CHAPTER 4: COUNTS and PROPORTIONS

MOTIVATION

Humans seem to enjoy categorising all manner of things, most especially fellow humans. Often we categorise on several characteristics simultaneously:

“she is young/old [age classification] and she suffers from diabetes/arthritis/high blood pressure [illness classification] and she requires insulin/physiotherapy/oral medication [treatment classification].”

Given enough classification criteria, we could presumably identify a particular subject to any degree of accuracy. But some criteria are likely to be redundant - they don’t add any new information. It would be nice to have methods for deciding which criteria are independent of others. It turns out that, for a range of problems, decisions concerning independence of classifications are identical to decisions concerning associations between nominal scaled variables (is exposure to heavy metals associated with malignant disease?), and differences in proportions (is the proportion of males who smoke greater than the proportion of females who smoke?). As epidemiological studies often lead to data in the form of cross-classified counts, we will take this opportunity to enlarge on our previous brief discussions of epidemiological study design.

The simplest analytical methods for dealing with categorised data involve probability models based on the Chi-square distribution, and this chapter revolves around this most useful of distributions.
§ 4.1 INTRODUCTION

I remarked in Chapter 1 that the measurement scale of a variable is an important consideration when choosing a method of analysis. In Chapters 3 we saw that when an outcome variable is measured on an interval or ratio scale and we wish to determine if that variable’s mean differs between two groups (if you prefer, is there an association between the mean of the variable and group membership?) we use a $t$ test. Note that the second variable, denoting group membership, is measured on a nominal scale. In Chapter 5 we will discuss linear correlation, a method for determining if an association exists between two variables, both of which are interval- or ratio-scaled. This course will not cover the so-called non-parametric methods for dealing with variables measured on an ordinal scale (giving rise to data in the form of ranks). Such methods are discussed in standard statistical texts to which you can refer if the need arises.

In this chapter we will explore how to tackle data that are measured on a nominal scale. Nominal (categorical) data naturally lend themselves to tests of association between variables and tests of equality of counts, rates and proportions.

§ 4.1.1 Organising Categorical Data

With nominal scaled variables, data are most often presented in the form of a table. Tables have $r$ rows and $c$ columns. A table which has more than one row and more than one column [$r > 1$ and $c > 1$] is called a contingency table or a cross-classification table. The rows represent categories of one variable (e.g., male and female for the sex variable) and the columns represent categories of a second variable. If either $r = 1$ or $c = 1$ then the table is dealing with only one variable. A unique row-column location is called a cell of the table. The particular entry in the $ij$th cell of the table is the frequency of occurrence of the event defined by the $i$th category of the row variable and the $j$th category of the column variable. All this sounds much more complicated than it really is. Just draw a few diagrams for yourself.

Fig 4.1 A 6 row by 5 column contingency table

In other words, a simple table with either one row only or one column only is just a way of presenting a frequency distribution of one variable. A cross-classification or contingency table (Fig 4.1) is just another way of showing a joint frequency.
distribution of 2 variables. It is notionally easy to extend contingency tables to reflect cross-classification of data on 3, 4 or more variables, but showing this on 2-dimensional paper gets a bit challenging! The notation for a three variable contingency table with \( r \) rows, \( c \) columns and say, \( d \) categories of the third variable would be \( R \times C \times D \).

Don’t confuse the number of rows and columns with the number of variables you are dealing with.

Example 4.1:

An 8\( \times \)7 table and a 2\( \times \)2 table are both dealing with two variables; it is just that each variable in the first table has more categories than the variables in the second table. The first table might be a cross-classification of students into faculty (8 possibilities) by social class (7 possibilities), whereas the second table might be a cross-classification of patients into smokers (yes/no) by heart disease (yes/no).

§ 4.2 Epidemiological Study Designs and the 2x2 Table

The 2\( \times \)2 table provides a simple model for discussing study designs that epidemiologists use to investigate the relationship between disease occurrence and risk factors. We will consider 3 designs and appropriate measures of effect.

§ 4.2.1 The Cross-Sectional Study

The data for such a study (sometimes also called a prevalence study) arise from a simple random sample of \( N \) subjects in the population, with no prior attention to disease or risk factor status. After selection, each subject is then cross-classified by disease and exposure status. The data are represented in a 2\( \times \)2 table:

<table>
<thead>
<tr>
<th></th>
<th>disease</th>
<th>no disease</th>
<th>( a+b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>exposed</td>
<td>a</td>
<td>b</td>
<td>( a+b )</td>
</tr>
<tr>
<td>non-exposed</td>
<td>c</td>
<td>d</td>
<td>( c+d )</td>
</tr>
<tr>
<td></td>
<td>a+c</td>
<td>b+d</td>
<td>( N = a+b+c+d )</td>
</tr>
</tbody>
</table>

where, for example, \( c \) is the number of subjects in the sample who, at the time of selection, had not been exposed to the risk factor but had the disease. The terms \( a+b \), \( c+d \), \( a+c \), and \( b+d \) are called marginal totals because (not unreasonably) they are found in the margins of the table.

What measure of effect might we use to summarise the data? A good choice might be a comparison of the proportions of those exposed and unexposed who have the disease, that is, the respective prevalences of disease in both groups. These are: \( a/(a+b) \) for the exposed, and \( c/(c+d) \) for the unexposed, so a prevalence difference could be constructed as \( [a/(a+b)] - [c/(c+d)] \). Alternatively, a prevalence ratio would
be: \[\frac{a}{a+b} / \frac{c}{c+d}\]. Inferential interest might centre on whether, in the population, the prevalence difference was equal to 0, or, equivalently, whether the prevalence ratio was equal to 1. A worked example is shown in Example 4.4 later in the chapter.

