Course goals

Emphasis: visualization methods
- Basic ideas: categorical vs. quantitative data
- Some novel displays: sieve diagrams, fourfold displays, mosaic plots, ...
- Some that extend more familiar ideas to the categorical data setting.

Emphasis: theory ⇒ practice
- Show what can be done, in both SAS and R (most in SAS)
- Framework for thinking about categorical data analysis in visual terms
- Provide software tools you can use

What is included, and what is not
- Some description of statistical methods— only as necessary
- Many software examples— only explained as necessary
- Too much material— some skipping may be required

Course structure, Parts 1–3

1. Overview and introduction
   - Categorical data? Graphics?
   - Discrete distributions
   - Testing association

2. Visualizing two-way and n-way tables
   - 2 × 2 tables; r × c tables: Fourfold & sieve diagrams
   - Observer agreement: Measures and graphs
   - Correspondence analysis

3. Mosaic displays and loglinear models
   - n-way tables: graphs and models
   - Mosaics software
   - Structured tables

Course structure, Parts 4–5

4. Logit models and logistic regression
   - Logit models; logistic regression models
   - Effect plots
   - Influence and diagnostic plots

5. Polytomous response models
   - Proportional odds models
   - Nested dichotomies
   - Generalized logits
What is categorical data?

- 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>

<table>
<thead>
<tr>
<th>Party</th>
<th>PQ</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>

- Ordered, quantitative factor

<table>
<thead>
<tr>
<th>Hair color among 592 students</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Party</th>
<th>PQ</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>

Overview: What is categorical data?

- Contingency tables (2 × 2 × ...)
  - Two-way

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

- Three-way

<table>
<thead>
<tr>
<th>Admission to graduate programs at UC Berkeley</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<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>Admit</td>
<td>512</td>
<td>353</td>
<td>120</td>
<td>138</td>
<td>53</td>
<td>22</td>
</tr>
<tr>
<td>Female</td>
<td>89</td>
<td>17</td>
<td>202</td>
<td>131</td>
<td>94</td>
<td>24</td>
</tr>
<tr>
<td>Rejected</td>
<td>313</td>
<td>207</td>
<td>205</td>
<td>279</td>
<td>138</td>
<td>351</td>
</tr>
<tr>
<td>Female</td>
<td>19</td>
<td>8</td>
<td>391</td>
<td>244</td>
<td>299</td>
<td>317</td>
</tr>
</tbody>
</table>

Overview: What is categorical data?

- Table and case-form

<table>
<thead>
<tr>
<th>Hair color among 592 students</th>
</tr>
</thead>
</table>

- 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

<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>9</td>
<td></td>
</tr>
</tbody>
</table>

<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>9</td>
<td></td>
</tr>
</tbody>
</table>
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 A and B associated?
- Mostly for two-way tables (possibly stratified)
- SAS: PROC FREQ
  - Pearson Chi-square
  - Fisher’s exact test (for small expected frequencies)
  - Mantel-Haenszel tests (ordered categories: test for linear association)
- R: `chisq.test()`, `mantelhaen.test()`, ...
- 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
- SAS: PROC LOGISTIC, CATMOD, GENMOD, INSIGHT (Fit YX)
  - estimate standard errors, covariances for model parameters
  - confidence intervals for parameters, predicted Pr{response}
- R: `glm()` family, car package, gnm package, ...
- 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
- 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.*

*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
  - 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: 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 2x2 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>
<thead>
<tr>
<th>Eye color</th>
<th>Hair color</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>Brown</td>
</tr>
<tr>
<td>Brown</td>
<td>68</td>
</tr>
<tr>
<td>Hazel</td>
<td>15</td>
</tr>
<tr>
<td>Green</td>
<td>5</td>
</tr>
<tr>
<td>Blue</td>
<td>20</td>
</tr>
</tbody>
</table>

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

Color coding:
- $<-4$<
- $<-2$<
- $<-1$<
- $0$>1
- $>2$>4

$n$ in each cell:
- $n <$ expected
- $n >$ expected

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

- **Comparisons**— Make visual comparisons easy
  - Visual grouping— connect with lines, make key comparisons contiguous
  - Baselines— compare *data to model* against a line, preferably horizontal

