Chapter 8

The analysis of count data

For the most part, this book concerns itself with measurement data and the corresponding analyses based on normal distributions. In this chapter we consider data that consist of counts. We begin in Section 8.1 by examining a set of data on the number of females admitted into graduate school at the University of California, Berkeley. A key feature of these data is that only two outcomes are possible: admittance or rejection. Data with only two outcomes are referred to as *binary (or dichotomous) data*. Often the two outcomes are referred to generically as success and failure. In Section 8.2, we expand our discussion by comparing two sets of dichotomous data; we compare Berkeley graduate admission rates for females and males. Section 8.3 examines *polytomous data*, i.e., count data in which there are more than two possible outcomes. For example, numbers of Swedish females born in the various months of the year involve counts for 12 possible outcomes. Section 8.4 examines comparisons between two samples of polytomous data, e.g., comparing the numbers of females and males that are born in the different months of the year. Section 8.5 looks at comparisons among more than two samples of polytomous data. The penultimate section considers a method of reducing large tables of counts that involve several samples of polytomous data into smaller more interpretable tables. The final section deals with a count data analogue of simple linear regression.

Sections 8.1 and 8.2 involve analogues of Chapters 2 and 4 that are appropriate for dichotomous data. The basic analyses in these sections simply involve new applications of the ideas in Chapter 3. Analyzing polytomous data requires techniques that are different from the methods of Chapter 3. Sections 8.3, 8.4, and 8.5 are polytomous data analogues of Chapters 2, 4, and 5. Everitt (1977) and Fienberg (1980) give more detailed introductions to the analysis of count data. Sophisticated analyses of count data frequently use analogues of ANOVA and regression called log-linear models. Christensen (1990b) provides an intermediate level account of log-linear models.

8.1 One binomial sample

The few distributions that are most commonly used in statistics arise naturally. The normal distribution arises for measurement data because the variability in the data often results from the mean of a large number of small errors and the central limit theorem indicates that such means tend to be normally distributed.

The binomial distribution arises naturally with count data because of its simplicity. Consider a number of trials, say n, each a success or failure. If each trial is independent of the other trials and if the probability of obtaining a success is the same for every trial, then the random number of successes has a binomial distribution. *The beauty of discrete data is that the probability models can often be justified solely by how the data were collected. This does not happen with measurement data.* The binomial distribution depends on two parameters, n, the number of independent trials, and the constant probability of success, say p. Typically, we know the value of n, while p is the unknown parameter of interest. Binomial distributions were examined in Section 1.4.

Bickel et al. (1975) report data on admissions to graduate school at the University of California, Berkeley. The numbers of females that were admitted and rejected are given below along with the total number of applicants.

Graduate admissions at Berkeley

	Admitted	Rejected	Total
Female	557	1278	1835

It seems reasonable to view the 1835 females as a random sample from a population of potential female applicants. We are interested in the probability p that a female applicant is admitted to graduate school. A natural estimate of the parameter p is the proportion of females that were actually admitted, thus our estimate of the parameter is

$$\hat{p} = \frac{557}{1835} = .30354.$$

We have a parameter of interest, p, and an estimate of that parameter, $\hat{p}$. If we can identify a standard error and an appropriate distribution, we can use the methods of Chapter 3 to perform statistical inferences.

The key to finding a standard error is to find the variance of the estimate. As we will see later,

$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}. \tag{8.1.1}$$

To estimate the standard deviation of $\hat{p}$, we simply use $\hat{p}$ to estimate p in (8.1.1) and take the square root. Thus the standard error is

$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.30354(1-.30354)}{1835}} = .01073.$$

The final requirement for using the results of Chapter 3 is to find an appropriate reference distribution for

$$\frac{\hat{p}-p}{\mathrm{SE}(\hat{p})}.$$

We can think of each trial as scoring either a 1, if the trial is a success, or a 0, if the trial is a failure. With this convention $\hat{p}$, the proportion of successes, is really the average of the 0–1 scores and since $\hat{p}$ is an average we can apply the central limit theorem. (In fact, $\mathrm{SE}(\hat{p})$ is very nearly $s/\sqrt{n}$, where s is computed from the 0–1 scores.) The central limit theorem simply states that for a large number of trials n, the distribution of $\hat{p}$ is approximately normal with a population mean that is the population mean of $\hat{p}$ and a population variance that is the population variance of $\hat{p}$. We have already given the variance of $\hat{p}$ and we will see later that $\mathrm{E}(\hat{p}) = p$. Thus for large n we have the approximation

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

The variance is unknown but by the law of large numbers it is approximately equal to our estimate of it, $\hat{p}(1-\hat{p})/n$. Standardizing the normal distribution (cf. Exercise 1.6.2) gives the approximation

$$\frac{\hat{p}-p}{\mathrm{SE}(\hat{p})} \sim N(0,1). \tag{8.1.2}$$

This distribution requires a sample size that is large enough for both the central limit theorem approximation and the law of large numbers approximation to be reasonably valid. For values of p that are not too close to 0 or 1, the approximation works reasonably well with sample sizes as small as 20.

We now have $Par = p$, $Est = \hat{p}$, $\mathrm{SE}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$ and the distribution in (8.1.2). As in Chapter 3, a 95% confidence interval for p has limits

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Here $1.96 = z(.975) = t(.975, \infty)$. Recall that a $(1-\alpha)100\%$ confidence interval requires the $(1-\alpha/2)$ percentile of the distribution. For the female admissions data, the limits are

$$.30354 \pm 1.96(.01073)$$

which gives the interval $(.28, .32)$. We are 95% confident that the population proportion of females admitted to Berkeley's graduate school is between .28 and .32. (As is often the case, it is not exactly clear what population these data relate to.)

We can also perform, say, an $\alpha = .01$ test of the null hypothesis $H_0 : p = 1/3$ versus the alternative $H_A : p \neq 1/3$. The test rejects H_0 if

$$\frac{\hat{p}-1/3}{\mathrm{SE}(\hat{p})} > 2.58$$

or if

$$\frac{\hat{p}-1/3}{\mathrm{SE}(\hat{p})} < -2.58.$$

Here $2.58 = z(.995) = t(.975, \infty)$. An α level two-sided test requires the $(1-\frac{\alpha}{2})100\%$ point of the distribution. The Berkeley data yield the test statistic

$$\frac{.30354 - .33333}{.01073} = -2.78$$

which is smaller than -2.58, so we reject the null hypothesis of $p = 1/3$ with $\alpha = .01$. In other words, we can reject, with strong assurance, the claim that one third of female applicants are admitted to graduate school at Berkeley. Since the test statistic is negative, we have evidence that the true proportion is less than one third. The test as constructed here is equivalent to checking whether $p = 1/3$ is within a 99% confidence interval.

There is an alternative, slightly different, way of performing tests such as $H_0 : p = 1/3$ versus $H_A : p \neq 1/3$. The difference involves using a different standard error. The variance of the estimate $\hat{p}$ is $p(1-p)/n$. In obtaining a standard error, we estimated p with $\hat{p}$ and took the square root of the estimated variance. Recalling that tests are performed *assuming that the null hypothesis is true*, it makes sense in the testing problem to use the assumption $p = 1/3$ in computing a standard error for $\hat{p}$. Thus an alternative standard error for $\hat{p}$ in this testing problem is

$$\sqrt{\frac{1}{3}\left(1-\frac{1}{3}\right)\bigg/1835} = .01100.$$

The test statistic now becomes

$$\frac{.30354 - .33333}{.01100} = -2.71.$$

Obviously, since the test statistic is slightly different, one could get slightly different answers for tests using the two different standard errors. Moreover, the results of this test will not always agree with a corresponding confidence interval for p because this test uses a different standard error than the confidence interval.

We should remember that the $N(0,1)$ distribution being used for the test is only a large sample approximation. (In fact, all of our results are only approximations.) The difference between the two standard errors is often minor compared to the level of approximation inherent in using the standard normal as a reference distribution. In any case, whether we ascribe the differences to the standard errors or to the quality of the normal approximations, the exact behavior of the two test statistics can be quite different when the sample size is small. Moreover, *when p is near 0 or 1, the sample sizes must be quite large to get a good normal approximation.*

The main theoretical results for a single binomial sample are establishing that $\hat{p}$ is a reasonable estimate of p and that the variance formula given earlier is correct. The data are $y \sim \text{Bin}(n, p)$. As seen in Section 1.4, $\text{E}(y) = np$ and $\text{Var}(y) = np(1-p)$. The estimate of p is $\hat{p} = y/n$. The estimate is unbiased because

$$\text{E}(\hat{p}) = \text{E}(y/n) = \text{E}(y)/n = np/n = p.$$

The variance of the estimate is

$$\text{Var}(\hat{p}) = \text{Var}(y/n) = \text{Var}(y)/n^2 = np(1-p)/n^2 = p(1-p)/n.$$

8.1.1 The sign test

We now consider an alternative analysis for paired comparisons based on the binomial distribution. Consider Burt's data on IQs of identical twins raised apart from Exercise 4.5.7 and Table 4.9. The earlier discussion of paired comparisons involved assuming and validating the normal distribution for the differences in IQs between twins. In the current discussion, we make the same assumptions as before except we replace the normality assumption with the weaker assumption that the distribution of the differences is symmetric. In the earlier discussion, we would test $H_0 : \mu_1 - \mu_2 = 0$. In the current discussion, we test whether there is a $50 : 50$ chance that y_1, the IQ for the foster parent raised twin, is larger than y_2, the IQ for the genetic parent raised twin. In other words, we test whether $\Pr(y_1 - y_2 > 0) = .5$. We have a sample of $n = 27$ pairs of twins. If $\Pr(y_1 - y_2 > 0) = .5$, the number of pairs with $y_1 - y_2 > 0$ has a $\text{Bin}(27, .5)$ distribution. From Table 4.9, 13 of the 27 pairs have larger IQs for the foster parent raised child. (These are the differences with a positive sign, hence the name sign test.) The proportion is $\hat{p} = 13/27 = .481$. The test statistic is

$$\frac{.481 - .5}{\sqrt{.5(1-.5)/27}} = -.20$$

which is nowhere near significant.

