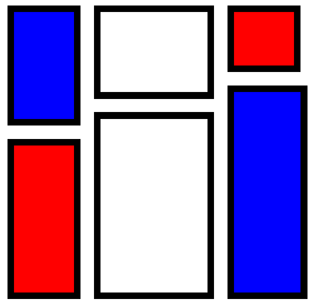
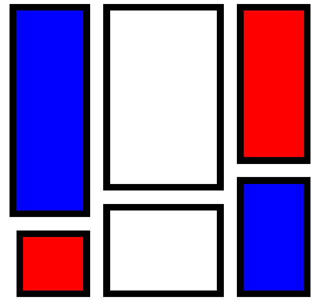
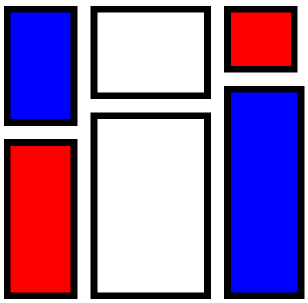
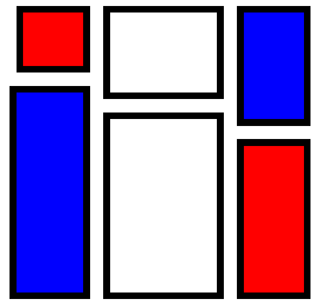
Extending

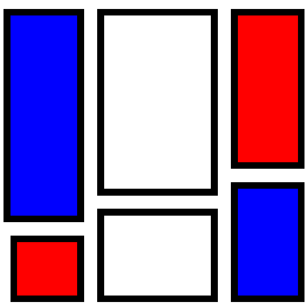


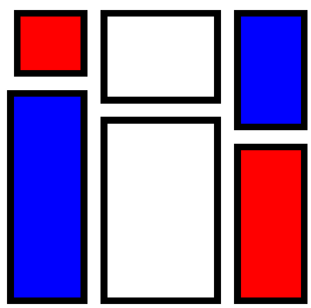




Mosaic







Displays

Marginal, Partial, and Conditional Views of Categorical Data

Michael Friendly
York University

July, 1998

Data Visualization in Statistics Workshop

Outline

- 1. Introduction - graphical methods for categorical data
- 2. Fourfold displays
- 3. Mosaic displays
  - Fitting loglinear models with mosaic displays
  - Examples
- 4. Mosaic matrices
  - Marginal views
  - Conditional views
- 5. Mosaic coplots for categorial data

Graphical Methods for Categorical Data

- Goals: develop graphical methods for categorical data which serve needs of
  - ***reconnaissance***—a preliminary overview of complex terrain;
  - ***exploration***—help detect patterns or unusual circumstances, or suggest hypotheses;
  - ***model building & diagnosis***—critique a fitted model as a reasonable statistical summary.
- Attempt to integrate these with methods for continuous data
- Categorical data needs a different visual representation: count $\sim$ area (Friendly, 1995)
- Static vs Dynamic displays

A dim idea: Two-way table display

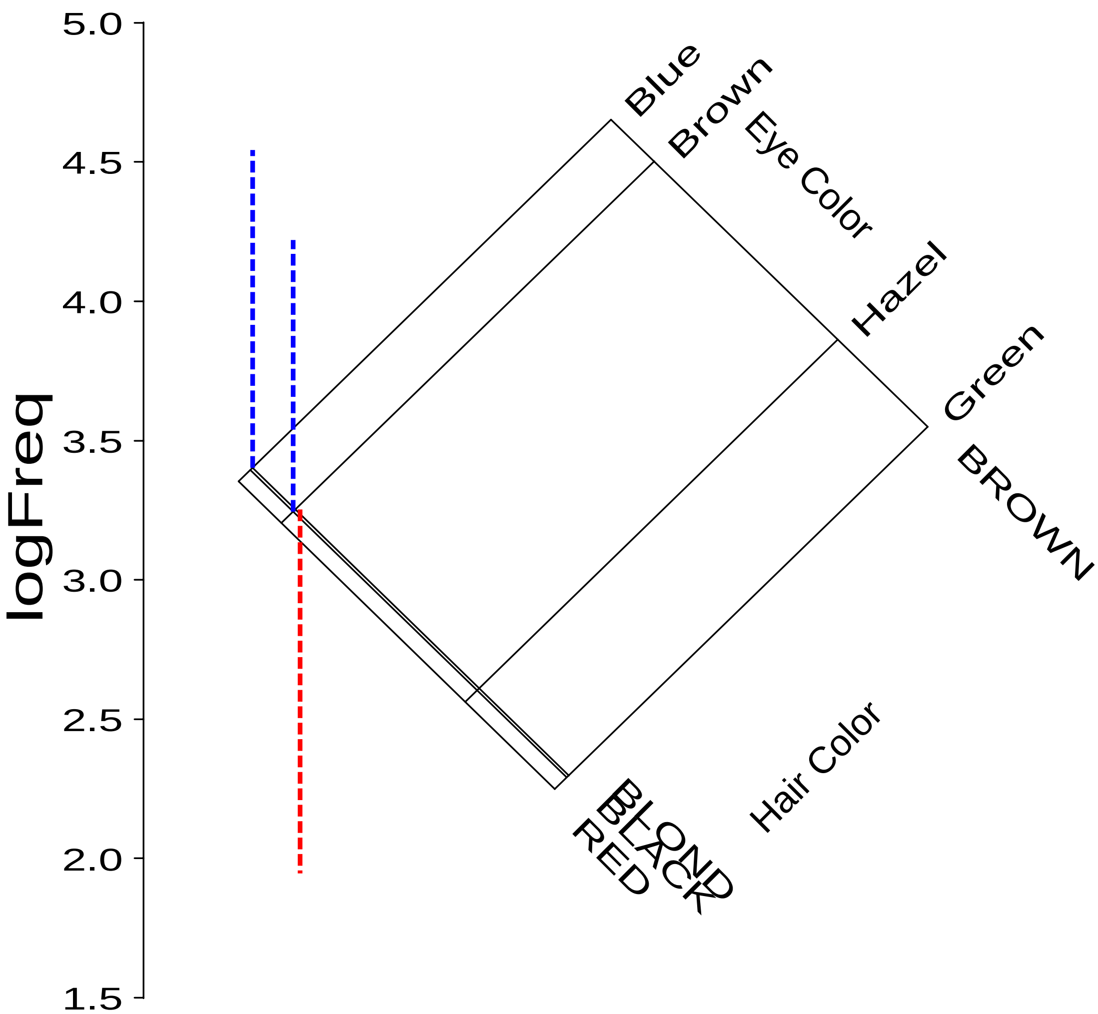
- For a two-way table, the saturated log-linear model is formally equivalent to a two-factor ANOVA model.

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

$$E(Y_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

- Suggests use of Tukey's two-way display of log(Freq)
  - Rows and columns ordered by mean log(Freq)
  - Vertical position shows fitted log(Freq) under independence
  - Residuals show deviations from independence

- But: main effect ordering not useful for counts - interest is on interactions



Hair2way

Graphic metaphor: count ~ area

- $A \perp B \longrightarrow p_{ij} = p_{i+} \times p_{+j}$
- $\therefore$ each cell can be drawn as a rectangle, with area = height $\times$ width = frequency.

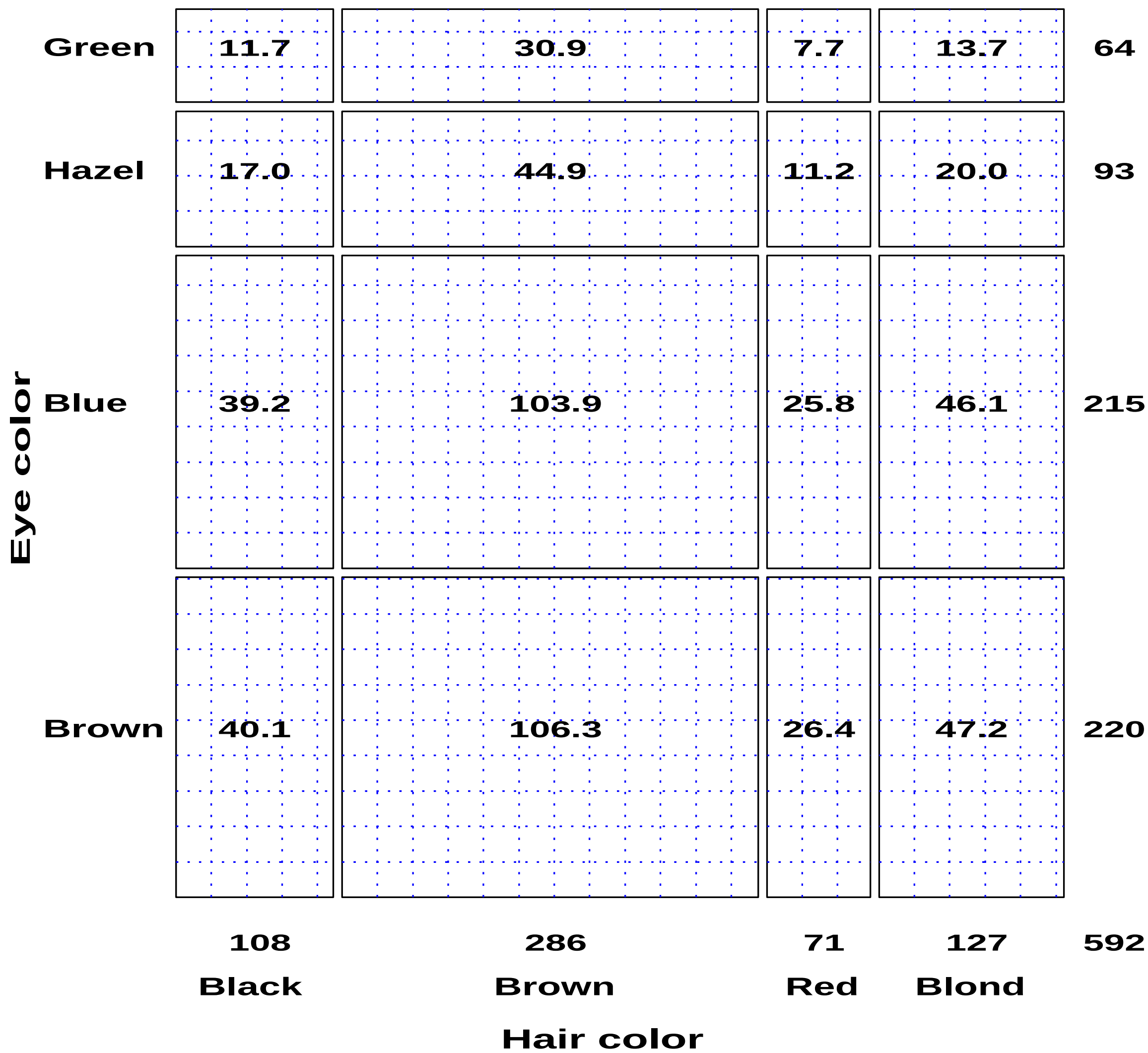


Figure 1: Expected frequencies under independence.

Fourfold display for $2 \times 2 \times 2$ tables

- Quarter circles: **area ∼ frequency**
- **Odds ratio**: θ = ratio of diagonally opposite cells
- **Standardize**: equal margins, same odds ratip (IPF)
- **Independence**: Adjacent segments equal
- **Confidence rings**: Overlap $\longleftrightarrow$ Accept $H_0 : \theta = 1$

Ex: Berkeley admissions data $\theta = \Pr \left(\frac{\text{Admit} | \text{Male}}{\text{Admit} | \text{Female}} \right) = 1.84$

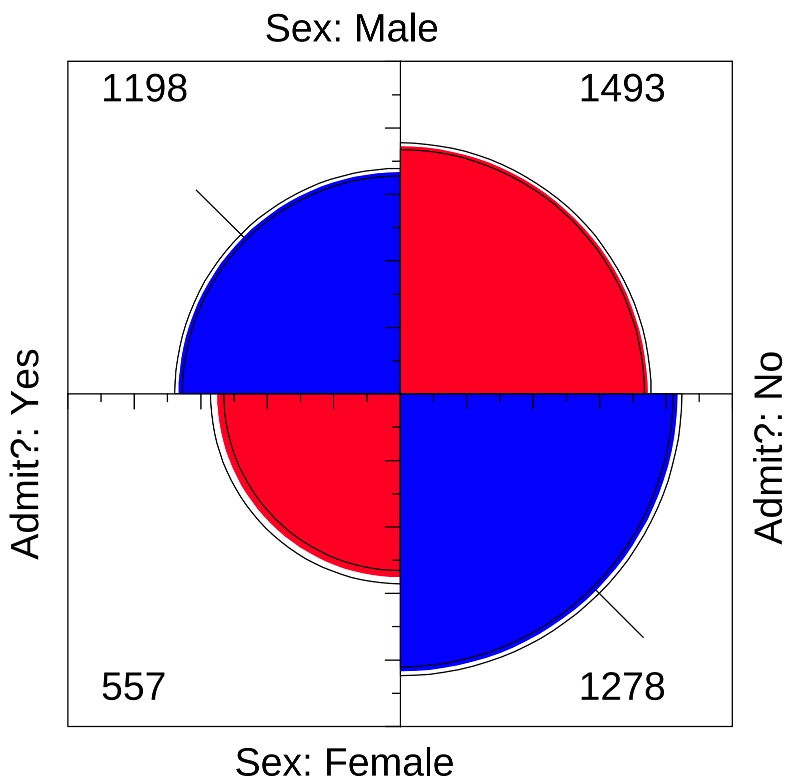


Figure 2: Berkeley admissions: Evidence for sex bias?

