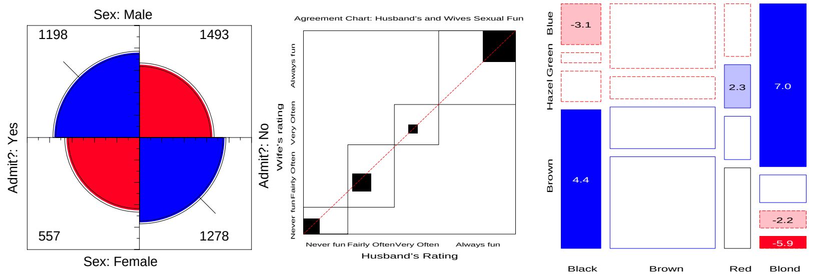
Part 2: Two-way and *n*-way tables



Topics:

- 2×2 tables and fourfold displays
- Sieve diagrams
- Observer agreement
- Mosaic displays and loglinear models for *n*-way tables
- Correspondence analysis

Visualizing Contingency tables

- Two-way tables
  - 2×2 tables — Visualize odds ratio (FFOLD macro)
  - $2 \times 2 \times k$ tables — Homogeneity of association
  - $r \times 3$ tables — Trilinear plots (TRIPLOT macro)
  - $r \times c$ tables — Visualize association (SIEVE program)
  - $r \times c$ tables — Visualize association (MOSAIC macro)
  - Square $r \times r$ tables — Visualize agreement (AGREE program)
- *n*-way tables
  - Fit loglinear models, visualize lack-of-fit — (MOSAIC macro)
  - Test & visualize partial association — (MOSAIC macro)
  - Visualize pairwise association — (MOSMAT macro)
  - Visualize conditional association — (MOSMAT macro)
  - Visualize loglinear structure — (MOSMAT macro)
- Correspondence analysis and MCA — (CORRESP macro)

Methods for 2×2 tables

- Bickel et al. (1975): data on admissions to graduate depatments at U. C. Berkeley in 1973.
- Aggregate data for the six largest departments:

Table 3: Admissions to Berkeley graduate programs

	Admitted	Rejected	Total	% Admitted
Males	1198	1493	2691	44.52
Females	557	1278	1835	30.35
Total	1755	2771	4526	38.78

- Evidence for gender bias?
  - $G^2_{(1)} = 93.7,\ \chi^2_{(1)} = 92.2,\ p < 0.0001$
  - Odds ratio, $\theta = \dfrac{\text{Odds(Admit | Male)}}{\text{Odds(Admit | Female)}} = \dfrac{1198/1493}{557/1276} = 1.84$
  - $\longrightarrow$ Males 84% more likely to be admitted.

Standard analysis: PROC FREQ

```
proc freq data=berkeley;
  weight freq;
  tables gender*admit / chisq;
```

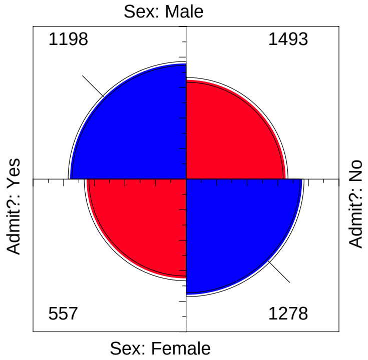
Output:

```
        Statistics for Table of gender by admit

Statistic                    DF       Value      Prob
------------------------------------------------------
Chi-Square                    1     92.2053    <.0001
Likelihood Ratio Chi-Square   1     93.4494    <.0001
Continuity Adj. Chi-Square    1     91.6096    <.0001
Mantel-Haenszel Chi-Square    1     92.1849    <.0001
Phi Coefficient                      0.1427
```

How to visualize and interpret?

Fourfold displays for 2 × 2 tables

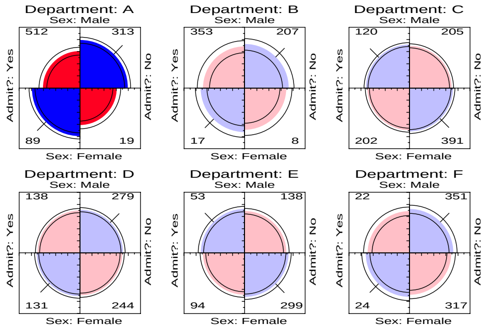
- ***Quarter circles***: radius $\sim \sqrt{n_{ij}} \Rightarrow$ **area** $\sim$ **frequency**
- ***Independence***: Adjoining quadrants $\approx$ align
- ***Odds ratio***: ratio of areas of diagonally opposite cells
- ***Confidence rings***: Visual test of $H_0 : \theta = 1 \leftrightarrow$ adjoining rings overlap



- Confidence rings do not overlap: $\theta \neq 1$

Fourfold displays for 2 × 2 × *k* tables

- Data in Table 3 had been pooled over departments
- Stratified analysis: one fourfold display for each department
- Each 2×2 table standardized to equate marginal frequencies
- Shading: highlight departments for which $H_a : \theta_i \neq 1$



- Only one department (A) shows association; $\theta_A = 0.349 \longrightarrow$ women $(0.349)^{-1} = 2.86$ times as likely as men to be admitted.

What happened here?

Simpson's paradox:

- Aggregate data are misleading because they falsely assume men and women apply *equally* in each field.
- But:
  - Large differences in admission rates across departments.
  - Men and women apply to these departments differentially.
  - Women applied in large numbers to departments with low admission rates.
- (This ignores possibility of structural bias against women: differential funding of fields to which women are more likely to apply.)
- Other graphical methods can show these effects.

The FOURFOLD program and the FFOLD macro

- The FOURFOLD program is written in SAS/IML.
- The FFOLD macro provides a simpler interface.
- Printed output: (a) significance tests for individual odds ratios, (b) tests of homogeneity of association (here, over departments) and (c) conditional association (controlling for department).

Plot by department:

berk4f.sas

```
%include catdata(berkeley);

%ffold(data=berkeley,
   var=Admit Gender,        /* panel variables    */
   by=Dept,                 /* stratify by dept   */
   down=2, across=3,        /* panel arrangement */
   htext=2);                /* font size          */
```

Aggregate data: first sum over departments, using the TABLE macro:

```
%table(data=berkeley, out=berk2,
   var=Admit Gender,        /* omit dept           */
   weight=count,            /* frequency variable */
   order=data);
%ffold(data=berk2, var=Admit Gender);
```

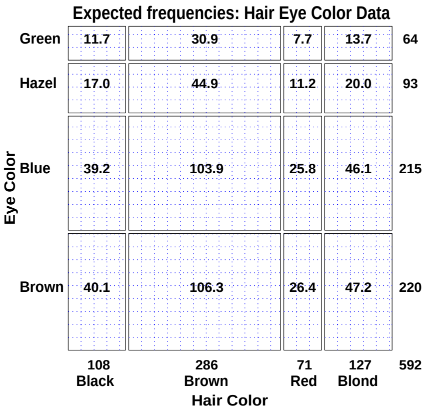
Two-way frequency tables

Table 4: Hair-color eye-color data

Eye	Hair Color				
Color	Black	Brown	Red	Blond	Total
Green	5	29	14	16	64
Hazel	15	54	14	10	93
Blue	20	84	17	94	215
Brown	68	119	26	7	220
Total	108	286	71	127	592

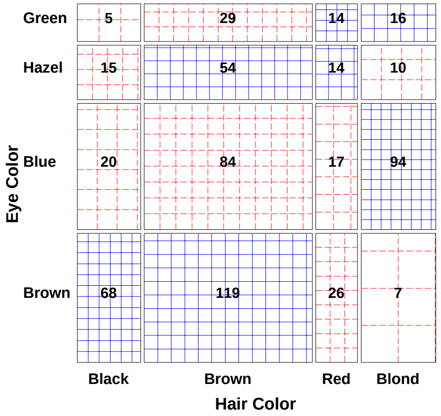
Two-way frequency tables: Sieve diagrams

