Analysis of Categorical Data

MIT 18.443

Dr. Kempthorne

Spring 2015

Outline

Analysis of Categorical Data

Counts Data: Two-Way Tables

Rosen and Jerdee (1974) Experimental Data

- 48 Male bank supervisors
- Each given a personnel file and decide whether to "Promote" the employee or "Hold File"
- By random assignment 24 evaluated Male employees and 24 evaluated Female employees with results:

	Male	Female
Promote	21	14
Hold File	3	10

Issue: Was there gender bias?
Resolution: Evaluate the chance of such extreme results if no bias

Counts Data: Two-Way Tables

McGarrell, E., Olivares, K., Crawford, K. & Kroovand, N. 2000)
Recidivism Study of 458 juvenile offenders in Indianapolis

- Study used an experimental design with random assignment of juveniles to experimental intervention:
  Family Group Counseling, 232 subjects
  Control group (diversion programs), 226 subjects).
- At six months, 46 subjects in the experimental intervention group re-offended, and 77 in the control group re-offended.

	Re-Offended	No Re-Offence	
FGC Group	46	186	232
Control Group	77	149	226

Issue: Are the recidivism rates significantly different?
Resolution: Evaluate the chance of such extreme results if no difference

Counts Data: Two-Way Tables

De Veaux, Velleman and Bock (2014): U. Texas Study

- 626 people treated for non-blood-related diseases
- Subjects categorized by two variables

Hepatitis C Status and Tatoo Status

	Hepatitis C	No Hepatitic C	Total
Tattoo, Parlor	17	35	52
Tattoo, Elsewhere	8	53	61
No Tattoo	22	491	513
Total	47	579	626

Issues: Is risk of hepatitis C related to having a tatoo?
Is risk related to where they got their tatoos?

Resolution: Evaluate the *independence* of the two factors

Chi-Square Tests: Three Problem Types

Chi-Square Goodness-of-Fit Test

- A single categorical variable is measured on one population.
- Does the sample distribution match the distribution predicted by a model?

Chi-Square Test of Homogeneity

- A single categorical variable is measured independently on two or more populations.
- Are the distributions for different populations the same?

Chi-Square Test of Independence

- Two categorical variables are measured on the same population.
- Are the two variables independent?

Chi-Square Tests

Assumptions and Conditions

- Sample of counts data
- Are individual members of counts independent of each other?
- Expected values are at least 5 in each cell.
- If generalizing from the data to some population, is sample representative?

  Is sample smaller than 10% of population?

Fisher's Exact Test

Fisher's Exact Test for Two-Way Tables of Counts

- Testing Independence row and column categories
- Exact p-values when cell counts are small

Example:

	Male	Female
Promote	21	14
Hold File	3	10

Under Null Hypothesis of Independence, model Males' data as:

- A random sample of size 24 from a finite population of 48 outcomes

  35 successes (Promote) and 13 failures (Hold File)
- Sampling *without* replacement
- The Females' data are the unsampled outcomes
- **Test Statistic:** X = number of Males promoted.

Fisher's Exact Test

Distribution of Test Statistic X

	Male	Female	Total
Promote	x	$35 - x$	35
Hold File	$24 - x$	$x - 11$	13
Total	24	24	48

- With $x = 21$ we get the observed data:

	Male	Female
Promote	21	14
Hold File	3	10

- Smallest value of x is 11 and Largest value is 24.
- $X \sim Hypergeometric(k = 24, m = 35, n = 13)$
- Values more extreme than $x = 21$:

$$R_x = \{21, 22, 23, 24\} \cup \{11, 12, 13, 14\}$$

- P-value $= P(X \in R_x)$

Definition: Hypergeometric Distribution

$$X \sim Hypergeometric(k = 24, m = 35, n = 13)$$

- Sampling without replacement from an Urn
- Urn has m white balls
- Urn has n black balls
- k is the number of balls drawn
- X is the number of white balls drawn

The pdf of X is given by:

$$P(X = x \mid m, n, k) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}$$

Note:

- $x \leq k$ and $x \leq m$ so $x \leq min(k, m) = 24$.
- $(k - x) \leq n$ and $(k - x) \leq k$ so $x \geq max(k - n, 0) = 11$

18.443 Statistics for Applications

Spring 2015