Multiple strata

- Multiple strata - one for each
- (Different rates of acceptance not visible)

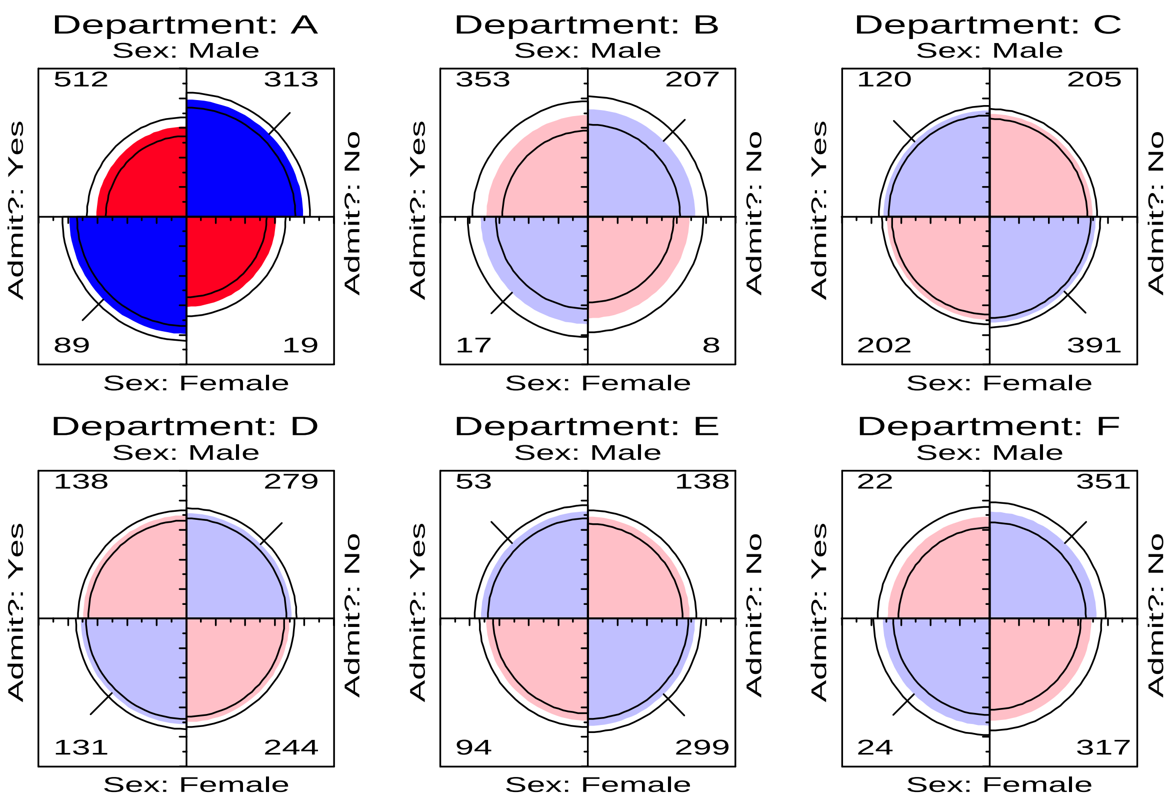


Figure 3: Berkeley admissions, by Department

Visualization principles

- **Controlled comparison** - compare, holding other things constant
  - Hold angles constant, vary radius $\longrightarrow$ corresponding cells in same position.
  - Equate row, col, or both margins, while keeping odds ratio fixed
- **Visual impact** - distinguish what should stand out ($\theta \neq 1$)

Mosaic displays

- Width $\sim$ one set of marginal probabilities, p_{i+}
- Height $\sim$ conditional probabilities, $p_{j\mid i}$
- area $\sim$ count, n_{ij}.
- **Independence**: Shown when cells align

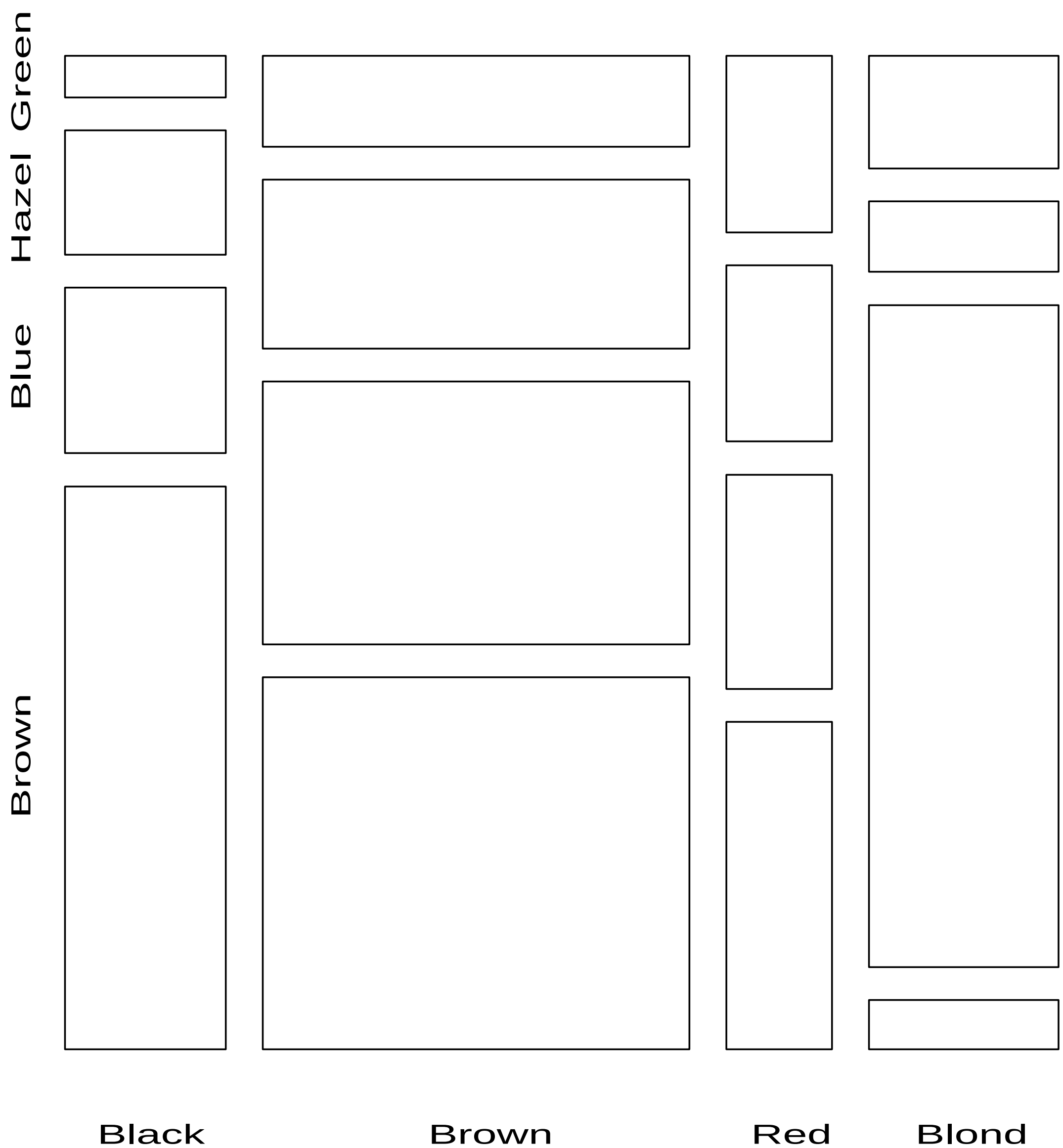


Figure 4: Basic mosaic display for hair-color and eye color data.

Enhanced mosaic display

- **Display residuals**, d_{ij} , by color and shading
  - Sign: color ($d_{ij} > 0$, $d_{ij} < 0$) Magnitude: $|d_{ij}| \sim$ darkenss
- **Reorder categories** - opposite corner pattern (CA scores)
- **Independence**: Cells are empty! ($d_{ij} \approx 0$: black)

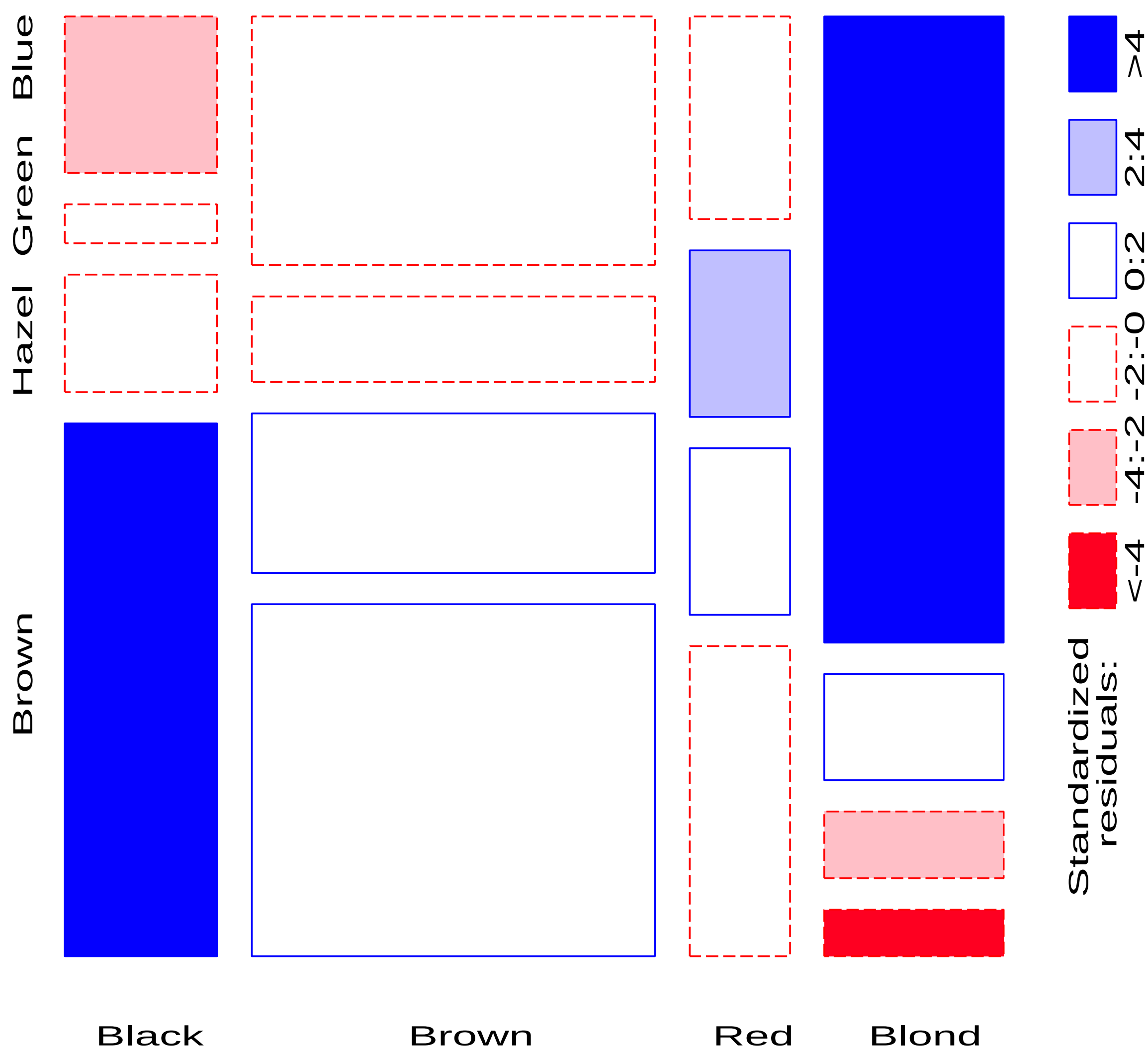


Figure 5: Extended mosaic, reordered and shaded.

Multi-way tables

- Generalizes to *n*-way tables (divide recursively)
- Can fit *any* log-linear model, e.g., [AB][C], [AC][BC], etc.
- Shows both the **DATA** (area) and **RESIDUALS** (shading)

Example: Joint Independence, $G^2(15) = 19.86$. (Do blue-eyed blonds have more fun?)

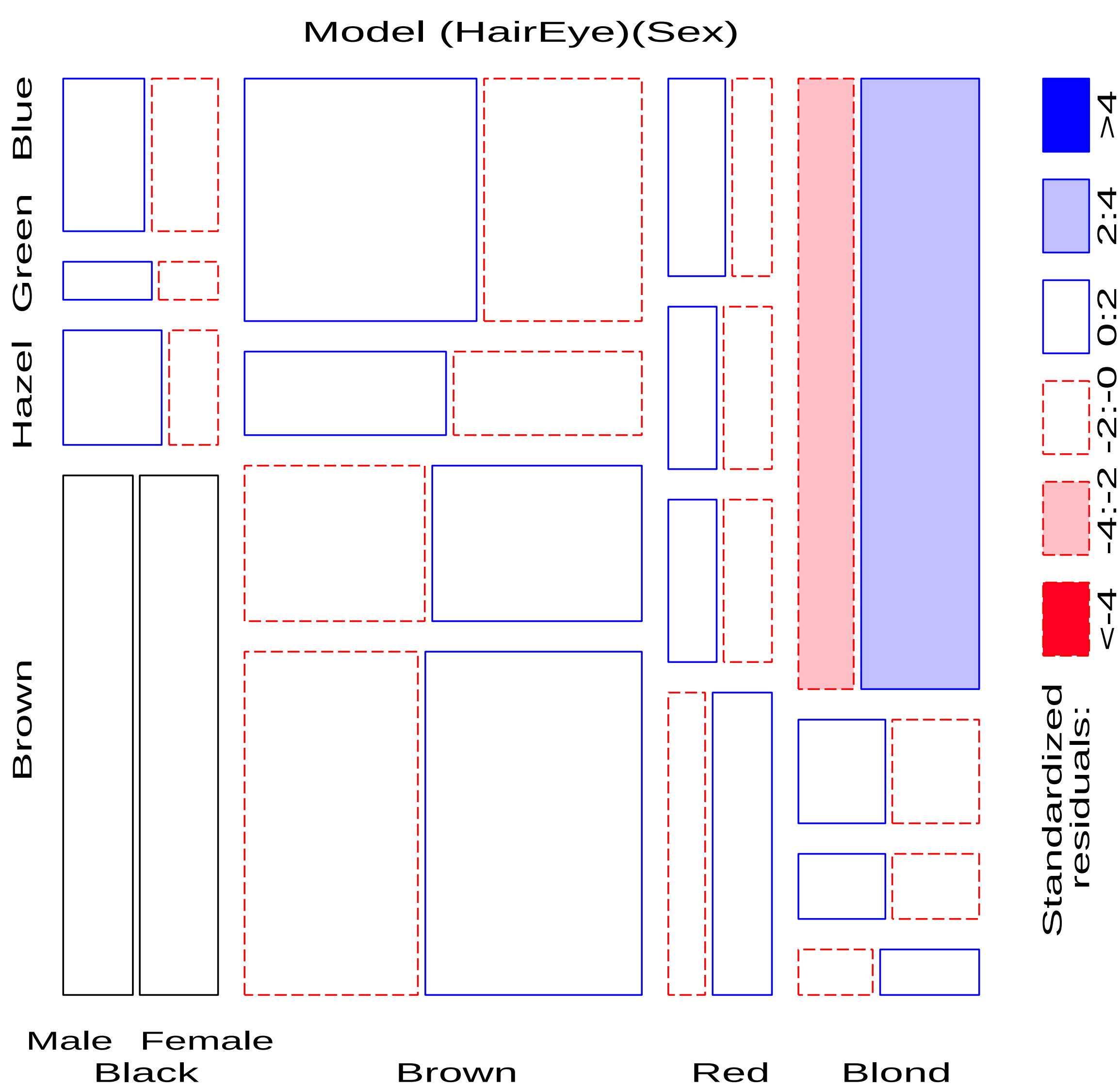


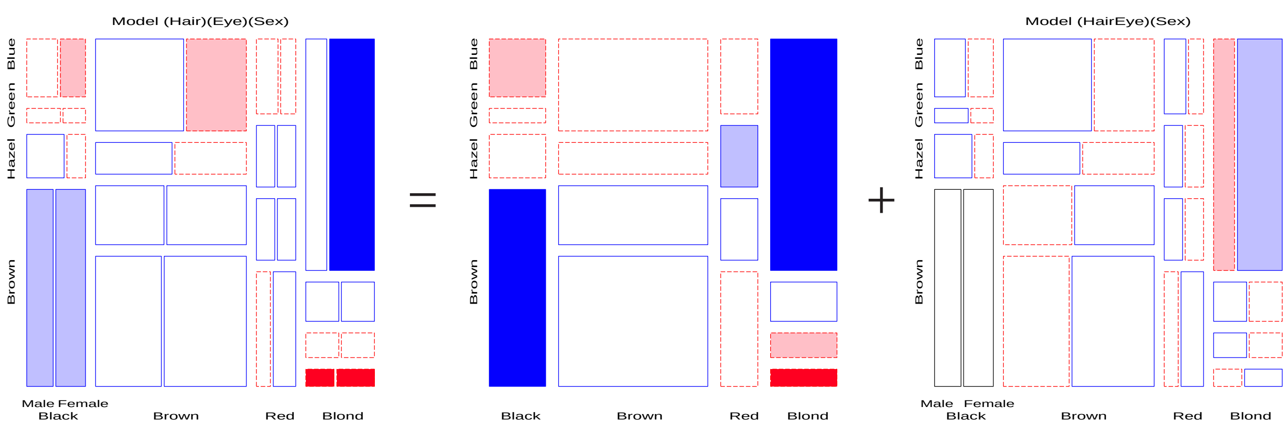
Figure 6: Three-way mosaic, Joint independence.

