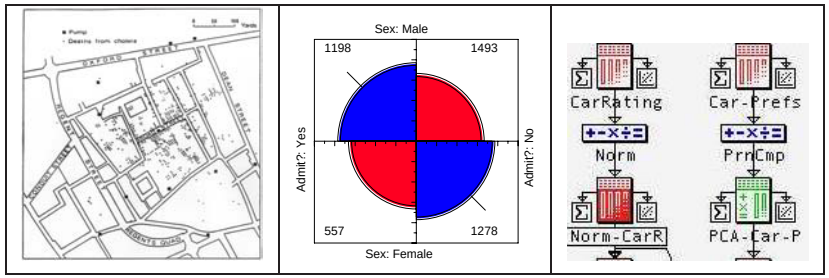
The Past, Present and Future of Statistical Graphics

(An Ideo-Graphic and Idiosyncratic View)



Michael Friendly

York University

http://www.math.yorku.ca/SCS/friendly.html

VIEWS, London, Nov, 2004

Outline

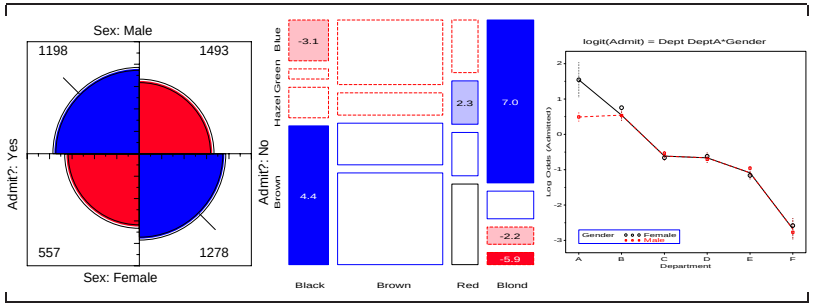
- Overview: Categorical Data *and* Graphics
- Methods for two-way frequency tables
  - Fourfold displays
  - Sieve diagrams
- Mosaic displays and loglinear models for *n*-way tables
  - Mosaic displays
  - Mosaic matrices
  - Software for mosaic displays
- Logistic and logit regression
  - Logit plots, effect plots
  - Diagnostic plots

Color version of these slides:
http://www.math.yorku.ca/SCS/Papers/views/

Part 3: Graphical Methods for Categorical Data

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

- Overview: Categorical Data *and* Graphics
- Methods for two-way frequency tables
- Mosaic displays and loglinear models for *n*-way tables
- Logistic and logit regression



Categorical Data Analysis

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

- **Non-parametric, randomization-based methods**
  - make minimal assumptions
  - useful for hypothesis-testing
  - SAS: `PROC FREQ`
    - Pearson Chi-square
    - Fisher's exact test (for small expected frequencies)
    - Mantel-Haenszel tests (ordered categories: test for *linear* association)
- **Model-based methods**
  - Must assume random sample (possibly stratified)
  - Useful for estimation purposes
  - Greater flexibility; fitting specialized models (e.g., symmetry)
  - More suitable for multi-way tables
  - SAS: `PROC LOGISTIC`, `CATMOD`, `GENMOD` , `INSIGHT` (Fit YX)
    - estimate standard errors, covariances for model parameters
    - confidence intervals for parameters, predicted Pr{response}

Graphical Methods for Categorical Data

- **Exploratory methods**
  - Minimal assumptions (like non-parametric methods)
  - Show the *data*, not just *summaries*
  - 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 *SSSG* - Make these methods *available* and *accessible* in SAS
  - **Practical power = Statistical power $\times$ 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 `www.math.yorku.ca/SCS/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

AGREE	Observer agreement chart
CORRESP	Plot PROC CORRESP results
FFOLD	Fourfold displays for $2 \times 2 \times k$ tables (macro)
FOURFOLD	Fourfold displays for $2 \times 2 \times k$ tables (SAS/IML)
SIEVEPLOT	Sieve diagrams
MOSAIC	Mosaic displays (macro)
MOSAICS	SAS/IML modules for mosaic displays
MOSMAT	Mosaic matrices (macro)
TABLE	Construct a grouped frequency table, with recoding
TRIPLOT	Trilinear plots for $n \times 3$ tables

Model-based methods

ADDVAR	Added variable plots for logistic regression
CATPLOT	Plot results from PROC CATMOD
HALFNORM	Half-normal plots for generalized linear models
INFLGLIM	Influence plots for generalized linear models
INFLOGIS	Influence plots for logistic regression
LOGODDS	Plot empirical logits and probabilities for binary data
POWERLOG	Power calculations for logistic regression
POWERRxC	Power calculations for two-way frequency table
POWER2x2	Power calculations for a 2×2 table
ROBUST	Robust fitting for linear models

Utility macros

DUMMY	Create dummy variables
LAGS	Calculate lagged frequencies for sequential analysis
PANELS	Arrange multiple plots in a panelled display
SORT	Sort a dataset by the value of a statistic or formatted value
Utility	Graphics utility macros: BARS, EQUATE, GDISPLA, GENSYM, GSKIP, LABEL, POINTS, PSCALE

VCD Archive (`vcdprog.zip`) available to purchasers at:
`support.sas.com/publishing/bbu/56571_sample.html`

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 6: 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, $\theta = \dfrac{\text{Odds(Admit}\,|\,\text{Male)}}{\text{Odds(Admit}\,|\,\text{Female)}} = \dfrac{1198/1493}{557/1276} = 1.84$
  - $\longrightarrow$ Males 84% more likely to be admitted.

