1
Introduction

This book is concerned with the analysis of cross-classified categorical data using log-linear models and with logistic regression. Log-linear models have two great advantages: they are flexible and they are interpretable. Log-linear models have all the modeling flexibility that is associated with analysis of variance and regression. They also have natural interpretations in terms of odds and frequently have interpretations in terms of independence. This book also examines logistic regression and logistic discrimination, which typically involve the use of continuous predictor variables. Actually, these are just special cases of log-linear models. There is a wide literature on log-linear models and logistic regression and a number of books have been written on the subject. Some additional references on log-linear models that I can recommend are: Agresti (1984, 1990), Andersen (1991), Bishop, Fienberg, and Holland (1975), Everitt (1977), Fienberg (1980), Haberman (1974a), Plackett (1981), Read and Cressie (1988), and Santner and Duffy (1989). Cox and Snell (1989) and Hosmer and Lemeshow (1989) have written books on logistic regression. One reason I can recommend these is that they are all quite different from each other and from this book. There are differences in level, emphasis, and approach. This is by no means an exhaustive list; other good books are available.

In this chapter we review basic information on conditional independence, random variables, expected values, variances, standard deviations, covariances, and correlations. We also review the distributions most commonly used in the analysis of contingency tables: the binomial, the multinomial, product multinomials, and the Poisson. Christensen (1996a, Chapter 1) contains a more extensive review of most of this material.

1.1 Conditional Probability and Independence

This section introduces two subjects that are fundamental to the analysis of count data. Both subjects are quite elementary, but they are used so extensively that a detailed review is in order. One subject is the definition and use of *odds*. We include as part of this subject the definition and use of *odds ratios*. The other is the use of independence and conditional independence in characterizing probabilities. We begin with a discussion of odds.

Odds will be most familiar to many readers from their use in sporting events. They are not infrequently confused with probabilities. (I once attended an American Statistical Association chapter meeting at which a government publication on the Montana state lottery was disbursed that presented probabilities of winning but called them odds of winning.) In log-linear model analysis and logistic regression, both odds and ratios of odds are used extensively.

Suppose that an event, say, the sun rising tomorrow, has a probability p. The odds of that event are

$$Odds = \frac{p}{1-p} = \frac{\Pr(\text{Event Occurs})}{\Pr(\text{Event Does Not Occur})}.$$

Thus, supposing the probability that the sun will rise tomorrow is .8, the odds that the sun will rise tomorrow are $.8/.2 = 4$. Writing 4 as $4/1$, it might be said that the odds of the sun rising tomorrow are 4 to 1. The fact that the odds are greater than one indicates that the event has a probability of occurring greater than one-half. Conversely, if the odds are less than one, the event has probability of occurring less than one-half. For example, the probability that the sun *will not* rise tomorrow is $1 - .8 = .2$ and the odds that the sun will not rise tomorrow are $.2/.8 = 1/4$.

The larger the odds, the larger the probability. The closer the odds are to zero, the closer the probability is to zero. In fact, for probabilities and odds that are very close to zero, there is essentially no difference between the numbers. As for all lotteries, the probability of winning big in the Montana state lottery was very small. Thus, the mistake alluded to above is of no practical importance. On the other hand, as probabilities get near one, the corresponding odds approach infinity.

Given the odds that an event occurs, the probability of the event is easily obtained. If the odds are O, then the probability p is easily seen to be

$$p = \frac{O}{O+1}.$$

For example, if the odds of breaking your wrist in a severe bicycle accident are .166, the probability of breaking your wrist is $.166/1.166 = .142$ or about $1/7$. Note that even at this level, the numerical values of the odds and the probability are similar.

Examining odds really amounts to a rescaling of the measure of uncertainty. Probabilities between zero and one half correspond to odds between zero and one. Probabilities between one half and one correspond to odds between one and infinity. Another convenient rescaling is the log of the odds. Probabilities between zero and one half correspond to log odds between minus infinity and zero. Probabilities between one half and one correspond to odds between zero and infinity. The log odds scale is symmetric about zero just as probabilities are symmetric about one half. One unit above zero is comparable to one unit below zero. From above, the log odds that the sun will rise tomorrow are $\log(4)$, while the log odds that it will not rise are $\log(1/4) = -\log(4)$. These numbers are equidistant from the center 0. This symmetry of scale fails for the odds. The odds of 4 are three units above the center 1, while the odds of $1/4$ are three-fourths of a unit below the center. For most mathematical purposes, the log odds are a more natural transformation than the odds.