§ 4.2.2 Fixed Cohort Study

Instead of taking a single random sample of the population, let us now imagine we take 2 samples: one purposively of subjects who are known \textit{a priori} to have been exposed to the risk factor and another of those who are known not to have been exposed. We then follow these two sample for a fixed length of time and count how many in each sample develop the disease of interest. This is a prospective, fixed cohort study. (For an example of a “dynamic” cohort study, wherein subjects may be followed for variable lengths of time, see Example 3.3.) The 2 \( \times \) 2 table for such a study looks just like that of the cross-sectional study, but now we can calculate the cumulative incidences of disease, rather than just the prevalences, in the exposed and unexposed groups. For example, the ratio: \[\frac{a}{a+b} / \frac{c}{c+d}\] would now be termed the \textit{cumulative incidence ratio} (also known, somewhat colloquially, as the “rate ratio” or the “relative risk”). Inferential interest would centre on whether, in the population, the cumulative incidence ratio was equal to 1.

§ 4.2.3 Case-Control Study

What if we did not have the time, resources or patience to wait for a sufficient number of exposed and unexposed subjects to cooperate with the study and develop disease such that a meaningful comparison could be made? This might especially be the case when the disease is relatively rare. We might then decide that a case-control design would be more appropriate. Again we take two samples, but this time the samples are of known diseased and non-diseased subjects. Indeed, if the disease is rare we might well take \textit{all} those from the defined population with disease (cases) and sample only from the non-diseased fraction of the population (controls). Optimally, the controls are a random sample of those without disease, but who, if they \textit{did} develop disease, would be detected by the study surveillance procedure and deemed as eligible to be a case. (The definition of controls and issues arising from their selection is the basis of much discussion among epidemiologists. One of the easiest ways to invalidate a case-control study is to introduce bias by selecting inappropriate controls.) Anyway, rather than \textit{prospectively} follow exposed and non-exposed subjects as in a cohort study, we now \textit{retrospectively} ascertain how many of the cases and controls \textit{had been} exposed or not exposed to the risk factor. The 2 \( \times \) 2 table looks just as it does for the cross-sectional study and the fixed cohort study, but we can no longer directly calculate the prevalence or incidence of disease from it. This is self-evident: the case-control design allows \textit{us} to \textit{choose} right from the start the number of diseased and non-diseased subjects. Rather, epidemiologists use a measure of effect called the \textit{odds ratio} to summarise the data. More properly, the odds ratio for a case-control study is an \textit{exposure odds ratio}. It works like this: for the cases, the \textit{odds} of being exposed is the ratio of the proportion of those exposed to the proportion of those not
exposed, this is \( \frac{a/(a+c)}{c/(a+c)} = \frac{a}{c} \). For the controls, the odds of being exposed is equal to \( \frac{b/(b+d)}{d/(b+d)} = \frac{b}{d} \). Therefore:

\[
\text{odds ratio} = \frac{a/c}{b/d} = \frac{ad}{bc} \quad \text{E4.1}
\]

Looking at the structure of the 2x2 table, it is easy to see why this is also sometimes called the cross-product ratio. One can show that, if disease is rare in both exposed and unexposed, the odds ratio from a case-control study approximates the cumulative incidence ratio (“relative risk”) from a fixed cohort study. The exposure odds ratio is itself a perfectly respectable measure of effect, but for most people it is somewhat more difficult to think in terms of the unfamiliar odds ratio metric than a relative risk metric, so people usually automatically interpret the odds ratio as a “relative risk”. One should remember that this is usually only appropriate if the disease is rare, but since this is often the reason a case-control design is chosen, it is often satisfactory. [Epidemiological texts discuss certain case-control designs that avoid the need for the “rare disease assumption”.] Anyway, for a case-control study, inference will centre on whether, in the population, the odds ratio is equal to 1.

**Example 4.2**

The following data show results from a multi-centre case-control study of glioma (malignant brain tumour) and its possible relation to cigarette smoking. [Note that this table, produced from the raw data by a computer program, has interchanged the definitions of rows and columns compared with the set-up in §4.2.1. This is perfectly all-right, but demonstrates that one should look carefully at a table before interpreting it. It is a good rule of thumb to ensure that exposed cases occupy the northwest [\( \Box \)] corner of a 2x2 table.]

<table>
<thead>
<tr>
<th>ever smoked cigarettes</th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>679</td>
<td>522</td>
<td>1201</td>
</tr>
<tr>
<td>Cases</td>
<td>1375</td>
<td>1119</td>
<td>2494</td>
</tr>
<tr>
<td>Controls</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2054</td>
<td>1641</td>
<td>3695</td>
</tr>
</tbody>
</table>

The odds ratio is \( \frac{679 \times 1119}{522 \times 1375} = 1.06 \). That is, the odds of exposure to smoking among those with glioma is 1.06 times the odds of exposure among controls. It is known that glioma is a rather uncommon disease (about 5 gliomas per 100,000 persons per year in South Australia), so we might say that the risk of glioma in smokers is 1.06 times that in non-smokers. This is a small increase in risk – from these data it does not appear that smoking is much of a risk factor for glioma.
§ 4.2.4  Inference and Epidemiological Study Designs

We have seen in Chapter 3 that there are two common approaches to inference. One approach entails the construction of a confidence interval and inspection of this interval to see which values of the population parameter are consistent, in a probabilistic sense, with the observed data of the sample. The construction of confidence intervals for epidemiological parameters such as the incidence ratio and odds ratio is somewhat beyond the scope of this introductory biostatistics text (as a consequence of non-Normal sampling distributions, calculation of confidence intervals for risk ratios and odds ratios involves transformations and numerical approximations), but you can consult any analytical epidemiological text for details. Just to illustrate the approach, let us look further at Example 4.2, the case-control study.

