Chapter 15
Bivariate Statistics—Discrete Data

In this chapter we discuss bivariate discrete distributions. Bivariate means that there are two factors (categorical variables) defining cells. The response values are frequencies, that is, counts or instances of observations, at each cell.

It is convenient to arrange such data in a contingency table, that is, a table with r rows representing the possible values of one categorical variable and c columns representing the possible values of the other categorical variable. Each of the rc cells of the table contains an integer, the number of observations having the levels of the two variables specified by the cell location. We give extra attention to the special case where $r = c = 2$, that is, a 2×2 contingency table.

This data type is different from situations with two categorical variables (factors) described in Chapters 12 through 14. In those chapters the response variables are one or more continuous measurements at each cell.

We introduce a relatively new form of graph, the mosaic plot, which along with related plots we access through the R package **vcd** "Visualizing Categorical Data". In simplest terms, a mosaic plot consists of a tiling of rectangles (hence the name) where the height, width, and area (product of height and width) are each interpretable.

15.1 Two-Dimensional Contingency Tables—Chi-Square Analysis

15.1.1 Example—Drunkenness Data

Table 15.1 shows the number of persons convicted of drunkenness in two London courts during the first six months of 1970. The data come from Cook (1971), later reprinted in Hand et al. (1994). The dataset is accessed as `data(drunk)`. There are

R.M. Heiberger, B. Holland, *Statistical Analysis and Data Display*,
Springer Texts in Statistics, DOI 10.1007/978-1-4939-2122-5_15

two rows, males and females. The five columns are five age categories. The question of interest is whether the age distribution of convicted offenders is the same for both genders, or equivalently, as illustrated in the bar chart of Figure 15.1, if the proportion of female offenders is the same for all age categories.

We show the mosaic plot of the same data in Figure 15.2.

Table 15.1 Persons convicted of drunkenness in two London courts during the first six months of 1970.

Age group	0–29	30–39	40–49	50–59	≥60
Number of males	185	207	260	180	71
Number of females	4	13	10	7	10

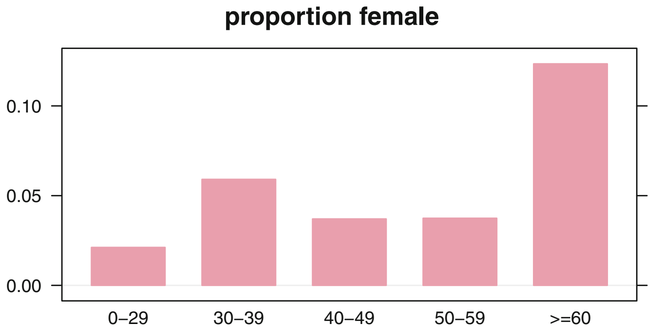


Fig. 15.1 Proportion female for drunkenness data.

The tabular display from the chi-square analysis (to be described in Section 15.1.2) is in Table 15.2.

The p-value .0042 for the chi-square test strongly suggests an association between age and gender of convicted offenders, that is, the proportion of females in each age group is not identical. From the "Cell Chi-square" values in Table 15.2 and their square roots displayed in Figure 15.3, it seems that only 1 of the 10 cells contributes appreciably to total chi-square. The cell for female offenders aged at least 60 contributes 10.34, or 68%, of the total chi-square value 15.25. We observe 10 female offenders aged at least 60, but under the null hypothesis of independence we expect only 3.8 offenders. No other cell is suggestive of dependence.

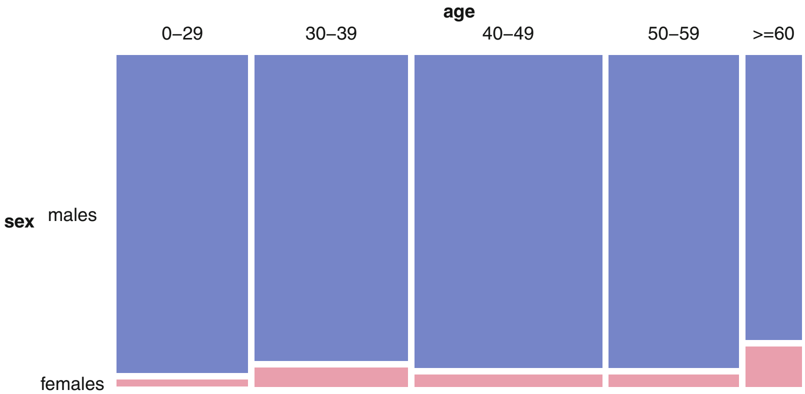


Fig. 15.2 Mosaic plot of the drunkenness data. The widths of the sets of tiles for each age group are proportional to the counts of all persons in that age groups. The heights of the bottom set of tiles (females) are the same as the heights of the bars in Figure 15.1 and are the proportions female within each age group. The area of each tile is proportional to the count of individuals of that sex and age group.

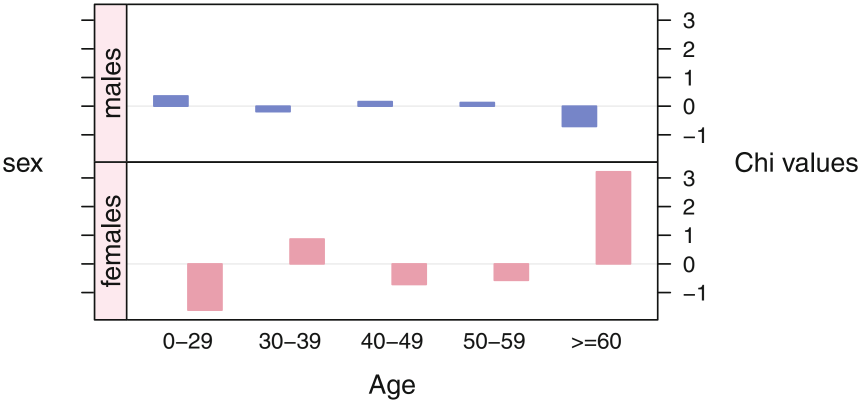


Fig. 15.3 Cell chi-deviations (residuals from Table 15.2, also the signed square root of the cell chi-square values $\chi_{ij} = (n_{ij} - e_{ij})/\sqrt{e_{ij}}$) for drunkenness data. The signed heights of these bars are the same as the signed heights of the bars in the association plot in Figure 15.4.

We must be careful not to overinterpret this finding as meaning that older females have a greater tendency toward this crime than older males. We believe that the finding may be an artifact of the demographic distribution. The population proportion of females under the age of 60 is roughly 50%. This proportion tends to increase after age 60 because of higher male mortality beginning at approximately that age. Therefore, it is possible that this study could have been improved by adjusting the responses to a per capita basis.

Table 15.2 Chi-square analysis of Drunkenness Data.

```
> drunk.chisq <- chisq.test(drunk)

> drunk.chisq

Pearson's Chi-squared test

data:  drunk
X-squared = 15.25, df = 4, p-value = 0.004217


> drunk.chisq$observed
         age
sex        0-29 30-39 40-49 50-59 >=60
  males     185   207   260   180   71
  females     4    13    10     7   10

> drunk.chisq$expected
         age
sex           0-29  30-39  40-49   50-59   >=60
  males    180.219 209.78 257.46 178.312 77.237
  females    8.781  10.22  12.54   8.688  3.763

> drunk.chisq$residuals   ## cell chi values
         age
sex           0-29    30-39    40-49    50-59    >=60
  males     0.3562 -0.1918   0.1586   0.1264 -0.7096
  females  -1.6135  0.8690  -0.7185  -0.5728  3.2148

> drunk.chisq$residuals^2 ## cell chi-square values
         age
sex            0-29     30-39     40-49     50-59      >=60
  males     0.12686   0.03679   0.02516   0.01599   0.50357
  females   2.60344   0.75512   0.51626   0.32814  10.33473
```

15.1.2 Chi-Square Analysis

When we work with two-dimensional tables, such as this example, we often want to test whether the row and column classifications are independent, that is, whether the probability of an entry's being in a particular row is independent of the entry's column.

When $r = 2$, the test is essentially asking whether the proportion of data in Row 1 is homogeneous across the c columns, i.e., whether c binomial populations have the same (unspecified) proportion parameter. Therefore, this is a generalization of the inferences comparing $c = 2$ population proportions discussed in Chapter 5 to $c \geq 2$ population proportions.

If the total number of observations n is sufficiently large and if none of the rc expected cell frequencies is less than somewhere between 3 and 5, the chi-square distribution may be used to test the hypothesis of independence. The logical idea behind this test is to compare, in each of the cells,

n_{ij} the observed frequency in the cell in row i and column j

with

e_{ij} the expected frequency calculated under the assumption
that the independence null hypothesis is true.

The test statistic is a function of the aggregate discrepancy between the n_{ij}'s and e_{ij}'s.

Define

$$\begin{cases} n_{i.} &= \sum_j n_{ij} \text{ the row totals} \\ n_{.j} &= \sum_i n_{ij} \text{ the column totals} \\ n = n_{..} &= \sum_{i,j} n_{ij} \text{ the grand total} \end{cases} \tag{15.1}$$

Under the null hypothesis of independence between rows and columns, the expected frequency for the cell in row i and column j is

$$e_{ij} = \frac{n_{i.}\, n_{.j}}{n} \tag{15.2}$$

The cell residuals, also called chi-deviations or scaled deviations,

$$\chi_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}} \tag{15.3}$$

are displayed in Figure 15.3. The squares, the cell chi-square values χ^2_{ij}, are displayed in Figure 15.3 and also displayed in Table 15.2.

The `assoc` function in the **vcd** package provides an alternative plot of the chi-deviations. In Figure 15.4 we see the association plot for the `drunk` dataset. The heights of the rectangles are proportional to the cell residual values χ_{ij}, the widths of the rectangles are proportional to the square root of the cell expected values $\sqrt{e_{ij}}$, and therefore the areas are proportional to the difference in observed and expected frequencies $n_{ij} - e_{ij}$.

The test statistic is the sum of the scaled deviations squared

$$\hat{\chi}^2 = \sum_{ij} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{15.4}$$

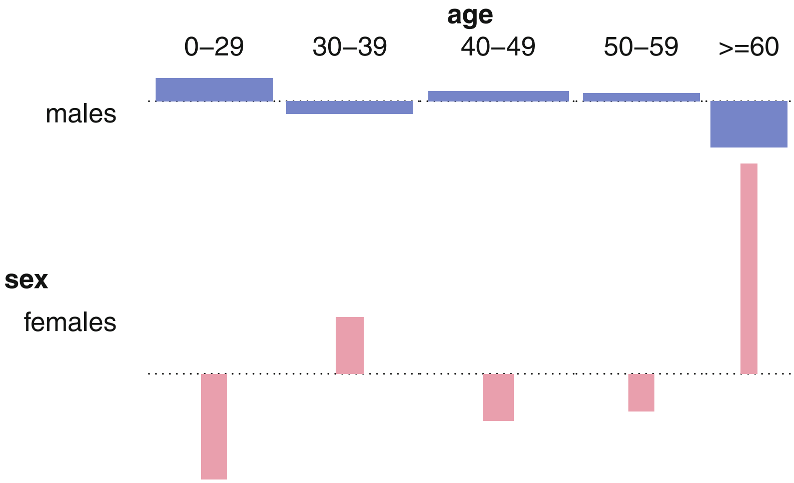


Fig. 15.4 Association plot for the `drunk` dataset. The heights of the rectangles are proportional to the cell residual values $\chi_{ij} = (n_{ij} - e_{ij})/\sqrt{e_{ij}}$ (the same heights as in Figure 15.3), the widths of the rectangles are proportional to the square root of the cell expected values $\sqrt{e_{ij}}$, and therefore the areas are proportional to the difference in observed and expected frequencies $n_{ij} - e_{ij}$.

where the sum is taken over all rows and columns. If the null hypothesis is true, $\hat{\chi}^2$ has, approximately, a chi-square distribution with $(r-1)(c-1)$ degrees of freedom, and the p-value is $1-\mathcal{F}_{\chi^2}(\hat{\chi}^2 \mid (r-1)(c-1))$, the chi-square tail probability associated with $\hat{\chi}^2$. Most authorities agree that the chi-square approximation is good if almost all e_{ij} are at least 5 and none is less than 3. The degrees-of-freedom formula derives from the fact that if all marginal totals are given, knowledge of $(r-1)(c-1)$ interior values uniquely determines the remaining $r + c - 1$ interior values.

