The Analysis of categorical data

Introduction: Normal, non-Normal and categorical data

The methods described so far in this course have been appropriate for the analysis of data that are recorded on a continuous scale, e.g. stature, serum concentrations, systolic blood pressure. Much of the discussion has also assumed that a Normal distribution is also appropriate, although there are numerous continuous outcomes that do not follow a Normal distribution. For example bilirubin concentrations are usually very skew, as are various survival times. Special methods and approaches for some of these variables are discussed elsewhere in the course.

Continuous variables might or might not be Normal. However, strictly speaking variables that are not continuous cannot be Normal, so the methods discussed hitherto are not appropriate and we need to consider new methods of analysis.

Types of categorical variable

A variable that is not continuous will be recorded as being in one of several *categories*, hence the term *categorical* variable. For example, the ABO blood group will be one of A, B, AB or O. Tumour stage is often recorded as stages I, II, III or IV. Both of these variables are categorical but there is a difference between them. A patient with a tumour in stage II is, in some sense worse, off than a patient with tumour stage I and better off than a patient with tumour stage IV. In other words there is an ordering between the categories and statistical analyses should usually take account of this. A categorical variable in which the categories are ordered, often referred to as an *ordinal variable*, occupies a half-way house between a an unordered categorical variable and a continuous variable. However, unlike a continuous variable, a patient with tumour stage IV cannot be said to be twice as bad as one with a stage II tumour.

Ordinal variables, especially those with large numbers of categories and collected on large samples, can sometimes be analysed as though they were continuous variables. However, for smaller samples and variables with only a few categories, special statistical methods are needed and these are beyond the scope of this course. In general the analysis of ordinal data requires specialist help.

The simplest form of categorical variable is a binary variable, being a categorical variable with just two categories. Examples are everywhere: does a grafted kidney exhibit delayed function (yes or no); does a patient die before discharge (yes or no); does the patient’s condition resolve, does the baby have a severe handicap, does the infant need treating for hypoglycaemia. In some cases the binary variable is a crude summary of some more informative underlying variable. For example, a severe handicap may be defined as a score being above or below some threshold. Whether or not a patient is hypoglycaemic will be based on their blood glucose concentration. In such cases you should always ask if the underlying variable ought to be analysed rather than the binary summary. However, as many clinical actions follow naturally from a binary summary (e.g. a patient either is or is not treated for hypoglycaemia), there may well be a good reason to analyse the binary variable.

Most of the rest of this document will be concerned with the analysis of binary variables.

Summarising a binary variable: what parameter are we dealing with?

A population of binary variables comprises patients (or volunteers or other experimental or observational subjects) some having one value for the variable and the remainder having the other value. A terminology that is sometimes used is that a patient either does or does not possess an 'attribute'. Not much can differ between different populations of this type. In fact the only thing which can change is some measure of how many in the population possess an attribute and how many do not. Most populations are conceived of as having infinite size, so the issue is addressed in terms of the proportion of the population possessing the attribute. This value is the sole parameter we need to define the population: in keeping with the convention that parameters are given Greek letters, it is usually written as π.

The parameter π can be thought of as the proportion of the population possessing the attribute in question. An equivalent interpretation is that it is the probability that a randomly chosen member of the population has the attribute. Of course, the proportion $1-\pi$ is simply the proportion not having the attribute or the probability of not having the attribute. In common with other population parameters, π is unknown, and our analyses will be directed towards estimating it and answering questions about it.

Estimating π is very straightforward. If we have a sample of n patients from a population, r of them will possess the attribute and n-r will not. The estimator of π is simply r/n, or if you wish to express it on a percentage scale, $100r/n$%.

For example, in an audit of in-hospital mortality from abdominal aortic aneurysm repair[†], 689 patients not on diuretics were sampled and 34 of these died before they could be discharged from hospital. Here n=689 and r = 34, so the estimate of the population proportion dying before discharge is 34/689 = 0.049, or 4.9%.

Why do we need new methods?

An immediate answer to this question is that the methods discussed in earlier chapters were derived assuming that a Normal distribution applied (as in the t-test), or at the least variables had a range of values (as in a histogram). However, this answer is rather superficial and unsatisfying. If we score 1 for those dying before discharge in the above example and 0 for those being discharged alive then the mean of these 0s and 1s is precisely the observed proportion of 1s in the sample. In other words, the summary suggested for the binary outcome, an observed proportion, is the same as the summary suggested for many continuous variables, namely the mean, provided that a simple and natural scoring system is adopted. So what is the deeper reason why we need new methods?

[†] Bayly, PJM *et al.* (2001) *Br J Surgery*, **88**, 687-692.

The answer is to do with the standard deviation of the variable. In the case of a continuous variable, the population was described in terms of a mean parameter μ (estimated by a sample mean m) and a quite independent parameter, σ the population standard deviation. In principle, any value of μ could go with any (positive) value of σ. For the case of a binary variable there is only one parameter, π, and this is clearly an analogue of the mean, μ, of a continuous population.

For a binary variable things have to different – the standard deviation must, in some way be related to the mean. This is indeed the case and the easiest way to see this is by considering how well r/n estimates π i.e. we consider the standard error rather than the standard deviation.

For a continuous variable this would be $\sigma/\sqrt{n}$, whereas for a binary variable it is $\sqrt{[\pi(1-\pi)/n]}$. The formulae share the presence of $\sqrt{n}$ on the denominator: the larger the sample, the smaller the standard error. However the numerators are rather different: that for the binary variable being defined in terms of π rather than being a separate parameter. The details of the formula are not important (although they are needed if you want to construct a confidence interval for π: see Appendix I), rather it is the fact that once π has been estimated there is no need for further estimation of spread. As this is not the case for continuous variables the formulae, and hence the methods we have used for them cannot be applied directly to binary variables.