Example 4.2 continued

A common type of case-control study is the unmatched design. In such a study, each control is chosen without reference to characteristics (for example, age and sex) of individual cases. For such an unmatched case-control study it can be shown that, even though the sampling distribution of the odds ratio is not Normal, the sampling distribution of the natural logarithm of the odds ratio is approximately Normally distributed with mean equal to \( \log(OR) \) and standard error approximately equal to \( \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \). Therefore an approximate 95% confidence interval for the \( \log(OR) \) is given by:

\[
\log (OR) \pm 1.96 \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}
\]

And so, the 95% confidence interval for the OR itself is achieved by exponentiating:

\[
e^{\log (OR) \pm 1.96 \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}
\]

If we plug in the values from our 2x2 table, we obtain the interval \([0.92, 1.22]\). Note that this interval is asymmetric about the point estimate of 1.06 (as expected for the estimate of a ratio), and that it includes the value of the OR specified by \( H_0 \), that is, 1. Therefore, we have insufficient evidence to reject the Null hypothesis which states that there is no increased risk of glioma among smokers.

The other approach to inference constructs a test of opposing hypotheses and yields a “P value” that summarises how unusual an observed result is, assuming the Null hypothesis is true. In epidemiological studies, whether we are interested in prevalence differences or ratios in a cross-sectional study, incidence ratios in a cohort study, or odds ratios in a case-control study, the Null hypothesis can generically be stated as: classification of subjects in rows is independent of classification in columns. It turns out that a single, simple hypothesis test — the chi-square test — can be used to test this Null hypothesis and we now turn our attention to developing such a test.
§ 4.3 ALL ABOUT CHI-SQUARE

§ 4.3.1 Definitions and Distributions

For a start, note that “chi” is pronounced “kye” (rhymes with “bye”). It is the 22nd (lower case) letter of the Greek alphabet, written as $\chi$, and chi-square is written as $\chi^2$. At this stage, don’t attempt to take the square root of a chi-square: think of $<\text{chi-square}>$ as an entity in itself.

The term chi-square ($\chi^2$) is used in a number of contexts:

- It is the name of a class of continuous probability distributions (in fact the $\chi^2$ distribution can be derived from the standard Normal distribution). Just like the Student’s $t$ distribution, there is a family of $\chi^2$ distributions, a unique member of the family depends on its sole parameter: the number of degrees of freedom. All values along the abscissa of a $\chi^2$ distribution are positive (unlike the Normal distribution and the $t$ distribution which have values from $-\infty$ to $+\infty$).

[Image: Chi-square distributions on 1, 3 and 6 df]

- Just as every value on the abscissa of the standard Normal ($z$) distribution is called a $z$-value and every value of a $t$ distribution is called a $t$-value, every value of the $\chi^2$ distribution is called a $\chi^2$ value. For example, just as a $z$-value of $\pm 1.96$ cuts off a total of 5% in the tails of the $z$ distribution, a $\chi^2$ of 3.841 cuts off 5% of the upper tail of the $\chi^2$ distribution which has 1 degree of freedom. (For reasons we need not dwell on here, when doing two-tailed $\chi^2$ tests we nevertheless refer to upper tail critical values in published tables.)

- You will remember from Chapter 3 that in testing hypotheses about means of interval-scaled variables, we manipulated the data by dividing differences in means by standard errors to derive a value of a $t$ test statistic which was then compared to...
a critical value of the Student’s $t$ distribution. In a similar manner, we can test hypotheses about associations between nominal-scaled variables by manipulating the counts in the cells of the contingency table to derive a $\chi^2$ test statistic which is then compared to a critical value of the appropriate $\chi^2$ distribution.

§ 4.3.2 Calculating a $\chi^2$ Statistic

In 1900 the great British statistician, Karl Pearson, showed that, *if* the assumption that the Null hypothesis of independence of row and columns classifications were true, a statistic calculated from the following expression E4.2:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

has, approximately, a **theoretical chi-square distribution** on $(r-1) \times (c-1)$ degrees of freedom (for $r > 1$ and $c > 1$). In E4.2:

- The inner summation is over the $c$ cells (one in each column) of the currently addressed row of the table, and the outer summation is over the $r$ rows taken one at a time, so that for a 2x2 table there will be 4 addends;
- $O_{ij}$ is the *observed count* in the $ij^{th}$ cell;
- $E_{ij}$ is the *expected count* in the $ij^{th}$ cell.

This $\chi^2$ statistic (and it *is* a statistic since it is a number derived from sample data) is a function of the squares of the deviations of the *observed counts* from the *expected counts* in each cell. The observed counts come from the raw data or observations we have collected. The expected counts are the counts which are generated *for each cell by the specifications of the Null hypothesis*. This will be made clearer in the worked examples below.

*If the sample data – the observed counts – deviate from the counts expected under the assumed true Null hypothesis, then this leads to a large $\chi^2$ value with an associated low probability, and hence rejection of the Null hypothesis.*

You cannot help but note that this situation is entirely analogous to that of testing differences in means – if the observed (sample) difference in means differs substantially from the difference expected under the Null hypothesis, usually given as zero, then the appropriate $t$ statistic will be large, leading to a low probability under the relevant $t$ distribution and consequent rejection of the Null hypothesis.

