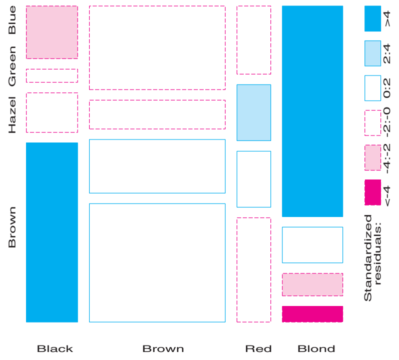
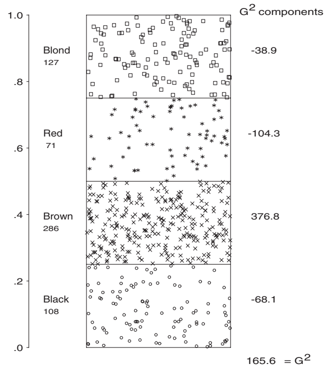
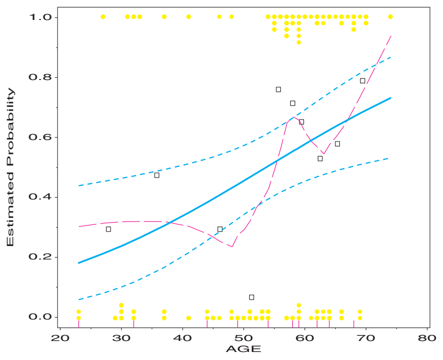
Chapter

1 Introduction







Categorical data consists of variables whose values comprise a set of discrete categories. Such data requires different statistical and graphical methods from those commonly used for quantitative data. The focus of this book is on visualization techniques and graphical methods designed to reveal patterns of relationships among categorical variables.

1.1 Data Visualization and Categorical Data 1
1.2 What Is Categorical Data? 2
1.3 Strategies for Categorical Data Analysis 5
1.4 Graphical Methods for Categorical Data 8
1.5 Visualization = Graphing + Fitting + Graphing 13

1.1 Data Visualization and Categorical Data

> Beauty is truth, truth beauty.—that is all
> Ye know on earth, and all ye need to know.
>
> John Keats, “Ode on a Grecian Urn”

“Data visualization” is an approach to data analysis that focuses on *insighful graphical display*. We can display the raw data, some summary statistics, or some indicators of the quality or adequacy of a fitted model. The word “insightful” suggests that the goal is (we hope) to reveal some aspects of the data that might not be perceived, appreciated, or absorbed by other means. The overall aims include both beauty and truth, though each of these is only as perceived by the beholder.

Methods for visualizing quantitative data have a long history. These methods are now widely used in both data analysis and data presentation, and in both popular and scientific media. However, graphical methods for categorical data have only recently developed and, consequently, are not as widely used. The goal of this book is to show, concretely, how data visualization can be usefully applied to categorical data.

"Categorical data" means different things in different contexts. The topic is introduced in Section 1.2, which contains some examples illustrating (a) types of categorical variables: binary, nominal, and ordinal; (b) data in case form vs. frequency form; (c) frequency data vs. count data; (d) univariate, bivariate, and multivariate data; and (e) the distinction between explanatory and response variables.

Methods for the analysis of categorical data also fall into two quite different categories, which are described and illustrated in Section 1.3. In the first category are the simple randomization-based methods typified by the classical Pearson χ^2, Fisher's exact test, and Cochran-Mantel-Haenszel tests. In the second category are the model-based methods represented by logistic regression, loglinear, and generalized linear models. Chapters 2 through 5 are mostly related to the randomization-based methods. Chapters 6 and 7 illustrate the model-based methods.

In Section 1.4, some important similarities and differences between categorical data and quantitative data are described, and the implications of these differences for visualization techniques are discussed. Section 1.5 outlines a strategy of data analysis focused on visualization.

1.2 What Is Categorical Data?

A *categorical variable* is a variable for which the possible measured or assigned values consist of a discrete set of categories. Here are some typical examples:

- Gender — Male, Female
- Marital Status — Never Married, Married, Separated, Divorced, Widowed
- Fielding Position (in baseball) — Pitcher, Catcher, 1st base, 2nd base,. . . , Left field
- Side Effects (in a pharmacological study) — None, Skin Rash, Sleep Disorder, Anxiety,. . .
- Political Preference — Left, Center, Right
- Treatment Outcome — No Improvement, Some Improvement, Marked Improvement
- Age — 0-9, 10-19, 20-29, 30-39,. . .
- Number of Children — 0, 1, 2, . . .

As these examples suggest, categorical variables differ in the number of categories: ***binary variables***, such as Gender, are distinguished from those that have more than two categories (called ***polytomous***). For example, Table 1.1 gives data about 4526 applicants to graduate departments at the University of California at Berkeley in 1973, classified by two binary variables, gender and admission status.

Table 1.1 Admissions to Berkeley graduate programs

	Admitted	Rejected	Total
Males	1198	1493	2691
Females	557	1278	1835
Total	1755	2771	4526

Some categorical variables, such as Political Preference and Treatment Outcome, may have ordered categories and are called ***ordinal***; other variables, such as Marital Status, have unordered categories and are called ***nominal***.[1] For example, Table 1.2 shows a 2 × 2 × 3 table of ordered outcomes (None, Some, or Marked Improvement) to an active treatment for rheumatoid arthritis in men and women compared to treatment with a placebo.

Table 1.2 Arthritis treatment data

		Improvement			
Treatment	Sex	None	Some	Marked	Total
Active	Female	6	5	16	27
	Male	7	2	5	14
Placebo	Female	19	7	6	32
	Male	10	0	1	11
Total		42	14	28	84

Finally, such variables differ in the fineness or level to which some underlying observation has been categorized for a particular purpose. From one point of view, *all* data may be considered categorical because the precision of measurement is necessarily finite, or an inherently continuous variable may be recorded only to limited precision. But this view is not helpful for the applied researcher because it neglects the phrase "for a particular purpose." Age, for example, might be treated as a quantitative variable in a study of native language vocabulary, or as an ordered categorical variable in terms of the efficacy or side-effects of treatment for depression, or even as a binary variable (Child vs. Adult) in an analysis of survival following an epidemic or a natural disaster.