- **count $\sim$ area**
  - When row/col variables are independent, $n_{ij} \sim n_{i+}n_{+j}$
  - $\Rightarrow$ each cell can be represented as a rectangle, with area = height × width $\sim$ frequency, n_{ij}



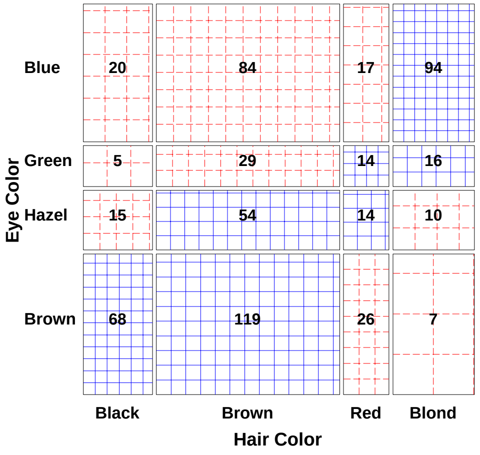
Sieve diagrams

- Height/width $\sim$ marginal frequencies, n_{i+}, n_{+j}
- Area $\sim$ expected frequency, $\sim n_{i+}n_{+j}$
- Shading $\sim$ observed frequency, n_{ij}, color: sign$(n_{ij} - \hat{m}_{ij})$.
- ***Independence***: Shown when density of shading is uniform.



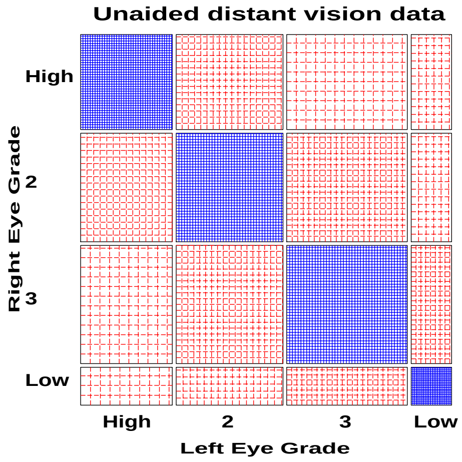
Sieve diagrams

- **Effect ordering**: Reorder rows/cols to make the pattern coherent



Sieve diagrams

- Vision classification data for 7477 women



Sieve diagrams: Example

sieve2.sas

```
proc iml;
  %include iml(sieve);
    *-- frequency table;
  tab = {1520   266   124    66,
          234  1512   432    78,
          117   362  1772   205,
           36    82   179   492 };
    *-- variable and level names;
  vnames = {'Right Eye Grade' 'Left Eye Grade'};
  lnames = { 'High' '2' '3' 'Low',
             'High' '2' '3' 'Low'};
  title  = {'Unaided distant vision data'};
    *-- Global options;
  font='hwpsl011';
  run sieve(tab, vnames, lnames, title );
quit;
```

Online weblet: `http://www.math.yorku.ca/SCS/Online/sieve/`

Observer Agreement

- **Inter-observer agreement** often used as to assess reliability of a subjective classification or assessment procedure
  - → square table, Rater 1 x Rater 2
  - Levels: diagnostic categories (normal, mildly impaired, severely impaired)
- **Agreement vs. Association:** Ratings can be strongly associated without strong agreement
- **Marginal homogeneity:** Different frequencies of category use by raters affects measures of agreement
- **Measures of Agreement:**
  - Intraclass correlation: ANOVA framework— multiple raters!
  - Cohen's κ: compares the observed agreement, $P_o = \sum p_{ii}$, to agreement expected by chance if the two observer's ratings were independent, $P_c = \sum p_{i+}\, p_{+i}$.

$$\kappa = \frac{P_o - P_c}{1 - P_c}$$

- Properties of Cohen's κ:
  - perfect agreement: $\kappa = 1$
  - minimum κ may be < 0; lower bound depends on marginal totals
  - Unweighted κ: counts only diagonal cells (same category assigned by both observers).
  - Weighted κ: allows partial credit for near agreement. (Makes sense only when the categories are *ordered*.)
- Weights: Cicchetti-Alison (inverse integer spacing) vs. Fleiss-Cohen (inverse square spacing)

```
         Integer Weights                      Fleiss-Cohen Weights
  1      2/3      1/3        0          1      8/9      5/9        0
2/3        1      2/3      1/3        8/9        1      8/9      5/9
1/3      2/3        1      2/3        5/9      8/9        1      8/9
  0      1/3      2/3        1          0      5/9      8/9        1
```

Cohen's κ: Example

The table below summarizes responses of 91 married couples to a questionnaire item,

> Sex is fun for me and my partner (a) Never or occasionally, (b) fairly often, (c) very often, (d) almost always.

```
                  --------- Wife's Rating --------
Husband's         Never   Fairly     Very   Almost
Rating              fun    often    Often   always   |    SUM
--------------------------------------------------------+-------
Never fun             7        7        2        3   |     19
Fairly often          2        8        3        7   |     20
Very often            1        5        4        9   |     19
Almost always         2        8        9       14   |     33
--------------------------------------------------------+-------
SUM                  12       28       18       33   |     91
```

Computing κ with SAS

- PROC FREQ: Use AGREE option on TABLES statement
  - Gives both unweighted and weighted κ (default: CA weights)
  - AGREE (wt=FC) uses Fleiss-Cohen weights
  - Bowker's (Bowker, 1948) test of symmetry: $H_0 : p_{ij} = p_{ji}$

kappa3.sas

```
title 'Kappa for Agreement';
data fun;
   do Husband = 1 to 4;
   do Wife    = 1 to 4;
      input count @@;
      output;
      end; end;
datalines;
 7      7     2      3
 2      8     3      7
 1      5     4      9
 2      8     9     14
;
proc freq;
  weight count;
  tables Husband * Wife / noprint agree;          /* default: CA weights*/
  tables Husband * Wife / noprint agree(wt=FC);
```

Computing κ with SAS

Output (CA weights):

```
                  Statistics for Table of Husband by Wife

                          Test of Symmetry
                        ------------------------
                        Statistic (S)     3.8778
                        DF                     6
                        Pr > S            0.6932

                            Kappa Statistics

    Statistic            Value       ASE      95% Confidence Limits
    ------------------------------------------------------------------
    Simple Kappa        0.1293    0.0686       -0.0051        0.2638
    Weighted Kappa      0.2374    0.0783        0.0839        0.3909

                            Sample Size = 91
```

Using Fleiss-Cohen weights:

```
    Weighted Kappa      0.3320    0.0973        0.1413        0.5227
```

Observer agreement: Multiple strata

- When the individuals rated fall into multiple groups, one can test for:
  - Agreement within each group
  - Overall agreement (controlling for group)
  - Homogeneity: Equal agreement across groups

Example: Diagnostic classification of mulitiple sclerosis by two neurologists, for two populations (Landis and Koch, 1977)

```
                 Winnipeg patients        New Orleans patients
  NO rater:
                Cert Prob  Pos Doubt      Cert Prob  Pos Doubt
                --------------------      --------------------
Winnipeg rater:
 Certain MS      38    5    0    1          5    3    0    0
 Probable        33   11    3    0          3   11    4    0
 Possible        10   14    5    6          2   13    3    4
 Doubtful MS      3    7    3   10          1    2    4   14
```

Analysis:

```
proc freq;
  tables strata * rater1 * rater2 / agree;
```

Observer agreement: Multiple strata

msdiag.sas

```
data msdiag;
  do patients='Winnipeg   ', 'New Orleans';
     do N_rating = 1 to 4;
        do W_rating = 1 to 4;
           input count @;
           output;
           end;
        end;
     end;
 label N_rating = 'New Orleans neurologist'
       W_rating = 'Winnipeg neurologist';
datalines;
38  5  0  1
33 11  3  0
10 14  5  6
 3  7  3 10
 5  3  0  0
 3 11  4  0
 2 13  3  4
 1  2  4 14
;

*-- Agreement, separately, and conrolling for Patients;
proc freq data=msdiag;
   weight count;
   tables patients * N_rating * W_rating / norow nocol nopct agree;
```