Before going on to consider the binary analogue to the t-test, it is instructive to consider the following fictitious and extreme example. It certainly does not prove the formula $\sqrt{[\pi(1-\pi)/n]}$, but it serves to illustrate why some of its features must be as they are.

Suppose $\pi = 0$, i.e. no member of the population has the attribute. In this case, no element of any sample can possess the attribute, so necessarily r=0 and hence $p = 0$, an estimate which is without error. Consequently, the standard error of this estimator must be zero, which explains the presence of π in the numerator of the above formula. A similar argument can be made when every element in the population has the attribute, i.e. $\pi = 1$. Now $p = 1$ for every sample and again there is no error, so the standard error must be zero. This explains the factor (1- π) in the numerator.

The binary analogue of the unpaired *t*-test: the χ^2 – test (or chi-squared test)

In addition to the 689 patients mentioned in the above example who were not taking diuretics, 241 patients were taking diuretics. Of the latter, 25 were not discharged alive. The sample proportions dying before discharge are, therefore, 25/241 = 10.4% on diuretics and 34/689 = 4.9% off diuretics. This information is conveniently arranged in a *2 × 2 contingency table*, viz.

	Dead	Alive	*Total*
Not on diuretics	34	655	*689*
On diuretics	25	216	*241*
Total	*59*	*871*	*930*

Table 1

The italicised row and column giving the totals are known as the *margins* of the 2 × 2 table. They will play an important role in our analysis.

A corresponding table of (row) percentages is

	Dead	Alive	*Total*
Not on diuretics	4.9	95.1	*100*
On diuretics	10.4	89.6	*100*
Total	*6.3*	*93.7*	*100*

Table 2

The population proportions dying before discharge in the groups taking or not taking diuretics are π_D and π_N, respectively. Does the difference in the corresponding sample proportions provide evidence that the population proportions differ? The test of this hypothesis is the χ^2 – *test* (sometimes written as the 'chi-squared test') and this will be described in what follows.

There are two unpaired groups, those taking or not taking diuretics, and the proportion is a parameter that is analogous to the mean of a continuous variable. Consequently the test of the null hypothesis that $\pi_D = \pi_N$, is the binary analogue of the unpaired *t*-test.

General rationale

In this example, 34 out of the 689 patients not on diuretics die before discharge, whereas 25 out of 241 on diuretics die before discharge. Are these numbers different? Clearly, if the proportions dying in the two groups were the same, we would not expect equal numbers to die in each group because one group is almost three time the size of the other. What we can do is to work out how many we would *expect* to die in each group *if the two population proportions (*i.e. π_D and π_N) were equal. We can arrange the results of this calculation in a table analogous to Table 1 and then try to decide if the disparity between the tables is too large to be reasonably ascribed to chance. In other words, we work out what we might expect if the *null hypothesis* is true and then see if the difference between this expectation and the real data should surprise us. That is, we adopt the same general approach as in the *t*-tests, or indeed any hypothesis test.

Calculating the 'expected' values.

If the proportions dying before discharge are the same in the two groups, then the numbers dying in the groups should be in the ratio of the sizes of the two groups. That is, the expected number dying in the diuretic group is 241 × *k* and the number in the no diuretic group is 689 × *k* where *k* is an estimate of the proportion dying that is common to the two groups. How do we get an estimate of *k*? If the two groups do share a common death rate then the best estimate of *k* comes from the data combined across the two groups. This has been done in the bottom (*Total*) row of Table 1, namely a total of 59 out of 930 patients died before discharge, so the estimate of *k* is 59/930.

We now have the numbers that we 'expect' to die before discharge in the two groups. The numbers we expect to survive can be calculated by repeating the above, but with *k* now calculated as the proportion surviving, namely 871/930. Alternatively, the number surviving in a group of size 689 must be 689 – the number dying, and similarly for the diuretic group (you can check to see that these two methods give the same answer).

We can now set out the table of expected values as table 3a below:

	Dead	Alive	*Total*
Not on diuretics	$689\times\frac{59}{930}$	$689\times\frac{871}{930}$	*689*
On diuretics	$241\times\frac{59}{930}$	$241\times\frac{871}{930}$	*241*
Total	*59*	*871*	*930*

Table 3a

Doing the arithmetic gives the following table.

	Dead	Alive	*Total*
Not on diuretics	43.71	645.29	*689*
On diuretics	15.29	225.71	*241*
Total	*59*	*871*	*930*

Table 3b

Notes:

i) The expected values add up along the rows and columns to give the same margins as in the table of observed values (Table 1). In fact, this is a consequence of the way the expected values have been calculated.

ii) The expected values are not whole numbers, so this table could not have been observed. It is a sort of 'average' table that would be obtained over many observations of samples that give these margins and have equal proportions dying before discharge in the two groups.

Measuring how far apart are the observed and expected tables.

If the basis on which Table 3b has been computed, i.e. the null hypothesis, is true, then we would expect Tables 1 and 3b to be 'close' in some sense. One way to assess this is informally, and this is most easily done by putting the two tables into the same frame, as in Table 4 below. The observed values are in plain type and the expected in italics

	Dead	Alive	*Total*
Not on diuretics	34 *43.71*	655 *645.29*	*689*
On diuretics	25 *15.29*	216 *225.71*	*241*
Total	*59*	*871*	*930*

Table 4

The two sets of figures do not look very close, but such an impression is hopelessly vague for any sensible attempt at inference. What is needed is a way of calculating the difference between the expected and observed tables.

An immediate reaction might be to compute the difference (*observed – expected*) in each cell of the table and add up the four differences so obtained. Unfortunately this won't do: by the method of construction the sum of the four differences will always be zero. This is a similar problem to that we encountered when attempting to define the standard deviation. A temptation is to adopt the solution used there and add up the squared differences, (*observed – expected*)^2.