1.2.1 Case Form vs. Frequency Form

In many circumstances, data is recorded about each individual or experimental unit. Data in this form is called case data or data in ***case form***. For example, the data in Table 1.2 was derived from the individual data listed in Appendix B.1. Whether or not the data variables and the questions we ask call for categorical or quantitative data analysis, we can always trace any observation back to its individual identifier or data record when the data is in case form.

Data in ***frequency form***, such as that shown in Table 1.2, has already been tabulated, by counting over the categories of the table variables. Data in frequency form may be analyzed by methods for quantitative data if there is a quantitative response variable (weighting each group by the cell frequency by using a `WEIGHT` or a `FREQ` statement). Otherwise, such data is generally best analyzed by methods for categorical data. In either case, however, an observation in a dataset in frequency form refers to all cases in the cell collectively, and it cannot be identified individually. Data in case form can always be reduced to frequency form, but the reverse is rarely possible.

[1] An ordinal variable may be defined as one whose categories are *unambiguously* ordered along a *single* underlying dimension. Both marital status and fielding position may be weakly ordered, but not on a single dimension, and not unambiguously.

1.2.2 Frequency Data vs. Count Data

In many cases, the observations that represent the classifications of events or variables are recorded from *operationally independent* experimental units or individuals, typically, a sample from some population. The tabulated data may be called ***frequency data***. The data in Tables 1.1 and 1.2 are examples of frequency data because each observation that is tabulated comes from a different person.

However, if several events or variables are observed for the same units or individuals, those events are not operationally independent, and it is useful to use the term ***count data*** in this situation. These terms (following Lindsey, 1995) are by no means standard, but the distinction is often important, especially in statistical models for categorical data. In a tabulation of the number of male children within families (Table 1.3), for example, the number of male children in a specific family would be a count variable, taking values 0, 1, 2, ... The number of independent families with a specific number of male children is a frequency variable. Count data also arises when a sequence of events is tabulated over time or, under different circumstances, in a number of individuals.

1.2.3 Univariate, Bivariate, and Multivariate Data

Table 1.1 is an example of a bivariate (two-way) contingency table, and Table 1.2 classifies the observations by three variables. Yet, the Berkeley admisssions data also recorded the department to which potential students applied (giving a three-way table), and in the arthritis data, the age of subjects was also recorded.

Therefore, any contingency table records the marginal totals, summed over all variables not represented in the table. For data in case form, this means simply ignoring (or not recording) one or more variables; the observations remain the same. However, data in frequency form results in smaller tables when any variable is ignored; the observations are the cells of the contingency table.

In the limiting case, only one table variable may be recorded or available, giving the categorical equivalent of univariate data. For example, Table 1.3 gives data about the distribution of the number of male children in families that have 12 children, as discussed in Example 2.10. This data was part of a large tabulation of the sex distribution of families in Saxony in the nineteenth century, but the data in Table 1.3 has only one discrete classification variable, that is, the number of males. Without further information, the only statistical questions concern the form of the distribution. The methods for fitting and graphing such discrete distributions are discussed in Chapter 2. The remaining chapters relate to bivariate and multivariate data.

Table 1.3 Number of Males in 6115 Saxony Families That Have 12 Children

Males	0	1	2	3	4	5	6	7	8	9	10	11	12
Families	3	24	104	286	670	1033	1343	1112	829	478	181	45	7

1.2.4 Explanatory vs. Response Variables

Many statistical models make a distinction between *response* (or *dependent*, or *criterion*) variables and *explanatory* (or *independent*, or *predictor*) variables. In the standard (classical) linear models for regression and analysis of variance (ANOVA), for instance, we treat one (or more) variables as responses, to be explained by the other, explanatory variables. The explanatory variables may be quantitative or categorical (e.g., `CLASS` variables), but this affects only the details of how the model is specified for `PROC GLM` or `PROC REG`. For example, the response variable, treatment outcome, must be considered quantitative, and the model attempts to describe how the *mean* of the distribution of responses changes with the values or levels of the explanatory variables, such as age or gender.

However, when the response variable is categorical, the standard linear models do not apply because they assume a normal (Gaussian) distribution for the model residuals. For example, in Table 1.2 the response is Improvement, and even if numerical scores were assigned to the categories None, Some, and Marked, it may be unlikely that the assumptions of the classical linear models could be met.

Hence, a categorical *response variable* generally requires analysis using methods for categorical data, but categorical explanatory variables may be readily handled by either method.

1.3 Strategies for Categorical Data Analysis

Methods of analysis for categorical data can be classified into two broad categories: those concerned with hypothesis testing *per se*, and those concerned with model building.

1.3.1 Hypothesis-Testing Approaches

In many studies, the questions of substantive interest translate readily into questions concerning hypotheses about association between variables. If a non-zero association exists, we may want to characterize the strength of the association numerically and understand the pattern or nature of the association. For example, in Table 1.1, the question "Is there evidence of gender-bias in admission to graduate school?" may be expressed in terms of an association between gender and admission status in a 2 × 2 contingency table of applicants who are classified by these two variables. If so, we can assess the strength of the association by a variety of measures, including the difference in proportions admitted for men and women or the ratio of the odds of admission for men compared to women, as described in Section 3.2.2.

