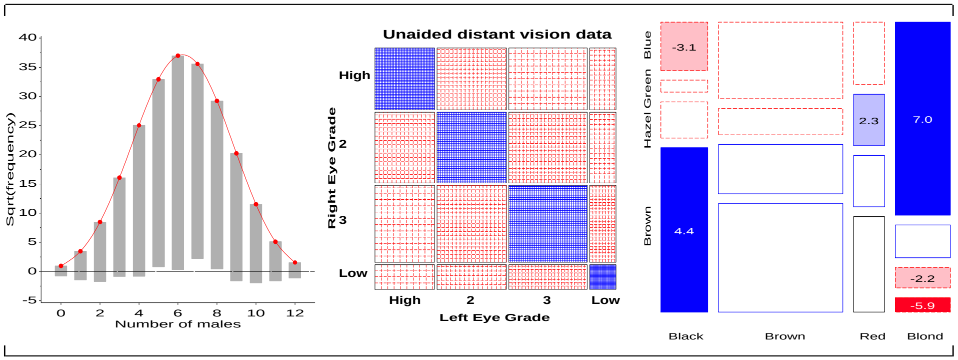
Visualizing Categorical Data with SAS and R

Michael Friendly

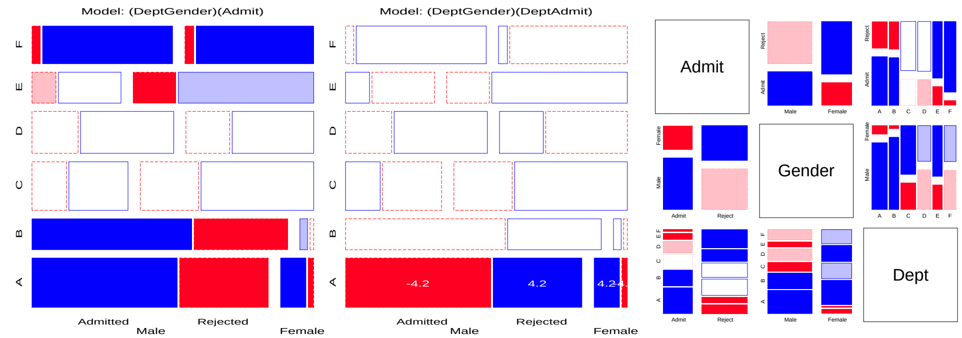
York University

Short Course, 2012

Web notes: `datavis.ca/courses/VCD/`



Part 3: Mosaic displays and loglinear models



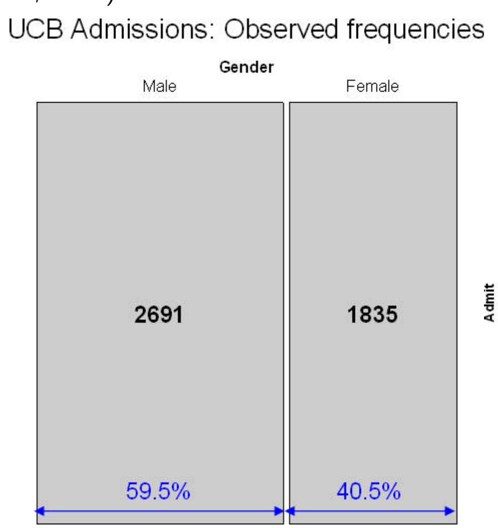
Topics:

- Mosaic displays
- loglinear models for *n*-way tables
- Visualizing loglinear models: SAS & R
- Models for square and structured tables
- Larger tables

Mosaic displays: Basic ideas

Hartigan and Kleiner (1981), Friendly (1994, 1999)

- Area-proportional display of frequencies in an *n*-way table
- Tiles (cells): recursive splits of a unit square—
  - V1: width ∼ marginal frequencies, n_{i++}
  - V2: height ∼ relative frequencies $| V1$, n_{ij+}/n_{i++}
  - V3: width ∼ relative frequencies $| (V1, V2)$, n_{ijk}/n_{ij+}
  - ⋯
- ⇒ area ∼ cell frequency, n_{ijk}

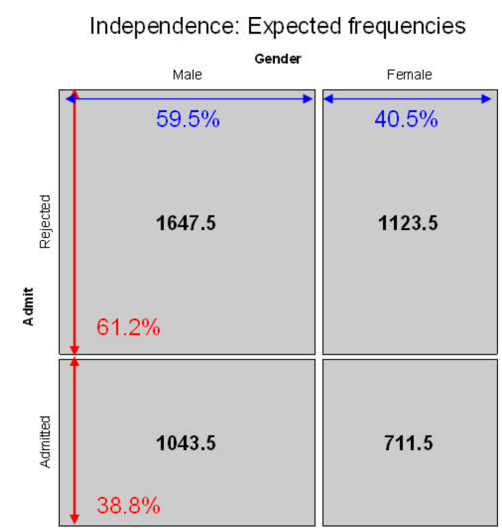


Mosaic displays: Basic ideas

- Independence: Two-way table
- Expected frequencies:

$$\widehat{m}_{ij} = \frac{n_{i+} n_{+j}}{n_{++}} = n_{++} \text{row \%} \text{col \%}$$

- ⇒ rows & columns align when variables are independent

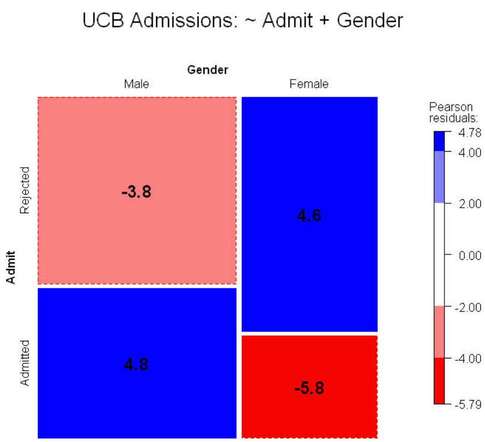


Mosaic displays: Residuals & shading

- Pearson residuals:

$$d_{ij} = \frac{n_{ij} - \widehat{m}_{ij}}{\sqrt{\widehat{m}_{ij}}}$$

- Pearson $\chi^2 = \Sigma\Sigma d_{ij}^2 = \Sigma\Sigma \frac{(n_{ij} - \widehat{m}_{ij})^2}{\widehat{m}_{ij}}$
- Other residuals: deviance (LR), Freeman-Tukey (FT), adjusted (ADJ), ...
- Shading:
  - Sign: $-$ negative in red; $+$ positive in blue
  - Magnitude: intensity of shading: $|d_{ij}| > 0, 2, 4, \ldots$
- $\Rightarrow$ Independence: rows align, or cells are empty!



Loglinear models: Overview

Modeling perspectives

- Loglinear models can be developed as an analog of classical ANOVA and regression models, where *multiplicative* relations (under independence) are re-expressed in *additive* form as models for log(frequency).

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B \equiv [A][B] \equiv \sim A + B$$

- More generally, loglinear models are also generalized linear models (GLMs) for log(frequency), with a Poisson distribution for the cell counts.

$$\log \mathbf{m} = \mathbf{X}\beta$$

- When one table variable is a response, a logit model for that response is equivalent to a loglinear model (discussed in Part 4).

$$\log(m_{1jk}/m_{2jk}) = \alpha + \beta_j^B + \beta_k^C \equiv [AB][AC][BC]$$

Loglinear models: Overview I

- **Two-way tables: Loglinear approach**
  For two discrete variables, A and B, suppose a multinomial sample of total size n over the IJ cells of a two-way $I \times J$ contingency table, with cell frequencies n_{ij}, and cell probabilities $\pi_{ij} = n_{ij}/n$.
  - The table variables are statistically independent when the cell (joint) probability equals the product of the marginal probabilities, $\Pr(A = i \;\&\; B = j) = \Pr(A = i) \times \Pr(B = j)$, or,

$$\pi_{ij} = \pi_{i+}\pi_{+j} \ .$$

  - An equivalent model in terms of expected frequencies, $m_{ij} = n\pi_{ij}$ is

$$m_{ij} = (1/n)\, m_{i+}\, m_{+j} \ .$$

  - This multiplicative model can be expressed in additive form as a model for $\log m_{ij}$,

$$\log m_{ij} = -\log n + \log m_{i+} + \log m_{+j} \ . \qquad (1)$$

Loglinear models: Overview II

- By anology with ANOVA models, the independence model (1) can be expressed as

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B \ , \qquad (2)$$

  where μ is the grand mean of $\log m_{ij}$ and the parameters λ_i^A and λ_j^B express the marginal frequencies of variables A and B, and are typically defined so that $\sum_i \lambda_i^A = \sum_j \lambda_j^B = 0$.

Dependence between the table variables is expressed by adding association parameters, λ_{ij}^{AB}, giving the *saturated model*,

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \equiv [AB] \equiv \sim A * B \ . \qquad (3)$$

- The saturated model fits the table perfectly ($\widehat{m}_{ij} = n_{ij}$): there are as many parameters as cell frequencies. Residual df = 0.
- A global test for association tests $H_0 : \boldsymbol{\lambda}_{ij}^{AB} = \mathbf{0}$.
- For ordinal variables, the λ_{ij}^{AB} may be structured more simply, giving tests for ordinal association.