Sequential plots & models

- Sequential construction, for given variable ordering

$$p_{ijk\ell\cdots} = \underbrace{\overbrace{p_i \times p_{j|i}}^{\{AB\}} \times p_{k|ij}}_{\{ABC\}} \times p_{\ell|ijk} \times \cdots \quad (1)$$

- Fit a model to each sequential marginal subtable: $\{A\}, \{AB\}, \{ABC\}, \ldots$.
- Sequential models of joint independence partition the G^2 for mutual independence: $G^2_{[A][B][C][D]} = G^2_{[A][B]} + G^2_{[AB][C]} + G^2_{[ABC][D]}$



Model	df	G^2
$[H][E]$	9	146.44
$[H,E][S]$	15	19.86
$[H][E][S]$	24	155.20

Visualization principles

- **Nested multiples**
  - Each mosaic shows its own marginal subtables (spacing $\rightarrow$ visual grouping)
  - Shows DATA + RESIDUALS
- **Association ordering** - sort the display by the effects to be observed
- **Visual impact** - distinguish what should stand out (patterns of residuals)
- **Decomposition** - show partitions of model fit in coherent ways

Example: Survival on the *Titanic*

Data from Dawson (1995) on the breakdown of 2201 passengers and crew:

Table 1: Survival on the Titanic

			Class			
Survived	**Age**	**Gender**	1st	2nd	3rd	Crew
No	Adult	Male	118	154	387	670
Yes			4	13	89	3
No	Child		0	0	35	0
Yes			0	0	17	0
No	Adult	Female	57	14	75	192
Yes			140	80	76	20
No	Child		5	11	13	0
Yes			1	13	14	0

Order of variables: Class, Gender, Age, Survival

Class $\times$ Gender:

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

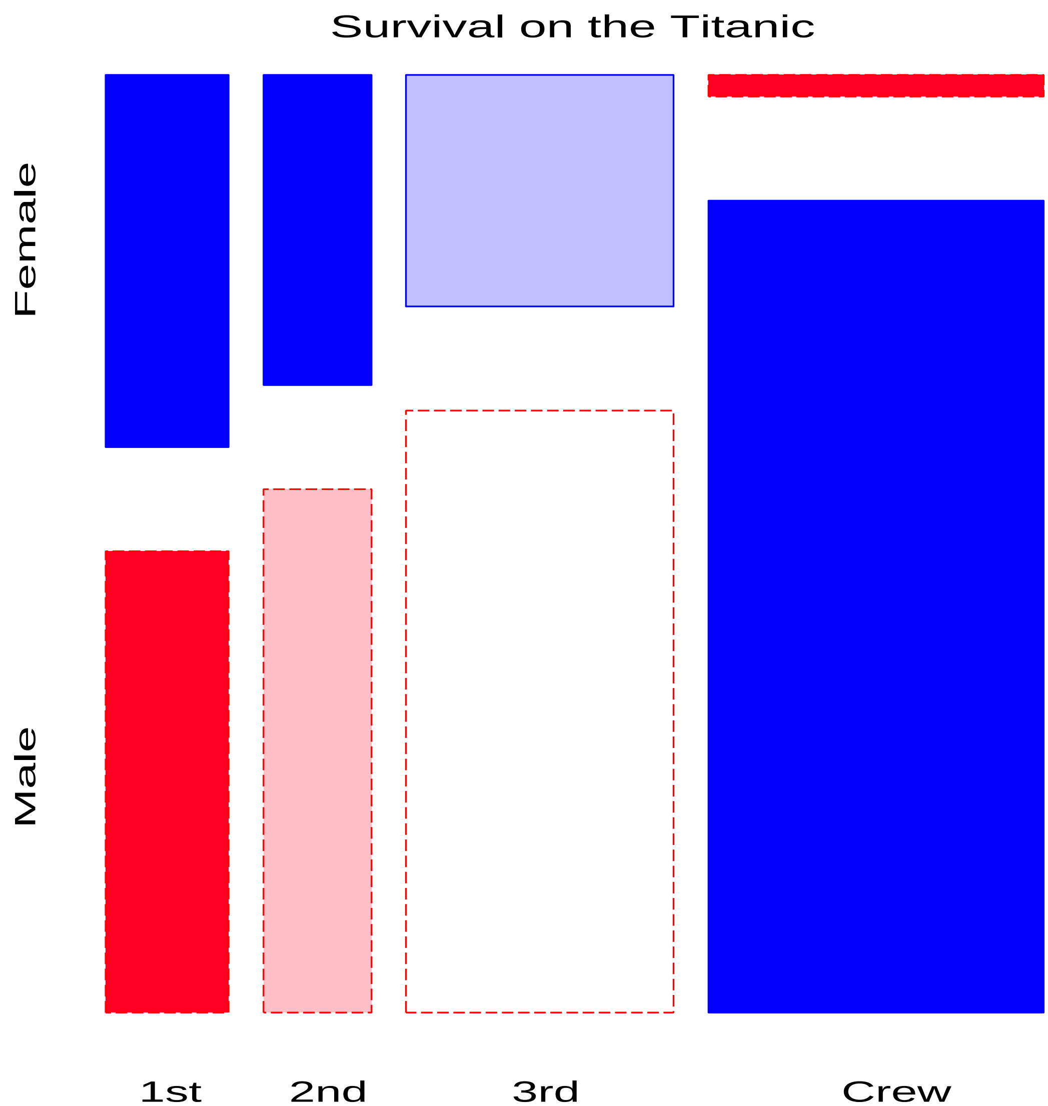


Figure 7: Titanic data: Class and Gender

3 way: $\{$Class, Gender$\} \perp$ 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?)

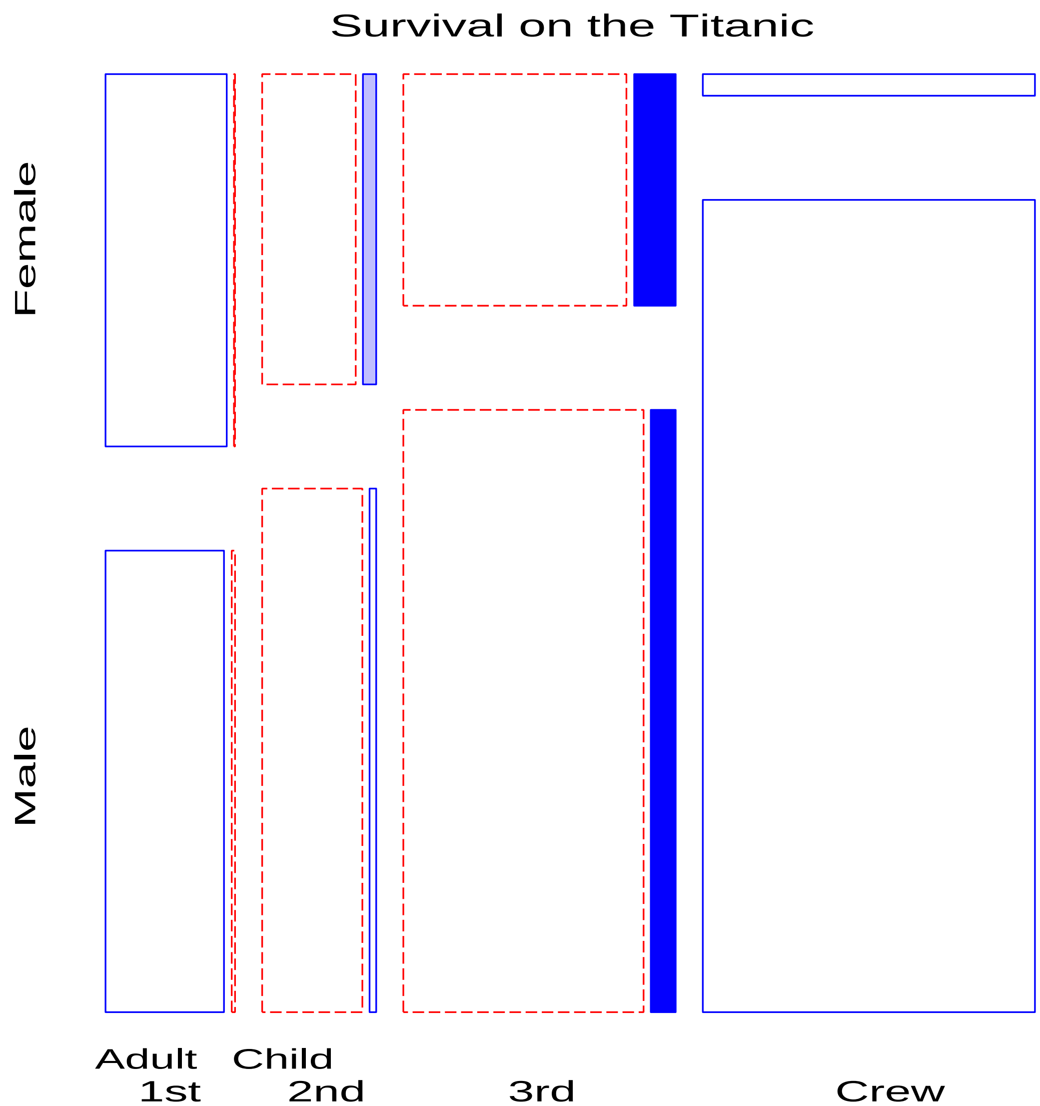


Figure 8: Titanic data: Class, Gender, Age

4 way: {Class, Gender, Age} $\perp$ Survival ?

- 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 SX terms

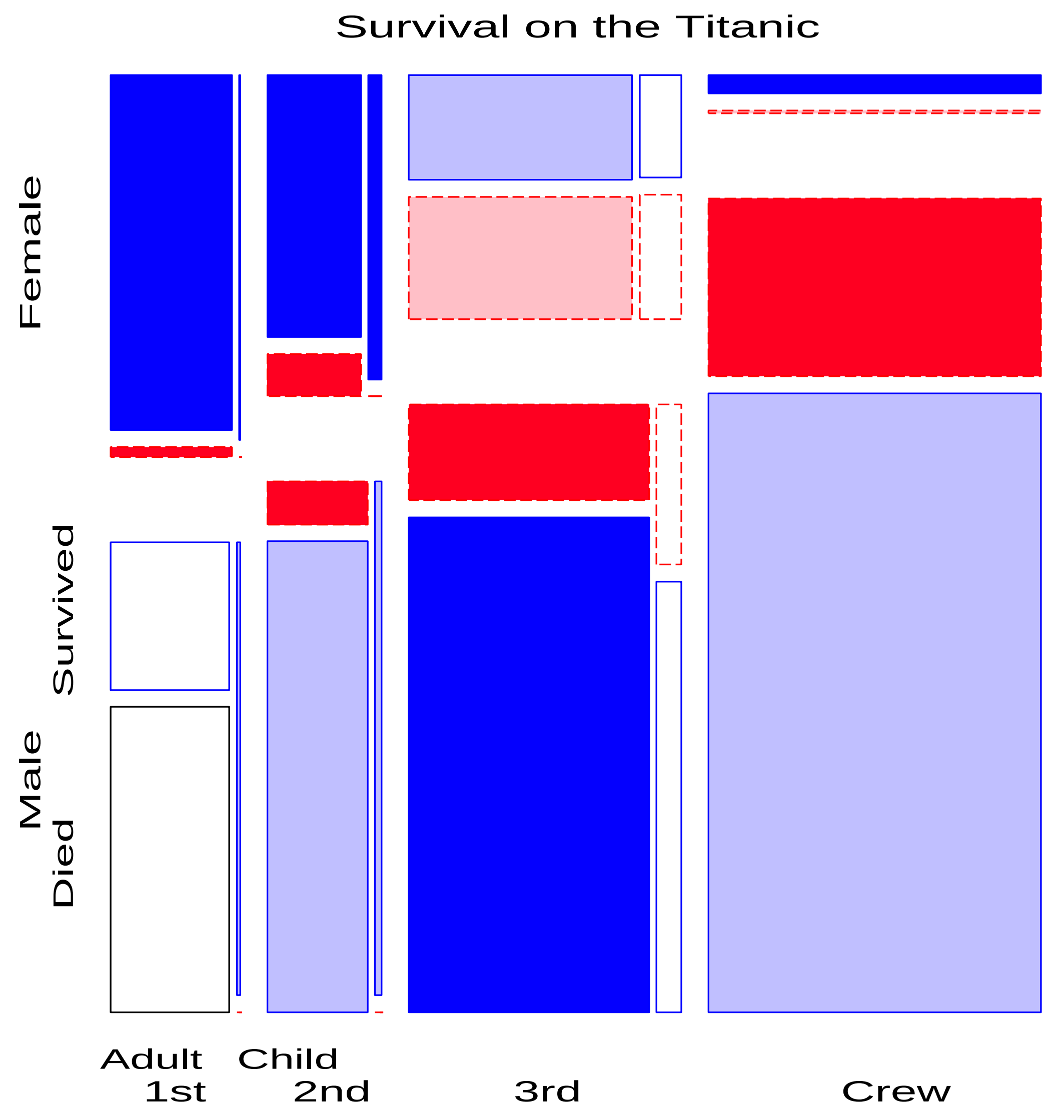


Figure 9: Class, Gender, Age, and Survival, Joint independence

Main effects of Class, Gender and Age on Survival:
$[CGA][CS][GS][AS]$

- Fit is much improved ($\Delta G^2(5) = 559.4$), but not good ($G^2(10) = 112.56$).
- $\Rightarrow$ Interactions among Class, Gender and Age on Survival.

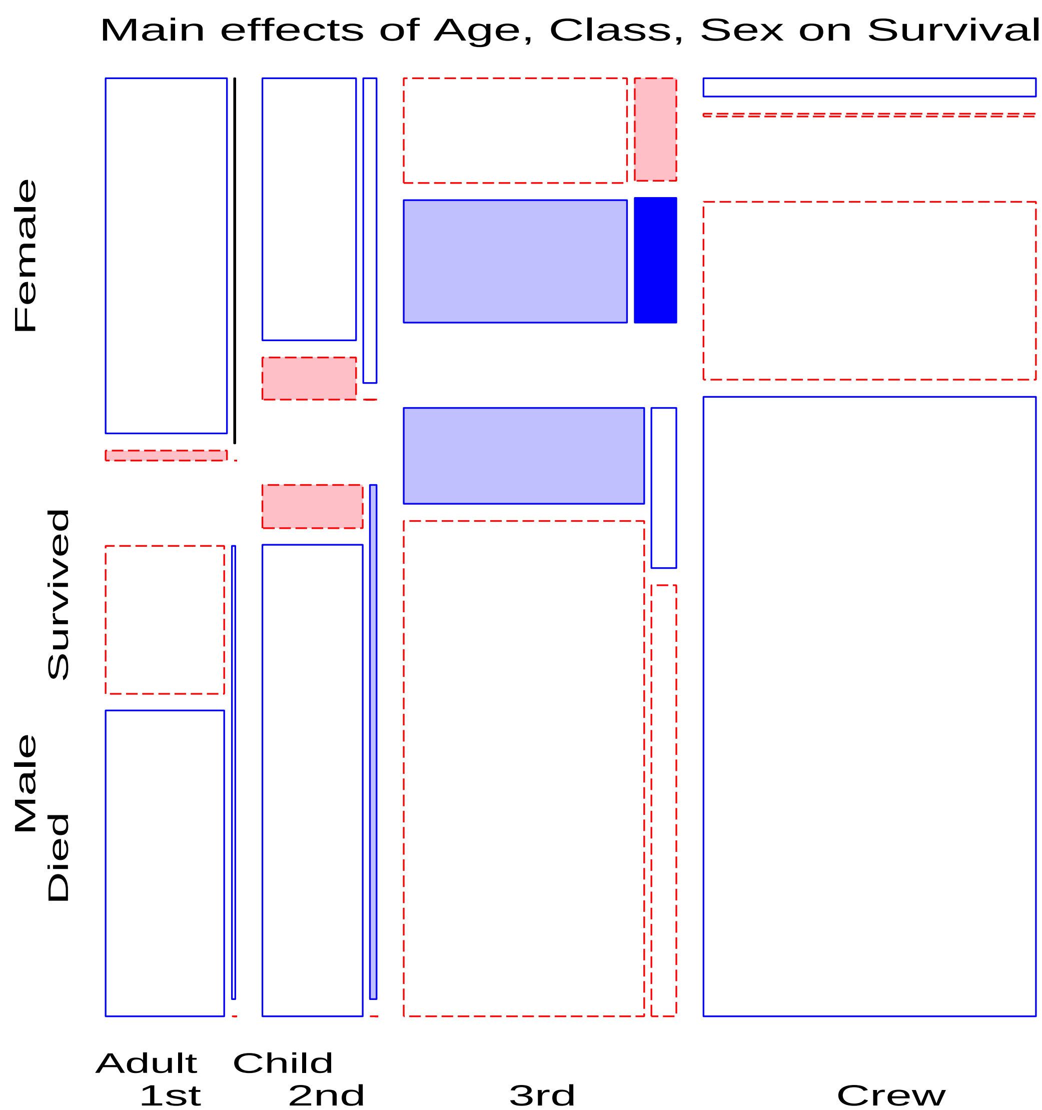


Figure 10: Main effects of Age, Gender and Class on Survival

"women and children first" $\longrightarrow$ model $[CGA][CS][\mathbf{GAS}]$

- Model improved slightly, but still not good ($G^2(9) = 94.54$).