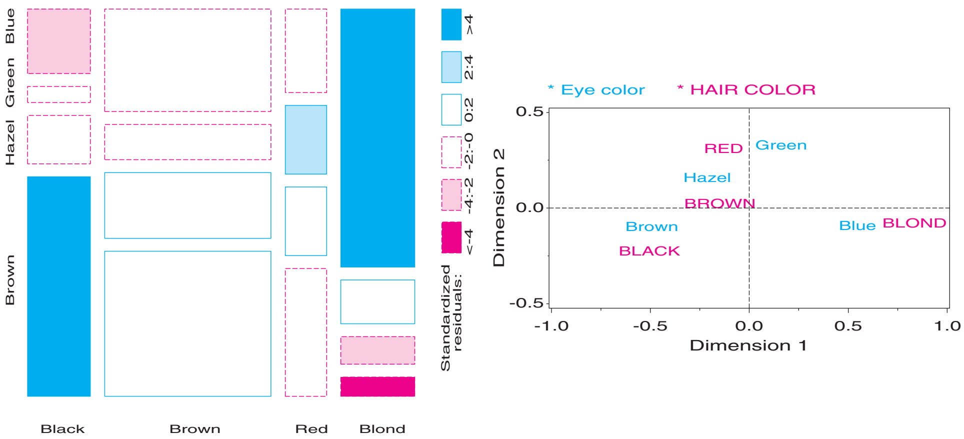
Similarly, in Table 1.2, questions about the efficacy of the treatment for rheumatoid arthritis can be answered in terms of hypotheses about the associations among the table variables: Treatment, Sex, and the Improvement categories. Although the main concern might be focused on the overall association between Treatment and Improvement, one would also want to know if this association is the same for men and women. A stratified analysis (Section 3.3) controls for the effects of background variables, such as Sex, and tests for *homogeneity of association* help determine if these associations are equal.

Questions involving tests of such hypotheses are answered most easily using the randomization-based methods provided by `PROC FREQ`. These include the familiar Pearson chi-square, the Cochran-Mantel-Haenszel test statistics, Fisher's exact test, and a wide variety of measures of strength of association. These tests make minimal assumptions, principally requiring that subjects or experimental units have been randomly assigned to the categories of experimental factors. The hypothesis testing approach is illustrated in Chapters 3 through 5, though the emphasis is on graphical methods that help to understand the nature of association between variables.

EXAMPLE 1.1 Hair color and eye color

Two graphical methods related to the hypothesis-testing approach are shown in Figure 1.1. The data concerns the relationship between hair color and eye color in a sample of nearly 600 students (see Table 3.2 and Appendix B.3). The standard analysis with `PROC FREQ` gives a Pearson χ^2 of 138.3 with 9 degrees of freedom (df), indicating substantial departure from independence. How do we understand the *nature* of this association between hair and eye color?

Figure 1.1 Graphical displays for hair color and eye color data.
Left: mosaic display; right: correspondence analysis 2-D solution.



The left panel of Figure 1.1 is a mosaic display (Chapter 4) constructed so that the size of each rectangle is proportional to the observed cell frequency. The shading reflects the cell contribution to the χ^2 statistic: shades of blue, when the observed frequency is substantially greater than the expected frequency under independence; shades of red, when the observed freqency is substantially less, as shown in the legend.

The right panel of this figure shows the results of a correspondence analysis (Chapter 5), where the deviations of the hair-color and eye-color points from the origin account for as much of the χ^2 as possible in two dimensions.

We observe that both the hair colors and the eye colors are ordered from dark-to-light in the mosaic display and along Dimension 1 in the correspondence analysis plot. The deviations between observed and expected frequencies have an opposite-corner pattern in the mosaic display, except for the combination of red hair and green eyes, which also stand out as the largest values on Dimension 2 in the correspondence analysis plot. Displays such as these provide a means to understand *how* the variables are related. □

1.3.2 Model-Building Approaches

In other situations, model-based methods provide tests of equivalent hypotheses about associations, but (at the cost of additional assumptions) offer additional advantages not provided by the simpler hypotheses-testing approaches. As in the analysis of quantitative data, linear statistical models relate the expected value of a response to a linear function of the table variables, and also assume that residuals or deviations from the model follow a known parametric form.

For a dichotomous response variable, for example, it is convenient to construct a model relating a function of the probability, π, of one event to a linear combination of the explanatory variables. Logistic regression uses the logit function,

$$\text{logit}(\pi) = \log_e \frac{\pi}{1-\pi}$$

which may be interpreted as the log odds of the given event.

Statistical inferences from model-based methods also provide tests of hypotheses, but they provide estimates of parameters in the model and associated confidence intervals and prediction intervals for the response as well. A particular advantage of the logit representation in the logistic regression model is that estimates of odds ratios (Section 3.2.2) may be obtained directly from the parameter estimates.

EXAMPLE 1.2 *Challenger* disaster

To illustrate, the graph in Figure 1.2 is based on a logistic regression model predicting the probability of a failure in one of the O-ring seals used in the NASA space shuttles prior to the disasterous launch of the *Challenger* in January, 1986.[2] The explanatory variable is the ambient temperature at the time of the flight. The sad story behind this data and the lessons to be learned for graphical data display are related in Example 6.5.

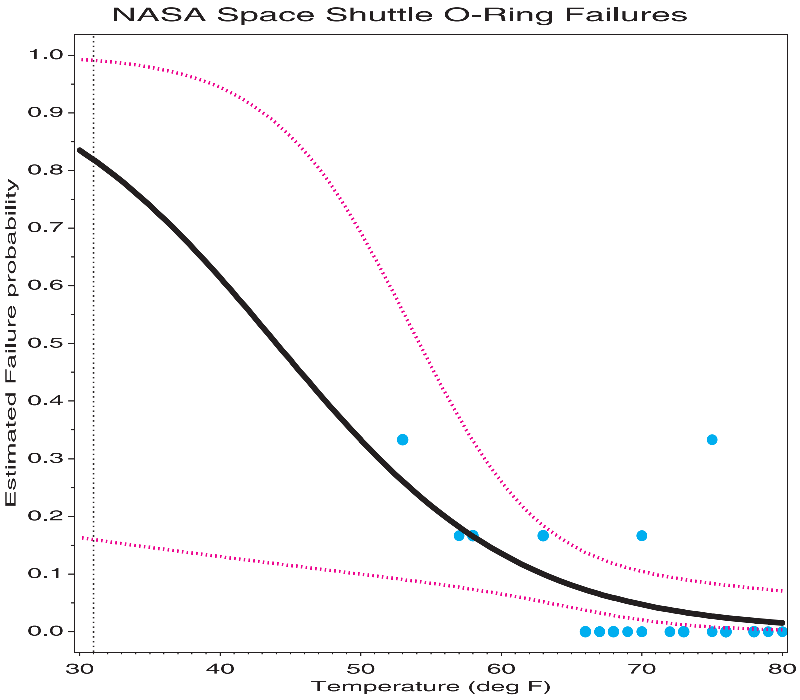
Here, we simply note that the fitted model, shown by the solid line in Figure 1.2, corresponds to the prediction equation (with standard errors shown in parentheses),

