Visualizing Categorical Data: Data, Stories, and Pictures

Article · April 2000

Source: CiteSeer

CITATIONS
23

READS
1,702

1 author:

Michael Friendly
York University
133 PUBLICATIONS **7,242** CITATIONS

SEE PROFILE

Visualizing Categorical Data: Data, Stories, and Pictures

Michael Friendly
York University, friendly@yorku.ca

Abstract

Categorical data—frequency data, and discrete data—are most often presented in tables, and analyses using loglinear models and logistic regression are most often presented in terms of parameter estimates. Over the past decade, I and others have developed novel visualization methods for categorical data, designed to provide exploratory and confirmatory graphic displays analogous to those used readily and easily for quantitative data. These graphical methods are described in *Visualizing Categorical Data*. The book also provides a large collection of macros designed to make these methods readily and easily used. This paper provides an overview of these graphical methods and macros, as told through data, their stories, and associated graphical displays.

KEYWORDS: categorical data, graphics, mosaic displays, mosaic matrices, correspondence analysis, loglinear models, logistic regression.

1 Introduction

Over the last decade a modest revolution has been brewing in the analysis of categorical data, as graphical methods and techniques of data visualization, so commonly used for quantitative data, have begun to be developed for frequency data and discrete data.

At SUGI 17 (Friendly, 1992a) I described some initial steps in the development of new graphical methods for categorical data, with the goals of (a) providing visualization techniques for data exploration and model fitting comparable in scope to those used for quantitative data, and (b) implementing these methods in readily available software. These goals have now been largely achieved. The methods are described and illustrated in a new book, *Visualizing Categorical Data* (*VCD*), now in production. The book includes nearly 40 general macros and programs (see Appendix A), covering most aspects of categorical data analysis.

This paper provides an overview of some of these graphical methods and macros, using examples from the book, as told through data, their stories, and associated graphical displays. (Most of the graphs are in color; see the CD version of the Proceedings.)

2 Disputed authorship: The Federalist Papers

In 1787–88, Alexander Hamilton, John Jay, and James Madison wrote a series of newspaper essays to persuade the voters of New York State to ratify the U.S. constitution. The essays were titled *The Federalist Papers* and all were signed with a pseudonym. Of the 77 papers published, the author(s) of 65 are known, but *both* Hamilton and Madison later claimed sole authorship of the remaining 12. Mosteller and Wallace (1984) investigated the use of statistical methods to identify authors of disputed works based on the frequency distributions of certain key function words, and concluded that Madison had indeed authored the 12 disputed papers.

Table 1 shows the distribution of the occurrence of one of these "marker" words, the word *may* in 262 blocks of text (each about 200 words long) from issues of the *Federalist Papers* and other essays known to be written by James Madison.

An important part of the analysis by Mosteller and Wallace was to establish the theoretical form of these frequency distributions, so that the known works could be compared in terms of estimated paramters, rather than through the entire distributions. A simple argument for the occurrence of rare events leads to a suggestion that the distribution of such words might be Poisson; however, numerical fitting led to the conclusion that the Negative Binomial gave better fits.

We concentrate here on visualization methods to determine the theoretical form of a discrete distribution.

Table 1: Number of occurrences (k) and number of blocks of text (n_k) of the word *may* in Federalist Papers and essays written by James Madison

k	0	1	2	3	4	5	6
n_k	156	63	29	8	4	1	1

2.1 Hanging rootograms

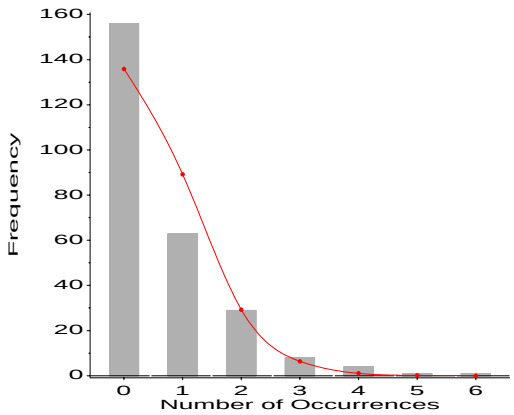

Figure 1: Histogram for Madison data, with Poisson fit

Discrete frequency distributions are often graphed as histograms, with a theoretical fitted distribution superimposed. Figure 1, for example, shows the data in Table 1 together with the fitted frequencies under a Poisson model. It is hard to compare the observed and fitted frequencies visually, because (a) we must assess deviations against a curvilinear relation, and (b) the largest frequencies dominate the display.

The hanging rootogram (Tukey, 1977) solves these problems by (a) shifting the histogram bars to coincide with the fitted curve, so that deviations may be judged by deviations from a horizontal line, and (b) plotting on a square-root scale, so that smaller frequencies are emphasized. Figure 2 shows more clearly that the observed frequencies differ systematically from those predicted under a Poisson model. In *VCD*, several macros are presented for fitting a variety

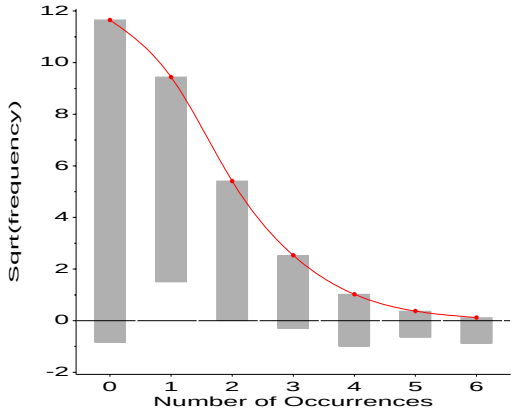

Figure 2: Suspended rootogram for Madison data

of discrete distributions. The `GOODFIT` macro carries out goodness-of-fit tests; the `ROOTGRAM` macro provides a variety of displays including those of Figure 1 and 2. For example, Figure 2 is produced as

```
%goodfit(data=madison, var=count, freq=blocks,
  dist=poisson, out=fit);
%rootgram(data=fit, var=count, obs=blocks);
```

2.2 Ord plots

A simple plot suggested by Ord (1967) may be used to diagnose the form of a discrete distribution. Ord showed that, for each of the Poisson, Binomial, Negative Binomial, and Logarithmic Series distributions, a plot of kp_k/p_{k-1} against k is linear, and these distributions were distinguished by the signs of the slope and intercept.