Apart from the degrees-of-freedom calculation, the form of this test, comparing observed and expected frequencies, is identical to the goodness-of-fit test described in Section 5.7.1. This methodology can be extended to contingency tables having more than two dimensions.

Further analysis is required to assess the nature of any lack of independence that is suggested by the chi-square test. One approach to this is discussed in the next paragraph. Another is a multivariate display technique called *correspondence analysis*. See Greenacre (1984) for an introduction to this topic.

The chi-square test of independence, shown in Table 15.2, is the default display from `chisq.test`. The result of the test also contains each cell's observed value n_{ij}, expected value e_{ij}, residual (square root of its contribution to the chi-square statistic) χ_{ij}.

Cells with a sizeable cell chi-square value have an appreciable discrepancy between their observed and expected frequency. Scrutiny of such cells leads to interpretation of the nature of the dependence between rows and columns. A cell chi-square is calculated as $(n - e)^2/e$, where n and e are, respectively, the cell's observed frequency and expected frequency under the null hypothesis. Under the model that observations are randomly assigned to cells with a Poisson distribution, the variance of n is also e. Hence $(n - e)^2/e$ has the form

$$\frac{(n - E(n))^2}{\text{var}(n)}$$

and, using the normal approximation, we interpret it as approximately a one-df chi-square statistic. Since the 95th and 99th percentiles of this chi-square distribution are 3.84 and 6.63, we recommend reporting the discrepancy between the observed and expected frequency for all cells with cell chi-square exceeding the higher value. Also, consideration should be given to reporting cells having chi-square between 3.84 and 6.63 when the discrepancy between the cell's observed and expected frequency can be meaningfully interpreted.

15.2 Two-Dimensional Contingency Tables—Fisher's Exact Test

An alternative to the approximate chi-square statistic discussed above is Fisher's exact test, which uses the exact hypergeometric distribution probabilities calculated for all tables at least as extreme in the alternative hypothesis direction as the existing table. Since it is exact, this procedure, when available, is preferable to the chi-square test, but it is extremely computer intensive for all but the smallest tables, even by contemporary standards.

Fisher's exact test is available in R as `fisher.test(x)`, where `x` is a two-dimensional contingency table in matrix form. For tables larger than 2 × 2, only the two-sided test is available.

15.2.1 Example—Do Juvenile Delinquents Eschew Wearing Eyeglasses?

Weindling et al. (1986) discuss a small study of juvenile delinquent boys and a control group of nondelinquents. The data in Table 15.3 also appear in Hand et al. (1994). All of these subjects failed a vision test. The boys were also classified according to whether or not they wore glasses.

Table 15.3 Wearing prescribed glasses and juvenile delinquency.

	Juvenile delinquents	Nondelinquents
Wears glasses	1	5
Doesn't wear glasses	8	2

Clearly, the data set is much too small to use the chi-square analysis of the previous section. Therefore, we request an analysis using Fisher's exact test, shown in Table 15.4. We are interested in the two-sided p-value, .0350. Since this falls between the two thresholds .01 and .05, we can say there is suggestive but inconclusive evidence that a smaller proportion of delinquents than nondelinquents wear glasses.

Table 15.4 Fisher's exact test of glasses data. Is the proportion of delinquents who wear glasses the same as the proportion of nondelinquents who wear glasses Fisher's exact test for glasses data. The p-value for the two-sided exact test is .035. Compare this to p = .0134 from the (uncorrected) chi-square approximation that R warns might be incorrect.

```
> fisher.test(glasses)

Fisher's Exact Test for Count Data

data:  glasses
p-value = 0.03497
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.0009526 0.9912282
sample estimates:
odds ratio
   0.06464
```

The calculation of the p-values for Fisher's exact test utilizes the hypergeometric probability distribution discussed in Appendix J to calculate the probability of obtaining the observed table and "more extreme" tables assuming that the table's marginal totals are fixed. We illustrate the calculations in Table 15.5. The observed table is shown in Column **1**. The remaining columns, indexed by the `[1,1]` cell count, show all possible tables with the same row and column margins. The probability of observing the counts in Table 15.3 (identical to Column **1**, marked "$*$", in Table 15.5) given this table's marginal totals is

$$\frac{\binom{9}{1}\binom{7}{5}}{\binom{16}{6}} = 0.0236$$

The one-sided p-value is the sum of this probability and the probability, 0.0009, of the more extreme table **0** on the same tail of the distribution.

Table 15.5 All possible 2×2 tables with the same margins as the observed glasses table. In these tables `glasses` denotes "wears glasses" and `del` denotes "delinquent." We show the probabilities of each under the assumption of a hypergeometric distribution for the `[1,1]` position and the common row and column cell margins. The observed table **1** is marked with "∗" and the more extreme tables in the same tail **0** and opposite tail **6** are marked "<". The mosaic plots of each of the tables are shown in Figure 15.5. The probabilities of each of the tables are graphed in Figure 15.6.

	`[1,1]` cell count													
glasses	**0**		**1**		**2**		**3**		**4**		**5**		**6**	
	del	no.del	del	no.del	del	no.del	del	no.del	del	no.del	del	no.del	del	no.del
glasses	0	6	1	5	2	4	3	3	4	2	5	1	6	0
no.glasses	9	1	8	2	7	3	6	4	5	5	4	6	3	7
probability	0.0009		0.0236		0.1573		0.3671		0.3304		0.1101		0.0105	
which	<		∗										<	

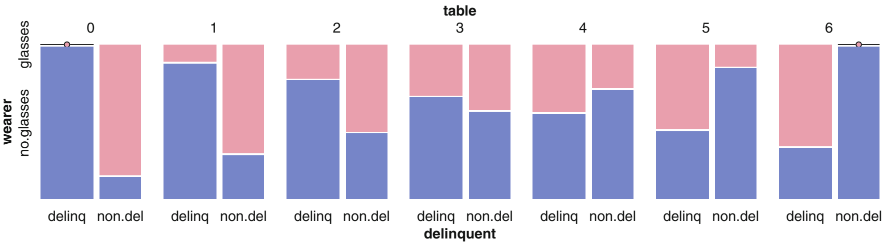


Fig. 15.5 All possible 2×2 tables with the same margins as the observed glasses table. The numerical values corresponding to this figure are in Table 15.5. The observed table is Table **1**. In each table we are comparing the proportion of `glasses` within `delinq` to the proportion of `glasses` within `non.del`. The widths of each bar are proportional to the count of people in that category of `delinquent`. The heights are percents of `wearer` within that category of `delinquent` and therefore add up to 100% in each column.

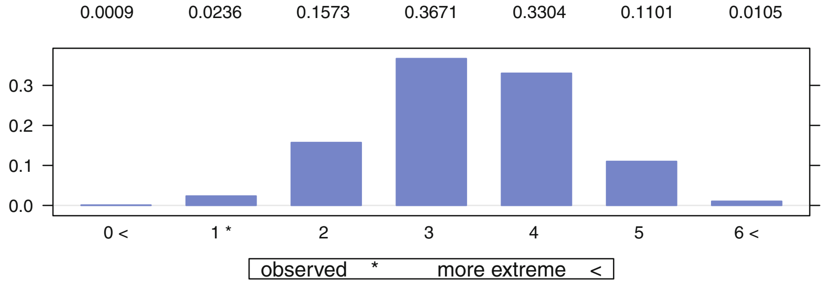


Fig. 15.6 Probabilities for all possible 2×2 tables with the same margins as the observed `glasses` table.

Table **6**, the more extreme on the opposite tail of the distribution, has probability

$$\frac{\binom{9}{6}\binom{7}{0}}{\binom{16}{6}} = 0.0105$$

The two-sided *p*-value is the sum of the probabilities of the observed table and both more extreme tables, 0.0236 + 0.0009 + 0.0105 = 0.0350. The R code for the probability calculations is included in file `HHscriptnames(15)`.

15.3 Simpson's Paradox

Simpson's paradox, a counterintuitive situation, occurs when the presence of a third variable is unexpectedly responsible for a change, or even reversal, of the relationship between two categorical variables. The following example taken from Blyth (1972), with dataset `data(blyth)`, illustrates this phenomenon.

The data, including the margin summed over location, are shown in Table 15.6 and Figure 15.7. A medical researcher selected 11,000 human subjects at location A and 10,100 subjects at location B. At A, 1,000 of the subjects were randomly assigned to the standard treatment (standard) and the remaining 10,000 subjects were assigned to a new treatment (new). At B, 10,000 of the subjects were randomly assigned to standard and the remaining 100 subjects were assigned to new. Eventually, each subject was classified as not-survived (not) or survived (survive).

Table 15.6. The intent is to show for each Location that the percentage surviving with the new Treatment is larger than with the standard treatment, but that

Table 15.6 Blyth's data illustrating Simpson's Paradox. Within each location (A and B), the new treatment has a higher survival rate than standard treatment. Summed over locations (A&B combined), new has a lower survival rate than standard. We show several different style graphs of the Counts and Proportions for this dataset in Figure 15.7.

		Location					
		A		B		A&B combined	
Summary	Survival	standard	new	standard	new	standard	new
Count	not	950	9000	5000	5	5950	9005
	survive	50	1000	5000	95	5050	1095
Percent	not	95	90	50	5	54	89
	survive	5	10	50	95	46	11

the reverse is true when the two Locations are combined. This portrayal fails for these data because it is not possible to distinguish the very small counts 5, 50, and 95 in three of the eight cells in Table 15.6 when they are displayed on a common numerical scale with counts ranging from 950 to 9,005 in the table's other cells.

The paradox is better communicated by Figure 15.7 panels b,c,d,f which graph the percentages themselves. We observe that in Location A, the percent surviving following the new Treatment was 10% as compared to 5% with the standard treatment. In Location B, the new Treatment improved the survival percentage to 95% from 50% for the standard Treatment. It seems that the new Treatment was very successful. Now look at the summary results for both Locations combined. The survival rate for the standard Treatment is 46%, but it is only 11% for the new Treatment. The combined results suggest that the new Treatment is a disaster!

