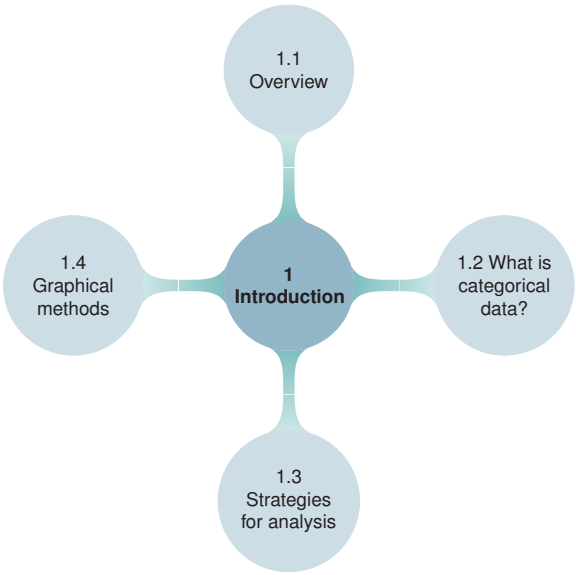


# 1

# Introduction

{ch:intro}

Categorical data consist of variables whose values comprise a set of discrete categories. Such data require different statistical and graphical methods than 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. This chapter outlines the basic orientation of the book and some key distinctions regarding the analysis and visualization of categorical data.

## 1.1 Data visualization and categorical data: Overview

{sec:viscat}

> Graphs carry the message home. A universal language, graphs convey information directly to the mind. Without complexity there is imaged to the eye a magnitude to be remembered. Words have wings, but graphs interpret. Graphs are pure quantity, stripped of verbal sham, reduced to dimension, vivid, unescapable.
>
> Henry D. Hubbard, in Foreword to Brinton (1939), *Graphic Presentation*

"Data visualization" can mean many things, from popular press infographics, to maps of voter turnout or party choice. Here we use this term in the narrower context of statistical analysis. As such, we refer to an approach to data analysis that focuses on *insightful* graphical display in the service of both *understanding* our data and *communicating* our results to others.

We may 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 (hopefully) to reveal some aspects of the data that might not be perceived, appreciated, or absorbed by other means. As

in the quote from Keats, the overall aims include both beauty and truth, though each of these are only as perceived by the beholder.

Methods for visualizing quantitative data have a long history and are now widely used in both data analysis and in data presentation, and in both popular and scientific media. Graphical methods for categorical data, however, have only a more recent history, and are consequently not as widely used. The goal of this book is to show concretely how data visualization may be usefully applied to categorical data.

"Categorical" means different things in different contexts. We introduce the topic in Section 1.2 with 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.

Statistical methods for the analysis of categorical data also fall into two quite different categories, described and illustrated in Section 1.3: (a) the simple randomization-based methods typified by the classical Pearson chi-squared ($\chi^2$) test, Fisher's exact test, and Cochran–Mantel–Haenszel tests, and (b) the model-based methods represented by logistic regression, loglinear, and generalized linear models. In this book, Chapters 3–6 are mostly related to the randomization-based methods; Chapters 7–9 illustrate the model-based methods.

In Section 1.4 we describe some important similarities and differences between categorical data and quantitative data, and discuss the implications of these differences for visualization techniques. Section 1.4.5 outlines a strategy of data analysis focused on visualization.

In a few cases we show R code or results as illustrations here, but the fuller discussion of using R for categorical data analysis is postponed to Chapter 2.

## 1.2 What is categorical data?

{sec:whatis}

A ***categorical variable*** is one for which the possible measured or assigned values consist of a discrete set of categories, which may be *ordered* or *unordered*. Some typical examples are:

- `Gender`, with categories "Male," "Female."
- `Marital status`, with categories "Never married," "Married," "Separated," "Divorced," "Widowed."
- `Fielding position` (in baseball), with categories "Pitcher," "Catcher," "1st base," "2nd base," . . ., "Left field."
- `Side effects` (in a pharmacological study), with categories "None," "Skin rash," "Sleep disorder," "Anxiety," . . ..
- `Political attitude`, with categories "Left," "Center," "Right."
- `Party preference` (in Canada), with categories "NDP," "Liberal," "Conservative," "Green."
- `Treatment outcome`, with categories "no improvement," "some improvement," or "marked improvement."
- `Age`, with categories "0–9," "10–19," "20–29," "30–39," . . . .
- `Number of children`, with categories $0, 1, 2, \ldots$.

As these examples suggest, categorical variables differ in the number of categories: we often distinguish ***binary variables*** (or ***dichotomous variables***) such as `Gender` from those with more than two categories (called ***polytomous variables***). For example, Table 1.1 gives data on $4,526$ applicants to graduate departments at the University of California at Berkeley in 1973, classified by two binary variables, gender and admission status.

Some categorical variables (`Political attitude`, `Treatment outcome`) may have ordered categories (and are called ***ordinal***), while other (***nominal***) variables like `Marital status`

{tab:berk220}

**Table 1.1:** Admissions to Berkeley graduate programs

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

have unordered categories.[1] For example, Table 1.2 shows a $2 \times 2 \times 3$ table of ordered outcomes ("none," "some," or "marked" improvement) to an active treatment for rheumatoid arthritis compared to a placebo for men and women.

{tab:arthrit0}

**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 with decade groups (0–10, 11–20, 20–30, ...) 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 natural disaster. In the analysis of data using categorical methods, continuous variables are often recoded into ordered categories with a small set of categories for some purpose.[2]

### 1.2.1 Case form vs. frequency form

{sec:case-freq}

In many circumstances, data is recorded on each individual or experimental unit. Data in this form is called case data, or data in ***case form***. The data in Table 1.2, for example, were derived from the individual data listed in the data set `Arthritis` from the vcd (Meyer et al., 2015) package. The following lines show the first five of $N = 84$ cases in the `Arthritis` data,

[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.

[2] This may be a waste of information available in the original variable, and should be done for substantive reasons, not mere convenience. For example, some researchers unfamiliar with regression methods often perform a "median-split" on quantitative predictors so they can use ANOVA methods. Doing this precludes the possibility of determining if those variables have non-linear relations with the outcome while also decreasing statistical power.

```
  ID Treatment  Sex Age Improved
1 57   Treated Male  27     Some
2 46   Treated Male  29     None
3 77   Treated Male  30     None
4 17   Treated Male  32   Marked
5 36   Treated Male  46   Marked
```

Whether or not the data variables, and the questions we ask, call for categorical or quantitative data analysis, when the data are in case form, we can always trace any observation back to its individual identifier or data record (for example, if the case with `ID` equal to 57 turns out to be unusual or noteworthy).

Data in ***frequency form*** has already been tabulated, by counting over the categories of the table variables. The same data shown as a table in Table 1.2 appear in frequency form as shown below.

```
   Treatment    Sex Improved Freq
1    Placebo Female     None   19
2    Treated Female     None    6
3    Placebo   Male     None   10
4    Treated   Male     None    7
5    Placebo Female     Some    7
6    Treated Female     Some    5
7    Placebo   Male     Some    0
8    Treated   Male     Some    2
9    Placebo Female   Marked    6
10   Treated Female   Marked   16
11   Placebo   Male   Marked    1
12   Treated   Male   Marked    5
```

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, with a weight variable). Otherwise, such data are generally best analyzed by methods for categorical data, where statistical models are often expressed as models for the frequency variable, in the form of an R formula like `Freq ~ .`.

In any case, an observation in a data set in frequency form refers to all cases in the cell collectively, and these cannot be identified individually. Data in case form can always be reduced to frequency form, but the reverse is rarely possible. In Chapter 2, we identify a third format, ***table form***, which is the R representation of a table like Table 1.2.

### 1.2.2 Frequency data vs. count data

{sec:freq-count}

In many cases the observations representing 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 Table 1.1 and Table 1.2 are both examples of frequency data because each tabulated observation 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, particularly in statistical models for categorical data.

For example, in a tabulation of the number of male children within families (Table 1.3, described in Section 1.2.3 below), the number of male children in a given family would be a *count* variable, taking values $0, 1, 2, \ldots$. The number of independent families with a given number of male children is a *frequency* variable. Count data also arise when we tabulate a sequence of events over time or under different circumstances in a number of individuals.

{tab:saxdata}

**Table 1.3:** Number of Males in 6115 Saxony Families of Size 12

| Males | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Families | 3 | 24 | 104 | 286 | 670 | 1,033 | 1,343 | 1,112 | 829 | 478 | 181 | 45 | 7 |

### 1.2.3 Univariate, bivariate, and multivariate data

{sec:uni-multi}

Another distinction concerns the number of variables: one, two, or (potentially) many shown in a data set or table, or used in some analysis. Table 1.1 is an example of a bivariate (two-way) contingency table and Table 1.2 classifies the observations by three variables. Yet, we will see later that the Berkeley admissions 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.

Any contingency table (in frequency or table form) therefore 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. Data in frequency form, however, result in smaller tables when any variable is ignored; the "observations" are the cells of the contingency table. For example, in the `Arthritis` data, ignoring `Sex` gives the smaller $2 \times 3$ table for `Treatment` and `Improved`.

```
  Treatment Improved Freq
1   Placebo     None   29
2   Treated     None   13
3   Placebo     Some    7
4   Treated     Some    7
5   Placebo   Marked    7
6   Treated   Marked   21
```

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 on the distribution of the number of male children in families with 12 children (discussed further in Example 3.2). These data were part of a large tabulation of the sex distribution of families in Saxony in the $19^{th}$ century, but the data in Table 1.3 have only one discrete classification variable, number of males. Without further information, the only statistical questions concern the form of the distribution. We discuss methods for fitting and graphing such discrete distributions in Chapter 3. The remaining chapters relate to bivariate and multivariate data.

### 1.2.4 Explanatory vs. response variables

{sec:exp-resp}

Most statistical models make a distinction between ***response variables*** (or *dependent*, or *criterion* variables) and ***explanatory variables*** (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., factors in R). This affects only the details of how the model is specified or how coefficients are interpreted for `lm()` or `glm()`. In these classical models, the response variable ("treatment outcome," for example), 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.

When the response variable is categorical, however, 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 variable is `Improvement`, and even if numerical scores were assigned to

the categories "none," "some," "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.

The distinction between response and explanatory variables also becomes important in the use of loglinear models for frequency tables (described in Chapter 9), where models can be specified in a simpler way (as equivalent logit models) by focusing on the response variable.

## 1.3 Strategies for categorical data analysis

{sec:strategies}

Data analysis typically begins with exploratory and graphical methods designed to expose features of the data, followed by statistical analysis designed to summarize results, answer questions, and draw conclusions. Statistical methods for the analysis of categorical data can be classified into two broad categories: those concerned with *hypothesis testing* per se versus those concerned with *model building*.

### 1.3.1 Hypothesis testing approaches

{sec:strategies-hyp}

In many studies, the questions of substantive interest translate readily into questions concerning hypotheses about ***association*** between variables, a more general idea than that of correlation (*linear* association) for quantitative variables. If a non-zero association exists, we may wish to characterize the strength of the association numerically and understand the pattern or nature of the association.

For example, in Table 1.1, a main question is: "Is there evidence of gender-bias in admission to graduate school?" Another way to frame this: "Are males more likely to be admitted?" These questions can be expressed in terms of an association between gender and admission status in a $2 \times 2$ contingency table of applicants classified by these two variables. If there is evidence for an association, we can assess its strength 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 4.2.2.

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 wish to know if this association is the same for men and women. A ***stratified analysis*** (Section 4.3) controls for the effects of background variables like Sex, tests for ***homogeneity of association***, and helps to determine if these associations are equal.

Questions involving tests of such hypotheses are answered most easily using a large variety of specific statistical tests, often based on randomization arguments. These include the familiar Pearson chi-squared test for two-way tables, the Cochran–Mantel–Haenszel test statistics, Fisher's exact test, and a wide range 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 4–6, though the emphasis is on graphical methods that help us to understand the nature of association between variables.

{ex:haireye0}

**EXAMPLE 1.1: Hair color and eye color**

The data set `HairEye` below records data on the relationship between hair color and eye color in a sample of nearly 600 students.

```
       Eye
Hair    Brown Blue Hazel Green
  Black    68   20    15     5
  Brown   119   84    54    29
  Red      26   17    14    14
  Blond     7   94    10    16
```

The standard analysis (with `chisq.test()` or `assocstats()`) gives a Pearson $\chi^2$ of 138.3 with nine degrees of freedom, indicating substantial departure from independence. Among the measures of strength of association, ***Cramer's V***, $V = \sqrt{\chi^2/N \min(r-1, c-1)} = 0.279$, indicates a substantial relationship between hair and eye color.[3]

```
                    X^2 df P(> X^2)
Likelihood Ratio 146.44  9        0
Pearson          138.29  9        0

Phi-Coefficient   : NA
Contingency Coeff.: 0.435
Cramer's V        : 0.279
```

The further (and perhaps more interesting question) is how do we understand the *nature* of this association between hair and eye color? Two graphical methods related to the hypothesis testing approach are shown in Figure 1.1.

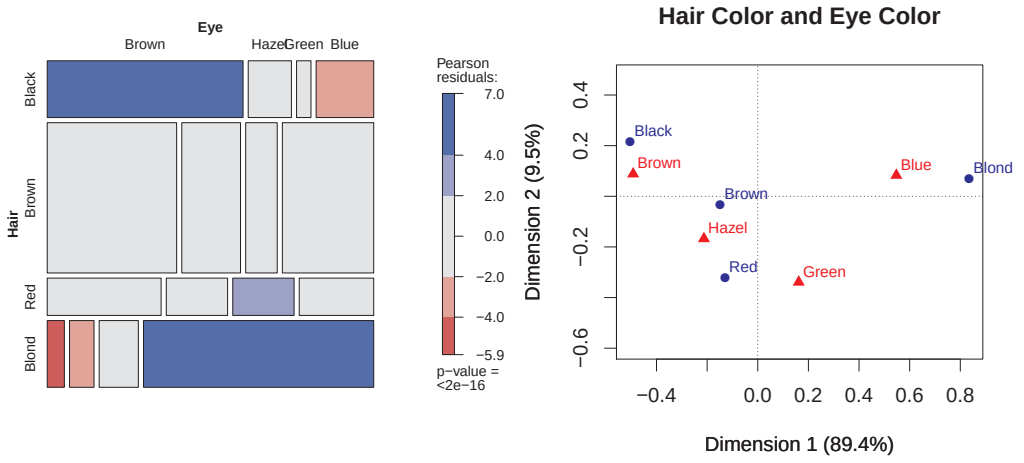


**Figure 1.1:** Graphical displays for the hair color and eye color data. Left: mosaic display; right: correspondence analysis plot. {fig:haireye02}

The left panel of Figure 1.1 is a ***mosaic display*** (Chapter 5), constructed so that the size of each rectangle is proportional to the observed cell frequency. The shading reflects the cell contribution to the $\chi^2$ statistic—shades of blue when the observed frequency is substantially greater than the expected frequency under independence, shades of red when the observed frequency is substantially less, as shown in the legend.

The right panel of this figure shows the results of a correspondence analysis (Chapter 6), where the deviations of the hair color and eye color points from the origin accounts for as much of the $\chi^2$ as possible in two dimensions.

[3] Cramer's V varies from 0 (no association) to 1 (perfect association).

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

Model-based methods provide tests of equivalent hypotheses about associations, but offer additional advantages (at the cost of additional assumptions) not provided by the simpler hypotheses-testing approaches. Among these advantages, model-based methods provide estimates, standard errors and confidence intervals for parameters, and the ability to obtain predicted (fitted/expected) values with associated measures of precision.

We illustrate this approach here for a dichotomous response variable, where it is often convenient to construct a model relating a function of the probability, $\pi$, of one event to a linear combination of the explanatory variables. Logistic regression uses the ***logit function***,

$$\text{logit}(\pi) \equiv \log_e \left( \frac{\pi}{1-\pi} \right) \ ,$$

which may be interpreted as the ***log odds*** of the given event. A linear logistic model can then be expressed as

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

Statistical inferences from model-based methods provide tests of hypotheses for the effects of the predictors, $x_1, x_2, \ldots$, but they also provide estimates of parameters in the model, $\beta_1, \beta_2, \ldots$ and associated confidence intervals. Standard modeling tools allow us to graphically display the fitted response surface (with confidence or prediction intervals) and even to extrapolate these predictions beyond the given data. A particular advantage of the logit representation in the logistic regression model is that estimates of odds ratios (Section 4.2.2) may be obtained directly from the parameter estimates.

{ex:nasa0}

**EXAMPLE 1.2: Space shuttle disaster**

To illustrate the model-based approach, 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 24 NASA space shuttles prior to the disastrous launch of the *Challenger* in January, 1986. The explanatory variable is the ambient temperature (in Fahrenheit) at the time of the flight. The sad story behind these data, and the lessons to be learned for graphical data display, are related in Example 1.10.

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),

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

A 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, convincing evidence for rejection.

The parameter estimate for temperature, $-0.116$, however, gives more information. Each $1^\circ$ 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 $10^\circ$ *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.

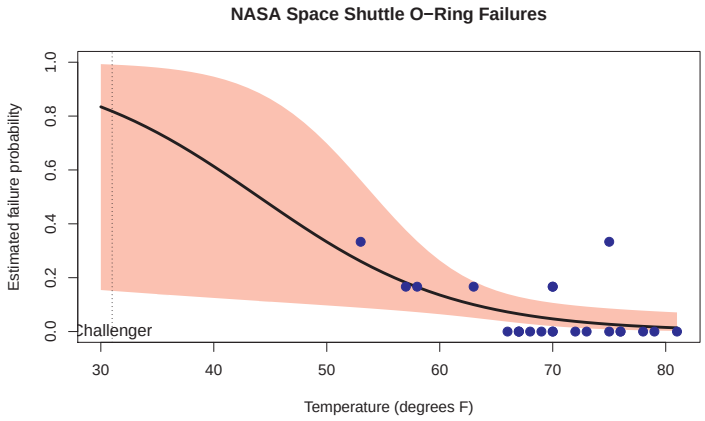


**Figure 1.2:** Space shuttle O-ring failure, observed and predicted probabilities. The dotted vertical line at $31^\circ$ shows the prediction for the launch of the *Challenger*.

{fig:spaceshuttle0}

When the *Challenger* was launched, the temperature was only $31^\circ$. The shaded region in Figure 1.2 shows 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 $31^\circ$ 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.

△

{ex:donner0}

**EXAMPLE 1.3: Donner Party**

In April–May of 1846 (three years before the California gold rush), the Donner and Reed families set out for California from the American Mid-west in a wagon train to seek a new life and perhaps their fortune in the new American frontier. 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 possessions.

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 disastrous: 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 that 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 grisly evidence that those who survived had cannibalized those who died.

Here we briefly examine how statistical models and graphical evidence can shed light on the question of who survived in the Donner party.

Figure 1.3 is an example of what we call a *data-centric, model-based* graph of a discrete (binary) outcome: lived (1) versus died (0). That is, it shows both the data and a statistical summary based on a fitted statistical model. The statistical model provides a smoothing of the discrete data.

The jittered points at the top and bottom of the graph show survival in relation to age of the person. You can see that there were more people who survived among the young, and more who died among the old. The blue curve in the plot shows the fitted probability of survival from a

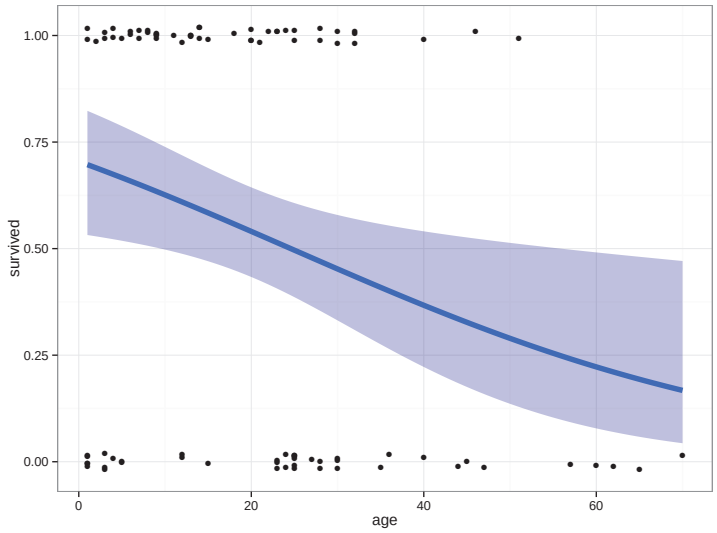


**Figure 1.3:** Donner party data, showing the relationship between age and survival. The blue curve and confidence band give the predicted probability of survival from a linear logistic regression model. {fig:donner0}

linear logistic regression model for these data with a 95% confidence band for the predictions. The prediction equation for this model can be given as:

$$\text{logit}(\text{survived}) = \underset{(0.372)}{0.868} - \underset{(0.015)}{0.0353}\ \text{age}$$

The equation above implies that the log odds of survival decreases by 0.0352 with each additional year of age or by $10 \times 0.0352 = 0.352$ for an additional decade. Another way to say this is that the odds of survival is multiplied by $\exp(0.353) = .702$ with each 10 years of age, a 30% decrease.

Of course, these visual and statistical summaries depend on the validity of the fitted model. For contrast, Figure 1.4 shows two other model-based smoothers that relax the assumption of the linear logistic regression model. The left panel shows the result of fitting a semi-parametric model with a natural cubic spline with one more degree of freedom than the linear logistic model. The right panel shows the fitted curve for a non-parametric, locally weighted scatterplot smoothing (loess) model. Both of these hint that the relationship of survival to age is more complex than what is captured in the linear logistic regression model. We return to these data in Chapter 7.

△

## 1.4 Graphical methods for categorical data

{sec:methods}

> 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. Graphical principles and strategies, and the relations between the visualization approach and traditional

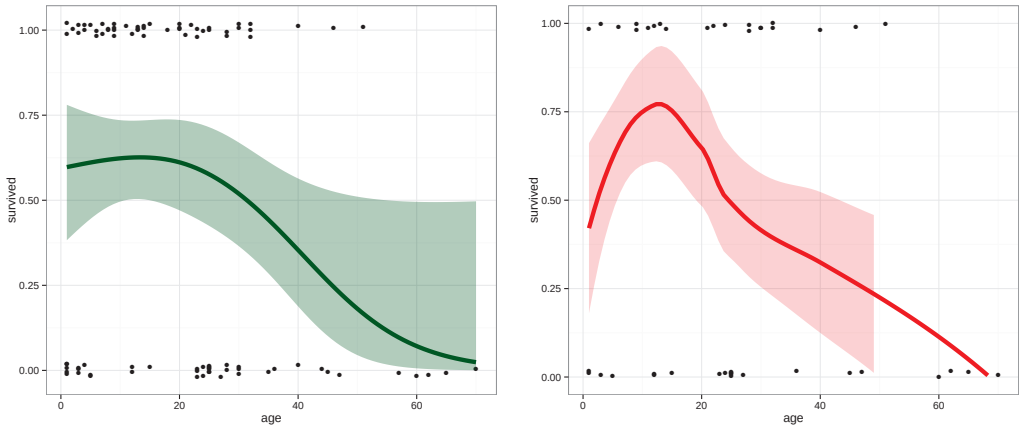


**Figure 1.4:** Donner party data, showing other model-based smoothers for the relationship between age and survival. Left: using a natural spline; right: using a non-parametric loess smoother.

{fig:donner0-other}

statistical methods, are described in a number of sources, including Chambers et al. (1983), Cleveland (1993b) and several influential books by Tufte (Tufte, 1983, 1990, 1997, 2006).

The fundamental idea of statistical graphics as a comprehensive system of visual signs and symbols with a grammar and semantics was first proposed in Jacques Bertin's *Semiology of Graphics* (1983). These ideas were later extended to a computational theory in Wilkinson's *Grammar of Graphics* (2005), and implemented in R in Hadley Wickham's ggplot2 (Wickham and Chang, 2015) package (Wickham, 2009, Wickham and Chang, 2015).

Another perspective on visual data display is presented in Section 1.4.1 focusing on the communication goals of statistical graphics. However, the discrete nature of categorical data implies that some familiar graphic 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. Section 1.4.3 discusses the principle of effect ordering for categorical variables in graphs and tables.

### 1.4.1 Goals and design principles for visual data display

{sec:intro-goals}

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 apperception 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), Spence and Lewandowsky (Lewandowsky and Spence, 1989, Spence, 1990, Spence and Lewandowsky, 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 (Friendly and Kwan, 2011). 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 to: (a) use ourselves, (b) present at a conference or meeting of our colleagues, (c) publish in a research report, or (d) communicate to a general audience (Friendly (1991, Ch. 1), Friendly and Kwan (2011)). Figure 1.5 illustrates a basic contrast between graphs for presentation purposes, designed to appeal persuasively to a large audience (one-to-many) and the use of perhaps many graphs we might make for ourselves for exploratory data analysis (many-to-one).

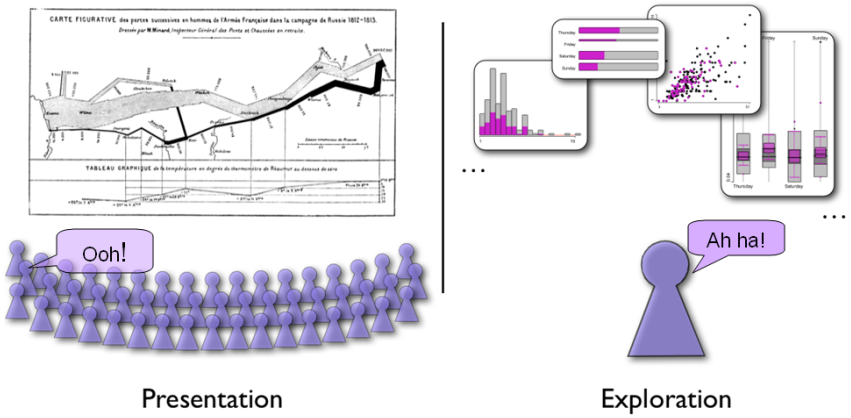


**Figure 1.5:** Different communication purposes require different graphs. For presentations, a single, carefully crafted graph may appeal best to a large audience; for exploratory analysis, many related images from different perspectives for a narrow audience (often you!). *Source*: Adapted from a blog entry by Martin Theus, `http://www.theusrus.de/blog/presentation-vs-exploration/`.

{fig:presentation-exploration}

Figure 1.6 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*.

We illustrate these ideas and distinctions in the examples below, most of which are treated again in later chapters.

{ex:arrests0}

**EXAMPLE 1.4: Racial profiling: Arrests for marijuana possession**

In a case study that will be examined in detail in Chapter 7 (Example 7.10), the *Toronto Star* newspaper studied a huge data base of arrest records by Toronto police for indications of possible racial profiling, i.e., differential treatment of those arrested on the basis of skin color. They focused on the charge of simple possession of a small amount of marijuana, for which enforcement procedures allowed police discretion. An officer could release an arrestee with a summons ("Form 9") to appear in court, or take the person to a police station for questioning ("Form 10") or booking ("Form 11.1"), or order the person held in jail for a bail hearing ("Show cause").

The statistical issue was whether the data on these arrests showed evidence of differential treatment in relation to skin color, particularly in the treatment of blacks vs. whites, controlling, of course, for other factors. Statistical tests on these data ($\chi^2$ tests, loglinear models, logistic regression) showed overwhelming evidence of differential treatment of blacks and whites. However, tables of these results do not reveal the nature of this association.

Figure 1.7 is an example of a graph designed for *analysis*—a mosaic display (Chapter 5) showing the frequencies of those arrested on this charge by skin color and release type. The size of each

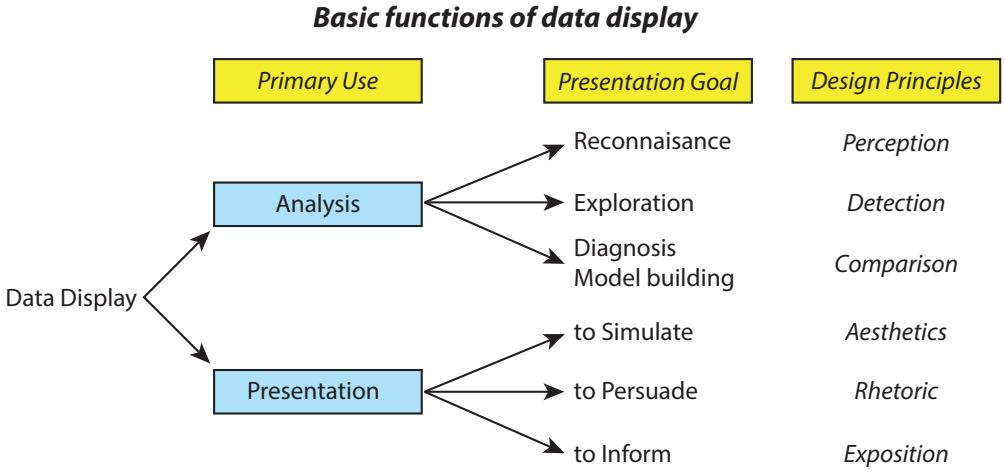


**Figure 1.6:** A taxonomy of the basic functions of data display by intended use, presentation goal, and design principles.

{fig:datadisp}

rectangle shows the frequency and these are shaded in relation to the association between skin color and release—blue for positive associations (more than expected if they were independent) to red for negative associations.

Once you know how to read such graphs, the pattern is clear: blacks were indeed more likely to be held for more severe treatment, whites were more likely to be released with a summons. But this is hardly a graph that would be clear to a general audience, and would require a good deal of explanation.

In contrast, Figure 1.8 shows a redesign of this as a *presentation graphic* prepared by the *Star* and published on December 11, 2002 in conjunction with a meeting between the newspaper and the Toronto Police Services Board to consider the issue of racial profiling. The police vehemently denied that racial profiling was taking place. The revision makes the point immediately obvious and compelling in the following ways:

- It announces the conclusion in the figure title: "Same charge, different treatment."
- The text box at the top provides the context for this conclusion
- Skin colors "Brown" and "Other," which appeared less frequently, were removed, and the release categories "Form 10" and "Form 11.1" were combined as "released at station."
- The graphic is still a mosaic display, however, it now shows explicitly the number of charges laid against whites and blacks and the percentage of each treatment.
- The labels for whites and blacks were enhanced by indicating what a reader should see for each.
- The legend for color is titled non-technically as "degree of likelihood."

Clear communication is not achieved without effort. The revised graph required several iterations and emails between the graphic designer and the statistical consultant (the first author of this book) in the few hours available before the newspaper went to press. The main question was, "what are we trying to show here?" Starting with the original Figure 1.7 mosaic, we asked, "what can we remove?" and "what can we add?" to make the message clearer.

△

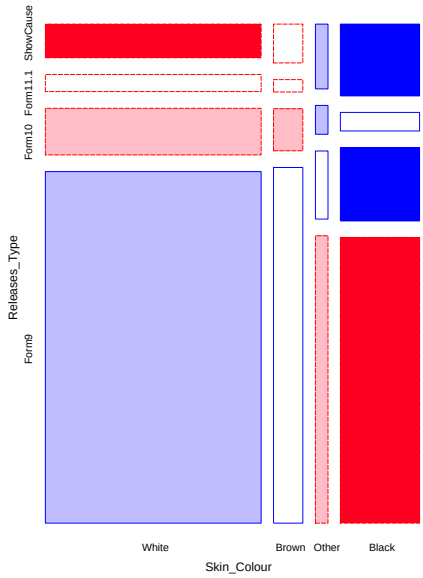


**Figure 1.7:** Mosaic display showing the relationship between skin color and release type for those arrested on a charge of simple possession of marijuana in Toronto, 1996–2002.

{fig:arrests0-mosaic}

### 1.4.2 Categorical data require different graphical methods

{sec:intro-catdata}

We mentioned earlier, and will see in greater detail in Chapter 7 and Chapter 9, that statistical models for discrete response data and for frequency data are close analogs of the linear regression and ANOVA models 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., effect plots, influence plots, added variable and partial residual plots, etc.) that have become common adjuncts in the analysis of quantitative data have been extended to generalized linear models including logistic regression (Section 7.5) and loglinear models (Section 11.6).

Unhappily, the familiar techniques for displaying raw data are often disappointing when applied to categorical data. The simple scatterplot, for example, widely used to show the relation between quantitative response and predictors, when applied to discrete variables, gives a display of the category combinations, with all identical values overplotted, and no representation of their frequency.

Instead, frequencies of categorical variables are often best represented graphically using *areas* rather than as position along a scale. Friendly (1995) describes conceptual and statistical models that give a rationale for this graphic representation. Figure 1.7 does this in the form of a modified bar chart (mosaic plot), where the widths of the horizontal bars show the proportions of whites and blacks in the data, and the divisions of each group give the percents of each release type. Consequently, the areas of each bar are proportional to the frequency in the cells of this $2 \times 3$ table.

As we describe later in this book, using the visual attribute

$$\textbf{area} \sim \textbf{frequency}$$

also allows creating novel graphical displays of frequency data for special circumstances.

Figure 1.9 shows two examples. The left panel gives a ***fourfold display*** of the frequencies of

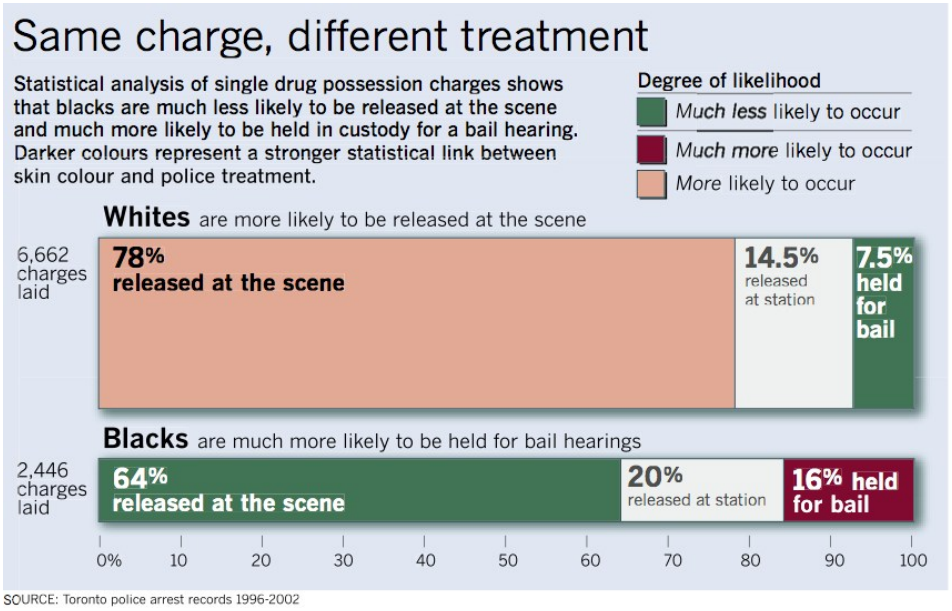


**Figure 1.8:** Redesign of Figure 1.7 as a presentation graphic. *Source*: Graphics department, *The Toronto Star*, December 11, 2002. Used by permission. {fig:arrests0-star}

admission and gender in the Berkeley data shown in Table 1.1. What should be seen at a glance is that males are more often admitted and females more often rejected (shaded blue); see Section 4.4 for details.

The right panel shows another specialized display, an ***agreement chart*** designed to show the strength of agreement in a square table for two raters (see Section 4.7.2). The example here (Example 4.18) concerns agreement of ratings of breast cancer from mammograms by two raters. The dark squares along the diagonal show exact agreement; the lighter diagonal rectangles allow 1-off agreement, and both are shown in relation to chance agreement (diagonal enclosing rectangles). What should be seen at a glance is that exact agreement is moderately strong and extremely strong if you allow the raters to differ by one rating category.

### 1.4.3 Effect ordering and rendering for data display

{sec:effect-order}

In plots of quantitative variables, standard methods (histograms, scatterplots) automatically position values along ordered scales, facilitating comparison ("which is less/more?") and detection of patterns, trends, and anomalies. However, by its nature, categorical data involves discrete variables such as education level, hair color, geographic region (state or province), or preference for a political party. With alphabetic labels for ordered categories (e.g., education: Low, Medium, High), it is unfortunately all too easy to end up with a nonsensical display with the categories ordered High, Low, Medium. Geographic regions (U.S. states) are often ordered alphabetically by default as are the names of political parties and other categorical variables. This may be useful for lookup, but for the purposes of comparison and detection, this is almost always a bad idea.

Instead, Friendly and Kwan (2003) proposed the principle of ***effect-order sorting*** for visual displays (tables as well as graphs):

**sort the data by the effects to be seen to facilitate comparison**

For quantitative data, this is often achieved by sorting the data according to means or medians of

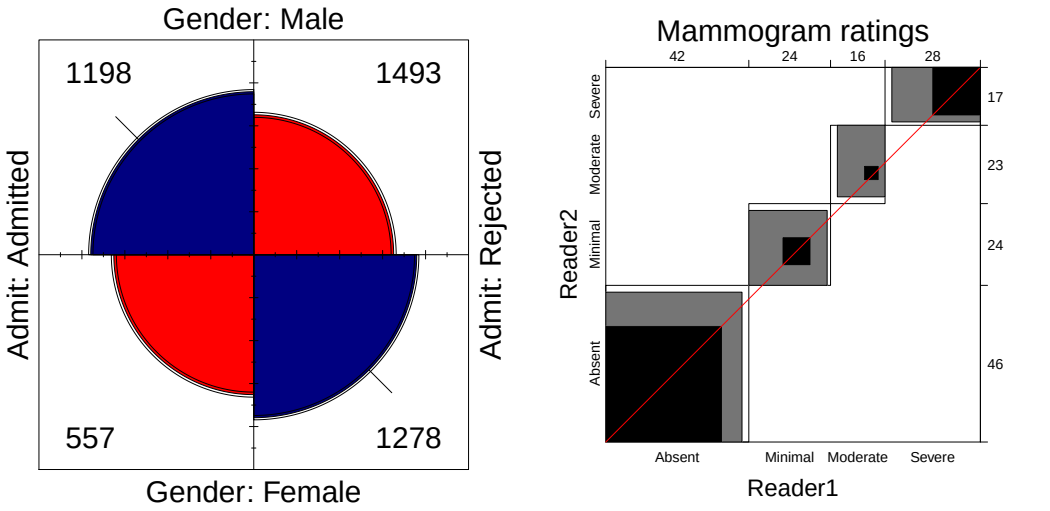


**Figure 1.9:** Frequencies of categorical variables shown as areas. Left: fourfold display of the relation between gender and admission in the Berkeley data; right: agreement plot for two raters assessing mammograms.

{fig:area-diagrams}

row and column factors, called ***main-effect ordering***. For categorical data, graphs and tables are often most effective when the categories are arranged in an order reflecting their association, called ***association ordering***.

Another important principle concerns the ***rendering*** of visual attributes of elements in graphical displays (Friendly, 2002). For example, categorical variables in plots (and tables) can be distinguished by any one or more of color, size, shape, or font. The examples below show the use of color to illustrate the precept:

**render the data by the effects to be seen to facilitate detection**

{ex:glass}

**EXAMPLE 1.5: British social mobility**

Bishop et al. (1975, p. 100) analyzed data on the occupations of 3500 British fathers and their sons from a study by Glass (1954), with five occupational categories: Professional, Managerial, Supervisory, Skilled manual, and Unskilled manual.

One would expect, of course, a strong association between a son's occupation and that of his father—the apple doesn't fall very far from the tree. Mosaic plots (detailed in Chapter 5) provide a natural way to show such relationships. Figure 1.10 shows two such plots. The left panel shows the result obtained when the table variables `father` and `son` are read as factors, and therefore ordered alphabetically by default. It is difficult to see any overall pattern, except for the large values in the diagonal cells (shaded blue) corresponding to equal occupational status.

In the right panel, the categories have been arranged in decreasing order of occupational status to show the association according to status. Now you can see a global pattern of shading color, where the tiles become increasingly red as one moves away from the main diagonal, reflecting a greater difference between the occupation of the father and son. The interpretation here is that most sons remain in their father's occupational class, but when they differ, there is little mobility across large steps.

In this example, `father` and `son` are clearly ordinal variables and should be treated as such in both graphs and statistical models. Correspondence analysis (Chapter 6) provides a natural way to depict association by assigning scores to the categories to optimally represent their relationships.

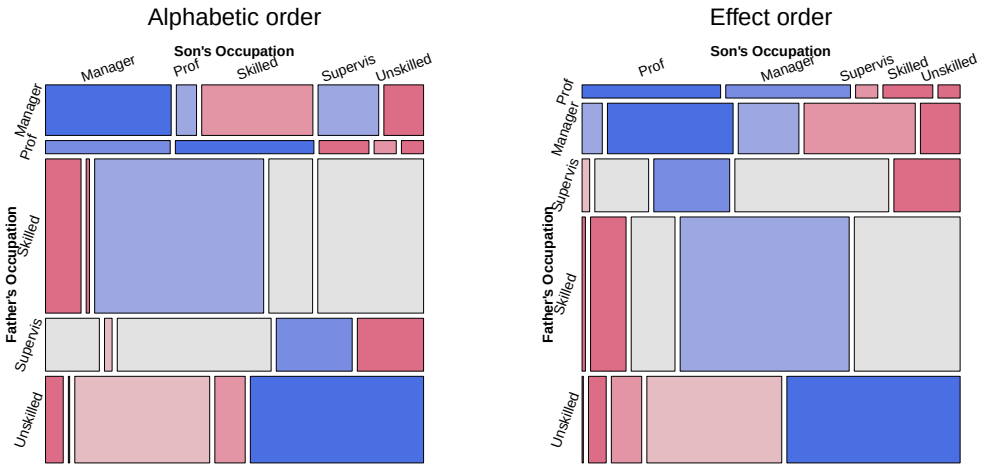


**Figure 1.10:** Mosaic plots for Glass' mobility table of occupational status. In these displays the area of each tile is proportional to frequency and shading color shows the departure from independence, using blue for positive, red for negative association. Left: default alphabetic ordering of categories; right: occupational categories ordered by status.

{fig:glass-mosaic}

Loglinear models provide special methods for ordinal variables (Section 10.1) and square frequency tables (Section 10.2).

△

The ideas of effect ordering and rendering with color shading to enhance perception can also be used in tabular displays, as illustrated in the next example.

{ex:barley}

**EXAMPLE 1.6: Barley data**

The classic `barley` dataset (in lattice (Sarkar, 2015)) from Immer et al. (1934) gives a $10 \times 2 \times 6$ table of yields of 10 varieties of barley in two years (1931, 1932) planted at 6 different sites in Minnesota. Cleveland (1993b) and many others have used this data to illustrate graphical methods, and one surprising finding not revealed in standard tabular displays is that the data for one site (Morris) may have had the values for 1931 and 1932 switched.[4]

To focus attention on this suspicious effect in a tabular display, you can calculate the *yield difference* $\Delta y_{ij} = y_{ij,1931} - y_{ij,1932}$. Table 1.4 shows these values in a $10 \times 6$ table with the rows and columns sorted by their means (main-effect ordering). In addition, the table cells have been colored according to the sign and magnitude of the year difference. The shading scheme uses blue for large positive values and red for large negative values, with a white background for intermediate values. The shading intensity values were determined as $|\Delta y_{ij}| > \{2, 3\} \times \widehat{\sigma}(\Delta y_{ij})$.

Effect ordering and color rendering have the result of revealing a new effect, shown as a regular progression in the body of the table. The negative values for Morris now immediately stand out. In addition, the largely positive other values show a lower-triangular pattern, with the size of the yield difference increasing with both row and column means. Against this background, one other cell, for Velvet grown at Grand Rapids, stands out with an anomalous negative value.

[4] This canonical story, like many others in statistics and graphics lore, turns out to be apocryphal on closer examination. Wright (2013) recently took a closer look at the original data and gives an expanded data set as `minnesota.barley.yield` in the agridat (Wright, 2015) package. With a wider range of years (1927–1936), other local effects like weather had a greater impact than the overall year effects seen in 1931–1932, and the results for the Morris site no longer stand out as surprising.

**Table 1.4:** Barley data, yield differences, 1931-1932, sorted by mean difference, and shaded by value {tab:barley2c}

| Variety | Site | | | | | | *Mean* |
|---|---|---|---|---|---|---|---|
| | Morris | Duluth | University Farm | Grand Rapids | Waseca | Crookston | |
| No. 475 | -22 | 6 | -5 | 4 | 6 | 12 | 0.1 |
| Wisconsin No. 38 | -18 | 2 | 1 | 14 | 1 | 14 | 2.4 |
| Velvet | -13 | 4 | 13 | -9 | 13 | 9 | 2.9 |
| Peatland | -13 | 1 | 5 | 8 | 13 | 16 | 4.8 |
| Manchuria | -7 | 6 | 0 | 11 | 15 | 7 | 5.5 |
| Trebi | -3 | 3 | 7 | 9 | 15 | 5 | 6.1 |
| Svansota | -9 | 3 | 8 | 13 | 9 | 20 | 7.3 |
| No. 462 | -17 | 6 | 11 | 5 | 21 | 18 | 7.4 |
| Glabron | -6 | 4 | 6 | 15 | 17 | 12 | 8.0 |
| No. 457 | -15 | 11 | 17 | 13 | 16 | 11 | 8.8 |
| ***Mean*** | -12.2 | 4.6 | 6.3 | 8.2 | 12.5 | 12.5 | 5.3 |

Although the use of color for graphs is now more common in some journals, color and other rendering details in tables are still difficult. The published version of Table 1.4 (Friendly and Kwan, 2003, Table 3) was forced to use only font shape (normal, italics) to distinguish positive and negative values.

△

Finally, effect ordering is also usefully applied to the variables in multivariate data sets, which by default, are often ordered in data displays according to their position in a data frame or alphabetically. {ex:1.7}

**EXAMPLE 1.7: Iris data**

The classic `iris` data set (Anderson, 1935, Fisher, 1936) gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris, *Iris setosa*, *versicolor*, and *virginica*. Such multivariate data are often displayed in ***parallel coordinate plot***s, using a separate vertical axis for each variable, scaled from its minimum to maximum.

The default plot, with variables shown in their data frame order, is shown in the left panel of Figure 1.11, and gives rise to the epithet *spaghetti plot* for such displays because of the large number of line crossings. This feature arises because one variable, sepal width, has negative relations in the species means with the other variables. Simple rearrangement of the variables to put sepal width last (or first) makes the relations among the species and the variables more apparent, as shown in the right panel of Figure 1.11. This plot has also been enhanced by using ***alpha-blending*** (partial transparency) of thicker lines, so that the density of lines is more apparent.

Parallel coordinate plots for categorical data are discussed in an online supplement on the web site for the book. A general method for reordering variables in multivariate data visualizations based on cluster analysis was proposed by Hurley (2004).

△

### 1.4.4 Interactive and dynamic graphics

{sec:intro-interactive}

Graphics displayed in print form, such as this book, are necessarily static and fixed at the time they are designed and rendered as an image. Yet, recent developments in software, web technology and

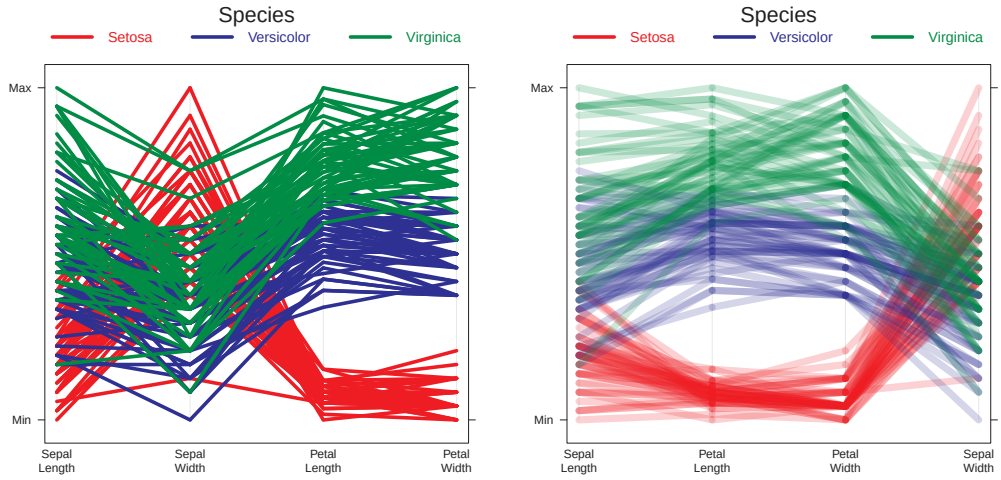


**Figure 1.11:** Parallel coordinates plots of the Iris data. Left: Default variable order; right: Variables ordered to make the pattern of correlations more coherent.

{fig:iris-parallel}

media alternative to print have created the possibility to extend graphics in far more useful and interesting ways, for both presentation and analysis purposes.

Interactive graphics allow the viewer to directly manipulate the statistical and visual components of graphical display. These range from

- graphical controls (sliders, selection boxes, and other widgets) to control details of an analysis (e.g., a smoothing parameter) or graph (colors and other graphic details), to
- higher-level interaction including zooming in or out, drilling down to a data subset, linking multiple displays, selecting terms in a model, and so forth.

The important effect is that the analysis and/or display is immediately re-computed and updated visually.

In addition, ***dynamic graphics*** use animation to show a series of views, as frames in a movie. Adding time as an additional dimension allows far more possibilities, for example showing a rotating view of a 3D graph or showing smooth transitions or interpolations from one view to another.

There are now many packages in R providing interactive and dynamic plots (e.g., rggobi (Temple Lang et al., 2014), iplots (Urbanek and Wichtrey, 2013)) as well as capabilities to incorporate these into interactive documents, presentations, and web pages (e.g., rCharts (Vaidyanathan, 2013), googleVis (Gesmann and de Castillo, 2015), ggvis (Chang and Wickham, 2015)). The animation (Xie, 2014) package facilitates creating animated graphics and movies in a variety of formats. The RStudio editor and development environment[5] provides its own manipulate (RStudio, Inc., 2011) package, as well as the shiny (RStudio, Inc., 2015) framework for developing interactive R web applications.

{ex:512paths}

**EXAMPLE 1.8: 512 paths to the White House**

Shortly before the 2012 U.S. presidential election (November 2, 2012) *The New York Times* published an interactive graphic, [6] designed by Mike Bostock and Shan Carter,[7] showing the effect

[5] `http://www.rstudio.com`.

[6] `http://www.nytimes.com/interactive/2012/11/02/us/politics/paths-to-the-white-house.html`.

[7] see: `https://source.opennews.org/en-US/articles/nyts-512-paths-white-house/`. for a description of their design process.

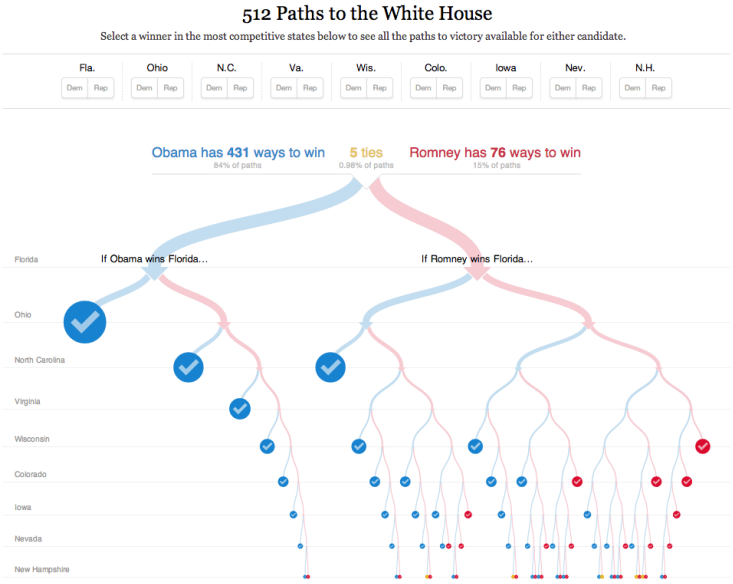


**Figure 1.12:** 512 paths to the White House. This interactive graphic allows the viewer to select a winner in any one or more of the nine most highly contested U.S. states and highlights the number of paths leading to a win by Obama or Romney, sorted and weighted by the number of Electoral College votes. *Source*: Mike Bostock & Shan Carter, *New York Times* interactive, November 2, 2012. Used by permission. {fig:nyt_512paths}

that a win for Barack Obama or Mitt Romney in the nine most highly contested states would have on the chances that either candidate would win the presidency.

With these nine states in play there are $2^9 = 512$ possible outcomes, each with a different number of votes in the Electorial College. In Figure 1.12, a win for Obama in Florida and Virginia was selected, with wins for Romney in Ohio and North Carolina. Most other selections also lead to a win by Obama, but those with the most votes are made most visible at the top. An R version of this chart was created using the rCharts package.[8] The design of this graphic as a ***binary tree*** was chosen here, but another possibility would be a ***treemap*** graphic (Shneiderman, 1992) or a mosaic plot.

△

### 1.4.5 Visualization = Graphing + Fitting + Graphing . . .

{sec:vis}

> Look here, upon this picture, and on this.
>
> Shakespeare, Hamlet

Statistical summaries, hypothesis tests, and the numerical parameters derived in fitted models are designed to capture a particular feature of the data. A quick analysis of the data from Table 1.1, for example, shows that 1198/2691 = 44.5% of male applicants were admitted, compared to 557/1835 = 30.4% of female applicants.

Statistical tests give a Pearson $\chi^2$ of 92.2 with 1 degree of freedom for association between admission and gender ($p < 0.001$), and various measures for the strength of association. Expressed

[8] http://timelyportfolio.github.io/rCharts_512paths/

in terms of the ***odds ratio***, males were apparently 1.84 times as likely to be admitted as females, with 99% confidence bounds $(1.56, 2.17)$. Each of these numbers expresses some part of the relationship between gender and admission in the Berkeley data. Numerical summaries such as these are each designed to compress the information in the data, focusing on some particular feature.

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 we should expect that a single graph shows all that may be seen. In most cases, your 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.

The precept here is that

**Visualization = Graphing + Fitting + Graphing** . . .

where the ellipsis indicates the often iterative nature of this process. Even for descriptive purposes, an initial fit of salient features can be removed from the data, giving residuals (departures from a model). Displaying the residuals may then suggest additional features to account for.

Simple examples of this idea include detrending time series graphs to remove overall and seasonal effects and plots of residuals from main-effect models for ANOVA designs. For categorical data, mosaic plots (Chapter 5) display the unaccounted-for association between variables by shading, as in Figure 1.10. Additional models and plots considered in Section 10.2 can reveal additional structure in square tables beyond the obvious effect that sons tend most often to follow in their fathers' footsteps.

{ex:donner0a}

**EXAMPLE 1.9: Donner Party**

The graphs in Figure 1.3 and Figure 1.4 suggest three different initial descriptions for survival in the Donner party. Yet they ignore all other influences, of which gender and family structure might also be important. A more complete understanding of this data can be achieved by taking these effects into account, both in fitted models and graphs. See Example 7.9 for a continuation of this story. △

{ex:nasa}

**EXAMPLE 1.10: Space shuttle disaster**

The space shuttle *Challenger* mentioned in Example 1.2 exploded 73 seconds after take-off on January 28, 1986. Subsequent investigation presented to the presidential commission headed by William Rogers determined that the cause was failure of the O-ring seals used to isolate the fuel supply from burning gases. The story behind the *Challenger* disaster is perhaps the most poignant missed opportunity in the history of statistical graphics. See Tufte (1997) for a complete exposition. It may be heartbreaking to find out that some important information was there, but the graphmaker 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, tables, and charts submitted by the engineers and concluded that there was insufficient evidence to cancel the flight.

The data relating O-ring failures to temperature were depicted as in Figure 1.13, our candidate for the most misleading graph in history. There had been 23 previous launches of these rockets giving data on the number of O-rings (out of 6) that were seen to have suffered some damage or

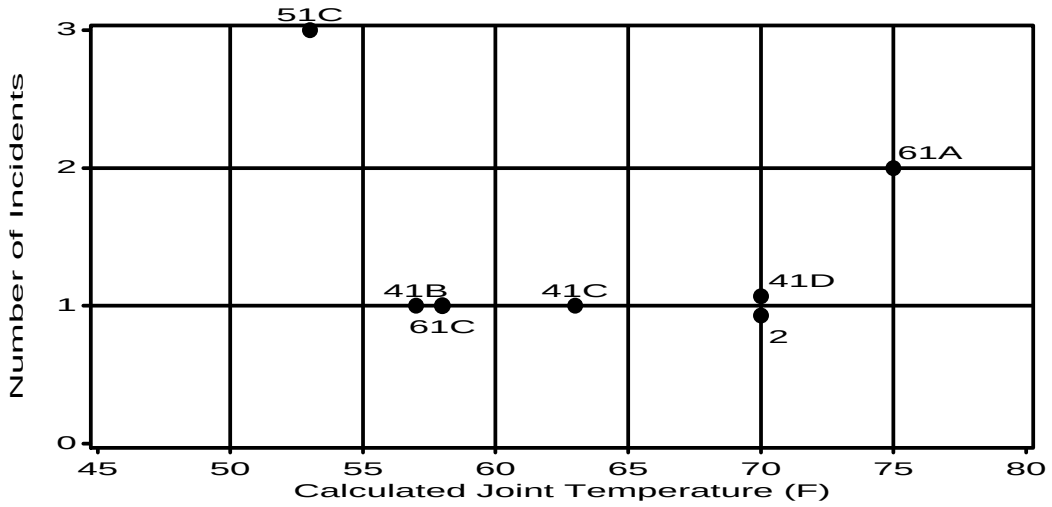


**Figure 1.13:** NASA Space Shuttle pre-launch graph prepared by the engineers at Morton Thiokol. {fig:nasa0}

failure. However, the engineers omitted the observations where no O-rings failed or showed signs of damage, believing that they were uninformative.

Examination of this graph seemed to indicate that there was no relation between ambient temperature and failure. Thus, the decision to launch the *Challenger* was made, in spite of the initial concerns of the Morton Thiokol engineers. Unfortunately, those 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 $31°$.

These data have been analyzed extensively (Dalal et al., 1989, Lavine, 1991). Tufte (1997) gives a thorough and convincing visual analysis of the evidence available prior to the launch. We consider statistical analysis of these data in Chapter 7, Example 7.4.

But, what if the engineers had simply made a better graph? At the very least, that would entail (a) drawing a smoothed curve to fit the points (to show the trend), and (b) removing the background grid lines (which obscure the data). Figure 1.14 shows a revised version of the same graph, highlighting the non-zero observations and adding a simple quadratic curve to allow for a possible non-linear relationship. For comparison, the excluded zero observations are also shown in grey. This plot, even showing only the non-zero points, should have caused any engineer to conclude that either: (a) the data were wrong, or (b) there were excessive risks associated with both high and low temperatures. But it is well-known that brittleness of the rubber used in the O-rings is inversely proportional to temperature cubed, so prudent interest might have focussed on the first possibility.[9]

△

[9] A coda to this story shows the role of visual explanation in practice as well (Tufte, 1997, pp. 50–53). The Rogers Commission contracted the reknowned theoretical physicist Richard Feynman to contribute to their investigation. He determined that the most probable cause of the shuttle failure was the lack of resiliancy of the rubber O-rings at low temperature. But how could he make this point convincingly? At a televised public hearing, he took a piece of the O-ring material, squeezed it in a C-clamp, and plunged it into a glass of ice water. After a few minutes, he released the clamp, and the rubber did not spring back to shape. He mildly said, "... there is no resilience in this particular material when it is at a temperature of 32 degrees. I believe this has some significance for our problem" (Feynman, 1988).

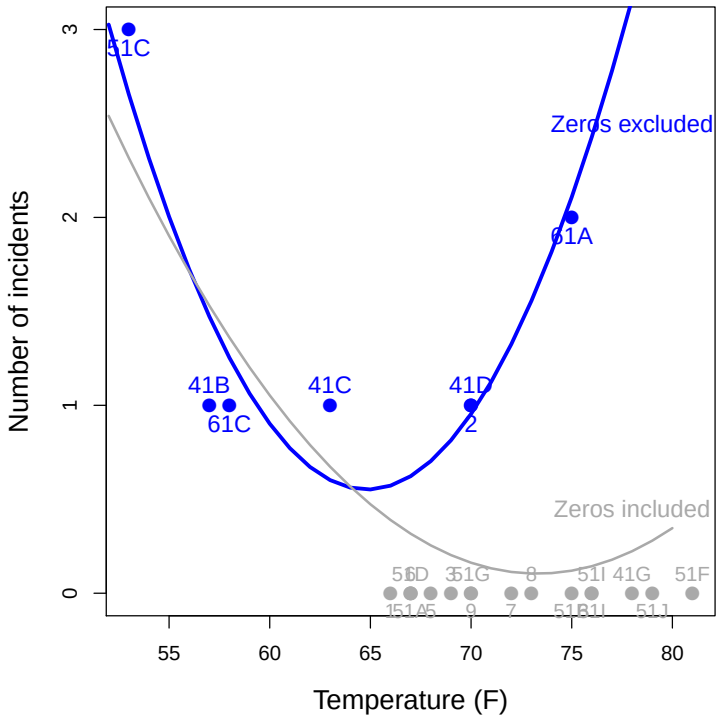


**Figure 1.14:** Re-drawn version of the NASA pre-launch graph, showing the locations of the excluded observations and with fitted quadratics for both sets of observations.

{fig:nasa}

### 1.4.6 The 80–20 rule

The Italian economist Vilfredo Pareto observed in 1906 that 80% of the land in Italy was owned by 20% of the population and this ratio also applied in other countries. It also applied to the yield of peas from peapods in his garden (Pareto, 1971). This idea became known as the ***Pareto principle*** or the ***80–20 rule***. The particular 80/20 ratio is not as important as the more general idea of the uneven distribution of results and causes in a variety of areas.

Common applications are the rules of thumb that: (a) in business 80% of sales come from 20% of clients; (b) in criminology 80% of crimes are said to be committed by 20% of the population. (c) In software development, it is said that 80% of errors and (d) crashes can be eliminated by fixing the top 20% most-reported bugs or that 80% of errors reside in 20% of the code.

The ***Pareto chart*** was designed to display the frequency distribution of a variable with a histogram or bar chart together with a cumulative line graph to highlight the most frequent category, and the ***Pareto distribution*** gives a mathematical form to such distributions with a parameter $\alpha$ (the *Pareto index*) reflecting the degree of inequality.

Applied to statistical graphics, the precept is that

**20% of your effort can generate 80% of your desired result in producing a given plot.**

This is good news for exploratory graphs you produce for yourself. Very often, the default settings will give a reasonable result, or you will see immediately something simple to add or change to make the plot easier to understand.

The bad news is the corollary of this rule:

**80% of your effort may be required to produce the remaining 20% of a finished graph.**

This is particularly important for presentation graphs, where several iterations may be necessary to get it right (or right enough) for your communication purposes. Some important details are:

**graph title** A presentation graphic can be more effective when it announces the main point or conclusion in the graphic title, as in Figure 1.8.

**axis and value labels** Axes should be labelled with meaningful variable descriptions (and perhaps the data units) rather than just plot defaults (e.g., "Temperature (degrees F)" in Figure 1.2, not `temp`). Axis values are often more of a challenge for categorical variables, where their text labels often overlap, requiring abbreviation, a smaller font, or text rotation.

**grouping attributes** Meaningfully different subsets of the data should be rendered with distinct visual attributes such as color, shape, and line style, and sometimes with more than one.

**legends and direct labels** Different data groups in a graphic display shown by color, shape, etc., usually need at least a graphic legend defining the symbols and group labels. Sometimes you can do better by applying the labels directly to the graphical elements,[10] as was done in Figure 1.14.

**legibility** A common failure in presentation graphs in journals and lectures is the use of text fonts too small to be read easily. One rule of thumb is to hold the graph at arms length for a journal and put it on the floor for a lecture slide. If you can't read the labels, the font is too small.

**plot annotations** Beyond the basic graphic data display, additional annotations can add considerable information to interpret the context or uncertainty, as in the use of plot envelopes to show confidence bands or regions (see Figure 1.3 and Figure 1.4).

**aspect ratio** Line graphs (such as Figure 3.1) are often easiest to understand when the ratio of height to width is such that line segments have an average slope near 1.0 (Cleveland et al., 1988). In R, you can easily manipulate a graph window manually with a mouse to observe this effect and find an aspect ratio that looks right.

Moreover, in graphs for biplots and correspondence analysis (Chapter 6), interpretation involves distances between points and angles between line segments. This requires an aspect ratio that equates the units on the axes. Careful software will do this for you,[11] and you should resist the temptation to re-shape the plot.

**colors** Whereas a good choice of colors can greatly enhance a graphical display, badly-chosen colors, ignoring principles of human perception, can actually spoil it. First, considering that graphs are often reproduced in black and white and a significant percentage of the human population is affected by color deficiencies, important information should not be coded by color alone without careful thought.

Second, color palettes should be chosen carefully to put the desired emphasis on the information visualized. For example, consider Figure 1.15 showing qualitative color palettes (appropriate for unordered categories) taken from two different color spaces: Hue-Saturation-Value (HSV) and Hue-Chroma-Luminance (HCL), where only the hue is varied. Whereas one would expect such a palette to be balanced with respect to colorfulness and brightness, the red colors in the

[10] For example, the `identify()` function allows points in a plot to be labeled interactively with a mouse. The directlabels (Hocking, 2013) package provides a general method for a variety of plots.

[11] For example using the graphics parameter `asp=1`, `eqsplot()` in MASS (Ripley, 2015), or the equivalents in lattice (`aspect="iso"`) and ggplot2 (`coord_equal`).

left (HSV) color wheel are generally perceived to be more more intense and flashy than the corresponding blue colors, and the highly saturated dark blue dominates the wheel. Consequently, areas shaded with these colors may appear more important than others in an uncontrolled way, distracting from the information to be conveyed. In contrast, the colors from the right (HCL) wheel are all balanced to the same gray level and in "harmony." These clearly should be preferred whenever categories of the same importance shall be compared.

Another related perception rule prescribes that lighter and darker colors should not be mixed in a display where areas should be compared since lighter colors look larger than darker ones. More background information on the choice of "good" colors for statistical graphics can be found in Zeileis et al. (2009).

**visual impact** Somewhat related, important features of a display should be visually distinguished from the less important. This may be achieved by different color or gray shading levels, or simply by contrasting filled with non-filled geometric shapes, or a different density of shading lines. One useful test for visual impact is to put a printed copy of a graph on the floor, rise up, and see what stands out.

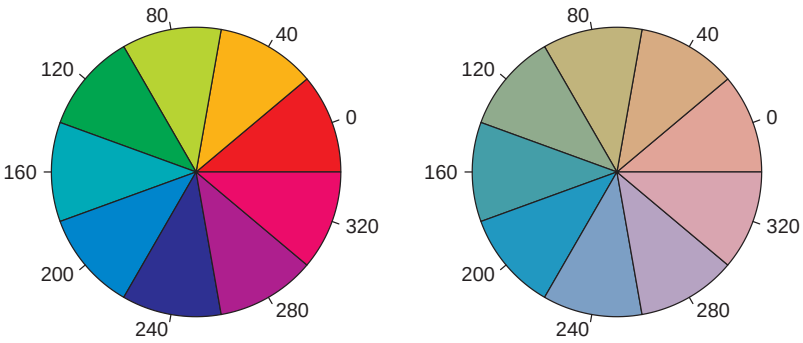


**Figure 1.15:** Qualitative color palette for the HSV (left) and HCL (right) spaces. The HSV colors are $(H, 100, 100)$ and the HCL colors $(H, 50, 70)$ for the same hues $H$. Note that in a monochrome version of this page, all pies in the right wheel will be shaded with the same gray, i.e., they will appear to be virtually identical. {fig:colors}

Nearly all of the graphs in this book were produced using R code in scripts saved as files. This has the advantages of reproducibility and enhancement: just re-run the code, or tweak it to improve a graph. If this is too hard, you can always use an external graphics editor (Gimp, Inkscape, Adobe Illustrator, etc.) to make improvements manually.

## 1.5 Chapter summary

- Categorical data differs from quantitative data because the variables take on discrete values (ordered or unordered, character or numeric) rather than continuous numerical values. Consequently, such data often appear in aggregated form representing category frequencies or in tables.

- Data analysis methods for categorical data are comprised of those concerned mainly with testing particular hypotheses versus those that fit statistical models. Model building methods have the advantages of providing parameter estimates and model-predicted values, along with measures of uncertainty (standard errors).

- Graphical methods can serve different purposes for different goals (data analysis versus presentation), and these suggest different design principles that a graphic should respect to achieve a given communication goal.
- For categorical data, some graphic forms (bar charts, line graphs, scatterplots) used for quantitative data can be readily adapted to discrete variables. However, frequency data often requires novel graphics using area and other visual attributes.
- Graphics can be far more effective when categorical variables are ordered to facilitate comparison of the effects to be seen and rendered to facilitate detection of patterns, trends or anomalies.
- The visualization approach to data analysis often entails a sequence of intertwined steps involving graphing and model fitting.
- Producing effective graphs for presentation is often hard work, requiring attention to details that support or detract from your communication goal.

## 1.6 Lab exercises

{sec:ch01-exercises}

{lab:1.1}

**Exercise 1.1** A web page, "The top ten worst graphs, " `http://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/` by Karl Broman lists his picks for the worst graphs (and a table) that have appeared in the statistical and scientific literature. Each entry links to graph(s) and a brief discussion of what is wrong and how it could be improved.

(a) Examine a number of recent issues of a scientific or statistical journal in which you have some interest. Find one or more examples of a graph or table that is a particularly bad use of display material to summarize and communicate research findings. Write a few sentences indicating how or why the display fails and how it could be improved.
(b) Do the same task for some popular magazine or newspaper that uses data displays to supplement the text for some story. Again, write a few sentences describing why the display is bad and how it could be improved.

{lab:1.2}

**Exercise 1.2** As in the previous exercise, examine the literature in recent issues of some journal of interest to you. Find one or more examples of a graph or table that you feel does a *good* job of summarizing and communicating research findings.

(a) Write a few sentences describing why you chose these displays.
(b) Now take the role of a tough journal reviewer. Are there any features of the display that could be modified to make them more effective?

{lab:1.3}

**Exercise 1.3** Infographics are another form of visual displays, quite different from the data graphics featured in this book, but often based on some data or analysis. Do a Google image search for the topic "Global warming" to see a rich collection.

(a) Find and study one or two images that attempt some visual explanation of causes and/or effects of global warming. Describe the main message in a sentence or two.
(b) What visual and graphic features are used in these to convey the message?

{lab:1.4}

**Exercise 1.4** The Wikipedia web page `en.wikipedia.org/wiki/Portal:Global_warming` gives a few data-based graphics on the topic of global warming. Read the text and study the graphs.

(a) Write a short figure title for each that would announce the conclusion to be drawn in a presentation graphic.

(b) Write a figure caption for each that would explain what is shown and the important graphical details for a reader to understand.

{lab:1.5}

**Exercise 1.5** The R Graph Gallery, `http://rgraphgallery.blogspot.com/`, contains a large collection of examples of graphs in R, tagged by type or content, together with the R code to produce them. Explore this collection for the terms (a) association plot (b) bar chart (c) categorical data (d) fluctuation diagram (e) mosaic plot Find one or two you particularly like and write a few sentences saying why you do.

# References


Anderson, E. (1935). The irises of the Gaspé peninsula. *Bulletin of the American Iris Society*, 35, 2–5.

Bertin, J. (1983). *Semiology of Graphics*. Madison, WI: University of Wisconsin Press. (trans. W. Berg).

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Brinton, W. C. (1939). *Graphic Presentation*. New York, NY: Brinton Associates.

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

Chang, W. and Wickham, H. (2015). *ggvis: Interactive Grammar of Graphics*. R package version 0.4.1.

Cleveland, W. S. (1993a). A model for studying display methods of statistical graphics. *Journal of Computational and Graphical Statistics*, 2, 323–343.

Cleveland, W. S. (1993b). *Visualizing Data*. Summit, NJ: Hobart Press.

Cleveland, W. S., McGill, M. E., and McGill, R. (1988). The shape parameter of a two-variable graph. *Journal of the American Statistical Association*, 83, 289–300.

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

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

Dalal, S., Fowlkes, E. B., and Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-*Challenger* prediction of failure. *Journal of the American Statistical Association*, 84(408), 945–957.

Feynman, R. P. (1988). *What Do You Care What Other People Think? Further Adventures of a Curious Character*. New York: W. W. Norton.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 8, 379–388.

Friendly, M. (1991). *SAS System for Statistical Graphics*. Cary, NC: SAS Institute, 1st edn.

Friendly, M. (1995). Conceptual and visual models for categorical data. *The American Statistician*, 49, 153–160.

Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56(4), 316–324.

Friendly, M. and Kwan, E. (2003). Effect ordering for data displays. *Computational Statistics and Data Analysis*, 43(4), 509–539.

Friendly, M. and Kwan, E. (2011). Comment (graph people versus table people). *Journal of Computational and Graphical Statistics*, 20(1), 18–27.

Gesmann, M. and de Castillo, D. (2015). *googleVis: R Interface to Google Charts*. R package version 0.5.8.

Glass, D. V. (1954). *Social Mobility in Britain*. Glencoe, IL: The Free Press.

Hocking, T. D. (2013). *directlabels: Direct labels for multicolor plots in lattice or ggplot2*. R package version 2013.6.15.

Hurley, C. B. (2004). Clustering visualizations of multidimensional data. *Journal of Computational and Graphical Statistics*, 13, 788–?806.

Immer, F. R., Hayes, H., and Powers, L. R. (1934). Statistical determination of barley varietial adaptation. *Journal of the American Society of Agronomy*, 26, 403–419.

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

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

Lavine, M. (1991). Problems in extrapolation illustrated with space shuttle O-ring data. *Journal of the American Statistical Association*, 86, 912–922.

Lewandowsky, S. and Spence, I. (1989). The perception of statistical graphs. *Sociological Methods & Research*, 18, 200–242.

Lindsey, J. K. (1995). *Modelling Frequency and Count Data*. Oxford, UK: Oxford University Press.

Meyer, D., Zeileis, A., and Hornik, K. (2015). *vcd: Visualizing Categorical Data*. R package version 1.3-3.

Pareto, V. (1971). *Manuale di economia politica ("Manual of political economy")*. New York: A.M. Kelley. Translated by Ann S. Schwier. Edited by Ann S. Schwier and Alfred N. Page.

Ripley, B. (2015). *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. R package version 7.3-40.

RStudio, Inc. (2011). *manipulate: Interactive Plots for RStudio*. R package version 0.98.977.

RStudio, Inc. (2015). *shiny: Web Application Framework for R*. R package version 0.11.1.

Sarkar, D. (2015). *lattice: Lattice Graphics*. R package version 0.20-31.

Shneiderman, B. (1992). Tree visualization with treemaps: A 2-D space-filling approach. *ACM Transactions on Graphics*, 11(1), 92–99.

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

Spence, I. and Lewandowsky, S. (1990). Graphical perception. In J. Fox and J. S. Long, eds., *Modern Methods of Data Analysis*, chap. 1, (pp. 13–57). Sage Publications, Inc.

Temple Lang, D., Swayne, D., Wickham, H., and Lawrence, M. (2014). *rggobi: Interface between R and GGobi*. R package version 2.1.20.

Tufte, E. (2006). *Beautiful Evidence*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1990). *Envisioning Information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1997). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press.

Urbanek, S. and Wichtrey, T. (2013). *iplots: iPlots - interactive graphics for R*. R package version 1.1-7.

Vaidyanathan, R. (2013). *rCharts: Interactive Charts using Javascript Visualization Libraries*. R package version 0.4.5.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer New York.

Wickham, H. and Chang, W. (2015). *ggplot2: An Implementation of the Grammar of Graphics*. R package version 1.0.1.

Wilkinson, L. (2005). *The Grammar of Graphics*. New York: Springer, 2nd edn.

Wright, K. (2013). Revisiting immer's barley data. *The American Statistician*, 67(3), 129–133.

Wright, K. (2015). *agridat: Agricultural Datasets*. R package version 1.11.

Xie, Y. (2014). *animation: A gallery of animations in statistics and utilities to create animations*. R package version 2.3.

Zeileis, A., Hornik, K., and Murrell, P. (2009). Escaping RGBland: Selecting colors for statistical graphics. *Computational Statistics & Data Analysis*, 53, 3259–3270.