- **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 *summarizes*
    - 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
  - Goals
    - *VCD* and R *vcd* package - Make these methods *available* and *accessible* in SAS & R
    - *Practical power* = *Statistical power* × *Probability of Use*
    - Today's goal: take-home knowledge
    - Tomorrow's goal: dynamic, interactive graphics for categorical data
### VCD Macros & SAS/IML programs

- Macros, datasets available at [datavis.ca/vcd/](http://datavis.ca/vcd/)

#### Discrete distributions

- **DISTPLOT** - Plots for discrete distributions
- **GOODFIT** - Goodness-of-fit for discrete distributions
- **ORDPLOT** - Ord plot for discrete distributions
- **POISPLOT** - Poissonness plot
- **ROOTGRAM** - Hanging rootograms

#### Two-way and \(n\)-way tables

- **AGREEPLOT** - Observer agreement chart
- **CORRESP** - Plot PROC CORRESP results
- **FFOLD** - Fourfold displays for \(2 \times 2 \times k\) tables
- **SIEVEPLOT** - Sieve diagrams
- **MOASAIC** - Mosaic displays
- **MOSMAT** - Mosaic matrices
- **TABLE** - Construct a grouped frequency table, with recoding
- **TRIPLOT** - Trilinear plots for \(n \times 3\) tables

---

### R software and the vcd package

- R software and the vcd package, available at [www.r-project.org](http://www.r-project.org)

#### Discrete distributions

- **distplot** - Plots for discrete distributions
- **goodfit** - Goodness-of-fit for discrete distributions
- **ordplot** - Ord plot for discrete distributions
- **poisplot** - Poissonness plot
- **rootgram** - Hanging rootograms

#### Two-way and \(n\)-way tables

- **agreementplot** - Observer agreement chart
- **fourfold** - Fourfold displays for \(2 \times 2 \times k\) tables
- **sieve** - Sieve diagrams
- **mosaic** - Mosaic displays
- **pairs.table** - Matrix of pairwise association displays
- **structable** - Manipulate high-dimensional contingency tables
- **triplot** - Trilinear plots for \(n \times 3\) tables

---

### R software: Other packages

#### Model-based methods

- **glm** - Fitting generalized linear models
- **gnm** - Fitting generalized non-linear models, e.g., RC(1) model
- **loglm** - MASS package: Fitting loglinear models
- **Rcmdr** - Menu-driven package for statistical analysis and graphics
- **effects** - Effects plots for generalized linear models

#### vcdExtra package

- **vcd-tutorial** - Vignette on working with categorical data and the vcd package
- **mosaic.glm** - Mosaic displays for GLMs and GNMs
- **mosaic3d** - 3D mosaic displays
- **glmlist** - Methods for working with lists of models
Discrete distributions

- **Counts of occurrences**: accidents, words in text, blood cells with some characteristic.
- **Data**: Basic outcome value, \( k \), \( k = 0, 1, \ldots \), and number of observations, \( n_k \), with that value.
- **Example**: distributions of key “marker” words: from, may, whilst, ... in *Federalist Papers* by James Madison, e.g., blocks of 200 words with *may*:

<table>
<thead>
<tr>
<th>Occurrences (k)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blocks (n_k)</td>
<td>156</td>
<td>63</td>
<td>29</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- **Example**: Saxony families with 12 children having \( k = 0, 1, \ldots 12 \) sons.

<table>
<thead>
<tr>
<th>( k )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_k )</td>
<td>3</td>
<td>24</td>
<td>104</td>
<td>286</td>
<td>670</td>
<td>1033</td>
<td>1112</td>
<td>829</td>
<td>478</td>
<td>181</td>
<td>45</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

Questions:
- What process gave rise to the distribution?
- Form of distribution: uniform, binomial, Poisson, negative binomial, geometric, etc.?
- Estimate parameters
- Visualize goodness of fit

For example:
- *Federalist Papers*: might expect a Poisson(\( \lambda \)) distribution.
- *Families in Saxony*: might expect a Bin(\( n, p \)) distribution with \( n = 12 \).
Perhaps \( p = 0.5 \) as well.