The substantive reason for this finding is that the subjects in Location A were much less healthy than those in B and the new Treatment was given mostly to subjects in A, where it could not be expected to fare as well as with subjects in B. That is, the factors Treatment and Location are not independent. When this is so, it can happen as here that

$$P(\texttt{survive} \mid \texttt{new}) < P(\texttt{survive} \mid \texttt{standard})$$

$$\text{while both}$$

$$P(\texttt{survive} \mid \texttt{new} \cap A) \geq P(\texttt{survive} \mid \texttt{standard} \cap A)$$

$$\text{and}$$

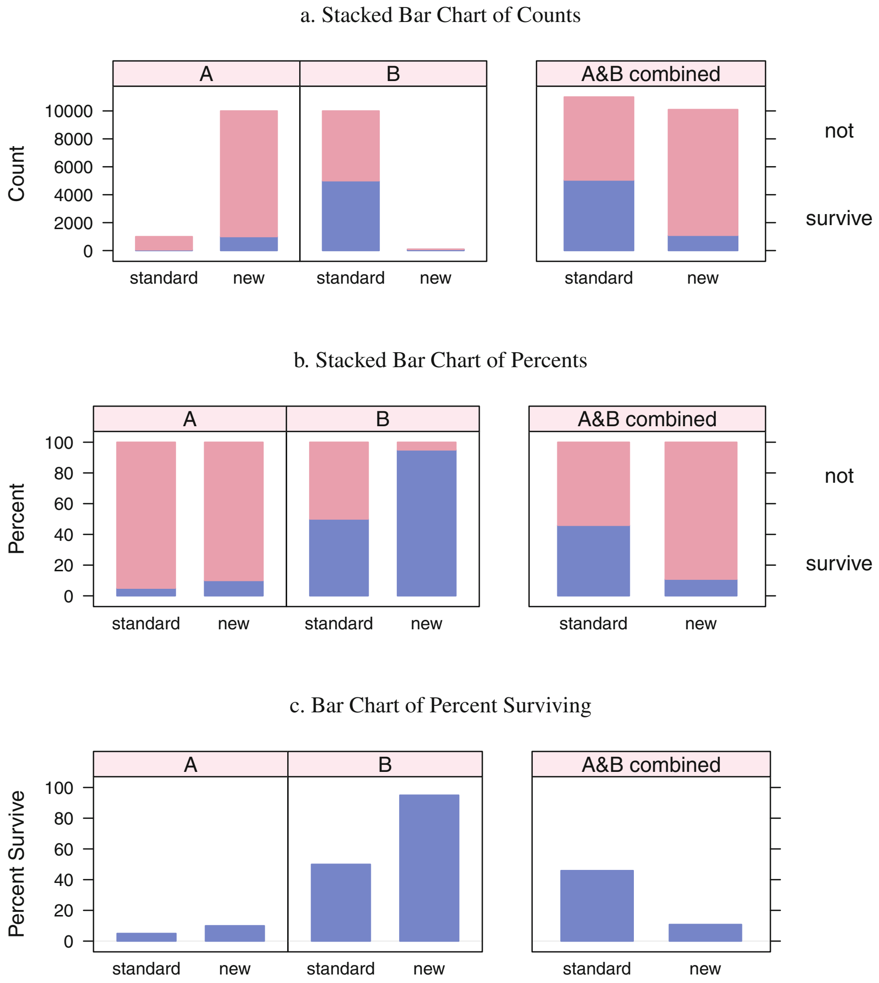
$$P(\texttt{survive} \mid \texttt{new} \cap B) \geq P(\texttt{survive} \mid \texttt{standard} \cap B)$$

corresponding to $.11 < .46$ but $.10 > .05$ and $.95 > .50$ in this example.

The Percent panels b,c,d,f of Figure 15.7 display the disparity better than the Count panels a,e because they transform the observed data from the scale reported by the client to the proportion scale, a scale in which the reversal is visible. In the proportion scale, most strongly in the mosaic plot in Panel d, we can easily see that the combined location information is almost the same as the B–standard and A–new information. In retrospect we can also see the same information in the Count panels. We can explain it by noting that there is almost no data in the A–standard and B–new cells; hence the combined information really is just the B–standard and A–new information plus a little noise.

When analyzing contingency table data, we should be alert to the possibility illustrated in this example that results for tables individually can differ from those when these tables are combined.

What is the resolution in situations such as this where individual results contradict combined results? Almost always, the individual results have more credence because combining such individuals cannot be adequately justified. In Blyth's example, the disparity between individual and combined results could have been attributable to different baseline health status of the patients at the two locations. Or it could be an artifact of the radically different treatment allocation patterns at the two locations.

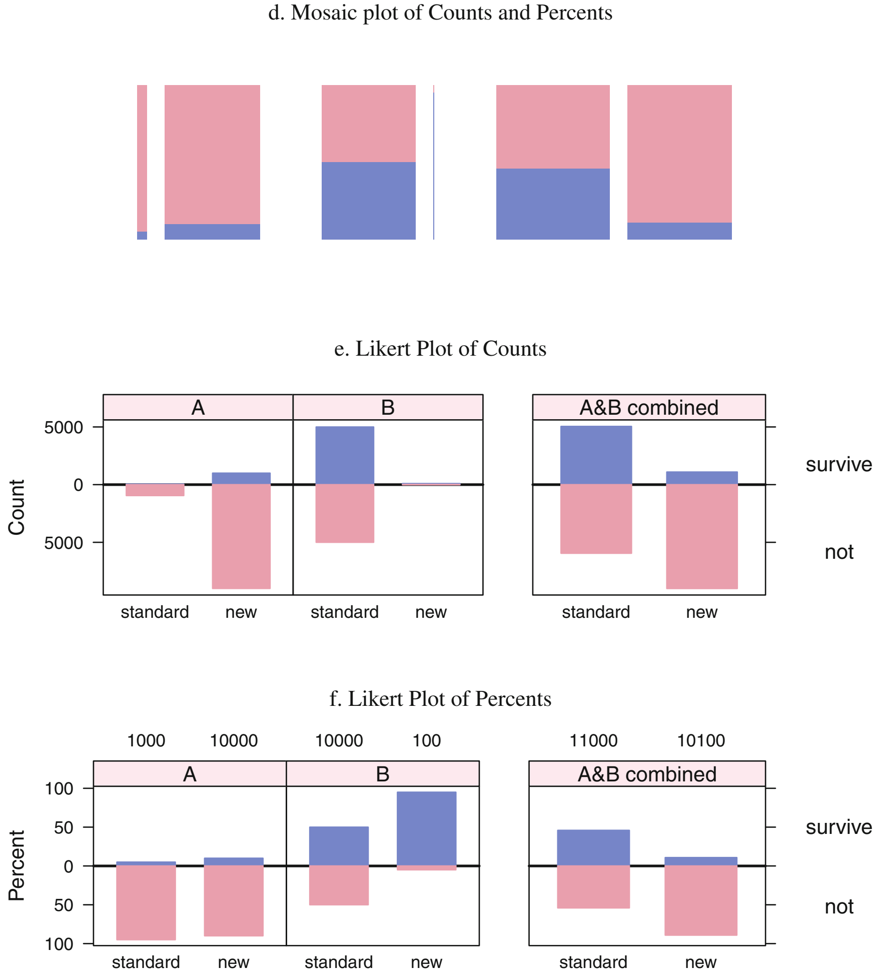


a. Stacked Bar Chart of Counts

b. Stacked Bar Chart of Percents

c. Bar Chart of Percent Surviving

Fig. 15.7 Blyth's data illustrating Simpson's paradox. Within each location (A and B), the new treatment has a higher survival rate than standard treatment. Summed over locations, new has a lower survival rate than standard. We note that the combined location information is almost the same as the B–standard and A–new information. The great disparity in counts among the four Location–Treatment groups makes it difficult to see the survival rates in the stacked bar chart in Panel a. The rates are easier to see in the stacked bar chart of percents in Panel b. They are less easy to see in the bar chart of only survive rates in Panel c—these bars are identical to the bottom bars in Panel b, but the visual sense of proportion is missing.


d. Mosaic plot of Counts and Percents

e. Likert Plot of Counts

f. Likert Plot of Percents

Fig. 15.7 continued The mosaic plot in Panel d shows three summaries: Counts of treatment assignments—as widths of each set of bars, Percents surviving conditional on Location–Treatment—as the relative heights of the bars in each Location–Treatment combination, and Counts of each Location–Treatment combination—as areas of each rectangle. In Panel e, the Likert plot of Counts, we again can barely see the B–new bars. In Panel f, the Likert plot of Percents, we print the observation counts for the Location–Treatment combinations above the bars. In the Likert plots (diverging stacked barcharts), the not survive values are shown diverging down from zero. In the other plots, the vertical range is from 0% to 100% and the not survive values are stacked above the survive values.

Simpson's paradox is the discrete analogue of the ecological fallacy discussed in Section 4.2. It is also related to the need to examine simple effects in the presence of interaction of qualitative factors, discussed in Chapters 12 to 14, since both problems refer to the importance of distinguishing overall conclusions from conclusions for subgroups.

15.4 Relative Risk and Odds Ratios

Analysis of data arranged in a 2×2 table is equivalent to comparing two proportions. However, analyzing the difference $p_1 - p_2$ via a CI or test as outlined in Chapter 5 is often not an appropriate way to compare them. Instead a measure of *relative* difference is appropriate. Consider two cases, the first with $p_1 = .02$ and $p_2 = .07$ and the second with $p_1 = .50$ and $p_2 = .55$. In both cases, $p_2 - p_1 = .05$. However, in the first case, p_2 is 250% more than p_1, but in the second case p_2 is only 10% more than p_1, and from this point of view it is inadequate to merely consider differences of proportions, particularly proportions close to either 0 or 1.

We discuss two additional measures for comparing two proportions. The first, the *relative risk*, is simply the ratio of the two proportions, $\hat{p}_1/\hat{p}_2$.

The *odds ratio* is a widely used measure of relative difference. It is more informative than a chi-square test for a 2 × 2 table because it measures the magnitude of difference between two proportions. Unlike the chi-square test, the odds ratio is minimally affected by the size of the sample.

Based on the definition of odds in Equation (3.2), if $\hat{p}$ is an estimated probability of success, the estimated *odds in favor* of success are $\hat{\omega} = \hat{p}/(1 - \hat{p})$. For comparing two estimated proportions, $\hat{p}_1$ and $\hat{p}_2$ in a 2 × 2 contingency table, the estimated ratio of two odds, the odds ratio, is

$$\hat{\Psi} = \hat{\omega}_2/\hat{\omega}_1 \tag{15.5}$$

A quick way to hand-calculate the estimated odds ratio is $\hat{\Psi} = (n_{11}n_{22})/(n_{21}n_{12})$, and for this reason the odds ratio is also known as the *cross-product ratio*.

If the odds ratio exceeds 1 so does the relative risk, and conversely.

15.4.1 Glasses (Again)

For example, reconsider the data of Section 15.2.1. The relative risk is

$$\frac{\left(\frac{8}{10}\right)}{\left(\frac{1}{6}\right)} = 4.8$$

This says that based on these data, nonwearers of glasses are four times more likely to become delinquent than wearers of glasses.