Figure 3 shows the Ord plot for the Madison data, which diagnoses the distribution as a Negative Binomial, based on the positive slope of the thicker line (found by weighted least squares). This plot is produced using the `ORDPLOT` macro, used as

```
%ordplot(data=madison, count=Count, freq=blocks);
```

2.3 Robust distribution plots

One disadvantage of the Ord plot is lack of resistance, since a single discrepant frequency, n_k, affects the points for both k and $k+1$. Robust distribution plots, following methods described by Hoaglin and Tukey (1985), are provided by the `DISTPLOT` macro.

Figure 4 shows the Negative Binomial distribution plot, produced using the `DISTPLOT` macro, as follows:

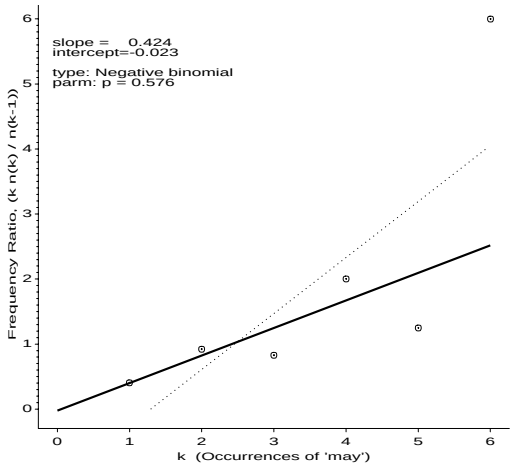

Figure 3: Ord plot for Madison data

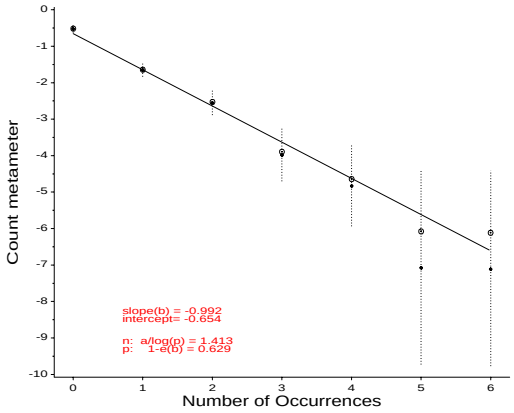

Figure 4: Robust distribution plot for Madison data for the negative binomial

```
%distplot(data=madison,count=count,freq=blocks,dist=negbin);
```

This plot has the property that the circled points are linear in k when the data follow the assumed distribution, as in the Ord plot. However, the ordinate "count metameter" depends only on n_k, and the confidence bars are calculated to take into account the variability of individual counts, n_k, in the observed distribution.

3 Gender bias in admission to Berkelely?

Bickel et al. (1975) analyzed data on admissions to graduate depatments at U. C. Berkeley in 1973. Aggregate data for the six largest departments are shown in Table 2, classified by admission and gender. The issue was whether these data showed evidence of gender bias in admissions.

Table 2: Admissions to Berkeley graduate programs

	Admitted	Rejected	Total
Males	1198	1493	2691
Females	557	1278	1855
Total	1755	2771	4526

3.1 Fourfold displays

Table 2 is an example of a 2×2 table. For such data, the *odds ratio*, $\theta = n_{11}n_{22}/n_{12}n_{21}$, is a natural measure of the strength of association between the two variables.

The ***fourfold display*** depicts these frequencies by quarter circles, whose radius is proportional to $\sqrt{n_{ij}}$, so the area is proportional to the cell count (Fienberg, 1975, Friendly, 1994a,c). The cell frequencies are usually scaled to equate the marginal totals, and so that the ratio of diagonally opposite segments depicts the odds ratio. Confidence rings for the observed θ allow a visual test of the hypothesis $H_0 : \theta = 1$ corresponding to no association. They have the property that the rings for adjacent quadrants overlap *iff* the observed counts are consistent with the null hypothesis.

Figure 5 shows the aggregate data from Table 2. The sample odds ratio, Odds (Admit|Male) / (Admit|Female) is 1.84 indicating that males were almost twice as likely to be admitted. The confidence rings in the figure do not overlap, showing that this association is highly significant. Does this constitute evidence for gender bias in admission?

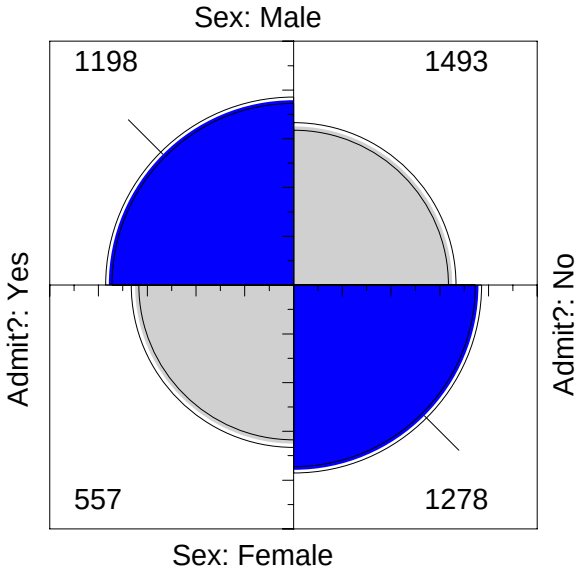

Figure 5: Fourfold display for Berkeley admissions data, margins equated

The admissions data shown in Figure 5 came from the six largest at Berkeley. To determine the source of the apparent sex bias in favor of males, we make a new plot, Figure 6, stratified by department.