A similar method could be used to test, say, whether there is a 50 : 50 chance that y_1 is at least 3 IQ points greater than y_2. This hypothesis translates into $\Pr(y_1 - y_2 \geq 3) = .5$. The test is then based on the number of differences that are 3 or more.

The point of the sign test is the weakening of the assumption of normality. If the normality assumption is appropriate, the t test of Section 4.1 is more powerful. When the normality assumption is not appropriate, some modification like the sign test should be used. In this book, the usual approach is to check the normality assumption and, if necessary, to transform the data to make the normality assumption reasonable. For a more detailed introduction to *nonparametric* methods such as the sign test, see, for example, Conover (1971).

8.2 Two independent binomial samples

In this section we compare two independent binomial samples. Consider again the Berkeley admissions data. Table 8.1 contains data on admissions and rejections for the 1835 females considered in Section 8.1 along with data on 2691 males. We assume that the sample of females is independent of the sample of males. Throughout, we refer to the females as the first sample and the males as the second sample.

TABLE 8.1. Graduate admissions at Berkeley

	Admitted	Rejected	Total
Females	557	1278	1835
Males	1198	1493	2691

We consider being admitted to graduate school a ‘success’. Assuming that the females are a binomial sample, they have a sample size of $n_1 = 1835$ and some probability of success, say, p_1. The observed proportion of female successes is

$$\hat{p}_1 = \frac{557}{1835} = .30354.$$

Treating the males as a binomial sample, the sample size is $n_2 = 2691$ and the probability of success is, say, p_2. The observed proportion of male successes is

$$\hat{p}_2 = \frac{1198}{2691} = .44519.$$

Our interest is in comparing the success rate of females and males. The appropriate parameter is the difference in proportions,

$$Par = p_1 - p_2.$$

The natural estimate of this parameter is

$$Est = \hat{p}_1 - \hat{p}_2 = .30354 - .44519 = -.14165.$$

With independent samples, we can find the variance of the estimate and thus the standard error. Since the females are independent of the males,

$$\text{Var}(\hat{p}_1 - \hat{p}_2) = \text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2).$$

Using the variance formula in equation (8.1.1),

$$\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}. \tag{8.2.1}$$

Estimating p_1 and p_2 and taking the square root gives the standard error,

$$\begin{aligned} \mathrm{SE}(\hat{p}_1 - \hat{p}_2) &= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ &= \sqrt{\frac{.30354(1-.30354)}{1835} + \frac{.44519(1-.44519)}{2691}} \\ &= .01439. \end{aligned}$$

For large sample sizes n_1 and n_2, both $\hat{p}_1$ and $\hat{p}_2$ have approximate normal distributions and they are independent, so $\hat{p}_1 - \hat{p}_2$ has an approximate normal distribution and the appropriate reference distribution is approximately

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\mathrm{SE}(\hat{p}_1 - \hat{p}_2)} \sim N(0, 1).$$

We now have all the requirements for applying the results of Chapter 3. A 95% confidence interval for $p_1 - p_2$ has endpoints

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

where the value $1.96 = z(.975)$ is obtained from the $N(0, 1)$ distribution. For comparing the female and male admissions, the 95% confidence interval for the population difference in proportions has endpoints

$$-.14165 \pm 1.96(.01439).$$

The interval is $(-.17, -.11)$. Thus we are 95% confident that the proportion of women being admitted to graduate school at Berkeley is between .11 and .17 less than that for men.

To test $H_0 : p_1 = p_2$, or equivalently $H_0 : p_1 - p_2 = 0$, against $H_A : p_1 - p_2 \neq 0$, reject an $\alpha = .10$ test if

$$\frac{(\hat{p}_1 - \hat{p}_2) - 0}{\mathrm{SE}(\hat{p}_1 - \hat{p}_2)} > 1.645$$

or if

$$\frac{(\hat{p}_1 - \hat{p}_2) - 0}{\mathrm{SE}(\hat{p}_1 - \hat{p}_2)} < -1.645.$$

Again, the value 1.645 is obtained from the $N(0, 1) \equiv t(\infty)$ distribution. With the Berkeley data, the observed value of the test statistic is

$$\frac{-.14165 - 0}{.01439} = -9.84.$$

This is far smaller than -1.645, so the test rejects the null hypothesis of equal proportions at the .10 level. The test statistic is negative, so there is evidence that the proportion of women admitted to graduate school is lower than the proportion of men.

Once again, an alternative standard error is often used in testing problems. The test assumes that the null hypothesis is true and under the null hypothesis $p_1 = p_2$, so in constructing a standard error for the test statistic it makes sense to pool the data into one estimate of this common proportion. The pooled estimate is a weighted average of the individual estimates,

$$\begin{aligned}\hat{p}_* &= \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} \\ &= \frac{1835(.30354) + 2691(.44519)}{1835 + 2691} \\ &= \frac{557 + 1198}{1835 + 2691} \\ &= .38776\,.\end{aligned}$$

Using $\hat{p}_*$ to estimate both p_1 and p_2 in equation (8.2.1) and taking the square root gives the alternative standard error

$$\begin{aligned}\mathrm{SE}(\hat{p}_1 - \hat{p}_2) &= \sqrt{\frac{\hat{p}_*(1-\hat{p}_*)}{n_1} + \frac{\hat{p}_*(1-\hat{p}_*)}{n_2}} \\ &= \sqrt{\hat{p}_*(1-\hat{p}_*)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]} \\ &= \sqrt{.38776(1-.38776)\left[\frac{1}{1835} + \frac{1}{2691}\right]} \\ &= .01475\end{aligned}$$

The alternative test statistic is

$$\frac{-.14165 - 0}{.01475} = -9.60.$$

Again, the two test statistics are slightly different but the difference should be minor compared to the level of approximation involved in using the normal distribution.

A final note. Before you conclude that the data in Table 8.1 provide evidence of sex discrimination, you should realize that females tend to apply to different graduate programs than males. A more careful examination of the complete Berkeley data shows that the difference observed here results from females applying more frequently than males to highly restrictive programs, cf. Christensen (1990b, p. 96).

8.3 One multinomial sample

In this section we investigate the analysis a single polytomous variable, i.e., a count variable with more than two possible outcomes. In particular, we assume that the data are a sample from a *multinomial* distribution, cf. Section 1.5. The multinomial distribution is a generalization of the binomial that allows more than two outcomes. We assume that each trial gives one of, say, q possible outcomes. Each trial must be independent and the probability of each outcome must be the same for every trial. The multinomial distribution gives probabilities for the number of trials that fall into each of the possible outcome categories. The binomial distribution is a special case of the multinomial distribution in which $q = 2$.

The first two columns of Table 8.2 give months and numbers of Swedish females born in each month. The data are from Cramér (1946) who did not name the months. We assume that the data begin in January.

TABLE 8.2. Swedish female births by month

Month	Females	$\hat{p}$	Probability	E	$(O-E)/\sqrt{E}$
January	3537	.083	1/12	3549.25	−0.20562
February	3407	.080	1/12	3549.25	−2.38772
March	3866	.091	1/12	3549.25	5.31678
April	3711	.087	1/12	3549.25	2.71504
May	3775	.087	1/12	3549.25	3.78930
June	3665	.086	1/12	3549.25	1.94291
July	3621	.085	1/12	3549.25	1.20435
August	3596	.084	1/12	3549.25	0.78472
September	3491	.082	1/12	3549.25	−0.97775
October	3391	.080	1/12	3549.25	−2.65629
November	3160	.074	1/12	3549.25	−6.53372
December	3371	.079	1/12	3549.25	−2.99200
Total	42591	1	1	42591.00	

With polytomous data such as those listed in Table 8.2, there is no one parameter of primary interest. One might be concerned with the proportions of births in January, or December, or in any of the twelve months. With no one parameter of interest, the methods of Chapter 3 do not apply. Column 3 of Table 8.2 gives the observed proportions of births for each month. These are simply the monthly births divided by the total births for the year. Note that the proportion of births in March seems high and the proportion of births in November seems low.