Lack of fit:
- Lack of fit tells us something about the process giving rise to the data
- Poisson: assumes constant small probability of the basic event
- Binomial: assumes constant probability and independent trials

Motivation:
- Models for more complex categorical data often use these basic discrete distributions
- Binomial (with predictors) \( \rightarrow \) logistic regression
- Poisson (with predictors) \( \rightarrow \) poisson regression, loglinear models
- \( \Rightarrow \) many of these are special cases of *generalized linear models*

VCD
- methods to fit, visualize, and diagnose discrete distributions:
  - **Fitting**: GOODFIT macro fits uniform, binomial, Poisson, negative binomial, geometric, logarithmic series distributions (or any specified multinomial)
  - **Hanging rootograms**: Sensitively assess departure between Observed, Fitted counts (ROOTGRAM macro)
  - **Ord plots**: Diagnose form of a discrete distribution (ORDPLOT macro)
  - **Poissonness plots**: Robust fitting and diagnostic plots for Poisson (POISPLOT macro)
  - **Robust distribution plots** (DISTPLOT macro)

Using SAS
- Fitting and graphing discrete distributions

26 / 73

27 / 73

28 / 73
SAS macros are high-level, general programs consisting of a series of DATA steps and PROC steps.
Keyword arguments substitute your data names, variable names, and options for the named macro parameters.
Use as:
\[ \%\text{macname}(\text{data}=\text{dataset}, \text{var}=\text{variables}, \ldots); \]
Most arguments have default values (e.g., \text{data}=_last_)
All \text{VCD} macros have internal and online documentation, [http://datavis.ca/sasmac/](http://datavis.ca/sasmac/)
Macros can be installed in directories automatically searched by SAS. Put the following options statement in your \text{AUTOEXEC.SAS} file:
\[ \text{options sasautos=('} 'c:\sasuser\macros' ' sasautos'); \]

E.g., the \text{GOODFIT} macro is defined with the following arguments:
\[ %\text{macro goodfit}(\text{data}=\_last_,\text{var}=,\text{freq}=,\text{dist}=,\text{parm}=,\text{sumat}=100000,\text{format}=,\text{out}=fit,\text{outstat}=stats); \]

Typical use:
\[ %\text{goodfit}(\text{data}=\text{madison}, \text{var}=\text{count}, \text{freq}=\text{blocks}, \text{dist}=\text{poisson}); \]

### Fitting discrete distributions

#### Distributions:
- Poisson, \( p(k) = e^{-\lambda} \frac{\lambda^k}{k!} \)
- Binomial, \( p(k) = \binom{n}{k} p^k (1-p)^{n-k} \)
- Negative binomial, \( p(k) = \binom{n+k-1}{k} p^k (1-p)^n \)
- Geometric, \( p(k) = p(1-p)^{k-1} \)
- Logarithmic series, \( p(k) = \theta^k / [-k \log(1-\theta)] \)

#### Estimate parameter(s):
- Poisson, \( \hat{\lambda} = \sum n_k / \sum n_k \) = mean
- Binomial, \( \hat{p} = \sum n_k / (n \sum n_k) \) = mean / n

#### Goodness of fit:
\[
\chi^2 = \sum_{k=1}^{K} \frac{(n_k - N\hat{p}_k)^2}{N\hat{p}_k} \sim \chi^2_{K-1}
\]

where \( \hat{p}_k \) is the estimated probability of each basic count, under the hypothesis that the data follows the chosen distribution.

---

E.g., the \text{GOODFIT} macro is defined with the following arguments:
\[ %\text{macro goodfit}(\text{data}=\_last_,\text{var}=,\text{freq}=,\text{dist}=,\text{parm}=,\text{sumat}=100000,\text{format}=,\text{out}=fit,\text{outstat}=stats); \]

Typical use:
\[ %\text{goodfit}(\text{data}=\text{madison}, \text{var}=\text{count}, \text{freq}=\text{blocks}, \text{dist}=\text{poisson}); \]
Fitting discrete distributions

The GOODFIT macro gives a table of observed and fitted frequencies, Pearson $\chi^2$ residuals (CHI) and likelihood-ratio deviance residuals (DEV).