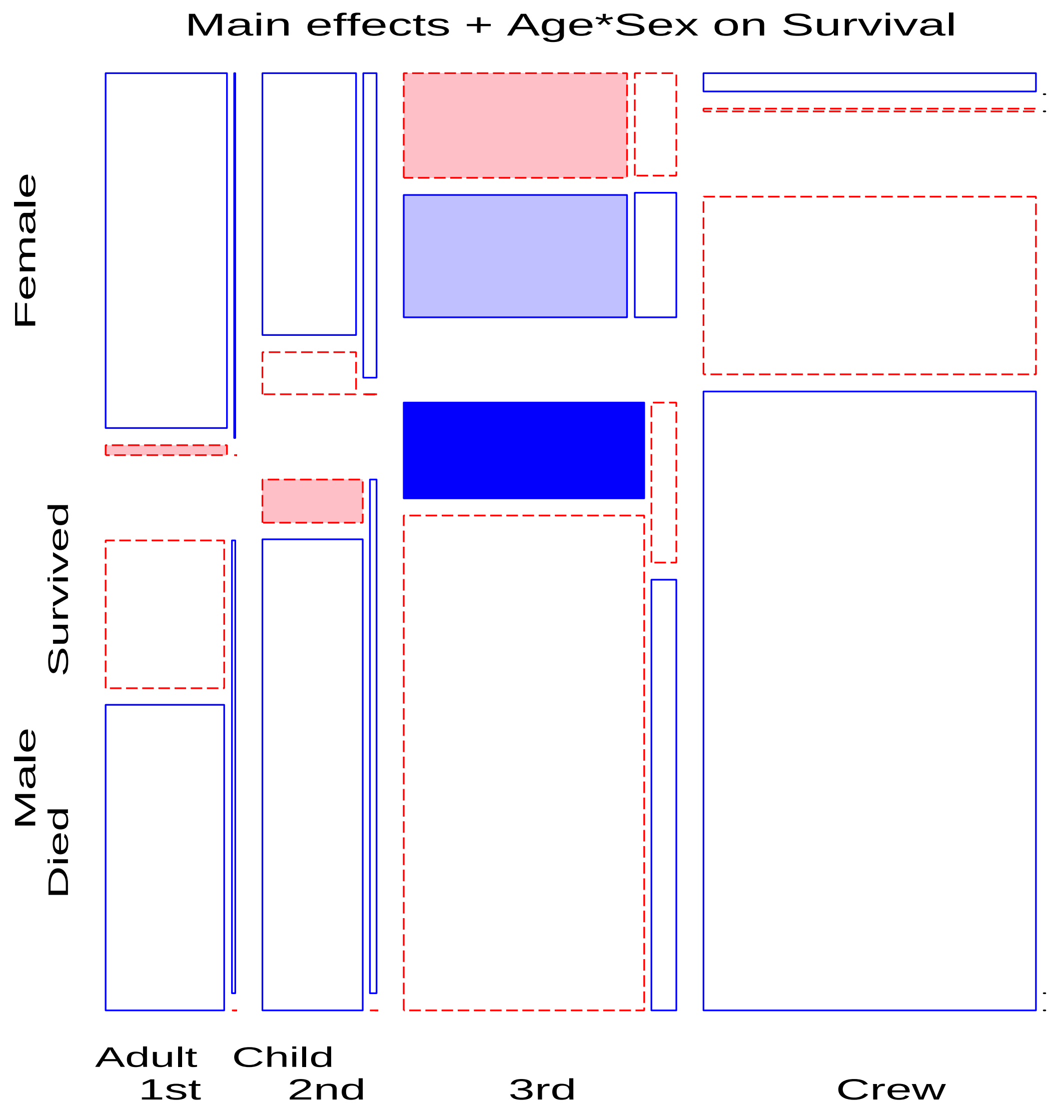


Figure 11: Main effects + Age*Gender on Survival

Class interacts with Age and Gender: $[CGA][\mathbf{CGS}][\mathbf{CAS}]$

- $G^2(4)$ now 1.69, a very good fit (too good?).

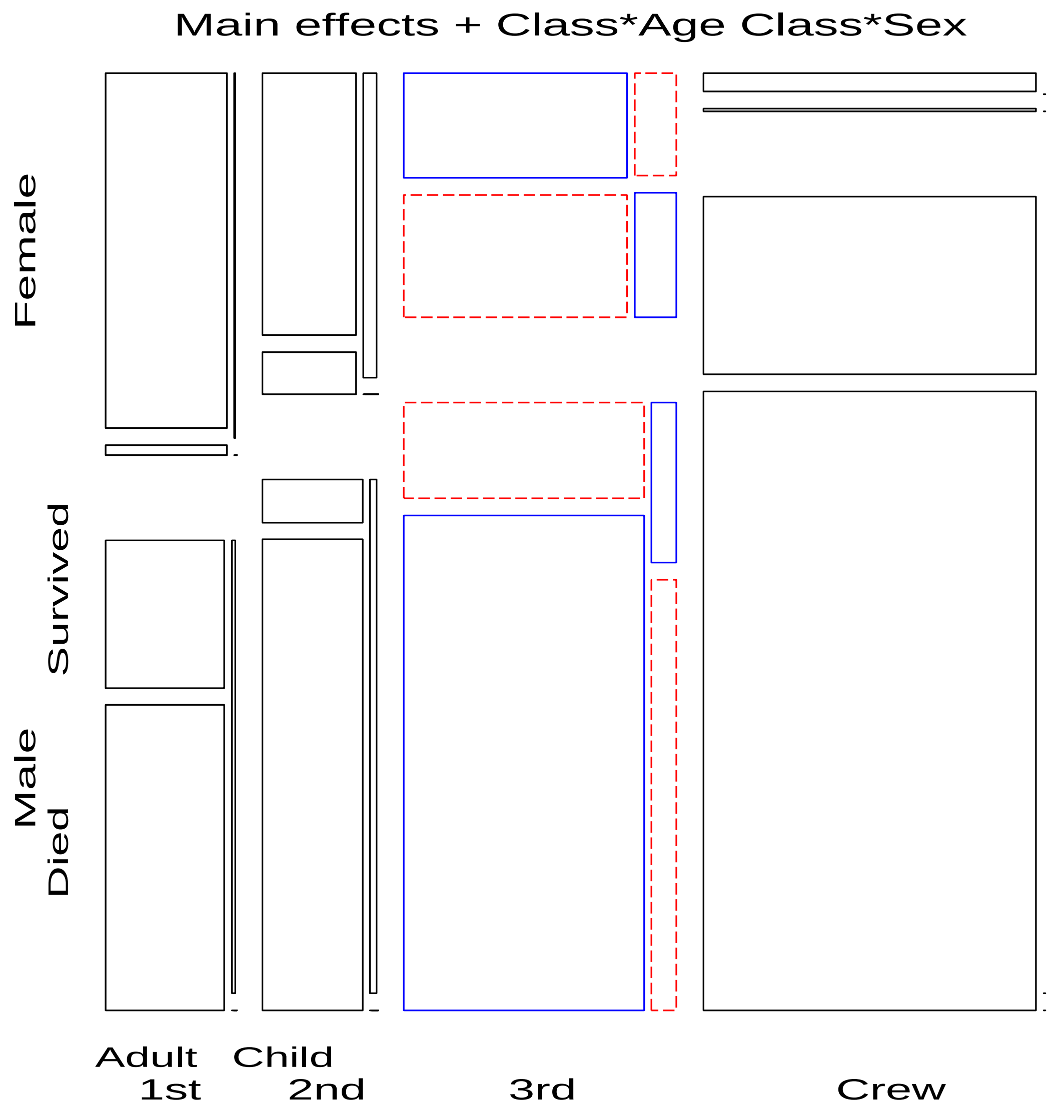


Figure 12: Main effects + Age*Gender + Class*Gender

Titanic Conclusions

- 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 $\times$ Gender: women in 3rd class did not have a significant advantage, while men in 1st class did (compared to men in other classes).
- Class $\times$ Age: no children in 1st or 2nd class died, but nearly two-thirds in 3rd class died;
- For adults, mortality $\uparrow$ as economic class $\downarrow$.
- Summary statement: “women and children (according to class), then 1st class men”.

Mosaic matrices

Quantitative data: scatterplot matrix shows $p \times (p-1)$ marginal views in a coherent display;

- Each scatterplot a projection of data
- Detect patterns not easily seen in separate graphs.
- Only shows bivariate relations.

Categorical data: Mosaic matrix shows all $p \times (p-1)$ marginal views

- Each mosaic shows bivariate relation
- Fit: marginal independence
- Direct visualization of the "Burt" matrix analyzed in MCA to account for all pairwise associations among p variables

$$B = Z^{\mathsf{T}} \operatorname{diag}(n) Z = \begin{bmatrix} N_{[1]} & N_{[12]} & \cdots \\ N_{[21]} & N_{[2]} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

where $N_{[i]} =$ diagonal matrix of one-way margin; $N_{[ij]} =$ two-way margin for variables i and j,

Example: Survival on the *Titanic*

- Strong associations of Class, Gender, Age with Survival.
- Each pair shown twice, splitting by the column variable first.

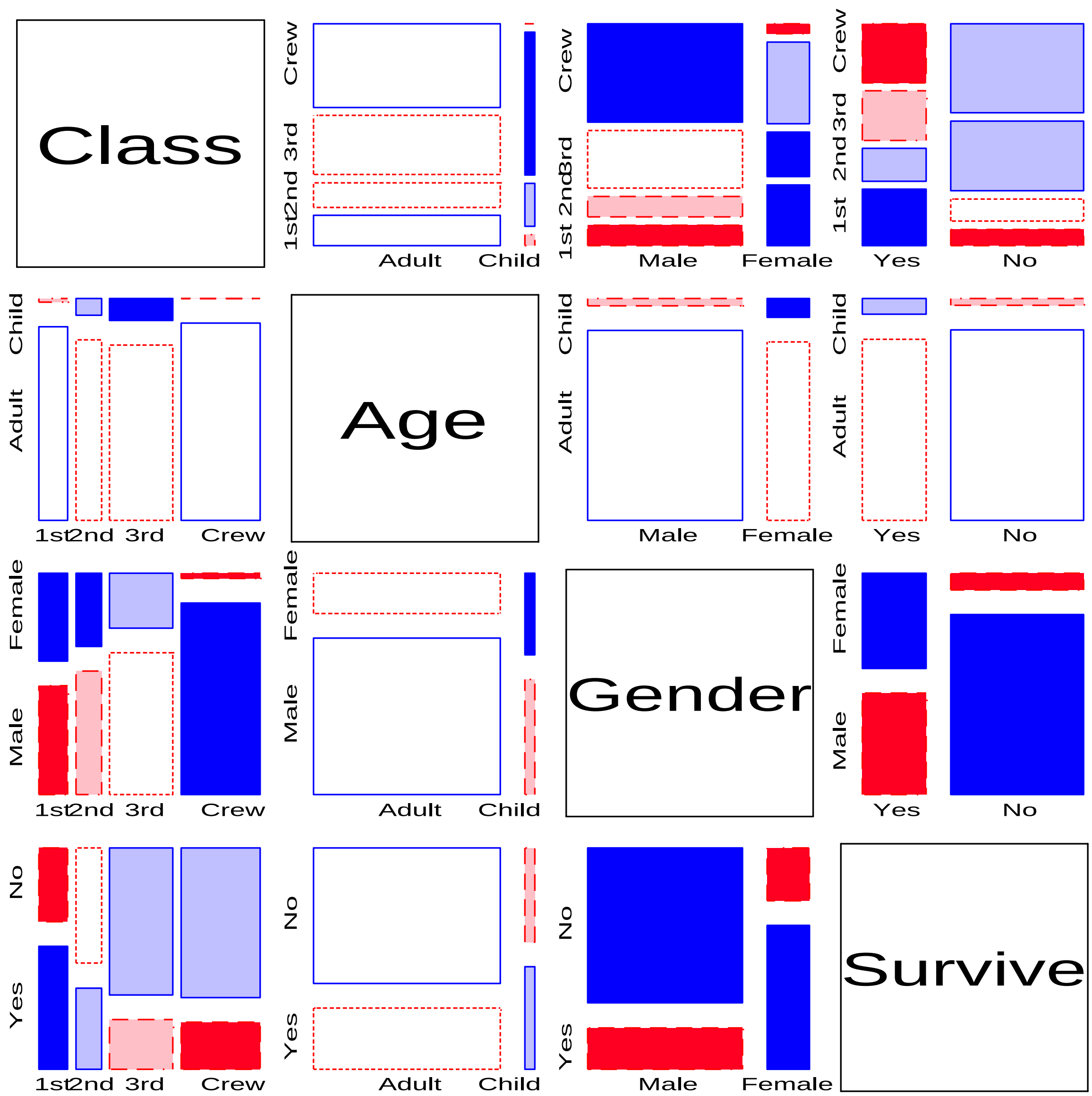
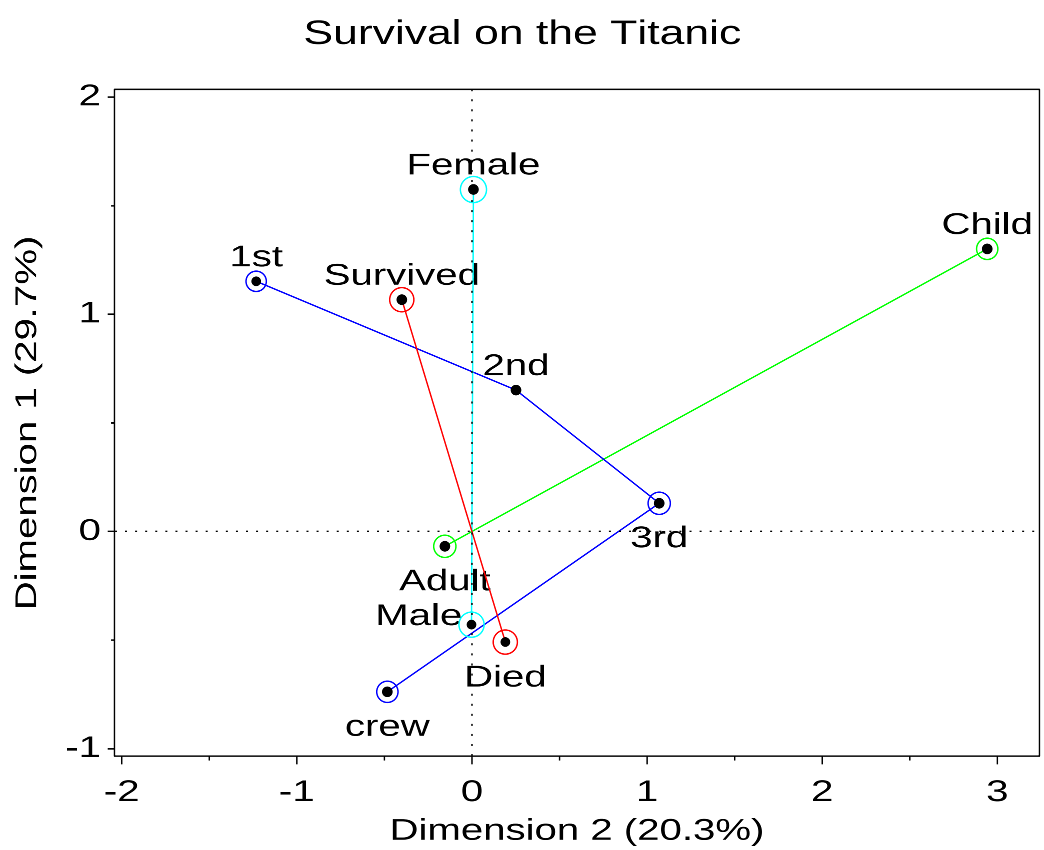


Figure 13: Mosaic matrix of *Titanic* data.