A simplistic, yet interesting, hypothesis is that the proportion of births is the same for every month. To test this null hypothesis, we compare the number of observed births to the number of births we would expect to see if the hypothesis were true. The number of births we expect to see in any month is just the probability of having a birth in that month times the total number of births. The equal probabilities are given in column 4 of Table 8.2 and the expected values are given in column 5. The entries in column 5 are labeled E for expected value and are computed as $(1/12)42591 = 3549.25$. *It cannot be overemphasized that the expectations are computed under the assumption that the null hypothesis is true.*

Comparing observed values with expected values can be tricky. Suppose an observed value is 2145 and the expected value is 2149. The two numbers are off by 4; the observed value is pretty close to the expected. Now suppose the observed value is 1 and the expected value is 5. Again the two numbers are off by 4 but now the difference between observed and expected seems quite substantial. A difference of 4 means something very different depending on how large both numbers are. To account for this phenomenon, we standardized the difference between observed and expected counts. We do this by dividing the difference by the square root of the expected count. Thus, when we compare observed counts with expected counts we look at

$$\frac{O-E}{\sqrt{E}} \tag{8.3.1}$$

where O stands for the observed count and E stands for the expected count. The values in (8.3.1) are called *Pearson residuals*, after Karl Pearson.

The Pearson residuals for the Swedish female births are given in column 6 of Table 8.2. As noted earlier, the two largest deviations from the assumption of equal probabilities occur for March and November. Reasonably large deviations also occur for May and to a lesser extent December, April, October, and February. In general, *the Pearson residuals can be compared to observations from a $N(0,1)$ distribution to evaluate whether a residual is large.* For example, the residuals for March and November are 5.3 and −6.5. These are not values one is likely to observe from a $N(0,1)$ distribution; they provide strong evidence that birth rates in March are really larger than 1/12 and that birth rates in November are really smaller than 1/12.

Births seem to peak in March and they, more or less, gradually decline until November. After November, birth rates are still low but gradually increase until February. In March birth rates increase markedly. Birth rates are low in the fall and lower in the winter; they jump in March and remain relatively high, though decreasing, until September. This analysis could be performed using the monthly proportions of column 2 but the results are clearer using the residuals.

A statistic for testing whether the null hypothesis of equal proportions is reasonable can be obtained by squaring the residuals and adding them together. This statistic is known as *Pearson's* χ^2 (chi-squared) statistic and is computed as

$$X^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E}.$$

For the female Swedish births,

$$X^2 = 121.24.$$

Note that small values of X^2 indicate observed values that are similar to the expected values, so small values of X^2 are consistent with the null hypothesis. Large values of X^2 occur whenever one or more observed values are far from the expected values. To perform a test, we need some idea of how large X^2 could reasonably be when the null hypothesis is true. It can be shown that for a problem such as this with 1) a fixed number of cells q, here $q = 12$, with 2) a null hypothesis consisting of known probabilities such as those given in column 4 of Table 8.2, and with 3) large sample sizes for each cell, the null distribution of X^2 is approximately

$$X^2 \sim \chi^2(q-1).$$

The degrees of freedom are only $q-1$ because the $\hat{p}$s *must* add up to 1. Thus, if we know $q-1 = 11$ of the proportions, we can figure out the last one. Only $q-1$ of the cells are really free to vary. From Appendix B.2, the 99.5th percentile of a $\chi^2(11)$ distribution is $\chi^2(.995, 11) = 26.76$. The observed X^2 value of 121.24 is much larger than this, so the observed value of X^2 could not reasonably come from a $\chi^2(11)$ distribution. In particular, an $\alpha = .005$ test of the null hypothesis is rejected easily, so the P value for the test is 'much' less than .005. It follows that there is overwhelming evidence that the proportion of female Swedish births is not the same for all months.

In this example, our null hypothesis was that the probability of a female birth was the same in every month. A more reasonable hypothesis might be that the probability of a female birth is the same on every day. The months have different numbers of days so under this null hypothesis they have different probabilities. For example, assuming a 365 day year, the probability of a female birth in January is 31/365 which is somewhat larger than 1/12. Exercise 8.8.4 involves testing this alternative null hypothesis.

We can use results from Section 8.1 to help in the analysis of multinomial data. If we consider only the month of December, we can view each trial as a success if the birth is in December and a failure otherwise. Writing the probability of a birth in December as p_{12}, from Table 8.2 the estimate of p_{12} is

$$\hat{p}_{12} = \frac{3371}{42591} = .07915$$

with standard error

$$\text{SE}(\hat{p}_{12}) = \sqrt{\frac{.07915(1-.07915)}{42591}} = .00131$$

and a 95% confidence interval has endpoints

$$.07915 \pm 1.96(.00131).$$

The interval reduces to $(.077, .082)$. Tests for monthly proportions can be performed in a similar fashion. Bonferroni adjustments can be made to all tests and confidence intervals to control the experimentwise error rate for multiple tests or intervals, cf. Section 6.2.

8.4 Two independent multinomial samples

Table 8.3 gives monthly births for Swedish females and males along with various marginal totals. We wish to determine whether monthly birth rates differ for females and males. Denote the females as population 1 and the males as population 2. Thus we have a sample of 42591 females and, by assumption, an independent sample of 45682 males.

TABLE 8.3. Swedish births: monthly observations (O_{ij}s) and monthly proportions by sex

	Observations			Proportions	
Month	Female	Male	Total	Female	Male
January	3537	3743	7280	.083	.082
February	3407	3550	6957	.080	.078
March	3866	4017	7883	.091	.088
April	3711	4173	7884	.087	.091
May	3775	4117	7892	.089	.090
June	3665	3944	7609	.086	.086
July	3621	3964	7585	.085	.087
August	3596	3797	7393	.084	.083
September	3491	3712	7203	.082	.081
October	3391	3512	6903	.080	.077
November	3160	3392	6552	.074	.074
December	3371	3761	7132	.079	.082
Total	42591	45682	88273	1.000	1.000

In fact, it is more likely that there is actually only one sample here, one consisting of 88273 births. It is more likely that the births have been divided into 24 categories depending on sex and birth month. Such data can be treated as two independent samples with (virtually) no loss of generality. The interpretation of results for two independent samples is considerably simpler than the interpretation necessary for one sample cross-classified by both sex and month. Thus we discuss such data as though they are independent samples. The alternative interpretation involves a multinomial sample with the probabilities for month and sex pairs all being independent.

The number of births in month i for sex j is denoted O_{ij}, where $i = 1, \ldots, 12$ and $j = 1, 2$. Thus, for example, the number of males born in December is $O_{12,2} = 3761$. Let $O_{i\cdot}$ be the total for month i, $O_{\cdot j}$ be the total for sex j, and $O_{\cdot\cdot}$ be the total over all months and sexes. For example, May has $O_{5\cdot} = 7892$, males have $O_{\cdot 2} = 45682$, and the grand total is $O_{\cdot\cdot} = 88273$.

Our interest now is in whether the population proportion of births for each month is the same for females as for males. We no longer make any assumption about what these proportions are, our null hypothesis is simply that the proportions are the same in each month. Again, we wish to compare the observed values, the O_{ij}s with expected values, but now, since we do not have hypothesized proportions for any month, we must estimate the expected values.

Under the null hypothesis that the proportions are the same for females and males, it makes sense to pool the male and female data to get an estimate of the proportion of births in each month. Using the column of monthly totals in Table 8.3, the estimated proportion for January is the January total divided by the total for the year, i.e.,

$$\hat{p}_1^0 = \frac{7280}{88273} = .0824714.$$

In general, for month i we have

$$\hat{p}_i^0 = \frac{O_{i\cdot}}{O_{\cdot\cdot}}$$

where the superscript of 0 is used to indicate that these proportions are estimated under the null hypothesis of identical monthly rates for males and females. The estimate of the expected number of females born in January is just the number of females born in the year times the estimated probability of a birth in January,

$$\hat{E}_{11} = 42591(.0824714) = 3512.54.$$

The expected number of males born in January is the number of males born in the year times the estimated probability of a birth in January,

$$\hat{E}_{12} = 45682(.0824714) = 3767.46.$$

In general,

$$\hat{E}_{ij} = O_{\cdot j}\,\hat{p}_i^0 = O_{\cdot j}\frac{O_{i\cdot}}{O_{\cdot\cdot}} = \frac{O_{i\cdot}O_{\cdot j}}{O_{\cdot\cdot}}.$$

Again, *the estimated expected values are computed assuming that the proportions of births are the same for females and males in every month, i.e., assuming that the null hypothesis is true*. The estimated expected values under the null hypothesis are given in Table 8.4. Note that the totals for each month and for each sex remain unchanged.

