QS101: Introduction to Quantitative Methods in Social Science

Week 14: Crosstabulations and Chi-Squared

Dr. Florian Reiche
Teaching Fellow in Quantitative Methods
Course Director BA Politics and Sociology
Deputy Director of Student Experience and Progression, PAIS

January 29, 2015

Crosstabulations

What is a Crosstabulation (cross tab)?

- A Crosstab (AKA contingency table) serves for the analysis of categorical variables
- It displays the number of subjects observed at all combinations of possible outcomes for the two variables

What does that look like?

Is there an association between gender and ice-cream flavour preference?

	Ice-Cream Flavours		
Gender	Chocolate	Vanilla	Total
Male	10	5	15
Female	8	12	20
Total	18	17	35

The row totals and the column totals are called *marginal distributions*.

Percentage Comparisons

To study how ice-cream flavour preference depends on gender, we convert the frequencies to percentages within each row.

	Ice-Cream Flavours			
Gender	Chocolate	Vanilla	Total	n
Male	66.6%	33.3%	100%	15
Female	40%	60%	100%	20

Percentage Comparisons (contd.)

- The two sets of percentages for males and females are called *conditional distributions* on ice-cream flavour.
- They refer to the sample data distribution of ice-cream flavour, conditional on gender.
- It is practice to form the conditional distribution for the response variable (here ice-cream flavour), within categories of the explanatory variable (here gender).

Good Practice for Cross Tabs

- We want to show the percentages of the response (dependent) variable, in the categories of the explanatory (independent) variable
- The dependent variable goes into the columns
- Clearly label the variable and the categories
- Include the total sample sizes on which the percentages are based

Independence and Dependence

- The question is now: Is there an association between ice-cream flavour and gender?
- Put more technically: are the population conditional distributions on one categorical variable identical at each category of the other variable?
- What would that look like?

Statistical Independence

	Ice-Cream Flavours		
Gender	Chocolate	Vanilla	Total
Male	8 (51.4%)	7 (48.6%)	15 (100%)
Female	10 (51.4%)	10 (48.6%)	20 (100%)

This table is hypothetical – you will never see it.

Queries

- Our initial table was a sample

Queries

- Our initial table was a sample
- We would expect variability depending on the sample we draw

Queries

- Our initial table was a sample
- We would expect variability depending on the sample we draw
- But what does the population look like?

Queries

- Our initial table was a sample
- We would expect variability depending on the sample we draw
- But what does the population look like?
- How plausible, given the sample, is it, that in the population gender and ice-cream flavour are independent?

We need a significance test!

- H_0: The variables are statistically independent
- H_1: The variables are statistically dependent

Chi-Squared Test of Independence

The Chi-Squared Test

- The Chi-Squared (χ^2) test compares the observed frequencies in the contingency table (our initial table) with values that satisfy the null hypothesis
- (The following table shows the observed frequencies, and the expected frequencies if H_0 was true in parentheses.

	Ice-Cream Flavours		
Gender	Chocolate	Vanilla	Total
Male	10 (8)	5 (7)	15
Female	8 (10)	12 (10)	20
Total	18	17	35

How did I calculate the expected values?

- Let f_o denote an observed frequency in a cell of the table.
- Let f_e denote an expected frequency.
- f_e is the count expected in a cell if the variables were independent.
- It equals the product of the row and the column totals for that cell, divided by the total sample size.
- E.g. $15 \times 18/35$

The χ^2 test statistic

$$\chi^2 = \Sigma \frac{f_o - f_e}{f_e} \qquad (1)$$

- We square the difference between the observed and expected frequency in a particular cell, and divide it by the expected frequency
- We sum the result from each cell up (That's what Σ does)
- If H_0 is true, then χ^2 is quite small
- The larger the χ^2 value...

The χ^2 test statistic

$$\chi^2 = \Sigma \frac{f_o - f_e}{f_e} \tag{2}$$

- We square the difference between the observed and expected frequency in a particular cell, and divide it by the expected frequency
- We sum the result from each cell up (That's what Σ does)
- If H_0 is true, then χ^2 is quite small
- The larger the χ^2 value, the greater the evidence against H_0: Independence

How do we interpret the magnitude of χ^2?

- The χ^2 distribution

How do we interpret the magnitude of χ^2?

- The χ^2 distribution
- Concentrated on the positive part of the real line (it cannot be negative!)

How do we interpret the magnitude of χ^2?

- The χ^2 distribution
- Concentrated on the positive part of the real line (it cannot be negative!)
- What is the minimal value and why?

How do we interpret the magnitude of χ^2?

- The χ^2 distribution
- Concentrated on the positive part of the real line (it cannot be negative!)
- What is the minimal value and why?
- It is skewed to the right

How do we interpret the magnitude of χ^2?

- The χ^2 distribution
- Concentrated on the positive part of the real line (it cannot be negative!)
- What is the minimal value and why?
- It is skewed to the right
- The precise shape depends on the *degrees of freedom* (df).

What are degrees of freedom?

- Given the marginal totals, the cell counts in a rectangular block of size $(r-1) \times (c-1)$ within the contingency table determine the other cell counts.

What are degrees of freedom?

- Given the marginal totals, the cell counts in a rectangular block of size $(r-1) \times (c-1)$ within the contingency table determine the other cell counts.
- More helpful: How many cells could I choose at freedom, before the marginal distributions determine the remaining cell values?

Where were we?

HERE!

- The χ^2 distribution
- Concentrated on the positive part of the real line (it cannot be negative!)
- It is skewed to the right
- The precise shape depends on the *degrees of freedom* (df).

The χ^2 Distribution

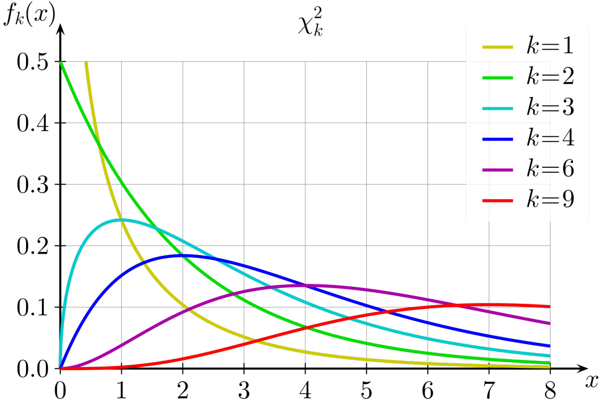

Figure: The χ^2 Distribution (k=df)

Sample Size Requirements

- The χ^2 test is a large sample test
- Ergo: the χ^2 distribution is the sampling distribution of the χ^2 test only if the sample size is large
- Rogh guideline: the expected frequency f_e in each cell should exceed 5

Queries

- How strong is the association if χ^2 is returned significant?
- With this alone, we cannot tell
- We have no idea whether all cells deviate greatly from independence, or only one or two cells do so
- Solution: Agresti and Finlay, Sections 8.3.-8.4. – HOMEWORK!