§ 4.3.3 A Few Extra Points before Seeing $\chi^2$ in Action

The number of degrees of freedom associated with a $\chi^2$ distribution is sometimes difficult to calculate. In general, it is equal to the number of cells in the table (*not* the sample size as with $t$ statistics) minus the number of independent restrictions placed on the expected cell counts. It is the figuring out of the number of restrictions or constraints that may be difficult. Luckily for us, the tests we will deal with have easy-
to-calculate degrees of freedom and some simple rules will be given in the examples below.

It turns out that if we want our calculated $\chi^2$ statistics to follow a $\chi^2$ distribution (so that the probabilities we look up in the tables are valid) then each expected cell count (the $E_{ij}$) should be at least 5. Other tests are available for dealing with tables plagued by small expected cell counts. One of these tests is called Fisher’s Exact Test and is discussed in more advanced texts.

In the common case of a 2x2 contingency table (and only in this case) some authors have recommended that the usual formula for calculating the $\chi^2$ statistic be modified by what is termed Yates’ correction for continuity. (Sir Frank Yates is a British statistician who has made major contributions in the field of experimental design.) This correction decreases the absolute value of each cell’s [observed minus expected] deviation by 0.5 before squaring the numerator:

$$\chi^2_{\text{(Yates)}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij} - 0.5)^2}{E_{ij}}$$

This continuity correction will tend to lessen the resulting $\chi^2$ statistic which in turn will mean a larger tail-end probability (draw a diagram and see!). The end result is that this correction makes it a little harder to get a significant $\chi^2$.

Why bother with this continuity correction? The rationale has been that the true distributions of categorical variables are discrete whereas the theoretical $\chi^2$ distribution against which we will compare our test statistic is continuous. Use of the continuity correction was believed to yield probabilities closer to those we would get if we used an exact probability model rather than the approximation to the $\chi^2$. However, recent research shows that use of the continuity correction “overcompensates” and leads to tests that are too conservative. We will not use the continuity correction.

§ 4.4 CHI-SQUARE TESTS in ACTION

§ 4.4.1 The One-Variable Case: A “Goodness of Fit” Test

This is best explained by means of an example. Example 4.3 looks at a common source of error in clinical measurement – digit preference.
Example 4.3

Here are data on systolic blood pressures of 100 patients in a hypertension clinic. The measurement apparatus has 1 mmHg gradations.

150 134 128 170 155 150 156 149 163 167
135 124 178 140 165 151 146 159 163 158
168 154 158 181 145 171 157 149 144 168
148 144 158 171 135 151 157 160 164 170
159 124 148 171 185 152 177 170 164 170
149 154 159 161 165 133 167 150 174 160
149 155 169 161 166 153 187 160 184 171
150 156 159 161 146 143 167 140 184 162
150 136 159 161 166 153 157 160 164 173
150 156 139 161 177 134 138 150 194 170

Consider the terminal digits of each measurement. It is well known that in taking measurements, be they height, weight, blood pressure, the human observer tends to round up or down to 0 or 5 or at least to the nearest even number. So, a real blood pressure of 149 mmHg might be recorded as 150 mmHg. If such a tendency were not present, then the terminal digits 0, 1, 2 ... 8, 9 should appear with equal, or uniform, frequency. So, if no digit preference were operating (Null hypothesis!) in our 100 measurements we would expect to find 0 ten times, 1 ten times, 2 ten times etc. The frequencies of occurrence of each terminal digit of our raw data represent the observed counts. We can now construct a one-variable table to represent the frequency distribution of the observed and expected counts:

<table>
<thead>
<tr>
<th>digits</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>observed</td>
<td>18</td>
<td>12</td>
<td>2</td>
<td>7</td>
<td>15</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>expected</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

I might point out that this is a $1 \times 10$ table; the observed and expected counts belong to the same single row.

It looks like our recording technique might be faulty: there seems to be over-representation of some digits, for example, 0 and 4, and under-representation of others, for example 2 and 3.

But is this imbalance just due to chance sampling from a population exhibiting biological variability? We can use the $\chi^2$ test to see how well our observed distribution fits the expected or theoretical distribution.

Put in terms of an hypothesis test:

$H_0$: there is no significant difference between observed and expected cell counts (which is another way of stating that terminal digits should appear with equal frequency - any apparent imbalances are just due to chance sampling variation)

versus

$H_1$: a difference exists between the observed and expected counts (not simply due to chance sampling variability).
Example 4.3 continued

We will use the 5% significance level. The $\chi^2$ statistic is given by E4.2 with $r = 1$ row and $c = 10$ columns and we sum over all 10 cells. Then our calculated statistic will be compared to the $\chi^2$ critical value.

$$\chi^2 = \frac{[(18-10)^2/10] + [(12-10)^2/10] + [(2-10)^2/10] + [(7-10)^2/10]}{10} + \frac{[(15-10)^2/10] + [(8-10)^2/10] + [(8-10)^2/10] + [(9-10)^2/10]}{10} + \frac{[(10-10)^2/10] + [(11-10)^2/10]}{10}$$

whence $\chi^2 = 17.6$.

This table has ten cells and one restriction placed on it: the total number of observations was fixed at 100, so after any 9 cells have been filled the tenth is already determined. Hence the number of degrees of freedom is $10 - 1 = 9$.

If we look up $\chi^2$ tables we find that for the $\chi^2$ distribution with 9 degrees of freedom the critical value cutting off the upper 0.05 tail probability is $\chi^2_{9\text{df};\alpha=0.05} = 16.92$. Since the value of our test statistic, $\chi^2 = 17.6$, exceeds this critical value we reject the Null hypothesis. We claim that our observations do not appear to fit the theoretical distribution (not surprisingly, called a uniform distribution) that would apply if there were no digit preference operating. We have evidence that our recording technique favours some terminal digits.