TABLE 8.4. Estimated expected Swedish births by month ($\hat{E}_{ij}$s) and pooled proportions

	Expectations			Pooled
Month	Female	Male	Total	proportions
January	3512.54	3767.46	7280	.082
February	3356.70	3600.30	6957	.079
March	3803.48	4079.52	7883	.089
April	3803.97	4080.03	7884	.089
May	3807.83	4084.17	7892	.089
June	3671.28	3937.72	7609	.086
July	3659.70	3925.30	7585	.086
August	3567.06	3825.94	7393	.084
September	3475.39	3727.61	7203	.082
October	3330.64	3572.36	6903	.078
November	3161.29	3390.71	6552	.074
December	3441.13	3690.87	7132	.081
Total	42591.00	45682.00	88273	1.000

The estimated expected values are compared to the observations using Pearson residuals, just as in Section 8.3. The Pearson residuals are

$$\tilde{r}_{ij} \equiv \frac{O_{ij} - \hat{E}_{ij}}{\sqrt{\hat{E}_{ij}}}.$$

A more apt name for the Pearson residuals in this context may be *crude standardized residuals*. It is the standardization here that is crude and not the residuals. The standardization in the Pearson residuals ignores the fact that $\hat{E}$ is itself an estimate. Better, but considerably more complicated, standardized residuals can be defined for count data, cf. Christensen (1990b, Section IV.9). For the Swedish birth data, the Pearson residuals are given in Table 8.5. Note that when compared to a $N(0,1)$ distribution, none of the residuals is very large; all are smaller than 1.51 in absolute value.

TABLE 8.5. Pearson residuals for Swedish birth months, ($\tilde{r}_{ij}$s)

Month	Female	Male
January	0.41271	−0.39849
February	0.86826	−0.83837
March	1.01369	−0.97880
April	−1.50731	1.45542
May	−0.53195	0.51364
June	−0.10365	0.10008
July	−0.63972	0.61770
August	0.48452	−0.46785
September	0.26481	−0.25570
October	1.04587	−1.00987
November	−0.02288	0.02209
December	−1.19554	1.15438

As in Section 8.3, the sum of the squared Pearson residuals gives Pearson's χ^2 statistic for testing the null hypothesis of no differences between females and males. Pearson's test statistic is

$$X^2 = \sum_{ij} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}.$$

For the Swedish birth data, computing the statistic from the 24 cells in Table 8.5 gives

$$X^2 = 14.9858.$$

For a formal test, X^2 is compared to a χ^2 distribution. The appropriate number of degrees of freedom for the χ^2 test is the number of cells in the table adjusted to account for all the parameters we have estimated as well as the constraint that the sex totals sum to the grand total. There are 12×2 cells but only $12-1$ free months and only $2-1$ free sex totals. The appropriate distribution is $\chi^2((12-1)(2-1)) = \chi^2(11)$. *The degrees of freedom are the number of data rows in Table 8.3 minus 1 times the number of data columns in Table 8.3 minus 1.* The 90th percentile of a $\chi^2(11)$ distribution is $\chi^2(.9, 11) = 17.28$, so the observed test statistic $X^2 = 14.9858$ could reasonably come from a $\chi^2(11)$ distribution. In particular, the test is not significant at the .10 level. Moreover, $\chi^2(.75, 11) = 13.70$, so the test has a P value between .25 and .10. There is no evidence of any differences in the monthly birth rates for males and females.

FIGURE 8.1. Monthly Swedish birth proportions by sex: solid line, female; dashed line, male.

Another way to evaluate the null hypothesis is by comparing the observed monthly birth proportions by sex. These observed proportions are given in Table 8.3. If the populations of females and males have the same proportions of births in each month, the observed proportions of births in each month should be similar (except for sampling variation). One can compare the numbers directly in Table 8.3 or one can make a visual display of the observed proportions as in Figure 8.1.

The methods just discussed apply equally well to the binomial data of Table 8.1. Applying the X^2 test given here to the data of Table 8.1 gives

$$X^2 = 92.2.$$

The statistic X^2 is equivalent to the test statistic given in Section 8.2 using the pooled estimate $\hat{p}_*$ to compute the standard error. The test statistic in Section 8.2 is -9.60, and if we square this we get

$$(-9.60)^2 = 92.2 = X^2.$$

The -9.60 is compared to a $N(0,1)$, while the 92.2 is compared to a $\chi^2(1)$ because Table 8.1 has 2 rows and 2 columns. A $\chi^2(1)$ distribution is obtained by squaring a $N(0,1)$ distribution, so P values are identical and critical values are equivalent.

Minitab commands

Minitab commands for generating the analysis of Swedish birth rates are given below. Column c1 contains the observations, the O_{ij}s. Column c2 contains indices from 1 to 12 indicating the month of each observation and c3 contains indices for the two sexes. The subcommand ‘colpercents’ provides the proportions discussed in the analysis. The subcommand ‘chisquare 3’ gives the observations, estimated expected values, and Pearson residuals along with the Pearson test statistic.

```
MTB > read 'swede2.dat' c1 c2 c3
MTB > table c2 c3;
```

```
SUBC> frequencies c1;
SUBC> colpercents;
SUBC> chisquare 3.
```

8.5 Several independent multinomial samples

The methods of Section 8.4 extend easily to dealing with more than two samples. Consider the data in Table 8.6 that was extracted from Lazerwitz (1961). The data involve samples from three religious groups and consist of numbers of people in various occupational groups. The occupations are labeled A, professions; B, owners, managers, and officials; C, clerical and sales; and D, skilled. The three religious groups are Protestant, Roman Catholic, and Jewish. This is a subset of a larger collection of data that includes many more religious and occupational groups. The fact that we are restricting ourselves to a subset of a larger data set has no effect on the analysis. As discussed in Section 8.4, the analysis of these data is essentially identical regardless of whether the data come from one sample of 1926 individuals cross-classified by religion and occupation, or four independent samples of sizes 348, 477, 411, and 690 taken from the occupational groups, or three independent samples of sizes 1135, 648, and 143 taken from the religious groups. We choose to view the data as independent samples from the three religious groups. The data in Table 8.6 constitutes a 3×4 table because, excluding the totals, the table has 3 rows and 4 columns.

TABLE 8.6. Religion and occupations

	Occupation				
Religion	A	B	C	D	Total
Protestant	210	277	254	394	1135
Roman Catholic	102	140	127	279	648
Jewish	36	60	30	17	143
Total	348	477	411	690	1926

We again test whether the populations are the same. In other words, the null hypothesis is that the probability of falling into any occupational group is identical for members of the various religions. Under this null hypothesis, it makes sense to pool the data from the three religions to obtain estimates of the common probabilities. For example, under the null hypothesis of identical populations, the estimate of the probability that a person is a professional is

$$\hat{p}_1^0 = \frac{348}{1926} = .180685.$$

For skilled workers the estimated probability is

$$\hat{p}_4^0 = \frac{690}{1926} = .358255.$$

Denote the observations as O_{ij} with i identifying a religious group and j indicating occupation. We use a dot to signify summing over a subscript. Thus the total for religious group i is

$$O_{i\cdot} = \sum_j O_{ij},$$

the total for occupational group j is

$$O_{\cdot j} = \sum_i O_{ij},$$

and

$$O_{\cdot\cdot} = \sum_{ij} O_{ij}$$

is the grand total. Recall that the null hypothesis is that the probability of being in an occupation group is the same for each of the three populations. Pooling information over religions, we have

$$\hat{p}_j^0 = \frac{O_{\cdot j}}{O_{\cdot\cdot}}$$

as the estimate of the probability that someone in the study is in occupational group j. *This estimate is only appropriate when the null hypothesis is true.*

The estimated expected count under the null hypothesis for a particular occupation and religion is obtained by multiplying the number of people sampled in that religion by the probability of the occupation. For example, the estimated expected count under the null hypothesis for Jewish professionals is

$$\hat{E}_{31} = 143(.180685) = 25.84.$$

Similarly, the estimated expected count for Roman Catholic skilled workers is

$$\hat{E}_{24} = 648(.358255) = 232.15.$$

In general,

$$\hat{E}_{ij} = O_{i\cdot}\hat{p}_j^0 = O_{i\cdot}\frac{O_{\cdot j}}{O_{\cdot\cdot}} = \frac{O_{i\cdot}O_{\cdot j}}{O_{\cdot\cdot}}.$$

Again, *the estimated expected values are computed assuming that the null hypothesis is true.* The expected values for all occupations and religions are given in Table 8.7.

TABLE 8.7. Estimated expected counts ($\hat{E}_{ij}$s)

Religion	A	B	C	D	Total
Protestant	205.08	281.10	242.20	406.62	1135
Roman Catholic	117.08	160.49	138.28	232.15	648
Jewish	25.84	35.42	30.52	51.23	143
Total	348.00	477.00	411.00	690.00	1926

The estimated expected values are compared to the observations using Pearson residuals. The Pearson residuals are

$$\tilde{r}_{ij} = \frac{O_{ij} - \hat{E}_{ij}}{\sqrt{\hat{E}_{ij}}}.$$

These crude standardized residuals are given in Table 8.8 for all occupations and religions. The largest negative residual is -4.78 for Jewish people with occupation D. This indicates that Jewish people were substantially underrepresented among skilled workers relative to the other two religious groups. On the other hand, Roman Catholics were substantially overrepresented among skilled workers, with a positive residual of 3.07. The other large residual in the table is 4.13 for Jewish people in group B. Thus Jewish people were more highly represented among owners, managers, and officials than the other religious groups. Only one other residual is even moderately large, the 2.00 indicating a high level of Jewish people in the professions. The main feature of these data seems to be that the Jewish group was different from the other two. A substantial difference appears in every occupational group except clerical and sales.