$$\text{logit(Failure)} = \underset{(3.06)}{5.09} - \underset{(0.047)}{0.116}\ \text{Temp}$$

An hypothesis test that failure probability is unassociated with temperature is equivalent to the test that the coefficient for temperature in this model equals 0; this test has a p-value of 0.014, which is convincing evidence for rejection. However, the parameter estimate for temperature, -0.116, gives more information. Each 1° increase in temperature decreases the log odds of failure by 0.116, with 95% confidence interval $(-0.208, -0.0235)$. The equivalent odds ratio is $\exp(-0.116) = 0.891$ (0.812–0.977). Equivalently, a 10° *decrease* in temperature corresponds to an odds ratio of a failure of $\exp(10 \times 0.116) = 3.18$, more than tripling the odds of a failure.

When the *Challenger* was launched, the temperature was only 31°. The dashed lines (red) in Figure 1.2 show 95% prediction intervals for failure probability. All previous shuttles (shown by the points in the figure) had been launched at much warmer temperatures, so the prediction interval (the dashed vertical line at the left of the graph) at 31° represents a considerable extrapolation beyond the available data. Nonetheless, the model-building approach does provide such predictions along with measures of their uncertainty. Figure 1.2 is a graph that might have saved lives. □

Figure 1.2 NASA Space Shuttle O-ring Failure, observed and predicted probabilities



[2] "Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure," vol. 84, no. 408, by Siddhartha R. Dalal, Edward B. Fowlkes, and Bruce Hoadley. Copyright © 1989 by *Journal of the American Statistical Association*. Reprinted by permission of Journal of the American Statistical Association via the Copyright Clearance Center.
Reprinted by permission, *Visual Explanations: Images and Quantities, Evidence and Narrative*, by Edward Tufte, Graphics Press 1997.

An additional advantage of the model-building approach is that it often provides greater flexibility and allows more detailed or specialized descriptions of the relations among variables to be tested. For instance, in square, two-way tables (such as those classifying the occupations of fathers and sons or attitudes of husbands and wives) specialized models that deal with symmetry or forms of lack of symmetry may be fit and tested. Such models are usually of much greater substantive interest than the hypothesis of general association. Similarly, specialized models for ordinal variables allow more detailed tests of the nature of association to be examined. Chapter 4, 6, and 7 illustrate many forms of these specialized models.

1.4 Graphical Methods for Categorical Data

> You can see a lot, just by looking.
>
> Yogi Berra

The graphical methods for categorical data described in this book are in some cases straightforward adaptations of more familiar visualization techniques developed for quantitative data. The graphical principles and strategies, and the relations between the visualization approach and traditional statistical methods are described in *SAS System for Statistical Graphics, First Edition*, Chapter 1, and Cleveland (1993b). Another perspective on visual data display is presented in Section 1.4.1. However, the discrete nature of categorical data implies that some familiar graphical methods need to be adapted, while in other cases, we require a new graphic metaphor for data display. These issues are illustrated in Section 1.4.2.

1.4.1 Goals and Design Principles for Visual Data Display

Designing good graphics is surely an art, but as 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 perception of patterns and their *meanings* must be extracted from the visual display.