- **Two-way tables: GLM approach**
  - In the GLM approach, the vector of cell frequencies, $\mathbf{n} = \{n_{ij}\}$ is specified to have a Poisson distribution with means $\mathbf{m} = \{m_{ij}\}$ given by

$$\log \mathbf{m} = \mathbf{X}\boldsymbol{\beta}$$

where $\mathbf{X}$ is a known design (model) matrix and $\boldsymbol{\beta}$ is a column vector containing the unknown λ parameters.
  - For example, for a 2×2 table, the saturated model (3) with the usual zero-sum constraints can be represented as

$$\begin{pmatrix} \log m_{11} \\ \log m_{12} \\ \log m_{21} \\ \log m_{22} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{pmatrix} \mu \\ \lambda_1^A \\ \lambda_1^B \\ \lambda_{11}^{AB} \end{pmatrix}$$

Note that only the linearly independent parameters are represented. $\lambda_2^A = -\lambda_1^A$, because $\lambda_1^A + \lambda_2^A = 0$, and so forth.
  - Advantages of the GLM formulation: easier to express models with ordinal or quantitative variables, special terms, etc. Can also allow for *over-dispersion*.

Three-way Tables I

- **Saturated model:** For a 3-way table, of size $I \times J \times K$ for variables A, B, C, the saturated loglinear model includes associations between all pairs of variables, as well as a 3-way association term, λ_{ijk}^{ABC}

$$\begin{aligned} \log\ m_{ijk} = \mu &+ \lambda_i^A + \lambda_j^B + \lambda_k^C \\ &+ \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC} \ . \end{aligned} \tag{4}$$

  - One-way terms ($\lambda_i^A, \lambda_j^B, \lambda_k^C$): differences in the *marginal frequencies* of the table variables.
  - Two-way terms ($\lambda_{ij}^{AB}, \lambda_{ik}^{AC}, \lambda_{jk}^{BC}$) pertain to the *partial association* for each pair of variables, *controlling* for the remaining variable.
  - The three-way term, λ_{ijk}^{ABC} allows the partial association between any pair of variables to vary over the categories of the third variable.
  - Such models are usually *hierarchical*: the presence of a high-order term, such as $\lambda_{ijk}^{ABC} \rightarrow$ *all* low-order relatives are automatically included.
  - Thus, a short-hand notation for a loglinear model lists only the high-order terms, i.e., model (4) $\equiv [ABC]$

Three-way Tables II

- **Reduced models:**
The usual goal is to fit the *smallest* model (fewest high-order terms) that is sufficient to explain/describe the observed frequencies.

Table: Log-linear Models for Three-Way Tables

Model	**Model symbol**	**Interpretation**
Mutual independence	$[A][B][C]$	$A \perp B \perp C$
Joint independence	$[AB][C]$	$(A\ B) \perp C$
Conditional independence	$[AC][BC]$	$(A \perp B) \mid C$
All two-way associations	$[AB][AC][BC]$	homogeneous assoc.
Saturated model	$[ABC]$	interaction

Symbolic notation (high-order terms):

$$[AB][C] \equiv \log\ m_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB}$$

$$[AB][AC] \equiv \log\ m_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC}$$

Three-way Tables III

- **Assessing goodness of fit**
  - Goodness of fit of a specified model may be tested by the likelihood ratio G^2,

$$G^2 = 2 \sum_i n_i \log(n_i / \widehat{m}_i) \ , \tag{5}$$

or the Pearson χ^2,

$$\chi^2 = \sum_i \frac{(n_i - \widehat{m}_i)^2}{\widehat{m}_i} \ , \tag{6}$$

with degrees of freedom = # cells - # estimated parameters.
  - E.g., for the model of mutual independence, $[A][B][C]$, df = $IJK - (I-1) - (J-1) - (K-1) = (I-1)(J-1)(K-1)$
  - The terms summed in (5) and (6) are the squared *cell residuals*
  - Other measures of balance goodness of fit against parsimony, e.g., *Akaike's Information Criterion* (smaller is better)

$$AIC = G^2 - 2df \ \text{ or } \ AIC = G^2 + 2\ \#\ \text{parameters}$$

Fitting loglinear models: SAS

SAS

- PROC CATMOD

```
%include catdata(berkeley);
proc catmod order=data data=berkeley;
   format dept dept. admit admit.;
   weight freq;                /* data in freq. form */
   model dept*gender*admit=_response_ ;
   loglin admit|dept|gender @2   / title='Model (AD,AG,DG)'; run;
   loglin admit|dept dept|gender / title='Model (AD,DG)';  run;
```

- PROC GENMOD

```
proc genmod data=berkeley;
   class dept gender admit;
   model freq = dept|gender dept|admit / dist=poisson;
run;
```

- mosaic macro usually fits loglin models internally and displays results
- You can also use PROC GENMOD for a more general model, and display the result with the mosaic macro.

Fitting loglinear models: R

R

- loglm() - data in contingency table form (MASS package)

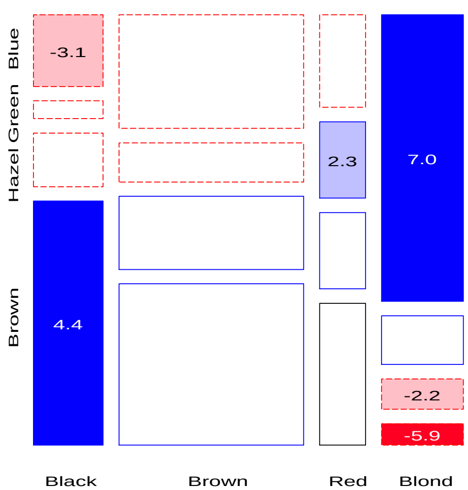
```
data(UCBAdmissions)
  ## conditional independence (AD, DG) in Berkeley data
mod.1 <- loglm(~ (Admit + Gender) * Dept, data=UCBAdmissions)
  ## all two-way model (AD, DG, AG)
mod.2 <- loglm(~ (Admit + Gender + Dept)^2, data=UCBAdmissions)
```

- glm() - data in frequency form

```
berkeley <- as.data.frame(UCBAdmissions)
mod.3 <- glm(Freq ~ (Admit + Gender) * Dept, data=berkeley,
                   family='poisson')
```

- loglm() simpler for nominal variables
- glm() allows a wider class of models
- gnm() fits models for structured association and generalized *non-linear* models
- vcdExtra package provides visualizations for all.

Mosaic displays: Hair color and eye color



We know that hair color and eye color are associated ($\chi^2(9) = 138.29$). The question is how?

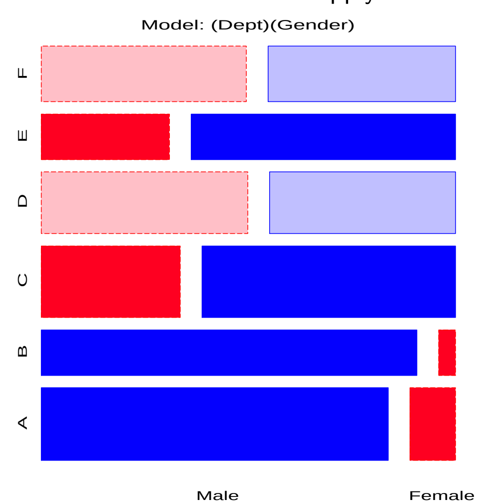
- Dark hair goes with dark eyes, light hair with light eyes
- Red hair, hazel eyes an exception?
- Effect ordering: Rows/cols permuted by CA Dimension 1

$\Rightarrow$ Opposite corner pattern

Mosaic displays: Marginal models

Berkeley data: Departments × Gender (ignoring Admit):

- Did departments differ in the total number of applicants?
- Did men and women apply differentially to departments?

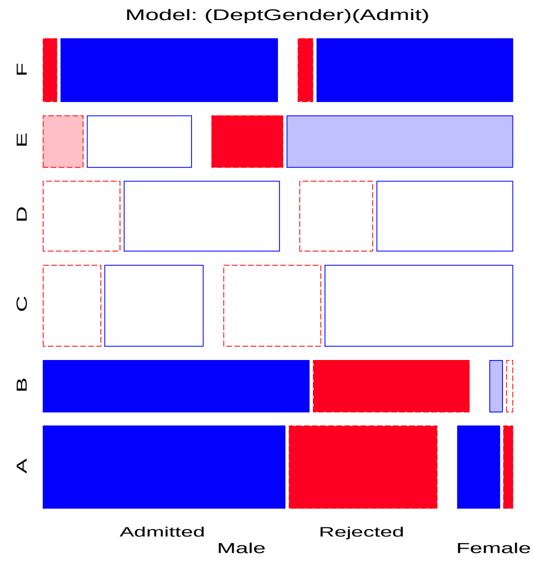


- Model [Dept] [Gender]: $G^2_{(5)} = 1220.6$.
- **Note**: Departments ordered A–F by overall rate of admission.