| Instances of 'may' in Federalist papers |
|----------|----------|----------|----------|----------|
| COUNT | BLOCKS | PHAT | EXP | CHI | DEV |
| 0 | 156 | 0.51867 | 135.891 | 1.72499 | 6.56171 |
| 1 | 63 | 0.34050 | 89.211 | -2.77509 | -6.62056 |
| 2 | 29 | 0.11177 | 29.283 | -0.05231 | -0.75056 |
| 3 | 8 | 0.02446 | 6.408 | 0.62890 | 1.88423 |
| 4 | 4 | 0.00401 | 1.052 | 2.87493 | 3.26912 |
| 5 | 1 | 0.00053 | 0.138 | 2.31948 | 1.98992 |
| 6 | 1 | 0.00006 | 0.015 | 8.01267 | 2.89568 |
| ====== | ====== | ====== | ====== | ====== |
| 262 | 0.99999 | 261.998 |

In addition, it provides the overall goodness-of-fit tests:

| Goodness-of-fit test for data set MADISON |
| Analysis variable: COUNT Number of Occurrences |
| Distribution: POISSON |
| Estimated Parameters: lambda = 0.6565 |
| Pearson chi-square = 88.92304707 |
| Prob > chi-square = 0 |
| Likelihood ratio G2 = 25.243121314 |
| Prob > chi-square = 0.0001250511 |
| Degrees of freedom = 5 |

The poisson model does not fit! Why?

What's wrong with histograms?

- Discrete distributions often graphed as histograms, with a theoretical fitted distribution superimposed.

%goodfit(data=madison, var=count, freq=blocks, dist=poisson);

Hang & root them → Hanging rootograms

Tukey (1972, 1977):
- shift histogram bars to the fitted curve → judge deviations vs. horizontal line.
- plot $\sqrt{freq}$ → smaller frequencies are emphasized.

%goodfit(data=madison, var=count, freq=blocks, dist=poisson, out=fit);
%rootgram(data=fit, var=count, obs=blocks);
Highlight differences → Deviation rootograms

- Emphasize differences between observed and fitted frequencies
- Draw bars to show the gaps (btype=dev)

%goodfit(data=madison, var=count, freq=blocks, dist=poisson, out=fit);
%rootgram(data=fit, var=count, obs=blocks, btype=dev);

Ord plots: Diagnose form of discrete distribution

- How to tell which discrete distributions are likely candidates?
- Ord (1967): for each of Poisson, Binomial, Negative Binomial, and Logarithmic Series distributions,
  - plot of $k_p / p_{k-1}$ against $k$ is linear
  - signs of intercept and slope → determine the form, give rough estimates of parameters

<table>
<thead>
<tr>
<th>Slope (b)</th>
<th>Intercept (a)</th>
<th>Distribution</th>
<th>Parameter estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>+</td>
<td>Poisson ($\lambda$)</td>
<td>$\lambda = a$</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>Binomial ($n, p$)</td>
<td>$p = b/(b-1)$</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>Neg. binomial ($n, p$)</td>
<td>$p = 1 - b$</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>Log. series ($\theta$)</td>
<td>$\theta = b$</td>
</tr>
</tbody>
</table>

- Fit line by WLS, using $\sqrt{n_k}$ as weights

Ord plots: Other distributions

- ORDPLOT macro
  %ordplot(data=madison, count=Count, freq=blocks);

- Diagnoses distribution as NegBin
- Estimates $\hat{p} = 0.576$
Robust distribution plots: Poisson

- Ord plots lack robustness
  - one discrepant frequency, \( n_k \) affects points for both \( k \) and \( k + 1 \)
- Robust plots for Poisson distribution (Hoaglin and Tukey, 1985)
  - For Poisson, plot count metameter \( \phi(n_k) = \log_e(k! n_k / N) \) vs. \( k \)
  - Linear relation \( \Rightarrow \) Poisson, slope gives \( \hat{\lambda} \)
  - CI for points, diagnostic (influence) plot

\[
\phi(n_k) \equiv \log\left(\frac{k! n_k}{N}\right) = -\lambda + (\log \lambda) k
\]

\( \Rightarrow \) if the distribution is Poisson, plotting \( \phi(n_k) \) vs. \( k \) should give a line with
  - intercept = \( -\lambda \)
  - slope = \( \log \lambda \)
- Nonlinear relation \( \rightarrow \) distribution is not Poisson
- Hoaglin and Tukey (1985) give details on calculation of confidence intervals and influence measures.