EXAMPLE 1.1.1. *N.F.L. Football*
On January 5, 1990, I decided how much of my meager salary to bet on the upcoming Superbowl. There were eight teams still in contention. *The Albuquerque Journal* reported *Harrah's Odds* for each team. The teams and their odds are given below.

Team	Odds
San Francisco Forty-Niners	even
Denver Broncos	5 to 2
New York Giants	3 to 1
Cleveland Browns	9 to 2
Los Angeles Rams	5 to 1
Minnesota Vikings	6 to 1
Buffalo Bills	8 to 1
Pittsburgh Steelers	10 to 1

These odds were designed for the benefit of Harrah's and were not really anyone's idea of the odds that the various teams would win. (This will become all too clear later.) Nonetheless, we examine these odds as though they determine probabilities for winning the Superbowl as of January 5, 1990, and their implications for my early retirement. The discussion of betting is quite general. I have no particular knowledge of how Harrah's works these things.

The odds on the Vikings are 6 to 1. These are actually the odds that the Vikings *will not* win the Superbowl. The odds are a ratio, $6/1 = 6$. The probabilities are

$$\Pr(\text{Vikings do not win}) = \frac{6}{6+1} = \frac{6}{7}$$

and

$$\Pr(\text{Vikings win}) = \frac{\frac{1}{6}}{\frac{1}{6}+1} = \frac{1}{1+6} = \frac{1}{7}.$$

Similarly, the odds on Denver are 5 to 2 or 5/2. The probabilities are

$$\Pr(\text{Broncos do not win}) = \frac{\frac{5}{2}}{\frac{5}{2}+1} = \frac{5}{5+2} = \frac{5}{7}$$

and

$$\Pr(\text{Broncos win}) = \frac{\frac{2}{5}}{\frac{2}{5}+1} = \frac{2}{5+2} = \frac{2}{7}.$$

San Francisco is even money, so their odds are 1 to 1. The probabilities of winning for all eight teams are given below.

Team	Probability of Winning
San Francisco Forty-Niners	.50
Denver Broncos	.29
New York Giants	.25
Cleveland Browns	.18
Los Angeles Rams	.17
Minnesota Vikings	.14
Buffalo Bills	.11
Pittsburgh Steelers	.09

There is a peculiar thing about these probabilities: They should add up to 1 but do not. One of these eight teams had to win the 1990 Superbowl, so the probability of one of them winning must be 1. The eight events are disjoint, e.g., if the Vikings win, the Broncos cannot, so the sum of the probabilities should be the probability that any of the teams wins. This leads to a contradiction. The probability that any of the teams wins is

$$.50 + .29 + .25 + .18 + .17 + .14 + .11 + .09 = 1.73 \neq 1.$$

All of the odds have been deflated. The probability that the Vikings win should not be .14 but $.14/1.73 = .0809$. The odds against the Vikings should be $(1 - .0809)/.0809 = 11.36$. Rounding this to 11 gives the odds against the Vikings as 11 to 1 instead of the reported 6 to 1. This has severe implications for my early retirement.

The idea behind odds of 6 to 1 is that if I bet \$100 on the Vikings and they win, I should win \$600 and also have my original \$100 returned. Of course, if they lose I am out my \$100. According to the odds calculated above, a fair bet would be for me to win \$1100 on a bet of \$100. (Actually, I should get \$1136 but what is \$36 among friends.) Here, "fair" is used in a

technical sense. In a fair bet, the expected winnings are zero. In this case, my expected winnings for a fair bet are

$$1136(.0809) - 100(1 - .0809) = 0.$$

It is what I win times the probability that I win minus what I lose times the probability that I lose. If the probability of winning is .0809 and I get paid off at a rate of 6 to 1, my expected winnings are

$$600(.0809) - 100(1 - .0809) = -43.4.$$

I don't think I can afford that. In fact, a similar phenomenon occurs for a bet on any of the eight teams. If the probabilities of winning add up to more than one, the true expected winnings on any bet will be negative. Obviously, it pays to make the odds rather than the bets.