Observer agreement: Multiple strata

Output, strata 1: (New Orleans patients):

```
           Statistics for Table 1 of N_rating by W_rating
                Controlling for patients=New Orleans

                        Test of Symmetry
                     ------------------------
                     Statistic (S)    9.7647
                     DF                    6
                     Pr > S           0.1349

                         Kappa Statistics

    Statistic          Value       ASE      95% Confidence Limits
    ------------------------------------------------------------
    Simple Kappa      0.2965    0.0785      0.1427        0.4504
    Weighted Kappa    0.4773    0.0730      0.3341        0.6204

                         Sample Size = 69
```

Observer agreement: Multiple strata

Output, strata 2: (Winnipeg patients):

```
           Statistics for Table 2 of N_rating by W_rating
                Controlling for patients=Winnipeg

                        Test of Symmetry
                     ------------------------
                     Statistic (S)    46.7492
                     DF                     6
                     Pr > S            <.0001

                         Kappa Statistics

    Statistic          Value       ASE      95% Confidence Limits
    ------------------------------------------------------------
    Simple Kappa      0.2079    0.0505      0.1091        0.3068
    Weighted Kappa    0.3797    0.0517      0.2785        0.4810

                         Sample Size = 149
```

Observer agreement: Multiple strata

Overall test:

```
           Summary Statistics for N_rating by W_rating
                     Controlling for patients

                    Overall Kappa Coefficients

    Statistic          Value       ASE      95% Confidence Limits
    ------------------------------------------------------------
    Simple Kappa      0.2338    0.0424      0.1506        0.3170
    Weighted Kappa    0.4123    0.0422      0.3296        0.4949
```

Homogeneity test: $H_0 : \kappa_1 = \kappa_2 = \dots$

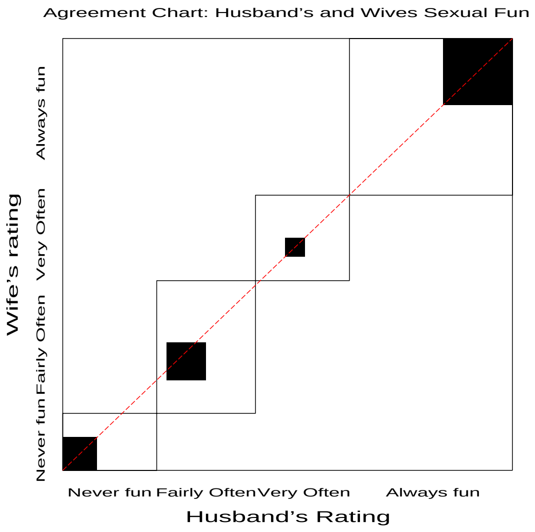
```
                 Tests for Equal Kappa Coefficients

          Statistic          Chi-Square     DF     Pr > ChiSq
          ------------------------------------------------------
          Simple Kappa           0.9009      1         0.3425
          Weighted Kappa         1.1889      1         0.2756

                    Total Sample Size = 218
```

Bangdiwala's Observer Agreement Chart

- The observer agreement chart Bangdiwala (1987) provides
  - a simple graphic representation of the strength of agreement, and
  - a measure of strength of agreement with an intuitive interpretation.
- Construction:
  - $n \times n$ square, n=total sample size
  - Black squares, each of size $n_{ii} \times n_{ii} \rightarrow$ observed agreement
  - Positioned within larger rectangles, each of size $n_{i+} \times n_{+i} \rightarrow$ maximum possible agreement
  - $\Rightarrow$ visual impression of the strength of agreement is

$$B_N = \frac{\text{area of dark squares}}{\text{area of rectangles}} = \frac{\sum_i^k \; n_{ii}^2}{\sum_i^k \; n_{i+} \, n_{+i}}$$

Husbands and wives: $B_N = .146$



Weighted Agreement Chart: Partial agreement

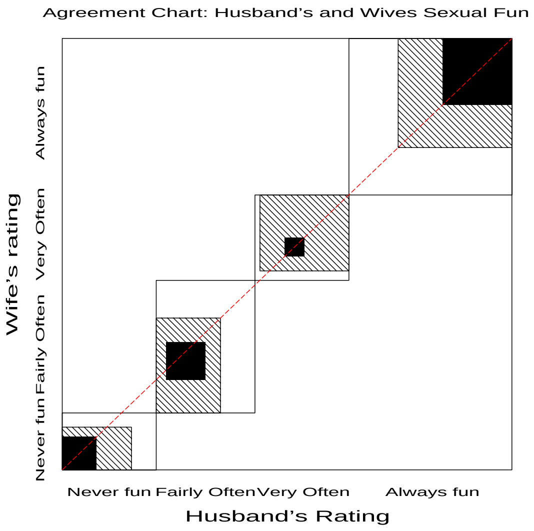
Partial agreement: include weighted contribution from off-diagonal cells, b steps from the main diagonal, using weights $1 > w_1 > w_2 > \cdots$.

$$\begin{array}{ccccc} & & n_{i-b,i} & & \\ & & \vdots & & \\ n_{i,i-b} & \cdots & n_{i,i} & \cdots & n_{i,i+b} \\ & & \vdots & & \\ & & n_{i-b,i} & & \end{array} \qquad \begin{array}{ccccc} & & w_2 & & \\ & & w_1 & & \\ w_2 & w_1 & 1 & w_1 & w_2 \\ & & w_1 & & \\ & & w_2 & & \end{array}$$

- Add shaded rectangles, size $\sim$ sum of frequencies, A_{bi}, within b steps of main diagonal
- $\Rightarrow$ weighted measure of agreement,

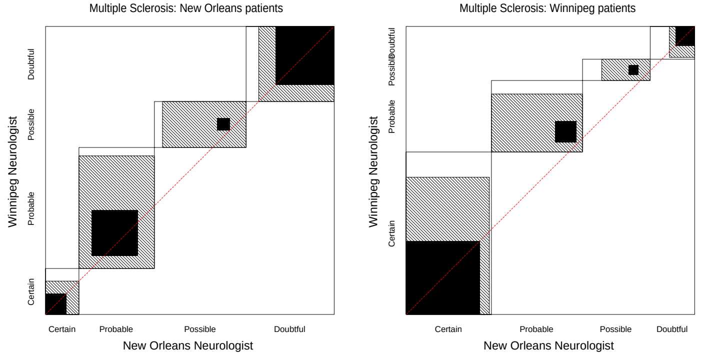
$$B_N^w = \frac{\text{weighted sum of agreement}}{\text{area of rectangles}} = 1 - \frac{\sum_i^k \; [n_{i+} n_{+i} - n_{ii}^2 - \sum_{b=1}^q \; w_b A_{bi}]}{\sum_i^k \; n_{i+} \, n_{+i}}$$

Husbands and wives: $B_N^w = .628$ with $w_1 = 8/9$



Marginal homogeneity and Observer bias

- Different raters may consistently use higher or lower response categories
- Test– ***marginal homogeneity***: $H_0 : n_{i+} = n_{+i}$
- Shows as departures of the squares from the diagonal line



- Winnipeg neurologist tends to use more severe categories

Testing marginal homogeneity

- Test marginal homogeneity using `PROC CATMOD`
  - Two tests available:
    - Equal marginal frequencies: `RESPONSE marginals;` statement
    - Equal mean scores: `RESPONSE means;` statement

agreemar.sas ···

```
title 'Classification of Multiple Sclerosis: Marginal Homogeneity';
proc format;
   value diagnos 1='Certain ' 2='Probable'  3='Possible'  4='Doubtful';

data ms;
 format win_diag no_diag diagnos.;
   do win_diag = 1 to 4;
   do no_diag  = 1 to 4;
      input count @@;
      if count=0 then count=1e-10;  /* avoid structural zeros */
      output;
      end; end;
datalines;
   5     3     0      0
   3    11     4      0
   2    13     3      4
   1     2     4     14
;
```

Testing marginal homogeneity

··· agreemar.sas ···

```
title2 'Testing equal marginal proportions';
proc catmod data=ms;
   weight count;
   response marginals;
   model win_diag * no_diag = _response_ / oneway;
   repeated neuro 2 / _response_= neuro;
```

Output:

```
              Testing equal marginal proportions
                     Analysis of Variance

        Source          DF   Chi-Square    Pr > ChiSq
        --------------------------------------------
        Intercept        3       222.62        <.0001
        Neuro            3        10.54        0.0145

        Residual         0            .             .
```

⇒ marginal proportions differ.