Mosaic displays for multiway tables

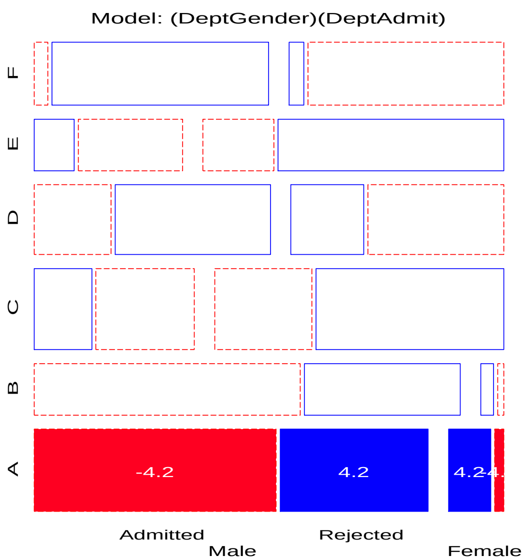
- Generalizes to *n*-way tables: divide cells recursively
- Can fit *any* log-linear model (e.g., 2-way, 3-way, . . .),
  - For a 3-way table: [A][B][C], [AB][C], [AB][AC], . . . , [ABC]
- Each mosaics shows:
  - **DATA** (size of tiles)
  - (some) **marginal** frequencies (spacing → visual grouping)
  - **RESIDUALS** (shading) — what associations have been omitted?
- Visual fitting:
  - Pattern of lack-of-fit (residuals) → “better” model— smaller residuals
  - “cleaning the mosaic” → “better” model— empty cells
  - best done interactively!

- E.g., Joint independence, [DG][A] (null model, Admit as response) [$G^2_{(11)} = 877.1$]:



Mosaic displays for multiway tables

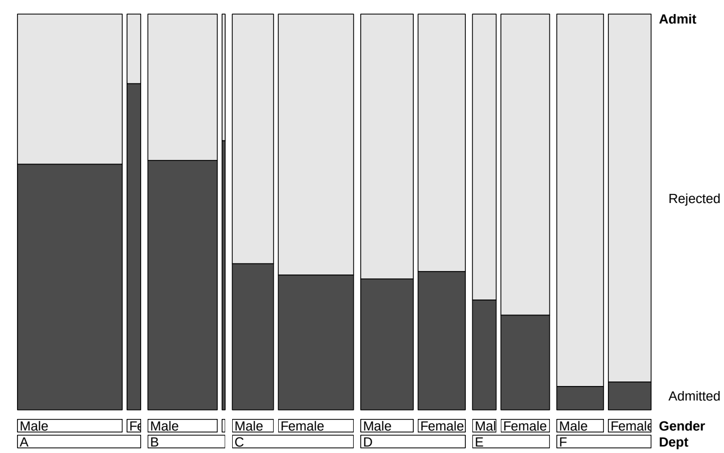
- Visual fitting:



- E.g., Add [Dept Admit] association → Conditional independence:
  - Fits poorly: ($G^2_{(6)} = 21.74$)
  - But, only in Department A!
- The GLM approach allows fitting a special term for Dept. A
- Technical note: These displays use *standardized residuals*: better statistical properties.

Other variations: Double decker plots

- Visualize dependence of one categorical (typically binary) variable on predictors
- Formally: mosaic plots with vertical splits for all predictor dimensions, highlighting the response by shading



Sequential plots and models

- Mosaic for an n-way table → hierarchical decomposition of association in a way analogous to sequential fitting in regression
- Joint cell probabilities are decomposed as

$$p_{ijk\ell\cdots} = \underbrace{\overbrace{p_i \times p_{j|i}}^{\{v_1 v_2\}} \times p_{k|ij}}_{\{v_1 v_2 v_3\}} \times p_{\ell|ijk} \times \cdots \times p_{n|ijk\cdots}$$

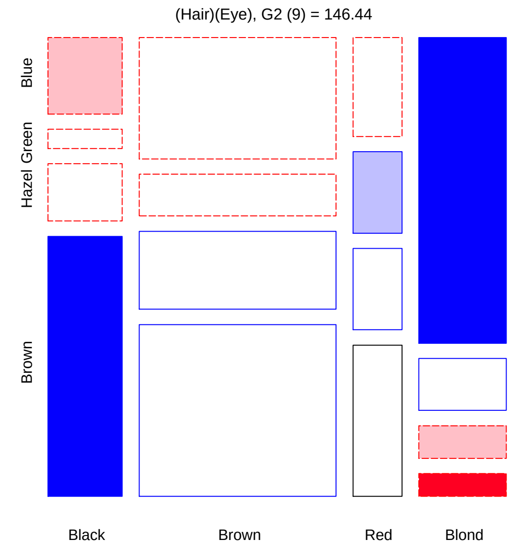
  - First 2 terms → mosaic for v_1 and v_2
  - First 3 terms → mosaic for v_1, v_2 and v_3
  - $\cdots$
- Sequential models of *joint independence* → additive decomposition of the total association, $G^2_{[v_1][v_2]\dots[v_p]}$ (mutual independence),

$$G^2_{[v_1][v_2]\dots[v_p]} = G^2_{[v_1][v_2]} + G^2_{[v_1 v_2][v_3]} + G^2_{[v_1 v_2 v_3][v_4]} + \cdots + G^2_{[v_1 \dots v_{p-1}][v_p]}$$

- As in regression, most useful when there is some substantive ordering of the variables

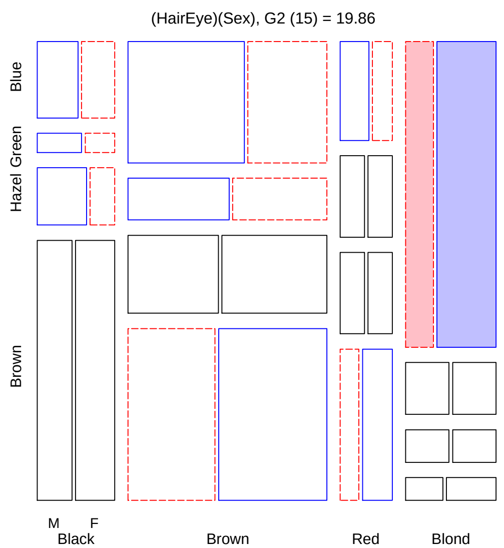
Sequential plots and models: Example

- Hair color x Eye color marginal table (ignoring Sex)



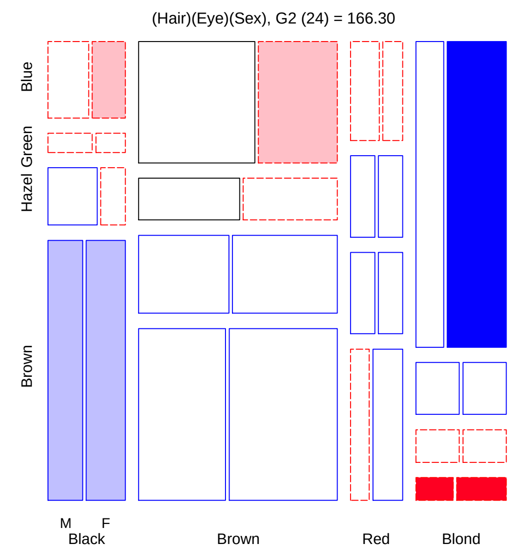
Sequential plots and models: Example

- 3-way table, Joint Independence Model [Hair Eye] [Sex]



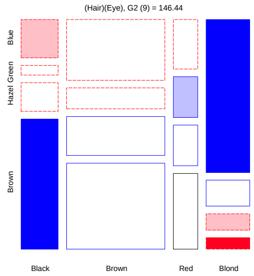
Sequential plots and models: Example

- 3-way table, Mutual Independence Model [Hair] [Eye] [Sex]