Not only odds but ratios of odds arise naturally in the analysis of logistic regression and log-linear models. It is important to develop some familiarity with *odds ratios*. The odds on San Francisco, Los Angeles, and Pittsburgh are 1 to 1, 5 to 1, and 10 to 1, respectively. Equivalently, the odds that each team will not win are 1, 5, and 10. Thus, L.A. has odds of not winning that are 5 times larger than San Francisco's and Pittsburgh's are 10 times larger than San Francisco's. The ratio of the odds of L.A. not winning to the odds of San Francisco not winning is $5/1 = 5$. The ratio of the odds of Pittsburgh not winning to San Francisco not winning is $10/1 = 10$. Also, Pittsburgh has odds of not winning that are twice as large as L.A.'s, i.e., $10/5 = 2$.

An interesting thing about odds ratios is that, say, the ratio of the odds of Pittsburgh not winning to the odds of L.A. not winning is the same as the ratio of the odds of L.A. winning to the odds of Pittsburgh winning. In other words, if Pittsburgh has odds of not winning that are 2 times larger than L.A.'s, L.A. must have odds of winning that are 2 times larger than Pittsburgh's. The odds of L.A. not winning are 5 to 1, so the odds of them winning are 1 to 5 or 1/5. Similarly, the odds of Pittsburgh winning are 1/10. Clearly, L.A. has odds of winning that are 2 times those of Pittsburgh. The odds ratio of L.A. winning to Pittsburgh winning is identical to the odds ratio of Pittsburgh not winning to L.A. not winning. Similarly, San Francisco has odds of winning that are 10 times larger than Pittsburgh's and 5 times as large as L.A.'s.

In logistic regression and log-linear model analysis, one of the most common uses for odds ratios is to observe that they equal one. If the odds ratio is one, the two sets of odds are equal. It is certainly of interest in a comparative study to be able to say that the odds of two things are the same. In this example, none of the odds ratios that can be formed is one because no odds are equal.

Another common use for odds ratios is to observe that two of them are the same. For example, the ratio of the odds of Pittsburgh not winning

relative to the odds of L.A. not winning is the same as the ratio of the odds of L.A. not winning to the odds of the Denver not winning. We have already seen that the first of these values is 2. The odds for L.A. not winning relative to Denver not winning are also 2 because $\frac{5}{1}/\frac{5}{2} = 2$. Even when the corresponding odds are different, odds ratios can be the same.

Marginal and *conditional probabilities* play important roles in logistic regression and log-linear model analysis. If $\Pr(B) > 0$, the conditional probability of A given B is

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

It is the proportion of the probability of B in which A also occurs. To deal with conditional probabilities when $\Pr(B) = 0$ requires much more sophistication. It is an important topic in dealing with continuous observations, but it is not something we need to consider.

If knowing that B occurs does not change your information about A, then A is *independent* of B. Specifically, A is independent of B if

$$\Pr(A|B) = \Pr(A).$$

This definition gets tied up in details related to the requirement that $\Pr(B) > 0$. A simpler and essentially equivalent definition is that A and B are independent if

$$\Pr(A \cap B) = \Pr(A)\Pr(B).$$

EXAMPLE 1.1.2. Table 1.1 contains probabilities for nine combinations of hair and eye color. The nine outcomes are all combinations of three hair colors, Blond (BlH), Brown (BrH), and Red (RH), and three eye colors, Blue (BlE), Brown (BrE), and Green (GE).

Table 1.1
Hair-Eye Color Probabilities

		Eye Color		
		Blue	Brown	Green
Hair Color	Blond	.12	.15	.03
	Brown	.22	.34	.04
	Red	.06	.01	.03

The (*marginal*) probabilities for the various hair colors are obtained by summing over the rows:

$$\begin{aligned} \Pr(\text{BlH}) &= .12 + .15 + .03 = .3 \\ \Pr(\text{BrH}) &= .6 \\ \Pr(\text{RH}) &= .1. \end{aligned}$$

Probabilities for eye colors come from summing the columns. Blue, Brown, and Green eyes have probabilities .4, .5, and .1, respectively. The conditional probability of Blond Hair given Blue Eyes is

$$\begin{aligned} \Pr(\text{BlH}|\text{BlE}) &= \Pr((\text{BlH}, \text{BlE}))/\Pr(\text{BlE}) \\ &= .12/.4 \\ &= .3\,. \end{aligned}$$

Note that $\Pr(\text{BlH}|\text{BlE}) = \Pr(\text{BlH})$, so the events BlH and BlE are independent. In other words, knowing that someone has blue eyes gives no additional information about whether that person has blond hair.