Testing marginal homogeneity

Test of mean scores is more powerful for ordered categories:

··· agreemar.sas

```
title2 'Testing equal means';
proc catmod data=ms;
   weight count;
   response means;
   model win_diag * no_diag = _response_ / oneway;
   repeated neuro 2 / _response_= neuro;
```

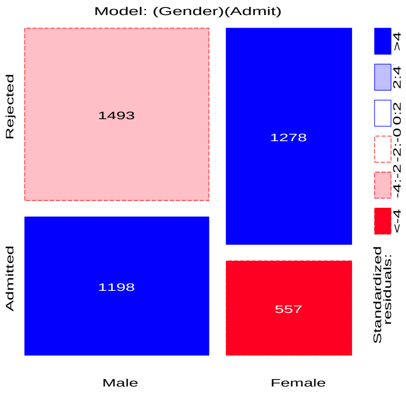
Output:

```
                     Testing equal means
                     Analysis of Variance

        Source          DF   Chi-Square    Pr > ChiSq
        --------------------------------------------
        Intercept        1       570.61        <.0001
        Neuro            1         7.97        0.0048

        Residual         0            .             .
```

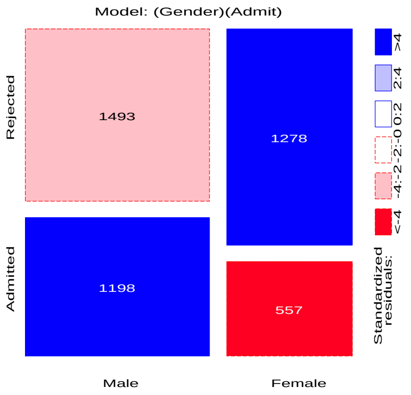
Mosaic displays and Log-linear Models

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

- ***Width*** ∼ one set of marginals, n_{i+}
- ***Height*** ∼ relative proportions of other variable, $p_{j\,|\,i} = n_{ij}/n_{i+}$
- ⇒ **area** ∼ **frequency**, $n_{ij} = n_{i+} p_{j\,|\,i}$



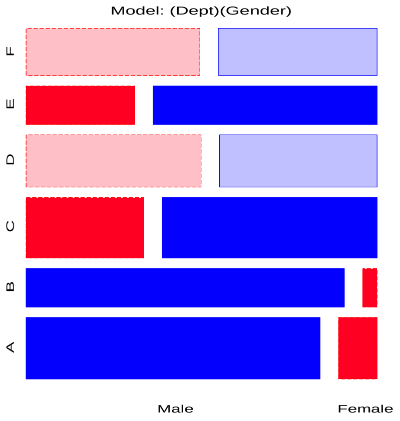
- ***Shading***: Sign and magnitude of Pearson χ^2 residual, $d_{ij} = (n_{ij} - \hat{m}_{ij}) / \sqrt{\hat{m}_{ij}}$ (or L.R. G^2)
  - Sign: − negative in red; + positive in blue
  - Magnitude: intensity of shading: $|d_{ij}| > 0, 2, 4, \ldots$
- ***Independence***: Rows ≈ align, *or* cells are empty!
- E.g., aggregate Berkeley data, independence model:



Mosaic displays

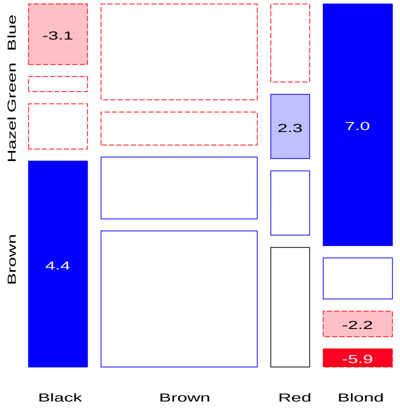
Departments × Gender:

- 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: Hair color and eye color



- 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

Mosaic displays for multiway tables

- Generalizes to n-way tables: divide cells recursively
- Can fit *any* log-linear model (e.g., 3-way),

Table 5: Log-linear Models for Three-Way Tables

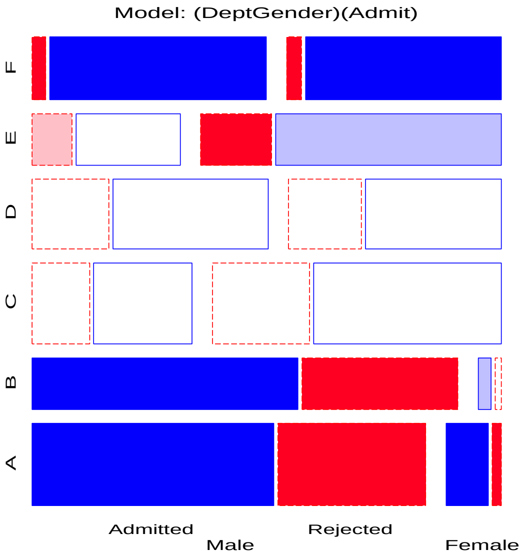
Model	**Model symbol**	**Independence interpretation**
Mutual independence	$[A][B][C]$	$A \perp B \perp C$
Joint independence	$[AB][C]$	$(A\ B) \perp C$
Conditional independence	$[AC][BC]$	$(A \perp B) \mid C$
All two-way associations	$[AB][AC][BC]$	(none)
Saturated model	$[ABC]$	(none)

e.g., the model for conditional independence ($A \perp C \mid B$):

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

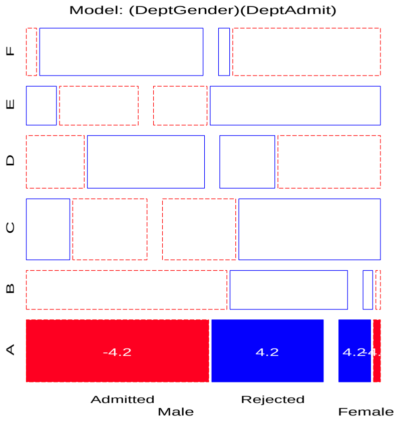
- Each mosaics shows:
  - **DATA** (size of tiles)
  - (some) **marginal** frequencies (spacing → visual grouping)
  - **RESIDUALS** (shading) — what associations have been omitted?

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



Mosaic displays for multiway tables

- Visual fitting:
  - Pattern of lack-of-fit (residuals) → “better” model— smaller residuals
  - “cleaning the mosaic” → “better” model— empty cells
  - best done interactively!



- E.g., Add [Dept Admit] association → Conditional independence:
  - Fits poorly, overall ($G^2_{(6)}$ = 21.74)
  - But, only in Department A!