Surprisingly, Figure 6 shows that, for five of the six departments, the odds of admission is approximately the same for both men and women applicants. Department A appears to differs from the others, with women approximately 2.86 ($= (313/19)/(512/89)$) times as likely to gain admission.

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 happen to 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.

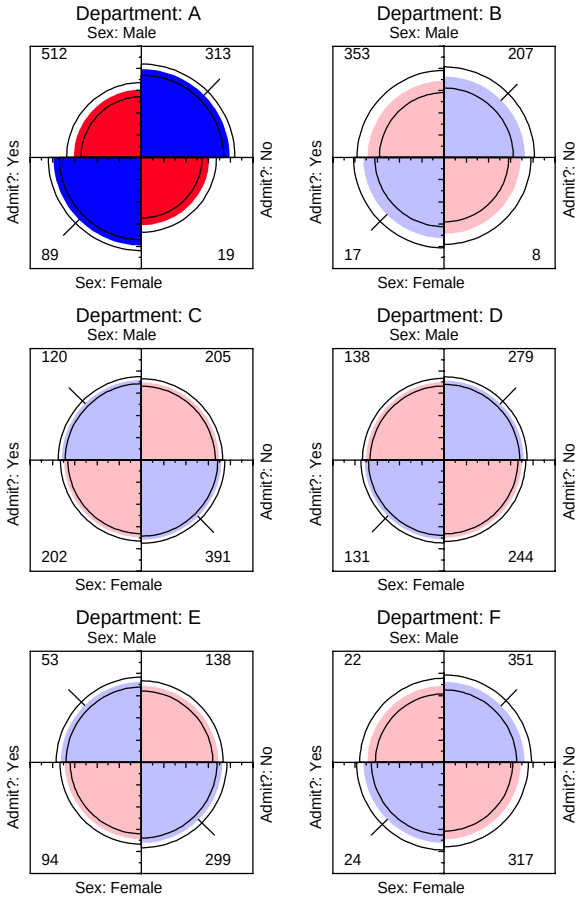

Figure 6: Fourfold display for Berkeley admissions data, by department

3.2 Mosaic displays

The ***mosaic display*** (Friendly, 1992b, 1994b, 1999, Hartigan and Kleiner, 1981) is a graphical method for visualizing an n-way contingency table and for building models to account for the associations among its variables. The frequencies in a contingency table are portrayed as a collection of rectangular "tiles" whose areas are proportional to the cell frequencies; the areas are colored and shaded to portray the residuals from a specified log-linear model.

Whereas goodness-of-fit statistics provide an overall summary of how well a model fits the data, the mosaic display reveals the pattern of lack of fit, and helps suggest an alternative model that may fit better.

The hypothesis that gender and admission are independent, *given* department, corresponds to the loglinear model [Admit Dept] [Gender Dept]. This model fits poorly ($G^2(6) = 21.74$), but the residuals in the mosaic (Figure 7) suggest that the lack of fit is due primarily to department A, where a *greater* proportion of women are admitted than men, as may also be seen in Figure 6.

3.3 Plots for logit models

Loglinear models treat all variables symmetrically, and do not distinguish between explanatory and response variables. When one variable can be regarded as a response variable, then the effects of the other variables may be expressed as an equivalent logit model.

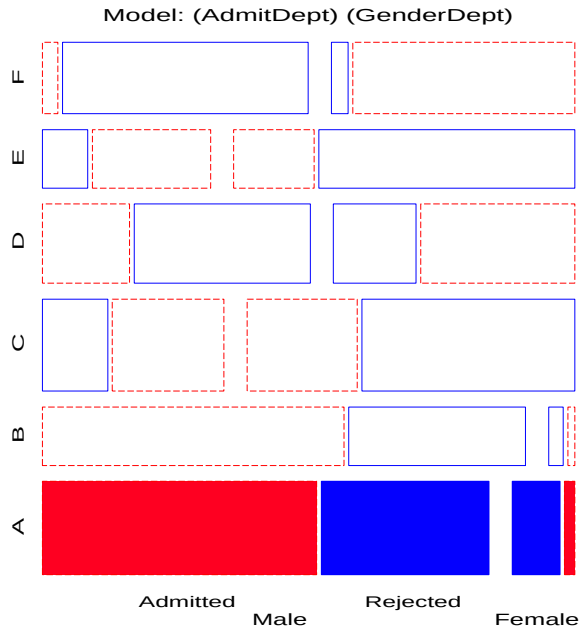

Figure 7: Three-way mosaic plot for Berkeley data: Conditional independence

For example, Figure 7 suggests a loglinear model which allows an association between admission and gender in Department A only,

$$\log\ m_{ijk} = \mu + \lambda_i^A + \lambda_j^D + \lambda_k^G + \lambda_{ij}^{AD} + \lambda_{jk}^{DG} + \delta_{j=1}\lambda_{ik}^{AG}\ , \quad (1)$$

where $\delta_{j=1}$ equals 1 for Department A ($j = 1$) and is zero otherwise. This model asserts that Admission and Gender are conditionally independent, given Department, except in Department A. It has one more parameter than the conditional independence model, $[AD][GD]$.

The loglinear model (1) has an equivalent logit formulation,

$$L_{ij} = \alpha + \beta_i^{\text{Dept}} + \delta_{j=1}\beta^{\text{Gender}}\ , \quad (2)$$

where $L_{ij} = \log(m_{ij1}/m_{ij2})$ is the log odds of admission for males as vs. females, β_i^{Dept} is the effect on admissions over departments, and $\delta_{j=1}\beta^{\text{Gender}}$ is the effect of gender in Dept. A. This model fits well, as shown in Figure 8.