Note that the process of constructing the test in Example 4.3 is entirely analogous to that of the tests in Chapter 3. That is, we:

- Assumed $H_0$ was true;
- Constructed the distribution of the counts that were expected if $H_0$ were true;
Calculated how likely it was that our sample data (observed counts) could have been drawn from such an expected distribution.

To put it another, more general, way we compared our data against a known probability model and reached a conclusion based on probability.

One can use the one sample “goodness of fit” chi-square test to see if observed counts fit a specific theoretical distribution. For example, does the number of patients preferring medications in tablet form, as opposed to other medication forms, follow a binomial distribution? Do the road fatalities observed over a twelve month period at a specific road intersection follow a Poisson distribution? The methodology is very similar to Example 4.3, except one would use the hypothesised distribution (eg the binomial) to work out the expected counts, and the degrees of freedom require a bit of thought. We won’t pursue these examples further.

§ 4.4.2 The Two-variable Case: Tests of Independence

We now continue our discussion of how to deal with cross-classified data, such as those arising from epidemiological studies. The analysis of data cross-classified by two categorical variables and presented in a contingency table will confront you for the rest of your career (just as the Normal distribution will!), so it is important that you become very familiar with the basics. There are more specialised techniques for dealing with these tables in areas such as epidemiology and genetics, but these techniques are just extensions of what is presented here.

The standard two variable case is a 2x2 table, that is, each variable is classified into two categories. Such a variable is termed binary or dichotomous. Examples of binary variables are sex (male/female), survival status (alive/dead), smoker (yes/no) and race (Caucasian/non-Caucasian). The last variable shows that, within the limits of common sense, a variable which in some circumstances might be classified into 4, 5 or 6 categories (Caucasian, Aboriginal, Malay, Eskimo, etc...) can be redefined (“collapsed”) into two categories depending on the needs of the researcher.

Here are examples of questions that the $\chi^2$ test can help answer – provided a suitable study has been designed!:

- Is there an association between sex and heart disease?
- Are smokers at increased risk of getting emphysema?
- Do more rural voters prefer conservative (versus liberal) health-care policies than urban voters?
- Is the proportion of musicians suffering from tendonitis higher than in typists?

Example 4.4

A random sample of adults (n = 61) is selected from the state electoral roll. When cross-classified by sex and smoking status, the following 2x2 table resulted:

Example 4.4 continued
Is there an association between sex and smoking status? Use the conventional $\alpha = 0.05$ significance level.

We first note that the prevalence of smoking in males is $12/36 = 33\%$, while in females, the prevalence is $5/25 = 20\%$. So, in this sample, there appears to be an association. Is there sufficient evidence to claim an association in the population?

The following statements illustrate how the Null hypothesis for the test in Example 4.4 may be proposed. They are all equivalent. Remember that the hypotheses are statements about underlying populations.

$H_0$: There is no association between sex and smoking status.

$H_0$: Sex and smoking status are independent.

$H_0$: The prevalence of smoking in females is the same as in males.

$H_0$: The difference in smoking prevalence rates between males and females is 0.

$H_0$: The prevalence ratio for smoking in males versus females is 1.

$H_0$: The same proportions of males and females are non-smokers.

$H_0$: Column classification is independent of row classification.

This last, and most general, of the expressions leads us now to consider what our expected cell counts should be, if $H_0$ were true. This relates directly to one of the basic laws of probability E2.5 we met in Chapter 2. Remember that two events $X$ and $Y$ are said to be independent if and only if the following relationship holds:

$$\text{Prob}(X \text{ and } Y) = \text{Prob}(X) \times \text{Prob}(Y);$$

that is, if the joint probability equals the product of the individual (unconditional) probabilities. To fix the methodology clearly in mind, we’ll deal with a concrete example: what should be the expected count of male smokers in Example 4.4?

Example 4.4 continued

If the Null Hypothesis were true, then smoking status and sex are independent (or more precisely, if you like, the event that one of the subjects is classified as a male is independent of the event that the same subject is classified as a smoker).

We’ll find the probability that a subject is a male smoker under the assumption that $H_0$ is true:

$$\text{Prob}(\text{male} \text{ and } \text{smoker}) = \text{Prob}(\text{male}) \times \text{Prob}(\text{smoker})$$

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>smokes</strong></td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td><strong>does not smoke</strong></td>
<td>24</td>
<td>20</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>36</td>
<td>25</td>
</tr>
</tbody>
</table>
Example 4.4 continued

So we need to find each of the two probabilities on the right hand side (RHS) of the equation. Remember these probabilities are population probabilities (since hypothesis testing is all about populations). In the absence of other information, we will have to make use of the observed marginal totals to estimate the required probabilities. This is important to remember because it has implications for our degrees of freedom, as we shall soon see. Our best estimate (and that is all it is) for the probability of being a male in the population is given by:

\[
\text{Prob(male)} = \frac{\text{number of males in the sample}}{\text{total size of the sample}} = \frac{36}{61}
\]

Similarly,

\[
\text{Prob(smoker)} = \frac{\text{number of smokers in the sample}}{\text{total size of the sample}} = \frac{17}{61}
\]

So, if independence is true (that is, if \( H_0 \) is true), then according to E2.5:

\[
\text{Prob(male and smoker)} = \left(\frac{36}{61}\right) \times \left(\frac{17}{61}\right)
\]

But this is the probability that any one person (chosen at random from a population where sex and smoking are not associated) is a male smoker. In our total sample of 61 we would therefore expect there to be \((36/61) \times (17/61) \times 61 = 10.03\) male smokers.

