Visualizing Categorical Data[*]

Michael Friendly
York University
Toronto, Ontario

Abstract

Graphical methods for quantitative data are well-developed, and widely used in both data analysis (e.g., detecting outliers, verifying model assumptions) and data presentation. Graphical methods for categorical data, however, are only now being developed, and are not widely used.

This paper outlines a general framework for data visualization methods in terms of communication goal (analysis vs. presentation), display goal, and the psychological and graphical design principles which graphical methods for different purposes should adhere to.

These ideas are illustrated with a variety of graphical methods for categorical data, some old and some relatively new, with particular emphasis on methods designed for large, multi-way contingency tables. Some methods (sieve diagrams, mosaic displays) are well-suited for detecting and patterns of association in the process of model building; others are useful in model diagnosis, or as graphical summaries for presentation of results.

Contents

[*] To appear as Chapter 20 (pp. 319–348) in *Cognition and Survey Research*, edited by M. G. Sirken, D. J. Hermann, S. Schechter, N. Schwartz, J. M. Tanur, and R Tourangeau, 1999, John Wiley and Sons, Inc. ISBN 0-471-24138-5. This copy reflects the submitted draft, but not editorial changes in the final copy. I am grateful to Douglas Hermann for helpful comments on the manuscript.

1 Introduction

For some time I have wondered why graphical methods for categorical data are so poorly developed and little used compared with methods for quantitative data. For quantitative data, graphical methods are commonplace adjuncts to all aspects of statistical analysis, from the basic display of data in a scatterplot, to diagnostic methods for assessing assumptions and finding transformations, to the final presentation of results. In contrast, graphical methods for categorical data are still in infancy. There are not many methods, and those that are available in the literature are not accessible in common statistical software; consequently they are not widely used.

What has made this contrast puzzling is the fact that the statistical methods for categorical data, are in many respects, discrete analogs of corresponding methods for quantitative data: log-linear models and logistic regression, for example, are such close parallels of analysis of variance and regression models that they can all be seen as special cases of generalized linear models.

Several possible explanations for this apparent puzzle may be suggested. First, it may just be that those who have worked with and developed methods for categorical data are just more comfortable with tabular data, or that frequency tables, representing sums over all cases in a dataset are more easily apprehended in tables than quantitative data. Second, it may be argued that graphical methods for quantitative data are easily generalized; for example, the scatterplot for two variables provides the basis for visualizing any number of variables in a scatterplot matrix; available graphical methods for categorical data tend to be more specialized.

However, a more fundamental reason may be that quantitative and categorical data display are best served by different visual metaphors. Quantitative data rely on the natural visual representation of magnitude by length or position along a scale; for categorical data, it will be seen that a count is more naturally displayed by an area or by the visual density of an area.

To make the contrast clear, Section 3 describes and illustrates several graphical methods for categorical data, some old and some relatively novel, with particular emphasis on methods designed for large, multi-way contingency tables. Some methods (sieve diagrams, mosaic displays) are well-suited for detecting and patterns of association in the process of model building; others are useful in model diagnosis, or as graphical summaries for presentation of results. A more substantive illustration follows in Section 4. A final section describes some ideas for effective visual presentation. But first I will outline a general framework for data visualization methods in terms of communication goal (analysis vs. presentation), display goal, and the psychological and graphical design principles which graphical methods for different purposes should adhere to.

2 Goals and Design Principles for Visual Data Display

Designing good graphics is surely an art, but equally surely, it is one that ought to be informed by science. In constructing a graph, quantitative and qualitative information is encoded by visual features, such as position, size, texture, symbols and color. This translation is reversed when a person studies a graph. The representation of numerical magnitude and categorical grouping, and the aperception of patterns and their *meaning* must be extracted from the visual display.

There are many views of graphs, of graphical perception, and of the roles of data visualization in discovering and communicating information. On the one hand, one may regard a graphical display as a ’stimulus’ – a package of information to be conveyed to an idealized observer. From this perspective certain questions are of interest: which form or graphic aspect promotes greater accuracy or speed of judgment (for a particular task or question)? What aspects lead to greatest memorability or impact? Cleveland Cleveland and McGill (1984, 1985), Cleveland (1993a), Lewandowsky and Spence Lewandowsky and Spence (1989), Spence (1990) have made important contributions to our understanding of these aspects of graphical display.

An alternative view regards a graphical display as an act of communication—like a narrative, or even a poetic text or work of art. This perspective places the greatest emphasis on the desired communication goal, and judges the effectiveness of a graphical display in how well that goal is achieved. Kosslyn (1985, 1989) has articulated this perspective most clearly from a cognitive perspective.

In this view, an effective graphical display, like good writing, requires an understanding of its purpose—what aspects of the data are to be communicated to the viewer. In writing we communicate most effectively when we know our audience and tailor the message appropriately. So too, we may construct a graph in different ways to use ourselves, to present at a conference or meeting of our colleagues, or to publish in a research report, or a communication to a general audience (Friendly, 1991, Ch. 1).

Figure 1 shows one organization of visualization methods in terms of the primary use or intended communication goal, the functional presentation goal, and suggested corresponding design principles.

The first distinction identifies ***Analysis*** or ***Presentation*** as the primary communication goal of a data graphic (with the understanding that a given graph may serve both purposes—or neither). Among graphical methods designed to help study or understand a body of data, I distinguish those designed for:

- ***reconnaissance***—a preliminary examination, or an overview of a possibly complex terrain;
- ***exploration***—graphs designed to help detect patterns or unusual circumstances, or to suggest hypotheses, analyses or models;
- ***diagnosis***—graphs designed to summarize or critique a numerical statistical summary.

Presentation graphics have different presentation goals as well. We may wish to stimulate, or to persuade, or simply to inform. As in writing, it is usually a good idea to know what it is you want to say with a graph, and tailor its message to that goal. (In what follows, my presentation goal is primarily didactic.)

3 Some Graphical Methods for Categorical Data

One-way frequency tables may be conveniently displayed in a variety of ways: typically as bar charts (though the bars should often be ordered by frequency, rather than by bar-label), dot charts Cleveland (1993b) or pie charts (when percent of total is important).

For two- (and higher-) way tables, however, the design principles of perception, detection, and comparison imply that we should try to show the observed frequencies in the cells in relation to what we would expect those frequencies to be under a reasonable null model — for example, the hypothesis that the row and column variables are unassociated.

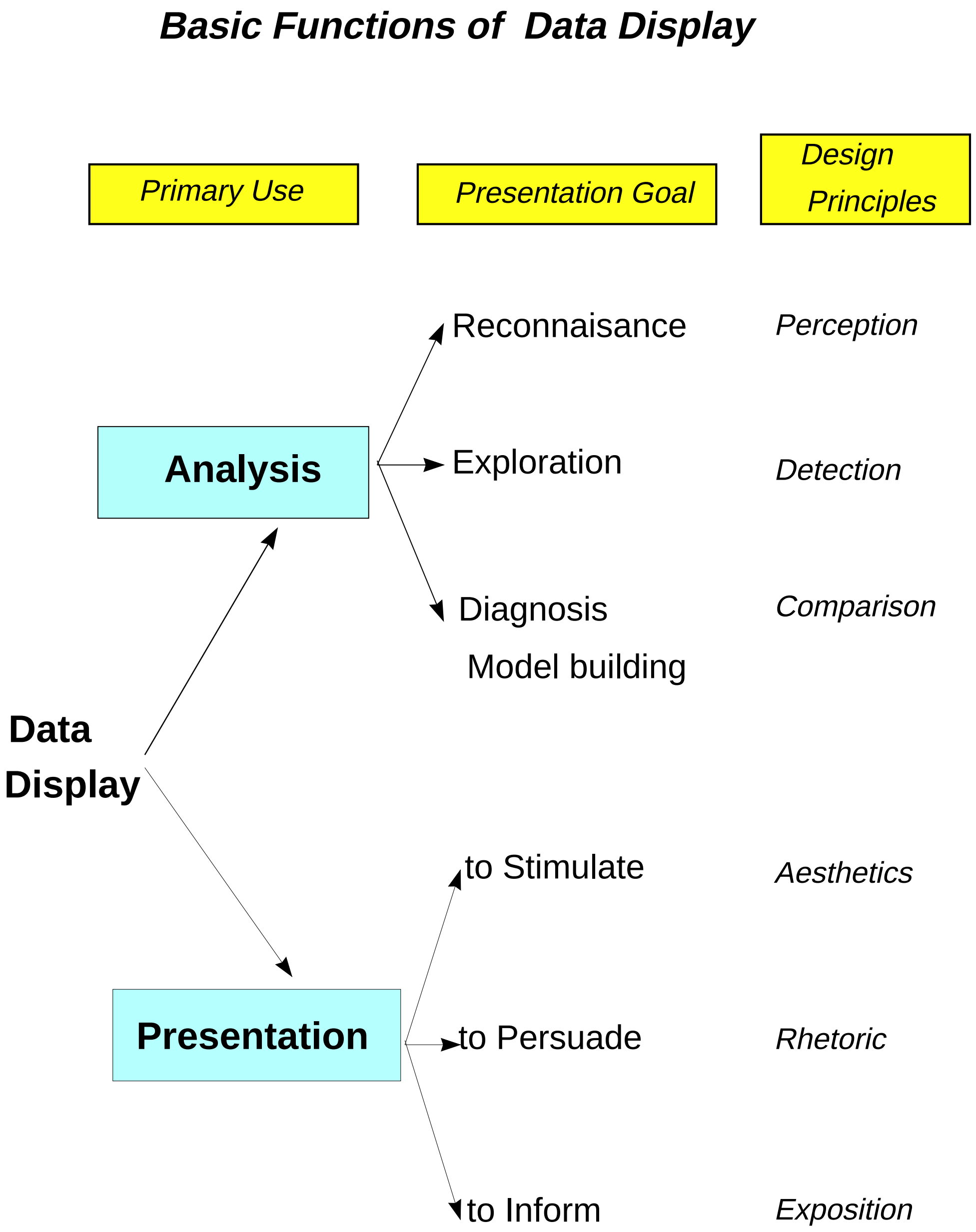

Figure 1: Basic functions of data display.

Table 1: Hair-color eye-color data.

	Hair Color				
Eye 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

To this end, several schemes for representing contingency tables graphically are based on the fact that when the row and column variables are independent, the estimated expected frequencies, m_{ij}, are products of the row and column totals (divided by the grand total): $m_{ij} = n_{i+}n_{+j}/n_{++}$. Then, each cell can be represented by a rectangle whose area shows the cell frequency, n_{ij}, or the deviation from independence.

3.1 Sieve diagrams

Table 1 shows data on the relation between hair color and eye color among 592 subjects (students in a statistics course) collected by Snee (1974). The Pearson χ^2 for these data is 138.3 with nine degrees of freedom, indicating substantial departure from independence. Assume the goal is to understand the *nature* of the association between hair and eye color.

For any two-way table, the expected frequencies under independence can be represented by rectangles whose widths are proportional to the total frequency in each column, n_{+j}, and whose heights are proportional to the total frequency in each row, n_{i+}; the area of each rectangle is then proportional to m_{ij}. Figure 2 shows the expected frequencies for the hair and eye color data. Each rectangle is ruled proportionally to the expected frequency, and we note that the visual densities are equal in all cells.

Riedwyl and Schüpbach (1983, 1994) proposed a ***sieve diagram*** (later called a ***parquet diagram***) based on this principle. In this display the area of each rectangle is proportional to expected frequency, as in Figure 2, but observed frequency is shown by the number of squares in each rectangle. Hence, the difference between observed and expected frequency appears as the density of shading, using color to indicate whether the deviation from independence is positive or negative. (In monochrome versions, positive residuals are shown by solid lines, negative by broken lines.) The sieve diagram for hair color and eye color is shown in Figure 3.