Logit models such as (2) are easily fit with `PROC CATMOD`. Figure 8 is produced from the output dataset produced by this procedure, using the `CATPLOT` macro:

```
data berkeley;
   set berkeley;
   dept1AG = (gender='F') * (dept=1);
proc catmod order=data data=berkeley;
   weight freq;
   population dept gender;
   direct dept1AG;
   response / out=predict;
   model admit = dept dept1AG / ml noiter noprofile ;
%catplot(data=predict, xc=dept, class=gender,
    type=FUNCTION, z=1.96, legend=legend1);
```

Such graphs often provide a clearer interpretation of a fitted model than can be obtained from parameter estimates.

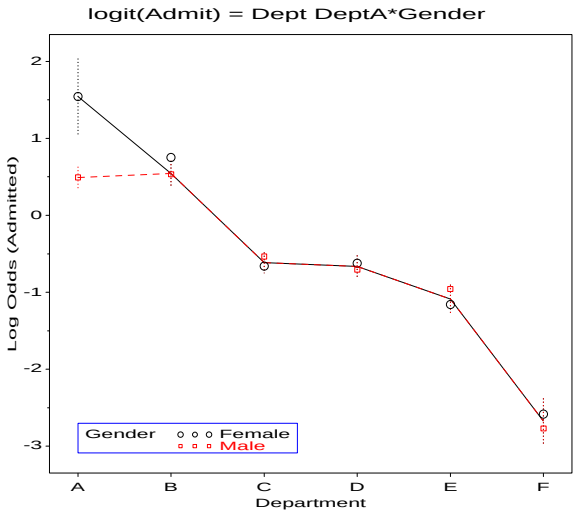

Figure 8: Observed and fitted logits for model (2)

4 The *Titanic* Story

There have been few marine disasters resulting in the staggering loss of life which occurred in the sinking of the *Titanic* on April 15, 1912 and (perhaps as a result) few that are so widely known by the public. There are two parts to the *Titanic* story. The first is concerns the analysis of survival of the passengers and crew; the second concerns data on the loading of the lifeboats.

4.1 Survival on the *Titanic*

Given the interest in the sinking of the *Titanic*, it is somewhat surprising that neither the exact death toll from this disaster nor the distributions of death among the passengers and crew are universally agreed. Dawson (1995, Table 2) presents the cross-classification of 2201 passengers and crew on the *Titanic* by Age, Gender, Class (1st, 2nd, 3rd, Crew) shown in Table 3 and describes his efforts to reconcile various historical sources. Let us see what we can learn from this dataset.

Table 3: Survival on the Titanic

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

Figure 9 shows the frequencies of the background variables, Class, Gender and Age by the sizes of the boxes. It also shows the association between Age and Class–Gender combinations by shading. There were no children among the crew, and the overall proportion of children was quite small (about 5 %). But among the

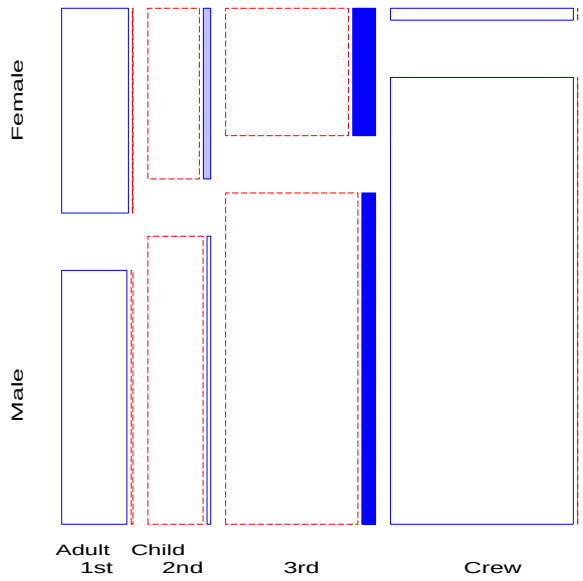

Figure 9: Titanic data, background variables

passengers, the proportion of children increases from first class to third class. The large positive residuals for children among the 3rd class passengers likely represents families traveling or emmigrating together.

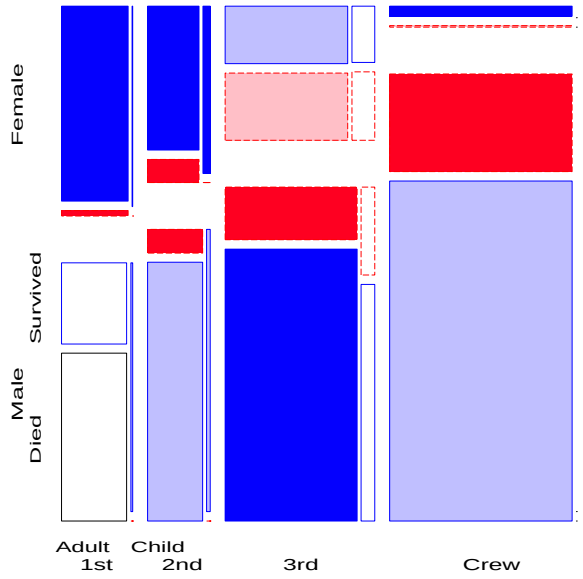

Figure 10: Titanic data, Joint independence: Survival $\perp$ {Class, Gender, Age}

Figure 10 shows an initial four-way mosaic for the full table, and fits the model $[CGA][S]$ which asserts that survival is independent of Class, Gender, and Age jointly. This is the minimal null model when the first three variables are explanatory. It is clear that greater proportions of women survived than men in all classes, but with greater proportions of women surviving in the upper two classes. Among males, the proportion who survived also increases with economic class. However, this model fits very poorly ($G^2(15) = 671.96$), and we may try to fit a more adequate model by adding associations between survival and the explanatory variables.