Sequential plots and models: Example

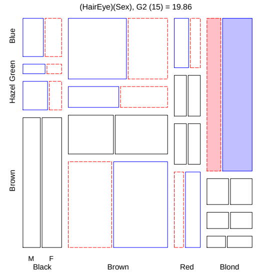
Marginal



[Hair] [Eye]
$G^2_{(9)} = 146.44$

+

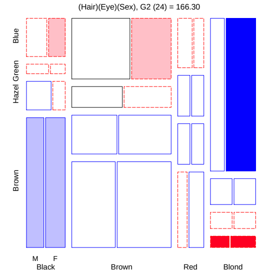
Joint



[Hair Eye] [Sex]
$G^2_{(15)} = 19.86$

=

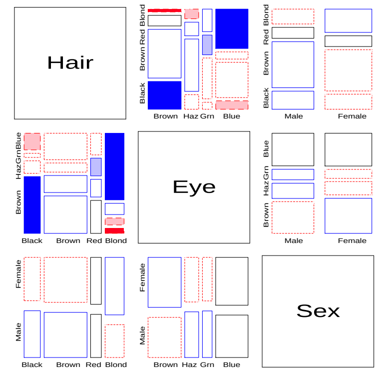
Total



[Hair] [Eye] [Sex]
$G^2_{(24)} = 166.30$

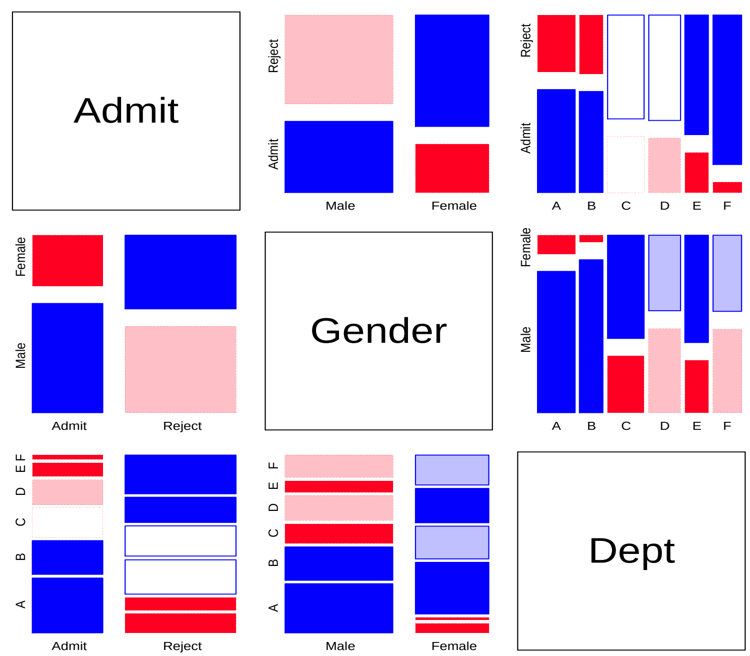
Mosaic matrices

- Analog of *scatterplot matrix* for categorical data (Friendly, 1999)
  - Shows all $p(p-1)$ pairwise views in a coherent display
  - Each pairwise mosaic shows bivariate (marginal) relation
  - Fit: marginal independence
  - Residuals: show marginal associations
  - Direct visualization of the "Burt" matrix analyzed in MCA for p categorical variables



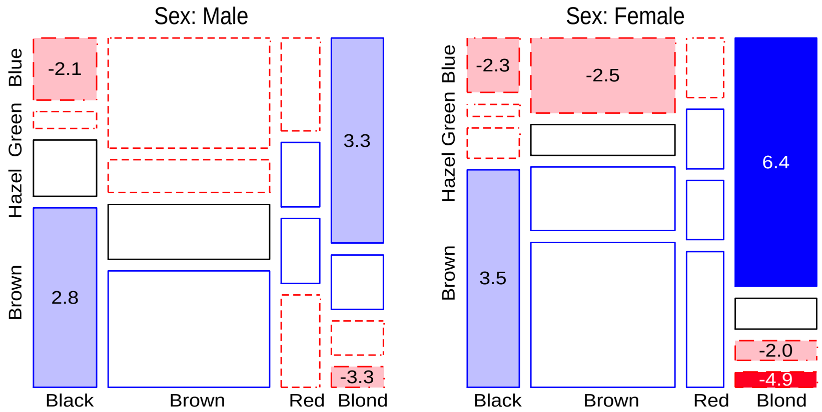
Hair, Eye, Sex data:

Berkeley data:



Partial association, Partial mosaics

- **Stratified analysis:**
  - How does the association between two (or more) variables vary over levels of other variables?
  - Mosaic plots for the main variables show *partial association* at each level of the other variables.
  - E.g., Hair color, Eye color *BY* Sex ↔ `TABLES sex * hair * eye;`



Partial association, Partial mosaics

Stratified analysis: conditional decomposition of G^2

- Fit models of partial (conditional) independence, $A \perp B \mid C_k$ at each level of (controlling for) C.
- $\Rightarrow$ partial G^2s add to the overall G^2 for conditional independence, $A \perp B \mid C$

$$G^2_{A \perp B \mid C} = \sum_k G^2_{A \perp B \mid C(k)}$$

Table: Partial and Overall conditional tests, *Hair* $\perp$ *Eye* | *Sex*

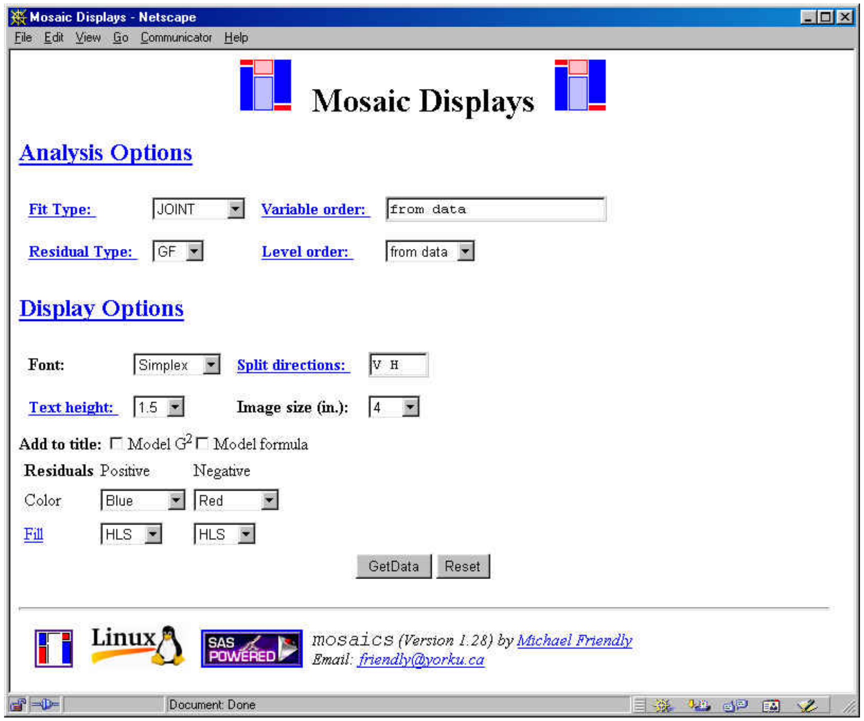
Model	df	G^2	p-value
[*Hair*][*Eye*] | Male	9	44.445	0.000
[*Hair*][*Eye*] | Female	9	112.233	0.000
[*Hair*][*Eye*] | Sex	18	156.668	0.000

Software for Mosaic Displays: Web applet

Demonstration web applet

Go to: `http://datavis.ca/online/mosaics/`

- Runs the *current* version of `mosaics.sas` via a cgi script (perl)
- Can:
  - run *sample* data,
  - *upload* a data file,
  - *enter* data in a form.
- Choose model *fitting* and *display* options (not all supported).
- Provides (limited) interaction with the mosaics via javascript



Software for Mosaic Displays: SAS

SAS software & documentation

http://datavis.ca/mosaics/mosaics.pdf - User Guide
http://datavis.ca/books/vcd/macros.html - Software

- **Examples**: Many in *VCD* and on web site
- **SAS/IML modules**: `mosaics.sas`— Most flexible
  - Enter frequency table directly in SAS/IML, or read from a SAS dataset.
  - Select, collapse, reorder, re-label table levels using SAS/IML statements
  - Specify structural 0s, fit specialized models (e.g., quasi-independence)
  - Interface to models fit using `PROC GENMOD`

Software for Mosaic Displays: SAS

- **Macro interface**: `mosaic` macro, `table` macro, `mosmat` macro
- `mosaic` **macro**— Easiest to use
  - Direct input from a SAS dataset
  - No knowledge of SAS/IML required
  - Reorder table variables; collapse, reorder table levels with `table` macro
  - Convenient interface to *partial mosaics* (`BY=`)
- `table` **macro**
  - Create frequency table from raw data
  - Collapse, reorder table categories
  - Re-code table categories using SAS formats, e.g., `1='Male' 2='Female'`
- `mosmat` **macro**
  - Mosaic matrices— analog of scatterplot matrix (Friendly, 1999)

`mosaic` macro example: Berkeley data

berkeley.sas

```
title 'Berkeley Admissions data';
proc format;
   value admit 1="Admitted" 0="Rejected"          ;
   value dept  1="A" 2="B" 3="C" 4="D" 5="E" 6="F";
        value $sex  'M'='Male'   'F'='Female';
data berkeley;
   do dept = 1 to 6;
      do gender = 'M', 'F';
         do admit = 1, 0;
            input freq @@;
            output;
   end; end; end;
/* -- Male --  - Female- */
/* Admit  Rej  Admit Rej */
datalines;
      512  313    89   19  /* Dept A */
      353  207    17    8  /*      B */
      120  205   202  391  /*      C */
      138  279   131  244  /*      D */
       53  138    94  299  /*      E */
       22  351    24  317  /*      F */
;
```

Data set berkeley:

dept	gender	admit	freq
1	M	1	512
1	M	0	313
1	F	1	89
1	F	0	19
2	M	1	353
2	M	0	207
2	F	1	17
2	F	0	8
3	M	1	120
3	M	0	205
3	F	1	202
3	F	0	391
4	M	1	138
4	M	0	279
4	F	1	131
4	F	0	244
5	M	1	53
5	M	0	138
5	F	1	94
5	F	0	299
6	M	1	22
6	M	0	351
6	F	1	24
6	F	0	317