On the other hand,

$$\begin{aligned} \Pr(\text{BrH}|\text{BlE}) &= .22/.4 \\ &= .55\,, \end{aligned}$$

while

$$\Pr(\text{BrH}) = .6,$$

so knowing that someone has blue eyes tells us that they are relatively less likely to have brown hair.

Now condition on blond hair,

$$\Pr(\text{BlE}|\text{BlH}) = .12/.3 = .4 = \Pr(\text{BlE})\,.$$

We again see that BlE and BlH are independent. In fact, it is also true that

$$\Pr(\text{BrE}|\text{BlH}) = \Pr(\text{BrE})$$

and

$$\Pr(\text{GE}|\text{BlH}) = \Pr(\text{GE})\,.$$

Knowing that someone has blond hair gives no additional information about any eye color.

EXAMPLE 1.1.3. Consider the eight combinations of three factors: economic status (High, Low), residence (Montana, Haiti), and beverage of preference (Beer, Other). Probabilities are given below.

	Beer		Other		Total
	Montana	Haiti	Montana	Haiti	
High	.021	.009	.049	.021	.1
Low	.189	.081	.441	.189	.9
Total	.210	.090	.490	.210	1.0

The factors in this table are completely independent. If we condition on either beverage category, then economic status and residence are independent. If we condition on either residence, then economic status and beverage

are independent. If we condition on either economic status, residence and beverage are independent. No matter what you condition on and no matter what you look at, you get independence. For example,

$$\begin{aligned} \Pr(\text{High}|\text{Montana, Beer}) &= .021/.210 \\ &= .1 \\ &= \Pr(\text{High})\,. \end{aligned}$$

Similarly, knowing that someone has low economic status gives no additional information relative to whether their residence is Montana or Haiti.

The phenomenon of complete independence is characterized by the fact that every probability in the table is the product of the three corresponding marginal probabilities. For example,

$$\begin{aligned} \Pr(\text{Low, Montana, Beer}) &= .189 \\ &= (.9)(.7)(.3) \\ &= \Pr(\text{Low})\Pr(\text{Montana})\Pr(\text{Beer})\,. \end{aligned}$$

EXAMPLE 1.1.4. Consider the eight combinations of socioeconomic status (High, Low), political philosophy (Liberal, Conservative), and political affiliation (Democrat, Republican). Probabilities are given below.

	Democrat		Republican		Total
	Liberal	Conservative	Liberal	Conservative	
High	.12	.12	.04	.12	.4
Low	.18	.18	.06	.18	.6
Total	.30	.30	.10	.30	1.0

For any combination in the table, one of the three factors, socioeconomic status, is independent of the other two, political philosophy and political affiliation. For example,

$$\begin{aligned} \Pr(\text{High, Liberal, Republican}) &= .04 \\ &= (.4)(.1) \\ &= \Pr(\text{High})\Pr(\text{Liberal, Republican})\,. \end{aligned}$$

However, the other divisions of the three factors into two groups do not display this property. Political philosophy is not always independent of socioeconomic status and political affiliation, e.g.,

$$\begin{aligned} \Pr(\text{High, Liberal, Republican}) &= .04 \\ &\neq (.4)(.16) \\ &= \Pr(\text{Liberal})\Pr(\text{High, Republican})\,. \end{aligned}$$

Also, political affiliation is not always independent of socioeconomic status and political philosophy, e.g.,

$$\begin{aligned} \Pr(\text{High}, \text{Liberal}, \text{Republican}) &= .04 \\ &\neq (.4)(.16) \\ &= \Pr(\text{Republican})\Pr(\text{High}, \text{Liberal})\,. \end{aligned}$$

EXAMPLE 1.1.5. Consider the twelve outcomes that are all combinations of three factors, one with three levels and two with two levels. The factors and levels are given below. They are similar to those in a study by Reiss et al. (1975) that was reported in Fienberg (1980).