Titanic: Multiple Correspondence Analysis

- 2D solution: 50% of all pairwise assoc (3D = 67%)
- Dim1: Gender, Survival; Dim2: Class, Age
- Binary factors: Distance from origin $\sim p_i^{-1}$
- Mosaic matrix: 100%, makes *form* of association explicit



Example: Berkeley admissions

- Admission, Gender: overall, more males admitted
- Dept A, B: highest admission rate; E, F lowest
- Males apply most to A, B, women more to C–F.

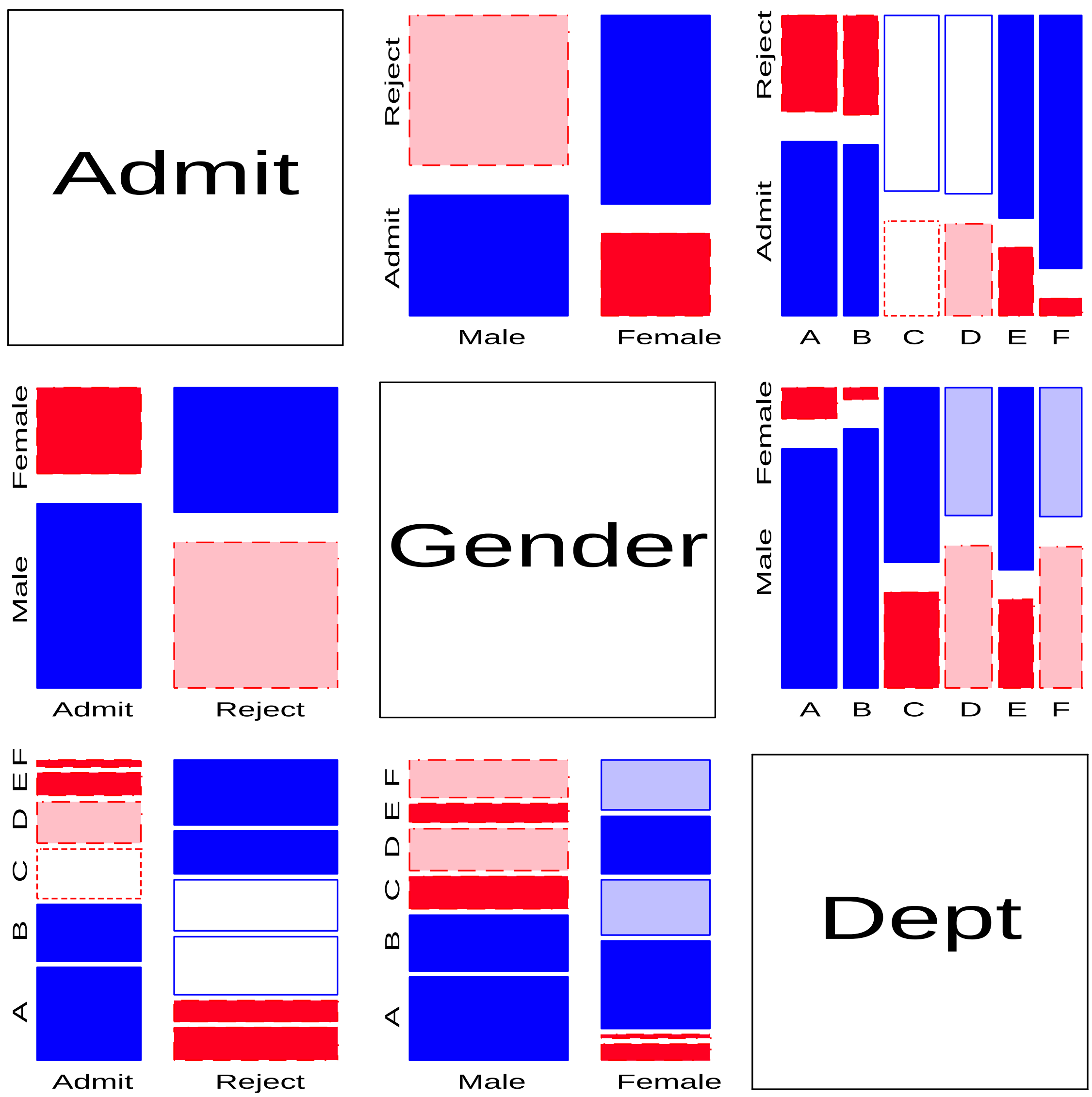


Figure 14: Mosaic matrix of Berkeley admissions.

Conditional plots for quantitative data

Iris data — scatterplot matrix

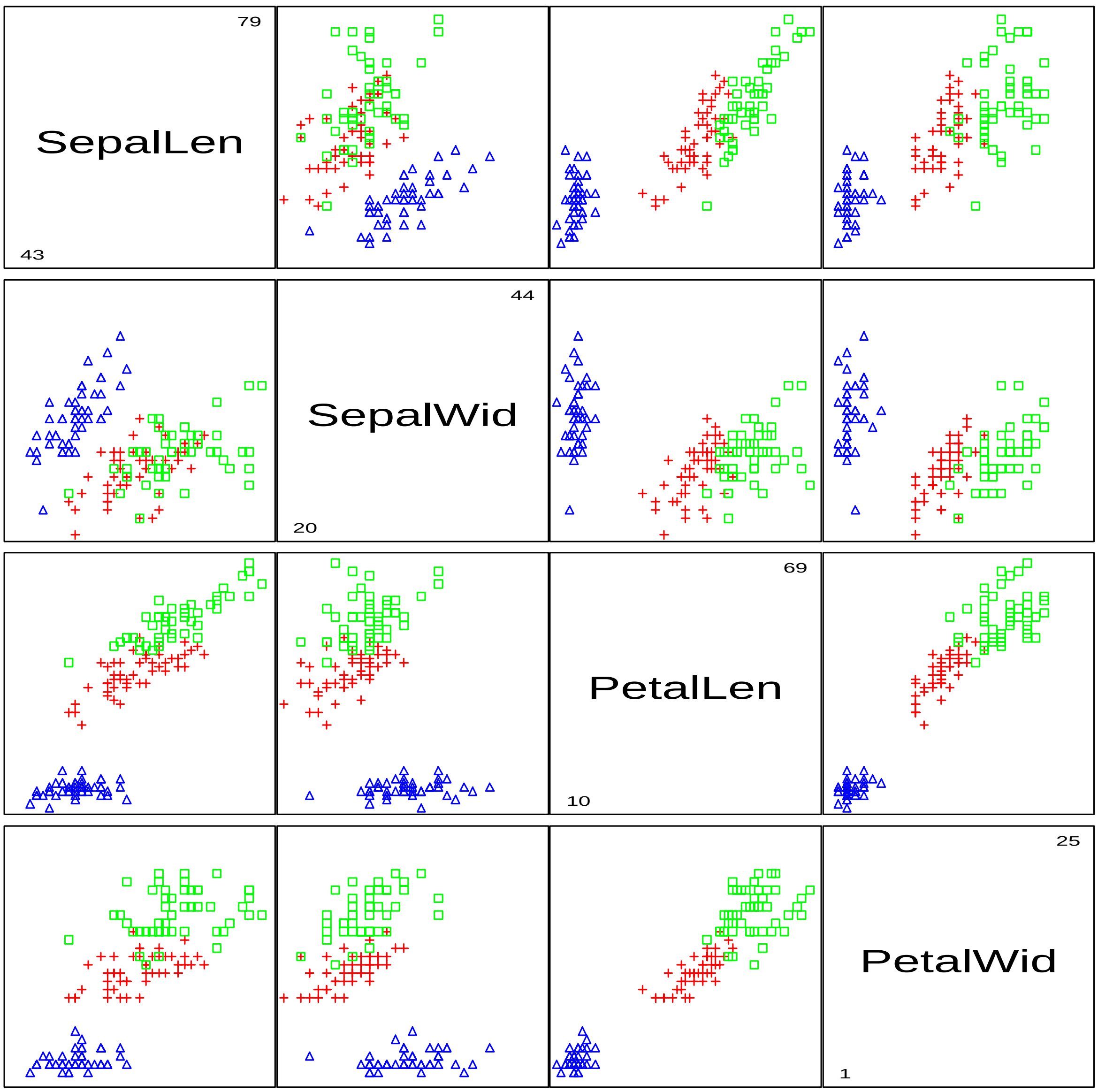


Figure 15: Scatterplot matrix for Iris data

Conditional plots for quantitative data

Iris data — conditional scatterplot matrix

- Plot $\widetilde{X_i} = X_i - \widehat{X_i}|\text{others}$ vs. $\widetilde{X_j} = X_j - \widehat{X_j}|\text{others} \quad \forall i, j$
- Removes species effect (correlated means)

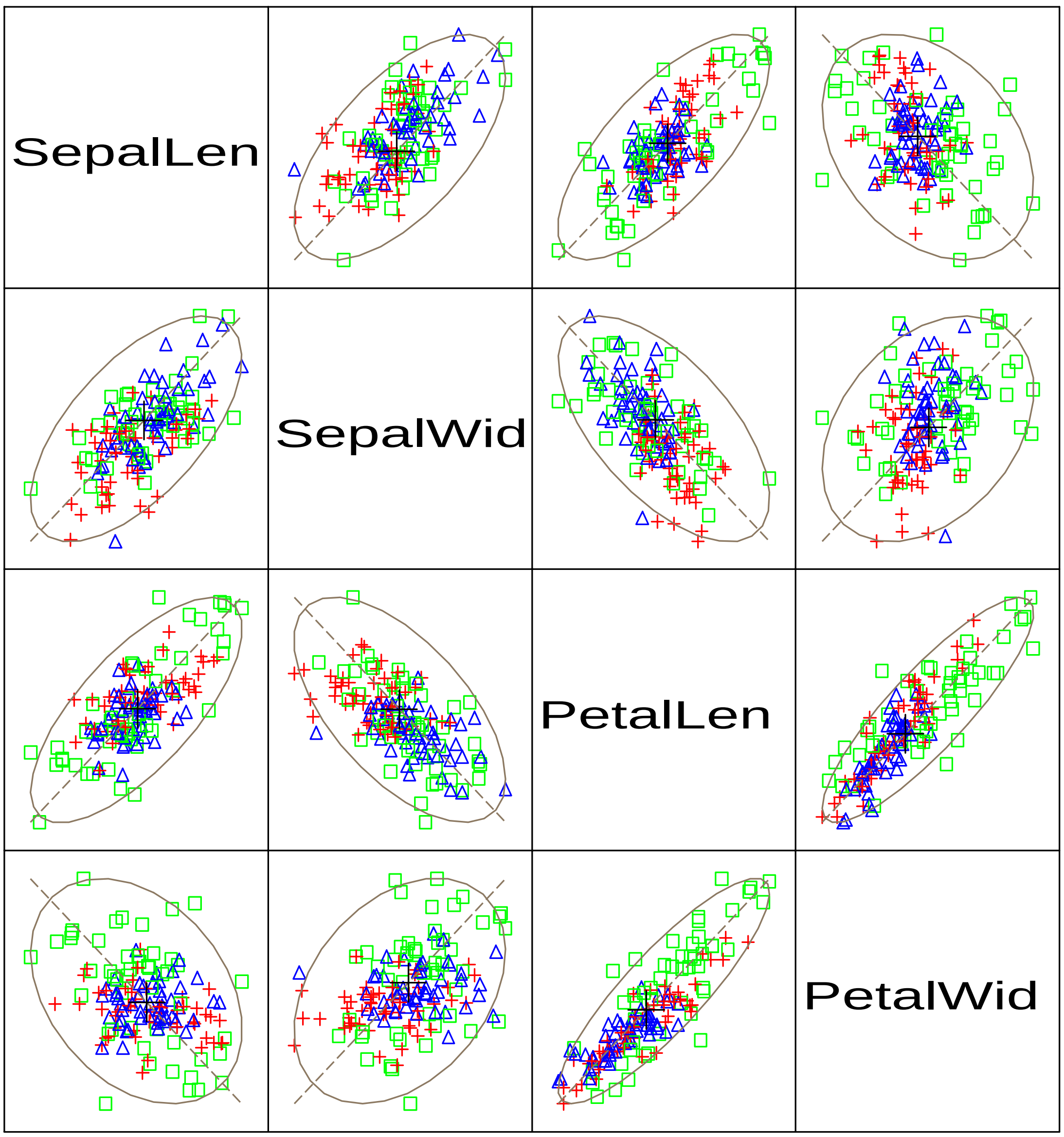


Figure 16: Conditional scatterplot matrix for Iris data

Conditional plots for quantitative data

$$\begin{aligned}\rho_{ij|\text{ others}} = 0 \quad &\Longleftrightarrow \quad \sigma^{ij} = 0 \\ &\Longleftrightarrow \quad X_i \perp X_j \mid \text{others}\end{aligned} \tag{2}$$

- Zero partial correlation plays same role for quantitative variables as two-way terms in graphical log-linear models.
- Conditional scatterplot matrix provides a visualization of the conditional independence relations.
- When Y is a response, panels in the row for Y are just the partial regression (added variable) plots. Other rows treat each variable in turn as a response.