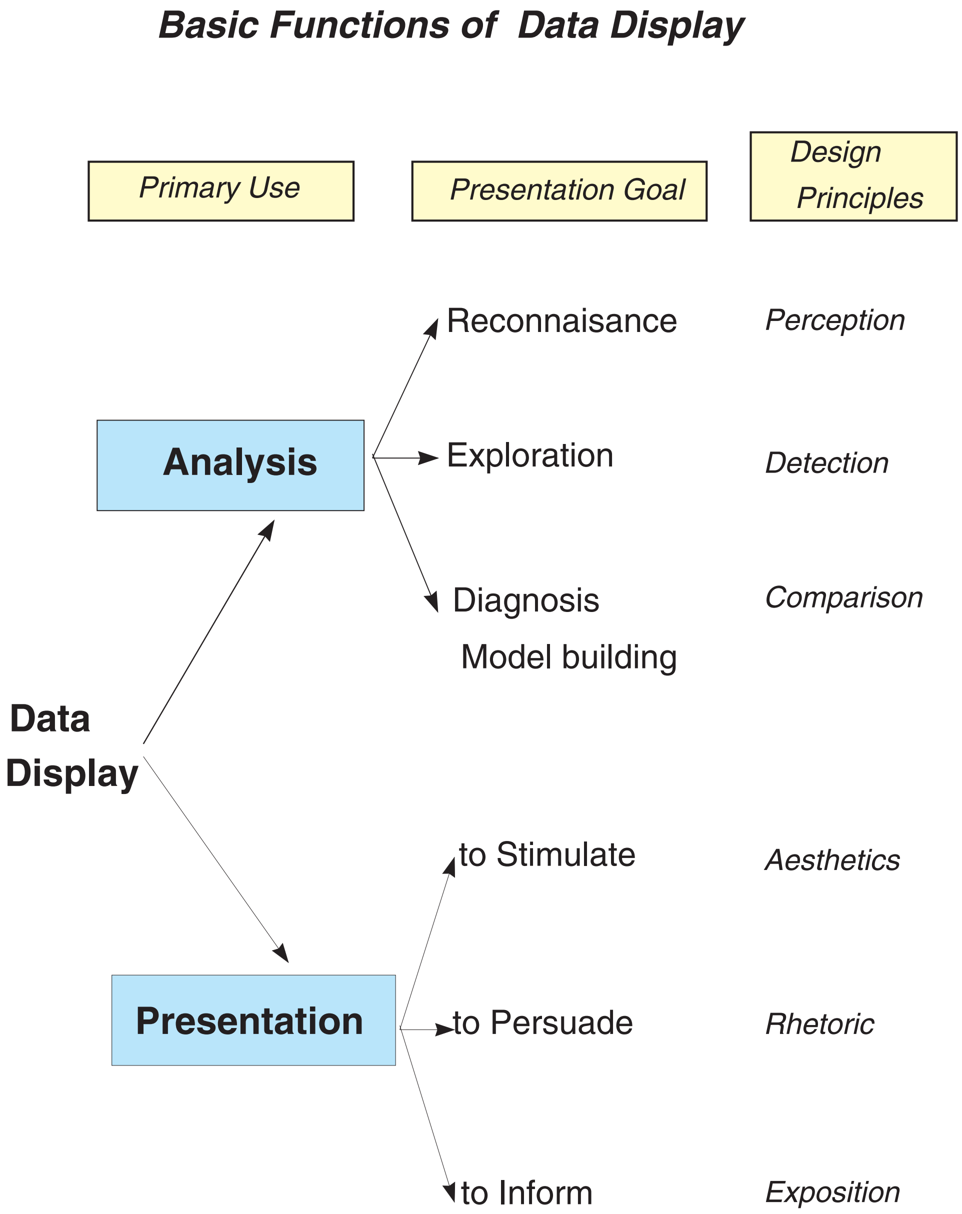
There are many views of graphs, of graphical perception, and of the roles of data visualization in discovering and communicating information. On the one hand, a graphical display may be regarded as a "stimulus"—a package of information to be conveyed to an idealized observer. From this perspective, certain questions are of interest: Which form or graphic aspect promotes greater accuracy or speed of judgment (for a specific task or question)? What aspects lead to greatest memorability or impact? Cleveland (Cleveland and McGill, 1984, 1985; Cleveland, 1993a), and 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 selected communication goal to be achieved, and judges the effectiveness of a graphical display in how well it meets that goal. Kosslyn (1985, 1989) and Tufte (1983, 1990, 1997) have articulated this perspective most clearly.

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: for personal use, to present at a conference or a meeting of our colleagues, or to publish in a research report or in a communication to a general audience (Friendly, 1991, Chapter 1).

Figure 1.3 shows one type of organization of visualization methods in terms of the primary use or intended communication goal, the functional presentation goal, and the suggested corresponding design principles that are applicable.

Figure 1.3 A taxonomy of the basic functions of data display by intended use and presentation goal



The first distinction identifies ***Analysis*** or ***Presentation*** as the primary use of a data graphic (with the understanding that a specific graph may serve both purposes—or, sadly, neither).

Analysis Graphs

Graphs used for data analysis should clearly show the data, but they should also “force us to notice what we never expected to see” (Tukey, 1977, p. vi).

Among graphical methods designed to help study or understand a body of data, it is possible to distinguish those methods designed for different purposes. As suggested in Figure 1.3, each presentation goal is associated with somewhat different design principles.

- ***reconnaissance*** — a preliminary examination or an overview of a possibly complex terrain. For this goal, we may be willing to sacrifice detail for a wider field of view. For example, with a large, multi-way contingency table, we might want to examine the collection of one-way and two-way marginal subtables visually.
- ***exploration*** — graphs designed to help detect patterns or unusual circumstances, or to suggest hypotheses, analyses, or models. For a binary response and a number of categorical or quantitative predictors, a collection of smoothed plots of the response against each predictor may suggest important variables that should be included in a model or extreme observations that should be examined.
- ***diagnosis*** — graphs designed to summarize or critique a numerical statistical summary.

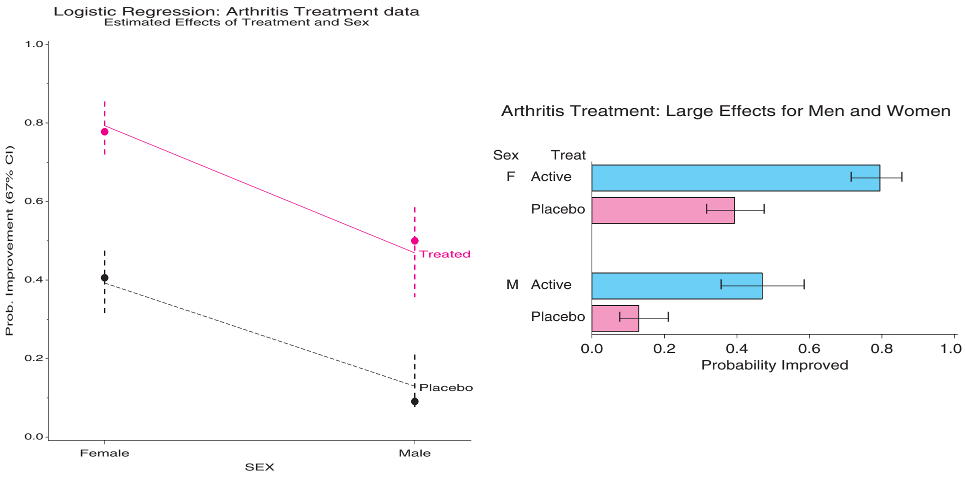
Presentation Graphs

Presentation graphics have different goals. You may want 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 to tailor its message to that goal.

It is often the case that a graph originally prepared as an aid to data analysis can be transformed to a graph intended for presentation by a simple re-design. Sometimes this entails removing detail useful for the analyst but which may detract from the major message; sometimes this may involve adding titles or annotation to make the message more immediately apparent. In still other cases, we may decide to change the graphic format to make visual comparisons easier for the intended audience.