POISPLOT macro: example

```sas
title "Instances of 'may' in Federalist papers";
data madison;
  input count blocks;
  label count='Number of Occurrences'
    blocks='Blocks of Text';
datalines;
0 156
1 63
2 29
3 8
4 4
5 1
6 1;
%poisplot(data=madison,count=count, freq=blocks);
```

POISPLOT macro: output

Curvilinear relation \( \rightarrow \) distribution is not Poisson

Instances of 'may' in Federalist papers

Influence plot for change in \( \lambda \)
Generalized robust distribution plots
Other distributions: Analogous plots, for suitable count metamer, \( \phi(n_k) \) vs. \( k \).
- Linear relation \( \Rightarrow \) correct distribution, slope gives parameter estimates
- CI reflect variability of the individual counts, \( n_k \)
- DISTPLOT macro

\[
\text{%distplot(data=madison, count=count, freq=blocks, dist=negbin);}
\]

\[
slope(b) = -0.992 \\
\text{intercept} = -0.654 \\
n: a / \log(p) = 1.413 \\
p: 1 - e(b) = 0.629
\]

Count metameter

<table>
<thead>
<tr>
<th>-10</th>
<th>-9</th>
<th>-8</th>
<th>-7</th>
<th>-6</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

Number of Occurrences

Using R

In R, discrete distributions are conveniently represented as one-way frequency tables.

```r
library(vcd)
data(Federalist)
Federalist
```

<table>
<thead>
<tr>
<th>n May</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1 2 3 4 5 6</td>
</tr>
<tr>
<td>156 63 29 8 4 1 1</td>
</tr>
</tbody>
</table>

The goodfit() function in vcd fits a variety of discrete distributions:

```r
# fit the poisson model
gf1 <- goodfit(Federalist, type="poisson")
gf1
```

Goodness-of-fit test for poisson distribution

<table>
<thead>
<tr>
<th>X^2 df</th>
<th>P(&gt; X^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio 25.24312 5</td>
<td>0.0001250511</td>
</tr>
</tbody>
</table>

```r
summary(gf1)
```

```
Observerved and fitted values for poisson distribution with parameters estimated by 'ML'

<table>
<thead>
<tr>
<th>count observed</th>
<th>fitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>135.89138870</td>
</tr>
<tr>
<td>1</td>
<td>89.21114067</td>
</tr>
<tr>
<td>2</td>
<td>29.28304617</td>
</tr>
<tr>
<td>3</td>
<td>6.40799484</td>
</tr>
<tr>
<td>4</td>
<td>1.05169381</td>
</tr>
<tr>
<td>5</td>
<td>0.13808499</td>
</tr>
<tr>
<td>6</td>
<td>0.01510854</td>
</tr>
</tbody>
</table>
```

The Poisson distribution

```r
# In a poisson, mean = var; this is 'over-dispersed'
mean(rep(0:6, times=Federalist))
```

```
[1] 0.6564885
```

```r
var(rep(0:6, times=Federalist))
```

```
[1] 1.007985
```

The negative binomial distribution, Nbin(r, p) allows the data to deviate from a true Poisson according to a parameter \( r > 0 \).

```r
## try negative binomial distribution (r, p)
gf2 <- goodfit(Federalist, type = "nbinomial")
gf2
```

Goodness-of-fit test for nbinomial distribution

<table>
<thead>
<tr>
<th>X^2 df</th>
<th>P(&gt; X^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio 1.964028 4</td>
<td>0.7423751</td>
</tr>
</tbody>
</table>

This has an acceptable fit to the Federalist data.
Discrete distributions with R and the vcd package

Compare the fits side-by-side:

```r
> plot(gf2, main="Federalist data: Negative binomial fit")
> plot(gf1, main="Federalist data: Poisson fit")
```