Standard analysis: PROC FREQ

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

Output:

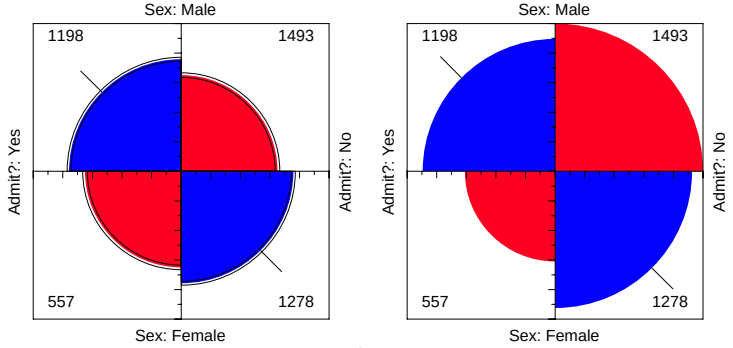
```
        Statistics for Table of gender by admit

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

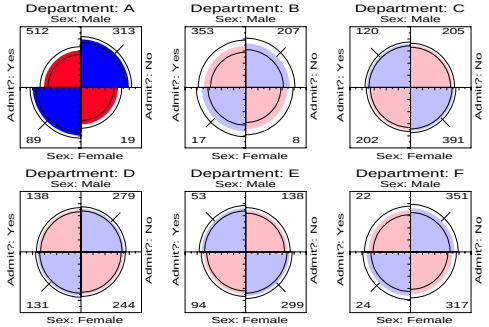
- ***Quarter circles***: radius $\sim \sqrt{n_{ij}} \Rightarrow$ **area** $\sim$ **frequency**
- ***Independence***: Adjoining quadrants $\approx$ align
- ***Odds ratio***: ratio of areas of diagonally opposite cells
- ***Confidence rings***: Visual test of $H_0 : \theta = 1 \leftrightarrow$ adjoining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



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

Fourfold displays for 2 × 2 × *k* tables

- Data in Table 6 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.

Sidebar: Using SAS macros

- 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:

```
%macname(data=dataset, var=variables, ...);
```

e.g.,

```
%boxplot(data=nations, var=imr, class=region, id=nation);
```

- Most arguments have default values (e.g., data=_last_)
- All *SAS System for Statistical Graphics, First Edition* and *VCD* macros have internal and/or online documentation,
  http://www.math/yorku.ca/SCS/sssg/
  http://www.math/yorku.ca/SCS/sasmac/
  http://www.math/yorku.ca/SCS/vcd/
- Macros can be installed in directories *automatically* searched by SAS. Put the following options statement in your AUTOEXEC.SAS file:

```
options sasautos=('c:\sasuser\macros' sasautos);
```


Sidebar: Using SAS macros

E.g., the FFOLD macro (fourfold displays) is defined with the following arguments:

ffold.sas ···