The odds ratio is

$$\hat{\Psi} = \frac{\left(\dfrac{\frac{5}{7}}{1 - \frac{5}{7}}\right)}{\left(\dfrac{\frac{1}{9}}{1 - \frac{1}{9}}\right)} = 20$$

This means that the odds that a nondelinquent wears glasses are estimated to be 20 times the odds that a delinquent wears glasses. Alternatively, the odds that a delinquent wears glasses are estimated as 1/20 times the odds that a nondelinquent wears glasses. If this ratio had been maintained for a larger sample, an implication might have been that police needn't pay much attention to boys wearing glasses.

Often investigators report the log of the odds ratio since the change in the reference group simply reverses the sign of the log odds: $\ln(20) = 2.996$ and $\ln\left(\frac{1}{20}\right) = -2.996$.

15.4.2 Large Sample Approximations

A useful property of both the odds ratio and the relative risk is that for large sample sizes, the log of the estimated odds ratio and the log of the estimated relative risk are approximately normally distributed.

15.4.2.1 Odds Ratio

From Agresti (1990) Equation (3.15) we find the log of the estimated odds ratio is approximately normally distributed

$$\ln(\hat{\Psi}) \sim N\left(\ln(\Psi),\, \sigma^2_{\ln(\hat{\Psi})}\right) \tag{15.6}$$

with mean equal to the log of the population odds ratio and estimated variance

$$\hat{\sigma}^2_{\ln(\hat{\Psi})} = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \tag{15.7}$$

These facts lead to large sample confidence intervals and hypothesis tests for odds ratios. A test of H_0: $\ln(\Psi) = 0$, or equivalently, $\Psi = 1$, is based on

$$z_{\text{calc}} = \frac{\ln(\hat{\Psi})}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}}$$

An approximate 100(1 – α)% confidence interval on $\ln(\Psi)$ is

$$\text{CI}\big(\ln(\Psi)\big) = \ln(\hat{\Psi}) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = (L, U) \qquad (15.8)$$

If we denote this interval by (L, U), the approximate confidence interval on the odds ratio Ψ_{21} is (e^L, e^U).

15.4.2.2 Relative Risk

From Agresti (1990) Equation (3.18) we find the log of the estimated relative risk is approximately normally distributed

$$\ln(\hat{p}_1/\hat{p}_2) \sim N\Big(\ln(p_1/p_2),\ \sigma^2_{\ln(\hat{p}_1/\hat{p}_2)}\Big) \qquad (15.9)$$

with mean equal to the log of the population relative risk and estimated variance

$$\hat{\sigma}^2_{\ln(\hat{p}_1/\hat{p}_2)} = \sqrt{\frac{1-p_1}{p_1\, n_1} + \frac{1-p_2}{p_2\, n_2}} \qquad (15.10)$$

These facts lead to large sample confidence intervals and hypothesis tests for relative risks. A test of H_0: $\ln(\hat{p}_1/\hat{p}_2) = 0$, or equivalently, $\hat{p}_1/\hat{p}_2 = 1$, is based on

$$z_{\text{calc}} = \frac{\ln(\hat{p}_1/\hat{p}_2)}{\sqrt{\frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}}}$$

An approximate 100(1 – α)% confidence interval on $\ln(\hat{p}_1/\hat{p}_2)$ is

$$\text{CI}\big(\ln(\hat{p}_1/\hat{p}_2)\big) = \ln(\hat{p}_1/\hat{p}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}} = (L, U) \qquad (15.11)$$

If we denote this interval by (L, U), the approximate confidence interval on the relative risk Ψ_{21} is (e^L, e^U).

15.4.3 Example—Treating Cardiac Arrest with Therapeutic Hypothermia

Holzer (2002) supervised a multicenter trial of patients who were randomly assigned to receive or not receive therapeutic hypothermia (lowered body temperature) to assist in recovery following resuscitation from cardiac arrest. Six months after cardiac arrest, patients were classified as having a favorable neurologic outcome or not. Of the 136 patients treated with hypothermia, 75 had a favorable neurological outcome. Of the 137 patients not treated with hypothermia, 54 had a favorable neurological outcome (see Table 15.7 and Figure 15.8 for the counts and Figure 15.9 for the odds and log of the odds). All patients received standard treatment for cardiac arrest apart from hypothermia and the treatment was blinded from the assessors of the outcome.

Table 15.7 Results of therapeutic hypothermia investigation Holzer (2002).

	Favorable neurological outcome		
	Yes	No	Total
Treated	75	61	136
Control	54	83	137

For a patient who receives the therapeutic hypothermia treatment, the estimated odds in favor of a favorable neurologic outcome are

$$\hat{\omega}_2 = \left(\frac{\dfrac{75}{136}}{1 - \dfrac{75}{136}} \right) \approx 1.23$$

For a patient who does not receive this treatment, the estimated odds in favor of a favorable neurological outcome are

$$\hat{\omega}_1 = \left(\frac{\dfrac{54}{137}}{1 - \dfrac{54}{137}} \right) \approx 0.65$$

The odds and log odds are plotted in Figure 15.9.

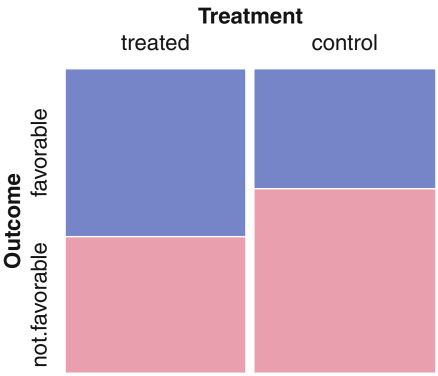


Fig. 15.8 `mosaic(Outcome ~ Treatment)`

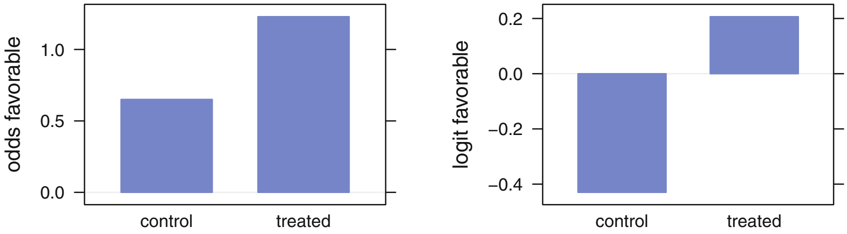


Fig. 15.9 Barchart of odds and barchart of logit.

The estimated odds ratio is

$$\hat{\Psi} = \frac{\hat{\omega}_2}{\hat{\omega}_1} = \frac{(75)(83)}{(54)(61)} \approx 1.8898$$

The reader can verify that the estimated standard deviation of the log of the odds ratio is 0.246, and that this leads to an approximate 95% confidence interval $(1.168, 3.058)$ for the population odds ratio.

For the hypothermia example with $\alpha = .05$, we have

$$\text{CI}\Big(\ln(\Psi)\Big) = \ln(1.8898) \pm 1.96\sqrt{\frac{1}{75} + \frac{1}{54} + \frac{1}{61} + \frac{1}{83}} \approx (0.1552, 1.1177)$$

We therefore have an approximate confidence interval on the odds ratio of

$$\text{CI}(\Psi) \approx (1.168, 3.058)$$

This means that the odds of a favorable neurological outcome for a patient receiving the therapeutic hypothermia treatment are estimated to be between 1.17 and 3.06 times the odds of a favorable neurological outcome for a patient not receiving this particular treatment. Further, the calculated z-statistic for a test of the one-sided alternative hypothesis that the population odds ratio exceeds 1 is 2.592, with corresponding p-value less than 0.01. Therefore, there is strong evidence that the therapeutic hypothermia treatment improves probability of a successful neurological outcome.

For fixed probability of favorable outcome for control patients of $54/137$ = .3942, corresponding to fixed odds of favorable outcome of $.3942/.6058$ = .6506, the point estimate for the odds of favorable outcome for treated patients is given by 1.8898×.6506 = 1.2295, and an estimated confidence interval for the odds by (1.168×.6506, 3.058×.6506) = (0.7599, 1.9895). The corresponding point and interval estimates for the probabilities of outcome for treated are

estimate	favorable		
point	1.2295/2.2295	=	0.5515
interval	(0.7599/1.7599, 1.9895/2.9895)	=	(0.4318, 0.6655)

The point estimate of the probability of favorable outcome for treated is exactly the observed proportion 75/136=0.5515, and the confidence interval of the proportion of favorable outcomes excludes the observed proportion for the control group.

We can extend this discussion by assuming any fixed probability of favorable outcome for treatment and then calculating the confidence interval for the probability of favorable outcome for control. We do so in Figure 15.10 for the set of fixed probabilities $p_1 = (0, .05, \ldots, 1)$.

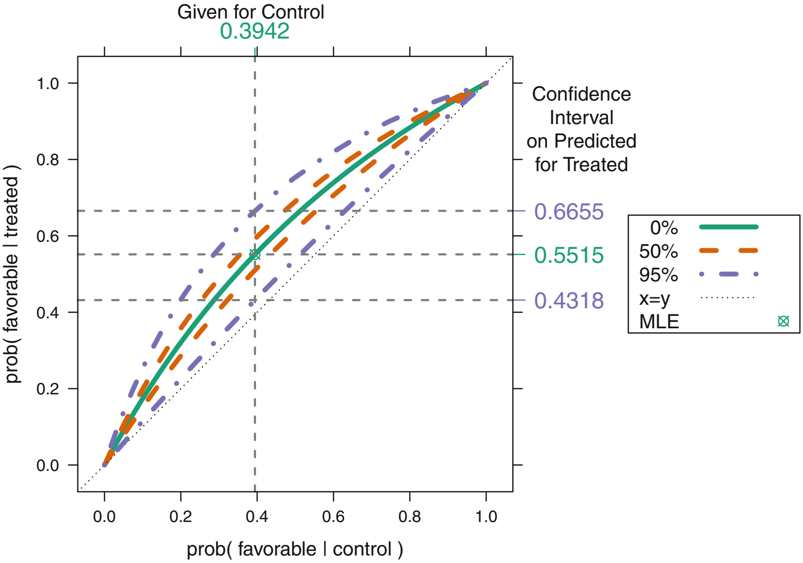


Fig. 15.10 Confidence intervals for P(favorable | treated) in the hypothermia example given the odds ratio. The confidence intervals are calculated from an assumed P(favorable | control) and the given odds ratio using the odds ratio formula in Equation 15.8. Details for P(favorable | control) = $54/137$ = .3942 and P(not.favorable | control) = .6058, corresponding to given odds of favorable outcome of .3942/.6058 = .6506, are shown in Section 15.4.3. Symmetrically, we can assume any fixed probability of favorable outcome for control and then calculate the confidence interval for the probability of favorable outcome for treatment.

15.5 Retrospective and Prospective Studies

Consider two possible experiments to assess whether vitamin C supplementation prevents occurrences of the common cold.

In the first experiment we select 100 people who have had a cold during the past two months and 100 people who have not had a cold during the past two months. We then ask these people whether or not they have taken a daily vitamin C supplement during this period. In the second experiment we select 200 volunteers, assigning them to take no vitamin C supplementation apart from that offered by the study. We randomly assign 100 of these subjects to receive the study's vitamin C supplement and the other 100 subjects to receive a placebo, indistinguishable by these subjects from vitamin C. Then, two months later, we ask the subjects whether or not they have had a cold since the experiment began.

