Categorical Data

Katie Pollard

BMI 206

In this unit we will learn …

- Estimating measures of association in 2-way tables
- Testing for association in 2-way tables
- A relationship between GLMs and chi-square tests

Relating Different Data Types

Covariate (independent variable)

Outcome (dependent variable)	Continuous or Both	Categorical
Continuous	Linear Regression / ANCOVA	ANOVA
Categorical	Generalized Linear Model Regression	Contingency Tables / Log-linear Model Regression

Relating Categorical Variables

rs80265967	Disease	No disease
A	1	6721
C	2	2

Association

rs17880490	Disease	No disease
G	360	1981
A	2	11

No association

* joint = product of marginals

1000 Genomes Allele Frequencies, hypothetical disease

Enrichment

Quantifies excess overlap in sets versus expectation under a null distribution (e.g., independence)

- Statistical tests use hypergeometric, binomial, multinomial distributions. Also simulation.

Example: Gene Ontology and RNA-seq

Sets of genes annotated with different ontology terms. For each term, test if genes differentially expressed in cancer vs. healthy are enriched.

Quantifying Enrichment

In a 2x2 table association can be measured in many ways:

- Difference in proportions
- Relative Risk = ratio of two proportions
- Odds Ratio = ratio of two odds
  where odds = $\pi/(1-\pi)$

Can compare rows or columns.

These generalize to IxJ tables.

Conditional Probabilities

Outcomes are independent if the conditional probability equals the marginal probability:

- $P(A \mid B) = P(A)$
- So, $P(A \text{ and } B) = P(A \mid B)\, P(A) = P(A)\, P(B)$

Testing for Independence

In a 2x2 table (generalizes to IxJ) independence can be tested by comparing observed counts to expected counts if no association:

- Pearson's chi-square test
- Binomial test
- Fisher's exact test

Log-linear models

In an IxJ table, expected cell counts (μ_{ij}) can be modeled as a linear function of the categorical variables:

$$\log \mu_{ij} = \mu + \mu^{i} + \mu^{j} + \mu^{ij}$$

- μ is the overall mean $E(n_{ij})=n\pi_{ij}$ (n are counts, π is prob)
- μ^{i} and μ^{j} are row and column effects
- μ^{ij} is interaction (association) of row and column

Independence corresponds to:

- All $\mu^{ij} = 0$.
- Equivalently, $\pi_{ij} = \pi_{i.}\, \pi_{.j}$ or $\mu_{ij} = n\, \pi_{i.}\, \pi_{.j}$ for all i,j.

Can easily extend to 3-way and higher tables…

Categorical Distributions

The distribution for contingency table data depends on the study design (i.e., what values are fixed in sampling):

- Nothing fixed = each cell is Poisson
- Total fixed, but no marginals = single Multinomial (with levels equal to number of cells)
- Row marginals fixed = product-Multinomial (multinomial per row with levels equal to number of columns; binomials if 2 columns)
- Column marginals fixed = product-Multinomial (multinomial per column with levels equal to number of rows; binomials if 2 rows)
- All marginals fixed = single Hypergeometric