3.2 Mosaic displays for n-way tables

The mosaic display, proposed by Hartigan and Kleiner (1981, 1984), and extended by Friendly (1992, 1994b) represents the counts in a contingency table directly by tiles whose area is proportional to the *observed* cell frequency. One important design goal is that this display should apply extend naturally to three-way and higher-way tables. Another design feature is to serve both exploratory goals (by

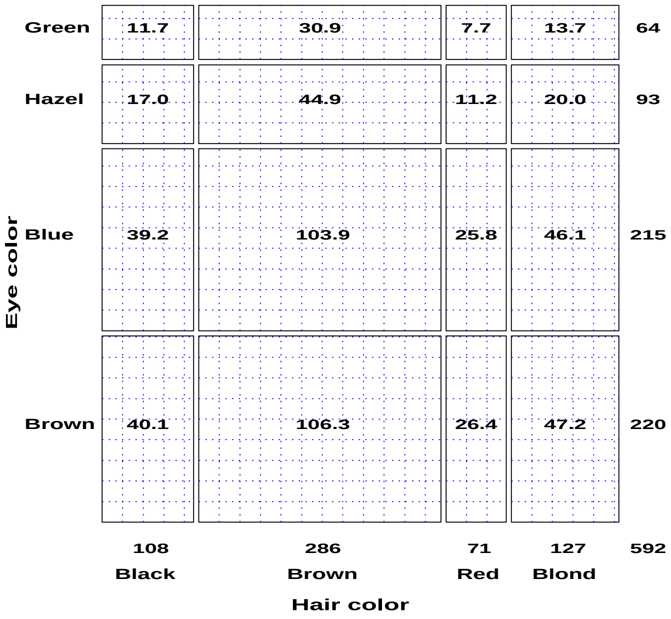

Figure 2: Expected frequencies under independence. Each box has area equal to its expected frequency, and is cross-ruled proportionally to the expected frequency.

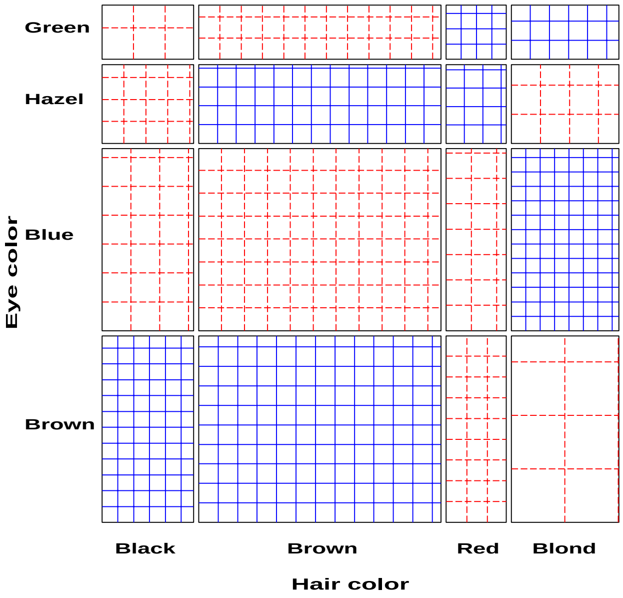

Figure 3: Sieve diagram for hair-color, eye-color data. Observed frequencies are equal to the number squares in each cell, so departure from independence appears as variations in shading density.

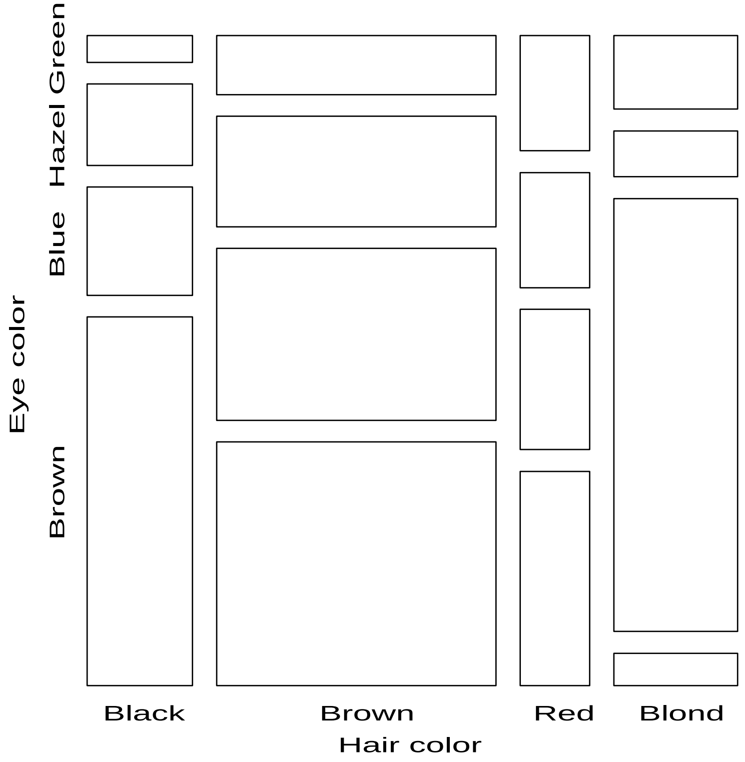

Figure 4: Condensed mosaic for Hair-color, Eye-color data. Each column is divided according to the conditional frequency of eye color given hair color. The area of each rectangle is proportional to observed frequency in that cell.

showing the pattern of observed frequencies in the full table), and model building goals (by displaying the residuals from a given log-linear model).

One form of this plot, called the ***condensed mosaic display***, is similar to a divided bar chart. The width of each column of tiles in Figure 4 is proportional to the marginal frequency of hair colors; the height of each tile is determined by the conditional probabilities of eye color in each column. Again, the area of each box is proportional to the cell frequency, and independence is shown when the tiles in each row all have the same height.

3.2.1 Enhanced mosaics

The enhanced mosaic display Friendly (1992, 1994b) achieves greater visual impact by using color and shading to reflect the size of the residual from independence and by reordering rows and columns to make the pattern of association more coherent. The resulting display shows both the observed frequencies and the pattern of deviations from a specified model.

Figure 5 gives the extended the mosaic plot, showing the standardized (Pearson) residual from independence, $d_{ij} = (n_{ij} - m_{ij})/\sqrt{m_{ij}}$ by the color and shading of each rectangle: cells with positive residuals are outlined with solid lines and filled with slanted lines; negative residuals are outlined with

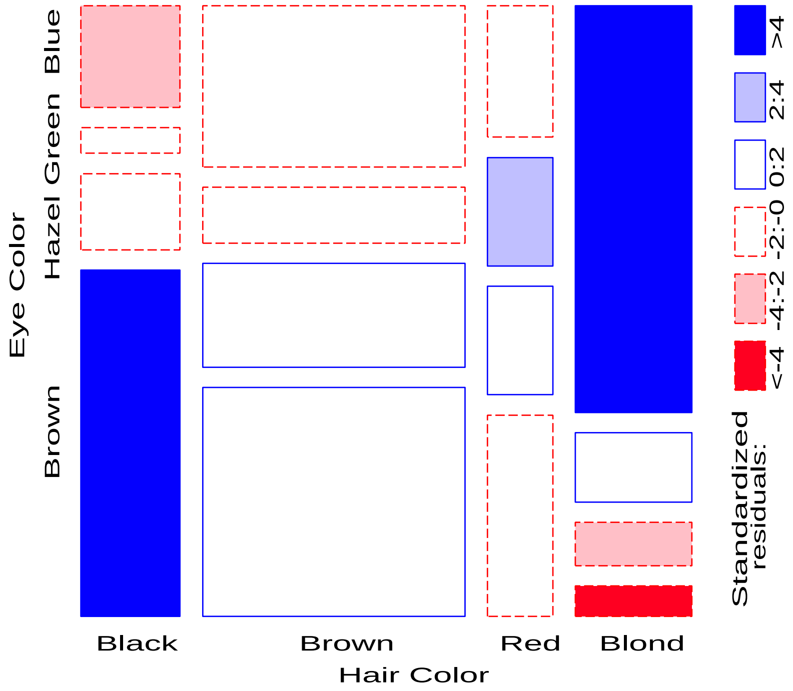

Figure 5: Condensed mosaic, reordered and shaded. Deviations from independence are shown by color and shading. The two levels of shading density correspond to standardized deviations greater than 2 and 4 in absolute value. This form of the display generalizes readily to multi-way tables.

broken lines and filled with grayscale. The absolute value of the residual is portrayed by shading density: cells with absolute values less than 2 are empty; cells with $|d_{ij}| \geq 2$ are filled; those with $|d_{ij}| \geq 4$ are filled with a darker pattern.[1] Under the assumption of independence, these values roughly correspond to two-tailed probabilities $p < .05$ and $p < .0001$ that a given value of $|d_{ij}|$ exceeds 2 or 4. For exploratory purposes, we do not usually make adjustments (e.g., Bonferroni) for multiple tests because the goal is to display the pattern of residuals in the table as a whole. However, the number and values of these cutoffs can be easily set by the user.

When the row or column variables are unordered, we are also free to rearrange the corresponding categories in the plot to help show the nature of association. For example, in Figure 5, the eye color categories have been permuted so that the residuals from independence have an opposite-corner pattern, with positive values running from bottom-left to top-right corners, negative values along the opposite diagonal. Coupled with size and shading of the tiles, the excess in the black-brown and blond-blue cells, together with the underrepresentation of brown-haired blonds and people with black hair and blue eyes is now quite apparent. Though the table was reordered based on the d_{ij} values, both dimensions in Figure 5 are ordered from dark to light, suggesting an explanation for the association. (In this example the eye-color categories could be reordered by inspection. A general method Friendly

[1] Color versions use blue and red at varying lightness to portray both sign and magnitude of residuals.

(1994b) uses category scores on the largest correspondence analysis dimension.)

3.2.2 Multi-way tables

Like the scatterplot matrix for quantitative data, the mosaic plot generalizes readily to the display of multi-dimensional contingency tables. Imagine that each cell of the two-way table for hair and eye color is further classified by one or more additional variables—sex and level of education, for example. Then each rectangle in Figure 5 can be divided vertically to show the proportion of males and females in that cell, and each of those portions can be subdivided again to show the proportions of people at each educational level in the hair-eye-sex group.

3.2.3 Fitting models

When three or more variables are represented in the mosaic, we can fit several different models of independence and display the residuals from that model. We treat these models as null or baseline models, which may not fit the data particularly well. The deviations of observed frequencies from expected, displayed by shading, will often suggest terms to be added to an explanatory model that achieves a better fit. Two examples are:

- ***Complete independence***: The model of complete independence asserts that all joint probabilities are products of the one-way marginal probabilities:

  $$\pi_{ijk} = \pi_{i++}\,\pi_{+j+}\,\pi_{++k} \tag{1}$$

  for all i, j, k in a three-way table. This corresponds to the log-linear model $[A]\,[B]\,[C]$. Fitting this model puts all higher terms, and hence all association among the variables, into the residuals, displayed in the mosaic.

- ***Joint independence***: Another possibility is to fit the model in which variable C is jointly independent of variables A and B,

  $$\pi_{ijk} = \pi_{ij+}\,\pi_{++k}. \tag{2}$$

  This corresponds to the log-linear model $[AB]\,[C]$. Residuals from this model show the extent to which variable C is related to the combinations of variables A and B but they do not show any association between A and B.