TABLE 8.8. Residuals ($\tilde{r}_{ij}$s)

Religion	A	B	C	D
Protestant	0.34	−0.24	0.76	−0.63
Roman Catholic	−1.39	−1.62	−0.96	3.07
Jewish	2.00	4.13	−0.09	−4.78

As in Sections 8.3 and 8.4, the sum of the squared Pearson residuals gives Pearson's χ^2 statistic for testing the null hypothesis that the three populations are the same. Pearson's test statistic is

$$X^2 = \sum_{ij} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}.$$

Summing the squares of the values in Table 8.8 gives

$$X^2 = 60.0.$$

The appropriate number of degrees of freedom for the χ^2 test is the number of data rows in Table 8.6 minus 1 times the number of data columns in Table 8.6 minus 1. Thus the appropriate reference distribution is $\chi^2((3-1)(4-1)) = \chi^2(6)$. The 99.5th percentile of a $\chi^2(6)$ distribution is $\chi^2(.995, 6) = 18.55$ so the observed statistic $X^2 = 60.0$ could not reasonably come from a $\chi^2(6)$ distribution. In particular, the test is significant at the .005 level, clearly indicating that the proportions of people in the different occupation groups differ with religious group.

As in the previous section, we can informally evaluate the null hypothesis by examining the observed proportions for each religious group. The observed proportions are given in Table 8.9. Under the null hypothesis, the observed proportions in each occupation category should be the same for all the religions (up to sampling variability). Figure 8.2 displays the observed proportions graphically. The Jewish group is obviously very different from the other two groups in occupations B and D and is very similar in occupation C. The Jewish proportion seems somewhat different for occupation A. The Protestant and Roman Catholic groups seem similar except that the Protestants are a bit underrepresented in occupation D and therefore are overrepresented in the other three categories. (Remember that the four proportions for each religion must add up to one, so being underrepresented in one category forces an overrepresentation in one or more other categories.)

TABLE 8.9. Observed proportions by religion

	Occupation				
Religion	A	B	C	D	Total
Protestant	.185	.244	.224	.347	1.00
Roman Catholic	.157	.216	.196	.431	1.00
Jewish	.252	.420	.210	.119	1.00

FIGURE 8.2. Occupational proportions by religion: solid – protestant, long dashes – catholic, short dashes – jewish.

8.6 Lancaster–Irwin partitioning

Lancaster–Irwin partitioning is a method for breaking a table of count data into smaller tables. When used to its maximum extent, partitioning is similar in spirit to looking at contrasts in analysis of variance. The basic idea is that a table of counts can be broken into two component tables, a reduced table and a collapsed table. Table 8.10 illustrates such a partition for the data of Table 8.6. In the reduced table, the row for the Jewish group has been eliminated, leaving a subset of the original table. In the collapsed table, the two rows in the reduced table, Protestant and Roman Catholic, have been collapsed into a single row.

In Lancaster–Irwin partitioning, we pick a group of either rows or columns, say rows. The reduced table involves all of the columns but only the chosen subgroup of rows. The collapsed table involves all of the columns and all of the rows *not* in the chosen subgroup, along with a row that combines (collapses) all of the subgroup rows into a single row. In Table 8.10 the chosen subgroup of rows contains the Protestants and Roman Catholics. The reduced table involves all occupational groups but only the Protestants and Roman Catholics. In the collapsed table the occupational groups are unaffected but the Protestants and Roman Catholics are combined into a single row. The other rows remain the same; in this case the other rows consist only of the Jewish row. As alluded to above, rather than picking a group of rows to form the partitioning, we could select a group of columns.

Lancaster–Irwin partitioning is by no means a unique process. There are as many ways to partition a table as there are ways to pick a group of rows or columns. In Table 8.10 we made a particular selection based on the residual analysis of these data from the previous section. The main feature we discovered in the residual analysis was that the Jewish group seemed to be different from the other two groups. Thus it seemed to be of interest to compare the Jewish group with a combination of the others and then to investigate what differences there might be among the other religious groups. The partitioning of Table 8.10 addresses precisely these questions.

Tables 8.11 and 8.12 provide statistics for the analysis of the reduced table and collapsed table. The reduced table simply reconfirms our previous conclusions. The X^2 value of 12.3

TABLE 8.10. A Lancaster–Irwin partition of Table 8.6

Reduced table

Religion	A	B	C	D	Total
Protestant	210	277	254	394	1135
Roman Catholic	102	140	127	279	648
Total	312	417	381	673	1783

Collapsed table

Religion	A	B	C	D	Total
Prot. & R.C.	312	417	381	673	1783
Jewish	36	60	30	17	143
Total	348	477	411	690	1926

indicates substantial evidence of a difference between Protestants and Roman Catholics. The percentage point $\chi^2(.995, 3) = 12.84$ indicates that the P value for the test is a bit greater than .005. The residuals indicate that the difference was due almost entirely to the fact that Roman Catholics have relatively higher representation among skilled workers. (Or equivalently, that Protestants have relatively lower representation among skilled workers.) Overrepresentation of Roman Catholics among skilled workers forces their underrepresentation among other occupational groups but the level of underrepresentation in the other groups was approximately constant as indicated by the approximately equal residuals for Roman Catholics in the other three occupation groups. We will see later that for Roman Catholics in the other three occupation groups, their distribution among those groups was almost the same as those for Protestants. This reinforces the interpretation that the difference was due almost entirely to the difference in the skilled group.

The conclusions that can be reached from the collapsed table are also similar to those drawn in the previous section. The X^2 value of 47.5 on 3 degrees of freedom indicates overwhelming evidence that the Jewish group was different from the combined Protestant–Roman Catholic group. The residuals can be used to isolate the sources of the differences. The two groups differed in proportions of skilled workers and proportions of owners, managers, and officials. There was a substantial difference in the proportions of professionals. There was almost no difference in the proportion of clerical and sales workers between the Jewish group and the others.

The X^2 value computed for Table 8.6 was 60.0. The X^2 value for the collapsed table is 47.5 and the X^2 value for the reduced table is 12.3. Note that $60.0 \doteq 59.8 = 47.5 + 12.3$. It is not by chance that the sum of the X^2 values for the collapsed and reduced tables is approximately equal to the X^2 value for the original table. In fact, this relationship is a primary reason for using the Lancaster–Irwin partitioning method. The approximate equality $60.0 \doteq 59.8 = 47.5 + 12.3$ indicates that the vast bulk of the differences between the three religious groups is due to the collapsed table, i.e., the difference between the Jewish group and the other two. Roughly 80% (47.5/60) of the original X^2 value is due to the difference between the Jewish group and the others. Of course the X^2 value 12.2 for the reduced table is still large enough to strongly suggest differences between Protestants and Roman Catholics.

Not all data will yield an approximation as close as $60.0 \doteq 59.8 = 47.5 + 12.3$ for the partitioning. The fact that we have an approximate equality rather than an exact equality is due to our choice of the test statistic X^2. Pearson's statistic is simple and intuitive; it compares observed values with expected values and standardizes by the size of the expected

TABLE 8.11. Reduced table

	Observations				
Religion	A	B	C	D	Total
Protestant	210	277	254	394	1135
Roman Catholic	102	140	127	279	648
Total	312	417	381	673	1783

	Estimated expected counts				
Religion	A	B	C	D	Total
Protestant	198.61	265.45	242.53	428.41	1135
Roman Catholic	113.39	151.55	138.47	244.59	648
Total	312.00	417.00	381.00	673.00	1783

	Pearson residuals			
Religion	A	B	C	D
Protestant	0.81	0.71	0.74	−1.66
Roman Catholic	−1.07	−0.94	−0.97	2.20

$X^2 = 12.3$, $df = 3$

TABLE 8.12. Collapsed table

	Observations				
Religion	A	B	C	D	Total
Prot. & R.C.	312	417	381	673	1783
Jewish	36	60	30	17	143
Total	348	477	411	690	1926

	Estimated expected counts				
Religion	A	B	C	D	Total
Prot. & R.C.	322.16	441.58	380.48	638.77	1783
Jewish	25.84	35.42	30.52	51.23	143
Total	348.00	477.00	411.00	690.00	1926

	Pearson residuals			
Religion	A	B	C	D
Prot. & R.C.	−0.57	−1.17	0.03	1.35
Jewish	2.00	4.13	−0.09	−4.78

$X^2 = 47.5$, $df = 3$

value. An alternative test statistic also exists called the likelihood ratio test statistic. The motivation behind the likelihood ratio test statistic is not as transparent as that behind Pearson's statistic, so we will not discuss the likelihood ratio test statistic in any detail. However, one advantage of the likelihood ratio test statistic is that the sum of its values for the reduced table and collapsed table gives *exactly* the likelihood ratio test statistic for the original table. For more discussion of the likelihood ratio test statistic, see Christensen (1990b, chapter II).

Further partitioning

We began this section with the 3×4 data of Table 8.6 that has 6 degrees of freedom for its X^2 test. We partitioned the data into two 2×4 tables, each with 3 degrees of freedom.