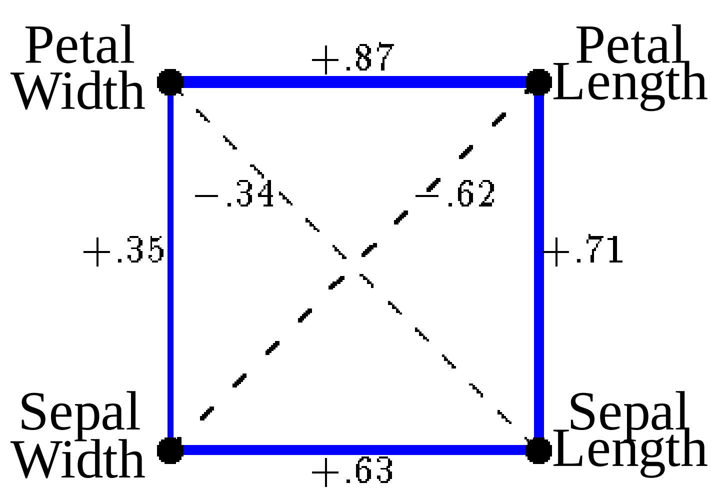


Figure 17: Independence graph for Iris Data

“Mixed” models: Categorical and Continuous Data

- **Marginal views**
  - X, Y pairs: scatterplot
  - A, B pairs: mosaic
  - X, A pairs: boxplot
- **Conditional views**
  - Fit graphical mixed model: $AB \, / \, / \, XY$ (Edwards, 1995) ?
  - Fit GLMs:

$$\begin{aligned} g(\mu_i) &= x^{\mathsf{T}}_{\text{others}}\beta \\ g(\mu_j) &= x^{\mathsf{T}}_{\text{others}}\beta \end{aligned}$$

    with identity link for X, Y, log link for A, B
  - Plot residuals as in marginal views

"Mixed" models: Categorical and Continuous Data

Iris data — Mixed scatterplot matrix

- Discrete: Species, SepalLen (divided into thirds)
- Continuous: PetalLen, PetalWid

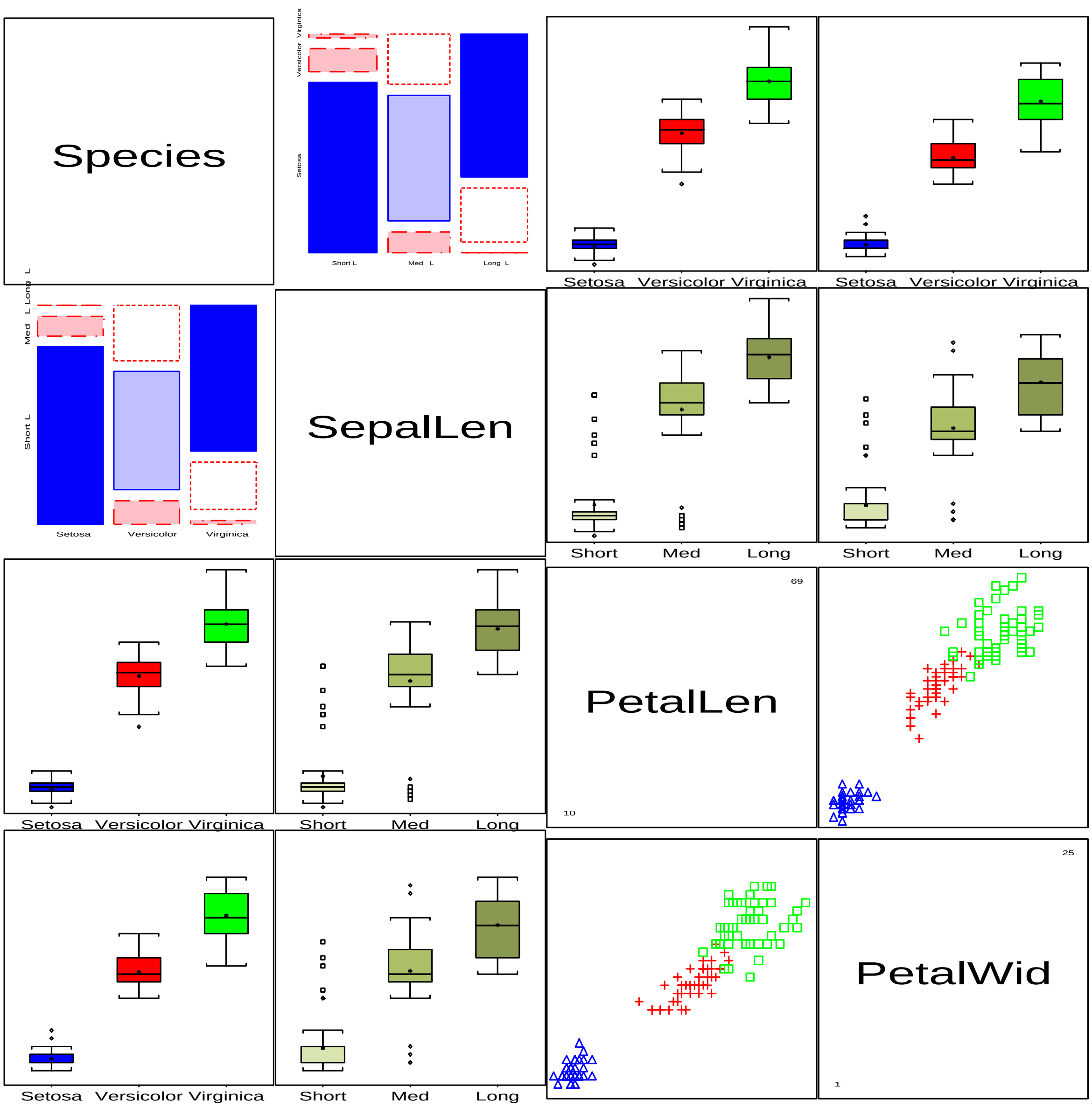


Figure 18: Mixed scatterplot matrix for Iris data

Example: A 5-way table

Heckman & Willis 1977 data:

Table 2: Labour force participation of married women 1967-1971

Employed?			**1968** Yes		No	
			1967			
1969	**1970**	**1971**	Yes	No	Yes	No
Yes	Yes	Yes	426	73	21	54
No			11	9	8	36
Yes	No		16	7	0	6
No			12	5	5	35
Yes	Yes	No	38	11	7	16
No			2	3	3	24
Yes	No		47	17	9	28
No			28	24	43	559

Marginal relations

- All years strongly associated: employment status persists
- Strength of association $\downarrow$ as lag $\uparrow$.

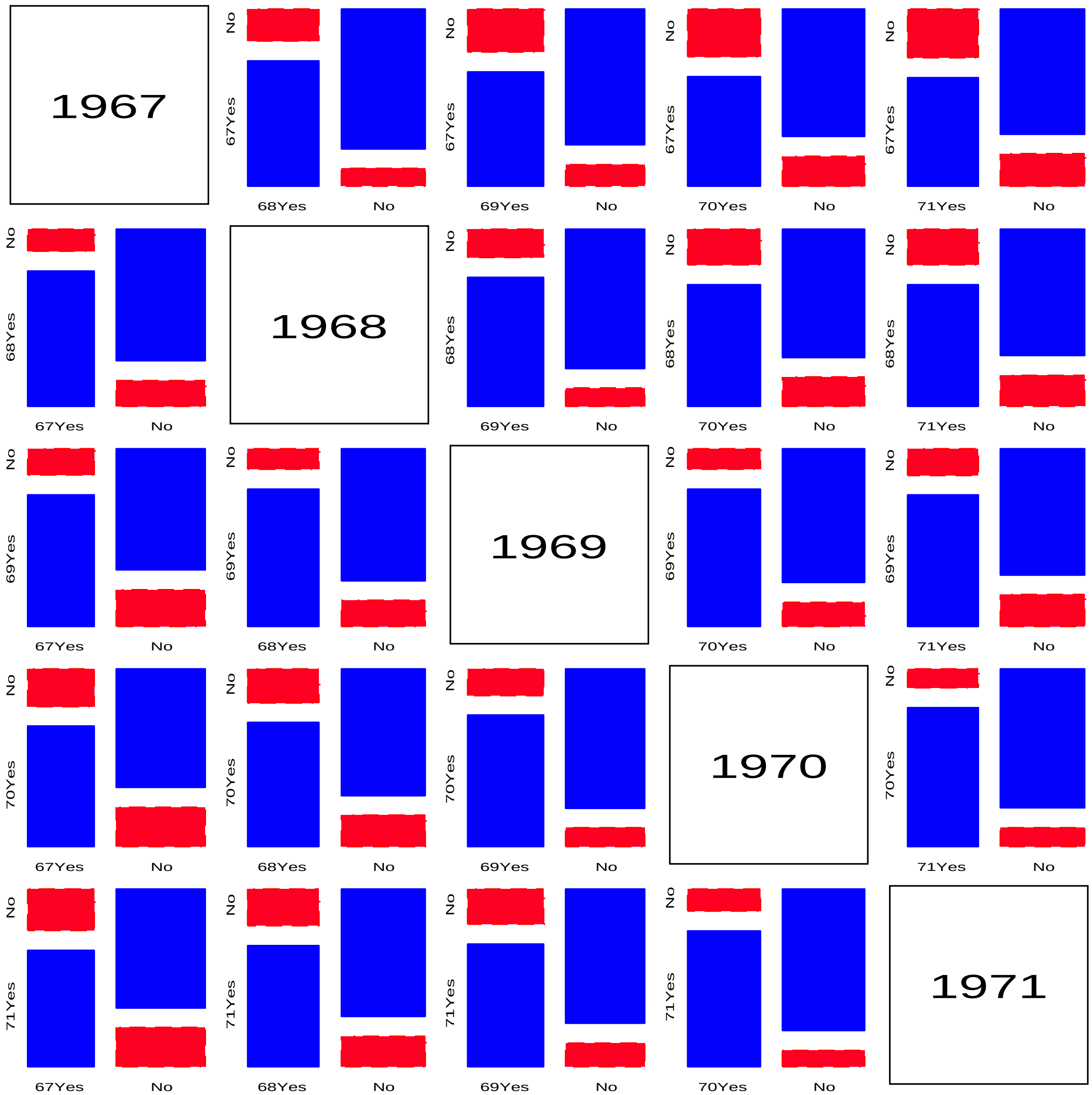


Figure 19: Mosaic matrix for pairwise associations

Conditional relations

- 3-way plots: row $\perp$ other $|$ col ?
- Employment status persists over several years.

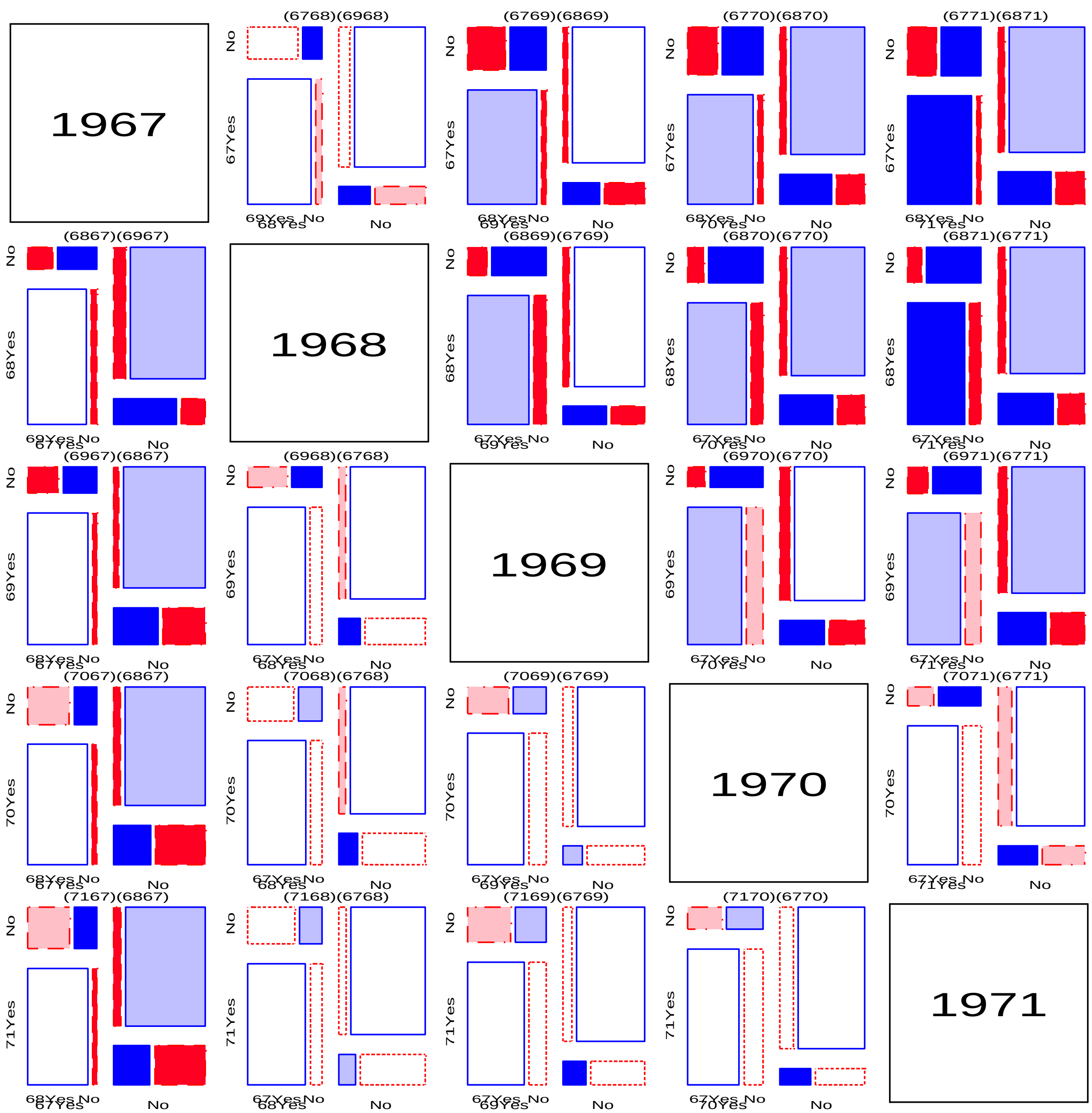


Figure 20: Mosaic matrix for conditional associations

Fitting Markov models

Table 3: Markov chain models fit to Heckman-Willis data

Order	Model	df	G^2	p
M1	[67,68][68,69][69,70][70,71]	22	210.225	0.000
M2	[67,68,69][68,69,70][69,70,71]	16	62.672	0.000
M3	[67,68,69,70][68,69,70,71]	8	9.023	0.340

i.e., $67 \perp 71 \mid \{68, 69, 70\}$ 5-way mosaics:

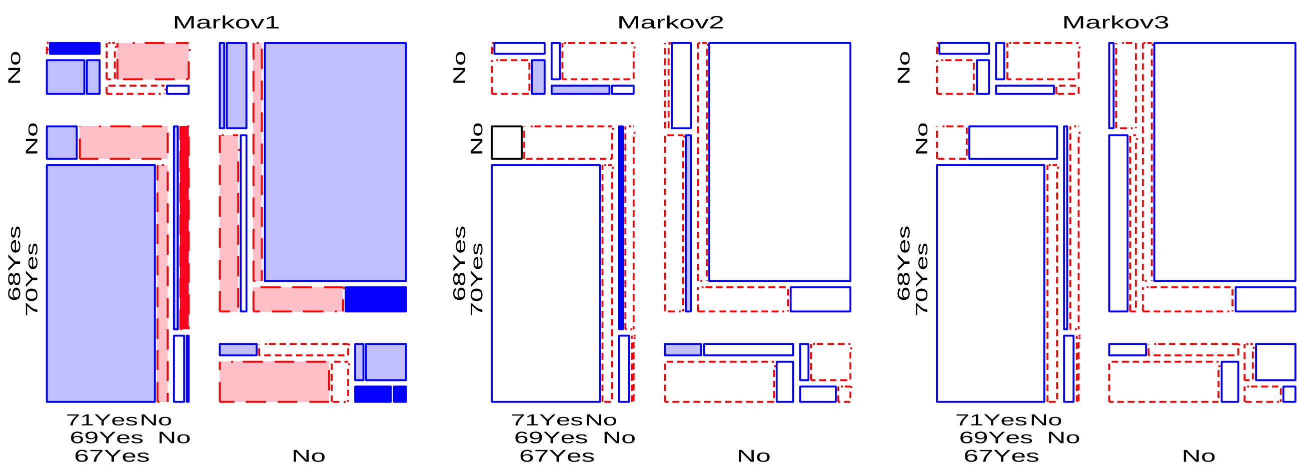


Figure 21: Markov chain models of order 1–3

Coplots for categorical data

- Conditional relations may also be visualized by stratifying the data on the given variables, rather than by partialling out.
- Quantitative variables: coplot display (Cleveland, 1993)
- Categorical variables: array of mosaics, stratified by given variables
- Each panel then shows the *partial* associations among the foreground variables
- the collection of such plots show how these change with the given variables.
- Models of independence fit to the strata separately decompose a model of conditional independence fit to the whole table.

$$G^2_{A \perp B \,|\, C} = \sum_{k}^{K} G^2_{A \perp B \,|\, C(k)} \tag{3}$$

- Collection of mosaic displays for the dependence of A and B for each of the levels of C provides a natural visualization of this decomposition.
- Adjusts automatically for differing marginals across strata—controlled comparison of foreground associations.