Now that you see how expected counts under the Null hypothesis are generated for the 2x2 table, or indeed any RxC table, you don’t need to wade through the logic every time. You can see that each cell’s expected count will simply be the product of the row and column marginal totals associated with that cell, divided by the total sample size.

Example 4.4 continued

Using the methods given above, the expected number of female non-smokers will be: 25 x 44/61 or about 18.03. (It is a good idea to carry several decimal points for expected counts through the intermediate calculations.) Here is the completed table with the expected counts for each cell italicised and in parentheses:

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smokes</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>(10.03)</td>
<td>(6.97)</td>
</tr>
<tr>
<td>Does not smoke</td>
<td>24</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>(25.97)</td>
<td>(18.03)</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>25</td>
</tr>
</tbody>
</table>

Note that the sums of expected counts agree with the marginals and total sample size, and this is a useful check on calculations. We will now calculate the \( \chi^2 \) statistic for this table using E4.2. If you plug in the observed and expected counts for all 4 cells (and you should!), you will find that:
Example 4.4 continued

\[ \chi^2 = 1.31 \] (to 2 decimal places)

This value of the chi-square statistic should be compared with the critical value of \( \chi^2 \) cutting off the upper 5% of the area under the \( \chi^2 \) distribution with 1 degree of freedom (\( \chi^2 = 3.841 \)).

Since 1.31 < 3.841 we are unable to reject the Null hypothesis and so we conclude, until and unless further evidence is produced, that sex and smoking status are independent; that is, there is no statistically significant association. A computer analysis gives the P value associated with \( \chi^2 = 1.31 \) as 0.25. This means that when sampling repeatedly from a population where there was no difference in smoking prevalence between the sexes, we might nevertheless observe in our sample a difference in prevalence as large as 13% (33% vs 20%) on every fourth occasion. Not an unusual occurrence, I would judge.

We could also form an approximate 95% confidence interval around the difference in prevalences. To do this we use (i) the result that the sampling distribution of a difference in proportions is approximately Normally distributed and (ii) the formula for the standard error of such a distribution given in E3.1. We need to substitute the sample proportions for the population proportions in the formula. We therefore calculate the limits of the 95% confidence interval as:

\[
(p_1 - p_2) \pm 1.96 \times \sqrt{\left\{ \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \right\}}
\]

which on our data (\( p_1=0.33, p_2=0.2, n_1 = 36, n_2=25 \)) yields \([-0.09, 0.35]\). Note that the confidence interval is symmetric about the point estimate of 0.13, and that the interval includes the value of the difference in prevalence specified by the Null Hypothesis, 0. We do not reject \( H_0 \).

Why is the test in Example 4.4 associated with only one degree of freedom? There are several ways of calculating the degrees of freedom – each method is really saying the same thing in a different way. We have 4 cells to play with, but we lose 3 degrees of freedom in estimating the expected counts from the observed marginal totals. When the marginal totals are “fixed” to allow computation of expected counts, only one cell is free to vary in value – the other 3 are then constrained to add up to the marginal totals.

If you prefer a simple, general rule (who doesn’t?) for calculating degrees of freedom for any two variable RxC table, it is:

degrees of freedom = \((r-1) \times (c-1)\)  \hspace{1cm} E4.4

In Example 4.4, \( r = 2 \) and \( c = 2 \), so \( df = (2-1) \times (2-1) = 1 \).

Finally, it is most important to understand that the \( \chi^2 \) test does not directly address the question of strength or degree of association between two variables. The test simply addresses the question of whether or not there is an association that cannot be
ascribed simply to chance. One should examine the measure of effect (incidence ratio, odds ratio, difference in proportions etc) to judge the strength of the association.

Just to tidy things away, let us revisit our case-control study.

**Example 4.2 continued**

If we use E4.2 with the data from the case-control study of glioma, we get $\chi^2 = 0.65$, which is certainly much less than the $\chi^2$ critical value for a test at the 5% level, 3.841. In fact, this value of the test statistic is associated with a P value of 0.42, so we are in no position to reject $H_0$. On more than 4 out of 10 occasions in repeated sampling, we could expect a population where there was no association between smoking and glioma to yield data with an odds ratio as great as that which we saw on this occasion, 1.06. Nothing unusual here! Note also that this decision accords with that we made earlier on the basis of the confidence interval.

§ 4.4.3 **A Note of Warning Concerning Independence**

This warning applies to chi-square tests of independent samples such as we have been considering. A common error in analysing frequency data is to ignore a fundamental assumption: that each observation must be independent of all other observations. One must not count the same individual twice, nor count occurrences of some phenomenon of interest twice in the same person, and treat this as two separate observations. This mistake is easy to make. It leads to a spurious inflation of the total sample size, and an associated hidden inflation of the Type I error. Null hypotheses may be rejected when they shouldn’t be.

These mistakes may occur at the design, execution and analysis phases of the study. Perhaps several researchers are collecting information which they intend to pool. Inadequate attention is given to territorial demarcation or to rigorous unique identification of study subjects. Without realising, two or more researchers may interview the same subject, so that subject’s data are recorded twice. This is bad news. Alternatively, one might be counting the number of, say, asthma attacks occurring in a sample. It is not correct to count two attacks in the same person during the study period and treat these as independent events. Statistical methods exist to deal with these types of data, but expert statistical assistance should be sought.
§ 4.4.4 Extensions to the $2 \times 2$ Table

The $2 \times 2$ situation can be extended to the $2 \times C$ (or the $R \times 2$) and the $R \times C$ situation. Here is an example of a $2 \times 4$ cross-classification that you should work through yourself. You could probably calculate the expected counts a little faster now by applying the rule: expected cell count equals product of marginal totals divided by total number (hopefully remembering the underlying rationale!). To keep you on track, I’ve filled in some of the expected cell counts in parentheses. Don’t forget to write down the opposing hypotheses and draw a diagram of the relevant distribution, labelling it with its degrees of freedom and the critical value of the $\chi^2$ distribution with which your test statistic will be compared.