This is almost what we do, but not quite. In fact we divide this squared difference by *expected* and then add these quantities over the cells of the table. The rationale for this is given in Appendix II. This measure of the difference between the tables is known as the χ^2 – statistic (or chi-squared statistic, from the Greek letter χ (chi), pronounced '*kye*'). This can be written symbolically as:

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

This formula is seen in many textbooks, with *O* and *E* standing for *observed* and *expected* counts respectively. Notice that it can never be negative and is zero only when the observed and expected tables coincide.

Computing (*observed – expected*)^2/*expected* for each cell in Table 4, we get

	Dead	Alive	*Total*
Not on diuretics	$\frac{(34-43.71)^2}{43.71} = 2.157$	$\frac{(655-645.29)^2}{645.29} = 0.146$	*689*
On diuretics	$\frac{(25-15.29)^2}{15.29} = 6.168$	$\frac{(216-225.71)^2}{225.71} = 0.418$	*241*
Total	*59*	*871*	*930*

Table 5

Adding these values gives χ^2 = 2.157+6.168+0.146+0.418 = 8.889.

This is the measure of the difference between the observed and expected tables that is used in the χ^2 test.

Should I be surprised?

The answer arrived at above, namely 8.889, shows that the two tables differ. But we knew that from looking at them, so it is tempting to react to the value of 8.889 by asking 'so what'?

We computed the statistic so that we could address the question of whether or not we should be surprised by the size of the difference between our observed table and the table we would 'expect' of the null hypothesis were true. If it is surprising, then we have evidence against the null hypothesis, the amount of evidence is measured by the probability of seeing a difference of, in this case, 8.889, if the null hypothesis is true, i.e. the P-value. So what we need is a way of turning 8.889 into a P-value.

In considering this question we focus attention solely on those tables which have the same marginal totals as the observed table, i.e. tables like

	Dead	Alive	*Total*
Not on diuretics	*?*	*?*	*689*
On diuretics	*?*	*?*	*241*
Total	*59*	*871*	*930*

Table 6

Clearly, it is sensible to keep the group sizes fixed, as this is part of the formulation of the problem. We also assume that the total numbers discharged dead or alive are fixed, as these give information on the size of the common proportion dying (remember we are assuming that the null hypothesis is true). However, even with this restriction, many tables are possible, with some being more likely than others. Which tables are more likely depends on the relative values of π_1 and π_2. If the null hypothesis is true, then tables such as the following are quite likely.

	Dead	Alive	*Total*
Not on diuretics	44	645	*689*
On diuretics	15	226	*241*
Total	*59*	*871*	*930*

Table 7a

	Dead	Alive	*Total*
Not on diuretics	47	642	*689*
On diuretics	12	229	*241*
Total	*59*	*871*	*930*

Table 7b

These tables are quite plausible if the null hypothesis is true because the proportions dying before discharge in the two groups are similar. For Table 7a they are 6.3% & 6.2% and for Table 7b they are 6.8% & 4.9%.

The χ^2 statistic can be computed for these tables, giving, respectively, 0.008 and 1.02. These are smaller than the value of 8.889 we obtained from the real data, but we are not that much clearer what they mean.

If we had some way of generating tables with these margins and *for which we knew the null hypothesis was true*, then we could generate many such tables, say 10000, and calculate χ^2 for each of them and then we would know what values of χ^2 occurred when the null hypothesis was true. If our observed value were larger then most of these, then we would be able to conclude that our data were unlikely to occur if the null hypothesis were true. In fact, the proportion of the 10000 χ^2 values exceeding our observed value of 8.889 would serve as a P-value for the test.

Is this what we do? Well, with modern computers we certainly could, but generally we adopt a solution that was worked out mathematically in the early 1900s. All good statistical packages will compute χ^2 and the corresponding P-value. The latter will have been computed using a mathematical approach which essentially anticipates the result of generating a large number of tables and hence determines the distribution of the χ^2 statistic if the null hypothesis is true. The technique uses what is known as an *asymptotic approximation* and the details need not concern us.

For our data, with $\chi^2 = 8.889$, this method gives P=0.003. So, if the population proportions dying before discharge in the two groups were the same, a table with proportions as disparate as those in Table 1 would arise by chance on 3 occasions in every 1000. In other words we have to conclude either that we have seen a very rare event, or that the null hypothesis is false. Therefore, as we do not accept that we have seen such a rare event, our data provides strong evidence against the null hypothesis.

The computational alternative, which is not generally used but which might give you further insight into how we derive the P-value, is outline briefly in Appendix III.

Larger tables

So far we have considered only tables made up from two classifications, each with two levels, giving rise to a 2 × 2 table. It is, of course, possible that factors classifying data can have more than two levels. An example from the audit of aortic aneurysm surgery used above is in Table 8.

Elective cases				
	Low volume	Medium volume	High volume	*Total*
Discharged dead	19	24	13	*56*
Discharged alive	261	319	175	*755*
Total	*280*	*343*	*188*	*811*

Table 8

The table shows the numbers of patients discharged alive or dead according to whether or not they had their operations in a unit that did a low, medium or high volume of aortic aneurysm surgery. There are, e.g., more cases in the low volume than the high volume class because the data are aggregated across many units, only a few of which did a high volume of this kind of operation. Also, the total number of cases is not the same as in Table 1 because this table considers only elective operations: urgent cases are excluded.

The same kind of analysis can be performed on a table like this. The null hypothesis would be that the pre-discharge death rates are the same in each of the volume groups. The table of expected values is calculated in a way that is entirely analogous to the 2 × 2 case. For example, the 56 deaths are allocated to the low, medium and high volume categories in the ratio 280:343:188. The results are shown in table 9.