Conclusions:
- Perhaps marker words like ‘may’ do not occur with constant probability in all blocks of text
- Perhaps the blocks of text were written under different circumstances

vcd includes Ord_plot() and distplot() functions. E.g.,

```r
> Ord_plot(Federalist, main = "Instances of ‘may’ in Federalist papers")
```

**Testing Association in Two-Way Tables**

**Typical analysis: Nominal factors**

- Pearson $\chi^2$ (or LR $\chi^2$) — when most expected frequencies $\geq 5$.
  ```r
  proc freq;
  weight count; /* if in frequency form */
  table factor * response / chisq;
  ```

- Exact tests — small tables, small sample sizes (e.g., Fisher’s)
  ```r
  proc freq;
  weight count; /* if in frequency form */
  table factor * response / chisq;
  exact pchi;
  ```

Is there a relation between Hi/Lo cholesterol diet and heart disease?

```sas
fats.sas

title 'Cholesterol diet and heart disease';
data fat;
  input diet $ disease $ count;
  datalines;
  LoChol No 6
  LoChol Yes 2
  HiChol No 4
  HiChol Yes 11;
proc freq data=fat;
  weight count;
  tables diet * disease / chisq nopercent nocol;
  exact pchi;
```
Standard output:

Table of diet by disease

<table>
<thead>
<tr>
<th>diet</th>
<th>disease</th>
<th>Frequency</th>
<th>Row Pct</th>
<th>Yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>HiChol</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td>26.67</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>LoChol</td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>75.00</td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>23</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>13</td>
</tr>
</tbody>
</table>

Statistics for Table of diet by disease

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>1</td>
<td>4.9597</td>
<td>0.0259</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>1</td>
<td>5.0975</td>
<td>0.0240</td>
</tr>
<tr>
<td>Continuity Adj. Chi-Square</td>
<td>1</td>
<td>3.1879</td>
<td>0.0742</td>
</tr>
</tbody>
</table>

WARNING: 50% of the cells have expected counts less than 5. (Asymptotic) Chi-Square may not be a valid test.

- The Pearson and LR $\chi^2$ tests are not valid
- The conservative continuity-adjusted test fails significance

Preview: Visualizing association in $2 \times 2$ tables

- Fourfold display: area $\sim$ frequency
- Color: blue (+), red (−)
- Confidence bands: significance of odds ratio
- Interp: Hi cholesterol $\rightarrow$ Heart disease

```
%fold(data=fat, var=diet disease);
```

Ordinal factors and Stratified analyses

- More powerful CMH tests
  - When either the row (factor) or column (response) levels are ordered, more specific (CMH = Cochran - Mantel - Haentzel) tests which take order into account have greater power to detect ordered relations.
  ```
  proc freq;
  weight count;
  table factor * response / chisq cmh;
  ```

- Control for other background variables
  - Stratified analysis tests the association between a main factor and response within levels of the control variable(s)
  - Can also test for homogeneous association across strata
  ```
  proc freq;
  weight count;
  table strata * factor * response / chisq cmh;
  ```
Example: Arthritis treatment

Data on treatment for rheumatoid arthritis (Koch and Edwards, 1988)

- **Ordinal response**: none, some, or marked improvement
- **Factor**: active treatment vs. placebo
- **Strata**: Sex

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Sex</th>
<th>None</th>
<th>Some</th>
<th>Marked</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>Female</td>
<td>6</td>
<td>5</td>
<td>16</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>Placebo</td>
<td>Female</td>
<td>19</td>
<td>7</td>
<td>6</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>Total</td>
<td>42</td>
<td>14</td>
<td>28</td>
<td>84</td>
<td></td>
</tr>
</tbody>
</table>

Overall analysis, ignoring sex

```sas
title 'Arthritis Treatment: PROC FREQ Analysis';
data arth;
  input sex$ treat$ @;
  do improve = 'None', 'Some', 'Marked';
    input count @;
  output;
  end;
datalines;
Female Active 6 5 16
Female Placebo 19 7 6
Male Active 7 2 5
Male Placebo 10 0 1
*-- Ignoring sex;
proc freq order=data;
  weight count;
  tables treat * improve / cmh chisq nocol nopercent;
run;
```

Notes:
- PROC FREQ orders character variables alphabetically (i.e., ‘Marked’, ‘None’, ‘Some’) by default.
- To treat the IMPROVE variable as ordinal, use ORDER=DATA on the PROC FREQ statement.