For example, with the data from Table 1 broken down by sex, fitting the joint independence model [HairEye][Sex] allows us to see the extent to which the joint distribution of hair-color and eye-color is associated with sex. For this model, the likelihood-ratio G^2 is 19.86 on 15 *df* ($p = .178$), indicating an acceptable overall fit. The three-way mosaic, shown in Figure 6, highlights two cells: among blue-eyed blonds, there are more females (and fewer males) than would be if hair color and eye color were jointly independent of sex. Except for these cells hair color and eye color appear unassociated with sex.

For higher-way tables, there are many more possible models that can be fit. However, they have the characteristic that, for any given table (or marginal subtable), the size of tiles in the mosaic always shows the same observed frequencies, while the shading (showing the sign and magnitude of the

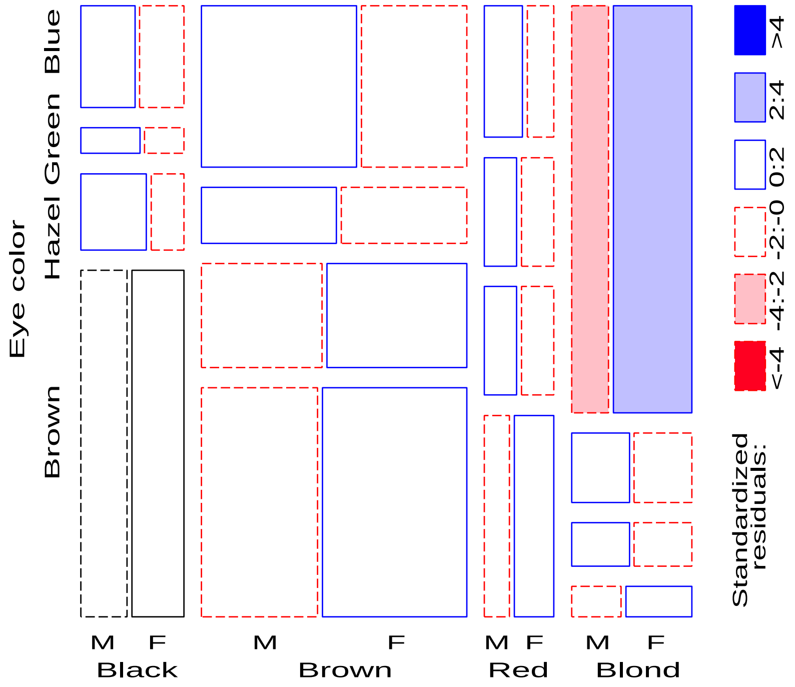

Figure 6: Three-way mosaic display for hair color, eye color, and sex. Residuals from the model of joint independence, $[HE]\,[S]$ are shown by shading. G^2 = 19.86 on 15 df. The only lack of fit is an overabundance of females among blue-eyed blonds.

residuals) varies from one model to another. Because a good-fitting model will have all or mostly small residuals, and these are shown unfilled, a search for an adequate explanatory model can be thought of as "cleaning the mosaic".

3.2.4 Color scales for residuals

The design of the mosaic display shares some features with the use of statistical maps to display quantitative information, and it is appropriate (and perhaps instructive) to consider my use of color and shading in a area-based display.

When we display *signed* magnitudes, such as residuals from a fitted model or differences between time points on a mosaic or a map, the goal is usually to convey both magnitude (how big a difference?) and direction (is it more or less?). For this use, a double-ended color scale, using opposing colors such as red and blue, with darker, more saturated colors at the extremes, is usually most effective. In the mosaic display I use the color scheme Cleveland (1993b) refers to as "two-hues, varying lightness", and which Carr (1994) uses quite effectively in the display of residuals on a map.

The choice of opposing colors and assignment to positive and negative, requires care and consideration of the intended audience, however. I use red for negative "deficit" values and blue for positive "excesses"; Carr (1994) in displays of rates of disease assigns red to positive "hot spots" and cooler blue to negative values. Various government agencies have other conventions about the assignment of color. Unfortunately, black and white reproduction of red/blue displays folds red and blue to approximately equal gray levels. In the mosaic displays, I usually prepare different versions—using gray level and pattern fills for figures to be shown in black and white. Color versions of the figures shown here are available on the WWW at http://www.math.yorku.ca/SCS/Papers/casm/.

3.3 Fourfold Display

A third graphical method based on area as the visual mapping of cell frequency is the "fourfold display" (Friendly, 1994a,c, Fienberg, 1975) designed for the display of 2×2 (or $2 \times 2 \times k$) tables. In this display the frequency n_{ij} in each cell of a fourfold table is shown by a quarter circle, whose radius is proportional to $\sqrt{n_{ij}}$, so the area is proportional to the cell count.

For a single 2×2 table the fourfold display described here also shows the frequencies by area, but scaled in a way that depicts the sample odds ratio, $\hat{\theta} = (n_{11}/n_{12}) \div (n_{21}/n_{22})$. An association between the variables ($\theta \neq 1$) is shown by the tendency of diagonally opposite cells in one direction to differ in size from those in the opposite direction, and the display uses color or shading to show this direction. Confidence rings for the observed θ allow a visual test of the hypothesis $H_0 : \theta = 1$. They have the property that the rings for adjacent quadrants overlap *iff* the observed counts are consistent with the null hypothesis.

As an example, Figure 7 shows aggregate data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex. At issue is whether the data show evidence of sex bias in admission practices Bickel et al. (1975). The figure shows the observed cell frequencies numerically in the corners of the display. Thus there were 2691 male applicants, of whom 1193 (44.4%) were admitted, compared with 1855 female applicants of whom 557 (30.0%) were admitted. Hence the sample odds ratio, Odds (Admit$|$Male) / (Admit$|$Female) is 1.84 indicating that males were almost twice as likely to be admitted.

The frequencies displayed graphically by shaded quadrants in Figure 7 are not the raw frequencies. Instead, the frequencies have been standardized (by iterative proportional fitting) so that all table margins are equal, while preserving the odds ratio. Each quarter circle is then drawn to have an area proportional to this standardized cell frequency. This makes it easier to see the association between admission and sex without being influenced by the overall admission rate or the differential tendency of males and females to apply. With this standardization the four quadrants will align when the odds ratio is 1, regardless of the marginal frequencies.

The shaded quadrants in Figure 7 do not align and the 99% confidence rings around each quadrant do not overlap, indicating that the odds ratio differs significantly from 1—putative evidence of gender bias. The width of the confidence rings gives a visual indication of the precision of the data—if we stopped here, we might feel quite confident of this conclusion.

3.3.1 Multiple strata

In the case of a $2 \times 2 \times k$ table, the last dimension typically corresponds to "strata" or populations, and it is typically of interest to see if the association between the first two variables is homogeneous across strata. The fourfold display is designed to allow easy visual comparison of the pattern of association between two dichotomous variables across two or more populations.

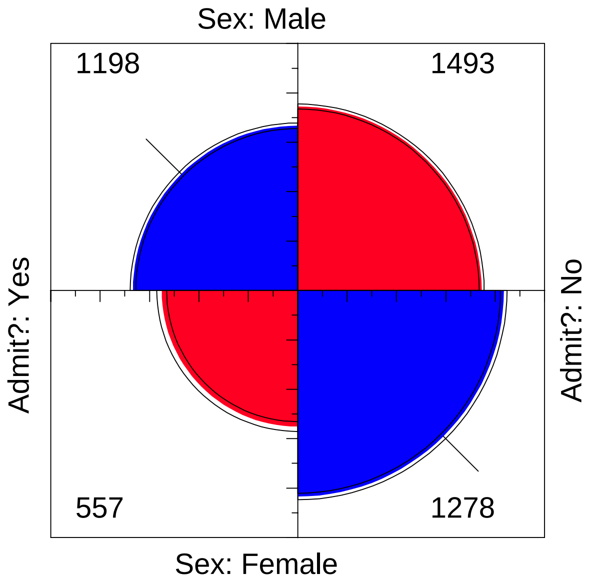

Figure 7: Four-fold display for Berkeley admissions: Evidence for sex bias? The area of each shaded quadrant shows the frequency, standardized to equate the margins for sex and admission. Circular arcs show the limits of a 99% confidence interval for the odds ratio.

For example, the admissions data shown in Figure 7 were obtained from a sample of six departments; Figure 8 displays the data for each department. The departments are labelled so that the overall acceptance rate is highest for Department A and decreases steadily to Department F. Again each panel is standardized to equate the marginals for sex and admission. This standardization also equates for the differential total applicants across departments, facilitating visual comparison.

Surprisingly, Figure 8 shows that, for five of the six departments, the odds of admission is essentially identical for men and women applicants. Department A appears to differs from the others, with women approximately 2.86 ($= (313/19)/(512/89)$) times *more* likely to gain admission. This appearance is confirmed by the confidence rings, which in Figure 8 are *joint* 99% intervals for θ_c, $c = 1, \ldots, 6$.

This result, which contradicts the display for the aggregate data in Figure 7, is a nice example of Simpson's paradox. The resolution of this contradiction can be found in the large differences in admission rates among departments. Men and women apply to different departments differentially, and in these data women apply in larger numbers to departments that have a low acceptance rate. The aggregate results are misleading because they falsely assume men and women are equally likely to apply in each field.[2]

[2] This explanation ignores the possibility of structural bias against women, e.g., lack of resources allocated to departments that attract women applicants.

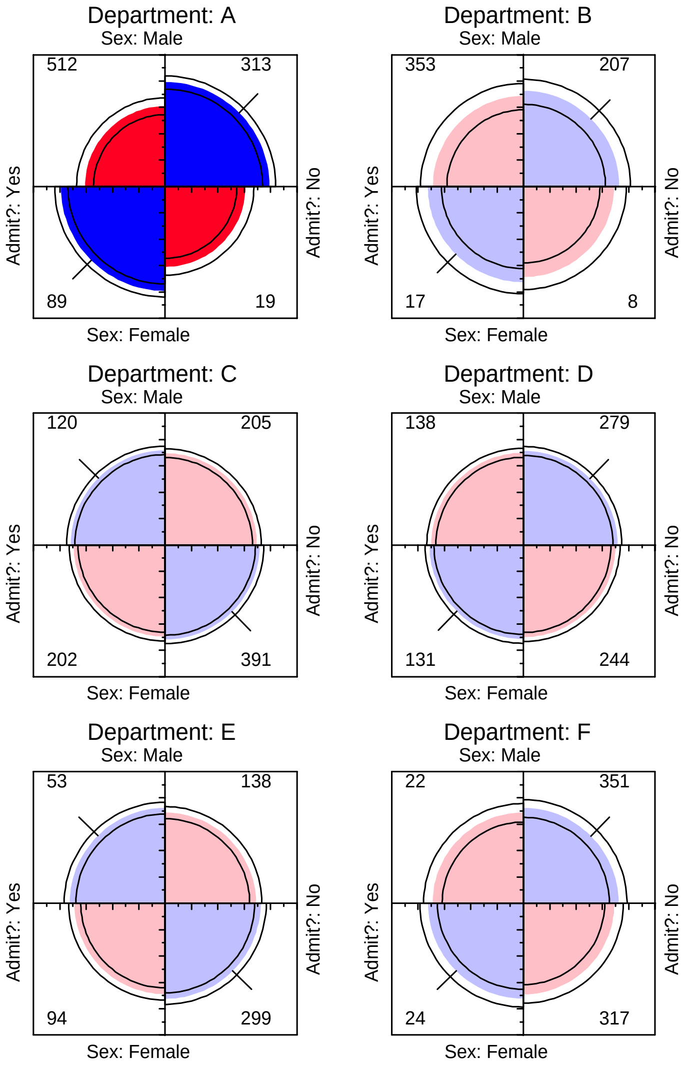

Figure 8: Fourfold display of Berkeley admissions, by department. In each panel the confidence rings for adjacent quadrants overlap if the odds ratio for admission and sex does not differ significantly from 1. The data in each panel have been standardized as in Figure 7.