Example 4.5

This contingency table cross-classifies 100 patients suffering from carcinoma of the colon (large intestine) by the type of symptom they first experienced (pain or rectal bleeding) and how advanced the cancer was when first diagnosed, according to Duke’s stages A, B, C, or D. (Increasing stage is associated with poorer prognoses: A represents local disease, through to D which represents distant metastases.) Is there an association between presenting symptom and stage of carcinoma of the colon at the $\alpha = 0.10$ level?

The observed counts were:

<table>
<thead>
<tr>
<th>initial symptom</th>
<th>stage of cancer (Duke’s Classification)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>pain</td>
<td>19</td>
</tr>
<tr>
<td>(21.06)</td>
<td></td>
</tr>
<tr>
<td>rectal bleeding</td>
<td>7</td>
</tr>
<tr>
<td>(6.46)</td>
<td></td>
</tr>
</tbody>
</table>

Hints:

- Some of the expected counts are calculated (in italics and parentheses).
- You should be able to see that the number of degrees of freedom for the $\chi^2$ distribution against which you will compare your test statistic is 3 (see E4.4).
- Your $\chi^2$ statistic should work out to be 5.5.
- The critical value for $\chi^2_{3df} \alpha=0.10$ is 6.25 (from tables).
- Sketch the Null distribution and note that the tail probability associated with $\chi^2 = 5.5$ is larger than 0.10. We cannot reject the Null hypothesis of no association.
§ 4.5 MATCHING

The basic concepts of matching, as discussed in Chapter 3 (§3.5), can be applied to studies involving categorical data; for such data the chi-square distribution is used as the main probability model. A chi-square test on paired or matched samples takes a slightly different form from that applied to independent samples and is usually called McNemar’s Test.

§ 4.5.1 Self-Matching (“Natural Pairing”)

Unlike the case of independent samples (see the warning in §4.4.3), it often happens that a subject is intentionally observed or interviewed twice. For example, patients with malignant disease have a blood count taken both before and after a blood transfusion, and on the basis of each test are classified as being anaemic or non-anaemic. Is there a change in such patients’ haematological status after the transfusion? (In the individual case, of course, the two blood counts themselves will give us the answer, but the oncologist may be interested in whether a more general policy of treating cancer patients with blood transfusions is worthwhile – this is where inferential statistics helps out.) Each patient is his or her own match or pair. The study design is often referred to as a before-after study or as a simple repeated measures study.

Example 4.6 develops a simple test to decide whether or not a statistically significant change has occurred. Once again, a statistically significant change is a change not due solely to the sampling variability.

<table>
<thead>
<tr>
<th>Example 4.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior to their scheduled open-heart surgery for coronary artery disease, 100 patients were classified as being either depressed or not depressed (“normal”). A psychiatrist was interested in whether there was any significant mood change three months after surgery (that is, had there been a change in the proportion of all patients who had been depressed?; put equivalently in yet another way, had there been a change in the proportion of all patients who had been normal?). The data are:</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>before surgery</th>
<th>after surgery</th>
<th>not depressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>depressed</td>
<td>63</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(12.5)</td>
</tr>
<tr>
<td>not depressed</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(12.5)</td>
</tr>
<tr>
<td></td>
<td>67</td>
<td>33</td>
</tr>
</tbody>
</table>
Example 4.6 continued

The table above shows the observed counts; the appropriate expected counts generated by the Null hypothesis are in parentheses. The rationale for the expected counts will be given shortly. (For the moment, don’t worry about the little r and s lurking around.)

Note carefully the difference in layout of this table of matched data from the previous tables we have seen of independent samples data. You can see that each of the 100 patients was cross-classified on his/her mood before and after surgery. For example 63 patients were depressed before surgery and felt the same afterwards. It must now be understood that since we are interested in change, the subjects who remained depressed or remained normal don’t supply any comparative information. We discard these concordant (tied) pairs and concentrate only on the 25 discordant (untied) pairs – the 21 persons who changed from depressed to normal and the 4 persons who changed from normal to depressed.

What about the expected counts for these two off-diagonal cells? If, as the Null hypothesis would specify, there was no overall mood change after surgery, then the individual mood changes that do occur should do so at random – it is just as likely that a subject might change from depressed to normal as from normal to depressed. Assuming $H_0$ is true, the probability of a discordant pair being a [normal $\rightarrow$ depressed] pair should equal the probability of being a [depressed $\rightarrow$ normal] pair. That is, each probability should be $\frac{1}{2}$. Since we have 25 untied pairs we expect $25/2 = 12.5$ of each type, and this is shown in parentheses in the above table.

We apply the usual chi-square test formula E4.2 to this 2x2 table. Remember that we are only interested in the observed and expected counts in the two off-diagonal cells – ignore the other two cells when applying the formula. The number of degrees of freedom is 1.

$$\chi^2 = \frac{(21-12.5)^2}{12.5} + \frac{(4-12.5)^2}{12.5} = 11.56$$

This test statistic exceeds the 5% critical value of 3.841, so we reject the Null hypothesis and claim there is a mood change after surgery.