We can continue to use the Lancaster–Irwin method to partition the reduced and collapsed tables given in Table 8.10. The process of partitioning previously partitioned tables can be continued until the original table is broken into a collection of 2×2 tables. Each 2×2 table has one degree of freedom for its chi-squared test, so partitioning provides a way of breaking a large table into one degree of freedom components. This is similar in spirit to looking at contrasts in analysis of variance. Contrasts break the sum of squares for treatments into one degree of freedom components.

What we have been calling the reduced table involves all four occupational groups along with the two religious groups Protestant and Roman Catholic. The table was given in both Table 8.10 and Table 8.11. We now consider this table further. It was discussed earlier that the difference between Protestants and Roman Catholics can be ascribed almost entirely to the difference in the proportion of skilled workers in the two groups. To explore this we choose a new partition based on a group of *columns* that includes all occupations other than the skilled workers. Thus we get the 'reduced' table in Table 8.13 with occupations A, B, and C and the 'collapsed' table in Table 8.14 with occupation D compared to the accumulation of the other three.

TABLE 8.13.

	Observations			
Religion	A	B	C	Total
Protestant	210	277	254	741
Roman Catholic	102	140	127	369
Total	312	417	381	1110

	Estimated expected counts			
Religion	A	B	C	Total
Protestant	208.28	278.38	254.34	741
Roman Catholic	103.72	138.62	126.66	369
Total	312.00	417.00	381.00	1110

	Pearson residuals		
Religion	A	B	C
Protestant	0.12	−0.08	0.00
Roman Catholic	−0.17	0.12	0.03

$X^2 = .065$, $df = 2$

Table 8.13 allows us to examine the proportions of Protestants and Catholics in the occupational groups A, B, and C. We are not investigating whether Catholics were more or less likely than Protestants to enter these occupational groups; we are examining their distribution *within* the groups. The analysis is based only on those individuals *that were in this collection of three occupational groups*. The X^2 value is exceptionally small, only .065. There is no evidence of any difference between Protestants and Catholics for these three occupational groups.

Table 8.13 is a 2×3 table. We could partition it again into two 2×2 tables but there is little point in doing so. We have already established that there is no evidence of differences.

Table 8.14 has the three occupational groups A, B, and C collapsed into a single group. This table allows us to investigate whether Catholics were more or less likely than Protestants to enter this group of three occupations. The X^2 value is a substantial 12.2 on one degree of freedom, so we can tentatively conclude that there was a difference between Protestants and Catholics. From the residuals, we see that *among people in the four occupational*

TABLE 8.14.

	Observations		
Religion	A & B & C	D	Total
Protestant	741	394	1135
Roman Catholic	369	279	648
Total	1110	673	1783

	Estimated expected counts		
Religion	A & B & C	D	Total
Protestant	706.59	428.41	1135
Roman Catholic	403.41	244.59	648
Total	1110.00	673.00	1783

	Pearson residuals		
Religion	A & B & C	D	
Protestant	1.29	−1.66	
Roman Catholic	−1.71	2.20	

$X^2 = 12.2$, $df = 1$

groups, Catholics were more likely than Protestants to be in the skilled group and less likely to be in the other three.

Table 8.14 is a 2×2 table so no further partitioning is possible. Note again that the X^2 of 12.3 from Table 8.11 is approximately equal to the sum of the .065 from Table 8.13 and the 12.2 from Table 8.14.

Finally, we consider additional partitioning of the collapsed table given in Tables 8.10 and 8.12. It was noticed earlier that the Jewish group seemed to differ from Protestants and Catholics in every occupational group except C, clerical and sales. Thus we choose a partitioning that isolates group C. Table 8.15 gives a collapsed table that compares C to the combination of groups A, B, and D. Table 8.16 gives a reduced table that involves only occupational groups A, B, and D.

TABLE 8.15.

	Observations		
Religion	A & B & D	C	Total
Prot. & R.C.	1402	381	1783
Jewish	113	30	143
Total	1515	411	1926

	Estimated expected counts		
Religion	A & B & D	C	Total
Prot. & R.C.	1402.52	380.48	1783
Jewish	112.48	30.52	143
Total	1515.00	411.00	1926

	Pearson residuals		
Religion	A & B & D	C	
Prot. & R.C.	−0.00	0.03	
Jewish	0.04	−0.09	

$X^2 = .01$, $df = 1$

Table 8.15 demonstrates no difference between the Jewish group and the combined Protestant–Catholic group. Thus the proportion of people in clerical and sales was the same for the Jewish group as for the combined Protestant and Roman Catholic group. Any differences between the Jewish and Protestant–Catholic groups must be in the proportions of people *within* the three occupational groups A, B, and D.

TABLE 8.16.

	Observations			
Religion	A	B	D	Total
Prot. & R.C.	312	417	673	1402
Jewish	36	60	17	113
Total	348	477	690	1515

	Estimated expected counts			
Religion	A	B	D	Total
Prot. & R.C.	322.04	441.42	638.53	1402
Jewish	25.96	35.58	51.47	113
Total	348.00	477.00	690.00	1515

	Pearson residuals		
Religion	A	B	D
Prot. & R.C.	−0.59	−1.16	1.36
Jewish	1.97	4.09	−4.80

$X^2 = 47.2$, $df = 2$

Table 8.16 demonstrates major differences between occupations A, B, and D for the Jewish group and the combined Protestant–Catholic group. As seen earlier and reconfirmed here, skilled workers had much lower representation among the Jewish group, while professionals and especially owners, managers, and officials had much higher representation among the Jewish group.

Table 8.16 can be further partitioned into Tables 8.17 and 8.18. Table 8.17 is a reduced 2×2 table that considers the difference between the Jewish group and others with respect to occupational groups B and D. Table 8.18 is a 2×2 collapsed table that compares occupational group A with the combination of groups B and D.

Table 8.17 shows a major difference between occupational groups B and D. Table 8.18 may or may not show a difference between group A and the combination of groups B and D. The X^2 values are 46.8 and 5.45 respectively. The question is whether an X^2 value of 5.45 is suggestive of a difference between religious groups when we have examined the data in order to choose the partitions of Table 8.6. Note that the two X^2 values sum to 52.25, whereas the X^2 value for Table 8.16, from which they were constructed, is only 47.2. The approximate equality is a very rough approximation. Nonetheless, we see from the relative sizes of the two X^2 values that the majority of the difference between the Jewish group and the other religious groups was in the proportion of owners, managers, and officials as compared to the proportion of skilled workers.

Ultimately, we have partitioned Table 8.6 into Tables 8.13, 8.14, 8.15, 8.17, and 8.18. These are all 2×2 tables except for Table 8.13. We could also have partitioned Table 8.13 into two 2×2 tables but we chose to leave it because it showed so little evidence of any difference between Protestants and Roman Catholics for the three occupational groups considered. The X^2 value of 60.0 for Table 8.6 was approximately partitioned into X^2 values of .065, 12.2, .01, 46.8, and 5.45 respectively. Except for the .065 from Table 8.13,

TABLE 8.17.

Religion	Observations B	D	Total
Prot. & R.C.	417	673	1090
Jewish	60	17	77
Total	477	690	1167

Religion	Estimated expected counts B	D	Total
Prot. & R.C.	445.53	644.47	1090
Jewish	31.47	45.53	77
Total	477.00	690.00	1167

Religion	Pearson residuals B	D	
Prot. & R.C.	−1.35	1.12	
Jewish	5.08	−4.23	

$X^2 = 46.8$, $df = 1$

TABLE 8.18.

Religion	Observations A	B & D	Total
Prot. & R.C.	312	1090	1402
Jewish	36	77	113
Total	348	1167	1515

Religion	Estimated expected counts A	B & D	Total
Prot. & R.C.	322.04	1079.96	1402
Jewish	25.96	87.04	113
Total	348.00	1167.00	1515

Religion	Pearson residuals A	B & D	
Prot. & R.C.	−0.56	0.30	
Jewish	1.97	−1.08	

$X^2 = 5.45$, $df = 1$

each of these values is computed from a 2×2 table, so each has 1 degree of freedom. The .065 is computed from a 2×3 table, so it has 2 degrees of freedom. The sum of the five X^2 values is 64.5 which is roughly equal to the 60.0 from Table 8.6.

The five X^2 values can all be used in testing. Not only does such testing involve the usual problems associated with multiple testing but we even let the data suggest the partitions. It is inappropriate to compare these X^2 values to their usual χ^2 percentage points to obtain tests. A simple way to adjust for both the multiple testing and the data dredging (letting the data suggest partitions) is to compare all X^2 values to the percentage points appropriate for Table 8.6. For example, the $\alpha = .05$ test for Table 8.6 uses the critical value $\chi^2(.95, 6) = 12.58$. By this standard, Table 8.17 with $X^2 = 46.8$ shows a significant difference between religious groups and Table 8.14 with $X^2 = 12.2$ nearly shows a significant difference between religious groups. The value of $X^2 = 5.45$ for Table 8.18 gives no evidence of a difference based on this criterion even though such a value would be highly suggestive if we could compare it to a $\chi^2(1)$ distribution. This method is similar in spirit to Scheffé's method from Section 6.4 and suffers from the same extreme conservatism.