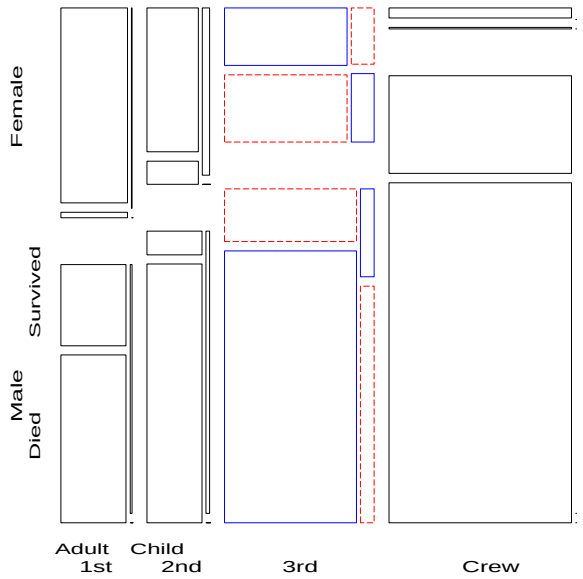

Figure 11: Titanic data, Model $[CGA][CGS][CAS]$

The rubric "women and children first" implies the model $[CGA][CS][GAS]$ in which Age and Gender interact in their influence on survival (independent of Class), but this model fits poorly ($G^2(9) = 94.54$). A more adequate model adds interactions of Class with *both* Age and Gender to give the model $[CGA][CGS][CAS]$, whose residuals are shown in Figure 11. The likelihood-ratio chi-square is now 1.69 with 4 df—a very good fit, indeed.

The import of these figures is clear. Regardless of Age and Gender, lower economic status was associated with increased mortality. But the differences due to Class were moderated by both Age and Gender. Although women on the *Titanic* were more likely overall to survive than men, women in 3rd class did not have a significant advantage, while men in 1st class did compared to men in other classes. Hence, although the phrase "women and children first" is mellifluous and appeals to a sense of Edwardian chivalry a more adequate description might be "women and children (according to class), then 1st class men."

4.2 Lifeboats on the *Titanic*

After the disaster, the British Board of Trade launched several inquiries, the most comprehensive of which resulted in the *Report on the Loss of the "Titanic" (S.S.)* by Lord Mersey (Mersey, 1912). Section 4 of this document contains a detailed account of the saving and rescue of the passengers and crew who survived. The report lists the time of launch and composition of the 18 boats (out of 20) actually launched, classified as "male passengers", "women and children", and "men of crew", as reported by witnesses.

Trilinear plots

Trilinear plots are quite useful for showing the relative proportions in each row of $n \times 3$ tables. Figure 12 shows the proportions of these three categories, classed by the side of the ship from which

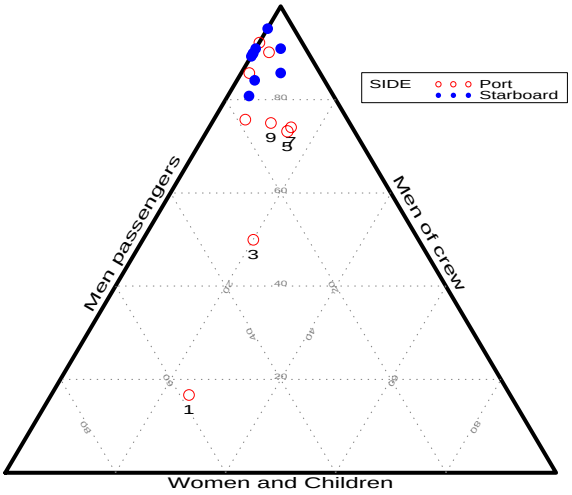

Figure 12: Lifeboats on the Titanic, trilinear plot

the lifeboat was launched. Boats with more than 10% male passengers are identified by number. The graph strongly suggests that the procedures for loading the lifeboats may have differed for the port and starboard side of the ship.

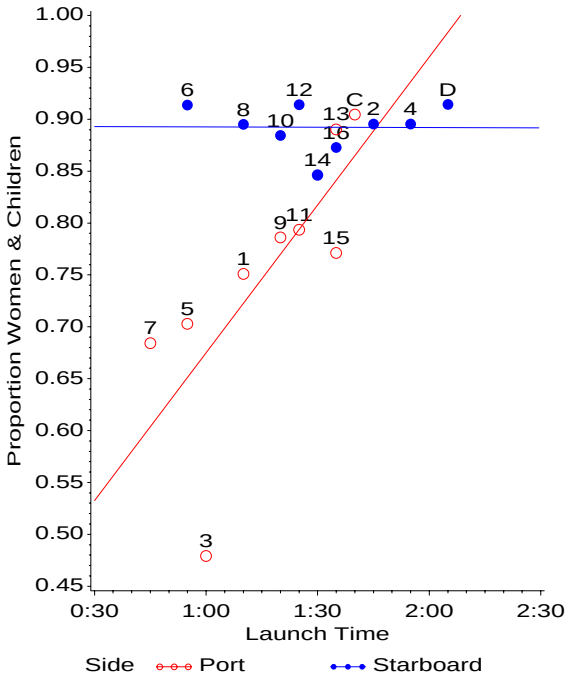

Figure 13: Lifeboats on the Titanic, logistic regression

Logistic regression

Figure 12 suggested a logistic regression model for the proportion of women and children on the lifeboats, using time of launch and side of boat as predictors. Graphical analysis led to a good-fitting model with separate slopes and intercepts for the port and starboard sides, with observed and fitted probabilities shown in Figure 13.

This graph (and others not shown here) bear eloquent witness to the suggestion that the regimes for loading the lifeboats differed substantially between the port and starboard side. On the starboard side, discipline and order were quickly established, women and children got consistent preference, and lifeboats were loaded close to their capacity. Loading on the port side, however, began with chaos, and general lack of effective control. The first few boats were only lightly loaded, and contained large numbers of men and crew; presumably whoever was nearby got on. The situation was brought under control over time. But alas, time ran out for the passengers and crew of the *Titanic*.

