

Analysing the humble crosstab (contingency table)

John Dawes

Ehrenberg-Bass institute for Marketing

Science 2006

Crosstabs

(and associated tests for dependence)

- Crosstabs are used to examine relationships between *categorical* variables
- Categorical variables - categories, not necessarily forming an ordering or metric difference
- Categorical variables: gender, bought / didn't buy; did/did not etc.

First let's use an elementary 'differences between proportions' test using the normal approximation.

- Suppose we survey 20 men and 20 women and ask them which they prefer - red or blue. 15 men say red and 5 women say red.
- Is there any statistically significant difference in the proportion of 'prefer red' in the two groups.
- We can use the Z test - you would have done this in undergrad stats
- We say the 'null hypothesis' is there is no difference. We compare the two proportions and estimate the probability of obtaining the result we got. If the probability is low, say under 0.05 we 'reject the null hypothesis'.
- We compare the two proportions using a formula. This spits out a "Z" value. The bigger the Z value the less likely the null hypothesis holds.

This is easier than it looks

$$Z = \frac{\text{Prop1} - \text{Prop2}}{\sqrt{P(1-P) * \left(\frac{1}{n1} + \frac{1}{n2}\right)}}$$

Prop 1 is 75% - eg 15 out of 20

Prop 2 is 25% - eg 5 out of 20

P is pooled proportion = 50%

1/n1 is just 1/20 and same for 1/n2

Answer: Z is 3.2

The 'critical value' for the Z distribution is -1.96 or 1.96

when testing at the 95% level of confidence. So 3.2 is way above, we 'reject the null hypothesis' and say there is a statistically significant difference in proportions.

suppose it was like this:

| | prefer red | prefer blue | column total |
|-----------|---------------|----------------|-----------------|
| Male | 15 | 5 | 20 |
| Female | 5 | 15 | 20 |
| Row Total | 20 | 20 | 40 |

Again, how likely is it that you would get 15 out of 20 males prefer red and only 5 of 20 females prefer red if there was no association between gender and colour preference ?

... the chi-square test

- The test for association when we are looking at a contingency table is the chi-square test. It creates 'expected' frequencies for each cell and creates a test statistic based on the differences between the actual frequencies and the expected frequencies.

| | prefer red | prefer blue | column total |
|-----------|------------|-------------|--------------|
| Male | 15 | 5 | 20 |
| Female | 5 | 15 | 20 |
| Row Total | 20 | 20 | 40 |

The 'expected' value is the pooled proportion we saw earlier. The pooled proportion is 50%, so each cell should have '10' in it.

The chi-square statistic is

$$\chi^2_{(r-1)(c-1)} = \sum \frac{(f_{obs} - f_{theo})^2}{f_{theo}}$$

This simply says the chi-square statistic for this contingency table with 2 rows and 2 columns is sum of [(observed-expected) squared divided by expected).

That is, $(15-10)^2/10 + (5-10)^2/10 + (5-10)^2/10 + (15-10)^2/10$

Answer: 10. Look up table for chi-square critical values for 1 d.f.

The probability for this is less than 0.005. Note the square root of 10 is the Z value we got earlier !!!!!

now some real data:

shopping for financial services - is there an association between
'value of product' and 'shopping around' ?

| | did not shop around | did shop around | Row total |
|--------------------|------------------------|--------------------|-----------|
| Low Value product | 188 | 50 | 238 |
| High value product | 68 | 37 | 105 |
| Column Total | 256 | 87 | 343 |

In other words, is the proportion of 'shop around' for low value products close enough to that for 'high value' that it is within random sampling variation ? Props = 21% and 35%

SPSS

VALUE * SHOPPED Crosstabulation

| | | | SHOPPED | | Total |
|-------|------------|-------------------|---------------------|-----------------|--------|
| | | | did not shop around | did shop around | |
| VALUE | low value | Count | 188 | 50 | 238 |
| | | Expected Count | 177.6 | 60.4 | 238.0 |
| | | % within VALUE | 79.0% | 21.0% | 100.0% |
| | | Std. Residual | .8 | -1.3 | |
| | | Adjusted Residual | 2.8 | -2.8 | |
| | high value | Count | 68 | 37 | 105 |
| | | Expected Count | 78.4 | 26.6 | 105.0 |
| | | % within VALUE | 64.8% | 35.2% | 100.0% |
| | | Std. Residual | -1.2 | 2.0 | |
| | | Adjusted Residual | -2.8 | 2.8 | |
| Total | | Count | 256 | 87 | 343 |
| | | Expected Count | 256.0 | 87.0 | 343.0 |
| | | % within VALUE | 74.6% | 25.4% | 100.0% |

The image shows two overlapping SPSS dialog boxes. The top one is the 'Data Editor' window for 'crosstabs preso data.sav', displaying a data table with columns 'value', 'shopped', and 'counts'. The bottom one is the 'Weight Cases' dialog box, where 'value' and 'shopped' are listed as variables, and 'counts' is selected as the frequency variable to weight cases by. The 'Weight cases by frequency variable' radio button is selected.

Chi-Square Tests

| | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|------------------------------------|--------------------|----|-----------------------|----------------------|----------------------|
| Pearson Chi-Square | 7.793 ^b | 1 | .005 | | |
| Continuity Correction ^a | 7.059 | 1 | .008 | | |
| Likelihood Ratio | 7.517 | 1 | .006 | | |
| Fisher's Exact Test | | | | .007 | .004 |
| Linear-by-Linear Association | 7.770 | 1 | .005 | | |
| N of Valid Cases | 343 | | | | |

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 26.63.

| | did not shop around | did shop around | Row total |
|---|------------------------|--------------------|-----------|
| Low Value product - obs <i>expected</i> | 188 <i>177.6</i> | 50 <i>60.4</i> | 238 |
| High Value product - obs <i>expected</i> | 68 <i>78.4</i> | 37 <i>26.6</i> | 105 |
| Column Total | 256 | 87 | 343 |

if there is no association between value and shopping around then the proportion of 'shopped around' should be the same for low value and high value. So we calculate what those proportions 'should be' assuming no association.

Expected value for each cell = (row total x column total) / overall total. Expected values shown in italics.

Note - the pooled proportion for 'did shop' is $87/343 = 25.4\%$, and if this was the case among low value the proportion would be $238 \times 0.254 = 60.4$ which is the expected value.

So we have a 'null hypothesis' that there is no association. If there is no association there should be little difference between observed and expected frequencies in each of the cells.

We use the Pearson chi-squared statistic

This is computed as:

$$\sum \frac{(\text{obs}-\text{exp})^2}{\text{obs}}$$

$$= (188-177.6)^2 / 177.6 + (50-60.4)^2 / 60.4 + \dots \text{etc.}$$

In this case the chi-squared statistic is 7.793

Is this big? It depends on how 'big' the contingency table is.

Here we have a 2x2 table - this equates to 1 degree of freedom.

It is extremely unlikely we would get a statistic like this with one d.f. if there were no association. The probability of getting 7.793 with 1 degree of freedom is 0.007.

“The P value summarises the evidence against the null hypothesis”