The first experiment is an example of a retrospective study, also called a case–control study. Subjects having a condition are called cases, subjects not having a condition are termed controls, and subjects are cross-classified with a risk factor (present or absent). In the above example, the risk factor is presence or absence of vitamin C supplementation. In retrospective studies, the subjects are selected after the events in question have already occurred, in this case having contracted one or more colds. Such studies are common in medical research because they generally assure a larger number of subjects than prospective studies.

The second experiment is an example of a prospective study, also known as a cohort study. Samples are taken from a population of subjects classified according to two risk factors (events) defined prior to initiating sampling, in this case assignment to vitamin C or placebo. Such studies often require that subjects be followed for a period of time until the subjects are determined to have a condition or not.

In the cold example, the analysis of the retrospective study can be done immediately, but analysis of the prospective study must wait two months to see if colds develop. Prospective studies often run over a long time period; 5 to 10 years is not unusual. It is not uncommon for subjects to withdraw or be lost to the study. For this reason, it is more difficult to obtain sizeable samples from prospective studies than from retrospective ones. Prospective studies are more informative than retrospective studies. Investigators have more control over the risk factor in prospective studies than in retrospective studies. In prospective studies investigators are often able to obtain information on important confounding variables that bear on the response. Such information is usually unavailable in retrospective studies. The experiment discussed in Section 15.4.3 is an example of a prospective study.

Odds ratios are particularly important in the analysis of experiments involving retrospective studies. In a retrospective study it is unlikely that the cases can be considered a random sample of all persons afflicted with the condition. In the context of our example, we cannot be sure that the 100 selected people with colds are representative of all people with colds. Therefore, in such a study we cannot estimate the proportion of people having the risk factor who have the condition, or the proportion of people without the risk factor who have the condition. Nevertheless, in a retrospective study we are able to measure the odds ratio and we can claim that the sample odds ratio estimates the population odds ratio.

15.6 Mantel–Haenszel Test

Analysts are often called on to interpret k 2×2 contingency tables, related to one another by the fact that each table has the same row and column categories. The k tables usually represent the k levels of a third (categorical) factor in addition to the two-level factors specified by rows and columns. For example, we look in Table 15.8 and Figure 15.11 at data studying the effectiveness of the Salk vaccine for polio protection for $k = 6$ different age groups. Each of the $k = 6$ 2×2 tables in the "**Observed**" column shows the response (`paralysis` or `no.paralysis`) for subjects who were or were not vaccinated (`vac` or `no.vac`). The complete discussion of the dataset and the table are in Section 15.7.

Table 15.8 Detail for calculation of the Cochran–Mantel–Haenszel test of the polio example. See the discussion in Section 15.7, where we find that the Mantel–Haenszel chi-square test without the continuity correction is 16.54.

Age Group	**Vaccination**	**Observed**		**Expected**		prob	**Chi-Square**		**[1,1] position for Mantel–Haenszel test**						
		no.par	par	no.par	par	no.par	chisq	p.chisq	O	E	O−E	var	n	dev	mh
0–4	no.vac	10	24	15.00	19.00	0.294	5.965	0.015	10	15.00	−5.00	4.25	68	−2.42	5.88
	vac	20	14	15.00	19.00	0.588									
5–9	no.vac	3	15	7.20	10.80	0.167	6.806	0.009	3	7.20	−4.20	2.65	45	−2.58	6.65
	vac	15	12	10.80	16.20	0.556									
10–14	no.vac	3	2	3.00	2.00	0.600	0.000	1.000	3	3.00	0.00	0.67	10	0.00	0.00
	vac	3	2	3.00	2.00	0.600									
15–19	no.vac	1	6	3.11	3.89	0.143	4.219	0.040	1	3.11	−2.11	1.12	18	−2.00	3.99
	vac	7	4	4.89	6.11	0.636									
20–39	no.vac	7	5	8.44	3.56	0.583	1.501	0.221	7	8.44	−1.44	1.44	27	−1.20	1.45
	vac	12	3	10.56	4.44	0.800									
40+	no.vac	3	2	3.33	1.67	0.600	0.600	0.439	3	3.33	−0.33	0.22	6	−0.71	0.50
	vac	1	0	0.67	0.73	1.000									

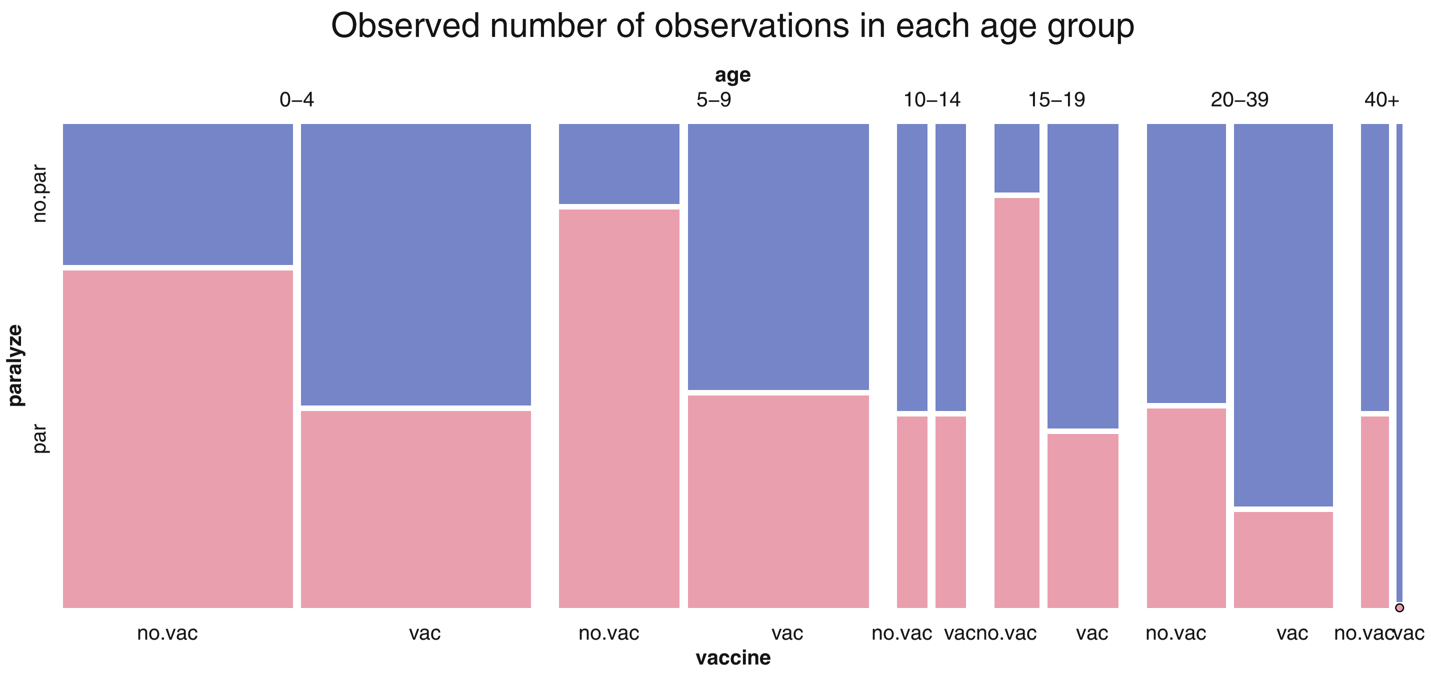


Fig. 15.11 Mosaic plot of the `salk` dataset. Within each age group we see that the observed proportion of non-paralysis (blue height as a fraction of total height for the bars) for the Vaccinated group is greater than or equal to the proportion of non-paralysis for the Not.Vaccinated group. The bar widths are proportional to the number of people in that age–vaccination grouping. See discussion in Section 15.7.

In earlier sections of this chapter we consider procedures for testing independence of the row and column categories for individual tables. We are now interested in testing the hypothesis that all k tables show the same pattern of relation of the rows to the columns: Either all tables show independence of rows and columns or all show the same dependency structure.

The Mantel–Haenszel test, also referred to as the Cochran–Mantel–Haenszel test, tests the hypothesis that row vs column independence holds *simultaneously in each table.* It is designed to be sensitive to an overall consistent pattern. It has low power for detecting association when patterns of association for some strata are in the opposite direction of other strata.

Let us now look at the algebra of the test statistic. Since we now have k 2 × 2 tables, we require a third subscript on the n's. Let the k^{th} table be

n_{11k}	n_{12k}	$n_{1.k}$
n_{21k}	n_{22k}	$n_{2.k}$
$n_{.1k}$	$n_{.2k}$	$n_{..k}$

Also define

$$e_{ijk} = \frac{(n_{i.k})\,(n_{.jk})}{n_{..k}}$$

to be the expected (i, j) cell count under independence in table k, and

$$V(n_{11k}) = \frac{n_{1.k}\ n_{2.k}\ n_{.1k}\ n_{.2k}}{n^2_{..k}\ (n_{..k} - 1)} = \frac{e_{11k}\ e_{22k}}{(n_{..k} - 1)}$$

to be the estimated variance of n_{11k} under the assumption of a hypergeometric distribution of a 2×2 table with fixed margins. Then we make a normal approximation and work with

$$n_{11k} \sim N\Big(e_{11k}, V(n_{11k})\Big)$$

The sum $\sum_k n_{11k}$ is also approximately normal with mean $\sum_k e_{11k}$ and variance $\sum_k V(n_{11k})$. We therefore use as the test statistic the quantity

$$M^2 = \frac{\left[\sum\limits_k n_{11k} - \sum\limits_k e_{11k}\right]^2}{\sum\limits_k V(n_{11k})} \qquad (15.12)$$

and the p-value of the test is $1 - \mathcal{F}_{\chi^2}(M^2 \mid 1)$, the corresponding tail percentage of the chi-square distribution with 1 df.

Sometimes we will wish to use a variant of M^2 with a continuity correction and then we use

$$M^2 = \frac{\left[\left|\sum_k n_{11k} - \sum_k e_{11k}\right| - .5\right]^2}{\sum_k V(n_{11k})} \tag{15.13}$$

Inspecting the form of M^2 tells us that significance can occur under either of two conditions:

1. Most or all tables must have the observed $(1,1)$ cell count at least as large as expected under the null hypothesis.
2. Most or all tables must have the observed $(1,1)$ cell count at most as small as expected under the null hypothesis.

Equivalently, most or all of the tables must have an odds ratio either

1. at least 1, or
2. at most 1.

Note that M^2 is **not** the same as the chi-square statistic one gets from the 2×2 table formed as the sum of the k tables.

15.7 Example—Salk Polio Vaccine

Chin et al. (1961), also in Agresti (1990), discuss 174 polio cases classified by age of subject, whether or not the subject received the Salk polio vaccine, and whether the subject was ultimately paralyzed by polio. The dataset is in `data(salk)`. We wish to learn if symptom status (paralysis or not) is independent of vaccination status after controlling for age.