Elective cases (expected values under the null hypothesis)				
	Low volume	Medium volume	High volume	*Total*
Discharged dead	19.33	23.68	12.98	*56*
Discharged alive	260.67	319.32	175.02	*755*
Total	*280*	*343*	*188*	*811*

Table 9

From these values χ^2 can be found from the usual formula $\chi^2 = \sum \frac{(O-E)^2}{E}$, giving a value of 0.011. The low value reflects the fact that the expected values are, in this case, very close to the observed values. From this value of χ^2 we can find P, in much the same way as for the 2 × 2 table, yielding P = 0.995[‡]. Therefore, in this table there is no evidence at all that the proportion dying before discharge differs between the volume categories.

Some items to remember about the χ^2 test.

Make sure the entries in the table you analyse are counts, not proportions

The data in our example in Table 1 comprises the counts in each cell. These can be analysed using the χ^2 test. It is useful to present the estimated proportions (or more usually percentages), as in Table 2. However, it is imperative that you do not apply the χ^2 test to the table of percentages.

The reason for this is to do with standard errors. When we discussed the *t*-test the way differences in means were compared to the standard error of the difference was quite explicit. Although nothing like as apparent, the χ^2 test does the same – differences in observed proportions are compared with the standard error of the difference. In order to do this properly it is essential that not only the proportions in the groups are presented but that the denominators are available. If you observe two events from a sample of size 10 then that is the same proportion as if you observed 200 events out of a sample of 1000. However, in the latter case the estimate of the true proportion is much more precise than in the former case[†]. Consequently when comparing with another proportion, the test will be much more sensitive to departures from the value for this population in the case with the larger sample size. Consequently it must be remembered (as it sometimes is not!) that by analysing the table of percentages, you are overlooking this vital aspect of your data.

Make sure the entries in the table you analyse count independent entities

This is a much more subtle problem than that above. It can be a particular nuisance to ophthalmologists (pairs of eyes) and orthopaedic surgeons or rheumatologists (pairs of knees, hips), who have pairs of things to deal with on each patient. However, it is

[‡] As the statistic is computed in a larger table by summing over more cells, the precise value of χ^2 that corresponds to a given P will be different for tables of different sizes or more precisely, for tables with different degrees of freedom: a table with *r* rows and *c* column has (*r*-1)(*c*-1) degrees of freedom.

[†] 95% confidence intervals for population proportion is (0.025,0.56) for 2/10 and (0.18,0.23) for 200/1000

most easily exemplified using an example from accident prevention in community paediatrics. The following is a fictitious dataset but which is based on a real study.

Suppose you want to compare the risk to children walking to school in different areas of a city. In particular you want to compare between two areas the number of uncontrolled major road crossings children have to make in their school journey. To do this you carry out a survey of children arriving at the school gates on a Monday morning in the two areas. You classify the data according to whether or not a child has had to make more than two uncontrolled crossings or not. The data might well be as follows:

	Two or fewer crossings	More than two crossings
Area 1	35	105
Area 2	45	92

Table 10

The usual χ^2 test gives a value of 2.076 and corresponding P value of 0.15. In this table 277 children have been surveyed.

So far all is well. However, suppose that the research workers go out and repeat the survey every day of the week. This could plausibly give a table like the following:

	Two or fewer crossings	More than two crossings
Area 1	171	526
Area 2	214	450

Table 11

This gives a χ^2 value of 9.96 and P=0.002. However, what is the total in the table? Arithmetically it is 1362 but what does it count? The schools in the two areas have not suddenly become about five times bigger, so this is not 1362 children, it is 1362 children-journeys. However, these are not *independent*. Most children will take the same route to school every day, so the entries in the table for the week will necessarily be about five times those for a single day. Absences and changes in route for some children will mean the tables are not exactly in the ratio 5:1, but they are likely to be of this order. However the χ^2 test does assume that five times the information has been collected (e.g. by surveying five times more schools) rather than that more or less the same information has been collected five times over. Therefore it is no surprise that the test statistic is larger and the P-value much more exciting.

Indeed, in the table collected over the week the observed and expected counts will each inevitably be about five times bigger than in the table based on the data for Monday. If the observed and expected values for Monday are denoted by O and E, then the χ^2 statistic for the full week will be (approximately):

$$\chi^2 \approx \sum \frac{(5O-5E)^2}{5E} = \sum \frac{5^2(O-E)^2}{5E} = 5\sum \frac{(O-E)^2}{E} = 5\times\chi^2 \text{ for data for Monday}$$

Given reasonably constant day to day behaviour of the pupils the χ^2 statistic based on the data for the full week is about five times that for the data for Monday. A consequence of this is that whatever the value of the statistic for Monday (other than the very unlikely value of 0) it can be made as large as the investigator wishes (and therefore as statistically significant as wished) simply by observing the same process for as long as necessary.

This is clearly illegitimate because the rates in the two areas are either the same or different, and whether or not this is the case should not be affected by choosing to replicate essentially the same observation arbitrarily often. In practice it can sometimes be awkward to detect when dependent observations are being entered into an analysis in this sort of way. A useful way that often spots problems is to make sure that the total number in any table (i.e. the number which usually appears in the bottom right-hand corner of the table) is equal to the total number of independent units in the analysis. In this example, the number of children at the different schools is the total number of independent units.

Tables with small expected values: Fisher's exact test

In the aortic aneurysm audit the mortality among patients in two age groups undergoing an elective procedure in centres with a low volume is described in the following table.

