Lecture 21: Distributions of Nominal Variables

Nominal Data
Some measurements are just types or categories
   Favorite color, college major, political affiliation, how you get to school, where you’re from
Minimal mathematical structure, but we can still do hypothesis testing
Hypotheses about frequencies or probabilities
   Are all categories equally likely?
   Do two groups differ in their distributions?
   Are two nominal variables related or independent?

Extending the Binomial Test
Binomial test
   Frequency of observations in yes/true category
   Compare to prediction of null hypothesis
Normal approximation
   Treat binomial distribution as Normal
   Convert frequency to z-score
   \[ z = \frac{f - \mu_{\text{freq}}}{\sigma_{\text{freq}}} \]

Multinomial test
Count observations in every category
   Observed frequencies, \( f^{\text{obs}} \)
Convert each to z-score
   \[ z_i = \frac{f_i^{\text{obs}} - \mu}{\sigma} \]
   \( H_0 \) predicts each \( z \) should be near 0

Chi-square statistic
   Sum of squared z-scores
   \[ \chi^2 = \sum z^2 \]
   Measures deviation from null hypothesis
p-value
   Probability of result greater than \( \chi^2 \)
   Uses chi-square distribution
   \( df = k - 1 \) (\( k \) is number of categories)
Counts are not independent; last constrained by rest
Details of z-score

**Expected frequencies: \( f_{\exp} \)**
- Frequency of each category predicted by \( H_0 \)
- Expected value or mean of sampling distribution
- Category probability times number of observations (\( n \))
- If all categories equally likely: \( p = \frac{1}{k}, \ f_{\exp} = \frac{n}{k} \)

**Standard error**
- Denominator of z formula
- Standard deviation of sampling distribution (adjusted for degrees of freedom)

Equals square root of expected frequency: \( \sqrt{f_{\exp}} \)

\[
  z_i = \frac{f_{i,\text{obs}} - f_{i,\text{exp}}}{\sqrt{f_{i,\text{exp}}}}
\]

\[
  \chi^2 = \sum \left( \frac{f_{i,\text{obs}} - f_{i,\text{exp}}}{f_{i,\text{exp}}} \right)^2
\]

**Example: Favorite Colors**
- Choices: Red, Yellow, Green, Blue, Purple
- Are they all equally popular?

**Null hypothesis:** For each color, \( f_{\exp} = \frac{n}{5} \)

**Deviances:** \( f_{i,\text{obs}} - f_{i,\text{exp}} \)

**Squared z-scores:** \( z^2 = \left( \frac{f_{i,\text{obs}} - f_{i,\text{exp}}}{f_{i,\text{exp}}} \right)^2 \)

**Chi-square statistic:** \( \chi^2 = \sum z^2 \)

**Critical value (\( df = 4, \ a = 5\% \)): 9.49**

**Independence of Nominal Variables**
- Are two nominal variables related?
  - Same question as correlation, but need different approach
  - Do probabilities for one variable differ between categories of another?
  - Experimental condition vs. success of learning; sex vs. political affiliation; origin vs. major

**Independent nominal variables**
- Probabilities for each variable unaffected by other
- Example: 80% from CO, 10% psych majors
  - 80% of psych majors are from CO
  - 80%·10% = 8% both psych and from CO

\[
p(x \& y) = p(x):p(y)
\]
Chi-square Test of Independence

Null hypothesis: Variables are independent

Use $H_0$ to calculate expected frequencies

Find observed marginal frequencies for each variable

Total count for each category, ignoring levels of other variable

Multiply marginal frequencies to get expected frequency for combination

\[ p_{x\&y}^{\exp} = p_x \cdot p_y = \frac{f_x^{\text{obs}}}{n} \cdot \frac{f_y^{\text{obs}}}{n} \]
\[ f_{x\&y}^{\exp} = p_{x\&y}^{\exp} \cdot n = \frac{f_x^{\text{obs}} \cdot f_y^{\text{obs}}}{n} \]

Same formula as before:

\[ \chi^2 = \sum \left( \frac{f_{\text{obs}} - f_{\exp}^{\text{exp}}}{f_{\exp}} \right)^2 \]
\[ df = (k_x - 1)(k_y - 1) \]

General Principles of Chi-square Tests

Can use any prediction about data as null hypothesis

Very general approach

Measure goodness of fit

Actually badness of fit

Deviation of data from prediction

Nominal data

Calculate z-score for each frequency within its sampling distribution

Observed minus expected frequency, divided by $\sqrt{f_{\exp}}$

Square zs and sum, to get $\chi^2$

Distribution of one variable; dependence between two variables

Compare GoF to chi-square distribution to get p-value

\[ p = p(\chi^2_{df} > \chi^2) \]

$df$ comes from number of parameters constrained by $H_0$

$k - 1$ for multinomial test; $(k_x - 1)(k_y - 1)$ for independence test