Each of the k=6 "Observed" subtables in Table 15.8, one for each of k=6 age ranges, shows two estimated probabilities of no paralysis, for subjects without vaccine and subjects with vaccine. In the "0–4" subtable, for example, we see $p_{\text{no.vac}}$(no.par)=.294 and p_{vac}(no.par)=.588. In all cases the observed proportion with vaccine is higher. The "chi-square" column shows the ordinary contingency table chi-square for each subtable. The four subtables with older subjects do not have many observations and do not strongly support the conclusion that vaccine is better. The Cochran–Mantel–Haenszel test provides a way of combining the information, properly weighted, from all six subtables to get a stronger conclusion. The "O", "E", and "O−E" columns show the [1,1] or [no.vac,no.par] position from the "Observed" and “Expected” tables. “O−E” is the weighted difference of the row probabilities $\text{O}_i{-}\text{E}_i = w_i(p_{\text{no.vac}}(\text{no.par}) - p_{\text{vac}}(\text{no.par}))$ [with weights $w_i = 1/(1/n_1 + 1/n_2)$ for the i^{th} subtable, where n_j is the total count on the j^{th} row]. While we choose to focus on the counts of the [1,1] cells, an identical conclusion would be reached if the focus were on any of the three other cells of the 2 × 2 table.

The “var” column shows the variance of “O−E” under the assumption of the hypergeometric distribution for O_i assuming both row and column margins of the i^{th} table are fixed. The “dev” column is a standardized deviation, $(\text{O}{-}\text{E})/\sqrt{\text{var}}$, and the “mh” column is the squared standardized deviation. We plot the standardized deviations in Figure 15.12. The squared standardized deviation is the Mantel–Haenszel statistic for the subtable. The MH statistic for a subtable is very close to the chi-square statistic.

The Cochran–Mantel–Haenszel (CMH) test for the set of all k=6 subtables is constructed as a weighted combination of the same components used for the subtable statistics. Since each “O−E” is a random variable with mean and variance, we use Equations (3.8) and (3.9) to combine them. The CMH statistic is constructed from the sum of the “O−E” for the subtables, divided by the standard deviation of the sum, which is the square root of the sum of the variances: $\sum(\text{O}{-}\text{E})/\sqrt{\sum(\text{var})}$. Then the whole is squared. Thus the CMH statistic for this example is

$$\frac{\left(\sum(\text{O}{-}\text{E})\right)^2}{\sum(\text{var})} = \frac{(-5 - 4.20 - 0 - 2.11 - 1.44 - .33)^2}{(4.25 + 2.65 + .67 + 1.12 + 1.44 + .22)} = 16.54$$

For each of the age ranges with a sufficiently large sample, Fisher’s exact test performed on the 2 × 2 tables, shown in Table 15.9, detects a positive association between symptom and vaccination status: Persons vaccinated had a significantly

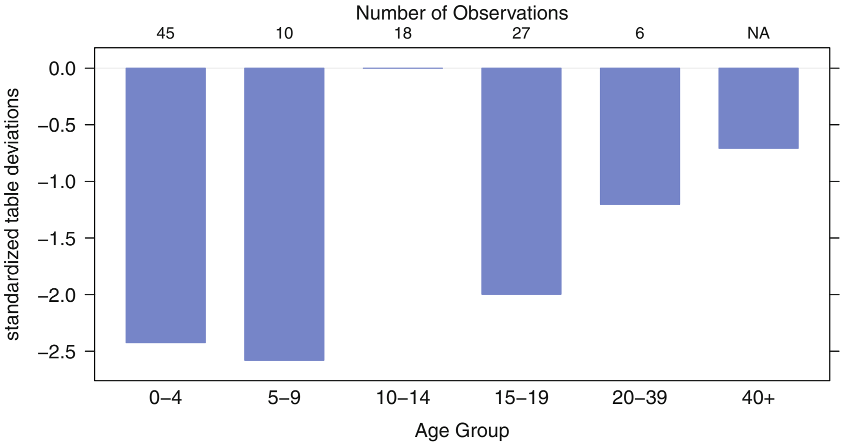


Fig. 15.12 Standardized deviations for individual table Mantel–Haenszel values.

lower incidence of paralysis than persons not vaccinated. In addition, each of the six 2×2 tables has an odds ratio $\big((\text{no.par}/\text{par})_{\text{no.vac}}\big)/\big((\text{no.par}/\text{par})_{\text{vac}}\big)$ of at most 1. Therefore, the Mantel–Haenszel test can be used. The Mantel–Haenszel test statistic has value 16.54, which is highly significant. This means that the relationship between symptom and vaccination status is consistent over all age ranges.

15.8 Example—Adverse Experiences

Evaluation of adverse experience data is a critical aspect of all clinical trials. This dotplot of incidence and relative risk, and the specific example, are taken from Amit et al. (2008). We proposed graphics for exploratory data analysis or signal identification, and for adverse experiences (AEs, also read as adverse events) that may result from a compound's mechanism of action or events that are of special interest to regulators.

Figure 15.13 is a two-panel display of the AEs most frequently occurring in the active arm of the study. The first panel displays their incidence by treatment group, with different symbols for each group. The second panel displays the relative risk of an event on the active arm relative to the placebo arm, with 95% confidence intervals (as defined in Equation 15.11) for a 2×2 table. The panels have the same vertical coordinates and different horizontal coordinates. R code for the construction of this plot is available as the `AEdotplot` function in the **HH** package.

Table 15.9 Run Fisher's exact test on each of the 2×2 tables in the "Observed" column of Table 15.8. Each of the six 2×2 tables has an odds ratio of at most 1. Therefore, the Mantel–Haenszel test can be used.

```
> data(salk)

> salk2 <- tapply(salk$Freq, salk[c(2,3,1)], c)

> class(salk2) <- "table"

> ## salk2  ## salk2 is structured as a set of 2x2 tables
> lt <- apply(salk2, 3, fisher.test, alternative="less")

> ## odds ratio and p-value
> sapply(lt, ‘[‘, c("estimate","p.value"))
         0-4     5-9      10-14  15-19   20-39  40+
estimate 0.2973  0.1669   1      0.1098  0.3645 0
p.value  0.01359 0.009521 0.7381 0.05656 0.2116 0.6667
```

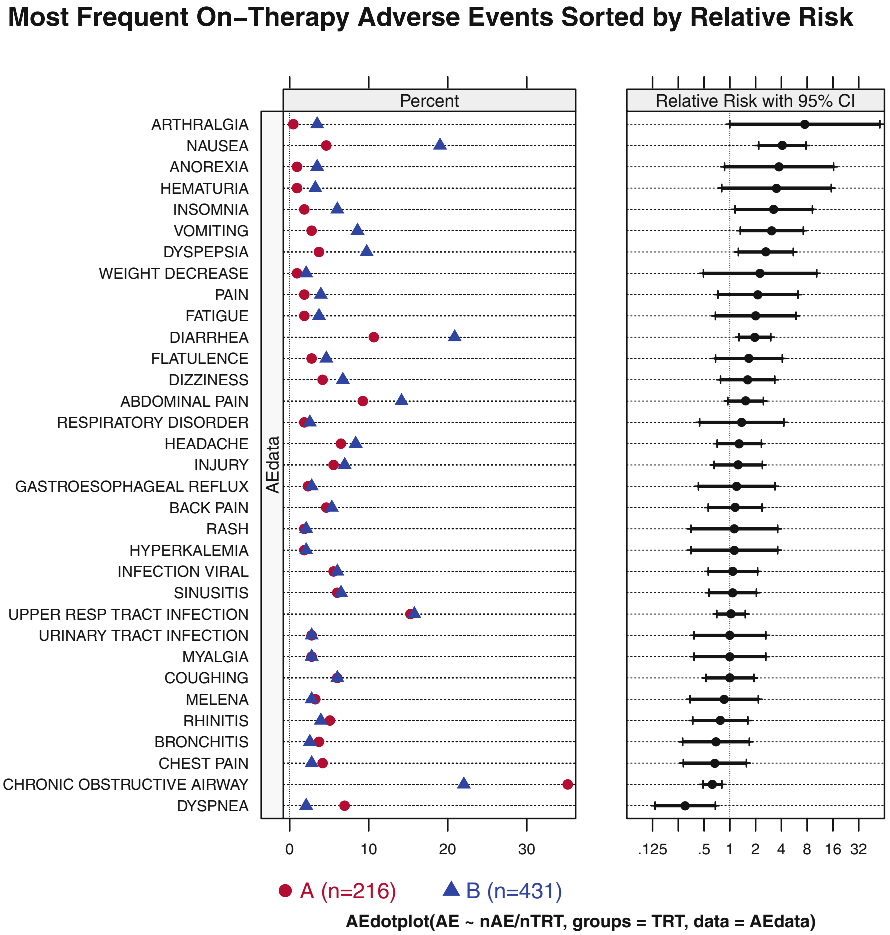


Fig. 15.13 Adverse Events dotplot for the clinical trial discussed in Amit et al. (2008). The left panel shows the observed percents of patients who reported each adverse event. The center panel shows the relative risk of Drug B relative to Drug A with 95% confidence intervals for a 2×2 table. By default, the AEs are ordered by relative risk so that events with the largest increases in risk for the active treatment are prominent at the top of the display.

If the display is not for regulatory purposes, intervals showing ±1 s.e. may be preferred. If confidence intervals are presented, multiple comparison issues should be given consideration, particularly if there is interest in assessing the statistical significance of differences of the relative risk for so many types of events. However, the primary goal of this display is to highlight potential signals by providing an estimate of treatment effect and the precision of that estimate. The criteria for including specific AEs in the display should be carefully considered. In Figure 15.13, the criterion used was that AEs have at least 2% incidence in the active arm.

The graphical presentation of this form of data has a very strong impact. The AEs are ordered by relative risk so that events with the largest increases in risk for the active treatment are prominent at the top of the display. Unlike with a table, it is immediately obvious to the reader which are the most serious AEs. This could be reversed to put the largest increases at the bottom, or the order could be defined by the actual risk in one of the treatment arms rather than the relative risk. Depending on the desired message, other sorting options include: magnitude of relative risk, incidence of event in a given treatment arm, or total incidence of events across all arms. We do not recommend ordering alphabetically by preferred term, which is the likely default with routine programming, because that makes it more difficult to see the crucial information of relative importance of the AEs.

15.9 Ordered Categorical Scales, Including Rating Scales

Ordered Categorical Scales, including Rating scales such as Likert scales, are very common in marketing research, customer satisfaction studies, psychometrics, opinion surveys, population studies, and numerous other fields. We recommend diverging stacked bar charts as the primary graphical display technique for Likert and related scales. We also show other applications where diverging stacked bar charts are useful. Many examples of plots of Likert scales are given. We discuss the perceptual issues in constructing these graphs.

An ordered categorical scale is an ordered list of mutually exclusive terms. Ordered categorical scales are used, for example, in questionnaires where each respondent is asked to choose one of the terms as a response to each of a series of questions. The usual summary is a table that shows the number of respondents who chose each term for each question. The summary table is a special case of a contingency table where the rows are individual questions and the columns are the ordered set of potential responses.

A rating scale is a form of psychometric scale commonly used in questionnaires. The most familiar rating scale is the Likert scale (Likert, 1932), which consists of a discrete number of choices per question among the sequence: “strongly disagree”, “disagree”, “no opinion”, “agree”, “strongly agree”. Likert-type scales may use other sequences of bipolar adjectives: “not important” to “very important”; “evil” to “good”. These scales sometimes have an odd number of levels, permitting a neutral choice. Sometimes they have an even number of levels, forcing the respondent to make a directional choice. Some ordered categorical scales are uni-directional—age ranges or population quantiles, for example—for which negative and neutral interpretations are not meaningful.