For example, Figure 1.4 shows two views of the results of fitting a logistic regression model to the arthritis treatment data (described in Section 6.4). The left panel shows the observed (points) and predicted probabilities of improvement (±1 standard error, giving approximate 67% confidence intervals) in the form of a line graph. The right panel shows a possible re-design of this graph for presentation purposes.

Figure 1.4 Two graphical displays for arthritis treatment data.
Left: initial analysis graph; right: re-design for presentation.



The line graph might be preferred for analysis purposes because it shows (a) the observed and fitted probabilities are quite similar, (b) there is a large effect of both treatment and sex, and (c) the effect of treatment is about the same for both men and women. The presentation version contains the same predicted probabilities and error bars as the original graph, but, for simplicity, omits the observed probabilities. The title explicitly announces the conclusion to be drawn from the graph.

1.4.2 Categorical Data Requires Different Graphical Methods

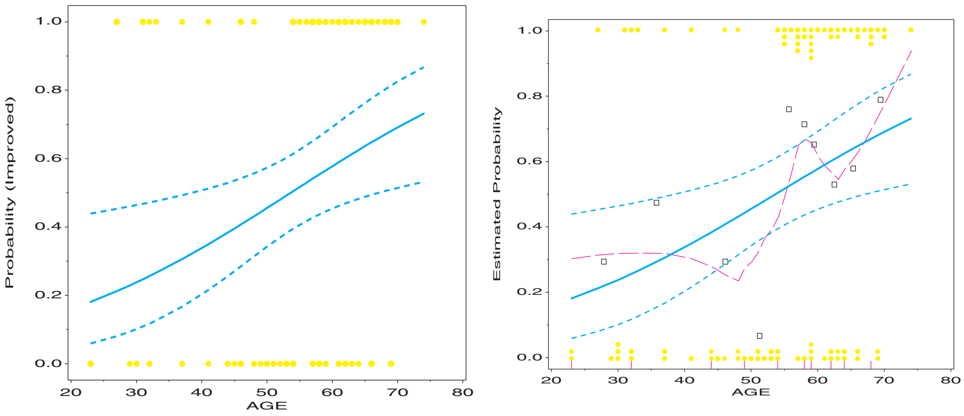
We will see in Chapters 6 and 7 that statistical models for discrete response data and for frequency data are close analogs of the linear regression and ANOVA models that are used for quantitative data. These analogies suggest that the graphical methods commonly used for quantitative data may be adapted directly to categorical data.

Happily, it turns out that many of the analysis graphs and diagnostic displays (e.g., influence plots, added variable and partial residual plots, etc.), which have become common adjuncts in the analysis of quantitative data, have been extended to generalized linear models including logistic regression and log-linear models.

Unhappily, the familiar techniques for displaying raw data are often disappointing when applied to categorical data. For example, the simple scatterplot is widely used, together with the fitted linear model, to show the relation between quantitative response and predictors. For the arthritis data in case form (Appendix B.1), the analogous plot for a logistic regression model (predicting Pr(Some or Marked) improvement from Age) shown in the left panel of Figure 1.5, is, well, underwhelming. First, the response Improve takes on only the values 0 and 1, and Age (in years) is also discrete, so, many points overplot in this graph.[3] Second, although this graph is enhanced with the curve of predicted probabilities under the fitted model (solid line) and 95% confidence bands (dashed lines), it is hard to appreciate how the data points relate to the fitted model. (Can you see that the probability of improvement increases with age?)

These problems may be reduced to some degree by smoothing and by jiggling the points to avoid overplotting. The right panel of Figure 1.5 shows a modest improvement. Here, the raw observations were offset by stacking down from 1 and up from 0 wherever duplicate observations occurred. In addition, the observations were grouped into tenths by age; the lower boundaries of the age categories are shown by the tick marks on the horizontal scale. The proportion of Improved responses in each age group is then plotted (squares), and

Figure 1.5 Graphical displays for Arthritis treatment data.
Left: raw data with logistic regression on age; right: stacked raw data, logistic regression, and smoothed lowess curve.



[3] Only 51 distinct points are shown for the 84 observations.

a non-parametric (lowess) smoothed curve is added to the plot. Although the smoothed curve is somewhat jagged, we now have a clearer visual impression that the probability of improvement increases with age, and we can see the large number of 1 responses among older people.

In Figure 1.5, the quantitative variable Age supports the use of the scatterplot as the graphic format for visual display. A more essential difference between quantitative data and categorical data arises when all variables are categorical, as in a contingency table like Table 1.2. Then, we find that a different visual representation is more natural and useful (Friendly, 1995, 1997).

For quantitative data, magnitude can be represented by length (in a bar chart) or by position along a scale (dotplots, scatterplots). When the data is purely categorical, design principles of perception, detection, and comparison (Friendly, 1999b) suggest that frequencies are most usefully represented as *areas*. In spite of the fact that (in magnitude estimation tasks) judgments of area are known to be less accurate than those of length (e.g., Cleveland and McGill, 1984), here are two fundamental reasons why area is a preferred visual representation for count data:

- multiplicative relations of probabilities and expected frequencies translate readily into height and width of rectangles, whose area then depicts a cell value.
- a concrete, physical model for categorical data (Friendly, 1995) based on count ∼ area yields a surprising range of correct, but novel, interpretations for statistical principles (maximum likelihood), estimation techniques (iterative proportional fitting, Newton–Raphson) and statistical concepts (power, why components of likelihood-ratio G^2 can be negative).

The first reason is illustrated in Figure 1.6, a sieve diagram (Section 3.5) for the Berkeley admissions data, broken down by department. In this display, each box has a height

Figure 1.6 Sieve diagram for Berkeley admissions data. Each box has area proportional to its expected frequency and is cross-ruled with boxes equal to the observed frequency.

