stored in **cats.dat**. You can load this data file by setting your working directory to the location of the file (see section 3.4.4) and executing:

```
catData<-read.delim("cats.dat", header = TRUE)
```

The resulting data look like this (heavily edited because you don't need to see all 200 rows to get the idea):