	Age < 65 yrs	Age ≥ 75 yrs	Total
Discharged dead	2	8	*10*
Discharged alive	72	71	*143*
Total	*74*	*79*	*153*

Table 12

Whether these data support the hypothesis that the death rate is the same in the two age groups could be assessed by the application of a χ^2 test.

If such a test is applied then the value of χ^2 is 3.45 and P = 0.063. The table of expected values is

	Age < 65 yrs	Age ≥ 75 yrs	Total
Discharged dead	4.84	5.16	*10*
Discharged alive	69.16	73.84	*143*
Total	*74*	*79*	*153*

Table 13

One of the expected values is less than 5 and in many packages this will prompt a warning. This will be to the effect that the expected values are too small. This is because the mathematics which allows a P-value be computed from the χ^2 statistic may not be reliable in these circumstances.

The usual response is to employ a method of analysis which does not use this mathematical approximation. The method is known at Fisher's Exact test and differs from the tests mentioned hitherto as it calculates a P-value directly from the data, rather than calculating a test statistic, such as a *t* value or a χ^2 value, and finding a P-value from that.

It is useful to introduce the technique by enumerating all the tables that could occur which have the marginal totals equal to those in Table 12, although it is not necessary to do this when using the method. There are eleven of these, all shown in Table 14. The tables are found by successively increasing and decreasing the entry in the top left cell. In order to maintain the same totals in the margins, the other cells must decrease or increase in concert. The process stops when one of the cells in the table goes to zero.

Probability					
	0	10	6	4	
0.001	74	69	68	75	0.194
	1	9	7	3	
0.011	73	70	67	76	0.099
	2	8	8	2	
0.049	72	71	66	77	*0.032*
	3	7	9	1	
0.131	71	72	65	78	*0.006*
	4	6	10	0	
0.223	70	73	64	79	*0.001*
	5	5			
0.253	69	74			

Table 14

The table we actually observed is outlined with a double line. The other ten possible tables are given, in order of the value of the top left cell.

While all these tables are possible, they are not equally likely. How likely different tables are depends on the relative values of the probabilities of death before discharge. If the two probabilities are the same, i.e. the null hypothesis is true, then the probabilities of the occurrence of each of the eleven possible tables can be calculated mathematically[†]. Each of these probabilities is given next to the corresponding table in Table 14. Note that these eleven values add up to 1, i.e. given these margins you are certain to have one of these tables. Given that the two groups are of similar size, tables with similar numbers of deaths, i.e. tables with 4, 5 and 6 in the top left cell and 6,5 and 4 in the top right cell, are the most likely if the null hypothesis is true.

The P-value testing the null hypothesis is found by adding up those of the eleven probabilities that are less than or equal to the probability of the observed table. These are the italicised probabilities in Table 14, so

$$P = 0.001 + 0.011 + 0.049 + 0.032 + 0.006 + 0.001 = 0.100.$$

[†] see, e.g., Armitage, P, Berry, G & Matthews, JNS (2002) Statistical Methods in Medical Research, 4th ed., Blackwells, Oxford, p.134-137.

(An alternative which has some merit is to take the sum of the probabilities that are smaller than that of observed *and in the same tail*, in this case 0.001+0.011=0.012 and add half the probability of the observed table: the P value is then double this quantity, i.e. 0.049+0.024 = 0.073: a rationale for this mid-P value is in Armitage, Berry and Matthews (2002), see previous footnote).

Some questions naturally arise about the use of Fisher's exact test.

'Exact' sounds attractively precise but what is 'exact' about the exact test? The exact test gives a P-value that is calculated on the basis of the correct, i.e. 'exact' probability distribution for the different tables. In other words there is no need to resort to the kind of asymptotic approximation that underpins the P-values calculated from the χ^2 statistic. However, this precision is rather spoilt because of the irritating lack of agreement among statisticians about the probabilities of which tables should be aggregated to give the right P-value, especially the correct two-sided value. Of course, this represents a disagreement over the underlying principles, which is solved once and for all by the practitioner deciding which principles he or she wishes to use. The user never knows quite when an asymptotic approximation is breaking down.

Why not use the Exact test all the time? If the relevant statistical science had been derived after the advent of today's computing power then we might all be doing this. For tables containing larger counts the list of all possible tables, in the style of Table 14, is very long and before powerful (and accurate) computers were available this would be present an insuperable arithmetical challenge. There is also a more subtle reason why 'exact' methods are not all their name might lead you to believe. Confidence intervals are often derived from related hypothesis tests, and confidence intervals derived from exact tests often are rather wider than they need to be for their nominal level of confidence. So while exact tests might be all we need in principle for tests, the corresponding confidence intervals are often not too good.

When should you use an exact test rather than χ^2 test? The traditional advice is that if any *expected* value in a 2 × 2 table is less than 5 you should use the exact test. In truth this is probably rather strict and the χ^2 method would probably be OK down to much smaller expected values. For larger tables the rule is that if over 20% of the cells have expected values below 5, or any cells with expected values below 1 then the χ^2 method may be unreliable. The problem here is what do you do? Fisher's exact test is for a 2 × 2 table. One approach is to amalgamate categories appropriately, so that the expected values in the new table are larger than these thresholds. An alternative that is becoming more widely available in statistical packages is the exact test for $r \times c$ table. The theory of this test has been known as long as that for the 2 × 2 case but the practice was computationally infeasible. If handled naively the task is still beyond the power of present-day computers. However, over the last 15 years or so advances in clever numerical algorithms have made exact tests on quite large tables entirely feasible. However, specialist advice ought to be sought before embarking on analyses using this software.

Measuring difference between proportions: confidence intervals

Measuring the difference between proportions: part a.

The emphasis thus far has been entirely on testing hypotheses, whereas in practice the presentation of confidence intervals measuring the likely difference between proportions is probably more important. However, before we can do this we need to decide how we are going to measure the difference between proportions.