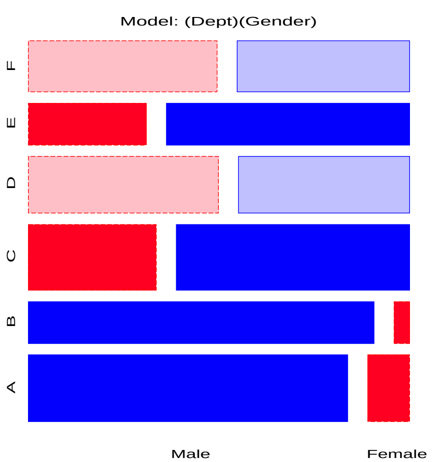
mosaic macro example: Berkeley data

mosaic9m.sas

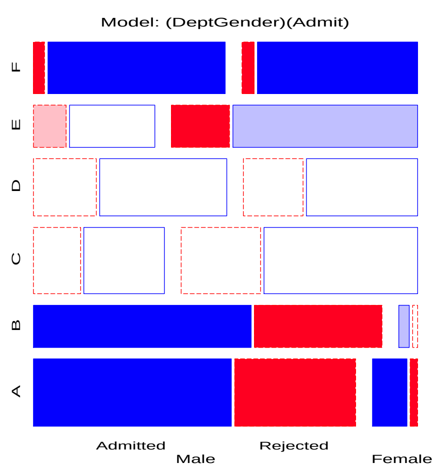
```
goptions hsize=7in vsize=7in;
%include catdata(berkeley);

*-- apply character formats to numeric table variables;
%table(data=berkeley,
    var=Admit Gender Dept,
    weight=freq,
    char=Y, format=admit admit. gender $sex. dept dept.,
    order=data, out=berkeley);

%mosaic(data=berkeley,
    vorder=Dept Gender Admit, /* reorder variables */
    plots=2:3,                /* which plots?      */
    fittype=joint,            /* fit joint indep.  */
    split=H V V, htext=3);    /* options           */
```

NB: The `fittype=` argument allows various types of sequential models: joint, conditional, etc.

mosaic macro example: Berkeley data



Two-way, Dept. by Gender

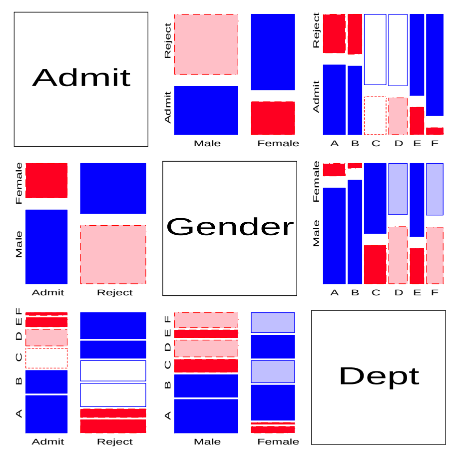


Three-way, Dept. by Gender by Admit

mosmat macro: Mosaic matrices

mosmat9m.sas

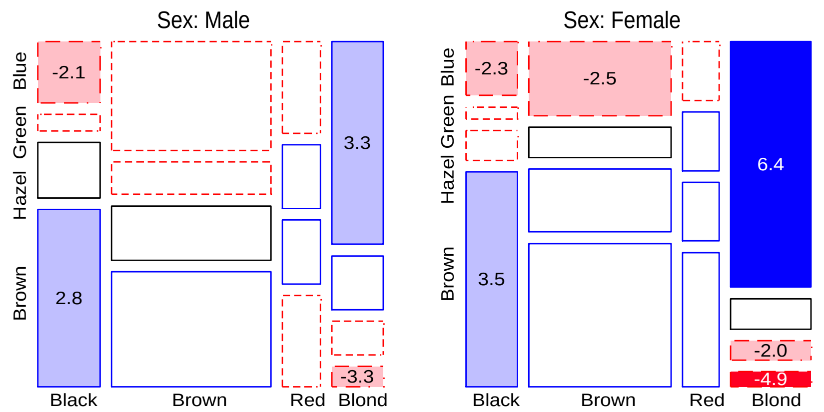
```
%include catdata(berkeley);
%mosmat(data=berkeley,
   vorder=Admit Gender Dept, sort=no);
```



Partial mosaics

mospart3.sas

```
%include catdata(hairdat3s);
%gdispla(OFF);
%mosaic(data=haireye,
    vorder=Hair Eye Sex, by=Sex,
    htext=2, cellfill=dev);
%gdispla(ON);
%panels(rows=1, cols=2);      /* make 2 figs -> 1 */
```



Using the vcd package in R

```
>library(vcd)         # load the vcd package & friends
>
>data(HairEyeColor)
>structable(Eye ~ Hair + Sex, data=HairEyeColor)
```

```
              Eye Brown Blue Hazel Green
Hair  Sex
Black Male           32   11    10     3
      Female         36    9     5     2
Brown Male           53   50    25    15
      Female         66   34    29    14
Red   Male           10   10     7     7
      Female         16    7     7     7
Blond Male            3   30     5     8
      Female          4   64     5     8
```

- The structable() function →'flat' representation of an *n*-way table, similar to mosaic displays
- Formula interface: Col factors ∼ row factors

Using the vcd package in R

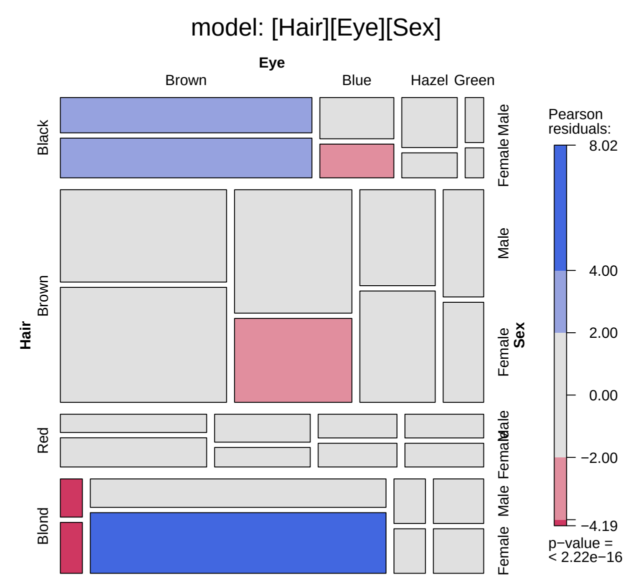
- The loglm() function fits a loglinear model, returns a loglm object
  - Fit the 3-way mutual independence model: Hair + Eye + Sex ≡ [Hair] [Eye] [Sex]
  - Printing the object gives a brief model summary (badness of fit)

```
>## Independence model of hair and eye color and sex.
>mod.1 <- loglm(~Hair+Eye+Sex, data=HairEyeColor)
>mod.1
```

```
Call:
loglm(formula = ~Hair + Eye + Sex, data = HairEyeColor)

Statistics:
                      X^2 df P(> X^2)
Likelihood Ratio 166.3001 24        0
Pearson          164.9247 24        0
```

- The mosaic() function plots the object.
- the vcdExtra package extends mosaic() to glm() models.


```
>mosaic(mod.1, main="model: [Hair][Eye][Sex]")
```



vcd package: Other models

```
>## Joint independence model.
>mod.2 <- loglm(~Hair*Eye+Sex, data=HairEyeColor)
>mod.2
```

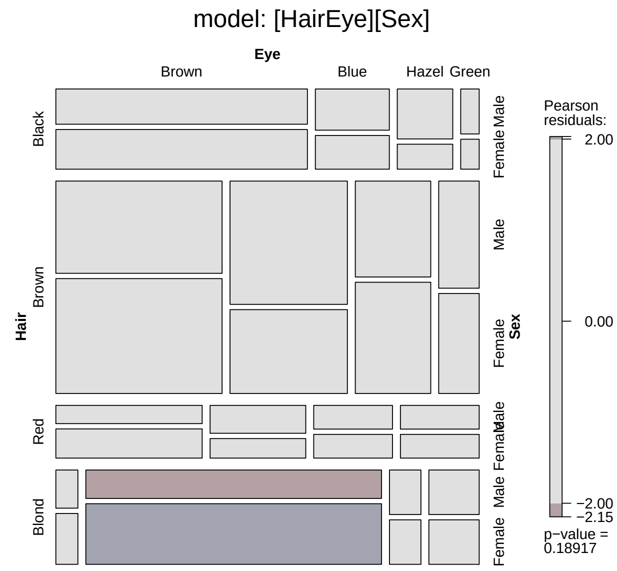
```
Call:
loglm(formula = ~Hair * Eye + Sex, data = HairEyeColor)

Statistics:
                      X^2 df  P(> X^2)
Likelihood Ratio 19.85656 15 0.1775045
Pearson          19.56712 15 0.1891745
```

```
>## Conditional independence model:  Hair*Eye + Sex*Eye
>mod.3 <- loglm(~(Hair+Sex)*Eye, data=HairEyeColor)
>mod.3
```