Factor	Levels
Attitude on Extramarital Coitus	Always Wrong, Not Always Wrong
Virginity	Virgin, Nonvirgin
Use of Contraceptives	Regular, Intermittent, None

The probabilities are

	Use of Contraceptives: Regular	
	Virgin	Nonvirgin
Always Wrong	3/50	12/50
Not Always	3/50	12/50
	Intermittent	
	Virgin	Nonvirgin
Always Wrong	1/80	2/80
Not Always	3/80	2/80
	None	
	Virgin	Nonvirgin
Always Wrong	3/40	1/40
Not Always	6/40	2/40

Consider the relationship between attitude and virginity given regular use of contraceptives. The probability of regular use is

$$\begin{aligned} \Pr(\text{Regular}) &= \frac{3}{50} + \frac{3}{50} + \frac{12}{50} + \frac{12}{50} \\ &= 30/50\,. \end{aligned}$$

The conditional probabilities given regular use are computed by dividing the entries in the 2×2 subtable for regular use by the (marginal) probability of regular use, 30/50, e.g., the probability for Always Wrong, Virgin given Regular is $(3/50)/(30/50) = .1$.

Conditional Probabilities Given
Regular Use of Contraceptives

	Virgin	Nonvirgin	Total
Always Wrong	.1	.4	.5
Not Always	.1	.4	.5
Total	.2	.8	1.0

Note that each entry is the product of the row total and the column total, e.g.,

$$\begin{aligned} &\Pr(\text{Always Wrong and Virgin}|\text{Regular}) \\ &= .1 \\ &= (.2)(.5) \\ &= \Pr(\text{Always Wrong}|\text{Regular})\Pr(\text{Virgin}|\text{Regular})\,. \end{aligned}$$

Because this is true for the entire 2×2 table, attitude and virginity are independent given regular use of contraceptives.

Similarly, the conditional probabilities given no use of contraceptives are

	Virgin	Nonvirgin	Total
Always Wrong	3/12	1/12	1/3
Not Always	6/12	2/12	2/3
Total	3/4	1/4	1

Again, it is easily seen that attitude and virginity are independent given no use of contraceptives.

Although we have independence given either no use or regular use, the probabilities of virginity and attitude change drastically. For regular use, nonvirginity is four times more probable than virginity. For no use, virginity is three times more probable. For regular use, attitudes are evenly split. For no use, the attitude that extramarital coitus is not always wrong is twice as probable as the attitude that it is always wrong.

If the conditional probabilities given intermittent use also display independence, we can describe the entire table as having attitude and virginity independent given use. Unfortunately, this does not occur. Conditional on intermittent use, the probabilities are

	Virgin	Nonvirgin	Total
Always Wrong	1/8	1/4	3/8
Not Always	3/8	1/4	5/8
Total	1/2	1/2	1

Virgins are three times as likely to think extramarital coitus is not always wrong, but nonvirgins are evenly split.

Conditional odds are readily obtained from the unconditional probabilities and other conditional probabilities. The odds that a virgin intermittent contraceptive user thinks that extramarital coitus is not always wrong are

$$\begin{aligned}
&\frac{\Pr(\text{Not Always}|\text{Virgin, intermittent use})}{\Pr(\text{Always Wrong}|\text{Virgin, intermittent use})} \\
&\quad = \frac{\Pr(\text{Not Always, Virgin}|\text{intermittent use})}{\Pr(\text{Always Wrong, Virgin}|\text{intermittent use})} \\
&\quad = \frac{\Pr(\text{Not Always, Virgin, intermittent use})}{\Pr(\text{Always Wrong, Virgin, intermittent use})} \\
&\quad = 3\,.
\end{aligned}$$

The reader should verify that all of these probability ratios give 3/1. Similarly, the odds that a nonvirgin intermittent contraceptive user thinks that extramarital coitus is not always wrong is

$$\begin{aligned}
\frac{\Pr(\text{Not Always}|\text{Nonvirgin, intermittent use})}{\Pr(\text{Always Wrong}|\text{Nonvirgin, intermittent use})} &= (1/4)\Big/(1/4) \\
&= (2/80)\Big/(2/80) \\
&= 1\,.
\end{aligned}$$

The odds for virgin intermittent users are different than for nonvirgin intermittent users; thus, independence does not hold. For nonusers, the odds for both virgins and nonvirgins are 2, so independence holds. For regular users, the odds for both virgins and nonvirgins are 1, so again independence holds. Rather than checking for equality of the odds for virgins and nonvirgins, we could look at the ratio of the odds. If the odds ratio is one, then the odds are equal and conditional independence given a particular use holds.