```
%macro ffold(
   data=_last_,      /* Name of input dataset                */
   var=,             /* Names of 2x2 factor variable         */
   by=,              /* Name(s) of BY variables              */
   count=count,      /* Name of the frequency variable       */
   std=,             /* How to standardize tables?           */
   config=,          /* margins to standardize               */
   down=,            /* number of panels down each page      */
   across=,          /* number of panels across each page    */
   order=,           /* DOWN|ACROSS - arrange multiple plots */
   ...
   );
```

Typical use:

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


Two-way frequency tables

- Example: data on hair-color and eye-color for 592 students

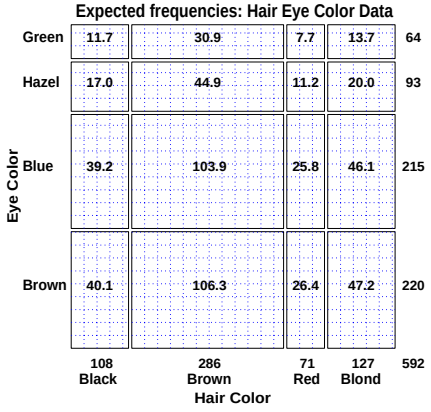
Table 7: Hair-color eye-color data

Eye Color	Hair 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

- Questions:
  - Is there an association between hair-color and eye-color?
  - ≡ Given hair-color, are some eye-colors more likely?
  - If associated, how to understand the *pattern* of the relation?
- Graphs: sieve diagrams, mosaic displays

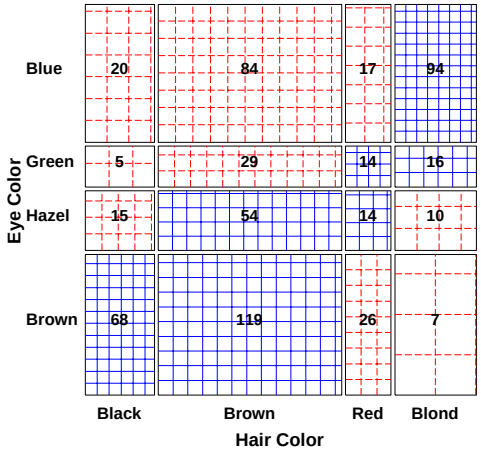
Two-way frequency tables: Sieve diagrams

- **count ∼ 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 $\times$ width $\sim$ frequency, n_{ij}



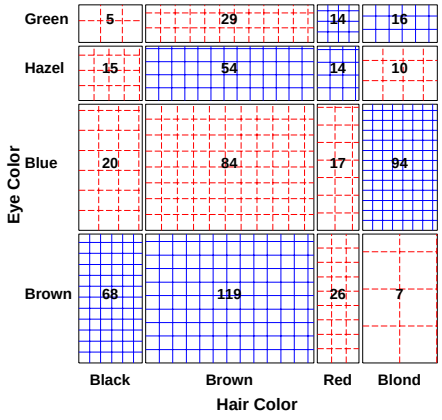
Sieve diagrams

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



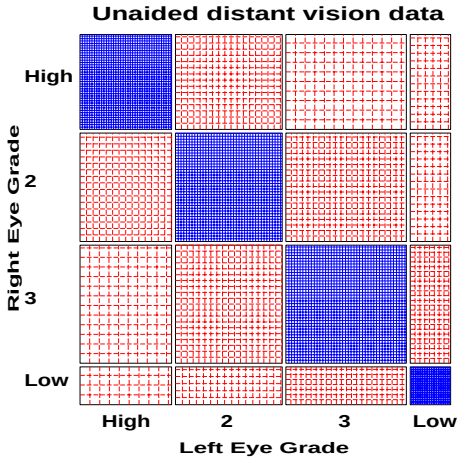
Sieve diagrams

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



Sieve diagrams

- Vision classification data for 7477 women

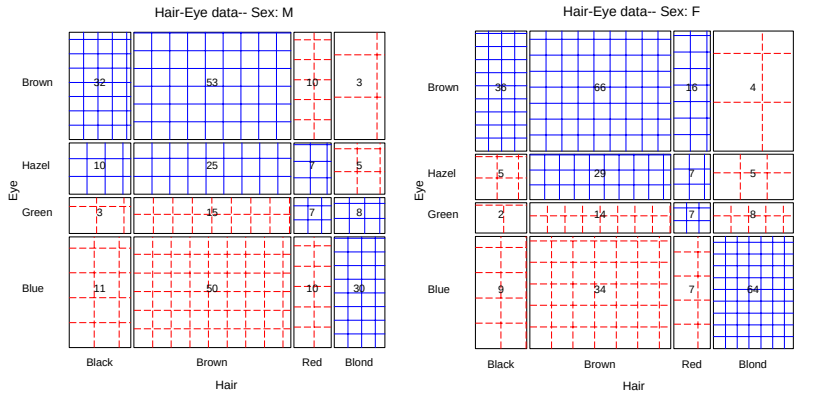


The `sieveplot` macro

- Sieve plots for a single table, or multiple two-way plots for 3- and higher-way tables.

haireyem.sas ···

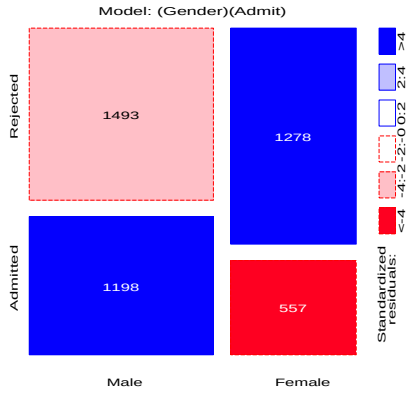
```
*-- Separate plots for males and females;
%sieveplot(data=haireye, var=Eye Hair Sex,
    by=Sex,
    title=Hair-Eye data--,
    filltype=obsp);
```



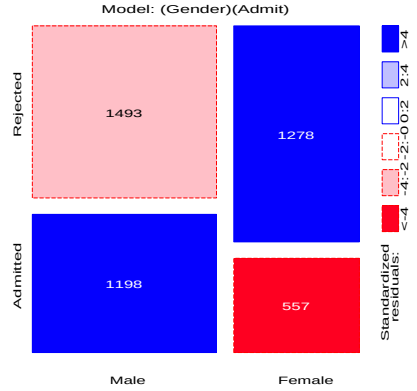
Mosaic displays and Log-linear Models

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

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



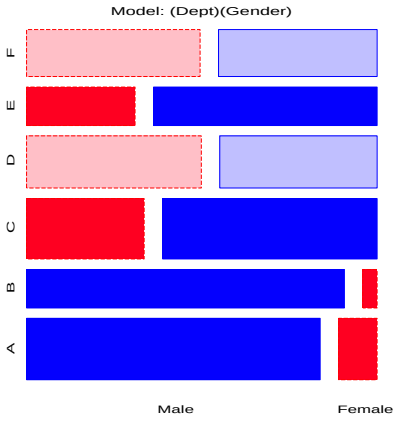
- ***Shading***: Sign and magnitude of Pearson χ^2 residual, $d_{ij} = (n_{ij} - \hat{m}_{ij})/\sqrt{\hat{m}_{ij}}$ (or L.R. G^2)
  - Sign: $-$ negative in red; $+$ positive in blue
  - Magnitude: intensity of shading: $|d_{ij}| > 0, 2, 4, \ldots$
- ***Independence***: Rows $\approx$ align, *or* cells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



Mosaic displays

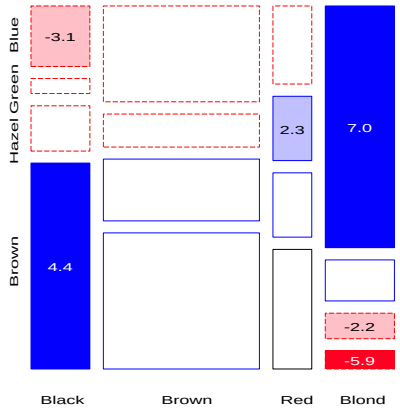