5 Sex and the Married Woman

A study of divorce patterns by Thornes and Collard (1979) analyzed two samples of about 500 people each, one still married, and another who had petitioned for divorce, giving the 2^4 table shown in Table 4. Each person was asked (a) whether they had made love with anyone else *before* their marriage, and (b) whether they had any sexual encounters with another person *after* marriage.

Table 4: Marital Status in Relation to Gender and Reported Premarital and Extramarital Sex

Extramarital Sex	Premarital Sex	Gender	**Marital Status** Divorced	Married
Yes	Yes	Women	17	4
No			54	25
Yes	No		36	4
No			214	322
Yes	Yes	Men	28	11
No			60	42
Yes	No		17	4
No			68	130
Total			494	542

5.1 Mosaic matrices

The ***mosaic matrix*** is a discrete analog for multivariate categorical data of the scatterplot matrix (Friendly, 1999). Like the scatterplot matrix, it contains all $p(p-1)$ pairwise plots for a p-variate dataset, but displays the relation of each pair of variables by a mosaic. Extensions of this idea include: (a) a conditional mosaic matrix, which fits a model of conditional independence between each row and column, controlling for one or more of the other variables—a generalization of partial regression plots, (b) mosaic displays of partial association, stratified by one or more variables—a discrete analog of coplots or Trellis displays.

Figure 14 shows the bivariate marginal relations among all pairs of variables in the marital status data, produced with the `MOSMAT` macro, as follows:

```
%include catdata(marital);
%mosmat(data=marital, var=Gender Pre Extra Marital,
   vorder=Marital Extra Pre Gender, devtype=LR ADJ);
```

Viewing Gender, Premarital sex and Extramarital sex as explanatory, and Marital status as the response, the mosaics in row 1 (and in column 1) shows how marital status depends on each predictor marginally. The remaining panels show the relations within the set of explanatory variables.

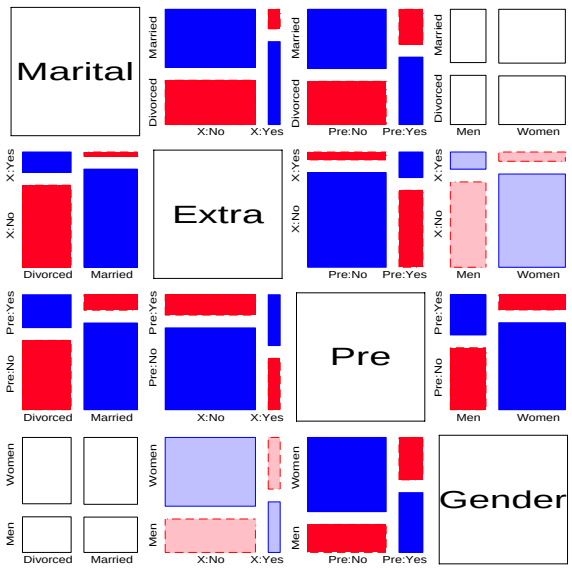

Figure 14: Mosaic matrix for marital status data. Each panel shows the bivariate marginal association.

Thus we see (row 1, column 4) that marital status is independent of gender, by design of the data collection. In the (1, 3) panel, we see that reported premarital sex is more often followed by divorce, while non-report is more prevalent among those still married. The (1, 2) panel shows a similar, but stronger relation between extramarital sex and marriage stability. These effects pertain to the associations of P and E with marital status—the terms [PM] and [EM] in a loglinear model.

Among the background variables, the (2, 3) panel shows a strong relation between premarital sex and subsequent extramarital sex, while the (2, 4) and (3, 4) panels show that men are far more likely to report premarital sex than women in this sample, and also more likely to report extramarital sex.

5.2 Correspondence analysis

Correspondence analysis is an analog of principal components analysis for frequency data, designed to display the association among categorical variables in a small number of dimensions, designed to account for the largest proportion of the Pearson χ^2. Multiple correspondence analysis extends this method to *n*-way tables, but displays only bivariate associations, analogous to the (marginal) mosaic matrix.

Figure 15 shows the 2D MCA solution for the marital status data. This graph was prepared by the `CORRESP` macro as follows:

```
%corresp(data=marital, tables=gender pre extra marital,
   weight=freq, options=mca, interp=vec, inc=1, pos=-,
   symbols=dot);
```

From the relations among the points we see that men and women who have reported premarital sex are far more likely to report extramarital sex than those who have not. (In the marginal [GP] [E] table, the conditional odds ratio of extramarital sex is 3.61 for men and 3.56 for women. Thus, extramarital sex depends on premarital sex, but not on gender.)

Figure 16 shows the 4-way mosaic with residuals for the model [GPE] [M], which asserts that marital status is independent of Gender, Premarital sex, and Extramarital sex jointly. From the pattern of residuals, we may see that among those reporting no premarital sex (bottom part of Figure 16), there is a similar pattern of cell sizes and deviations for marital status in relation to gender and extramarital sex: People who did not report premarital sexual experience are more likely to remain married if they report no extramarital sex and more likely to be divorced if they did. Among those who do report premarital sex (top part of Figure 16), there is also a similar pattern of sign of deviations, positive for those who are divorced, negative for those who are married.