The main ways of doing this are the absolute difference, the relative risk and the odds ratio. Therefore before the methods for providing confidence intervals for these quantities can be explained a brief digression to introduce odds is needed.

Odds and Probabilities

The parameter for a binary variable, π, is the proportion of the population which possesses the attribute in question and always lies between 0 and 1. Another interpretation is that π is the probability that a randomly selected member of the population has the attribute. So, e.g., if two in every three members of the population do not have blue eyes, then the probability that a member of the population does not have blue eyes is 2/3.

Non-blue eyes
Blue eyes
Non-blue eyes

Another way to express this is to observe that for every member of the population with blue eyes, there are two with eyes of some other colour. In other words there is a chance of 2 to 1 of not being blue-eyed. These are two ways of expressing the same degree of uncertainty: the former is a probability of 2/3, the latter is an odds of 2:1, or simply 2. Notice that while a probability must be between 0 and 1, an odds can take any positive value.

In general, if an event has probability π, then the same event has odds π:(1-π) or simply $\pi/(1-\pi)$. If you set $\pi=2/3$, then the odds is $(2/3)/(1-2/3)=2$, as required. If you know the probability of an event, then you can find the odds. Equally, if you know the odds of an event is η, then the probability is $\eta/(1+\eta)$. In other words the odds and the probability of an event are equivalent ways of quantifying the same degree of uncertainty: if you know one, then you can find the other. Probabilities are always between 0 and 1, odds are bigger than 0.

Odds are quite widely used, especially in epidemiology. They arise naturally through logistic regression and are fundamental to the utility of case-control studies. In the latter application, it is useful to note that if an event is very unlikely, or rare, then the value of the probability is very similar to that of the odds. Why the odds is important in these two areas is beyond our current scope. However, it is important to be able to describe the difference between two odds as well as between two probabilities.

Measuring the difference between proportions: part b.

The three main ways of describing a difference between parameters π_1 and π_2 are:

i)	the absolute difference	$D = \pi_1 - \pi_2$;
ii)	the relative risk	$R = \pi_1/\pi_2$;
iii)	the odds ratio	$OR = \{\pi_1/(1-\pi_1)\}/\{\pi_2/(1-\pi_2)\}$

The null values for these three quantities, i.e. the values which correspond to no difference between the groups are $D = 0$, $R = 1$ and $OR = 1$. The null hypotheses $D = 0$, $R = 1$ and $OR = 1$ are all the same. When the two groups are independent, or unpaired, the χ^2 test is the one which is required.

However, confidence intervals for each of these measures of difference are computed differently. In fact, the differences between the methods for *R* and *OR* are slight, so only the confidence the *OR* is discussed here. Those interested in confidence intervals for *R* should refer to Altman DG (1991, *Practical Statistics for Medical Research*, Chapman & Hall, London, pp.266-267) or to Armitage, Berry & Matthews (2002) p.126-127.

***Confidence intervals for* D**

The data from Table 1, repeated below as Table 15 for convenience, are used to illustrate the computation of a confidence interval for *D*.

	Dead	Alive	*Total*
Not on diuretics	34	655	*689*
On diuretics	25	216	*241*
Total	*59*	*871*	*930*

Table 15 (Table 1 reprinted)

The proportion discharged dead	= 34/689 = 0.0493	(not on diuretics)
	= 25/241 = 0.1037	(on diuretics)

Therefore $D = 0.1037 - 0.0493 = 0.054$ (to three d.p.).

A confidence interval for the difference can be found by extending the ideas in Appendix I. The standard error of *D* can be found as:

$$\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}\ .$$

This cannot be computed as the πs are unknown. An estimate of the standard error can be found by replacing the parameters by estimates, in this case 0.0493 and 0.1037, with corresponding *n*s 689 and 241 respectively. This gives an estimate of the standard error of:

$$\sqrt{\frac{0.0493\times0.9507}{689}+\frac{0.1037\times0.8963}{241}}=0.0213.$$

The 95% confidence interval is found as $D \pm 1.96 \times$ standard error and this is computed as = 0.054 ± 1.96 × 0.0213 = 0.012, 0.096.

In other words, while the estimate of *D* is 0.054, there is a 95% chance it is between 0.012 and 0.096. The interval excludes the null value for *D*, namely 0, as it must when the test of the null hypothesis of no difference between the groups has $P<0.05$.

The value 1.96 used above is what makes this interval a 95% interval. Use of other values will give other confidence coefficients: e.g. a 99% interval results from using 2.58 in place of 1.96.

Confidence intervals for OR

The odds ratio

[note: in this section it avoids tedious circumlocution if we use *OR* to stand both for the *OR*, which strictly speaking is a parameter, and its estimate]

Perhaps the first thing to do is to compute the odds ratio itself. The probability of death before discharge on the non-diuretic group is 34/689, so the odds is 34/(689-34) = 34/655 = 0.0519. In the diuretic group the corresponding figure is 25/216 = 0.1157. Consequently the odds in the diuretic group is 0.1157/0.0519 = 2.230 times larger than the odds in the non-diuretic group: i.e. the odds ratio is *OR* = 2.230.

[note: the *OR* could have been calculated directly as (25/216)/(34/655) = (25×655)/(34×216), i.e. the ratio of the products of the diagonals in the 2 × 2 table]

Notice that if the odds of death is 2.23 times larger in the diuretic than the non-diuretic group, then the odds of death is 1/2.23 = 0.448 times larger in the non-diuretic than in the diuretic group. In other words, if you compare the groups one way round the *OR* might be *x*, whereas if you arbitrarily choose to compare them the other way round the *OR* is simply 1/*x*. There is therefore a connection between any odds ratio greater than 1 and its reciprocal, which is between 0 and 1. The fact that all values from 1 up to infinity are, in some sense equivalent to the values between 0 and 1, gives the distribution of the *OR* a natural skewness.