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

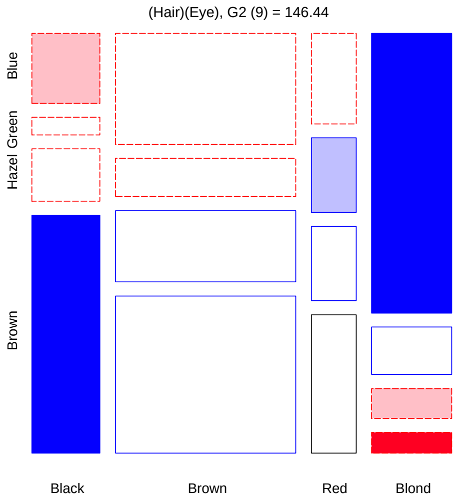
$$p_{ijk\ell\cdots} = \underbrace{\overbrace{p_i \times p_{j|i}}^{\{v_1 v_2\}} \times p_{k|ij}}_{\{v_1 v_2 v_3\}} \times p_{\ell|ijk} \times \cdots \times p_{n|ijk\cdots}$$

  - 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, $G^2_{[v_1][v_2]\ldots[v_p]}$ (mutual independence),

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

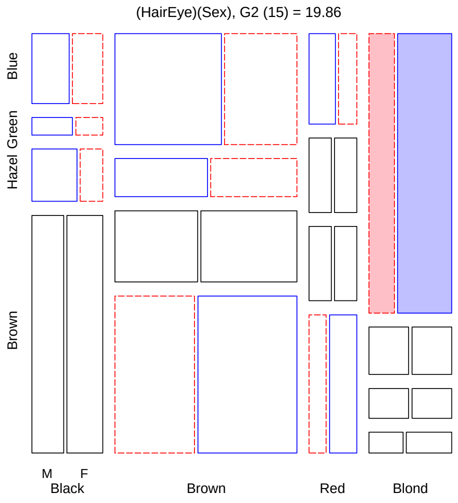
Sequential plots and models: Example

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



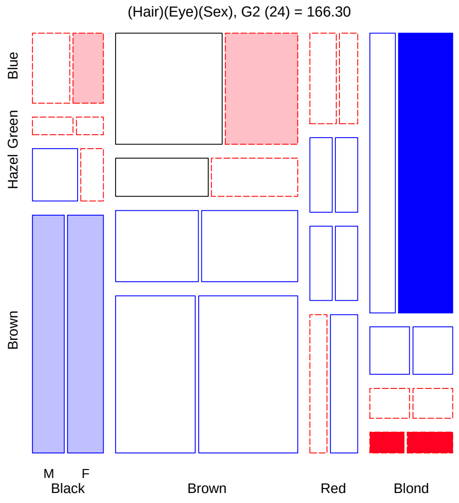
Sequential plots and models: Example

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

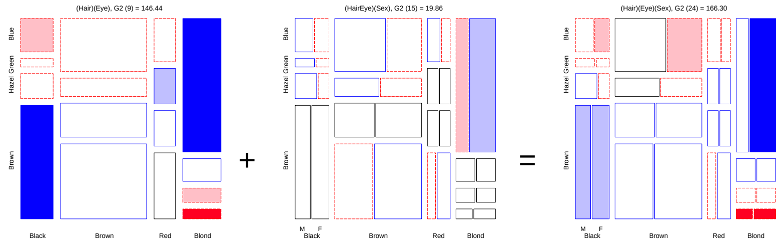


Sequential plots and models: Example

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



Sequential plots and models: Example



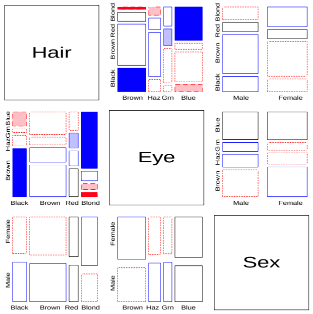
[Hair] [Eye] $G^2_{(9)} = 146.44$

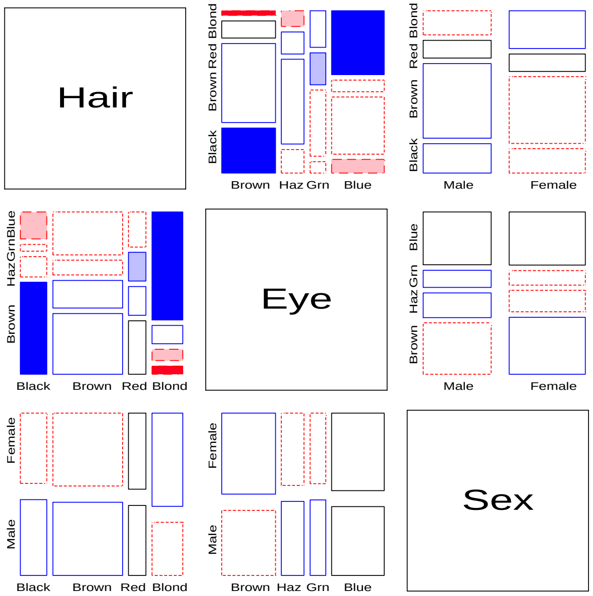
[Hair Eye] [Sex] $G^2_{(15)} = 19.86$

[Hair] [Eye] [Sex] $G^2_{(24)} = 166.60$

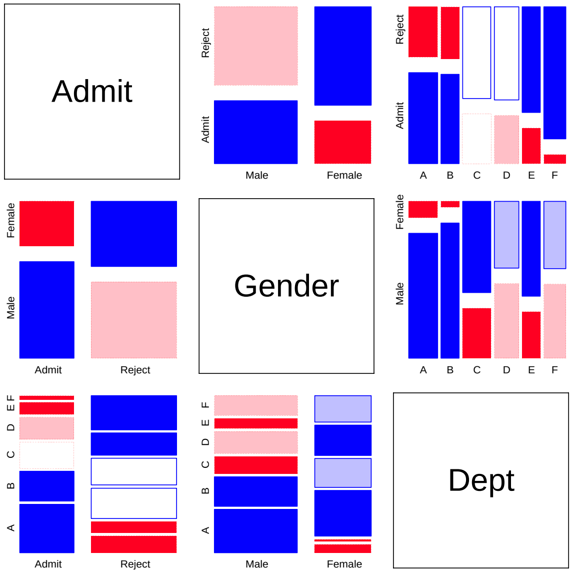
Mosaic matrices

- 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 multiple correspondence analysis for p categorical variables



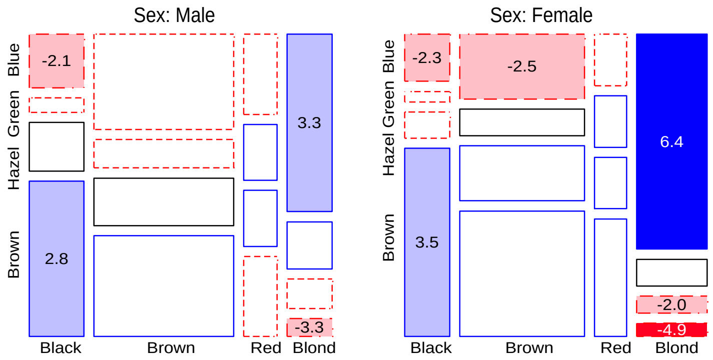


Berkeley data:



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:**
  - For models of partial independence, $A \perp B$ at each level of (controlling for) C $A \perp B \,|\, C_k$, partial G^2s add to the overall G^2 for conditional independence,

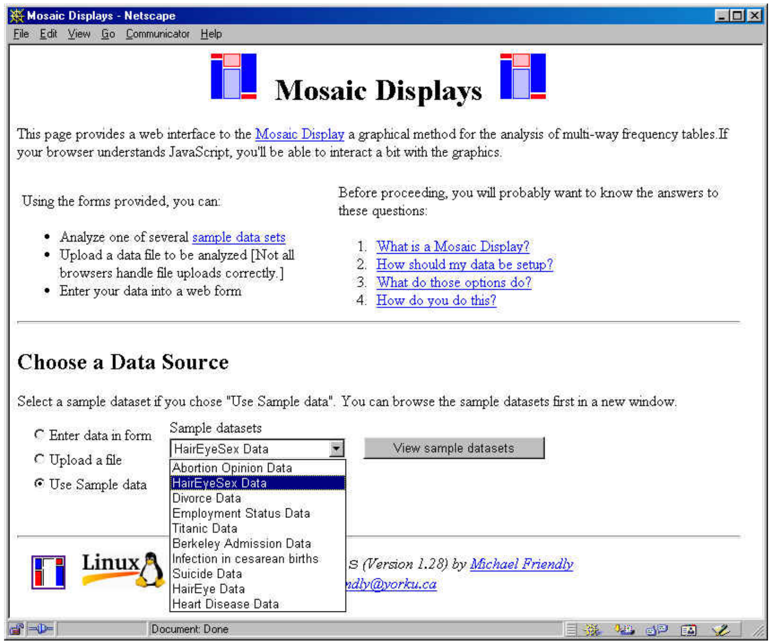
$$G^2_{A \perp B \,|\, C} = \sum_k G^2_{A \perp B \,|\, C(k)}$$

Table 6: 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

- **Demonstration web applet**: `http://www.math.yorku.ca/SCS/Online/mosaics/`
  - Runs the *current* version of mosaics via a cgi-script
  - Can run *sample data*, *upload* a data file, *enter* data in a form.
  - Choose model fitting and display options (not all supported).