For concreteness we present in Section 15.9.1 a dataset from a survey for which a natural display is a coordinated set of diverging stacked bar charts. We introduce the dataset by showing and discussing a multi-panel plot of the entire dataset. Then we move to the construction and interpretation of individual panels.

15.9.1 Display of Professional Challenges Dataset

Our primary data example is from an *Amstat News* article (Luo and Keyes, 2005) reporting on survey responses to a question on job satisfaction. A total of 565 respondents replied to the survey. Each person answered one of five levels of agreement or disagreement with the question "Is your job professionally challenging?" The respondents were partitioned into nonoverlapping subsets by several different criteria. For each of the criteria, the original authors were interested in comparing the percent agreement by that criterion's groups.

In Figure 15.14, we show the complete results of the survey as a coordinated set of diverging stacked bar charts. In this section we concentrate on the appearance of the plot for its function of representing the meaning of the dataset.

There are six panels in the plot. The top panel shows "All Survey Respondents". The remaining panels show different partitions of the 565 respondents. In the second panel from the top, for example, the criterion name "Employment sector" is in the left strip label. The respondents self-identify to one of the five employment groups named in the left tick labels. The number of people in each group is indicated as the right-tick label. Each stacked bar is 100% wide. Each is partitioned by the percent of that employment group who have selected the agreement level indicated in the legend below the body of the plot. The legend is ordered by the values of the labels. Darker colors indicate stronger agreement. Gray indicates the neutral position, in this example, "No Opinion". The bar for the neutral position is split, half to the left side of the vertical zero reference line and half to the right side. The reference line is placed behind the bars to prevent it from artificially splitting the neutral bar into two pieces. The default color palette has red on the left for disagreement and blue on the right for agreement. See Section 15.9.2 for a discussion of color palettes.

The intent of this plot is to compare percents within subgroups of the survey population; consequently we made all bars have equal vertical thickness. The panel heights are proportional to the number of bars in the panel. The *x*-axis labels are displayed with positive numbers on both sides. The bars within each panel have been sorted by the percent agreeing (totaled over all levels of agreement). We usually prefer horizontal bars, as shown here, because the group labels and the names of the groups are easier to read when they are displayed horizontally on the *y*-axis.

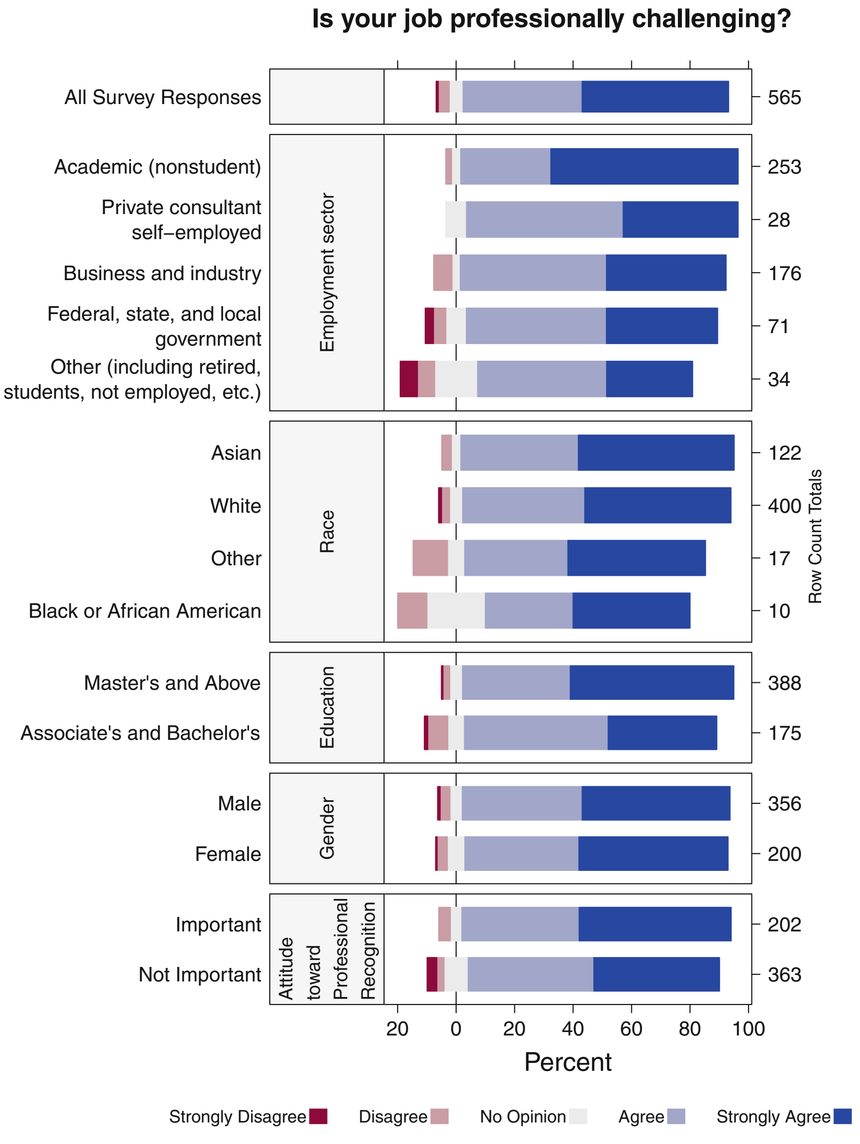


Fig. 15.14 Survey responses to a question on job satisfaction (Luo and Keyes, 2005). A total of 565 respondents replied to the survey. Each person answered one of five levels of agreement or disagreement with the question “Is your job professionally challenging?” Each panel of the plot shows a breakdown of the respondents into categories defined by the criterion listed in its left strip label.

Table 15.10 The respondents have been divided into five employment categories. The rows (employment categories) are displayed in the original order: alphabetical plus other. Columns are displayed sequentially, with disagreement to the left and agreement to the right.

	Strongly Disagree	Disagree	No Opinion	Agree	Strongly Agree
Employment Sector					
Academic (nonstudent)	0	5	8	78	162
Business and industry	0	11	5	88	72
Federal, state, an local government	2	3	5	34	27
Private consultant/self-employed	0	0	2	15	11
Other (including retired, students, not employed, etc.)	2	2	5	15	10

15.9.2 Single-Panel Displays

In this section, we look at the data for just the Employment panel of the full plot in Figure 15.14. Table 15.10 shows the respondents divided into five employment categories and the counts for each agreement level within each employment category.

Diverging stacked bar charts are easily constructed from Likert scale data. Each row of the table is mapped to a stacked bar in a bar chart. Usually the bars are horizontal. The counts (or percentages) of respondents on each row who agree with the statement are shown to the right of the zero line in one color; the counts (or percentages) who disagree are shown to the left in a different color. Agreement levels are coded from light (for closer to neutral) to dark (for more distant from neutral). The counts (or percentages) for respondents who neither agree nor disagree are split down the middle and are shown in a neutral color. The neutral category is omitted when the scale has an even number of choices. Our default color palette is the (Red–Blue) palette constructed by the `diverge_hcl` function in the **colorspace** package in R. The colors in the diverging palettes have equal intensity for equal distances from the center. The base colors Red and Blue have been chosen to avoid ambiguity for those with the most prevalent forms of color vision deficiencies.

It is difficult to compare lengths without a common baseline; see pages 54–57 of Robbins (2013) and the reference therein to Cleveland and McGill (1984). We are primarily interested in the total count (or percent) to the right or left of the zero line; the breakdown into strongly or not is of lesser interest so that the primary comparisons do have a common baseline of zero.

Figure 15.15 shows a direct translation of the counts in Table 15.10 to a plot. The strongest message in this presentation is that the sample has a very large percentage of academics. It is harder to compare relative proportions in the employment categories because the total counts in each row are quite disparate.

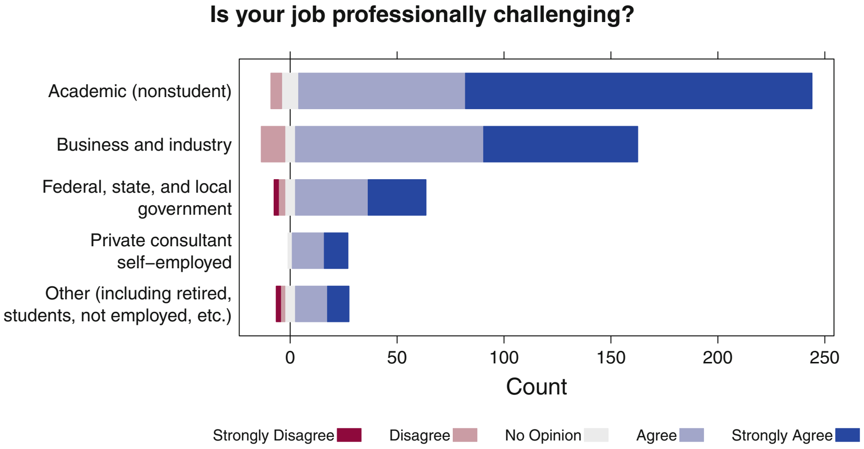


Fig. 15.15 This plot is a direct translation of the numerical values in Table 15.10 to graphical form. Blue is agree, red is disagree, gray is no opinion. The strongest message in this presentation is that the sample has a very large percentage of academics.

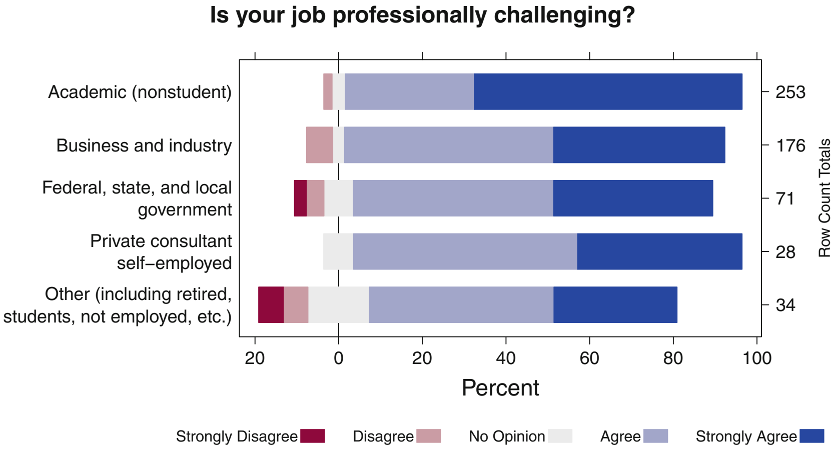


Fig. 15.16 In this variant of the plot, we display the row percents. We don't want to lose information about the uneven selection of respondents from the employment sectors so we display the counts as the right axis labels. Now we see that "Academic (nonstudent)" stands out as the largest percentage of dark blue on the graph.

Figure 15.16 displays the percents within each row. Now it is easy to see that a large majority of the people in all employment categories have a positive answer to the survey question. We don't want to lose the disparity in row totals, so we use the row count totals as the right-axis tick labels.