Departments $\times$ Gender:

- 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 8: Log-linear Models for Three-Way Tables

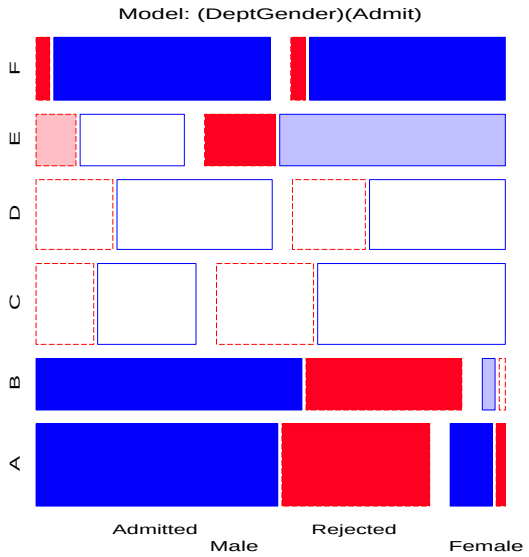
Model	Model symbol	Independence interpretation
Mutual independence	$[A][B][C]$	$A \perp B \perp C$
Joint independence	$[AB][C]$	$(A\,B) \perp C$
Conditional independence	$[AC][BC]$	$(A \perp B) \,\vert\, C$
All two-way associations	$[AB][AC][BC]$	(none)
Saturated model	$[ABC]$	(none)

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

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

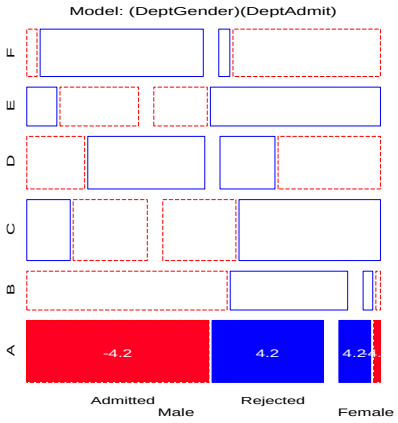
- Each mosaics shows:
  - **DATA** (size of tiles)
  - (some) **marginal** frequencies (spacing → visual grouping)
  - **RESIDUALS** (shading) — what associations have been omitted?

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



Mosaic displays for multiway tables

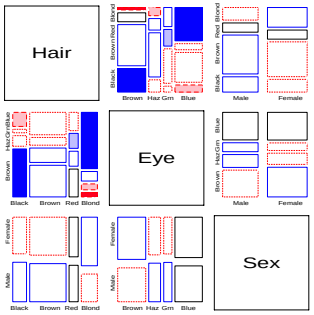
- Visual fitting:
  - Pattern of lack-of-fit (residuals) → "better" model— smaller residuals
  - "cleaning the mosaic" → "better" model— empty cells
  - best done interactively!

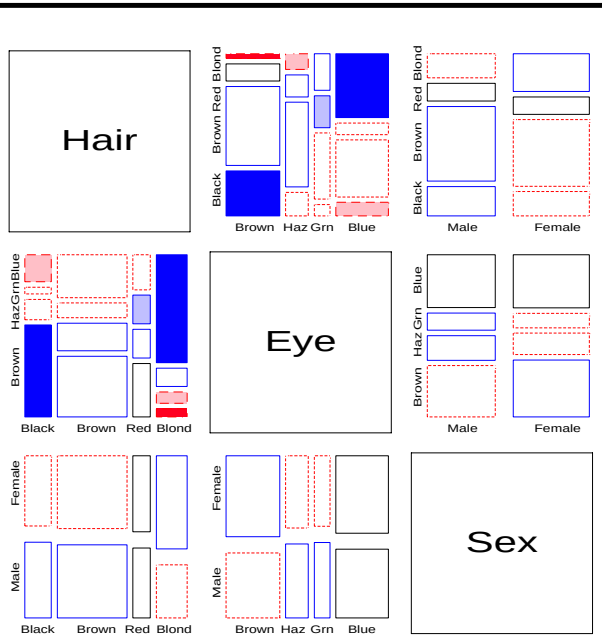


- E.g., Add [Dept Admit] association → Conditional independence:
  - Fits poorly, overall ($G^2_{(6)}$ = 21.74)
  - But, only in Department A!

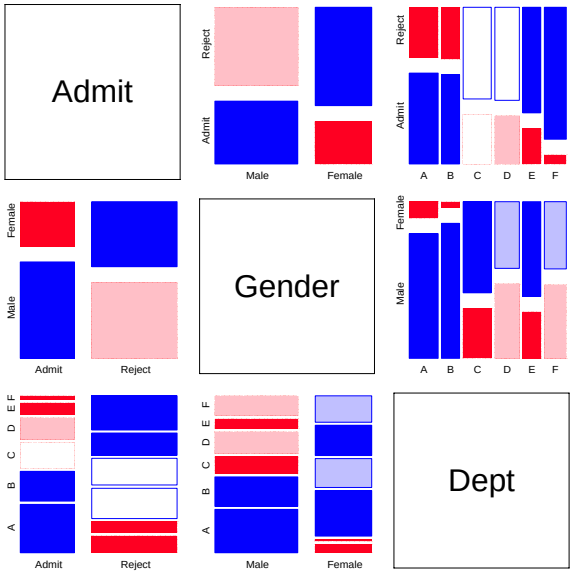
Mosaic matrices

- 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



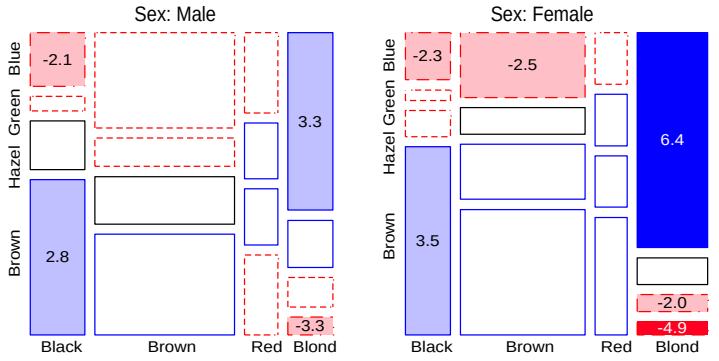


Berkeley data:



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,

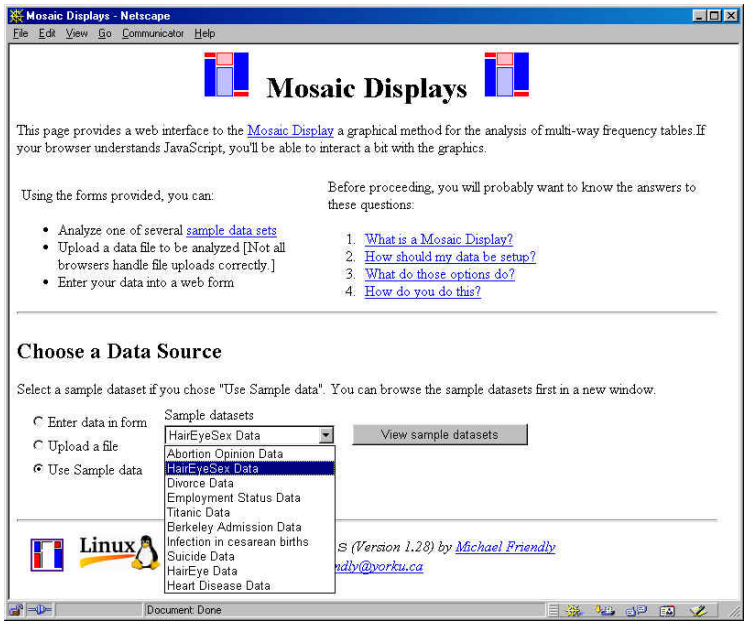
$$G^2_{A \perp B \,|\, C} = \sum_k G^2_{A \perp B \,|\, C(k)}$$

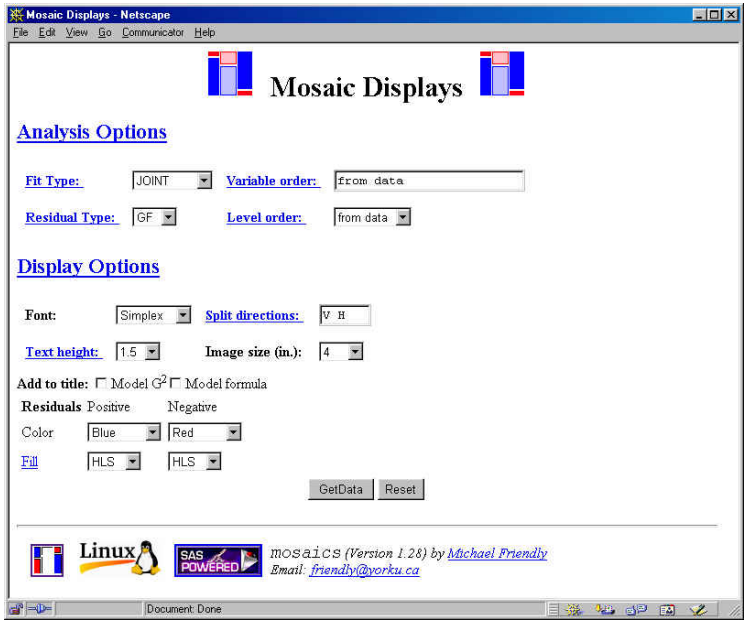
Table 9: Partial and Overall conditional tests, $Hair \perp Eye \,|\, Sex$

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

- **SAS software**: `http://www.math.yorku.ca/SCS/mosaics.html`
  - SAS/IML modules, `mosaics.sas`
  - `mosaic` macro– input from dataset, n-way mosaics, partial mosaics
  - `mosmat` macro– mosaic matrices
- **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).
  - Interactively query cell frequencies, residuals, etc.





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` program
- **Macro interface**: `mosaic` macro, `table` macro, `mosmat` macro
- `mosaic` **macro**
  - Direct input from a SAS dataset
  - No knowledge of SAS/IML required
  - Reorder table variables; collapse, reorder variable levels with `table` macro
  - Convenient interface to *partial mosaics* (BY=)
  - Specialized models (quasi-independence) can be fit using `PROC GENMOD`

`mosaic` macro example: Berkeley data

Input the frequency data:

berkeley.sas