Mosaic Displays - Netscape
File Edit View Go Communicator Help
Linux
View sample datasets
Document: Done

Mosaic Displays

This page provides a web interface to the Mosaic Display a graphical method for the analysis of multi-way frequency tables.If your browser understands JavaScript, you'll be able to interact a bit with the graphics.

Using the forms provided, you can:

- Analyze one of several sample data sets
- Upload a data file to be analyzed [Not all browsers handle file uploads correctly.]
- Enter your data into a web form

Before proceeding, you will probably want to know the answers to these questions:

1. What is a Mosaic Display?
2. How should my data be setup?
3. What do those options do?
4. How do you do this?

Choose a Data Source

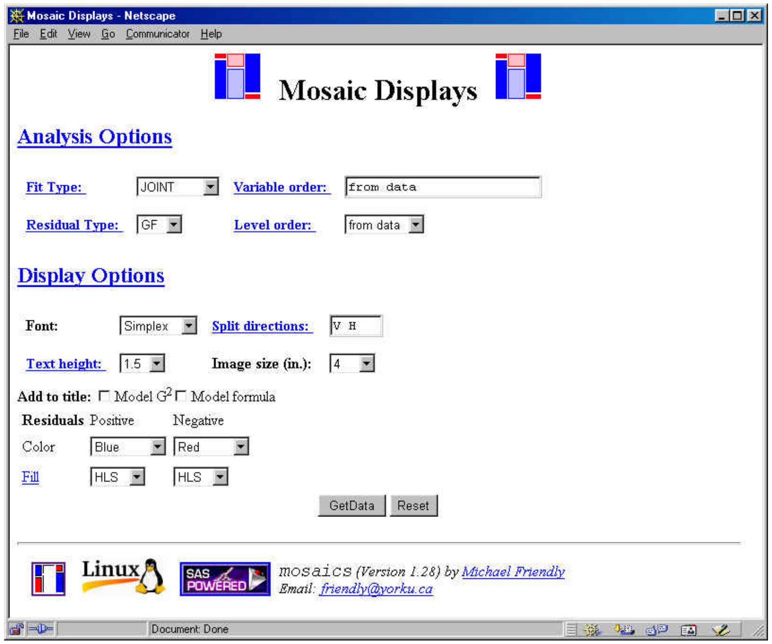
Select a sample dataset if you chose "Use Sample data". You can browse the sample datasets first in a new window.

○ Enter data in form
○ Upload a file
● Use Sample data

Sample datasets: HairEyeSex Data

Abortion Opinion Data
HairEyeSex Data
Divorce Data
Employment Status Data
Titanic Data
Berkeley Admission Data
Infection in cesarean births
Suicide Data
HairEye Data
Heart Disease Data

s (Version 1.28) by Michael Friendly
ndly@yorku.ca


Mosaic Displays - Netscape
File Edit View Go Communicator Help
GetData
Reset
Linux
SAS POWERED
Document: Done

Mosaic Displays

Analysis Options

Fit Type: JOINT Variable order: from data

Residual Type: GF Level order: from data

Display Options

Font: Simplex Split directions: V H

Text height: 1.5 Image size (in.): 4

Add to title: □ Model G^2 □ Model formula

Residuals	Positive	Negative
Color	Blue	Red
Fill	HLS	HLS

mosaics (Version 1.28) by Michael Friendly
Email: friendly@yorku.ca

Software for Mosaic Displays

- **SAS software & documentation**:
  http://www.math.yorku.ca/SCS/mosaics.html
  http://www.math.yorku.ca/SCS/vcd/
- **Examples**: Many in *VCD* and on web site
- **SAS/IML modules**: mosaics.sas SAS/IML program
  - Enter frequency table directly in SAS/IML, or read from a SAS dataset.
  - Most flexible:
    - 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

- **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`:

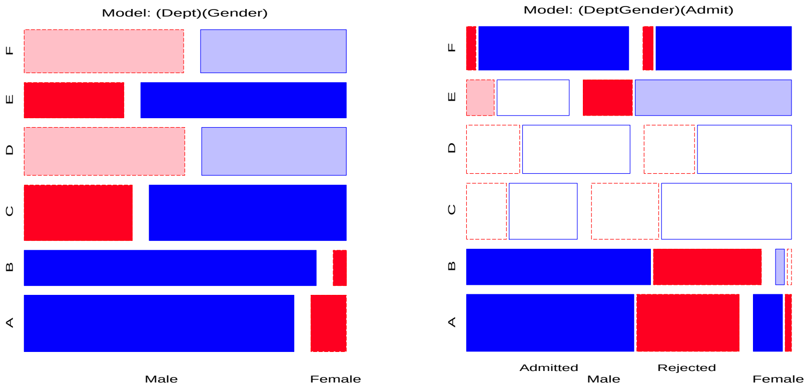
```
          dept     gender     admit     freq

            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           */
```

`mosaic` macro example: Berkeley data



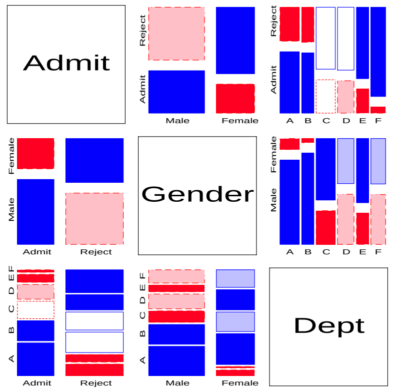
Two-way, Dept. by Gender

Three-way, Dept. by Gender by Admit

`mosmat` macro: Mosaic matrices

mosmat9m.sas

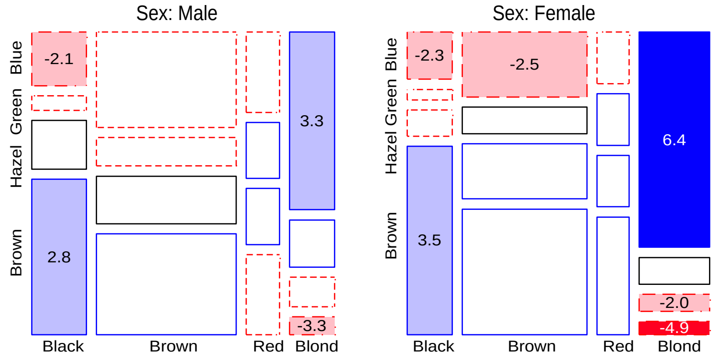
```
%include catdata(berkeley);
%mosmat(data=berkeley,
    vorder=Admit Gender Dept, sort=no);
```



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

```
R># load the vcd library & friends
R>library(vcd)
R>
R>data("HairEyeColor")
R>structable(HairEyeColor)
```

```
                Eye Brown Blue Hazel Green
Hair  Sex
Black Male             32   11    10     3
      Female           36    9     5     2
Brown Male             38   50    25    15
      Female           81   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
```