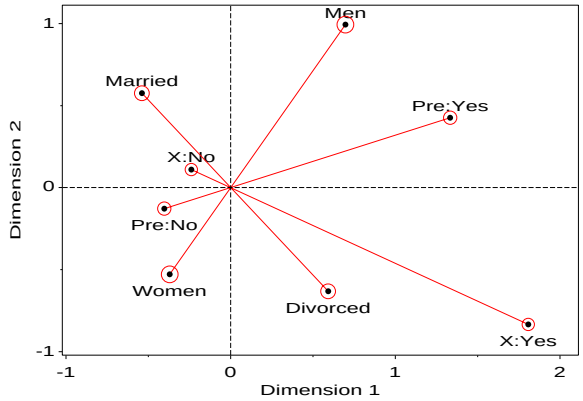

Figure 15: 2D multiple correspondence analysis display for marital status data

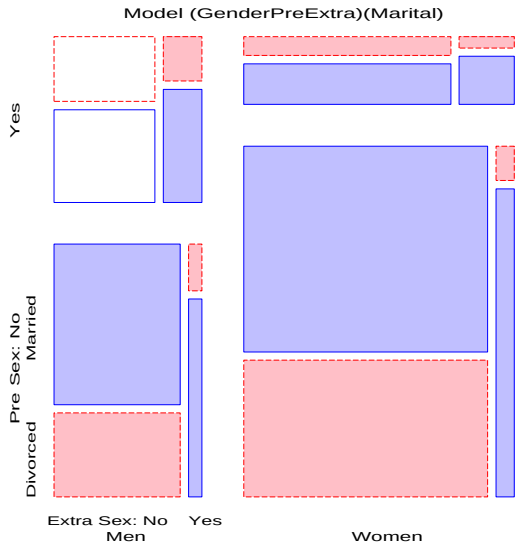

Figure 16: Four-way mosaic for the model [GPE] [M]

The bottom line on these analyses is, if you're going to fool around, do it early; if you didn't fool around early, don't do it later, if you want to stay married.

6 The *Challenger* Disaster

The space shuttle *Challenger* exploded 31 seconds after take-off on January 28, 1986. Subsequent investigation determined that the cause was failure of the O-ring seals used to isolate the fuel sup-

ply from burning gases. The story behind the *Challenger* disaster is perhaps the most poignant missed opportunity in the history of statistical graphics. It may be heartbreaking to find out that some important information was there, but the graph maker missed it.

Engineers from Morton Thiokol, manufacturers of the rocket motors, had been worried about the effects of unseasonably cold weather on the O-ring seals and recommended aborting the flight. NASA staff analysed the data on the relation between ambient temperature and the number of O-ring failures (out of 6), but they had excluded observations where no O-rings failed, believing that they were uninformative. Figure 17 shows a graph which led to this conclusion, perhaps the most misleading graph in history!

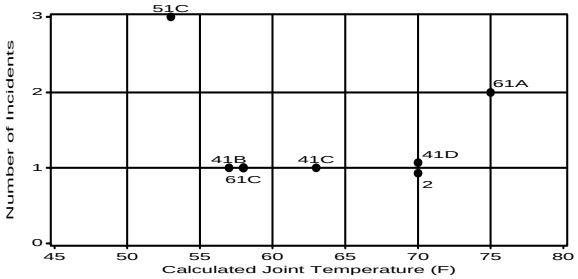

Figure 17: NASA Space shuttle, pre-launch graph

Unfortunately, the 0-failure observations had occurred when the launch temperature was relatively warm ($65 - 80°$F) and were indeed informative. The coldest temperature at any previous launch was $53°$; when *Challenger* was launched on January 28, the temperature was a frigid $31°$.

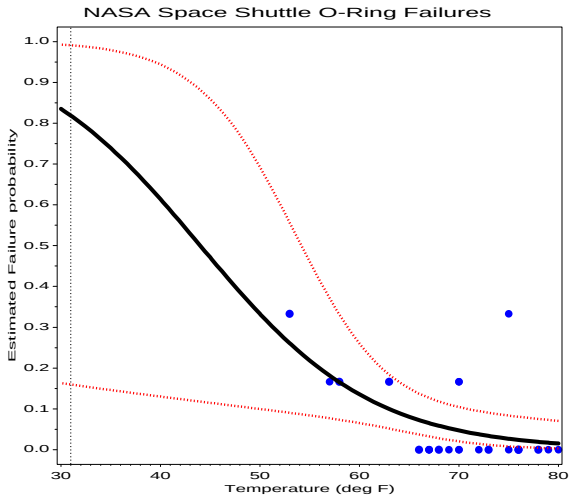

Figure 18: NASA Space Shuttle O-ring Failure, Observed and Predicted probabilities

Figure 18 shows observed and predicted failure probabilities from a logistic regression model, together with a confidence band for the predictions. There's hardly any data at low temperatures, and the width of the gives a visual cue to this uncertainty. Nevertheless, the failure probabilities are uncomfortably high at low temperatures. A graph like this might have led to a different decision about the launch of the *Challenger*.

7 The Donner Party

In April–May of 1846, three years before the gold rush, the Donner and Reed families set out for California from the American midwest in a wagon train. By mid July, a large group had reached a site in present-day Wyoming; George Donner was elected to lead what was to be called the "Donner Party," which eventually numbered 87 people in 23 wagons, along with their oxen, cattle, horses, and worldly posessions.

They were determined to reach California as quickly as possible. Lansford Hastings, a self-proclaimed trailblazer (retrospectively, of dubious distinction), proposed that the party follow him through a shorter path through the Wasatch Mountains. Their choice of "Hastings's Cutoff" proved disasterous: Hastings had never actually crossed that route himself, and the winter of of 1846 was to be one of the worst on record.

In October, 1846, heavy snow stranded them in the eastern Sierra Nevada, just to the east of a pass which bears their name today. The party made numerous attempts to seek rescue, most turned back by blizzard conditions. Relief parties in March–April 1847 rescued 40, but discovered grizzly evidence that those who survived had cannabalized those who died.