CMH tests for ordinal variables

Three types of test:

**Non-zero correlation**
- Use when both row and column variables are ordinal.
- CMH $\chi^2 = (N - 1)r^2$, assigning scores (1, 2, 3, ...)
- Most powerful for linear association

**Row Mean Scores Differ**
- Use when only column variable is ordinal
- Analogous to the Kruskal-Wallis non-parametric test (ANOVA on rank scores)
- Ordinal variable must be listed last in the TABLES statement

**General Association**
- Use when both row and column variables are nominal.
- Similar to overall Pearson $\chi^2$ and Likelihood Ratio $\chi^2$. 
## Sample CMH Profiles

### Only general association:

<table>
<thead>
<tr>
<th>b1</th>
<th>b2</th>
<th>b3</th>
<th>b4</th>
<th>b5</th>
<th>Total</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>0</td>
<td>15</td>
<td>25</td>
<td>15</td>
<td>0</td>
<td>55</td>
</tr>
<tr>
<td>a2</td>
<td>5</td>
<td>20</td>
<td>5</td>
<td>20</td>
<td>5</td>
<td>55</td>
</tr>
<tr>
<td>a3</td>
<td>25</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>20</td>
<td>55</td>
</tr>
</tbody>
</table>

### Linear Association:

<table>
<thead>
<tr>
<th>b1</th>
<th>b2</th>
<th>b3</th>
<th>b4</th>
<th>b5</th>
<th>Total</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>31</td>
</tr>
<tr>
<td>a2</td>
<td>2</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>31</td>
</tr>
<tr>
<td>a3</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td>a4</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>31</td>
</tr>
</tbody>
</table>

### Output:

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Alternative Hypothesis</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nonzero Correlation</td>
<td>1</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>2</td>
<td>Row Mean Scores Differ</td>
<td>2</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td>General Association</td>
<td>8</td>
<td>91.797</td>
<td>0.000</td>
</tr>
</tbody>
</table>

### Visualizing Association: Sieve diagrams

#### General Association

- a1
- a2
- a3

#### Linear Association

- a1
- a2
- a3
- a4

### Stratified analysis

#### Overall analysis
- ignores other variables (like sex), by collapsing over them
- risks losing important interactions (e.g., different associations for M & F)

#### Stratified analysis
- controls for the effects of one or more background variables
- list stratification variable(s) `first` on the TABLES statement
  ```
  proc freq;
  tables age * sex * treat * improve;
  ```

#### Looking forward: Loglinear models
- allow more general hypotheses to be stated and tested
- closer connection between testing and visualization (how are variables associated)
### Stratified analysis

The statements below request a stratified analysis with CMH tests, controlling for sex.

```
... arthfreq.sas ...
```

```sas
proc freq order=data;
weight count;
tables sex * treat * improve / cmh chisq nocol nopercent;
run;
```

→ separate tables (partial tests) for Females and Males

#### STATISTICS FOR TABLE 1 OF TREAT BY IMPROVE CONTROLLING FOR SEX=Female

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>2</td>
<td>11.296</td>
<td>0.004</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>2</td>
<td>11.731</td>
<td>0.003</td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
<td>1</td>
<td>10.935</td>
<td>0.001</td>
</tr>
</tbody>
</table>

- Strong association between TREAT and IMPROVE for females

#### STATISTICS FOR TABLE 2 OF TREAT BY IMPROVE CONTROLLING FOR SEX=Male

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>2</td>
<td>4.907</td>
<td>0.086</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>2</td>
<td>5.855</td>
<td>0.054</td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
<td>1</td>
<td>3.713</td>
<td>0.054</td>
</tr>
</tbody>
</table>

**WARNING:** 67% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

- Weak association between TREAT and IMPROVE for males
- Sample size \(N = 29\) for males is small

### Stratified tests

- Individual *(partial)* tests are followed by a *conditional* test, controlling for strata (SEX)
- These tests do not require large sample size in the individual strata— just a large total sample size.
- They assume, but do not test that the association is the same for all strata.

