Categorical data and contingency tables

Katie Pollard

BMI 206
September 28, 2016

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

- Refers to counts of observations in sets
- Not applicable to quantitative data
- Expectation is relative to a null distribution, e.g.,
  - Independence
  - Background level of dependence
- 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.

Measures of Association

In a 2x2 table (generalizes to IxJ) association can be measured in many ways:

- Difference in proportions between rows (columns)
- Relative Risk = ratio of two proportions
- Odds Ratio = ratio of two odds where odds = p/(1-p)

Testing for independence:

- Pearson's chi-square (Poisson, product-multinomial)
- Fisher's exact test (small counts, fixed marginals)

2x2 Table Examples

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

Categorical Distribution Mathematics

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…

Code Examples