3.3.2 Visualization principles

An important principle in the display of large, complex datasets is ***controlled comparison***—we want to make comparisons against a clear standard, with other things held constant. The fourfold display differs from a pie chart in that it holds the angles of the segments constant and varies the radius, whereas the pie chart varies the angles and holds the radius constant. An important consequence is that we can quite easily compare a series of fourfold displays for different strata, since corresponding cells of the table are always in the same position. As a result, an array of fourfold displays serve the goals of comparison and detection better than an array of pie charts. Moreover, it allows the observed frequencies to be standardized by equating either the row or column totals, while preserving the odds ratio. In Figure 8, for example, the proportion of men and women, and the proportion of accepted

Table 2: National Assessment of Educational Proficiency, 1992 Mathematics. Counts of Achievement Levels by Program, Poverty and Ethnicity.

		Not Poor				Poor			
		White	Asian	Black	Hispanic	White	Asian	Black	Hispanic
Academic Program	Advanced	23403	1711	0	1050	6814	542	0	633
	Proficient	224723	15145	4591	5716	96442	5654	3308	7500
	Basic	679319	30590	32937	26709	439109	27276	64451	42726
	BelowBasic	102849	3645	23983	20596	135919	7974	132421	52327
Non Academic Program		White	Asian	Black	Hispanic	White	Asian	Black	Hispanic
	Advanced	1866	0	0	0	0	0	0	0
	Proficient	24317	2518	1039	570	716	557	0	482
	Basic	208710	13142	14271	16338	151369	5478	9229	23551
	BelowBasic	244240	7149	60868	35233	254734	3315	134874	81356

applicants were equated visually in each department. This provides a clear standard which also greatly facilitates controlled comparison.

Another principle is ***visual impact***—we want the important features of the display to be easily distinguished from the less important Tukey (1993). In Figure 8 distinguishes the one department for which the odds ratio differs significantly from 1 by shading intensity, even though the same information can be found by inspection of the confidence rings.

4 Example: NAEP 1992 Grade 12 Mathematics

The previous section used simple examples to illustrate these somewhat novel techniques themselves. This section illustrates the use of these graphical methods in attempting to understand a large, complex dataset of the sort often found in public policy research, and which often defies standard visualization methods.

Table 2 is a four-way classification of over 3.5 million examinees from the 1992 National Assessment of Educational Proficiency (Mullis et al., 1993, Wainer, 1997), classified by their high school program (Academic, Non Academic), their economic status (Not Poor vs. Poor), and ethnic group (White, Asian, Hispanic, Black). Students in the background groupings are then classified by their performance on the NAEP Mathematics test, recorded here as Advanced, Proficient, Basic or Below Basic.

One glance at Table 2 shows two things: First, very few students achieve Advanced level, none in Non Academic Programs (save a few Not Poor white students); as a result, the analysis below combines Advanced and Proficient, labelled “Proficient+”. Second, the usefulness of this table as a data display is nil, except as a record of the results.

Mosaic displays are constructed sequentially, with the variables arranged in a particular order, and it usually makes sense to order the variables in a quasi-causal, or predictor-response fashion. Here I consider Ethnicity and Poverty as (partial) determinants of academic Program, and all three of these as potential predictors of achievement level.

4.1 Analysis of [Ethnicity, Poverty, Achievement Level]

For simplicity, we begin with an analysis of the marginal table of Ethnicity, Poverty and Level, collapsing over Program. Figure 9 shows the mosaic for (the marginal table of) Ethnicity and Poverty, fitting the independence model. If Poverty were unrelated to Ethnic group, the tiles would all be equally wide in each column. There is, of course, a pronounced association between Poverty and Ethnic Group ($G^2(3) = 193.95$), as shown by the shading pattern of the residuals: Asians and Whites are more frequently NotPoor and Hispanics and Blacks are more frequently Poor, than would be the case if these variables were independent.[3]

Figure 10 shows the relation between Achievement Level and Ethnicity and Poverty, fitting a model [(EthPov)(Level)] which says that Level is independent of the combinations of Ethnicity and Poverty jointly. The shading pattern shows how violently this model is contradicted by the data ($G^2(14) = 553,568$):

- Among NotPoor Whites, an over-abundance are in Basic level or higher; for Poor Whites, however, frequencies significantly greater than expected under this model of joint independence occur only in the Basic level.
- Among Asians, there are greater than expected frequencies in all but the Below Basic category, independent of Poverty.
- Among both Hispanics and Blacks, there are greater than expected frequencies in the Below Basic level.

It is depressingly striking how few 12th grade children are classified in the Advanced or Proficient categories.

4.2 Analysis of the Full Table

With some understanding among the relations among Poverty, Ethnicity and Achievement Level, we proceed to fitting sequential models of joint independence to the full table. The analysis goal is not to provide an adequate model, but simply to remove the associations we have already seen, thereby revealing the associations which remain.[4]

Figure 11 shows the relation between Academic Program and the combinations of Ethnicity and Poverty; residuals show how Program is associated with the categories of the other two variables:

- NotPoor Whites are more likely to be in Academic programs, while the reverse is true for Poor Whites.

[3] Blacks and Hispanics are interchanged from the original table, in accord with association ordering, described below.

[4] All of these models fit very badly, partly due to the enormous sample size, but they are to be regarded only as baseline models. When a model of joint independence, say, (Ethnicity, Poverty)(Program), is fit, the association between Ethnicity and Poverty is fit *exactly*, and so does not appear in the residuals.

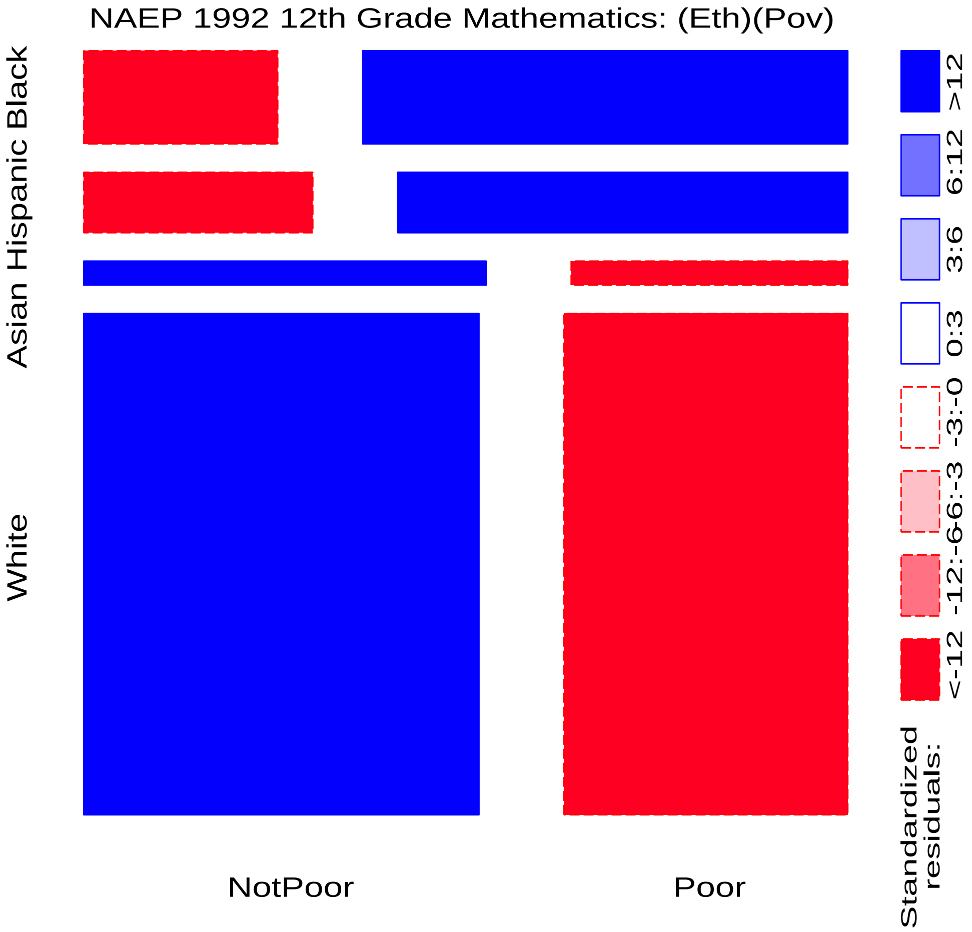

Figure 9: Mosaic display of Ethnicity and Poverty. Residuals show the pattern of (marginal) association between these two variables.

- Asians are more likely to be in Academic programs. This result holds regardless of Poverty.
- Hispanics and Blacks are more likely to be in NonAcademic programs, also independent of Poverty.

Figure 12 shows the relation between NAEP Achievement Level and the combinations of Ethnicity, Poverty, and academic Program jointly. The associations among the three background variables just observed (Figure 11) have been eliminated and the residuals now show how Achievement Level is associated with the combinations of the other three variables. This is admittedly a complex display, but a few moments study reveals the following:

- NotPoor Whites in Academic programs are more likely to achieve Basic level or above; NotPoor Whites in NonAcademic programs are more likely to be classified Below Basic level.
- Approximately the same pattern obtains for Poor Whites. (More detailed analysis shows them

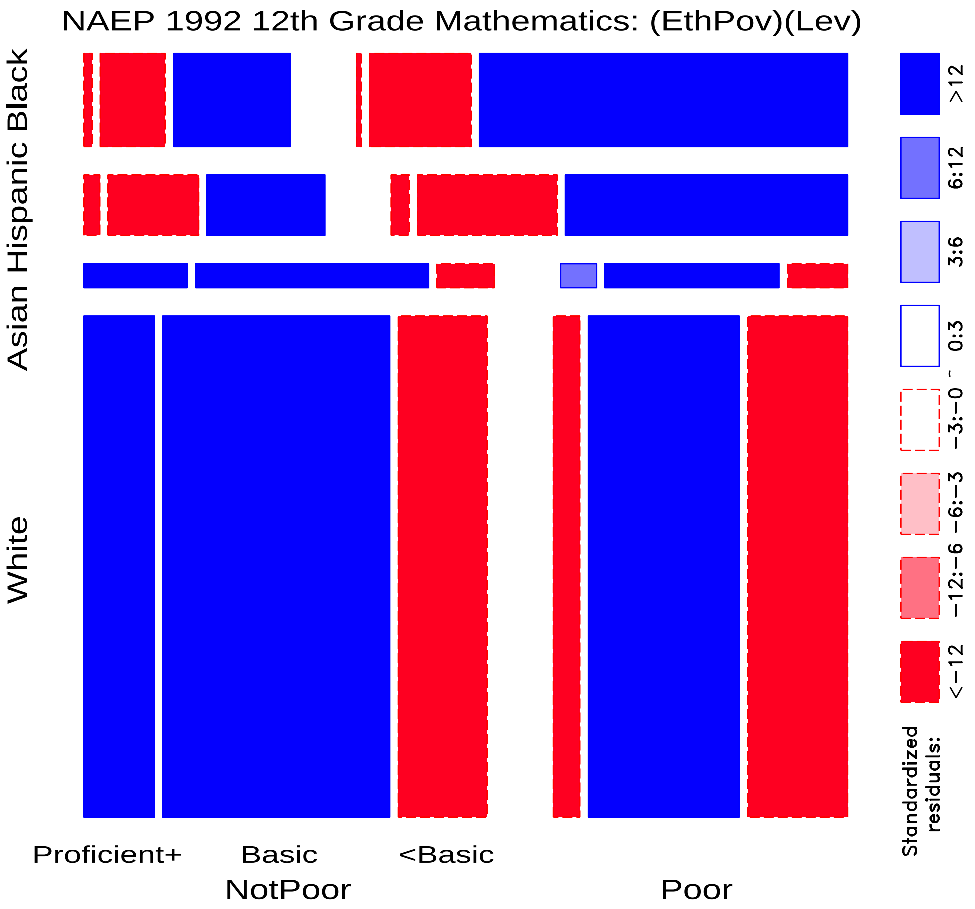

Figure 10: NAEP data, Ethnicity and Poverty vs. Achievement Level. Residuals display the dependence of Level on both Ethnicity and Poverty.

less likely than NotPoor Whites to achieve Advanced level.)

