Categorical Data Analysis: Course Overview

Michael Friendly
Psych 6136
September 4, 2017

Course goals

This course is designed as a broad, applied introduction to the statistical analysis of categorical (or discrete) data, with an emphasis on:

**Emphasis: visualization methods**
- exploratory graphics: see patterns, trends, anomalies in your data
- model diagnostic methods: assess violations of assumptions
- model summary methods: provide an interpretable summary of your data

**Emphasis: theory ⇒ practice**
- Understand how to translate research questions into statistical hypotheses and models
- Understand the difference between simple, non-parametric approaches (e.g., $\chi^2$ test for independence) and model-based methods (logistic regression, GLM)
- Framework for thinking about categorical data analysis in visual terms

Course outline

1. Exploratory and hypothesis testing methods
   - Week 1: Overview; Introduction to R
   - Week 2: One-way tables and goodness-of-fit test
   - Week 3: Two-way tables: independence and association
   - Week 4: Two-way tables: ordinal data and dependent samples
   - Week 5: Three-way tables: different types of independence
   - Week 6: Correspondence analysis

2. Model-based methods
   - Week 7: Logistic regression I
   - Week 8: Logistic regression II
   - Week 9: Multinomial logistic regression models
   - Week 10: Log-linear models
   - Week 11: Loglinear models: Advanced topics
   - Week 12: Generalized Linear Models: Poisson regression
   - Week 13: Course summary & additional topics

Textbooks

**Main texts:**

**Supplementary readings:**
For those who desire a more in-depth treatment of categorical data analysis:
What is categorical data?

A **categorical variable** is one for which the possible measured or assigned values consist of a **discrete set of categories**, which may be **ordered** or **unordered**.

Some typical examples are:
- **Gender**, with categories “Male”, “Female”.
- **Marital status**, with categories “Never married”, “Married”, “Separated”, “Divorced”, “Widowed”.
- **Treatment outcome**, with categories “no improvement”, “some improvement”, or “marked improvement”.
- **Number of children**, with categories 0, 1, 2, . . . .

Categorical data structures: 1-way tables

**Simplest case: 1-way frequency distribution**
- **Unordered factor**

<table>
<thead>
<tr>
<th>Hair color</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>108</td>
</tr>
<tr>
<td>Brown</td>
<td>286</td>
</tr>
<tr>
<td>Red</td>
<td>71</td>
</tr>
<tr>
<td>Blond</td>
<td>127</td>
</tr>
</tbody>
</table>

Hairstyles among 592 students

<table>
<thead>
<tr>
<th>Party</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>BQ</td>
<td>104</td>
</tr>
<tr>
<td>Cons</td>
<td>392</td>
</tr>
<tr>
<td>Green</td>
<td>126</td>
</tr>
<tr>
<td>Liberal</td>
<td>404</td>
</tr>
<tr>
<td>NDP</td>
<td>174</td>
</tr>
</tbody>
</table>

Voting intentions in Harris-Decima poll, 08/21/08

**Questions:**
- Are all hair colors equally likely?
- Do blondes have more fun?
- Is there a difference in voting intentions between Liberal and Conservative?

Categorical data structures: 1-way tables

Even here, simple graphs are better than tables

**Bar Graphs**

- **Count**
  - Hair color:
    - Black, Brown, Red, Blond
  - # of sons in Saxony families with 12 children

**Questions:**
- What is the form of this distribution?
- Is it useful to think of this as a **binomial distribution**?
- If so, is Pr(male) = .5 reasonable?
- How could so many families have 12 children?

But these don’t really provide answers to the questions. Why?
What is categorical data?

Categorical data structures

When a particular distribution is in mind,
- better to plot the data together with the fitted frequencies
- better still: a hanging rootogram—plot frequencies on sqrt scale, and hang the bars from the fitted values.

Categorical data structures: 1-way tables

Contingency tables (2 x 2 x …)

- Two-way

Admission to graduate programs at UC Berkeley

Questions:
- Is admission associated with gender?
- Does admission rate vary with department?

Categorical data structures: Larger tables

Table and case-form

- The previous examples were shown in table form
  - # observations = # cells in the table
  - variables: factors + COUNT
- Each has an equivalent representation in case form
  - # observations = total COUNT
  - variables: factors
- Case form is required if there are continuous variables
- Case form is tidy
Categorical data: Analysis methods

Methods of analysis for categorical data fall into two main categories:

Non-parametric, randomization-based methods
- Make minimal assumptions
- Useful for hypothesis-testing:
  - Are men more likely to be admitted than women?
  - Are hair color and eye color associated?
  - Does the binomial distribution fit these data?
- Mostly for two-way tables (possibly stratified)
- R:
  - Pearson Chi-square: `chisq.test()`
  - Fisher's exact test (for small expected frequencies): `fisher.test()`
  - Mantel-Haenszel tests (ordered categories: test for linear association): `CMHtest()`
- SAS: PROC FREQ — can do all the above
- SPSS: Crosstabs

Model-based methods
- Must assume random sample (possibly stratified)
- Useful for estimation purposes: Size of effects (std. errors, confidence intervals)
- More suitable for multi-way tables
- Greater flexibility; fitting specialized models
  - Symmetry, quasi-symmetry, structured associations for square tables
  - Models for ordinal variables
- R: `glm()` family, Packages: car, gnm, vcd, ...
  - estimate standard errors, covariances for model parameters
  - confidence intervals for parameters, predicted Pr\{response\}
- SAS: PROC LOGISTIC, CATMOD, GENMOD, INSIGHT (Fit YX), ...
- SPSS: Hiloglinear, Loglinear, Generalized linear models

