Categorical Data Analysis: Course Overview

Michael Friendly

Psych 6136

January 12, 2015
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 $\Rightarrow$ 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</th>
<th>Black</th>
<th>Brown</th>
<th>Red</th>
<th>Blond</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>108</td>
<td>286</td>
<td>71</td>
<td>127</td>
</tr>
</tbody>
</table>

Hair color among 592 students

<table>
<thead>
<tr>
<th>Party</th>
<th>BQ</th>
<th>Cons</th>
<th>Green</th>
<th>Liberal</th>
<th>NDP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>104</td>
<td>392</td>
<td>126</td>
<td>404</td>
<td>174</td>
<td>1200</td>
</tr>
<tr>
<td>%</td>
<td>8.7</td>
<td>32.6</td>
<td>10.5</td>
<td>33.7</td>
<td>14.5</td>
<td>100</td>
</tr>
</tbody>
</table>

Voting intention in Harris-Decima poll, 8/21/08

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

But these don’t really provide answers to the questions. Why?
Categorical data structures

Simplest case: 1-way frequency distribution

- Ordered, quantitative factor

<table>
<thead>
<tr>
<th>nMales</th>
<th># of sons in Saxony families with 12 children</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>104</td>
</tr>
<tr>
<td>3</td>
<td>286</td>
</tr>
<tr>
<td>4</td>
<td>670</td>
</tr>
<tr>
<td>5</td>
<td>1033</td>
</tr>
<tr>
<td>6</td>
<td>1343</td>
</tr>
<tr>
<td>7</td>
<td>1112</td>
</tr>
<tr>
<td>8</td>
<td>829</td>
</tr>
<tr>
<td>9</td>
<td>478</td>
</tr>
<tr>
<td>10</td>
<td>181</td>
</tr>
<tr>
<td>11</td>
<td>45</td>
</tr>
<tr>
<td>12</td>
<td>7</td>
</tr>
</tbody>
</table>

Questions:

- What is the *form* of this distribution?
- Is it useful to think of this as a *binomial distribution*?
- If so, is $\Pr(\text{male}) = 0.5$ reasonable?
- How could so many families have 12 children?
Categorical data structures: 1-way tables

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: 2x2 tables

Contingency tables \((2 \times 2 \times \ldots)\)

- Two-way

<table>
<thead>
<tr>
<th></th>
<th>Gender</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Admit</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Admitted</strong></td>
<td></td>
<td>1198</td>
<td>557</td>
</tr>
<tr>
<td><strong>Rejected</strong></td>
<td></td>
<td>1493</td>
<td>1278</td>
</tr>
</tbody>
</table>

Admission to graduate programs at UC Berkeley

- Three-way, stratified by another factor

... by Department

<table>
<thead>
<tr>
<th></th>
<th>Dept</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Admit Gender</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Admitted Male</strong></td>
<td>512</td>
<td>353</td>
<td>120</td>
<td>138</td>
<td>53</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td><strong>Female</strong></td>
<td>89</td>
<td>17</td>
<td>202</td>
<td>131</td>
<td>94</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td><strong>Rejected Male</strong></td>
<td>313</td>
<td>207</td>
<td>205</td>
<td>279</td>
<td>138</td>
<td>351</td>
<td></td>
</tr>
<tr>
<td><strong>Female</strong></td>
<td>19</td>
<td>8</td>
<td>391</td>
<td>244</td>
<td>299</td>
<td>317</td>
<td></td>
</tr>
</tbody>
</table>

Questions:

- Is admission associated with gender?
- Does admission rate vary with department?
What is categorical data?

Categorical data structures: Larger tables

Contingency tables (larger)

- Two-way

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

- Three-way

<table>
<thead>
<tr>
<th>Sex</th>
<th>Hair</th>
<th>Eye</th>
<th>Brown</th>
<th>Blue</th>
<th>Hazel</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Black</td>
<td>32</td>
<td>11</td>
<td>10</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Brown</td>
<td>53</td>
<td>50</td>
<td>25</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Red</td>
<td>10</td>
<td>10</td>
<td>7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Blond</td>
<td>3</td>
<td>30</td>
<td>5</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>Black</td>
<td>36</td>
<td>9</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Brown</td>
<td>66</td>
<td>34</td>
<td>29</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Red</td>
<td>16</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Blond</td>
<td>4</td>
<td>64</td>
<td>5</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>
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
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
Categorical data: Analysis methods

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 \( \text{Pr}\{\text{response}\} \)
- 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 \sim Gender + Dept
  - Party \sim Age + Education + Urban

\Rightarrow Logit models, logististic 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]

\Rightarrow Loglineal 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.*

*Getting information from a table is like extracting sunlight from a cucumber.*

**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

**Basic functions of data display**

- **Primary Use**
  - Analysis
  - Presentation

- **Presentation Goal**
  - Reconnaissance
  - Exploration
  - Diagnosis
  - Model building
  - to Simulate
  - to Persuade
  - to Inform

- **Design Principles**
  - Perception
  - Detection
  - Comparison
  - Aesthetics
  - Rhetoric
  - Exposition
Graphical methods: Quantitative data

Quantitative data (amounts) are naturally displayed in terms of magnitude $\sim$ 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 \(\text{count} \sim \text{area}\) (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.

“Corrgrams: Exploratory displays for correlation matrices” (Friendly, 2002)
**Effect ordering and high-lighting for tables**

**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*: [Hair][Eye] $\chi^2 (9) = 138.29$

**Color coding:**

-4<br/>
-2<br/>
-1<br/>
0<br/>
>1<br/>
>2<br/>
>4

**n in each cell:**

- $n < \text{expected}$
- $n > \text{expected}$
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

1. Values encoded by size and shape
2. Sorted and grouped by themes and country regions

Watch: [Youtube video of Bertifier](http://www.youtube.com/watch?v=1234567890)
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](image1.png) ![Suspended rootogram](image2.png)

- 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

```
<table>
<thead>
<tr>
<th></th>
<th>Prestige</th>
<th>Educ</th>
<th>Income</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prestige</td>
<td>14.8</td>
<td>6.38</td>
<td>611</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>87.2</td>
<td>15.97</td>
<td>25879</td>
<td>97.51</td>
</tr>
</tbody>
</table>
```
e.g., mosaic matrix for quantitative data: all pairwise mosaic plots
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