```
title 'Berkeley Admissions data';
data berkeley;
   do dept = "A", "B", "C", "D", "E", "F";
      do gender = 'Male   ', 'Female';
         do admit = 'Admitted', 'Rejected';
            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 */
;
```


`mosaic` macro example: Berkeley data

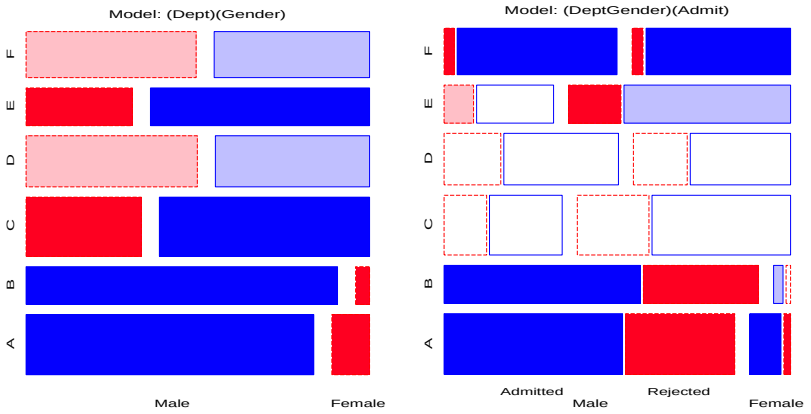
Call the `mosaic` macro

mosaic9m.sas

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

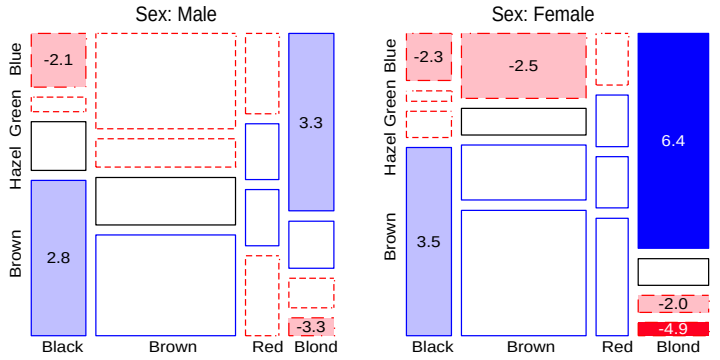
Three-way, Dept. by Gender by Admit

Partial mosaics

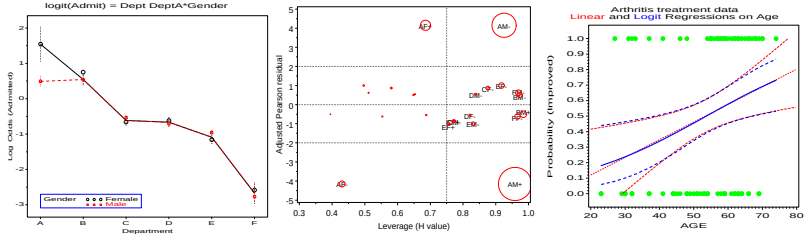
mospart3.sas