As usual in statistics, there is an alternative simpler expression to give the matched chi-square. It entails the off-diagonal observed counts r and s (in Example 4.6, r = 21 and s = 4) as shown in the diagram.

$$\chi^2 = \frac{(r-s)^2}{r+s} \quad \text{E4.5}$$

Plug in $r = 21$ and $s = 4$ and convince yourself that it yields $\chi^2 = 11.56$ as before. For completeness, although we will not use it, a formula that does include the continuity correction is given by:

$$\chi^2_{(Yates)} = \frac{[(r-s) - 1]^2}{r+s} \quad \text{E4.6}$$
§ 4.5.2 Artificial Matching

I will use a type of case-control design to illustrate this situation. Quite often in case-control studies, each case (diseased subject) is intentionally and individually matched with a non-diseased control on the basis of sex, age, marital status, race and/or other factors which the researcher believes may confound the true association – should one exist – between the risk factor(s) of interest and the disease. This type of matching is an artificial pairing as opposed to the self-pairing seen in Example 4.6. In matched studies, it is the pair which is classified on the risk factor status (see Example 4.7 below). Epidemiologists still use the odds ratio as the measure of effect for such a study, but its calculation differs from the cross-product ratio for the unmatched study (see E4.1).

\[
\text{odds ratio}_{\text{matched}} = \frac{r}{s} \quad \text{E4.7}
\]

where \( r \) is the number of discordant pairs where the case is exposed but the control is unexposed, and \( s \) is the number of discordant pairs where the control is exposed but the case is unexposed. The interpretation is just like that of an unmatched study: an OR \( > 1 \) means that cases are more likely to have been exposed than controls (and, if we are sure of the validity of our results, we say that the exposure is a risk factor for the disease); OR \( < 1 \) means that cases are less likely than controls to have been exposed (and we may say that the exposure is a “protective factor” for the disease). If we wish to test \( H_0: \text{OR} = 1 \) we use McNemar’s Test.

**Example 4.7**

Each member of a sample (n = 60) of subjects with breast cancer (a case) is individually matched to a subject without breast disease (a control) on the basis of age, marital status, history of breast-feeding and parity, all of which the researcher feels might confound any relationship between the disease and the risk factor of interest, use of oral contraceptives as a teenager. Exposure to the oral contraceptive pill was ascertained by patient recall (does this raise a problem?). Then each of the 60 case-control pairs was classified on whether or not the oral contraceptive pill had been used before the age of 20.

The Null hypothesis for this study is:

\( H_0: \) There is no association between teenage use of oral contraceptives and breast cancer.

(Of course, by now, you could easily express this in a variety of equivalent ways.)

Here is the data setup. Examine it carefully, and note that we are dealing with 60 pairs, not 120 individuals:
Example 4.7 continued

<table>
<thead>
<tr>
<th>cases</th>
<th>used OC</th>
<th>not use OC</th>
</tr>
</thead>
<tbody>
<tr>
<td>used OC</td>
<td>25</td>
<td>19</td>
</tr>
<tr>
<td>not used OC</td>
<td>11</td>
<td>5</td>
</tr>
</tbody>
</table>

For example, there were 25 pairs whose case and control members were exposed to the oral contraceptive. Once again, under the assumed true Null hypothesis of no association in the population, an untied pair should have an equal chance of being a [case used OC, control not used OC] pair as of being a [case not used OC, control used OC] pair. Since we have r + s = 19+11 = 30 pairs, the expected discordant counts are each 15, as shown.

The observed data show that there are more discordant pairs in which the case was exposed to the pill than discordant pairs where the control was exposed (19 versus 11). We calculate the odds ratio, using E4.7, as 19/11 = 1.73. This suggests that the pill might be a risk factor for breast cancer, but is this imbalance just due to sampling variability (that is, just due to chance)? We use the matched samples chi-square (McNemar’s) test to help us decide.

You should verify that the chi-square statistic E4.5 is calculated as 2.13, which does not exceed the critical value: \( \chi^2_{1 \text{df}, \alpha=0.05} = 3.841 \). That is, there is a greater than 5% chance that a population which has no association between the disease and risk factor might generate a sample such as ours. In fact, using a computer, I find that \( \chi^2 = 2.13 \) is associated with a P value of 0.14. We have failed to reject the Null hypothesis at the 5% level.

Although a statistics package on a computer will deliver more satisfactory, exact confidence limits for the matched odds ratio, approximate 95% confidence limits are given by:

\[
(r/s) \cdot e^{\pm 1.96 \sqrt{(1/r)+(1/s)}}
\]

Substituting \( r=19 \) and \( s=11 \), we obtain the approximate 95% confidence interval for the population odds ratio as [0.82, 3.63]. Using a computer, I also calculated an exact 95% CI: [0.78, 4.02]. Note that whichever method is used: (i) the confidence limits are asymmetrical around the point estimate of 1.73 (this asymmetry is usual when dealing with ratio estimators) and (ii) the confidence interval includes 1, the value specified by \( H_0 \). Again, we fail to reject \( H_0 \).
§ 4.6 SUMMARY

When approaching data arrayed in a contingency table, always take the time to ensure you interpret the data correctly. You will need to know something of the design of the study that generated the data, choose an appropriate measure of effect to summarise the data, and understand the results of any inferential test you conduct.

The chi-square distribution often provides a useful, approximate, probability model for data measured on a categorical scale. Be cautious when interpreting any test result which is based on tables with empty or near-empty cells. Special methods, not discussed in this course, must be applied in such cases.