For plots such as Figure 15.16 with a single panel, and also for multiple-panel plots where the rows are distinct in each panel, we can still do better. Figure 15.17 shows the same scaling as Figure 15.16, but this time the row order is data-

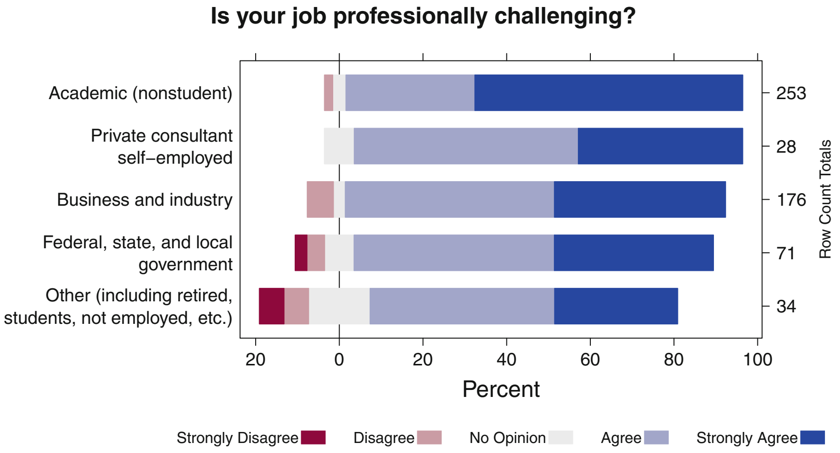


Fig. 15.17 In our third presentation of the same table, we now sort the rows of the table by the total percent agree (dark blue + light blue + $\frac{1}{2}$ gray). Data-dependent ordering is usually more meaningful than alphabetical ordering for unordered categories.

dependent. Rows are now ordered by the total percent of positive responses. This allows the reader to recognize groupings among rows (in this example, groupings of employment categories) that show similar responses.

15.9.3 Multiple-Panel Displays

Our illustration in Figures 15.15–15.17 is a single panel. It displays information on a single question for a partition of the respondents into several groups based on employment. Figure 15.14 is a multiple-panel display containing Figure 15.17 combined with other partitions of the same set of respondents.

15.9.3.1 One Question with Multiple Subsets of the Sample

Figure 15.14 shows responses to the same question for the same population of respondents partitioned into several series of groups based on additional characteristics.

The partitions have different numbers of groups. In order to retain the same vertical spacing between parallel bars, the vertical space allocated for the panels must differ. The different panels have been labeled by the name of the partitioning characteristic. In this example the panels are identified by a left strip label. Within each panel the bars have been sorted by the total percent of positive responses.

15.9.3.2 One or More Subpopulations with Multiple Questions

Figure 15.18 is differently structured. The data are the responses to a survey sponsored by the New Zealand Ministry of Research Science and Technology (New Zealand Ministry of Research Science and Technology, 2006). Here we have two different sets of questions that have been asked of the same set of respondents. Sometimes the respondents can be subdivided. If we had the data separately for the Male and Female subsets, we could have a plot similar to Figure 15.18 but with two columns of panels, left and right, one for each subset.

15.9.3.3 Common Structure at Multiple Times—Population Pyramids

Population pyramids are used in demographic studies and in epidemiological studies. The pyramid is a pair of back-to-back bar charts, one for males and one for females. We display the population pyramid as a Likert-type scale with two levels, male and female, for each age range. Figure 15.19 shows five pyramids at ten-year intervals years 1939–1979, with the *y*-labels on both the left and right axes. A **shiny** app that cycles through all years 1900–1979 is available by entering

```
shiny::runApp(system.file("shiny/PopulationPyramid",
                                    package="HH"))
```

The data is from the `USAage` dataset in the **latticeExtra** package.

15.10 Exercises

15.1. Hand et al. (1994) revisit a dataset attributed to Karl Pearson, `data(crime)`, that examines the relationship between type of `crime` committed and whether the perpetrator was a `drinker` or abstainer. Investigate whether these two classifications are independent, and if they are not, discuss the nature of the dependence.

15.2. Senie et al. (1981), also in Hand et al. (1994), investigate whether the frequency of breast self-examination is related to age group. The data appear accessed as `data(selfexam)`. Do you agree that there is a relationship? If so, describe it.

15.3. Sokal and Rohlf (1981), later in Hand et al. (1994), concern an experiment to determine the preference of invading ant colonies on two species of acacia tree. A total of 28 trees were made available for the study, 15 of species A and 13 of species B. Initially each tree was treated with insecticide to remove all existing colonies. Then 16 ant colonies were invited to invade any of the trees they chose. By construction, the 2 × 2 data, in `data(acacia)`, have both margins fixed. Use

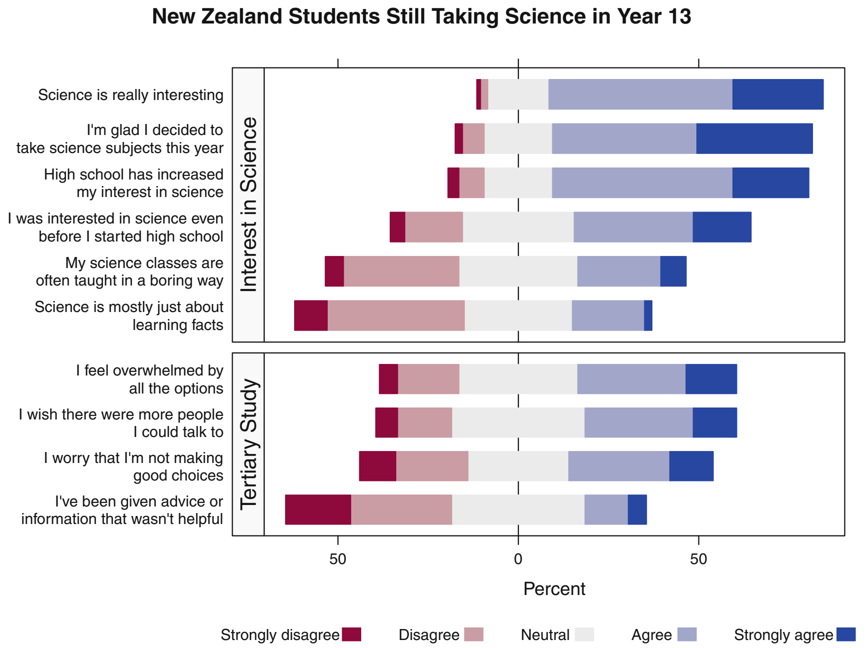


Fig. 15.18 Two sets of questions have been asked of all respondents. Each set is presented in its own panel.

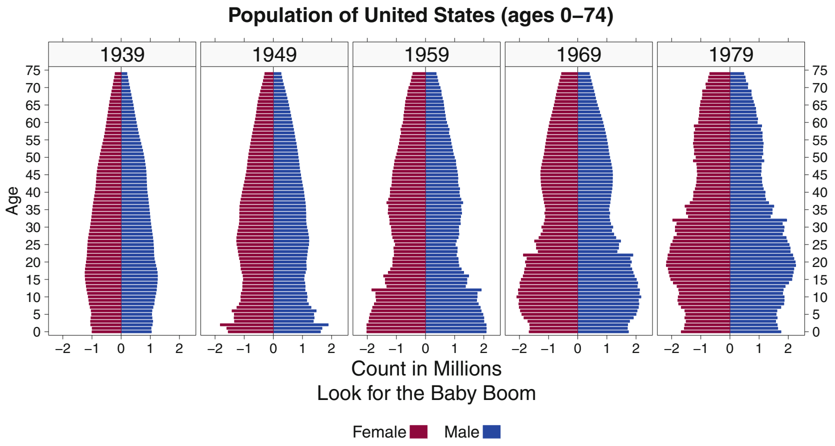


Fig. 15.19 Five population pyramids at ten-year intervals years 1939–1979. We can see the baby boom start at the bottom of the population graph for 1949 and work its way up over time. We have placed the age tick labels on both the left and right axes of the set of panels.

Fisher's exact test to determine if the ants have a significant preference for one species over the other.

15.4. Fleiss (1981) presents the dataset in `data(mortality)` concerning mortality following 37,840 live births to nonwhite mothers in New York City in 1974. In the file, rows are birth weights, ≤2500 grams or >2500 grams, and columns are outcomes one year after birth, dead or alive. 2500 grams is 5.5 pounds. Construct and carefully interpret the sample odds ratio for these data and construct a 95% confidence interval for the population odds ratio.

15.5. Wynder et al. (1958), later in Fleiss (1981), report a retrospective study of factors associated with cancer of the oral cavity. In this study there were 34 women with this cancer and 214 women, matched by age, without it. It was found that 24% of the cases but 66% of the controls were nonsmokers. The dataset is available as `data(oral)`. Construct and carefully interpret the sample odds ratio for these data and then construct a 95% confidence interval on the population odds ratio.

15.6. Braungart (1971) refers to `data(political)`, also found in Bishop et al. (1975), in which 271 college students of the 1960s who admitted to extreme political leanings were cross classified according to the style of parental decision making they received, authoritarian or democratic, and their political leaning, left or right. Construct and carefully interpret the sample odds ratio for these data and also construct a 95% confidence interval on the population odds ratio.

Table 15.11 Jury pool composition data for the Rotorua and Nelson districts.

	Rotorua		Nelson		Combined	
	Maori	Non-Maori	Maori	Non-Maori	Maori	Non-Maori
Jury pool	79	258	1	56	80	314
Others	8,810	23,751	1,328	32,602	10,138	56,353

15.7. Westbrooke (1998) discusses claims that Maori are underrepresented on juries in districts in New Zealand. Jury pool composition data for the Rotorua and Nelson districts are shown in Table 15.11, and accessed as `data(jury)`, along with totals for these two districts combined.

From this table it is easy to verify the following:

- The population of Rotorua is 27.0% Maori, but this district's jury pool is only 23.4% Maori.
- The population of Nelson is 3.9% Maori, but this district's jury pool is only 1.7% Maori.

- However, the combined population of these two districts is 15.3% Maori, but the combined jury pools of these districts is 20.3% Maori.

Discuss whether Maori are indeed underrepresented on the juries of these two districts.

15.8. Brochard et al. (1995) report a prospective study of patients with chronic obstructive pulmonary disease, assessing the effect of noninvasive ventilation therapy on reducing the need for subsequent invasive intubation. A total of 85 patients were recruited at five `centers`. The data are in Table 15.12 and dataset `data(intubate)`. The data in the file are arranged differently than in the table: The first column is the center number; the second column entries are `yes` if received ventilation therapy and `no` if didn't receive this therapy; the third column entries are `yes` if required invasive intubation and `no` if didn't require invasive intubation; and the entries in column 4 are the number of patients in the categories specified in columns 1–3.

Compute the odds ratio at each `center`. Use the Mantel–Haenszel test to produce a carefully stated conclusion of the combined data.

Table 15.12 Chronic Obstructive Pulmonary Disease.

	Center									
	1		2		3		4		5	
Ventilation therapy	invasive intubation	not inv	invasive intubation	not inv	invasive intubation	not inv	invasive intubation	not inv	invasive intubation	not inv
ventilation	3	6	2	3	1	7	0	5	5	11
not vent	9	0	5	1	4	5	3	1	10	4