SO5032 Lecture 1

Brendan Halpin

January 23, 2023

Outline

SO5032 Lecture 1

SO5032 Lecture 1

Lecture 1

Association between categorical variables

- Association between categorical variables: departure from independence
- Visible in patterns of percentages
- Three main questions (cf Agresti/Finlay p265)
    - Is there evidence of association?
    - What is the form of the association?
    - How strong is the association?

The χ^2 test

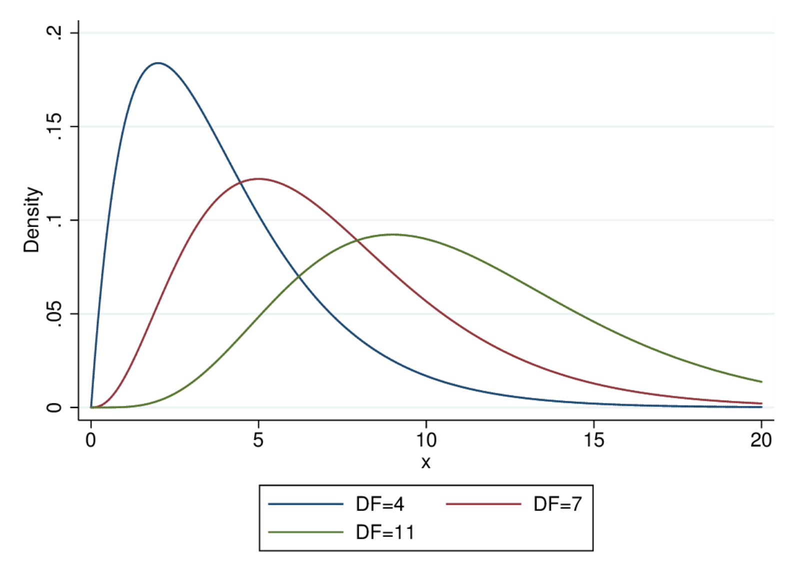
- Compare observed values with expected values under independence:

$$E = \frac{RC}{T}$$

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

- For frequency data, and for large samples the χ^2 statistic has a χ^2 distribution with $df = (r - 1)(c - 1)$
- Interpretation: chance of getting a χ^2 this big or bigger if H_0 (independence) is true in the population

The χ^2 distribution



Limitations of χ^2

- Large sample required: most expected counts 5+
- For frequency or count data, not rates or percentages
- Tests for *evidence* of association, not strength (see Agresti/Finlay Table 8.14, p 268)
- Looks for unpatterned association, may miss weak systematic association between ordinal variables

Pattern of association

- The form association takes is interesting
- We can see it by examining percentages
- Or residuals: $O - E$
- But residuals depend on sample and expected value size

Pearson residuals

- "Pearson residuals" are better:

$$\frac{O - E}{\sqrt{E}}$$

- Square and sum these residuals to get the χ^2 statistic

sociology
UNIVERSITY OF LIMERICK

Adjusted Residuals

- The sum of squared Pearson residuals has a χ^2 distribution, but individually they are not normally distributed
- Adjusted residuals scale to have a standard normal distribution if independence holds:

$$AdjRes = \frac{O - E}{\sqrt{E(1 - \pi_r)(1 - \pi_c)}}$$

- Adjusted residuals outside the range -2 to +2 indicate cells with unusual observed values ($< c5\%$ chance)
- Adjusted residuals outside the range -3 to +3 indicate cells with very unusual observed values

sociology UNIVERSITY OF LIMERICK

Measures of association

- Evidence, pattern, now strength of association
- A number of measures
  - Difference of proportions
  - Odds ratio
  - Risk ratio (ratio of proportions)
- Focus on 2 by 2 pairs, but can be extended to bigger tables

Difference of proportions

No association

	Favour	Oppose	Total
White	360	240	600
Black	240	160	400
Total	600	400	1000

Maximal association

	Favour	Oppose	Total
White	600	0	600
Black	0	400	400
Total	600	400	1000

Difference in proportions

- Difference in proportions (i): $\frac{360}{600} - \frac{240}{400} = 0.6 - 0.6 = 0$
- Difference in proportions (ii): $\frac{600}{600} - \frac{0}{400} = 1 - 0 = 1$
- Range: -1 through 0 (no association) to +1

Relative risk

- “Relative risk” of ratio or proportions is also popular
- The ratio of two percentages:

$$RR = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}$$

where n_{1+} indicates the row-1 total *etc.*

- Range = 0 through 1 (no association) to ∞

sociology
UNIVERSITY OF LIMERICK

Odds ratios

- Odds differ from proportions/percentages:
  - Percentage: $\pi_i = \frac{f_i}{Total}$
  - Odds: $O_i = \frac{f_i}{Total - f_i} = \frac{\pi_i}{1 - \pi_i}$
- Odds ratios are the ratios of two odds:

$$OR = \frac{n_{11}/n_{12}}{n_{21}/n_{22}}$$

- Range: 0 though 1 (no association) to ∞

sociology
UNIVERSITY OF LIMERICK

Odds ratios

- Odds ratio (i): $\frac{\frac{360}{240}}{\frac{240}{160}} = \frac{1.5}{1.5} = 1$
- Odds ratio (ii): $\frac{\frac{600}{0}}{\frac{0}{400}} = \frac{\infty}{0} = \infty$
- Range: 0 through 1 (no association) to $+\infty$

sociology UNIVERSITY OF LIMERICK

Comparing measures

- Difference of proportions is simple and clear
- Ratio of proportions/Relative Risk is also simple
- Odds ratio is less intuitive but turns out to be mathematically more tractable
- DP and RR less consistent across different base levels of “risk”

Ordinal Data

- χ^2 may miss ordinal association
- Symmetric ordinal measures based on concordant and discordant pairs: γ (gamma), Kendall's τ (tau).