Chapter 3

Simple Inference for Categorical Data

Categorical Data

For a single variable:

- Just look at counts
- Number in Group A, B, C, etc.
- Could use graphics like pie and bar charts
- Can infer proportions of the groups in the broader population

Example: Locations in Water Data Set

- Get frequencies of north and south in the data
- Create bar charts or pie charts to see proportions

Categorical Data

For multiple categorical variables:

- Data is cross-classified
- Counts in cells of contingency tables
- Looking for indications of association between categorical variables
- For instance, Democrats more or less likely to be in favor of some political issue

Example: Groupings in Water Data

- Suppose we only know if the towns have mortalities and concentrations above or below the median
- Cross-classify the counts based on these groupings using **proc freq**
- See the expected counts
- Is there significant association between mortality and concentration groups?

Association

- Akin to the concept of correlation
- The null hypothesis is no association
- Under this hypothesis, expect counts to simply be products of proportions, e.g.

$$E(N_{Democrat\ in\ Favor}) = n_{total}\frac{n_{Democrat}}{n_{total}}\frac{n_{in\ Favor}}{n_{total}}$$

- Hypothesis tests based on deviations from expected counts

Chi-Square and Exact Tests

- Pearson's Chi-Square basis for several measures
- Add **chisq** option to get these measures (also includes Fisher's exact test for 2x2)
- See **Details>Statistical Computations>Chi-Square Tests and Statistics** for details of statistics and related measures

Fisher's Exact Test

- For 2x2 table, n_{11} determines other cells (given row, column, and table sums)
- Under null, expect n_{11} to follow a hypergeometric distribution exactly
- Left-tail alternative says we should expect fewer in cell (1,1), $\boldsymbol{P(\mathbf{F} \leq n_{11})}$.
- Right-tail says we should expect more, $\boldsymbol{P(\mathbf{F} \geq n_{11})}$.

Fisher's Exact Test

- Two-sided alternative says there is some association
- P-value in 2x2 table based on less likely cell (1,1) counts
- Extends to RxC tables by summing probabilities of less likely table configurations

Example: Sandflies

- Male and female flies
- Traps set at different heights
- Data given as counts
- Use **weight** statement in **proc freq**
- Proceed as before

Risks and Risk Differences

- Risks are row (or column) proportions
- Can estimate these from cell counts and row (or column) sums
- Can obtain confidence intervals for the risks
- Can compare differences of risks (2x2 case)
- Looking at significant differences of probability in the broader population
- Use **riskdiff** option to **tables** statement

Example: Grouped Water Data

- Column risk and risk difference estimates and intervals
- Flip table to get same for the rows

Exercise: Back to Sandflies

- Obtain row and column percentages
- Comment on significance of association
- Do there appear to be any differences within rows or columns?
- Obtain confidence intervals for row and column risks and comment on significant differences

Exercise: Acacia Ants

- Two species of acacia trees (A and B)
- Test whether one species is more likely to be invaded by ants
- Obtain expected counts. Is independence reasonable?
- Test for association
- Obtain risk estimates to see if difference is significant

Exercise: Oral Cancer

- Geographic regions of India (3 regions)
- Types of oral cancer (9 types)
- Get a contingency table (just frequencies)
- Do counts suggest association?
- Do measures suggest association?
- Do tests reject hypothesis of independence?
- Need **exact** option to get Fisher's exact test

Some Additional Topics

- Measures and tests of agreement
- Relative risk
- Odds ratios (and odds)