```
%include catdata(hairdat3s);

%mosaic(data=haireye,
    vorder=Hair Eye Sex,
    by=Sex,
    cellfill=dev);
```



Model-based methods for categorical data



Topics:

- Logit models
  - Plots for logit models
  - Diagnostic plots for generalized linear models
  - Effect plots for generalized linear models

Logit models

For a binary response, each loglinear model is equivalent to a logit model (logistic regression, with categorical predictors)

- Admit ⊥ Gender | Dept (conditional independence, ↔ loglinear [AD][DG])

$$\log m_{ijk} = \mu + \lambda_i^A + \lambda_j^D + \lambda_k^G + \lambda_{ij}^{AD} + \lambda_{jk}^{DG}$$

$$\leftrightarrow L_{jk} = \log(m_{1jk}/m_{2jk}) = (\lambda_1^A - \lambda_2^A) + (\lambda_{1j}^{AD} - \lambda_{2j}^{AD}) = \alpha + \beta_j^{\text{Dept}}$$

- All two-way associations ↔ loglinear [AD][DG][AG])

$$\log m_{ijk} = \mu + \lambda_i^A + \lambda_j^D + \lambda_k^G + \lambda_{ij}^{AD} + \lambda_{jk}^{DG} + \lambda_{ik}^{AG}$$

$$\leftrightarrow L_{jk} = \log(m_{1jk}/m_{2jk}) = \alpha + \beta_i^{\text{Gender}} + \beta_j^{\text{Dept}}$$

where,

- $L_{ij} = \log(m_{ij1}/m_{ij2})$: log odds of admission,
- β_i^{Gender}: effect on admissions of gender (bias!),
- β_j^{Dept}: effect on admissions of department,

Logit models

- **Fitting procedures**
  - `PROC CATMOD`
  - `PROC LOGISTIC`
  - `PROC GENMOD / dist=poisson`
  - SAS/INSIGHT (Fit Y X) Options ⟶ Distribution poisson
- **Visualization procedures**
  - `CATPLOT` macro - plot predicted, observed log odds from `CATMOD`
  - `INFLGLIM` macro - influence plots for generalized linear models
  - `HALFNORM` macro - half-normal plot of residuals for generalized linear models
- **SAS craft**
  - All SAS procedures ⟶ output dataset with obs., fitted values, residuals, diagnostics, etc.
  - New model ⟶ new output dataset
  - Plotting steps remain the same

Plots for logit models

- **Model:** Admit ∼ Gender + Dept ↔ [AD] [AG] [DG]

catberk2.sas ···