- Asians in Academic programs are more likely to achieve Basic level or above; Asians in Non-Academic programs are most likely to achieve just Basic level. Both statements apply independently of Poverty.

The patterns for Blacks and Hispanics diverge here for the first time:

- Blacks are most likely to achieve Below Basic level, independent of Program and Poverty, except that NotPoor Blacks in Academic programs are more likely to achieve in Basic level.
- Hispanics are similar, except that NotPoor Hispanics in Academic programs are more likely to achieve Advanced standing than the model of joint independence predicts.

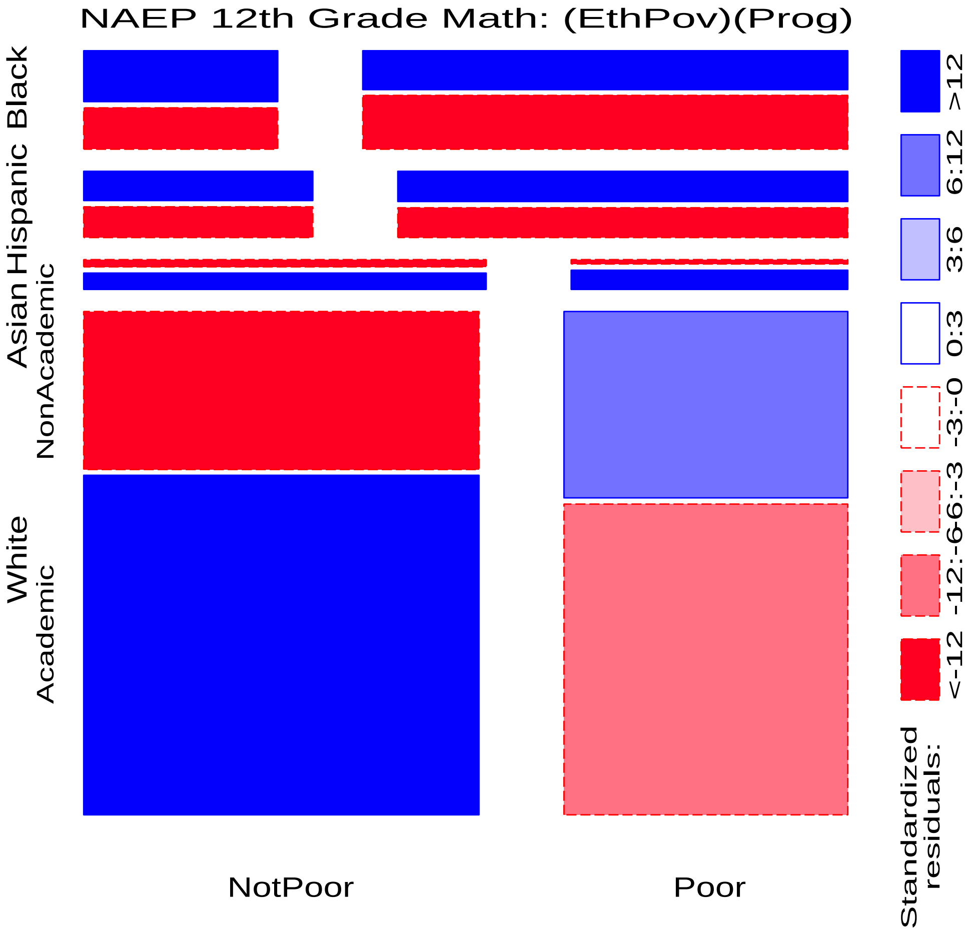

Figure 11: NAEP data, Ethnicity and Poverty vs. Academic Program. Residuals display how Academic Program depends on both Ethnicity and Poverty.

4.3 Other displays

The NAEP data is a $2 \times 2 \times 4 \times 4$ table, which we can regard as a collection of 2×2 tables, showing the relation between Poverty and Academic Program for each of the 16 Ethnicity by Achievement Level groups. Figure 13 shows one fourfold display of these data. In each panel lower left quadrant represents the most advantaged students, and the upper right represents the least advantaged. Thus, a positive association between Poverty and Non Academic Program is shown by tick marks in the $45°$ direction, and we see that nearly all associations are positive.

In order to adjust for the great imbalance in the numbers of examinees in the different Poverty-Ethnic group combinations, the data in Table 2 were standardized to equate the column totals in that table. These standardized frequencies are displayed directly in Figure 13, scaled so that the largest such frequency has unit radius. As a result, the distributions across the columns and rows may be readily compared.

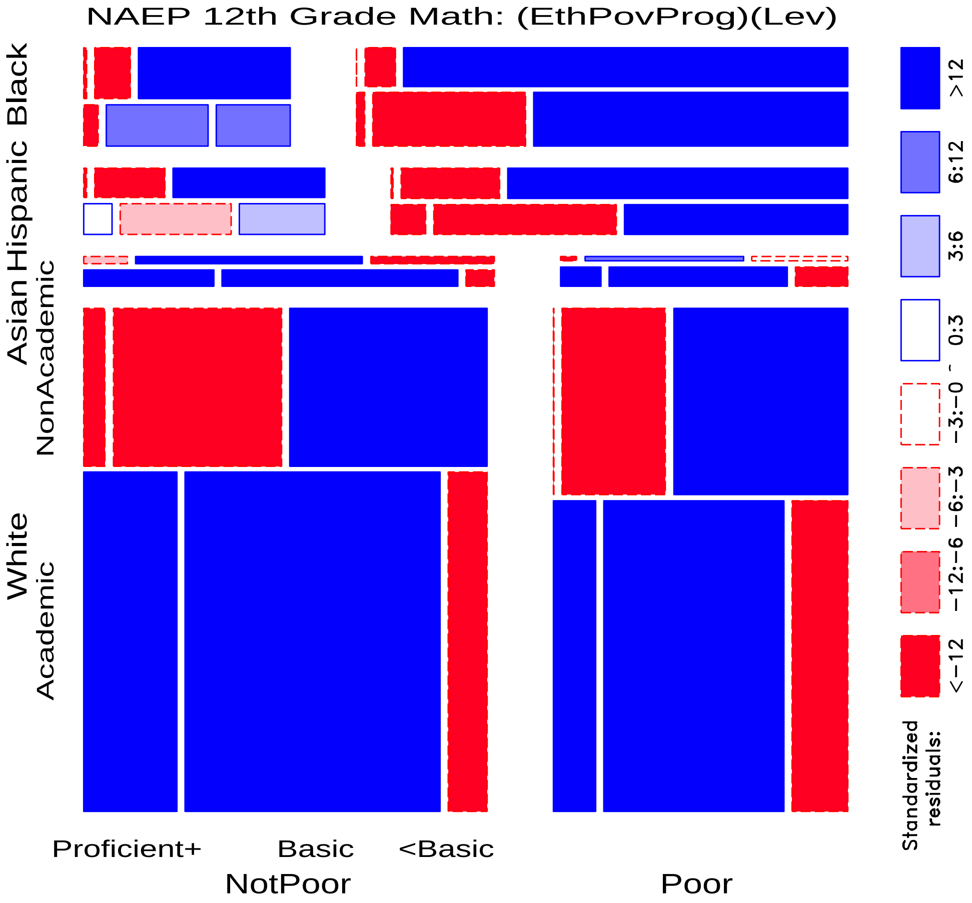

Figure 12: NAEP data, Ethnicity, Poverty and Academic Program vs. Achievement.

It is immediately apparent that the greatest proportion of all students performed at the lowest achievement levels, and that, at the very lowest level the largest proportion are in Non Academic programs, particularly among Hispanics and Blacks. Second, those who achieve Proficient level are much more likely Not Poor, and taking Academic programs. Finally (and sadly) the dearth of students achieving Advanced level stands out clearly.

In closing this section, I'll comment on more traditional statistical analyses and graphical displays of such data. One statistical analysis would be to fit a more "meaningful" log-linear model than the baseline models used earlier. When this is done, we find that the only (barely) tenable model (short of the saturated model, which must fit perfectly) is the all-three-way-interaction model, symbolized as

$$(EthPovProg)(EthPovLev)(EthProgLev)(PovProgLev)$$

If Ethnicity, Poverty and Program are all regarded as predictors of achievement Level, this model says that achievement level depends on the combinations of Ethnicity and Poverty, the combinations of Ethnicity and academic Program, and the combinations of Poverty and Program. Unfortunately, this

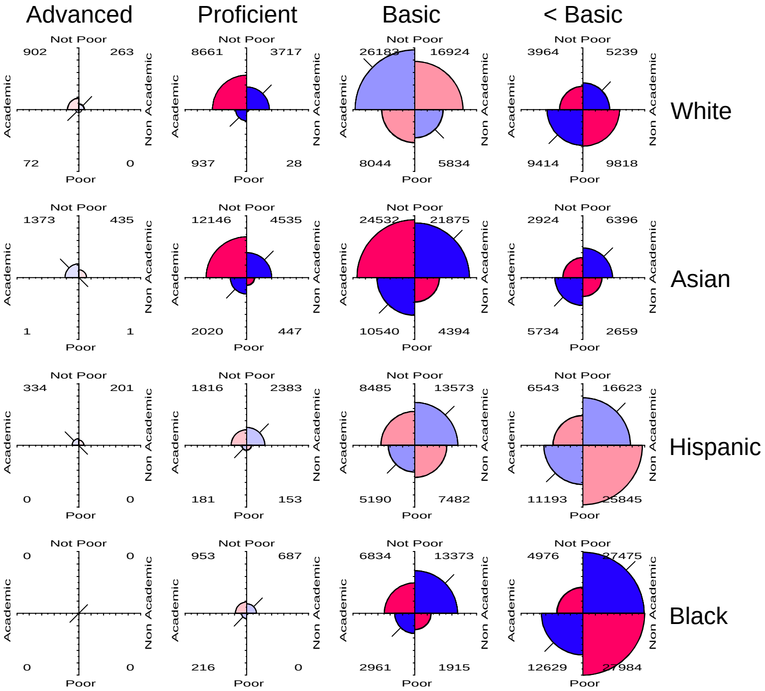

Figure 13: NAEP data, fourfold display.

analysis is not helpful in understanding *how* achievement Level depends on those factors, although tables of parameter estimates from this model could be interpreted to reveals some of the effects shown in the mosaic displays.