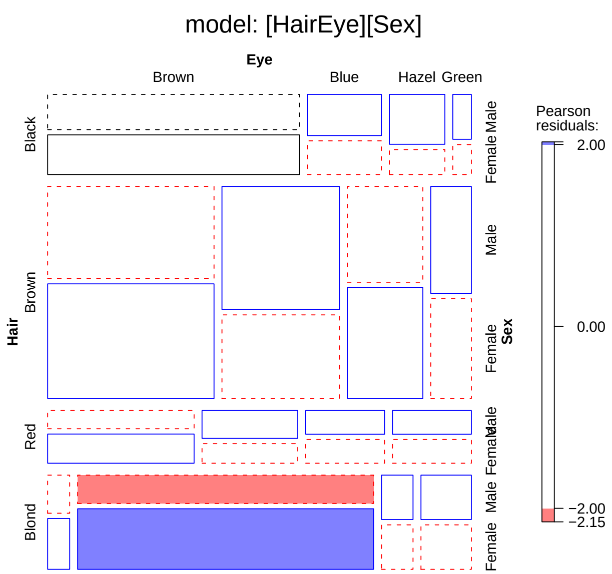
```
Call:
loglm(formula = ~(Hair + Sex) * Eye, data = HairEyeColor)

Statistics:
                      X^2 df  P(> X^2)
Likelihood Ratio 18.32715 12 0.1061122
Pearson          18.04110 12 0.1144483
```



```
>mosaic(mod.2,  main="model: [HairEye][Sex]")
```




```
>mosaic(mod.2,  main="model: [HairEye][Sex]", gp=shading_Friendly)
```



Testing differences between models

- For nested models, $M_1 \subset M_2$ (M_1 nested within, a special case of M_2), the difference in LR G^2, $\Delta = G^2(M_1) - G^2(M_2)$ is a specific test of the difference between them. Here, $\Delta \sim \chi^2$ with $df = df_1 - df_2$.
- R functions are object-oriented: they do different things for different types of objects.

```
>anova(mod.1, mod.2)
```

```
LR tests for hierarchical log-linear models

Model 1:
 ~Hair + Eye + Sex
Model 2:
 ~Hair * Eye + Sex

           Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   166.30014 24
Model 2    19.85656 15  146.44358         9         0.0000
Saturated   0.00000  0   19.85656        15         0.1775
```

More structured tables

Ordered categories

Tables with ordered categories may allow more parsimonious tests of association

- Can represent λ_{ij}^{AB} by a small number of parameters
- $\rightarrow$ more focused and *more powerful* tests of lack of independence (recall: CMH tests)
- Allow one to "explain" the pattern of association in a compact way.

Square tables

For square $I \times I$ tables, where row and column variables have the same categories:

- Can ignore diagonal cells, where association is expected and test remaining association (*quasi-independence*)
- Can test whether association is *symmetric* around the diagonal cells.
- Can test substantively important hypotheses (e.g., mobility tables)

All of these require the GLM approach for model fitting

Ordered categories I

- **Ordinal scores**
  - In many cases it may be reasonable to assign numeric scores, $\{a_i\}$ to an ordinal row variable and/or numeric scores, $\{b_j\}$ to an ordinal column variable.
  - Typically, scores are equally spaced and sum to zero, $\{a_i\} = i - (I+1)/2$, e.g., $\{a_i\} = \{-1, 0, 1\}$ for I=3.
- **Linear-by-Linear (Uniform) Association**: When *both* variables are ordinal, the simplest model posits that any association is *linear* in both variables.

$$\lambda_{ij}^{AB} = \gamma\, a_i b_j$$

  - Only adds one additional parameter to the independence model ($\gamma = 0$).
  - It is similar to CMH test for linear association
  - For integer scores, the local log odds ratios for *any* contiguous 2×2 table are all equal, $\log \theta_{ij} = \gamma$
  - This is a model of *uniform association* — simple interpretation!

Ordered categories II

For a two way table, there are 4 possibilities, depending on which variables are ordinal, and assigned scores:

B$\rightarrow$ / A$\downarrow$	Nominal	Col scores b_j, j=1,…J
Nominal	**General association** df: (I-1)(J-1) parm: λ_{ij}^{AB}	**Row effects** df: I-1 parm: $\alpha_i\, b_j$
Row scores a_i, i=1, … I	**Col effects** df: J-1 parm: $a_i\, \beta_j$	**Uniform association** df: 1 parm: $\gamma\, a_i\, b_j$

Ordered categories III

- **Row Effects and Column Effects**: When only one variable is assigned scores, we have the *row effects model* or the *column effects model*.
  - E.g., in the row effects model, the row variable (A) is treated as nominal, while the column variable (B) is assigned ordered scores $\{b_j\}$.

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \alpha_i b_j$$

    where the row parameters, α_i, are defined so they sum to zero.
  - This model has $(I-1)$ more parameters than the independence model.
  - A Row Effects + Column Effects model allows both variables to be ordered, but not necessarily with linear scores.
- **Fitting models for ordinal variables**
  - Create *numeric* variables for category scores
  - `PROC GENMOD`: Use as quantitative variables in `MODEL` statement, but *not* listed as `CLASS` variables
  - R: Create numeric variables with `as.numeric(factor)`

Ordered categories: RC models

- **RC(1) model**: Generalizes the uniform association, R, C and R+C models by relaxing the assumption of specified order and spacing.

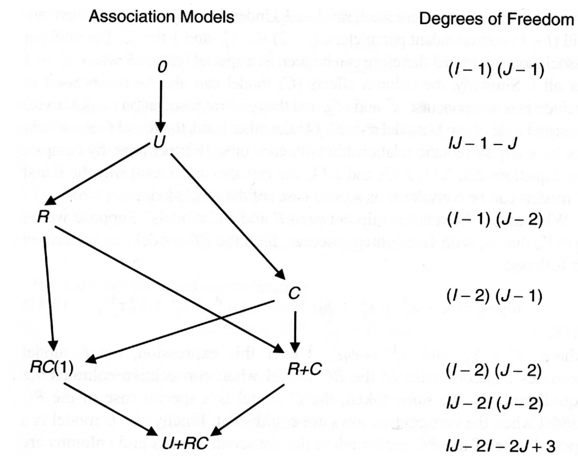
$$RC(1) : \log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \phi \mu_i \nu_j$$

  - The row parameters (μ_i) and column parameters (ν_j) are estimated from the data.
  - ϕ is the measure of association, similar to γ in the uniform association model
- **RC(2) . . . RC(M) models**: Allow two (or more) log-multiplicative association terms; e.g.:

$$RC(2) : \log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \phi_1 \mu_{i1} \nu_{j1} + \phi_2 \mu_{i2} \nu_{j2}$$

  Related to CA, but provide hypothesis tests, std. errors, etc.
- **Fitting RC models**
  - SAS: no implementation
  - R: Fit with `gnm(Freq ~ R + C + Mult(R, C))`

Relations among models



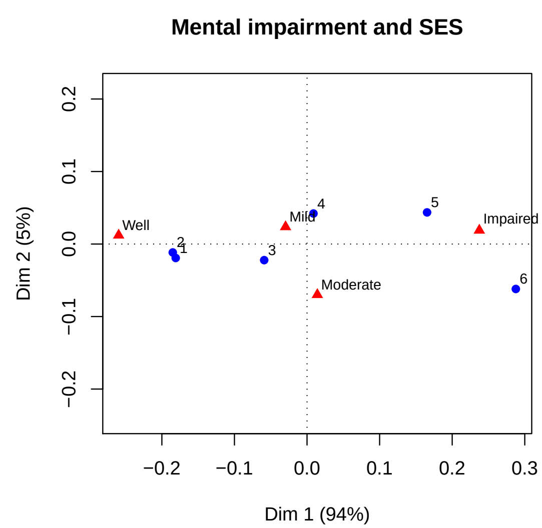
- Structured models: different ways to account for association
- Ordered by: df (# of parameters)
- Arrows show nested models (compare directly: $\Delta\chi^2$)
- All can be compared using AIC (or BIC)

Example: Mental impairment and parents' SES

- Srole et al. (1978) Data on mental health status of ∼1600 young NYC residents in relation to parents' SES.
  - Mental health: Well, mild symptoms, moderate symptoms, Impaired
  - SES: 1 (High) – 6 (Low)

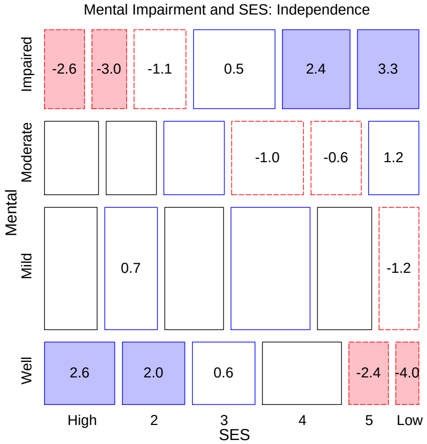
Mental health	Parents' SES					
	High	2	3	4	5	Low
1: Well	64	57	57	72	36	21
2: Mild	94	94	105	141	97	71
3: Moderate	58	54	65	77	54	54
4: Impaired	46	40	60	94	78	71

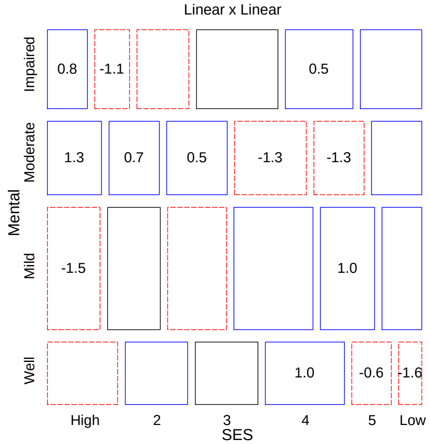
Before fitting models, it is often useful to explore the relation amongs the row/column categories. Correspondence analysis is a good idea!



