The Analysis of categorical data

So far, data have been continuous, usually Normally distributed

Often data are not continuous – i.e. they are recorded on a discrete scale, or are categorical.

Examples are:

ABO blood group: four categories, A, B, AB and O

Tumour stage: often I, II, III, IV

Both categorical, but not the same: latter is *ordinal* – specific statistical methods are complicated

Binary data

A categorical variable with two levels is known as a binary variable: often usefully clinically

Is patient hypoglycaemic?

Does the kidney graft function immediately?

Sometimes you should analyse the underlying (?continuous) variable, such as blood glucose, sometimes the dichotomised version is a more sensible focus.

Binary data: what are the parameters?

Actually only one.

Population comprises two sorts of individuals: those with and those without an attribute; the 0s and the 1s, etc.

The only thing to measure is how many are 1s and how many are not

Parameter is

π = proportion of 1s

Aliter:

π = probability a randomly selected individual is a 1

An example

Audit of hospital mortality from abdominal aortic aneurysm repair

Sample comprises 689 patients, not taking diuretics. Of these, 34 died before they could be discharged

Estimate of π = probability of dying before discharge

= 34/689 = 0.049 or 4.9%

So a simple sample proportion is an estimator of π

Why do we need new methods?

If we score each patient as 1 (died) or 0 (survived) then the proportion is the ‘ordinary’ mean of these 0s and 1s

Need for new methods is based on aspects of spread. Actually there is only one parameter i.e. no separate parameter for spread.

Standard error (continuous)	Standard error (binary)
$\frac{\sigma}{\sqrt{n}}$	$\sqrt{\frac{\pi(1-\pi)}{n}}$

Implications

New formula for SE can be used for confidence interval (see appendix I)

Analogues of *t*-tests etc need to take the mean/spread dependence into account.

Comparing Two independent samples

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

2 × 2 table, with marginal totals

p_1 = estimate of π_1 = 34/689 = 4.9%: p_2 = 25/241 = 10.4%

Comparing two groups: the χ^2 test

In the above example, do the data provide evidence for the equality of π_1 and π_2?

Approach has three components

Assume the null hypothesis is true: i.e. $\pi_1 = \pi_2$

i) construct the table you would 'expect' given this assumption

ii) find out how far this is from the observed table

iii) work out how surprising this distance is

The expected table

A group of 689 and another of 241 patients

Divide these up between dead and alive in the ratio 59:871

	Dead	Alive	*Total*
Not on diuretics	$689 \times \frac{59}{930} = 43.71$	$689 \times \frac{871}{930} = 645.29$	*689*
On diuretics	$241 \times \frac{59}{930} = 15.29$	$241 \times \frac{871}{930} = 225.71$	*241*
Total	*59*	*871*	*930*

Discrepancy: observed vs expected

	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*

Compute $(O\text{-}E)^2/E$ in each cell and add up – gives

$$\chi^2 = \Sigma \frac{(O-E)^2}{E} = 8.889.$$

This is the χ^2 (chi -pronounced '*kye*'-squared) statistic.

The answer is 8.889 – so what?

We need to know:

Is this sort of difference one that occurs all the time in tables in which the null hypothesis is true,

or does it mean something unusual has been observed?

If the latter, we may not be eager to believe the null hypothesis

Suppose we could generate our own tables, each with the margins of the observed table *and* with null hypothesis true

Simulated tables

Simulation 1		χ^2 value
44	645	0.007886
15	226	

Simulation 2		χ^2 value
43	646	0.047619
16	225	

Simulation 3		χ^2 value
44	645	0.007886
15	226	

Simulation 4		χ^2 value
47	642	1.019848
12	229	

Simulation 5		χ^2 value
43	646	0.047619
16	225	

Simulation 6		χ^2 value
41	648	0.692663
18	223	

Simulation 7		χ^2 value
50	639	3.728548
9	232	

Simulation 8		χ^2 value
44	645	0.007886
15	226	

Simulation 9		χ^2 value
43	646	0.047619
16	225	

Simulation 10		χ^2 value
41	648	0.692663
18	223	

Simulation 11		χ^2 value
46	643	0.494001
13	228	

Simulation 12		χ^2 value
37	652	4.24507
22	219	

Simulation 13		χ^2 value
47	642	1.019848
12	229	

Simulation 14		χ^2 value
43	646	0.047619
16	225	

Simulation 15		χ^2 value
39	650	2.091814
20	221	

Simulation 16		χ^2 value
49	640	2.637122
10	231	

Simulation 17		χ^2 value
42	647	0.275878
17	224	

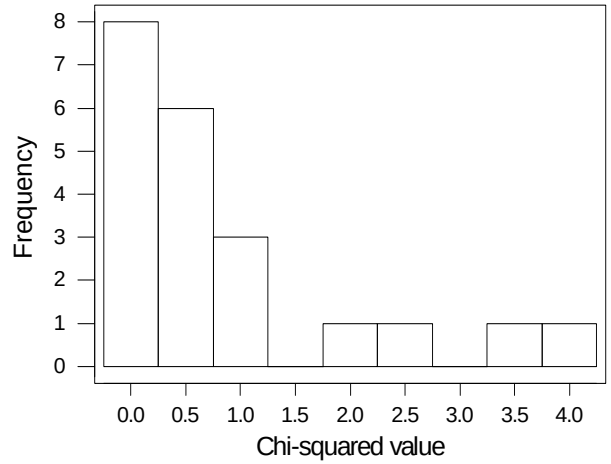
Simulation 18		χ^2 value
46	643	0.494001
13	228	