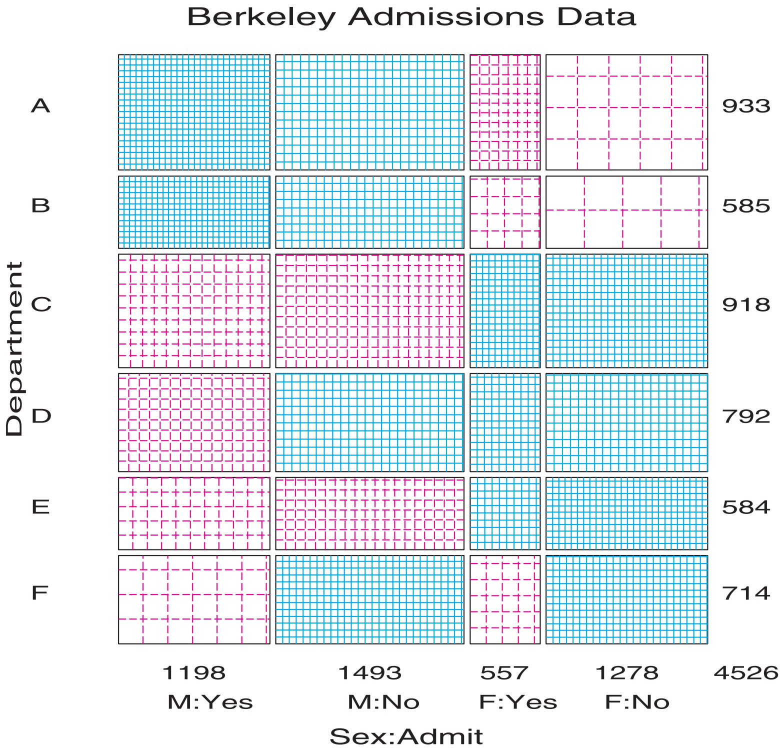
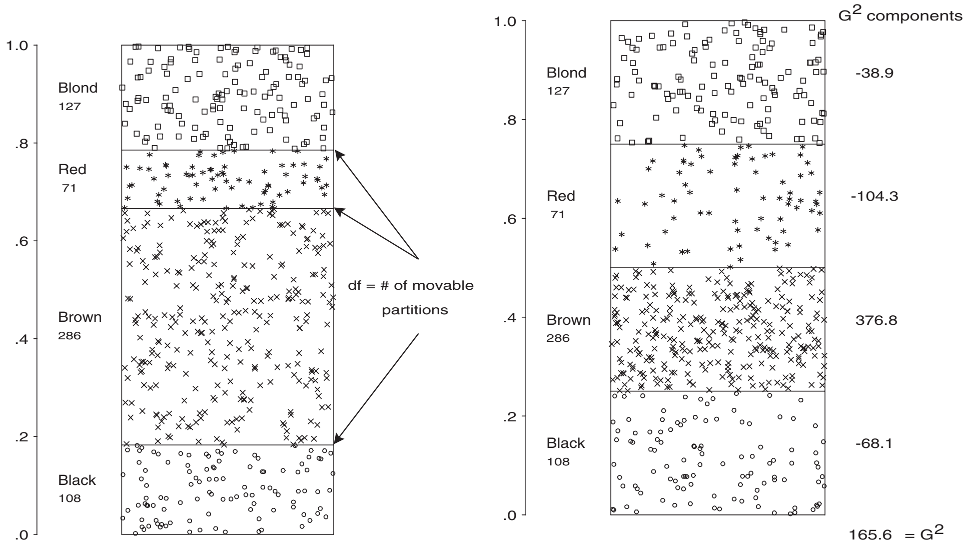


Figure 1.7 Conceptual model for categorical data



proportional to the marginal total for the corresponding department and a width proportional to the column marginal total, so the area is proportional to the expected frequency under independence. The observed frequency in each cell is shown by the number of cross-ruled boxes, so departures from independence are shown visually as variations in shading density.

The second point is illustrated in Figure 1.7, using data (see Table 3.2) on $n = 592$ individuals classified by hair color. In the conceptual model (Friendly, 1995, Sall, 1991), categorical observations are likened to molecules of an ideal gas confined to chambers separated by movable partitions. In both panels of the figure, the number of symbols in each box exactly equals the number of observations in each hair-color category.

When the location of the partitions are unconstrained, as shown in the left panel of Figure 1.7, the forces balance in each chamber by moving to the positions of minimum potential energy, so that the height of each chamber is $p_i = n_i/n$, which is the maximum likelihood estimate of the probability π_i in each cell.

To test the hypothesis that all hair colors are equally likely, imagine forcing the partitions to move to the positions where $\pi_i = \frac{1}{4}$, as shown in the right panel. The change in energy in each compartment is then $-(\log p_i - \log \pi_i) = -\log(p_i/\pi_i)$, the change in negative log-likelihood. Sum these and multiply by 2 to get the likelihood ratio G^2. This gives a concrete interpretation of G^2 as a measure of the effort to maintain belief in the hypothesis in the face of the data.

This concrete model supplies neat explanations of many other results for categorical data, extends readily to multiway tables, and provides a rationale for the graphic representation of counts by area or by visual density. It also serves as the basis for the mosaic display described in Chapter 4.

1.5 Visualization = Graphing + Fitting + Graphing

Look here, upon this picture, and on this.

William Shakespeare, *Hamlet*

Statistical summaries, hypothesis tests, and the numerical parameters derived in fitted models are designed to capture a particular feature of the data. An analysis of the data from Table 1.1, for example, shows that 44.5% of male applicants were admitted to Berkeley, compared to 30.4% of female applicants, giving a Pearson chi-square of 92.2 with 1 degree of freedom for association between admission and gender ($p < 0.001$). Expressed in terms of the odds ratio, males were apparently 1.84 times as likely to be admitted as females, with 99% confidence bounds 1.562–2.170. Each of these numbers expresses some part of the relationship between gender and admission in the Berkeley data.