Confidence interval

A confidence interval for the *OR* is computed using a simple formula. However, there are several points of detail to the procedure which can trip up the careless analyst.

The key point is that the simple formula is not for the *OR* but for the natural logarithm of *OR*, i.e. $\log_e OR$ or ln*OR*.

The standard error of $\log_e OR$ is the square root of the sum of the reciprocals of the entries in the 2 × 2 table underlying the odds ratio: i.e.

$$se(\log_e OR) = \sqrt{\frac{1}{34} + \frac{1}{655} + \frac{1}{25} + \frac{1}{216}} = 0.2749$$

It is important to emphasise that the logarithm must be the *natural* logarithm, for otherwise this formula for the standard error is wrong.

The natural logarithm of *OR* is 0.8019 and the 95% confidence interval for $\log_e OR$ is

$$0.8019 \pm 1.96 \times 0.2749 = 0.2631, 1.3407$$

The confidence interval for the *OR* is found simply by taking the natural antilog of these values (sometimes written as the *exponential* of the values). This gives the values 1.301 and 3.822.

This means that the *OR* is estimated to be 2.23, with a 95% confidence interval 1.30 and 3.82. Notice that the null value for an odds ratio, namely 1, is excluded from the 95% confidence interval – something which follows from the fact that the P-value for the comparison of the groups is < 0.05.

With the confidence intervals met hitherto, the single-value estimate is the mid-point of the ends of the confidence interval. However, this is not the case for the *OR*: the mid-point of the ends of the confidence interval is ½(1.30+3.82)= 2.56, which is larger than the *OR*.

In fact, a moment's thought will demonstrate that this apparent lack of symmetry is a natural consequence of the *OR* scale. Consider a table in which the *OR* is 1, i.e. there is no difference between the groups. If the confidence interval were symmetric, say from ½ to 1½ then the confidence interval for 1/*OR* would be from 2/3 to 2 (i.e. the reciprocals of ½ and 1½). However, 1/*OR* is simply the odds ratio of group B relative to group A rather than group A to group B and because the two groups are indistinguishable the intervals for *OR* and 1/*OR* should be the same. It follows that the interval for *OR* must be of the form $(1/x, x)$ for some $x>1$, for example (½,2). Such intervals are not symmetric about 1 in the sense that they cannot be expressed as $1 \pm y$ for some y (although they are symmetric about 0 on the log scale).

Appendix I: confidence interval for a proportion.

We have seen that an estimator of π is *r/n*. However, earlier in the course the value of estimating parameters not with single numbers but with intervals has been emphasised. The 95% confidence interval for a proportion is constructed along lines very similar to that for a mean, namely:

$$\frac{r}{n} - 1.96 \times \sqrt{\frac{\pi(1-\pi)}{n}}, \frac{r}{n} + 1.96 \times \sqrt{\frac{\pi(1-\pi)}{n}}\ .$$

The problem with this formula is that π is unknown, so we replace it in this formula with an estimate, namely *r*/*n*, giving the new formula as:

$$\frac{r}{n} - \frac{1.96}{n} \times \sqrt{\frac{r(n-r)}{n}}, \frac{r}{n} + \frac{1.96}{n} \times \sqrt{\frac{r(n-r)}{n}}\ .$$

The formula works well if the proportion is not too small or large (between 0.1 and 0.9, say) and *n* is larger than 10, or so. In the extreme case when *r*=*n*, then the formula gives an interval with width of zero, which is not correct. For such extreme outcomes more sophisticated methods are needed.

Appendix II: Why is $\chi^2 = \sum \frac{(O-E)^2}{E}$? (Not Assessed)

The reason we divide (*observed – expected*)^2 by *expected* before summing to obtain the χ^2 statistic is quite subtle. It is to do with how much importance we should attach to the discrepancy between counts and their expected values as the size of the count changes. We will not show precisely why we do this, but will give an indication that shows that it is plausibly a sensible thing to do.

To think about this it is easiest to consider an experiment in which we want to assess whether a coin is fair, i.e. whether it is equally likely to come down heads (H) or tails (T). In experiment 1 we toss a fair coin 30 times and we would ‘expect’ it to come down H 15 times. In experiment 2 we toss the same coin 3000 times, so expect it to come down H 1500 times. However, in the first experiment there is a chance that we will see only ten or fewer H. This event has probability 0.0494, i.e. nearly 5%. It is therefore quite unlikely that we will see so few H if the coin really is fair but it is certainly not a very rare event. What would be a similarly discrepant difference for experiment 2? Two possible events spring to mind.

a) In experiment 1 we saw 5 fewer H than we expected to. So an equally discrepant observation in experiment 2 might be to see 1495 or fewer H. If these two events should be weighted equally in assessing evidence of the fairness of the coin then we would want to give (*observed – expected*)^2 the same weight in both cases when computing the χ^2 statistic. However, in tossing a coin 3000 times there is bound to be more ‘noise’ in the observed number of H than in tossing the coin 30 times, so it perhaps ought to be expected that the number of H may be very likely to differ from 1500 by more than 5, even if the coin is fair. In fact, the chance of seeing 1495 or fewer H in 3000 tosses of a fair coin is 0.435, i.e. nearly 50%. So a discrepancy of the same absolute amount is much less evidence against the fairness of the coin when based on a larger number of tosses. In calculating a χ^2 statistic, the contribution of the (*observed – expected*)^2 from larger *expected* values ought to be increasingly downweighted for increasing expected values.

b) In experiment we saw 10 H in 30 tosses, so an alternative for an an equally discrepant observation in experiment 2 might be if we saw 1000 or fewer heads, i.e. we work out the discrepancy *pro rata* rather than absolutely. If these two events count equally when assessing the fairness then we would weight them equally in computing the χ^2 statistics, so we would have divided (*observed – expected*)^2 by *expected*^{2}. However, in 3000 tosses of a fair coin, the proportion of H gets closer and closer to its true value, so the chance of seeing 1000 or fewer heads will be very, very small indeed (it is, in fact, about 5×10^{-76}). Therefore, a given proportionate squared discrepancy (*observed – expected*)^2/*expected*^{2}, should clearly count more heavily in the calculation of χ^2 if it is based on a larger expected value.