Simulation 19		χ^2 value
43	646	0.047619
16	225	

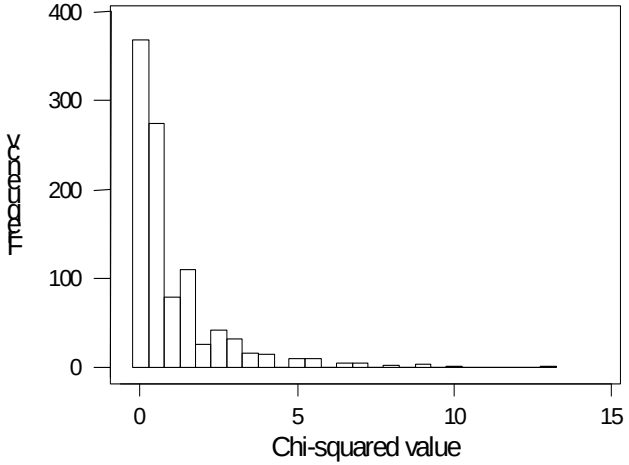
Simulation 20		χ^2 value
47	642	1.019848
12	229	

Next to each simulated table is the corresponding χ^2 value

Distribution of χ^2 (I)



20 χ^2 values



1000 χ^2 values

This allows us to assess what 8.889 means and gives P=0.002

Distribution of χ^2 (II)

- Do you have to do this every time you analyse a table?
- No (although you can if you want to)
- Can use a mathematical approximation instead, allowing a χ^2 value to be turned into a P-value

Larger Tables

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*

Observed table

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*

Expected table, constructed in same way as for 2 × 2 table

The value of χ^2 =0.011 and P = 0.995

Things to remember about the χ^2 test.

- Make sure the elements in the cells are counts, not proportions
- Make sure each independent unit appears once in the table
- Make sure the expected values are not too small

  (<5 in a 2×2 table)

Use counts in the χ^2 test.

Observe 2 items having an attribute from a sample of size 10

Observe 200 items having an attribute from a sample of size 1000

Both give the same proportion, namely 20%

However, in the latter case, the value of 20% is a much more precise estimate of π than in the former.

Comparing groups: it plainly is easier to detect a difference when one is 20 out of 1000 rather than 2 out of 10.

This cannot follow if the table contains only proportions

Independent counts

Survey of children going to school on Monday

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

$\chi^2 = 2.076$ and $P = 0.15$.
In this table 277 children have been surveyed.

If we repeat every day of the week we get

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

$\chi^2 = 9.96$ and $P = 0.002$. But inevitable that χ^2 will increase by about 5 times

Fisher's Exact test.

Aortic aneurysm audit gave mortality among patients in two age groups undergoing an elective procedure is:

	Age < 65 yrs	Age ≥ 75 yrs	Total
Discharged dead	2 *4.84*	8 *5.16*	*10*
Discharged alive	72 *69.16*	71 *73.84*	*143*
Total	*74*	*79*	*153*

Expected values are the lower values.

Expected values smaller than 5 casts doubt on the approximation that gives P from χ^2 (χ^2 = 3.45, P=0.063)

There is a widely used alternative: Fisher's Exact test

Fisher's Exact test II

Enumerate all tables with same margins as observed

Probability					
0.001	0	10	6	4	
	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			

Probabilities are calculated assuming the proportions in the groups are equal

P value is sum of all probabilities ≤ observed values
P = 0.001+0.011+0.049+0.032+0.006+0.001 = 0.100

Fisher's Exact test III

Why not always use this method?

- Too many tables when the counts are large
- No need, χ^2 test adequate
- Exact test does not need approximation but corresponding confidence intervals are often too wide.

When to use?

- 20% of expected values < 5 or any < 1, although this is probably stricter than necessary
- For tables larger than 2×2, analogue of Fisher's Exact test has only just become computationally feasible

Measuring difference: 95% confidence intervals

Three ways of finding a difference between parameters π_1 and π_2:

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

Notice that the null values (value corresponding to no difference between populations) are i) 0, ii) 1 and iii) 1.

Will not consider ii) further, as it is very similar to iii)

95% confidence intervals for *D*

In main example

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.).

The SE of the difference of proportions is:

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

Replacing πs by *estimated* proportions *p* we can evaluate this SE giving 0.0213

95% confidence interval is then $0.054 \pm 1.96 \times 0.0213$ = 0.012, 0.096

95% confidence intervals for *OR*

Need a digression to explain odds, before we can talk about ratios thereof

Return to main example

Probability of death before discharge = 34/689 (non-diuretic group)

Odds of death before discharge = 34/(689-34) = 34/655 = 0.0519

In general, probability of p $\rightarrow$ odds $p/(1-p)$

E.g. probability of 25%, i.e. 1 in 4 (¼) , gives odds of 1 to 3 (= ¼/¾)

Odds of death before discharge = 25/216 = 0.1157 (diuretic group)

So odds is in diuretic group is 0.1157/0.0519 = 2.230 times larger than in the non-diuretic group

i.e. the odds ratio is *OR* = 2.230

95% confidence intervals for *OR*

Way to get 95% confidence interval is easy but has several trip-wires

First, find OR, in our example it is 2.230

Based on a 2 × 2 table

34	655
25	216

OR = (25×655)/(34×216) = 2.230

Second, get *natural* log of OR, i.e. ln(2.230) = 0.8019

SE of ln OR is the square root of the sum of reciprocals of the table, viz.

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

95% confidence intervals for *OR*

Now use this to get 95% confidence interval for *ln OR*

That is

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

Now take natural antilogs, to get confidence interval for *OR*

This gives 1.301 and 3.822

i) Make sure you calculate confidence interval for log of OR

ii) Make sure you use natural logs, otherwise the SE formula is wrong