Example: Berkeley admissions

Admit $\perp$ Dept $|$ Gender ?

- Strong association between Admission and Department—different rates of admission,
- *Pattern* of association is qualitatively similar for both men and women
- association is quantitatively stronger for men than women—larger differences in admission rates across departments.

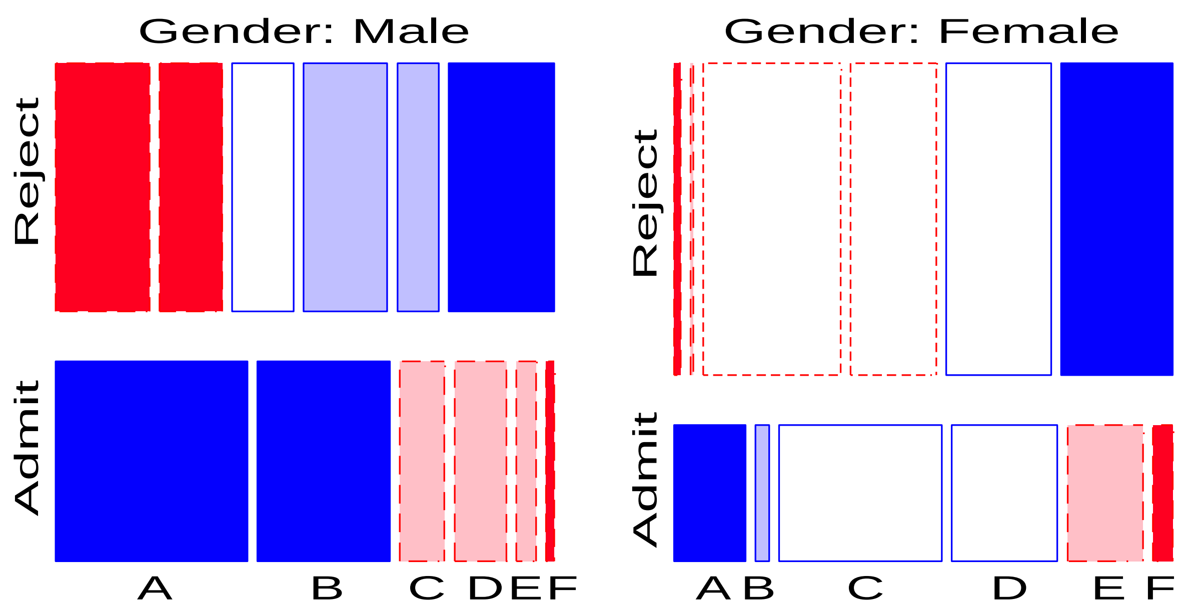


Figure 22: Mosaic coplot of Berkeley admissions, given Gender.

Example: Berkeley admissions

Admit $\perp$ Gender $|$ Dept ?

- No association, except in Dept. A, where females *more* likely to gain admission
- Changes in % admitted, and % female may also be seen.

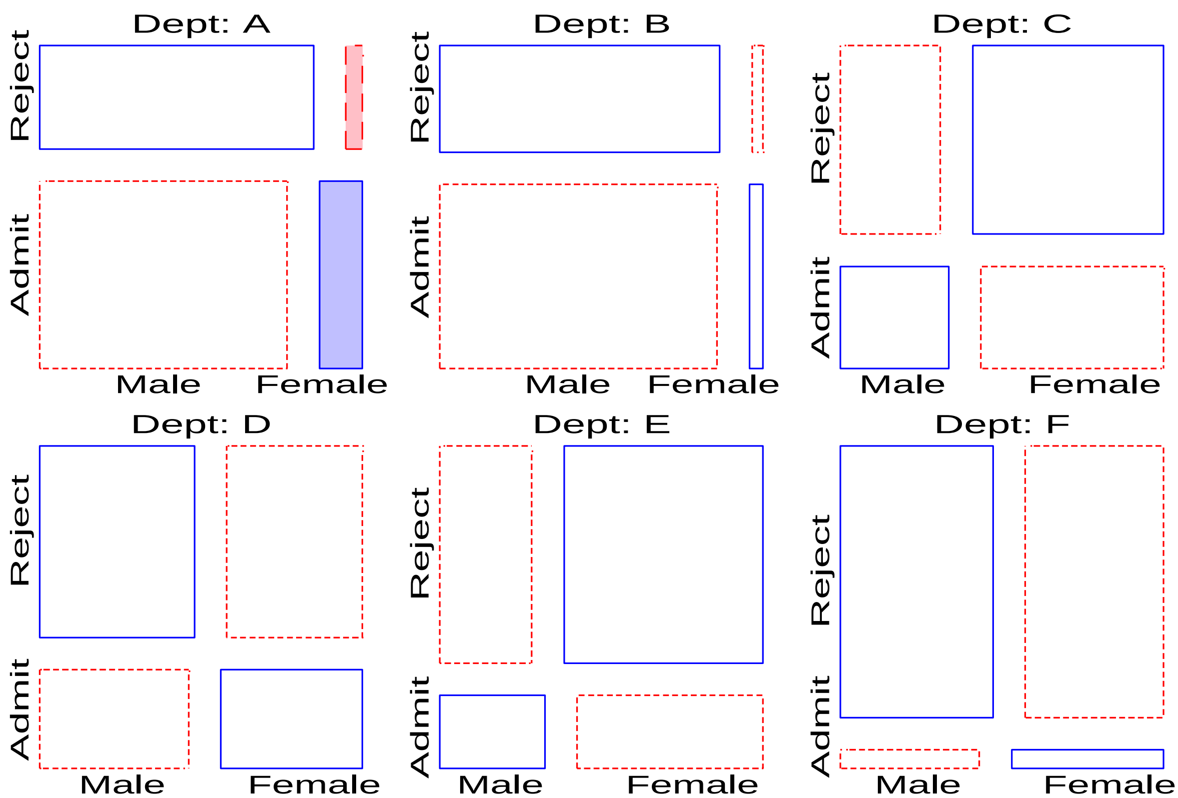


Figure 23: Mosaic coplot of Berkeley admissions, given Department.

Breakdown of G^2 for model Admit $\perp$ Gender $\mid$ Dept:

Table 4: Partial tests of independence of Gender and Admission, by Department

Dept	df	G^2	p
A	1	19.054	0.000
B	1	0.259	0.611
C	1	0.751	0.386
D	1	0.298	0.585
E	1	0.990	0.320
F	1	0.384	0.536
Total	6	21.735	0.001

Effect Ordering for Data Displays

- Where data values are labelled by factors, the ordering of levels has considerable impact on graphical displays.
- With unordered factors, sort the data by effects to be observed.
- Sorting brings similar items together, making them easier to compare.

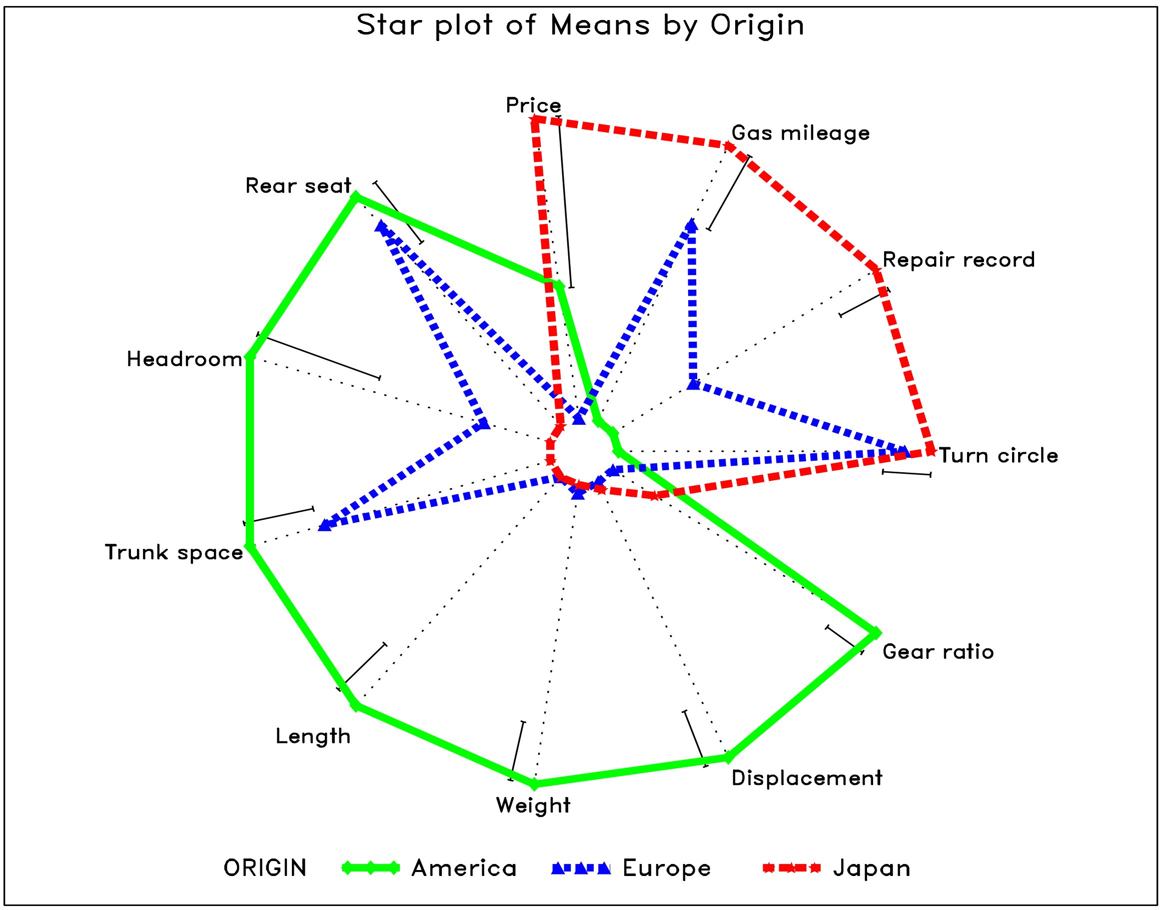
➡ For quantitative data, sort boxplots, dotplots and tables by means, medians, or row and column effects ("***main effects ordering***")

➡ Multivariate glyph plots, stars, faces, parallel coordinates plots - order variables by PCA / biplot dimensions ("***correlation ordering***")

➡ Multivariate plots of means - order variables by canonical discriminant dimensions ("***discriminant ordering***").

Star plot of Means for MANOVA

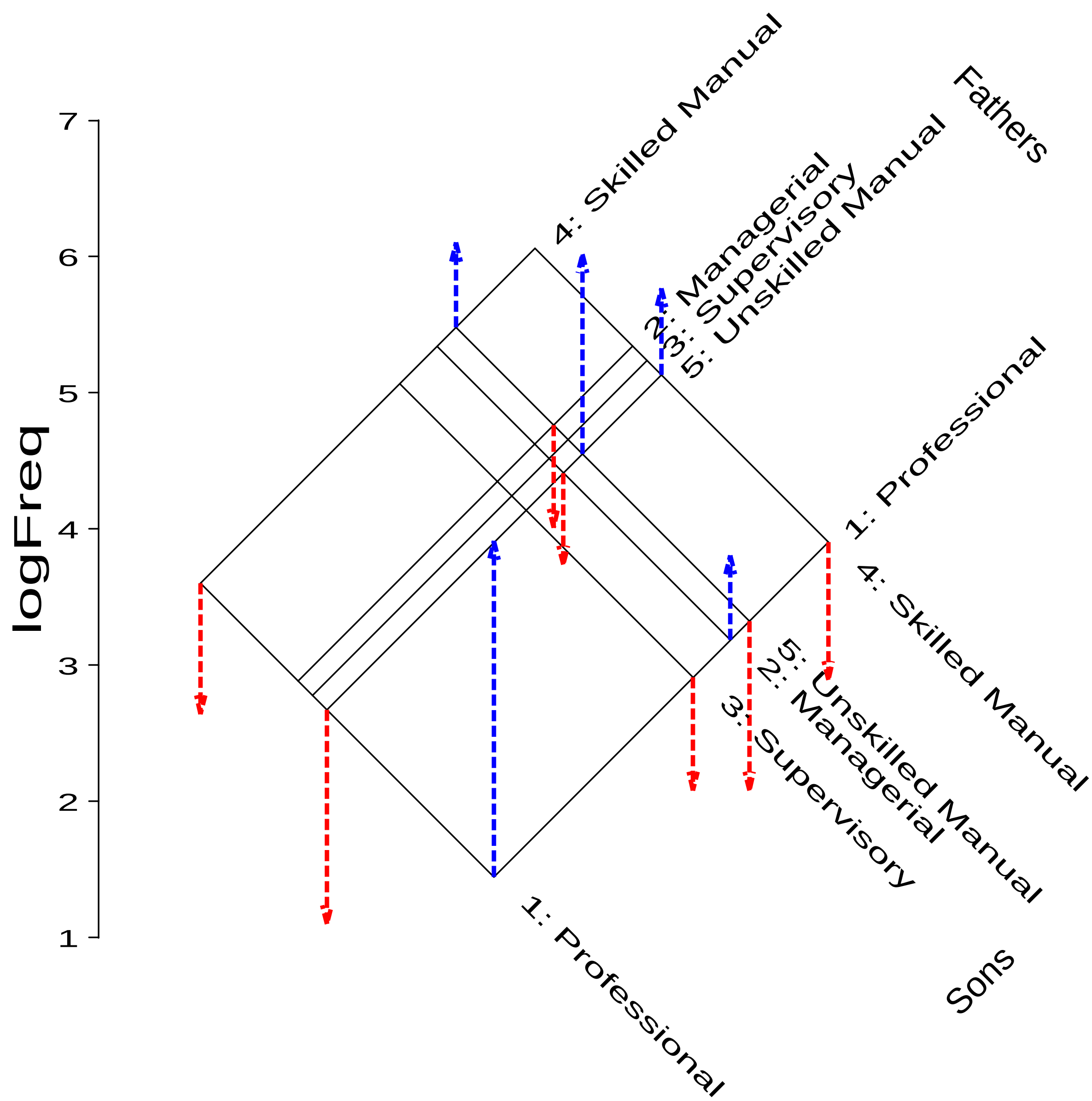
- Display means for 2 or more groups on *m* measures
- **Error bars** display Least Significant Difference
- **Effect ordering:** variables ordered by discriminant dim1



Effect Ordering for Categorical Data Displays

- Two-way display of log(Freq) shows the **local** pattern of association
- The ordering of rows and columns by marginal mean log(F) conceals the **global** structure.