#### SUMMARY STATISTICS FOR TREAT BY IMPROVE CONTROLLING FOR SEX

<table>
<thead>
<tr>
<th>Cochran-Mantel-Haenszel Statistics (Based on Table Scores)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistic</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

### Homogeneity of association

- Is the association between the primary table variables the same over all strata?
- \(2 \times 2\) tables: \(\rightarrow\) Equal odds ratios across all strata?
  - PROC FREQ: MEASURES option on TABLES statement \(\rightarrow\) Breslow-Day test
    ```sas
    proc freq;
    tables strata * factor * response / measures cmh ;
    ...
    loglin strata | factor | response @2;
    ```
  - Larger tables: Use PROC CATMOD to test for *no three-way association*
    - \(\equiv\) same association for the primary factor & response variables \(\forall\) strata
    - \(\equiv\)loglinear model: [Strata Factor] [Strata Response] [Factor Response]
    ```sas
    proc catmod;
    ...
```
Homogeneity of association: Example

- Arthritis data: homogeneity ↔ no 3-way sex * treatment * outcome association
  - ≡ loglinear model: [SexTreat] [SexOutcome] [TreatOutcome]
  - ≡ loglin sex|treat|improve@2 for PROC CATMOD
- Zero frequencies: PROC CATMOD treats as "structural zeros" by default; recode if necessary.

```sas
data arth;
  if count=0 then count=1E-20; /*-- sampling zeros;*/
  proc catmod order=data;
    model sex * treat * improve = _response_ / ml ;
    loglin sex|treat|improve@2 / title='No 3-way association';
  run;
  loglin sex treat|improve / title='No Sex Associations';
```

Output:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Chi-Square</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEX</td>
<td>1</td>
<td>14.13</td>
<td>0.0002</td>
</tr>
<tr>
<td>TREAT</td>
<td>1</td>
<td>1.32</td>
<td>0.2512</td>
</tr>
<tr>
<td>SEX*TREAT</td>
<td>2</td>
<td>2.93</td>
<td>0.0871</td>
</tr>
<tr>
<td>IMPROVE</td>
<td>2</td>
<td>13.36</td>
<td>0.0013</td>
</tr>
<tr>
<td>SEX*IMPROVE</td>
<td>2</td>
<td>6.51</td>
<td>0.0386</td>
</tr>
<tr>
<td>TREAT*IMPROVE</td>
<td>2</td>
<td>13.36</td>
<td>0.0013</td>
</tr>
<tr>
<td>LIKELIHOOD RATIO</td>
<td>5</td>
<td>1.70</td>
<td>0.4267</td>
</tr>
</tbody>
</table>

- But, associations of SEX*TREAT and SEX*IMPROVE are both small.
- Suggests stronger model of homogeneity, [Sex] [TreatOutcome], tested by loglin sex treat|improve; statement.

Visualizing Association: Mosaic displays

Baseline Model: [Sex Treat] [Improve] G2 (6) = 22.60
Reduced Model: [Sex] [Treat Improve] G2 (5) = 9.81

- Fits reasonably well
- How to interpret?
Summary: Part 1

- **Categorical data**
  - Table form vs. case form
  - Non-parametric methods vs. model-based methods
  - Response models vs. association models

- **Graphical methods for categorical data**
  - Frequency data more naturally displayed as count \( \sim \) area
  - Sieve diagram, fourfold & mosaic display: compare observed vs. expected frequency
  - Graphical principles: Visual comparison, effect-ordering, small multiples

- **Discrete distributions**
  - Fit: GOODFIT; Graph: hanging rootograms to show departures
  - Ord plot: diagnose form of distribution
  - POISPLT, DISTPLOT for robust distribution plots

- **Testing association**
  - Pearson \( \chi^2 \), L.R. \( \chi^2 \) (large samples) vs. Fisher exact test (small samples)
  - CMH tests more powerful for ordinal factors
  - Three-way+ tables: Stratified analysis, homogeneity of association
  - Visualize with Sieve diagram, fourfold & mosaic display