Alternatively, a logistic regression might attempt to model achievement level in relation to the predictors. As an example, consider a model for the probability that a student achieves at least Basic level. The logistic model of equivalent complexity to the log-linear model can be expressed as

$$\begin{aligned} \text{Pr (Basic or better)} \sim & \text{Poverty} + \text{Ethnicity} + \text{Program} + \\ & \text{Poverty} \times \text{Ethnicity} + \text{Poverty} \times \text{Program} + \text{Ethnicity} \times \text{Program} \end{aligned} \tag{3}$$

where $\sim$ means "is modelled by". These results provide a sensible and relatively simple way to display these data, which may serve a presentation goal better than the novel displays shown previously. The idea is to plot the predicted probabilities or predicted log-odds from the model. Figure 14 shows these results in one possible format, plotting the log odds directly, but with a probability scale on the right for those who are more comfortable with probabilities than logits. This figure shows most of the relations between Achievement Level and the predictors discussed earlier, but shows none of the relations among the predictors. The interaction terms in model (3) stem from the fact that the profiles in Figure 14 are not parallel. For students in Non Academic programs, the effect of Poverty increases as achievement level decreases and is far greater for Hispanics than for Asians. For students in Academic programs, the effect of Poverty is larger for both Asians and Hispanics than for Whites and Blacks. This graph displays these more subtle aspects of the fitted logistic model, but the dominant message is clear: An Academic program greatly increases the likelihood of at least Basic achievement; being Poor greatly decreases it.

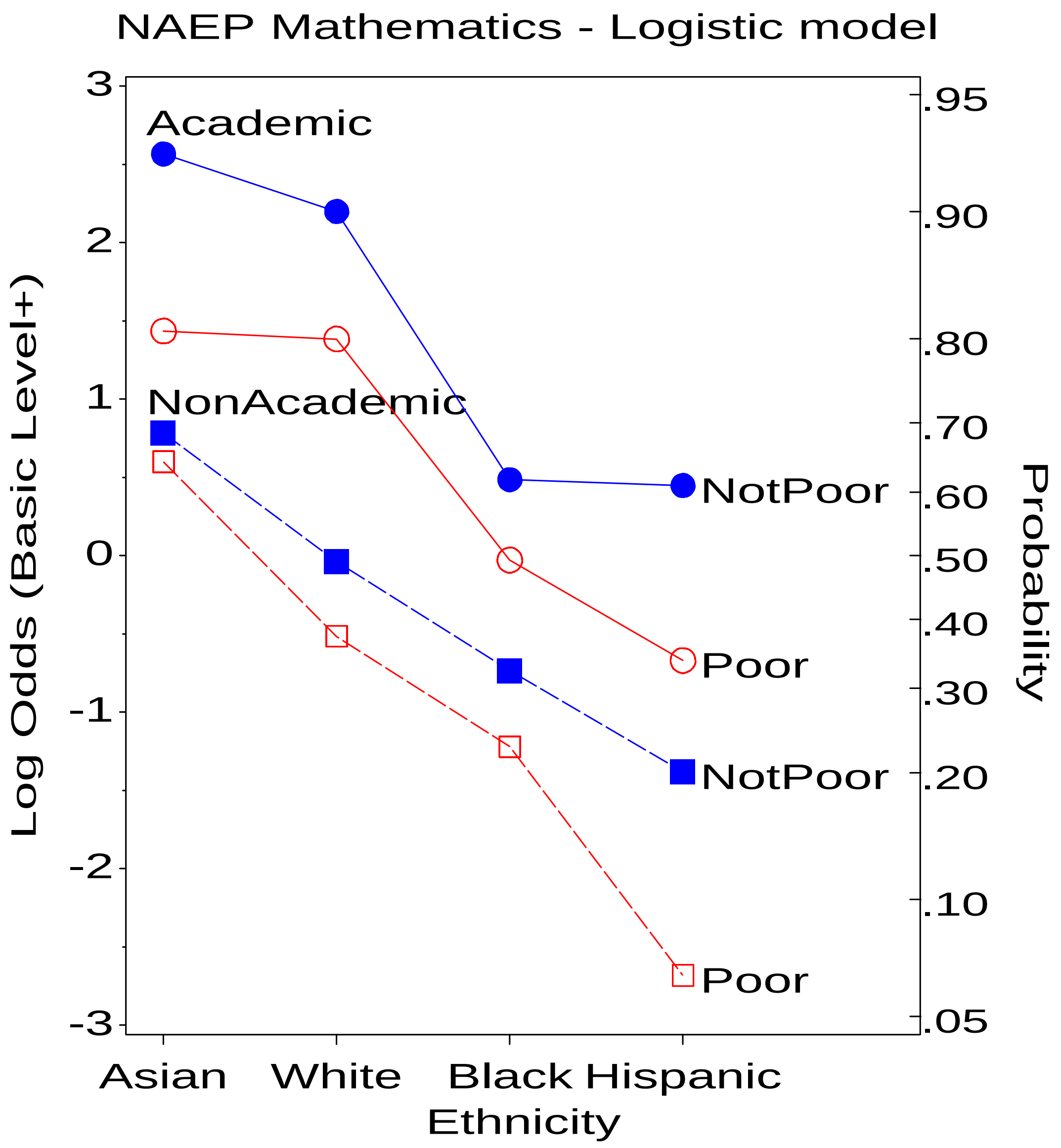

Figure 14: NAEP data, Logistic regression.

5 Effect Ordering for Data Displays

One reason why graphs of quantitative data are effective is that graphing values on quantitative axes automatically orders those values so that “less” and “more” are given visually ordered positions in the display. However, when data values are classified by “factors”, the ordering of the levels of the factor variables has considerable impact on graphical display. Ordered factors (such as age, level of education, etc.) are usually (though not always) most sensibly arranged in their natural order for all presentation goals.

Unordered factors (disease classification, geographic region) deserve more careful thought. For a geographic classification (states, provinces) is it common to arrange the units alphabetically or (as is common in Canada) from east to west. When the goal of presentation is detection or comparison (as opposed to table lookup), this is almost always a bad idea.

Instead, I suggest a general rule for arranging the levels of unordered factors in visual displays — tables as well as graphs: ***sort the data by the effects to be observed***. Sorting has both global and local effects: globally, a more coherent pattern appears, making it easier to spot exceptions; locally, effect-ordering brings similar items together, making them easier to compare. See de Falguerolles et al. (1997) for related ideas.

The use of this principle is illustrated by the following:

- ***Main-effects ordering***: For quantitative data where the goal is to see "typical" values, sort the units in boxplots, dotplots and tables by means, medians, or by row and column effects. (If the goal is to see differences in variability, sort by standard deviation or interquartile range.)

  For example, Figure 15 shows a Trellis dotplot display Cleveland (1993b) of data on barley yields in a three-factor design: Year by Site by Variety. All three factors have been ordered in the panels by median overall yield. With this arrangement, the plot shows a startling anomaly which is not apparent in conventional plots: for all sites and for all varieties, yields in 1931 were greater than in 1932, except at the Morris site, where this difference is reversed. Wainer (1993) and Carr and Olsen (1996) have presented similar arguments for the effectiveness of such orderings on detection.

- ***Discriminant ordering***: For multivariate data, where the goal is to compare different groups in their means on a (possibly large) number of variables, arrange the *variables* in tables or visual displays according to the weights of those variables on the dimensions which best discriminate among the groups (canonical discriminant dimensions).

  Figure 16 shows the means for various characteristics of automobiles (1972 data from Chambers et al. (1983)), classified by region of manufacture. The variables have been scaled so that longer rays represent "better" for all variables, and arranged around the circle according to their positions on the largest discriminant dimension. The figure makes it immediately clear that American cars are mainly larger and heavier, Japanese cars are better in price, mileage and repair record, while European cars have intermediate and mixed patterns. The same idea of ordering variables could be used in a profile plot or parallel coordinates plot.

- ***Correlation ordering***: A related idea is that in the display of multivariate data by glyph plots, star plots, parallel coordinate plots and so forth, the variables should be ordered according to the largest principal component or biplot Gabriel (1980, 1981) dimension(s). This arrangement brings similar variables together, where similarity is defined in terms of patterns of correlation.

- ***Association ordering***: For categorical data, where the goal is to understand the pattern of association among variables, order the levels of factors according to their position on the largest correspondence analysis dimension Friendly (1994b).

6 Mosaic Matrices and Coplots for Categorical Data

A second reason for the wide usefulness of graphs of quantitative data has been the recognition that combining multiple views of data into a single display allows detection of patterns which could not

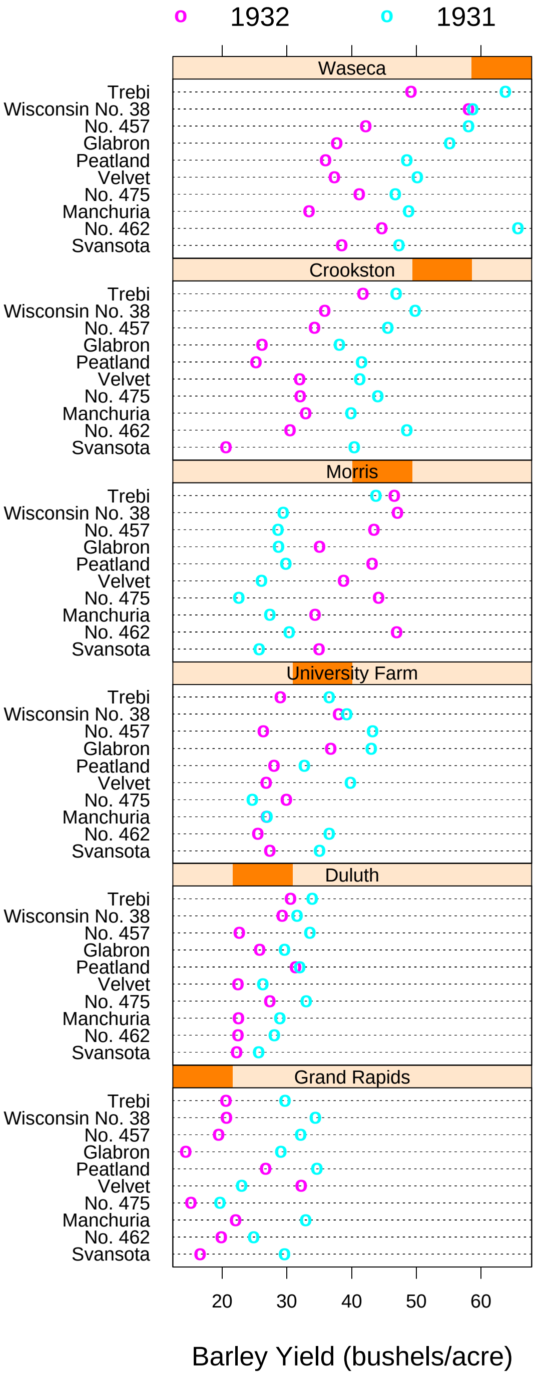

Figure 15: Three factor data, each factor ordered by median overall yield.

readily be discerned from a series of separate graphs. The scatterplot matrix shows all pairwise (marginal) views of a set of variables in a coherent display, whose design goal is to show the interdependence among the collection of variables as a whole. The conditioning plot, or *coplot* Cleveland (1993b) shows a collection of (conditional) views of several variables, conditioned by the values of