8.7 Logistic regression

Logistic regression is a method of modeling the relationships between probabilities and predictor variables. We begin with an example.

EXAMPLE 8.7.1. Woodward et al. (1941) reported data on 120 mice divided into 12 groups of 10. The mice in each group were exposed to a specific dose of chloracetic acid and the observations consist of the number in each group that lived and died. Doses were measured in grams of acid per kilogram of body weight. The data are given in Table 8.19, along with the proportions of mice who died at each dose. We could analyze these data using the methods discussed earlier in this chapter; we have samples from twelve populations and we could test to see if the populations are the same. In addition though, we can try to model the relationship between dose level and the probability of dying. If we can model the probability of dying as a function of dose, we can make predictions about the probability of dying for any dose levels that are similar to those in the original data. □

TABLE 8.19. Lethality of chloracetic acid

Dose	Group	Died	Survived	Total	$\hat{p}_i$
.0794	1	1	9	10	.1
.1000	2	2	8	10	.2
.1259	3	1	9	10	.1
.1413	4	0	10	10	.0
.1500	5	1	9	10	.1
.1588	6	2	8	10	.2
.1778	7	4	6	10	.4
.1995	8	6	4	10	.6
.2239	9	4	6	10	.4
.2512	10	5	5	10	.5
.2818	11	5	5	10	.5
.3162	12	8	2	10	.8

Logistic regression as applied to this example is somewhat like fitting a simple linear regression to one-way ANOVA data as discussed in Section 7.12. In Section 7.12 we considered data on the ASI indices given in Table 7.10. These data have seven observations on each of five plate lengths. The data can be analyzed as either a one-way ANOVA or as a simple linear regression, and in Section 7.12 we examined relationships between the two approaches. In particular, we mentioned that the estimated regression line could be obtained by fitting a line to the sample means for the five groups. The analysis of the lethality data takes a similar approach. Instead of fitting a line to sample means, we perform a regression on the observed proportions. Unfortunately, a standard regression is inappropriate because the observed proportions do not have constant variance. For $i = 1, \ldots, q$, $\hat{p}_i$ is the observed proportion from N_i binomial trials, so as discussed in Section 8.1, $\text{Var}(\hat{p}_i) = p_i(1 - p_i)/N_i$. One approach is to use the variance stabilizing transformation from Sections 2.3 and 7.10

on the $\hat{p}_i$s and then apply standard regression methods. As alluded to in Section 2.3, there are better methods available and this section briefly introduces some of them.

We begin with a reasonably simple analysis of the chloracetic acid data. This analysis involves not only a transformation of the $\hat{p}_i$s but incorporating weights into the simple linear regression procedure. Weighted regression is a method for dealing with nonconstant variances in the observations. If the variances are not constant, *observations with large variances should be given relatively little weight, while observations with small variances are given increased weight.* The details of weighted regression are discussed in Section 15.7. The discussion given there requires one to know the material in Chapter 13 and the first five sections of Chapter 15, but considerable insight can be obtained from Examples 15.7.1 and 15.7.2. These examples merely require the background from Section 7.12.

In weighted regression for binomial data we take the observations on the dependent variable as

$$\log[\hat{p}_i/(1-\hat{p}_i)].$$

We then fit the model

$$\log[\hat{p}_i/(1-\hat{p}_i)] = \beta_0 + \beta_1 x_i + \varepsilon_i$$

with weights

$$w_i = N_i \hat{p}_i(1-\hat{p}_i).$$

The regression estimates from this method minimize the weighted sum of squares

$$\sum_{i=1}^{q} w_i \left(\log[\hat{p}_i/(1-\hat{p}_i)] - \beta_0 - \beta_1 x_i\right)^2 .$$

There are a couple of serious drawbacks to this procedure. First, *the weights are really only appropriate if all the samples sizes N_i are large.* The weights rely on large sample variance formulae and the law of large numbers. Second, *the values* $\log[\hat{p}_i/(1-\hat{p}_i)]$ *are not always defined.* If we have an observed proportion with $\hat{p}_i = 0$ or 1, $\log[\hat{p}_i/(1-\hat{p}_i)]$ is undefined. Either we are trying to take the log of zero or we are trying to divide by zero. With $\hat{p}_i = y_i/N_i$, so that y_i is the number of 'successes,' this problem occurs whenever y_i equals 0 or N_i. The problem is often dealt with by adding or subtracting a small number to y_i. Generally, *the size of the small number should be chosen to be small in relation to the size of N_i.* In many applications, *all* of the N_is are 1. *In any case with $N_i = 1$, $\hat{p}_i$ is always either 0 or 1, so* $\log[\hat{p}_i/(1-\hat{p}_i)]$ *is always undefined.* These drawbacks are not as severe with another method of analysis that we will examine later.

EXAMPLE 8.7.1 CONTINUED. We now return to the chloracetic acid data. In this example $N_i = 10$ for all i, so the sample sizes are all reasonably large. For dose $x = .1413$, the number of deaths was 0, so the observed proportion was zero. We handle this problem by treating the observed count as .5, so the observed proportion is taken as .5/10 = .05. A computer program for regression analysis will typically give output such as the following tables.

Raw parameter table

Predictor	$\hat{\beta}_k$	SE($\hat{\beta}_k$)	t	P
Constant	−3.1886	0.5914	−5.39	0.000
Dose	13.181	2.779	4.74	0.000

Analysis of variance: weighted simple linear regression

Source	df	SS	MS	F	P
Regression	1	15.282	15.282	22.50	0.000
Error	10	6.791	0.679		
Total	11	22.074			

The estimates of the regression parameters are appropriate but everything involving variances in these tables is wrong! The problem is that with binomial data the variance depends solely on the probability and we have already accounted for the variance in defining the weights. Thus there is no separate parameter σ^2 to deal with but standard regression output is designed to adjust for such a parameter. To obtain appropriate standard errors, we need to divide the reported standard errors by $\sqrt{MSE}$. The adjusted table is given below.

Adjusted parameter table

Predictor	$\hat{\beta}_k$	SE($\hat{\beta}_k$)	t	P
Constant	−3.1886	0.7177	−4.44	0.000
Dose	13.181	3.373	3.91	0.000

The table provides clear evidence of the need for both parameters. To predict the probability of death for rats given a dose x, the predicted probability $\hat{p}$ satisfies

$$\log[\hat{p}/(1-\hat{p})] = \hat{\beta}_0 + \hat{\beta}_1 x = -3.1886 + 13.181x.$$

Solving for $\hat{p}$ gives

$$\hat{p} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 x)} = \frac{\exp(-3.1886 + 13.181x)}{1+\exp(-3.1886 + 13.181x)}.$$

For example, if $x = .3$, $-3.1886 + 13.181(.3) = .7657$ and $\hat{p} = e^{.7657}/(1+e^{.7657}) = .68$.

The only interest in the ANOVA table is in the error line. As we have seen, $\sqrt{MSE}$ is needed to adjust the standard errors. In addition, the SSE provides a lack of fit test similar in spirit to that discussed in Section 7.12. To test for lack of fit compare SSE to a $\chi^2(dfE)$ distribution. Large values of SSE indicate lack of fit. In this example $SSE = 6.791$, which is smaller than $dfE = 10$, so the χ^2 test gives no evidence of lack of fit. A line seems to fit these data adequately. □

Minitab commands

In Minitab let c1 contain the doses, c2 contain the number of deaths, and c3 contain the number of trials (10 in each case). The commands for this analysis are given below.

```
MTB > let c5=c2/c3
MTB > let c6=1-c5
MTB > let c7=loge(c5/c6)
MTB > let c8=c3*c5*c6
MTB > regress c7 on 1 c1;
SUBC> weights c8.
```

The logistic model and maximum likelihood

When we have a one-way ANOVA with treatments that are quantitative levels of some factor, we can fit either the one-way ANOVA model

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

or the simple linear regression model

$$y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij}.$$

We can think of the regression as a model for the μ_is, i.e.,

$$\mu_i = \beta_0 + \beta_1 x_i.$$

Logistic regression uses a very similar idea. The binomial situation here has 'observations' $\hat{p}_i = y_i/N_i$ where $y_i \sim \text{Bin}(N_i, p_i)$, $i = 1, \ldots, q$. In logistic regression, we model the parameters p_i. In particular, the model is

$$\log[p_i/(1-p_i)] = \beta_0 + \beta_1 x_i. \tag{8.7.1}$$

The question is then how to fit this model. The weighted regression approach was discussed earlier. The weighted regression estimates are the values of β_0 and β_1 that minimize the function

$$\sum_{i=1}^{q} w_i \left(\log[\hat{p}_i/(1-\hat{p}_i)] - \beta_0 - \beta_1 x_i\right)^2.$$

An alternative method for estimating the parameters is to maximize something called the likelihood function.

Recall from Section 1.4 that the probability function for an individual binomial, say, $y_i \sim \text{Bin}(N_i, p_i)$ is

$$\Pr(y_i = r_i) = \binom{N_i}{r_i} p_i^{r_i}(1-p_i)^{N_i - r_i}.$$

We are dealing with q *independent* binomials, so probabilities for the entire collection of random variables are obtained by multiplying the probabilities for the individual events.