```
%include catdata(berkeley);
proc catmod order=data
            data=berkeley;
   weight freq;
   response / out=predict;
   model admit = dept gender / ml;
  run;
```

PROC CATMOD output: Overall tests and goodness of fit

Maximum Likelihood Analysis of Variance

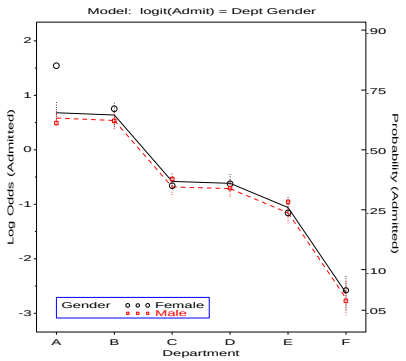
Source	DF	Chi-Square	Pr > ChiSq
Intercept	1	262.49	<.0001
dept	5	534.78	<.0001
gender	1	1.53	0.2167
Likelihood Ratio	5	20.20	0.0011

- No effect of Gender
- Model doesn't fit well— Why? How to modify?

Plots for logit models

- Fit: PROC CATMOD; plot: CATPLOT macro

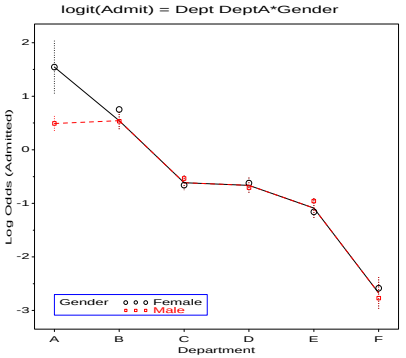
```
proc catmod order=data data=berkeley;
   weight freq;
   response / out=predict;
   model admit = dept gender / ml;
%catplot(data=predict, xc=dept, class=gender,
    type=FUNCTION, z=1.96, legend=legend1);
```



Fitting and graphing other models

- Change MODEL statement ⟶ new fitted values
- Plotting step remains the same
- Admit ⊥ Gender | Dept, except for Dept. A ↔ Admit ∼ Dept + $\delta_{j=1}$ Gender

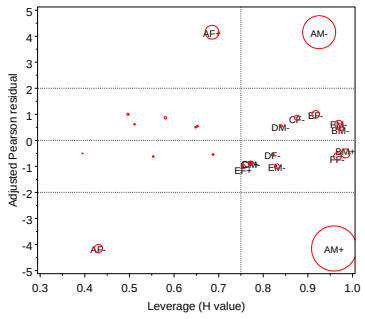
```
proc catmod order=data data=berkeley;
   response / out=predict;
   model admit = dept dept1AG / ml;
%catplot(data=predict, xc=dept, class=gender,
    type=FUNCTION, z=1.96, legend=legend1);
```



Diagnostic plots for Generalized Linear Models

INFLGLIM macro: Influence plots for generalized linear models

- Which cells have undue impact on fitted model?
- Fit: PROC GENMOD; calculates additional diagnostic measures (Hat value, Cook's D, etc.)
- Plot: measures of residual (GY=$\Delta\chi^2$, χ^2 residual) vs. leverage (GX=hat value), bubble size (area, radius) ∼ Cook's D.



INFLGLIM macro: Example

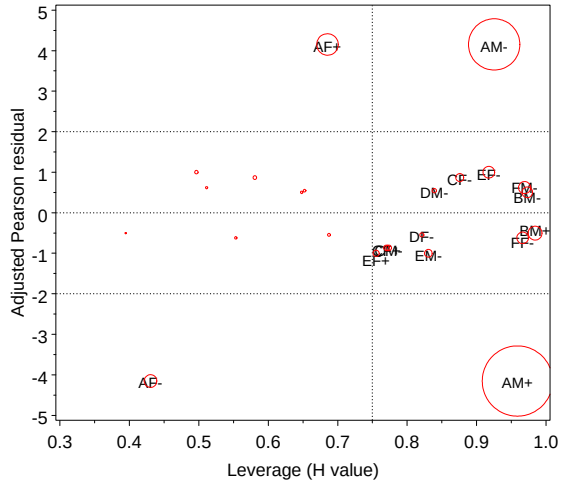
- Berkeley data, loglinear model [AD][GD] $\leftrightarrow L_{ij} = \alpha + \beta_j^{\text{Dept}}$

genberk1.sas

```
%include catdata(berkeley);
*-- make a cell ID variable, joining factors;
data berkeley;
  set berkeley;
  cell = trim(put(dept,dept.)) ||
         gender ||
         trim(put(admit,yn.));