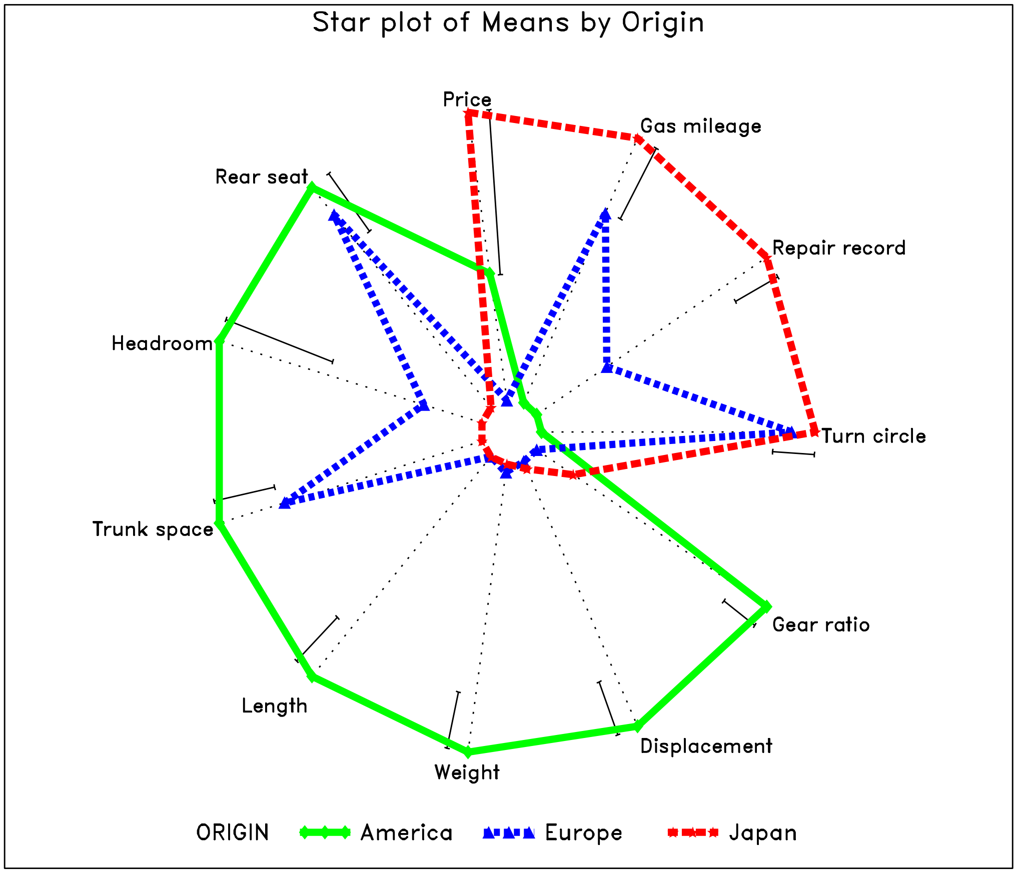

Figure 16: The star-shaped figures show the means for various characteristics of automobiles, classified by region of manufacture. The variables have been arranged around the circle according to their positions on the largest discriminant dimension. Error bars show the (univariate) least significant difference.

one or more other variables. The design goal is to visualize how a relationship depends on or changes with one or more additional factors. These ideas can be readily extended to categorical data.

One analog of the scatterplot matrix for categorical data is a matrix of mosaic displays showing some aspect of the bivariate relation between all pairs of variables. The simplest case shows, for each pair of variables, the marginal relation, summed over all other variables. For example, Figure 17 shows the pairwise marginal relations among the variables Admit, Gender and Department in the Berkeley data shown earlier in fourfold displays (Figure 7 and Figure 8). The panel in row 2, column 1 shows that Admission and Gender are strongly associated marginally, as we saw in Figure 7, and overall, males are more often admitted. The diagonally-opposite panel (row 1, column 2) shows the same relation, splitting first by gender.

The panels in the third column (and third row) illuminate the explanation for the paradoxical result (Figure 8) that, within all but department A, the likelihood of admission is equal for men and women. The (1,3) panel shows the marginal relation between Admission and Department; departments A and B have the greatest overall admission rate, departments E and F the least. The (2, 3) panel shows that men apply in much greater numbers to departments A and B, while women apply in greater numbers

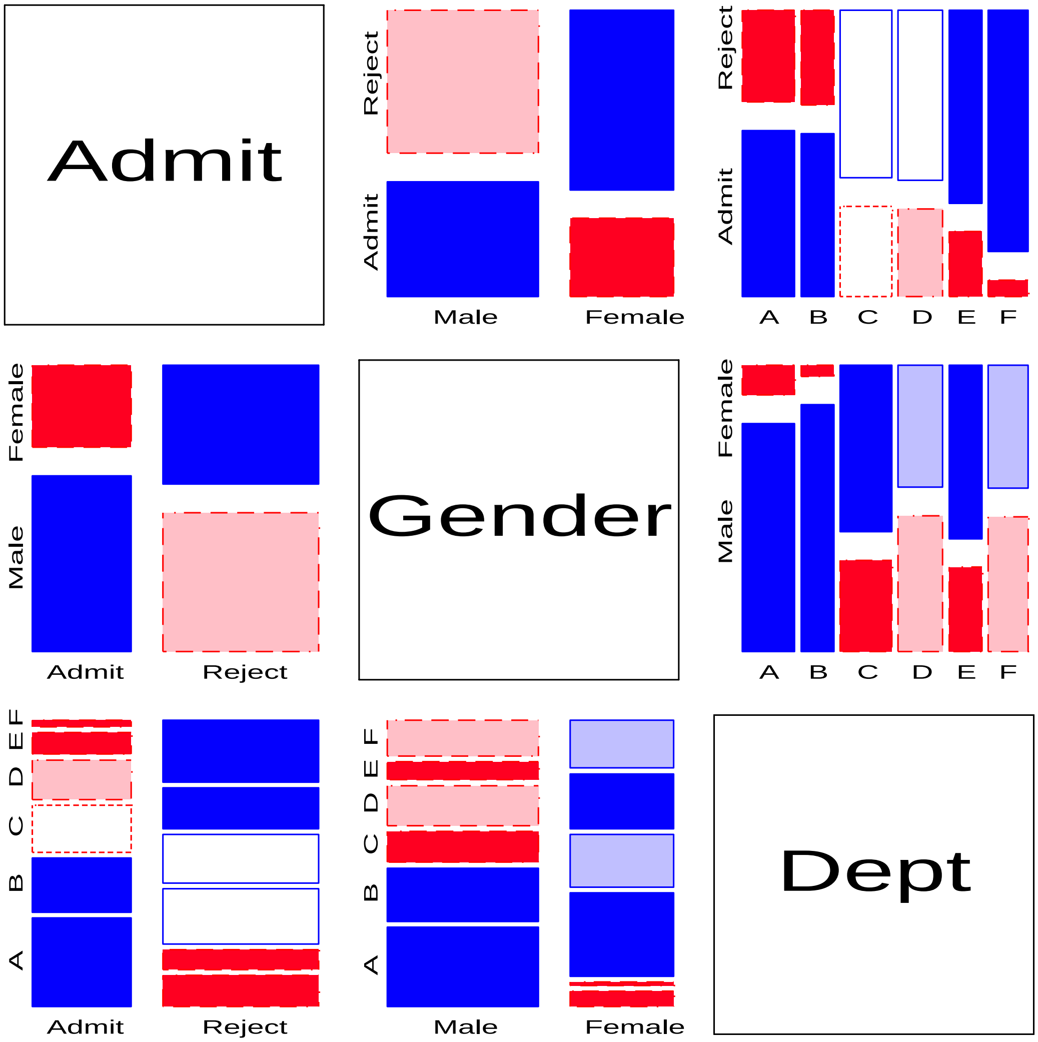

Figure 17: Mosaic matrix of Berkeley admissions. Each panel shows the marginal relation, fitting an independence model between the row and column variable, collapsed over other variable(s)

to the departments with the lowest overall rate of admission.

Several further extensions are now possible. First, we need not show the marginal relation between each pair of variables in the mosaic matrix. For example, Figure 18 shows the pairwise *conditional* relations among these variables, in each case fitting a model of conditional independence with the remaining variable controlled. Thus, the (1,2) and (2,1) panels shows the fit of the model [Admit,Dept] [Gender, Dept], which asserts that Admission and Gender are independent, given (controlling for) department. Except for Department A, this model fits quite well, again indicating lack of gender bias.

Second, the framework of the scatterplot matrix can now be used as a general method for displaying marginal or conditional relations among a mixture of quantitative and categorical variables. For marginal plots, pairs of quantitative variables are shown as a scatterplot, while pairs of categorical variables are shown as a mosaic display. Pairs consisting of one quantitative and one categorical variable

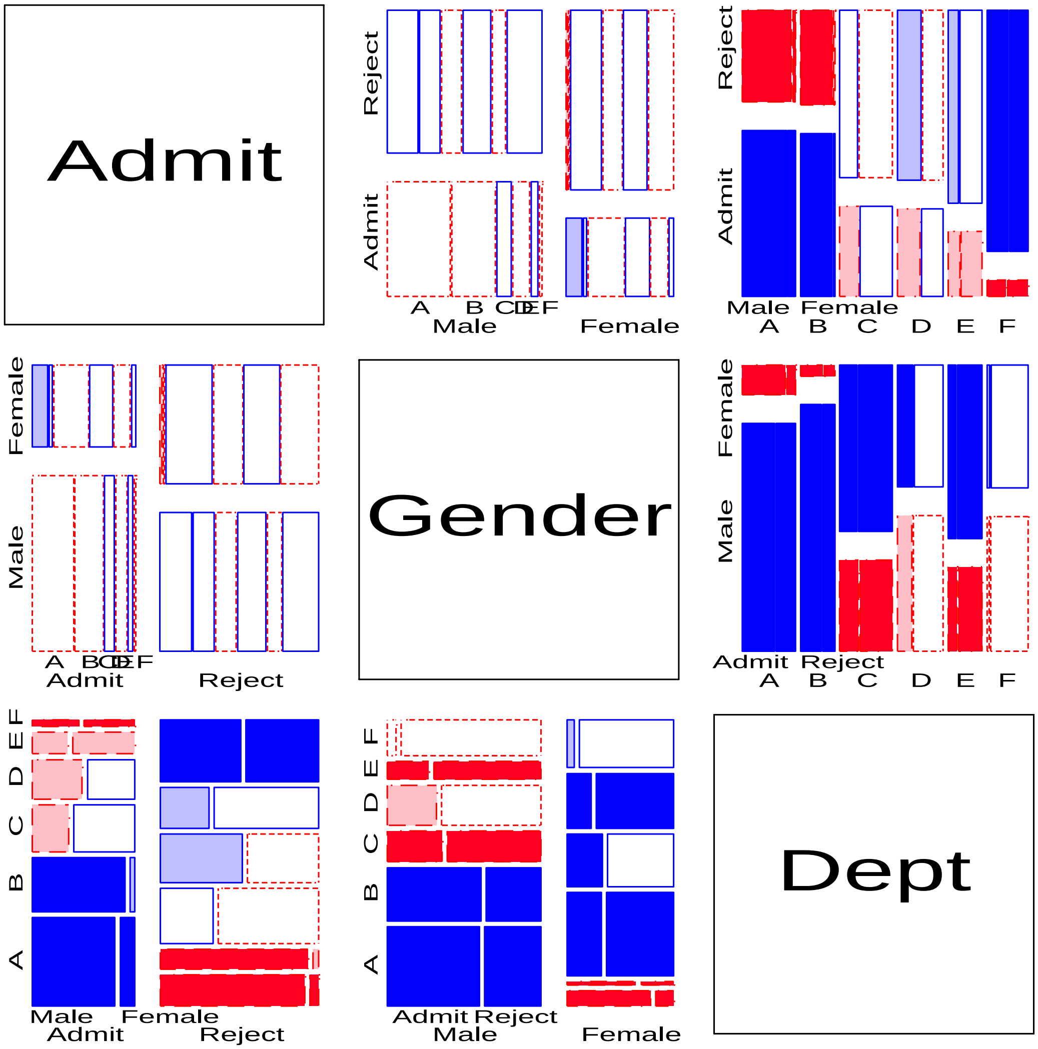

Figure 18: Conditional mosaic matrix of Berkeley admissions. Each panel shows the conditional relation, fitting a model of conditional independence model between the row and column variable, controlling for other variable(s)

can be shown as a set of boxplots for each level of the categorical variable. For conditional plots, we fit a model predicting the row variable from the column variable, partialing out (or conditioning on) all other variables from each. For quantitative variables, this is just the partial regression plot.