Categorical data: Response vs. Association models

Response models
- Sometimes, one variable is a natural discrete response.
- Q: How does the response relate to explanatory variables?
  - Admit ~ Gender + Dept
  - Party ~ Age + Education + Urban

⇒ Logit models, logistic regression, generalized linear models

Association models
- Sometimes, the main interest is just association among variables
- Q: Which variables are associated, and how?
  - Berkeley data: [Admit Gender]? [Admit Dept]? [Gender Dept]
  - Hair-eye data: [Hair Eye]? [Hair Sex]? [Eye, Sex]

⇒ Loglinear models

This is similar to the distinction between regression/ANOVA vs. correlation and factor analysis

Graphical methods: Tables and Graphs

If I can’t picture it, I can’t understand it. — Albert Einstein

Getting information from a table is like extracting sunlight from a cucumber. — Farquhar & Farquhar, 1891

Tables vs. Graphs
- Tables are best suited for look-up and calculation—
  - read off exact numbers
  - show additional calculations (e.g., % change)
- Graphs are better for:
  - showing patterns, trends, anomalies,
  - making comparisons
  - seeing the unexpected!
- Visual presentation as communication:
  - what do you want to say or show?
  - design graphs and tables to ‘speak to the eyes’
Graphical methods: Communication goals

Different audiences require different graphs:

- **Presentation**: A single, carefully crafted graph to appeal to a wide audience
- **Exploration, analysis**: Many related graphics from different perspectives, for a narrow audience (often: you!)

**Graphical methods: Presentation goals**

Different presentation goals appeal to different design principles

<table>
<thead>
<tr>
<th>Basic functions of data display</th>
<th>Presentation Goal</th>
<th>Design Principles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis</td>
<td>Perception</td>
<td>Reconnaissance</td>
</tr>
<tr>
<td>Exploration</td>
<td>Detection</td>
<td>Exploration</td>
</tr>
<tr>
<td>Model building</td>
<td>Comparison</td>
<td>Diagnosis</td>
</tr>
<tr>
<td>to Simulate</td>
<td>Aesthetics</td>
<td>Model building</td>
</tr>
<tr>
<td>to Persuade</td>
<td>Rhetoric</td>
<td>to Simulate</td>
</tr>
<tr>
<td>to Inform</td>
<td>Exposition</td>
<td>to Persuade</td>
</tr>
</tbody>
</table>

**Graphical methods: Quantitative data**

Quantitative data (amounts) are naturally displayed in terms of **magnitude ~ position along a scale**

- Scatterplot of Income vs. Experience
- Boxplot of Income by Gender

**Graphical methods: Categorical data**

Frequency data (counts) are more naturally displayed in terms of **count ~ area** (Friendly, 1995)

- Fourfold display for $2 \times 2$ table
- Mosaic plot for 3-way table
Principles of Graphical Displays

**Effect ordering** (Friendly and Kwan, 2003) — In tables and graphs, sort unordered factors according to the effects you want to see/show.

**Table:** Hair color - Eye color data: Effect ordered

<table>
<thead>
<tr>
<th>Eye color</th>
<th>Black</th>
<th>Brown</th>
<th>Red</th>
<th>Blond</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>68</td>
<td>119</td>
<td>26</td>
<td>7</td>
</tr>
<tr>
<td>Hazel</td>
<td>15</td>
<td>54</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>Green</td>
<td>5</td>
<td>29</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>Blue</td>
<td>20</td>
<td>84</td>
<td>17</td>
<td>94</td>
</tr>
</tbody>
</table>

Model: Independence: $\chi^2 (9) = 138.29$

Color coding:

<table>
<thead>
<tr>
<th>n in each cell:</th>
<th>$n &lt; \text{expected}$</th>
<th>$n &gt; \text{expected}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$&lt; -4$</td>
<td>$&gt; 4$</td>
</tr>
<tr>
<td></td>
<td>$&lt; -2$</td>
<td>$&gt; 2$</td>
</tr>
<tr>
<td></td>
<td>$&lt; -1$</td>
<td>$&gt; 1$</td>
</tr>
<tr>
<td></td>
<td>$0$</td>
<td></td>
</tr>
</tbody>
</table>

"Corrgrams: Exploratory displays for correlation matrices" (Friendly, 2002)

Clustered heat map: Showing patterns in tables

The clustered heat map is one method for making large tables more visually understandable.

- Social statistics from UN survey
- Rows and columns are sorted, using cluster analysis
- Standardized data values are encoded using color

Bertifier: Turning tables into graphics

Bertifier: A web app implementing Bertin’s idea of the reorderable matrix. See: [http://www.aviz.fr/bertifier](http://www.aviz.fr/bertifier)

A table: Attitudes and attributes by country
- Values encoded by size and shape
- Sorted and grouped by themes and country regions

Watch: [Youtube video of Bertifier](https://www.youtube.com/watch?v=example_video_id)
Visual comparisons

Comparisons — Make visual comparisons easy

- Visual grouping — connect with lines, make key comparisons contiguous
- Baselines — compare data to model against a line, preferably horizontal
- Frequencies often better plotted on a square-root scale

Standard histogram with fit

Suspended rootogram

Small multiples — combine stratified graphs into coherent displays (Tufte, 1983)

- e.g., scatterplot matrix for quantitative data: all pairwise scatterplots

Graphical methods: Categorical data

Exploratory methods

- Minimal assumptions (like non-parametric methods)
- Show the data, not just summaries
- But can add summaries: smoothed curve(s), trend lines, ...
- Help detect patterns, trends, anomalies, suggest hypotheses

Plots for model-based methods

- Residual plots - departures from model, omitted terms, ...
- Effect plots - estimated probabilities of response or log odds
- Diagnostic plots - influence, violation of assumptions
References I