- Essentially 1D
- Both variables are ordered
- High SES goes with better mental health status
- Can we treat either or both as equally-spaced?
- GLM approach allows testing/comparing hypotheses vs. eye-balling
- Parameter estimates quantify effects.

Visual assessment of various loglin/GLM models: mosaic displays





- Residuals from the independence model show an opposite-corner pattern. This is consistent with both:
  - Linear × linear model: equi-spaced scores for both Mental and SES
  - Row effects model: equi-spaced scores for SES, ordered scores for Mental

Statistical assesment:

Table: Mental health data: Goodness-of-fit statistics for ordinal loglinear models

Model	G^2	df	$\Pr(> G^2)$	AIC	AIC-best
Independence	47.418	15	0.00003	65.418	35.523
Col effects (SES)	6.829	10	0.74145	34.829	4.934
Row effects (mental)	6.281	12	0.90127	30.281	0.386
Lin x Lin	9.895	14	0.76981	29.895	0.000

- Both the Row Effects and Linear × linear models are significantly better than the Independence model
- AIC indicates a slight preference for the Linear × linear model
- In the Linear × linear model, the estimate of the coefficient of $a_i b_j$ is $\hat{\gamma} = 0.0907 = \widehat{\log \theta}$, so $\hat{\theta} = \exp(0.0907) = 1.095$.
- $\mapsto$ each step down the SES scale increases the odds of being classified one step *poorer* in mental health by 9.5%.
- Compare with purely exploratory (CA) interpretation: mental health increases with SES

Fitting these models with PROC GENMOD:

mentgen2.sas

```
%include catdata(mental);
data mental;
   set mental;
   m_lin = mental;     *-- copy m_lin and s_lin for;
   s_lin = ses;        *-- use non-CLASS variables;

title 'Independence model';
proc genmod data=mental;
   class mental ses;
   model count = mental ses / dist=poisson obstats residuals;
   format mental mental. ses ses.;
   ods output obstats=obstats;
%mosaic(data=obstats, vorder=Mental SES, resid=stresdev,
title=Mental Impairment and SES: Independence, split=H V);
```

Row Effects model:

mentgen2.sas

```
proc genmod data=mental;
   class mental ses;
   model count = mental ses mental*s_lin / dist=poisson obstats;
   ...
```

Linear × linear model:

mentgen2.sas

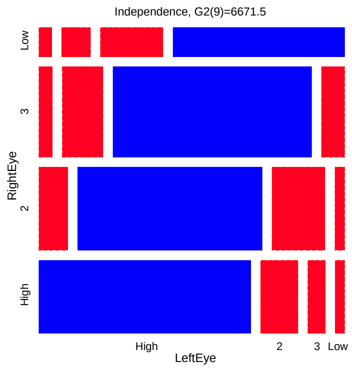
```
proc genmod data=mental;
   class mental ses;
   model count = mental ses m_lin*s_lin / dist=poisson obstats;
```


Fitting these models with glm() in R (see: mental-glm.R for plots)

```
library(vcdExtra)
data(Mental)
# Integer scores for rows/cols
Cscore <- as.numeric(Mental$ses)
Rscore <- as.numeric(Mental$mental)

indep <- glm(Freq ~ mental+ses, family = poisson, data=Mental)

# column effects model (ses)
coleff <- glm(Freq ~ mental + ses + Rscore:ses,
              family = poisson, data = Mental)

# row effects model (mental)
roweff <- glm(Freq ~ mental + ses + mental:Cscore,
              family = poisson, data = Mental)

# linear x linear association
linlin <- glm(Freq ~ mental + ses + Rscore:Cscore,
              family = poisson, data = Mental)

# compare models
AIC(indep, coleff, roweff, linlin)
```

Square tables

- Tables where two (or more) variables have the same category levels:
  - Employment categories of related persons (mobility tables)
  - Multiple measurements over time (panel studies; longitudinal data)
  - Repeated measures on the same individuals under different conditions
  - Related/repeated measures are rarely independent, but may have simpler forms than general association
- E.g., vision data: Left and right eye acuity grade for 7477 women



Square tables: Quasi-Independence

- Related/repeated measures are rarely independent— most observations often fall on diagonal cells.
- Quasi-independence ignores diagonals: tests independence in remaining cells ($\lambda_{ij} = 0$ for $i \neq j$).
- The model dedicates one parameter (δ_i) to each diagonal cell, fitting them exactly,

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \delta_i \, I(i = j)$$

  where $I(\bullet)$ is the indicator function.
- This model may be fit as a GLM by including indicator variables for each diagonal cell: fitted exactly

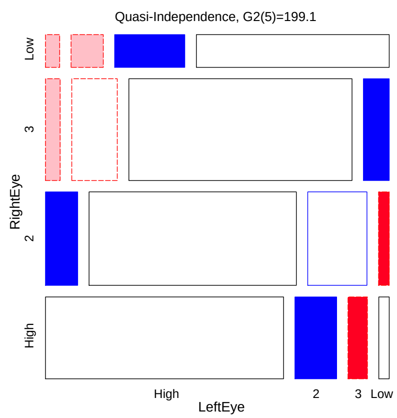
```
diag     4 rows     4 cols

         1     0     0     0
         0     2     0     0
         0     0     3     0
         0     0     0     4
```


- Using PROC GENMOD

··· mosaic10g.sas

```
title 'Quasi-independence model (women)';
proc genmod data=women;
    class RightEye LeftEye diag;
    model Count = LeftEye RightEye diag /
        dist=poisson link=log obstats residuals;
    ods output obstats=obstats;
%mosaic(data=obstats, vorder=RightEye LeftEye, ...);
```

Mosaic:



Square tables: Symmetry

- Tests whether the table is symmetric around the diagonal, i.e., $m_{ij} = m_{ji}$
- As a loglinear model, symmetry is

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \;,$$

  subject to the conditions $\lambda_i^A = \lambda_j^B$ and $\lambda_{ij}^{AB} = \lambda_{ji}^{AB}$.
- This model may be fit as a GLM by including indicator variables with equal values for symmetric cells, and indicators for the diagonal cells (fit exactly)

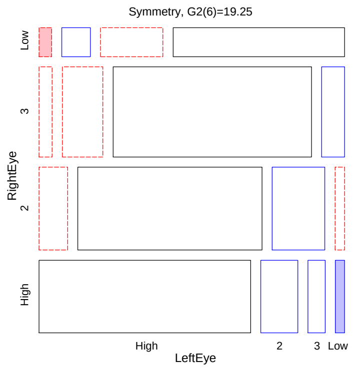
```
symmetry     4 rows     4 cols)

         1    12    13    14
        12     2    23    24
        13    23     3    34
        14    24    34     4
```

- Using PROC GENMOD

··· mosaic10g.sas

```
proc genmod data=women;
class symmetry;
model Count = symmetry /
dist=poisson link=log obstats residuals;
   ods output obstats=obstats;
%mosaic(data=obstats, vorder=RightEye LeftEye, ...);
```

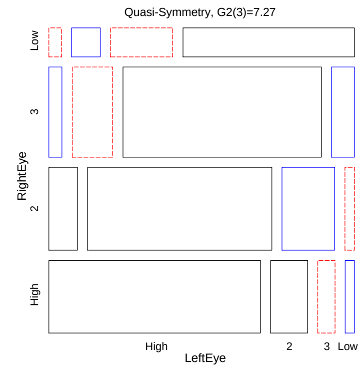
Mosaic:



- **Quasi-Symmetry**
  - Symmetry is often too restrictive: $\mapsto$ equal marginal frequencies ($\lambda_i^A = \lambda_i^B$)
  - PROC GENMOD: Use the usual marginal effect parameters + symmetry:

··· mosaic10g.sas

```
proc genmod data=women;
class LeftEye RightEye symmetry;
model Count = LeftEye RightEye symmetry /
dist=poisson link=log obstats residuals;
   ods output obstats=obstats;
```



Comparing models

Table: Summary of models fit to vision data

Model	G^2	df	Pr(> G^2)	AIC	AIC - min(AIC)
Independence	6671.51	9	0.00000	6685.51	6656.23
Linear*Linear	1818.87	8	0.00000	1834.87	1805.59
Row+Column Effects	1710.30	4	0.00000	1734.30	1705.02
Quasi-Independence	199.11	5	0.00000	221.11	191.83
Symmetry	19.25	6	0.00376	39.25	9.97
Quasi-Symmetry	7.27	3	0.06375	33.27	3.99
Ordinal Quasi-Symmetry	7.28	5	0.20061	29.28	0.00