One of the things that students initially find confusing about statistical theory is that we often use the same symbols for random variables and for observations from those random variables. I am about to do the same thing. I want to write down the probability of the data that we actually saw. If we saw y_i, the probability of seeing that is

$$\binom{N_i}{y_i} p_i^{y_i}(1-p_i)^{N_i - y_i}.$$

If all together we saw $y_1, \ldots, y_q$, the probability of obtaining all those values is the product of the individual probabilities, i.e.,

$$\prod_{i=1}^{q} \binom{N_i}{y_i} p_i^{y_i}(1-p_i)^{N_i - y_i}. \tag{8.7.2}$$

This probability of getting the observed data is called the *likelihood function*. In the likelihood function we know all of the N_is and y_is but we do not know the p_is. Thus the likelihood is a function of the p_is. It is not too difficult to show that the maximum value of the likelihood function is obtained by taking $p_i = \hat{p}_i = y_i/N_i$ for all i. The observed proportions $\hat{p}_i$ are the values of the parameters that maximize the probability of getting the observed data. We say that such values are *maximum likelihood estimates (mles)* of the parameters p_i.

The model (8.7.1) specifies the p_is in terms of β_0 and β_1. We can solve (8.7.1) for p_i by writing

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}. \tag{8.7.3}$$

If we now substitute this formula for p_i into equation (8.7.2) we get the likelihood as a function of β_0 and β_1. The maximum likelihood estimates of β_0 and β_1 are simply the values of β_0 and β_1 that maximize the likelihood function. Equations (8.7.1) and (8.7.3) are equivalent ways of writing the model. Equation (8.7.3) is actually the logistic regression model and equation (8.7.1) is the corresponding *logit* model.

Computer programs are available for finding maximum likelihood estimates. Such programs typically give standard errors that are valid for large samples. If the large sample approximations are appropriate, the parameters, estimates, and standard errors can be used as in Chapter 3 with a $N(0,1)$ reference distribution. For the approximations to be valid, it is typically enough that the total number of trials in the entire data be large; the individual sample sizes N_i need not be large.

Maximum likelihood theory also provides a test of lack of fit similar to the weighted regression χ^2 test using the SSE. In maximum likelihood theory the test examines the value of the likelihood (8.7.2) when using the mles of β_0 and β_1 in equation (8.7.3) to determine the p_is, and compares that value to the likelihood when using the observed proportions $\hat{p}_i$ as the p_is. Using the observed proportions involves less structure so the likelihood value will be greater using them. The lack of fit test statistic is -2 times the log of the ratio of the likelihood using the estimated β_ks to the likelihood using the $\hat{p}_i$s. This test statistic is properly called the (generalized) likelihood ratio test statistic but is often simply called the *deviance*. (The likelihood ratio test was also mentioned in the previous section.) The deviance is compared to a $\chi^2(q-2)$ distribution where q is the number of independent binomials and 2 is the number of regression parameters in the logistic model. Unlike the standard errors for the β_is, *all the sample sizes N_i must be large for the lack of fit test to be valid!*

EXAMPLE 8.7.2. Maximum likelihood for the chloracetic acid data gives the following results.

Predictor	$\hat{\beta}_k$	SE($\hat{\beta}_k$)	t	P
Constant	−3.570	0.7040	−5.07	0.000
Dose	14.64	3.326	4.40	0.000

These are similar to the weighted regression results. The deviance of the maximum likelihood fit is 10.254 with $12-2=10$ degrees of freedom for the lack of fit test. The sample sizes are all reasonably large, so a χ^2 test is appropriate. The test statistic is approximately equal to the degrees of freedom, so a test would not be rejected. A simple line seems to fit the data adequately. The maximum likelihood results were obtained using the computer program GLIM. □

We will not analyze more sophisticated count data in this book but we should mention that *both the maximum likelihood methods and the weighted regression methods extend to much more general models*, such as those treated in the remainder of the book. Both methods work when there are many predictors, so we can perform multiple logistic regression which is similar in spirit to multiple regression as treated in Chapters 13, 14, and 15. By modifying the matrix approach to ANOVA problems discussed in Section 16.5, the methods introduced here can be applied to models that are structured like analysis of variance and even analysis of covariance. Christensen (1990b) contains a more complete discussion of logistic regression and logit models. It also contains references to additional work.

8.8 Exercises

EXERCISE 8.8.1. Reiss et al. (1975) and Fienberg (1980) reported that 29 of 52 virgin female undergraduate university students who used a birth control clinic thought that extramarital sex is not always wrong. Give a 99% confidence interval for the population proportion of virgin undergraduate university females who use a birth control clinic and think that extramarital sex is not always wrong.

In addition, 67 of 90 virgin females who did not use the clinic thought that extramarital sex is not always wrong. Give a 99% confidence interval for the difference in proportions between the two groups and give a .05 level test that there is no difference.

EXERCISE 8.8.2. Pauling (1971) reports data on the incidence of colds among French skiers who where given either ascorbic acid or a placebo. Of 139 people given ascorbic acid, 17 developed colds. Of 140 people given the placebo, 31 developed colds. Do these data suggest that the proportion of people who get colds differs depending on whether they are given ascorbic acid?

EXERCISE 8.8.3. Quetelet (1842) and Stigler (1986, p. 175) report data on conviction rates in the French Courts of Assize (Law Courts) from 1825 to 1830. The data are given in Table 8.20. Test whether the conviction rate is the same for each year. Use $\alpha = .05$. (Hint: Table 8.20 is written in a nonstandard form. You need to modify it before applying the methods of this chapter.) If there are differences in conviction rates, use residuals to explore these differences.

TABLE 8.20. French convictions

Year	Convictions	Accusations
1825	4594	7234
1826	4348	6988
1827	4236	6929
1828	4551	7396
1829	4475	7373
1830	4130	6962

EXERCISE 8.8.4. Use the data in Table 8.2 to test whether the probability of a birth in each month is the number of days in the month divided by 365. Thus the null probability for January is 31/365 and the null probability for February is 28/365.

EXERCISE 8.8.5. Snedecor and Cochran (1967) report data from an unpublished report by E. W. Lindstrom. The data concern the results of cross-breeding two types of corn (maize). In 1301 crosses of two types of plants, 773 green, 231 golden, 238 green-golden, and 59 golden-green-striped plants were obtained. If the inheritance of these properties is particularly simple, Mendelian genetics suggests that the probabilities for the four types of corn may be 9/16, 3/16, 3/16, and 1/16, respectively. Test whether these probabilities are appropriate. If they are inappropriate, identify the problem.

EXERCISE 8.8.6. In France in 1827, 6929 people were accused in the courts of assize and 4236 were convicted. In 1828, 7396 people were accused and 4551 were convicted. Give a 95% confidence interval for the proportion of people convicted in 1827. At the .01 level, test the null hypothesis that the conviction rate in 1827 was greater than or equal to 2/3. Does the result of the test depend on the choice of standard error? Give a 95% confidence interval for the difference in conviction rates between the two years. Test the hypothesis of no difference in conviction rates using $\alpha = .05$ and both standard errors.

EXERCISE 8.8.7. Table 8.21 contains additional data from Lazerwitz (1961). These consist of a breakdown of the Protestants in Table 8.6 but with the addition of four more occupational categories. The additional categories are E, semiskilled; F, unskilled; G, farmers; H, no occupation. Analyze the data with an emphasis on partitioning the table.

TABLE 8.21. Occupation and religion

Religion	A	B	C	D	E	F	G	H
White Baptist	43	78	64	135	135	57	86	114
Black Baptist	9	2	9	23	47	77	18	41
Methodist	73	80	80	117	102	58	66	153
Lutheran	23	36	43	59	46	26	49	46
Presbyterian	35	54	38	46	19	22	11	46
Episcopalian	27	27	20	14	7	5	2	15

EXERCISE 8.8.8. Stigler (1986, p. 208) reports data from the *Edinburgh Medical and Surgical Journal* (1817) on the relationship between heights and chest circumferences for Scottish militia men. Measurements were made in inches. We concern ourselves with two groups of men, those with 39 inch chests and those with 40 inch chests. The data are given in Table 8.22. Test whether the distribution of heights is the same for these two groups.

TABLE 8.22. Heights and chest circumferences

	Heights					
Chest	64–65	66–67	68–69	70–71	71–73	Total
39	142	442	341	117	20	1062
40	118	337	436	153	38	1082
Total	260	779	777	270	58	2144

EXERCISE 8.8.9. Use weighted least squares to fit a logistic model to the data of Table 8.20 that relates probability of conviction to year. Is there evidence of a trend in the conviction rates over time? Is there evidence for a lack of fit?

EXERCISE 8.8.10. Is it reasonable to fit a logistic regression to the data of Table 8.22? Why or why not? Explain what such a model would be doing. Whether reasonable or not, fitting such a model can be done. Use weighted least squares to fit a logistic model and discuss the results. Is there evidence for a lack of fit?