%inflglim(data=berkeley,
   class=dept gender admit,
   resp=freq,
   model=admit|dept gender|dept,
   dist=poisson,
   id=cell,
   gx=hat, gy=streschi);
```


INFLGLIM macro: Example



- All cells which do not fit ($|r_i| > 2$) are for department A.
- Males applying to dept A have large leverage $\Rightarrow$ large influence (Cook's D)

Effect plots: Racial profiling? example

- Data on Toronto police treatment of people arrested for "simple posession" of marijuana (*Toronto Star* series on racial profiling)
- Outcome: Released (with summons) vs. {taken to police station for further questioning; held for bail hearing; ...}
- **Q**: How does release depend on race, age, sex and other variables?
  - `colour`: Black, White (others ignored)
  - `checks`: Number of police databases in which person's name appears on arrest (0–6)

Released	Colour	Year	Age	Sex	Employed	Citizen	Checks
Yes	White	2002	21	Male	Yes	Yes	3
No	Black	1999	17	Male	Yes	Yes	3
Yes	White	2000	24	Male	Yes	Yes	3
No	Black	2000	46	Male	Yes	Yes	1
Yes	Black	1999	27	Female	Yes	Yes	1
Yes	Black	1998	16	Female	Yes	Yes	0
Yes	White	1999	40	Male	No	Yes	0
Yes	White	1998	34	Female	Yes	Yes	1
Yes	Black	2000	23	Male	Yes	Yes	4
Yes	White	2001	30	Male	Yes	Yes	3
...							

Effect plots: Racial profiling? example

Fitted model:

arrests.sas ···

```
proc logistic data=arrests descending;
   class Colour Sex Employed Citizen Year /descending;
   model Released = Employed Citizen Checks
                    Colour|Year Colour|Age;
```

Output:

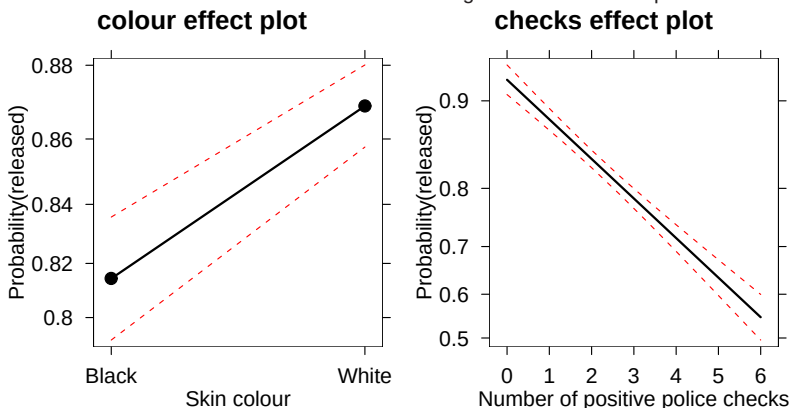
Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Employed	1	75.1924	<.0001
Citizen	1	26.5260	<.0001
Checks	1	198.3283	<.0001
Colour	1	21.8131	<.0001
Year	5	4.1158	0.5329
Colour*Year	5	21.3541	0.0007
Age	1	3.8176	0.0507
Age*Colour	1	13.4217	0.0002

How can we understand these effects?

Effect plots: Racial profiling? example

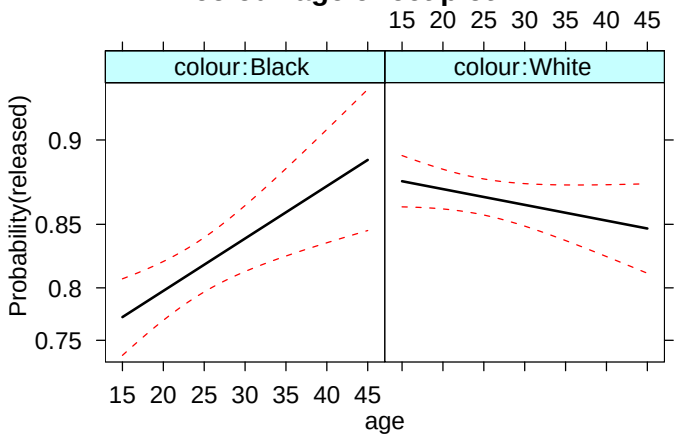
Main effects: Predicted values calculated at average levels of all other predictors



- Controlling for all other factors, there is evidence that blacks are treated more harshly than whites.
- Is this the entire story?

Effect plots: Racial profiling? example

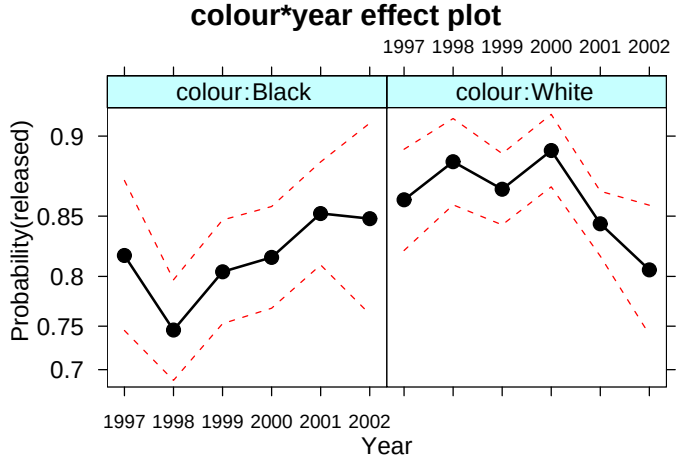
Colour × Age: Predicted values at average levels of all other predictors



- Release treatment of blacks and whites differed strongly according to age.

Effect plots: Racial profiling? example

Colour × Year: Predicted values at average levels of all other predictors



- Release treatment of blacks and whites changed over years, in different patterns.