- Only the quasi-symmetry models provide an acceptable fit: When vision is unequal, association is symmetric!
- The ordinal quasi-symmetry model is most parsimonious
- AIC is your friend for model comparisons

Using the gnm package in R

- Diag() and Symm(): structured associations for square tables
- Topo(): more general structured associations
- mosaic.glm() in vcdExtra

```
library(vcdExtra)
library(gnm)
women <- subset(VisualAcuity, gender=="female", select=-gender)

indep <- glm(Freq ~ right + left, data = women, family=poisson)
mosaic(indep, residuals_type="rstandard", gp=shading_Friendly,
       main="Vision data: Independence (women)"  )

quasi.indep <- glm(Freq ~ right + left + Diag(right, left),
       data = women, family = poisson)

symmetry <- glm(Freq ~ Symm(right, left),
       data = women, family = poisson)

quasi.symm <- glm(Freq ~ right + left + Symm(right, left),
       data = women, family = poisson)

# model comparisons: for *nested* models
anova(indep, quasi.indep, quasi.symm, test="Chisq")
anova(symmetry, quasi.symm, test="Chisq")
```

Survival on the *Titanic*

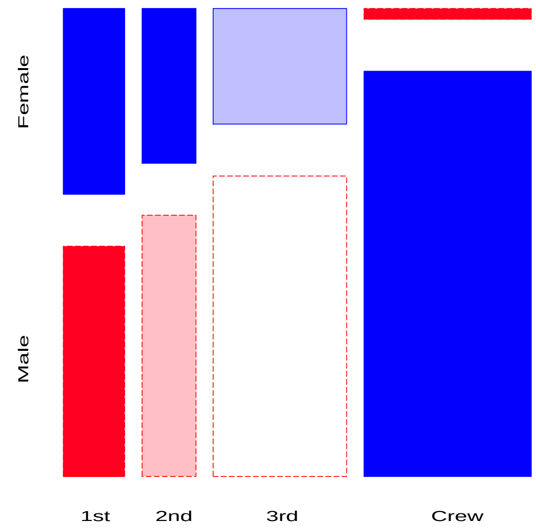
Survival on the *Titanic*: 2201 passengers, classified by Class, Gender, Age, survived. Data from:

- Mersey (1912), *Report on the loss of the "Titanic" S.S.*
- Dawson (1995)

			Class			
Gender	Age	Survived	1st	2nd	3rd	Crew
Male	Adult	Died	118	154	387	670
Female			4	13	89	3
Male	Child		0	0	35	0
Female			0	0	17	0
Male	Adult	Survived	57	14	75	192
Female			140	80	76	20
Male	Child		5	11	13	0
Female			1	13	14	0

Order of variables in mosaics: Class, Gender, Age, Survival

Survival on the *Titanic*: Background variables

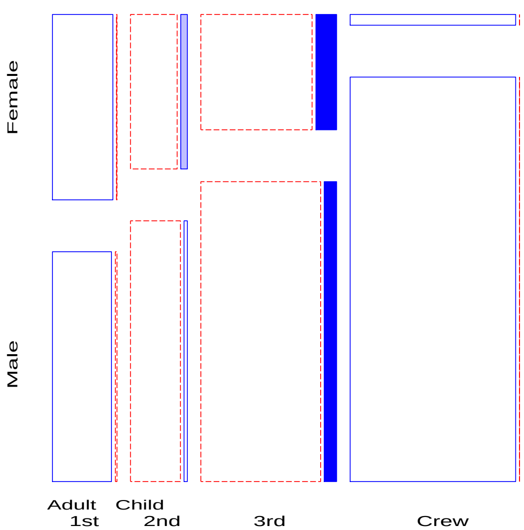


Class × Gender:

- % males decreases with increasing economic class,
- crew almost entirely male

Sequential mosaics: understand associations among background variables

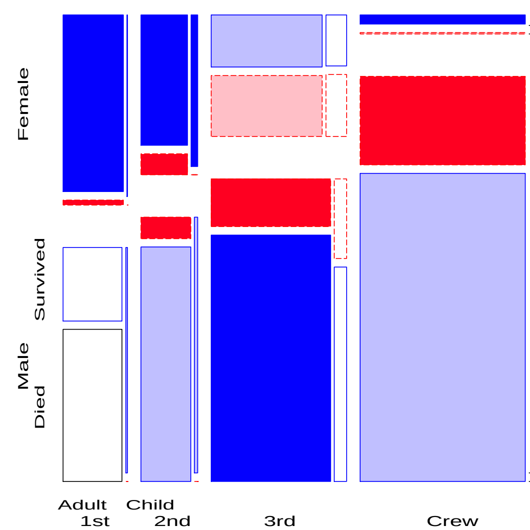
Survival on the *Titanic*: Background variables



3 way: {Class, Gender} ⊥ Age ?

- Overall proportion of children quite small (about 5 %).
- % children smallest in 1st class, largest in 3rd class.
- Residuals: greater number of children in 3rd class (families?)

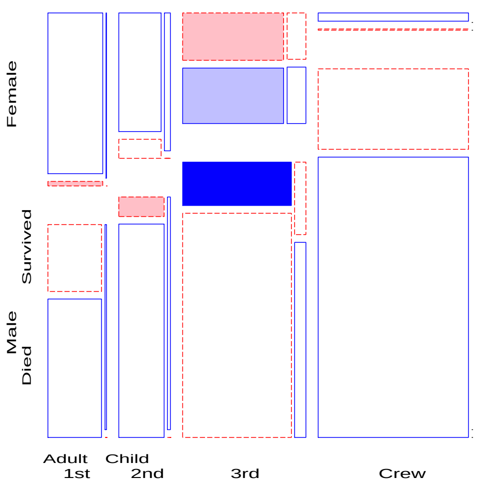
Survival on the *Titanic*: 4 way table



4 way: {Class, Gender, Age} ⊥ Survival ?

- Joint independence: [CGA][S]
- Minimal null model when C, G, A are explanatory
- More women survived, but greater % in 1st & 2nd
- Among men, % survived increases with class.
- Fits poorly [$G^2_{(15)} = 671.96$] ⇒ Add *S*-assoc terms

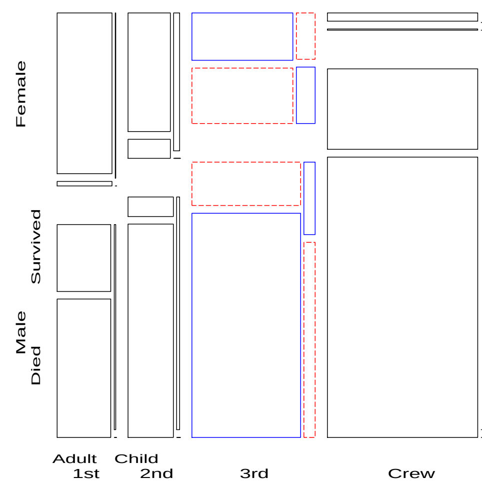
Survival on the *Titanic*: Better models



women and children first $\longrightarrow$

- model [*CGA*][**CS**][**GAS**] (Age and Gender affect survival, independent of Class)
- Model improved slightly, but still not good ($G^2_{(9)} = 94.54$).

Survival on the *Titanic*: Better models



Class interacts with Age & Gender on survival:

- Model [*CGA*][**CGS**][**CAS**]
- $G^2_{(4)}$ now 1.69, a very good fit.
- Perhaps too good? (Overfitting?) $\rightarrow$ check AIC!

Titanic Conclusions

Mosaic displays allow a detailed explanation:

- Regardless of Age and Gender, lower economic status $\longrightarrow$ increased mortality.
- Differences due to Class were moderated by both Age and Gender.
- Women more likely *overall* to survive than men, but:
  - Class × Gender: women in 3rd class *did not* have a significant advantage
  - men in 1st class *did*, compared to men in other classes.
- Class × Age:
  - no children in 1st or 2nd class died, but
  - nearly two-thirds of children in 3rd class died.
  - For adults, mortality ↑ as economic class ↓.
- Summary statement:
  "**women and children (according to class), then 1st class men**".

Summary: Part 3

- **Mosaic displays**
  - Recursive splits of unit square $\rightarrow$ area $\sim$ observed frequency
  - Fit *any* loglinear model $\rightarrow$ shade tiles by residuals
  - $\Rightarrow$ see *departure* of the data from the model
  - SAS: `mosaic` macro, `mosmat` macro; R: `mosaic()`
- **Loglinear models**
  - Loglinear approach: analog of ANOVA for $\log(m_{ijk\cdots})$
  - GLM approach: linear model for $\log(\mathbf{m}) = \mathbf{X}\beta \sim$ Poisson()
  - SAS: `PROC CATMOD`, `PROC GENMOD`; R: `loglm()`, `glm()`
  - Visualize: `mosaic, mosmat` macro; R: `mosaic()`
  - Complex tables: sequential plots, partial plots are useful
- **Structured tables**
  - Ordered factors: models using ordinal scores $\rightarrow$ simpler, more powerful
  - Square tables: Test more specific hypotheses about pattern of association
  - SAS: `PROC GENMOD`; R: `glm()`, `gnm()`