Finally, an analog of the coplot for categorical data is an array of plots of the dependence among two or more variables, conditioned (or stratified) by the values of one or more *given* variables. Each such panel then shows the *partial* associations among the foreground variables; the collection of such plots show how these as the given variables vary. Figure 8 and Figure 13 are two examples of this idea, using the fourfold display to represent the association in 2 × 2 tables.

Figure 19 and Figure 20 show two further examples, using the mosaic display to show the partial relations [Admit][Dept] given Gender, and [Admit][Gender] given Dept, respectively. Figure 20 shows

tl t.
A

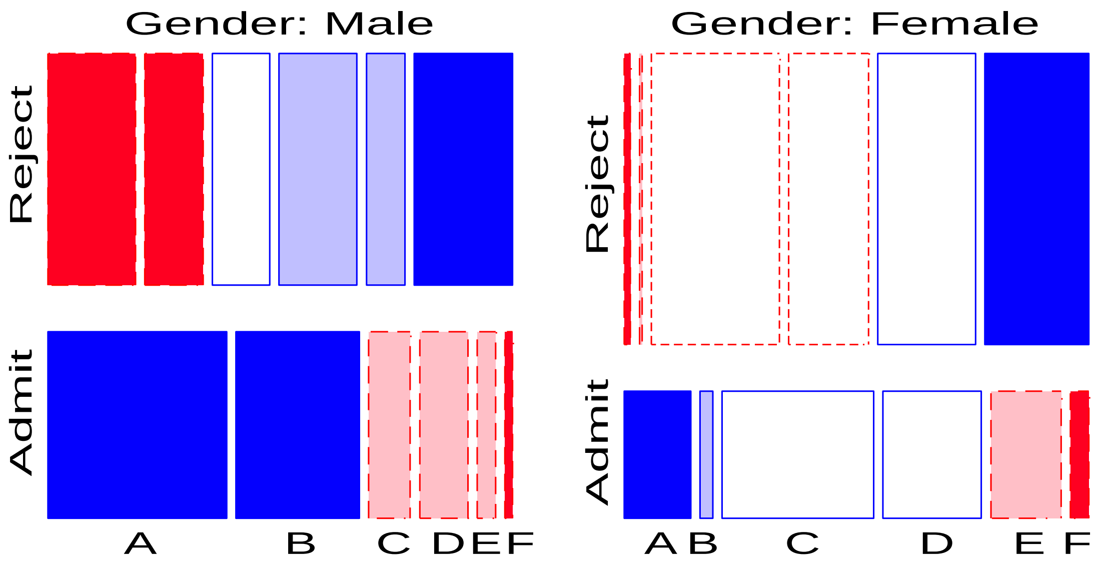

Figure 19: Mosaic coplot of Berkeley admissions, given Gender. Each panel shows the partial relation, fitting a model of independence model between Admission and Department.

Figure 19 shows that there is a very strong association between Admission and Department—different rates of admission, but also shows two things not seen in other displays: First, the *pattern* of association is qualitatively similar for both men and women; second the association is quantitatively stronger for men than women—larger differences in admission rates across departments.

About the Author

Michael Friendly is Associate Professor of Psychology and Coordinator of the Statistical Consulting Service at York University. He is an Associate Editor of the *Journal of Computational and Graphical Statistics*, and has been working on the development of graphical methods for categorical data. For further information, see `http://www.math.yorku.ca/SCS/friendly.html`, which contains links to most of the software developed for the graphical methods described here. There are also Web-based interfaces to sieve diagrams and mosaic displays accessible at `http://www.math.yorku.ca/SCS/Online/`.

Preparation of this chapter was aided by the National Sciences and Engineering Research Council of Canada, Grant OGP0138748.

References

Bickel, P. J., Hammel, J. W., and O'Connell, J. W. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187:398–403, 1975. 11

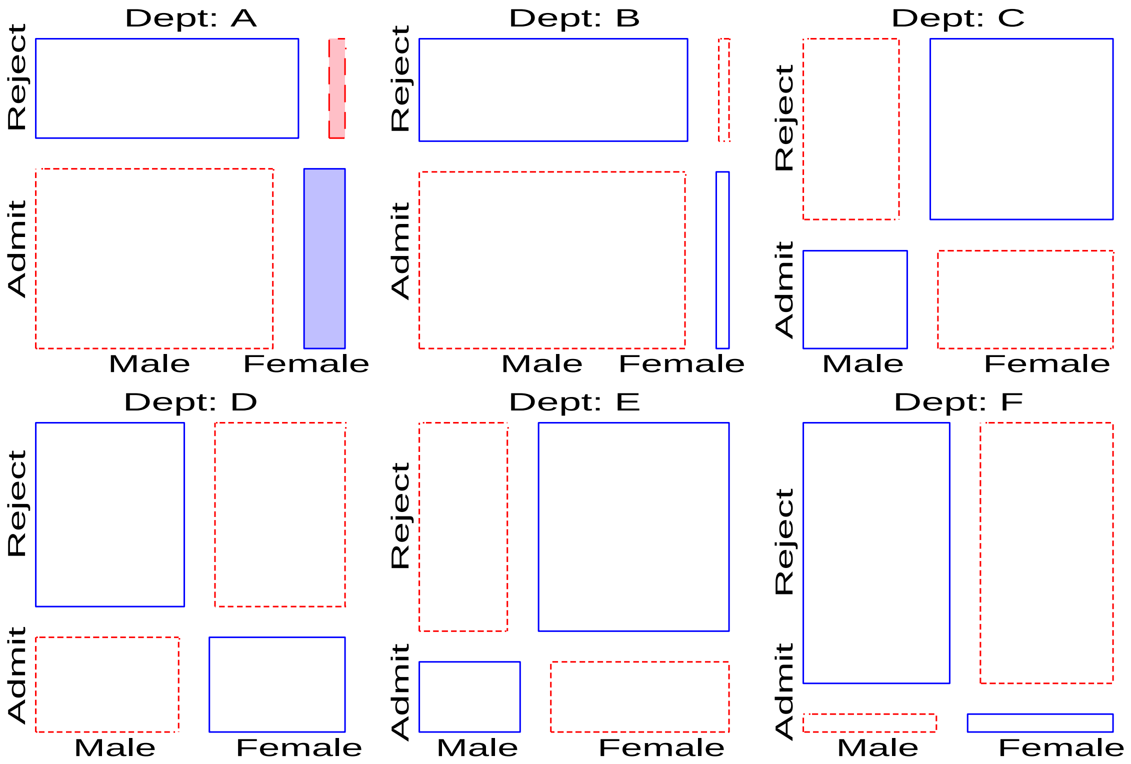

Figure 20: Mosaic coplot of Berkeley admissions, given Department Each panel shows the partial relation, fitting a model of independence model between Admission and Gender.

Carr, D. Color perception, the importance of gray and residuals, on a chloropleth map. *Statistical Computing & Statistical Graphics Newsletter*, 5(1):16–21, 1994. 10, 11

Carr, D. and Olsen, A. Simplifying visual appearance by sorting: An example using 159 AVHRR classes. *Statistical Computing & Statistical Graphics Newsletter*, 7(1):10–16, 1996. 22

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. *Graphical Methods for Data Analysis*. Wadsworth, Belmont, CA, 1983. 22

Cleveland, W. S. A model for studying display methods of statistical graphics. *Journal of Computational and Statistical Graphics*, 2:323–343, 1993a. 3

Cleveland, W. S. *Visualizing Data*. Hobart Press, Summit, NJ, 1993b. 3, 10, 22, 23

Cleveland, W. S. and McGill, R. Graphical perception: Theory, experimentation and application to the development of graphical methods. *Journal of the American Statistical Association*, 79:531–554, 1984. 3

Cleveland, W. S. and McGill, R. Graphical perception and graphical methods for analyzing scientific data. *Science*, 229:828–833, 1985. 3

de Falguerolles, A., Fredrich, F., and Sawitzki, G. A tribute to J. Bertin's graphical data analysis. In Bandilla, W. and Faulbaum, F., editors, *SoftStat '97, Advances in Statistical Software*, pages 11–20. Lucius and Lucius, Stutgart, 1997. 22

Fienberg, S. E. Perspective canada as a social report. *Social Indicators Research*, 2:153–174, 1975. 11

Friendly, M. *SAS System for Statistical Graphics*. SAS Institute Inc, Cary, NC, 1st edition, 1991. 3

Friendly, M. Mosaic displays for loglinear models. In *ASA, Proceedings of the Statistical Graphics Section*, pages 61–68, Alexandria, VA, 1992. 5, 7

Friendly, M. A fourfold display for 2 by 2 by K tables. Technical Report 217, York University, Psychology Dept, 1994a. 11

Friendly, M. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89:190–200, 1994b. 5, 7, 8, 22

Friendly, M. SAS/IML graphics for fourfold displays. *Observations*, 3(4):47–56, 1994c. 11

Gabriel, K. R. Biplot. In Johnson, N. L. and Kotz, S., editors, *Encyclopedia of Statistical Sciences*, volume 1, pages 263–271. John Wiley and Sons, New York, 1980. 22

Gabriel, K. R. Biplot display of multivariate matrices for inspection of data and diagnosis. In Barnett, V., editor, *Interpreting Multivariate Data*, chapter 8, pages 147–173. John Wiley and Sons, London, 1981. 22

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, NY, 1981. 5

Hartigan, J. A. and Kleiner, B. A mosaic of television ratings. *The American Statistician*, 38:32–35, 1984. 5

Kosslyn, S. M. Graphics and human information processing: A review of five books. *Journal of the American Statistical Association*, 80:499–512, 1985. 3

Kosslyn, S. M. Understanding charts and graphs. *Applied Cognitive Psychology*, 3:185–225, 1989. 3

Lewandowsky, S. and Spence, I. The perception of statistical graphs. *Sociological Methods & Research*, pages 200–242, 1989. Beverly Hills, CA: Sage Publications. 3

Mullis, I. V. S., Dossey, J. A., Owen, E. H., and Phillips, G. W. NAEP 1992: Mathematics report card for the nation and the states. Technical Report 23-ST02, National Center for Education Statistics, Washington, D. C., 1993. 14

Riedwyl, H. and Schüpbach, M. Siebdiagramme: Graphische darstellung von kontingenztafeln. Technical Report 12, Institute for Mathematical Statistics, University of Bern, Bern, Switzerland., 1983. 5

Riedwyl, H. and Schüpbach, M. Parquet diagram to plot contingency tables. In Faulbaum, F., editor, *Softstat '93: Advances In Statistical Software*, pages 293–299. Gustav Fischer, New York, 1994. 5

Snee, R. D. Graphical display of two-way contingency tables. *The American Statistician*, 28:9–12, 1974. 5

Spence, I. Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance*, 16:683–692, 1990. 3

Tukey, J. W. Graphic comparisons of several linked aspects: Alternative and suggested principles. *Journal of Computational and Statistical Graphics*, 2(1):1–33, 1993. 14

Wainer, H. Tabular presentation. *Chance*, 6(3):52–56, 1993. 22

Wainer, H. Some multivariate displays for NAEP results. *Psychological Methods*, 2(1):34–63, 1997. 14