It follows from these remarks that basing χ^2 on the sum of (*observed – expected*)^2 would give relatively too much weight to cells with larger counts, whereas using the sum of (*observed – expected*)^2/*expected*^{2} would give too little weight to these cells. Using (*observed – expected*)^2/*expected* is intermediate between these and is, indeed, the right thing to do.

Appendix III: deriving P-values using the computer.

The table of expected values, such as Table 3b, gives an average table *assuming the null hypothesis is true* but it does not give any indication of how the observed values vary between the tables which can be observed if the null hypothesis is true. The computer can, in fact, be used to generate tables with the same margins as in the above example, with observed values varying as they would under the null hypothesis[†].

The result of asking the computer for 20 such simulated tables is shown below

Simulation 1			χ^2 value	Simulation 11			χ^2 value
	44	645	0.007886		46	643	0.494001
	15	226			13	228	
Simulation 2			χ^2 value	Simulation 12			χ^2 value
	43	646	0.047619		37	652	4.24507
	16	225			22	219	
Simulation 3			χ^2 value	Simulation 13			χ^2 value
	44	645	0.007886		47	642	1.019848
	15	226			12	229	
Simulation 4			χ^2 value	Simulation 14			χ^2 value
	47	642	1.019848		43	646	0.047619
	12	229			16	225	
Simulation 5			χ^2 value	Simulation 15			χ^2 value
	43	646	0.047619		39	650	2.091814
	16	225			20	221	
Simulation 6			χ^2 value	Simulation 16			χ^2 value
	41	648	0.692663		49	640	2.637122
	18	223			10	231	
Simulation 7			χ^2 value	Simulation 17			χ^2 value
	50	639	3.728548		42	647	0.275878
	9	232			17	224	
Simulation 8			χ^2 value	Simulation 18			χ^2 value
	44	645	0.007886		46	643	0.494001
	15	226			13	228	
Simulation 9			χ^2 value	Simulation 19			χ^2 value
	43	646	0.047619		43	646	0.047619
	16	225			16	225	
Simulation 10			χ^2 value	Simulation 20			χ^2 value
	41	648	0.692663		47	642	1.019848
	18	223			12	229	

Next to each simulated table is the value for the χ^2 statistic for that table. If we bring these values together in a histogram, we begin to see how the χ^2 value varies if the null hypothesis is true.

[†] This can be thought of as follows. Assume the computer can be made to toss a biased coin, with chance of heads being 59/930. Ask the computer to generate 689 tosses of this coin, which gives the number of deaths in the non-diuretic group. Asking for a further 241 tosses gives the number in the diuretic group. As the coin has the same probability of coming down heads in the two sets of simulations, the null hypothesis must be true. If the number of deaths in the two groups adds to 59, then you have a simulated table. Otherwise, discard the table and start again. This is a valid way to proceed and shows that the exercise can be accomplished. However, it is not the way used in practice as it is so inefficient – most of the tables generated have to be discarded.

The histogram of these twenty values is shown below in Table A1a.

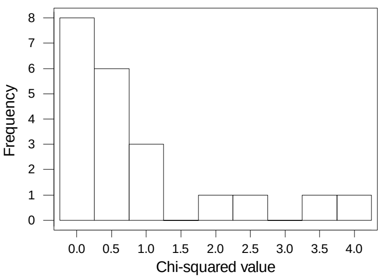

Table A1a Twenty simulated χ^2 values

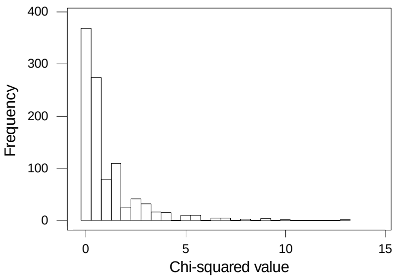

Table A1b One thousand simulated χ^2 values

The range of values obtained from the twenty tables goes up to about 3.7, so the observed value of 8.889 looks as though it may well be unusual if the null hypothesis were true. However, 20 simulated tables may not give a wholly representative picture of the distribution, so the exercise is repeated to give 1000 simulated tables and the histogram of the corresponding 1000 χ^2 values is shown in Table A1b. These values go up to 12.92, so even if the null hypothesis were true, large values of χ^2 do occasionally occur. In fact, only 2 of these 1000 values exceed 8.889, so the observed value is certainly unusual. We can quantify how unusual by computing the proportion of the simulated values which exceed the observed value, namely 2/1000, and this is the P-value testing the null hypothesis of equal pre-discharge death rates in the two groups, i.e. P=0.005. This is not all that different from the value 0.003 obtained using the mathematical approximation.

The precise P-values obtained by simulation vary according to the simulations obtained. However, especially when based on large numbers of simulations, they vary little from run to run. A second determination of the above P-value based on 50000 tables gave P=0.002.

The key point of the exercise is to realise that the simulations, being based on the assumption that the null hypothesis is true, give us an idea of the distribution of values of the test statistics *if the null hypothesis is true*. It therefore allows us to assess the degree of conflict between our data and the assumption that the null hypothesis is true.