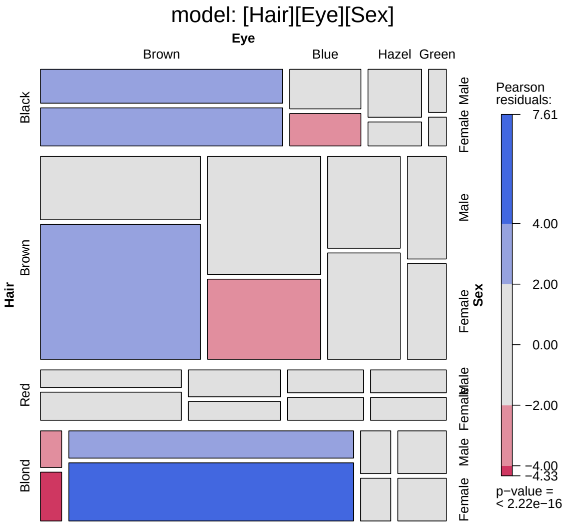
Using the vcd package in R

- The `loglm()` function fits a loglinear model, returns a `loglm` object
- The `mosaic()` function plots the object

```
R>## Independence model of hair and eye color and sex.
R>mod.1 <- loglm(~1+2+3, data=HairEyeColor)
R>mod.1
```

```
Call:
loglm(formula = ~1 + 2 + 3, data = HairEyeColor)

Statistics:
                     X^2 df P(> X^2)
Likelihood Ratio 175.7934 24        0
Pearson          171.8144 24        0
```

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

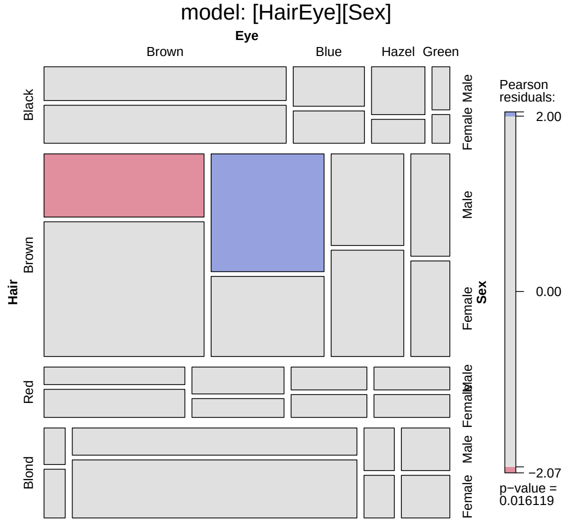


Joint independence

```
R>## Joint independence model.
R>mod.2 <- loglm(~1*2+3, data=HairEyeColor)
R>mod.2
```

```
Call:
loglm(formula = ~1 * 2 + 3, data = HairEyeColor)

Statistics:
                      X^2 df   P(> X^2)
Likelihood Ratio 29.34982 15 0.01449443
Pearson          28.99286 15 0.01611871
```

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



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.

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

```
LR tests for hierarchical log-linear models

Model 1:
 ~`1` + `2` + `3`
Model 2:
 ~`1` * `2` + `3`

           Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   175.79340 24
Model 2    29.34982 15  146.44358         9        0.00000
Saturated   0.00000  0   29.34982        15        0.01449
```

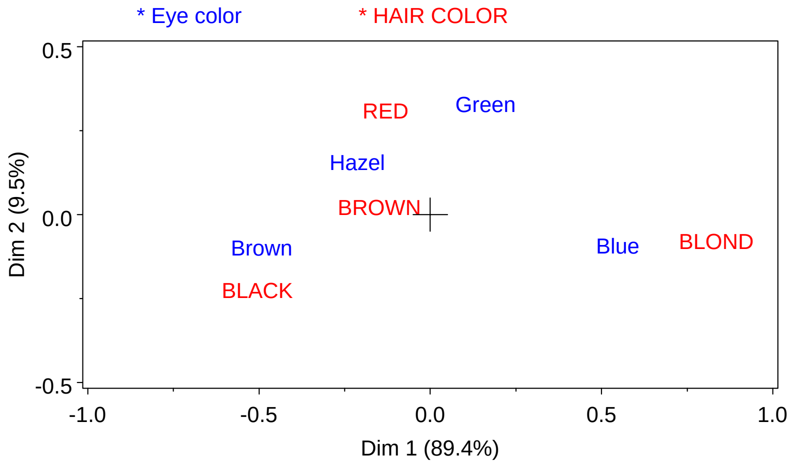
Correspondence analysis and MCA

- **Correspondence analysis (CA):** Analog of PCA for frequency data:
  - account for maximum % of χ^2 in few (2-3) dimensions
  - finds scores for row (x_{im}) and column (y_{jm}) categories on these dimensions
  - uses Singular Value Decomposition of residuals from independence, $d_{ij} = (n_{ij} - \widehat{m}_{ij}) / \sqrt{\widehat{m}_{ij}}$

$$\frac{d_{ij}}{\sqrt{n}} = \sum_{m=1}^{M} \lambda_m \, x_{im} \, y_{jm}$$

  - *optimal scaling*: each pair of scores, x_{im}) and column (y_{jm}, have highest possible correlation ($= \lambda_m$).
  - plots of the row (x_{im}) and column (y_{jm}) scores show associations
- **MCA:** Extends CA to n-way tables, but only uses bivariate associations (like mosaic matrix)

Hair color, Eye color data:



- Interpretation: row/column points “near” each other are positively associated
- Dim 1: 89.4% of χ^2 (dark ↔ light)
- Dim 2: 9.5% of χ^2 (RED/Green vs. others)

PROC CORRESP and the CORRESP macro

- Two forms of input dataset:
  - dataset in *contingency table* form – column variables are levels of one factor, observations (rows) are levels of the other.

```
Obs      Eye       BLACK     BROWN     RED     BLOND

 1       Brown       68       119       26        7
 2       Blue        20        84       17       94
 3       Hazel       15        54       14       10
 4       Green        5        29       14       16
```

  - Raw category responses (*case form*), or cell frequencies (*frequency form*), classified by 2 or more factors (e.g., output from PROC FREQ)

```
Obs      Eye       HAIR       Count

  1      Brown     BLACK        68
  2      Brown     BROWN       119
  3      Brown     RED          26
  4      Brown     BLOND         7
  ...
 15      Green     RED          14
 16      Green     BLOND        16
```

PROC CORRESP and the CORRESP macro

- PROC CORRESP
  - Handles 2-way CA, extensions to *n*-way tables, and MCA
  - Many options for scaling row/column coordinates and output statistics
  - OUTC= option ⟶ output dataset for plotting (PROC CORRESP doesn't do plots itself)
- **CORRESP macro**
  - Uses PROC CORRESP for analysis
  - Produces labeled plots of the category points in either 2 or 3 dimensions
  - Many graphic options; can equate axes automatically
  - See: `http://www.math.yorku.ca/vcd/corresp.html`

Example: Hair and Eye Color

- **Input the data** in contingency table form

corresp2a.sas ···

```
data haireye;
  input  EYE $ BLACK BROWN RED BLOND ;
  datalines;
        Brown     68    119     26      7
        Blue      20     84     17     94
        Hazel     15     54     14     10
        Green      5     29     14     16
;
```

Example: Hair and Eye Color

- **Using PROC CORRESP directly**— labeled printer plot

```
proc corresp data=haireye outc=coord short;
  id eye;                         /* row variable  */
  var black brown red blond;      /* col variables */
proc plot data=coord vtoh=2;      /* plot step     */
  plot dim2 * dim1 = '*' $eye
   / box haxis=by .1 vaxis=by .1; /* plot options */
```

- **Using the CORRESP macro**— labeled high-res plot

```
%corresp (data=haireye,
    id=eye,                     /* row variable  */
    var=black brown red blond,  /* col variables */
    dimlab=Dim);                /* options       */
```

Example: Hair and Eye Color

Printed output:

```
                The Correspondence Analysis Procedure

                 Inertia and Chi-Square Decomposition

  Singular  Principal Chi-
  Values    Inertias  Squares Percents   18   36   54   72   90
                                      ----+----+----+----+----+---
  0.45692   0.20877   123.593  89.37% *************************
  0.14909   0.02223    13.158   9.51% ***
  0.05097   0.00260     1.538   1.11%
            -------   -------
            0.23360    138.29 (Degrees of Freedom = 9)

                          Row Coordinates
                                Dim1          Dim2

                 Brown        -.492158      -.088322
                 Blue         0.547414      -.082954
                 Hazel        -.212597      0.167391
                 Green        0.161753      0.339040

                         Column Coordinates
                                Dim1          Dim2

                 BLACK        -.504562      -.214820
                 BROWN        -.148253      0.032666
                 RED          -.129523      0.319642
                 BLOND        0.835348      -.069579
```