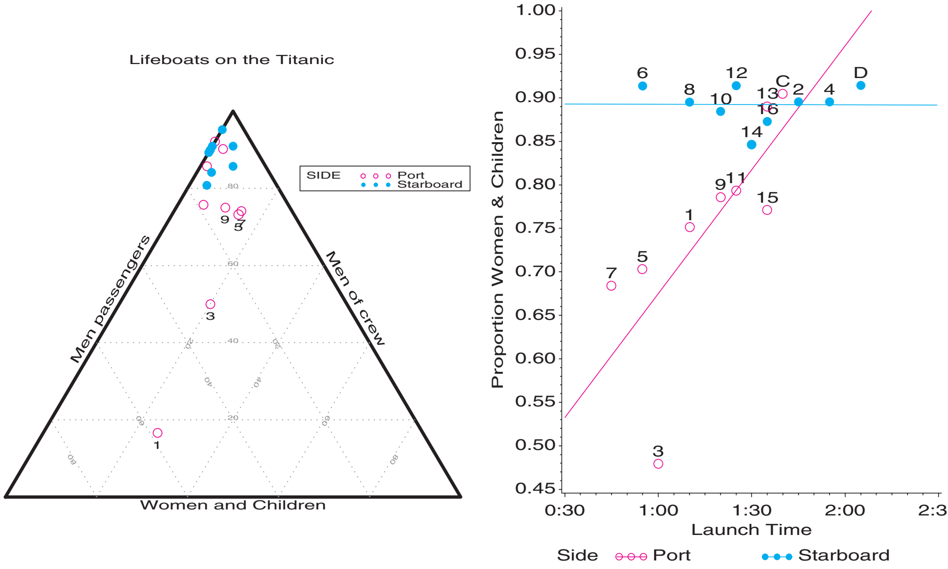
Numerical summaries, even for such a small dataset as this, are designed to compress the information in the data. In contrast, the visualization approach to data analysis is designed to (a) *expose* information and structure in the data, (b) *supplement* the information available from numerical summaries, and (c) *suggest* more adequate models. In general, the visualization approach seeks to serve the needs of both summarization and exposure.

This approach recognizes that both data analysis and graphing are iterative processes. You should not expect that any one model captures all features of the data, any more than you should expect that a single graph shows all that may be seen. In most cases, the initial steps should include some graphical display guided by understanding of the subject matter of the data. What you learn from a graph may then help suggest features of the data to be incorporated into a fitted model. Your desire to ensure that the fitted model is an adequate summary may then lead to additional graphs.

EXAMPLE 1.3 Lifeboats on the *Titanic*

One example is shown in Figure 1.8, described in more detail in Example 3.18. The left panel shows a trilinear plot of the composition of lifeboats on the *Titanic*. Each point in the plot shows the relative proportions of male passengers and identifies the lifeboats that have 10% or more men, women and children, and men-of-crew reported in each of the 18 lifeboats launched from the port and starboard sides of that ill-fated vessel. Trilinear plots are described in Section 3.8, but essentially, the points near the top apex represent boats that are almost all filled with women and children.

Figure 1.8 Two graphical displays for *Titanic* lifeboat data.
Left: trilinear plot, right: logistic regression.



That graph suggested that the procedures for loading the lifeboats might have differed for the port and starboard side of the ship. This led to fitting a logistic regression model to predict the proportion of women and children loaded over a period of time, with separate slopes and intercepts for the port and starboard sides. The panel on the right of Figure 1.8 shows predicted proportions from this model, with simple linear regression lines for the two sides. Even without further details about the data or the analysis, the graph clearly shows that passengers on the two sides of the *Titanic* were subject to different regimes for loading the lifeboats. □

This interplay between graphing and fitting can be expressed as

$$\textbf{Visualization} = \textbf{Graphing} + \textbf{Fitting} + \textbf{Graphing} + \cdots \ ,$$

where the ellipsis ($\cdots$) reminds us that there are often additional steps.

Sometimes a visualization is sufficiently strong (as, perhaps in Figure 1.4 or in the panel on the right side of Figure 1.8), that hypothesis tests and model parameters serve an ancillary role in understanding the effects in the data. p-values are useful in the conventions of scientific communication, but perhaps less convincing evidence than a graph whose conclusions hit you between the eyes (sometimes called the Intraocular Traumatic Test).

In other cases, graphing serves as a supporting member of the data-analytic cast. Model-based methods rely on assumptions about the data. Diagnostic plots for logistic regression (Section 6.6) and log-linear models (Section 7.7) may provide comfort that the assumptions on which these inferences depend are reasonable for the data at hand, or the plots may suggest that some modification of the model would help us to rest more easily.

In any event, it is well to remember that data analysis requires both summarization and exposure, and the needs of both are served by the combination of graphing and fitting.

1.5.1 Static vs. Dynamic Graphics

The confines of a book and of the software that are described here for visualizing categorical data limit this presentation to static displays that are produced by SAS programs. Many of these static graphics are made considerably easier to use and more flexible when you use SAS macros as illustrated in the following chapters and described in Appendix A.

The most productive use of these methods requires the addition of two aspects of interactive graphics that presently are being developed by me (Friendly, 1996) and by others (Theus and Lauer, 1999; Young, 1994).

- The first aspect relates to interactive methods for choosing variables, parameters, and options for analysis and graphical displays. The development tools for this form of interactivity are provided in SAS/AF and most of the macro programs described here may be easily wrapped in an interactive front-end.
- A second aspect of dynamic graphics is related to the ability to interact with multiple, linked views of a dataset, so that, for example, selecting a subset of cases in one view highlights them in all other views. SAS/INSIGHT is a prototype for this type of interaction in SAS software. JMP software provides another route.

I look forward to the possible development of more interactive methods and the extension of multiple, linked data views for categorical data to be used with SAS software. Nevertheless, it is necessary to understand the various forms of graphic displays that are particularly useful for discrete data before learning how to employ them interactively.