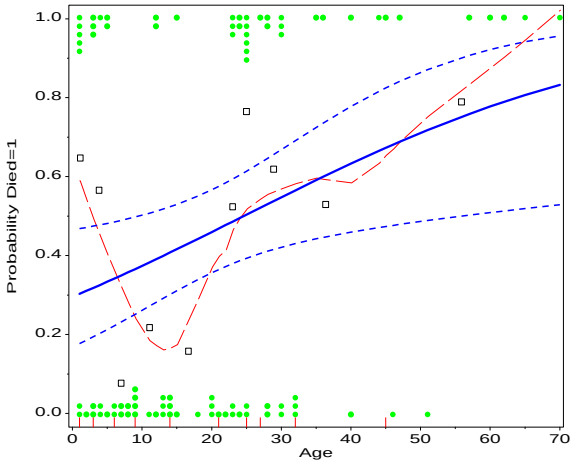

Figure 19: Donner Party, empirical logit probability plot

We examine here how survival in the Donner Party varied with Age and Gender. (The data were obtained from Kristin Johnson's Donner Party web site, `www.metrogourmet.com/crossroads/KJhome.htm`). At issue is whether a linear logistic model is satisfactory for these data. For such purposes, smoothing techniques are often crucial in visualizing the relation between a discrete response and predictors.

Figure 19 shows a plot of the observations (circles), and estimated probabilities of death (squares) vs. Age, based on grouping the ages into deciles, one of several plots produced by the `LOGODDS` macro. The thick solid line shows the estimated probability under a linear logistic model (with 95% prediction intervals). The dashed curve, produced using the `LOWESS` macro, suggests however that the relation with Age is quadratic: The very youngest and the oldest were most likely to perish.

Figure 20 shows the observations (women: filled circles; men: open circles) estimated probabilities under a quadratic model, $\Pr(\text{Death}) \sim \text{Age} + \text{Age}^2 + \text{Male}$. The statistical evidence for the term in Age^2 is equivocal (Wald $\chi^2 = 2.84, p = 0.09$; LR $G^2(1) = 4.40, p = 0.03$). The visual evidence from Figure 19–20

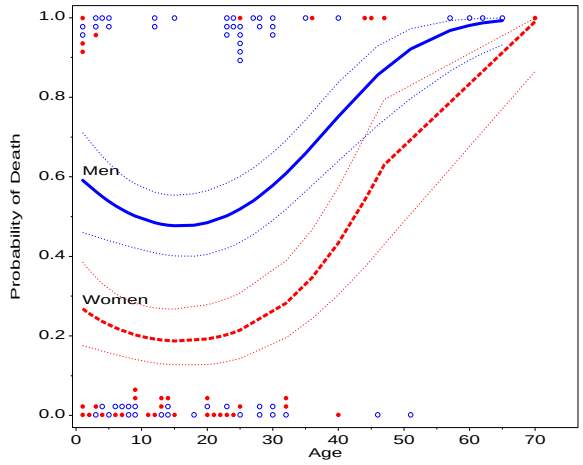

Figure 20: Donner Party, fitted logistic model, $\Pr(\text{Death}) \sim \text{Age} + \text{Age}^2 + \text{Male}$

(and other graphcs not shown here) in favor of the quadratic model is more compelling, and makes better sense—a linear model would predict greatest survival among the youngest members of the Donner Party.

As Yogi Berra said, "You can see a lot, just by looking." It is hoped that the tools and techniques described in *Visualizing Categorical Data* contributes to the greater use of graphical methods in the analysis of frequency and discrete data.

A Macros and Programs

The following macros and programs are described and illustrated in *VCD*. All require SAS/STAT and SAS/GRAPH; many require SAS/IML. They will be available on the web at `www.math.yorku.ca/SCS/vcd/`.

`ADDVAR`	Added variable plots for logistic regression
`AGREE`	Observer agreement chart (SAS/IML)
`BIPLOT`	Generalized biplot displays
`CATPLOT`	Plot results from `PROC CATMOD`
`CORRESP`	Plot `PROC CORRESP` results
`DISTPLOT`	Plots for discrete distributions
`DUMMY`	Create dummy variables
`FOURFOLD`	Fourfold displays for $2 \times 2 \times k$ tables (SAS/IML)
`GOODFIT`	Goodness-of-fit for discrete distributions
`HALFNORM`	Half-normal plots for generalized linear models
`INFLGLIM`	Influence plots for generalized linear models
`INFLOGIS`	Influence plots for logistic regression
`LAGS`	Calculate lagged frequencies for sequential analysis
`LOGODDS`	Plot empirical logits for binary data
`MOSAIC`	Mosaic displays (macro)
`MOSAICS`	SAS/IML modules for mosaic displays
`MOSMAT`	Mosaic matrices (macro)
`ORDPLOT`	Ord plot for discrete distributions
`PANELS`	Arrange multiple plots in a panelled display
`POISPLOT`	Poissonness plot
`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
`ROOTGRAM`	Hanging rootograms
`SIEVE`	Sieve diagrams (SAS/IML)
`SORT`	Sort a dataset by a statistic or formatted value
`TABLE`	Construct a grouped frequency table, with recoding
`TRIPLOT`	Trilinear plots for $n \times 3$ tables
`Utility`	Graphics utility macros: `BARS`, `EQUATE`, `GDISPLA`, `GENSYM`, `GSKIP`, `LABEL`, `POINTS`, `PSCALE`.

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.

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

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

Friendly, M. Graphical methods for categorical data. *Proceedings of the SAS User's Group International Conference*, 17:1367–1373, 1992a.

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

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

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

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

Friendly, M. Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Statistical Graphics*, 8:373–395, 1999.

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*, pp. 268–273. Springer-Verlag, New York, NY, 1981.

Hoaglin, D. C. and Tukey, J. W. Checking the shape of discrete distributions. In Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors, *Exploring Data Tables, Trends and Shapes*, chapter 9. John Wiley and Sons, New York, 1985.

Mersey, L. Report on the loss of the "Titanic" (S. S.). Parliamentary command paper 6352, 1912.

Mosteller, F. and Wallace, D. L. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Springer-Verlag, New York, NY, 1984.

Ord, J. K. Graphical methods for a class of discrete distributions. *Journal of the Royal Statistical Society, Series A*, 130:232–238, 1967.

Thornes, B. and Collard, J. *Who Divorces?* Routledge & Kegan, London, 1979.

Tukey, J. W. *Exploratory Data Analysis*. Addison Wesley, Reading, MA, 1977.