Example: Hair and Eye Color

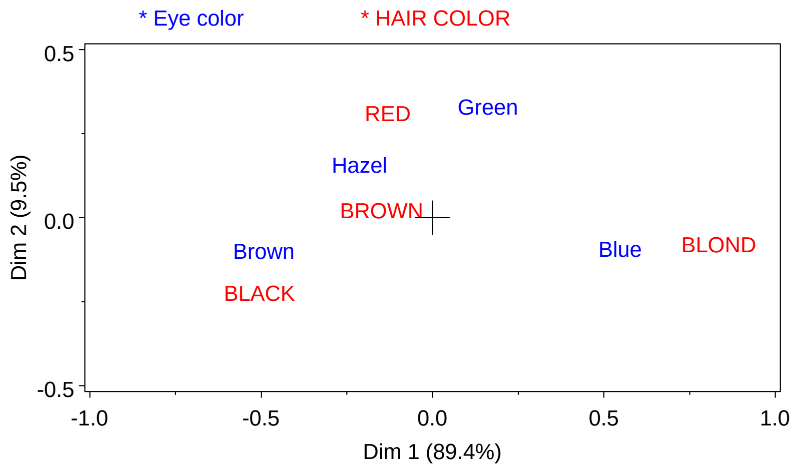
Output dataset(selected variables):

```
 Obs    _TYPE_      EYE        DIM1        DIM2

  1     INERTIA                 .           .
  2     OBS         Brown    -0.49216    -0.08832
  3     OBS         Blue      0.54741    -0.08295
  4     OBS         Hazel    -0.21260     0.16739
  5     OBS         Green     0.16175     0.33904
  6     VAR         BLACK    -0.50456    -0.21482
  7     VAR         BROWN    -0.14825     0.03267
  8     VAR         RED      -0.12952     0.31964
  9     VAR         BLOND     0.83535    -0.06958
```

Row and column points are distinguished by the _TYPE_ variable: OBS vs. VAR

Example: Hair and Eye Color

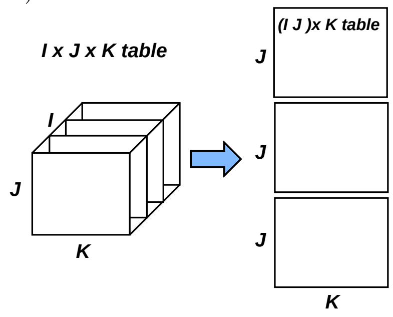
Graphic output from CORRESP macro:



- Top legend produced with Annotate data set and the INANNO= option to the CORRESP macro

Multi-way tables

- **Stacking approach:** van der Heijden and de Leeuw (1985)—
  - three-way table, of size $I \times J \times K$ can be sliced and stacked as a two-way table, of size $(I \times J) \times K$



  - The variables combined are treated “interactively”
  - Each way of stacking corresponds to a loglinear model
    - $(I \times J) \times K \rightarrow$ [AB][C]
    - $I \times (J \times K) \rightarrow$ [A][BC]
    - $J \times (I \times K) \rightarrow$ [B][AC]

Multi-way tables: Stacking

- PROC CORRESP: Use TABLES statement and option CROSS=ROW or CROSS=COL. E.g., for model [A B] [C],

```
proc corresp cross=row;
   tables A B, C;
   weight count;
```

- CORRESP **macro**: Can use / instead of ,

```
%corresp(
   options=cross=row,
   tables=A B/ C,
   weight count);
```

Example: Suicide Rates

Suicide rates in West Germany, by Age, Sex and Method of suicide

Sex	Age	POISON	GAS	HANG	DROWN	GUN	JUMP
M	10-20	1160	335	1524	67	512	189
M	25-35	2823	883	2751	213	852	366
M	40-50	2465	625	3936	247	875	244
M	55-65	1531	201	3581	207	477	273
M	70-90	938	45	2948	212	229	268
F	10-20	921	40	212	30	25	131
F	25-35	1672	113	575	139	64	276
F	40-50	2224	91	1481	354	52	327
F	55-65	2283	45	2014	679	29	388
F	70-90	1548	29	1355	501	3	383

- CA of the [Age Sex] by [Method] table:
  - Shows associations between the Age-Sex combinations and Method
  - Ignores association between Age and Sex

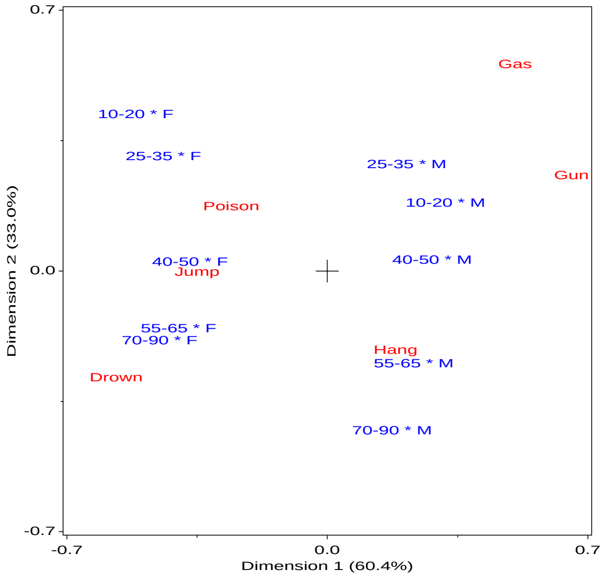
Example: Suicide Rates

suicide5.sas · · ·

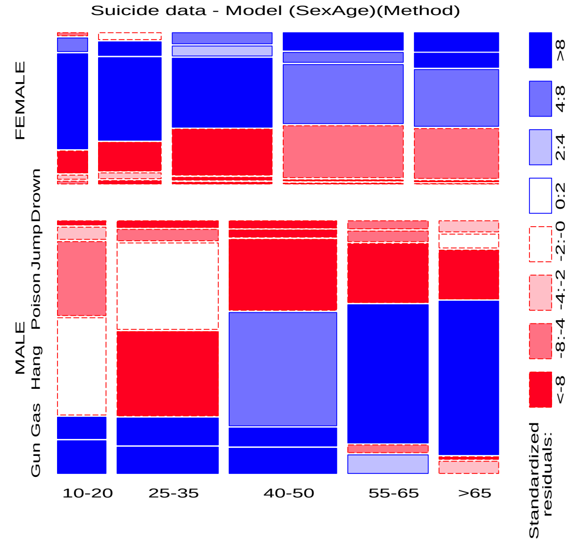
```
%include catdata(suicide);
   *-- equate axes!;
axis1 order=(-.7 to .7 by .7) length=6.5 in label=(a=90 r=0);
axis2 order=(-.7 to .7 by .7) length=6.5 in;
%corresp(data=suicide,  weight=count,
    tables=%str(age sex, method),
    options=cross=row short,
    vaxis=axis1, haxis=axis2);
```

Output:

```
                  Inertia and Chi-Square Decomposition

     Singular  Principal  Chi-
     Values    Inertias   Squares Percents   12   24   36   48   60
                                          ----+----+----+----+----+---
     0.32138   0.10328    5056.91   60.41% *************************
     0.23736   0.05634    2758.41   32.95% **************
     0.09378   0.00879     430.55    5.14% **
     0.04171   0.00174      85.17    1.02%
     0.02867   0.00082      40.24    0.48%
               -------    -------
               0.17098    8371.28  (Degrees of Freedom = 45)
```



Compare with mosaic display:


Suicide data - Model (SexAge)(Method)
FEMALE
MALE
Gun Gas Hang Poison Jump Drown
>65
>8
4:8
2:4
0:2
-2:-0
-4:-2
-8:-4
<-8
Standardized residuals: