

Chapter 1

Introduction



{ch:intro}

Categorical data consists 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}


Beauty is truth; truth, beauty.
That is all ye know on Earth, all ye need to know.

John Keats, *Ode on a Grecian urn*

“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 which 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” ~~data~~  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 χ^2 , 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 7–8 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 1.5 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, ...

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

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

status 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.

b: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` package. The following lines show the first five of $N = 84$ cases in the `Arthritis` data,

```
data("Arthritis", package="vcd")
head(Arthritis, 5)

##      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
```

¹An 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.


²This may be wasteful 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.

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  observation back to its individual identifier or data record (for example, if the case with `ID==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.

```
as.data.frame(xtabs(~Treatment+Sex+Improved, data=Arthritis))
```


##	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  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 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 Table 1.1 and Table 1.2 are both examples of frequency data because each observation 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, 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, 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.

ab: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	1033	1343	1112	829	478	181	45	7

:uni-multi}

1.2.3 Univariate, bivariate, and multivariate data

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 × 3 table for Treatment and Improved.

```
as.data.frame(xtabs(~Treatment + Improved, data=Arthritis))

##      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 19th 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.

However, 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 8), 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}

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

{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 2×2 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, and tests for **homogeneity of association** help 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-square 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

Chapter 4–6, though the emphasis is on graphical methods which help to understand the nature of association between variables.

{ex:haireye0}

EXAMPLE Hair color and eye color

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

```
library(vcd)
(HairEye <- margin.table(HairEyeColor, c(1, 2)))

##           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 χ^2 of 138.3 with nine degrees of freedom, indicating substantial departure from independence. Among the measures of strength of association, the **phi coefficient**, $\phi = \sqrt{\chi^2/N} = 0.483$, indicates a substantial relationship between hair and eye color.

```
assocstats(HairEye)

##           X^2 df P(> X^2)
## Likelihood Ratio 146.44  9      0
## Pearson          138.29  9      0
##
## Phi-Coefficient   : 0.483
## 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.

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 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 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 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. \triangle

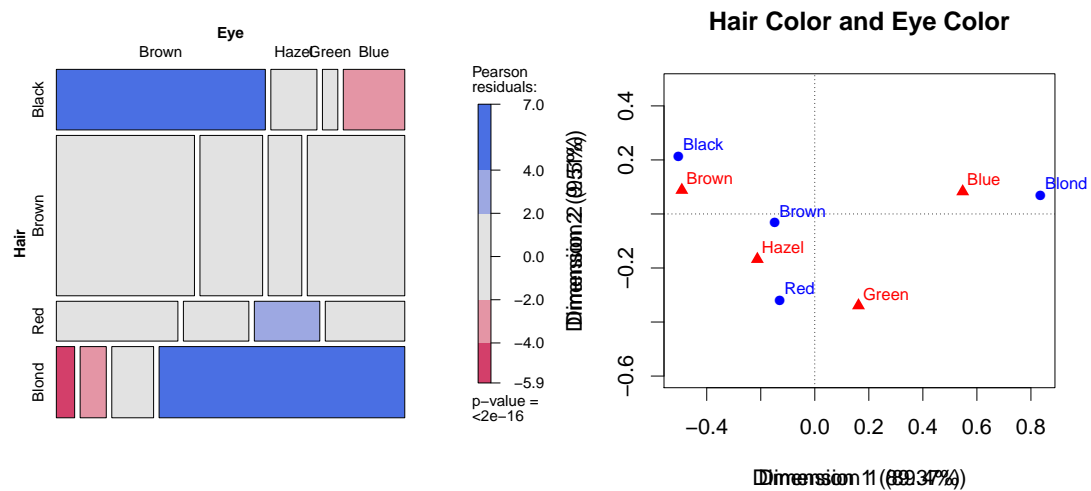


Figure 1.1: Graphical displays for the hair color and eye color data. Left: mosaic display; right: correspondence analysis plot.

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) 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, π , 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

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

Statistical inferences from model-based methods provide tests of hypotheses for the effects of the predictors, x_1, x_2, \dots , but they also provide estimates of parameters in the model, β_1, β_2, \dots 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.

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