E.g., British Social Mobility: Occupations of Fathers and Sons (Glass, 1954)



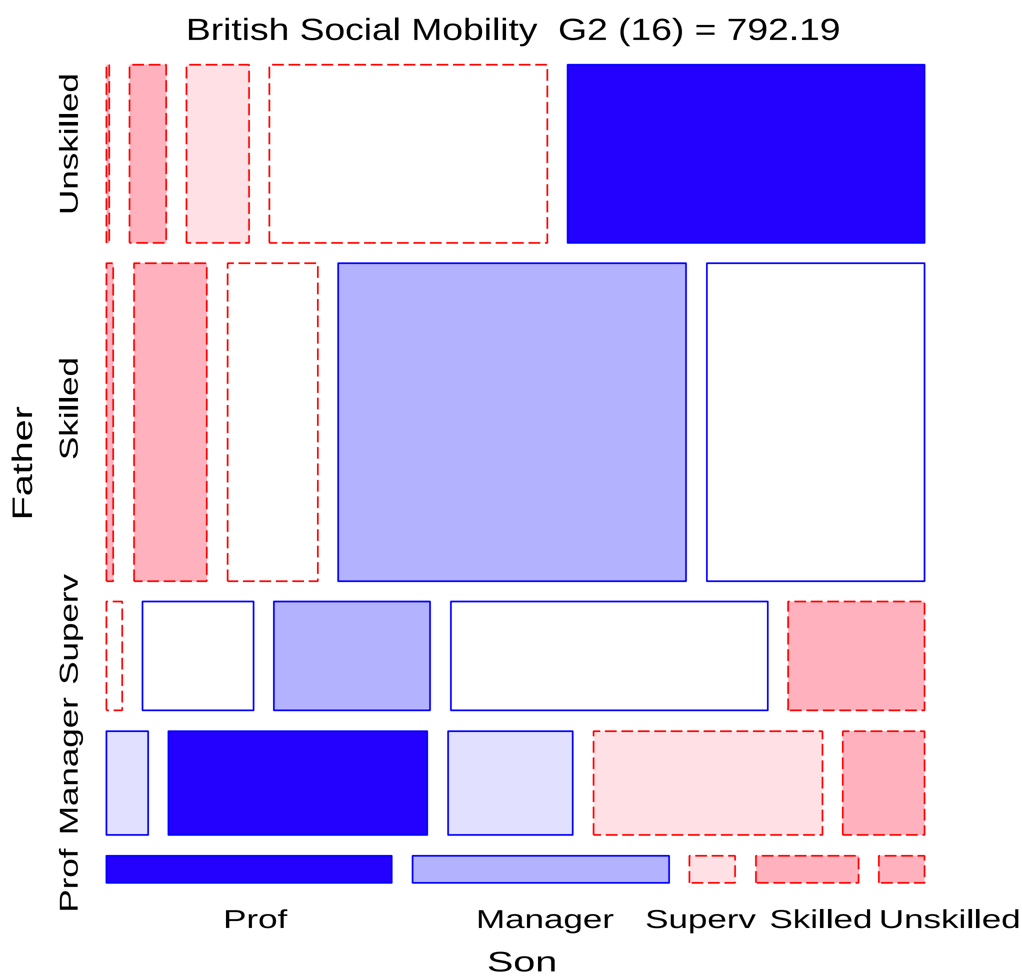
Effect Ordering for Categorical Data Displays

- Mosaic display orders rows and columns by largest Correspondence Analysis dimension.

Residuals, ordered by Row and Column Scores on CA Dimension 1

	Prof	Manager	Superv	Skilled	Unskilled	RowDim1
Prof	23.32	6.35	-2.17	-4.78	-4.82	2.09
Manager	3.36	12.61	2.37	-3.38	-7.41	0.54
Superv	-1.18	0.66	5.10	0.79	-4.44	0.05
Skilled	-4.69	-4.20	-0.93	3.93	0.39	-0.17
Unskilled	-4.48	-7.03	-3.72	-1.41	10.49	-0.36
ColDim1	2.22	0.62	0.04	-0.15	-0.34	

- Residuals from independence are displayed in the context of this global structure.



Mosaic matrices: Structure of Log-linear Models

- Show relations among variables in log-linear models (Theus and Lauer, 1998).
- Display *expected* frequencies under a given model
- E.g., $[A]\,[B]\,[C] \longrightarrow$ all pairs marginally *and* conditionally independent

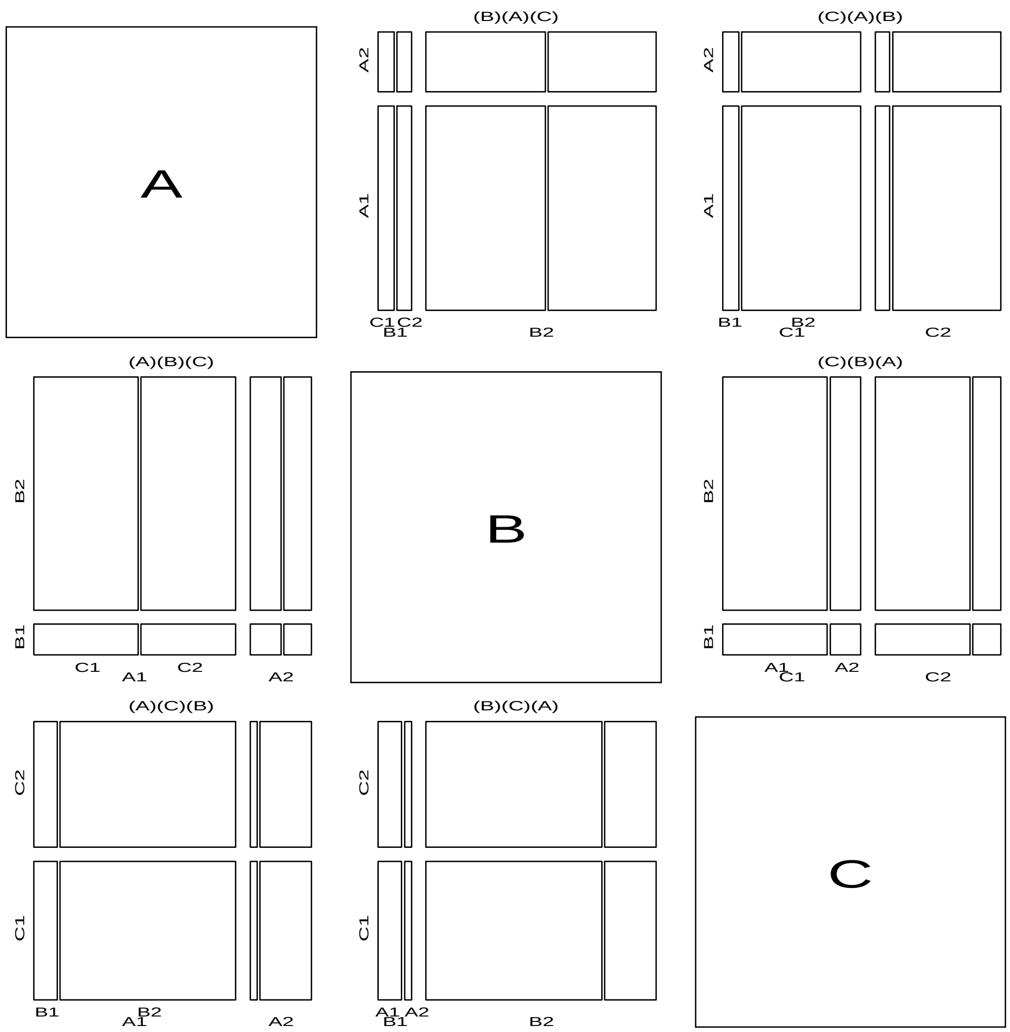


Figure 24: Mosaic matrix for mutual independence.

Joint Independence

- $[A\,B][C] \longrightarrow \{A, B\} \perp C$ and also $A \perp B \mid C$, but $A \not\perp B$.

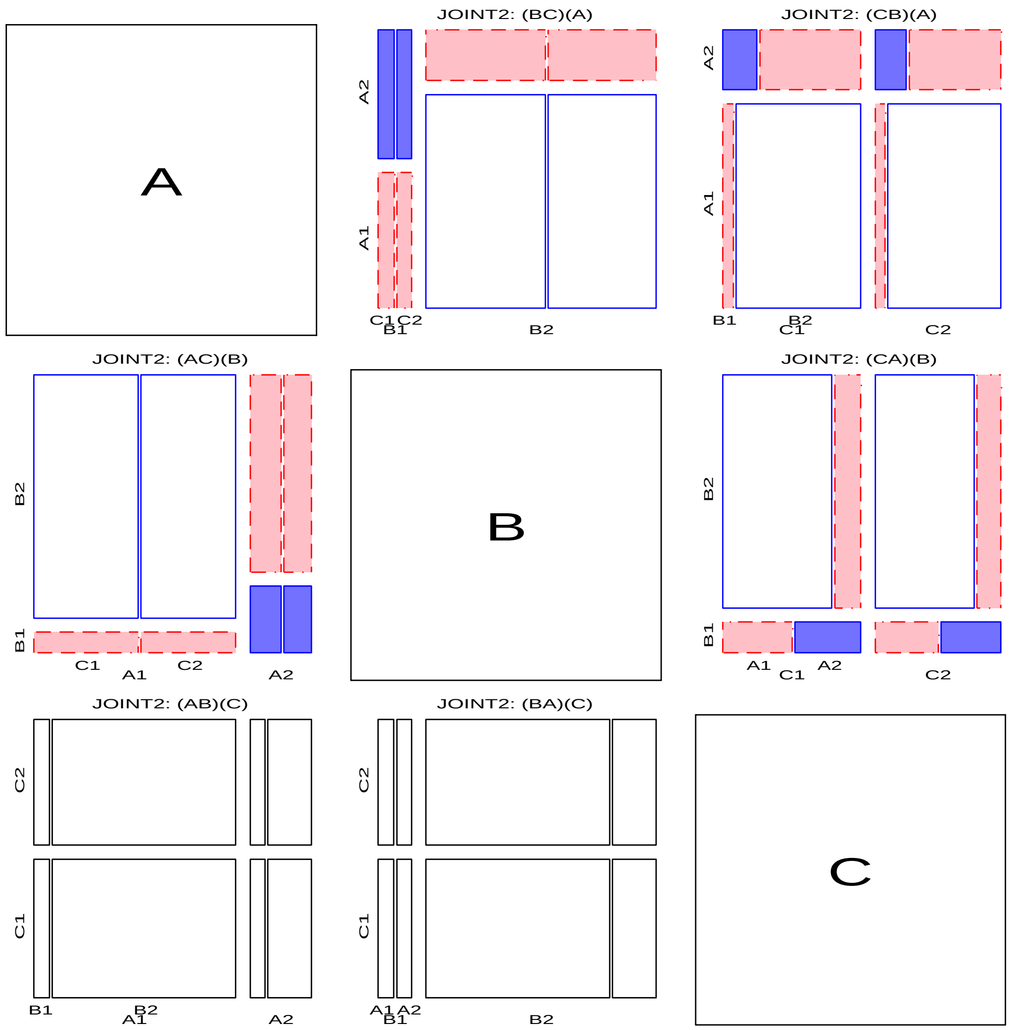


Figure 25: Mosaic matrix for joint independence.

Conditional Independence

- $[AC][BC] \longrightarrow A \perp B \;\forall\; C_i$, but no pair is marginally independent

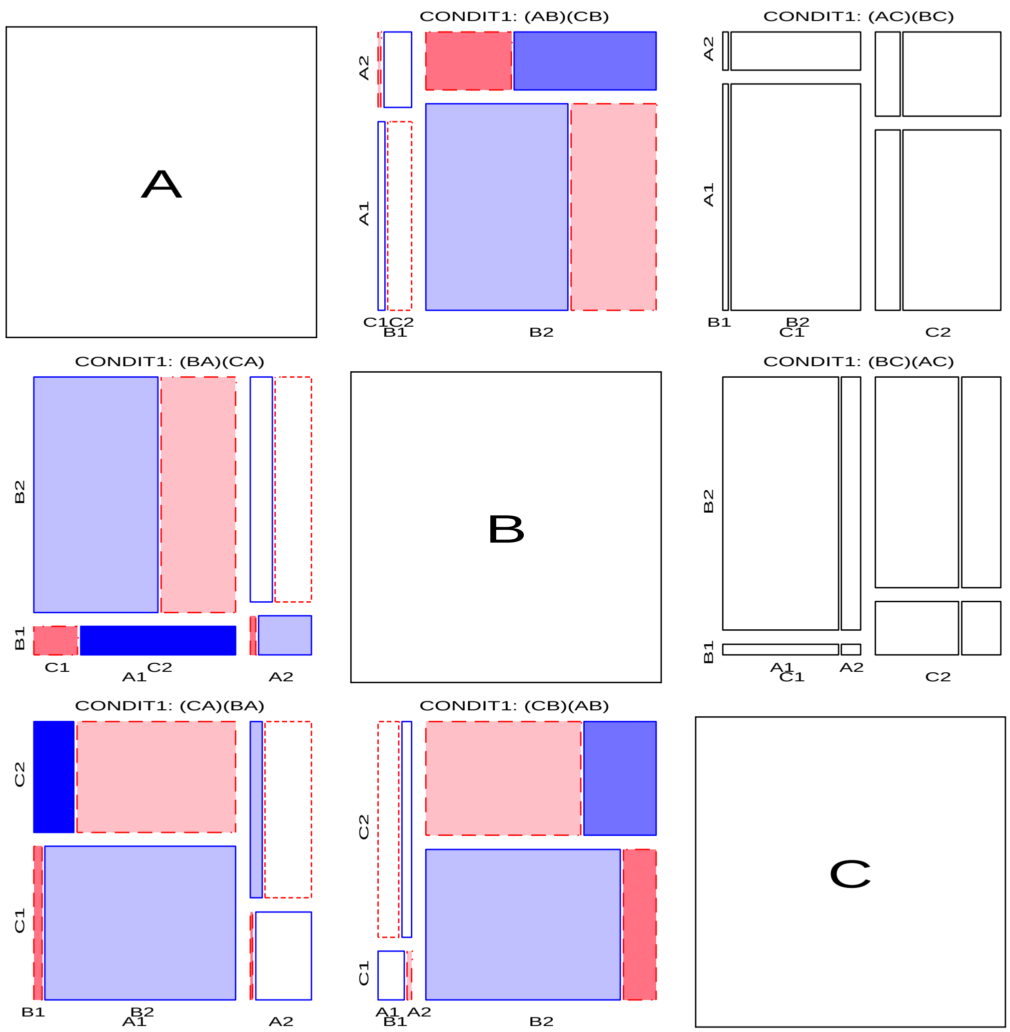


Figure 26: Mosaic matrix for conditional independence

Further Info

- A large collection of documents and programs for graphical data analysis on WWW:
  - http://www.math.yorku.ca/SCS/friendly.html
  - ftp://hotspur.psych.yorku.ca/pub/sas/mosaics
- Static implementations:
  - SAS/IML: MOSAICS: http://www.math.yorku.ca/SCS/mosaics.html
  - SAS/INSIGHT (not exemplary)
  - S-Plus: Jay Emerson http://www.stat.yale.edu/~emerson/JCGS/
- Dynamic/interactive implementations:
  - CGI: http://www.math.yorku.ca/SCS/Online/mosaics/
  - Java: Martin Theus– Mondrian http://www.research.att.com/~theus/Mondrian/Mondrian.html
  - Java: David McClelland, "Seeing Statistics"
  - Mac: Heike Hoffman, Antony Unwin, Martin Theus– Manet http://www1.math.uni-augsburg.de/Manet/
  - XlispStat
    - Ernest Kwan: mosaics.lsp
    - Forrest Young: Vista (5.10) http://forrest.psych.unc.edu/research/

References

Cleveland, W. S. *Visualizing Data*. Hobart Press, Summit, NJ, 1993. 34

Dawson, R. J. M. The "unusual episode" data revisited. *Journal of Statistics Education*, 3(3), 1995. 13

Edwards, D. *Introduction to Graphical Modelling*. Springer-Verlag, New York, 1995. 28

Friendly, M. Mosaic displays for loglinear models. In *ASA, Proceedings of the Statistical Graphics Section*, pages 61–68, Alexandria, VA, 1992.

Friendly, M. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89:190–200, 1994.

Friendly, M. Conceptual and visual models for categorical data. *Amer. Statistician*, 49:153–160, 1995. 2

Friendly, M. Conceptual models for visualizing contingency table data. In Greenacre, M. and Blasius, J., editors, *Visualization of Categorical Data*, chapter 2, pages 17–35. Academic Press, San Diego, CA, 1997.

Hartigan, J. A. and Kleiner, B. Mosaics for contingency tables. In Eddy, W. F., editor, *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pages 286–273. Springer-Verlag, New York, 1981.

Heckman, J. J. and Willis, R. J. A beta-logistic model for the analysis of sequential labor force participation by married women. *Journal of Political Economy*, 85:27–58, 1977. 30

Theus, M. and Lauer, S. R. W. Visualizing loglinear models. *Journal of Computational and Statistical Graphics*, 7:??–??, 1998. In press. 42