1.2 Random Variables and Expectations

A *random variable* is simply a function from a set of outcomes to the real numbers. A *discrete random variable* is one that takes on values in a countable set. The *distribution* of a discrete random variable is a list of the possible values for the random variable along with the probabilities that the values will occur. The *expected value* of a random variable is a number that characterizes the middle of the distribution. For a random variable y with a discrete distribution, the expected value is

$$\mathrm{E}(y) = \sum_{\text{all } r} r\Pr(y = r)\,.$$

Distributions with the same expected value can be very different. For example, the expected value indicates the middle of a distribution but does not indicate how spread out it is. The *variance* is a measure of how spread out a distribution is from its expected value. Let $\mathrm{E}(y) = \mu$, then the variance of y is

$$\mathrm{Var}(y) = \sum_{\text{all } r} (r - \mu)^2 \mathrm{Pr}(y = r) .$$

One problem with the variance is that it is measured on the wrong scale. If y is measured in meters, $\mathrm{Var}(y)$ involves the terms $(r - \mu)^2$; hence, it is measured in meters squared. To get things back on a comparable scale, we consider the *standard deviation* of y

$$\text{Std. dev.}(y) = \sqrt{\mathrm{Var}(y)} .$$

Standard deviations and variances are useful as measures of the relative dispersions of different random variables. The actual numbers themselves do not mean much. Moreover, there are other equally good measures of dispersion that can give results that are inconsistent with these. One reason standard deviations and variances are so widely used is because they are convenient mathematically. Of particular importance in applied work is the fact that the commonly used normal (Gaussian) distributions are completely characterized by their expected values (means) and variances. With these two numbers, one knows everything about a normal distribution. Normal distributions are widely used in statistics, so variances and their cousins, standard deviations, are also widely used.

The *covariance* is a measure of the linear relationship between two random variables. Suppose y_1 and y_2 are random variables. Let $\mathrm{E}(y_1) = \mu_1$ and $\mathrm{E}(y_2) = \mu_2$. The covariance between y_1 and y_2 is

$$\mathrm{Cov}(y_1, y_2) = \sum_{\text{all } (r,s)} (r - \mu_1)(s - \mu_2) \mathrm{Pr}(y_1 = r, y_2 = s) .$$

It is immediate that

$$\mathrm{Var}(y_1) = \mathrm{Cov}(y_1, y_1) .$$

In an attempt to get a handle on what the numerical value of the covariance means, it is often rescaled into a *correlation coefficient*.

$$\mathrm{Corr}(y_1, y_2) = \mathrm{Cov}(y_1, y_2) \Big/ \sqrt{\mathrm{Var}(y_1)\mathrm{Var}(y_2)} .$$

A perfect increasing linear relationship is indicated by a 1. A perfect decreasing linear relationship gives a -1. The absence of any linear relationship is indicated by a value of 0.

Exercise 1.6.5 contains important results on the expected values, variances, and covariances of linear combination of random variables.

1.3 The Binomial Distribution

There are a few distributions that are used in the vast majority of statistical applications. The reason for this is that they tend to occur naturally. The normal distribution is one. It occurs in practice because the central limit theorem dictates that other distributions will approach the normal. Two other distributions, the binomial and the multinomial, occur in practice because they are so simple. A fourth distribution, the Poisson, also occurs in nature because it is the distribution arrived at in another limit theorem. In this section, we discuss the binomial. Subsequent sections discuss the multinomial and the Poisson.

If you have independent identical trials and are counting how often something (anything) occurs, the appropriate distribution is the binomial. What could be simpler? Typically, the outcome of interest is referred to as a success. If the probability of a success is p in each of N independent identical trials, then the number of successes n has a binomial distribution with parameters N and p. Write

$$n \sim \text{Bin}(N, p)\,.$$

The distribution of n is

$$\Pr(n = r) = \binom{N}{r} p^r (1-p)^{N-r}$$

for $r = 0, 1, \ldots, N$. Here,

$$\binom{N}{r} = \frac{N!}{r!(N-r)!}$$

and for any positive integer m, $m! = m(m-1)(m-2)\cdots(2)(1)$.

Given the distribution, we can find the mean (expected value) and variance. By definition, the mean is

$$\text{E}(n) = \sum_{r=0}^{N} r \binom{N}{r} p^r (1-p)^{N-r}\,.$$

By writing n as the sum of N independent $\text{Bin}(1, p)$ random variables and using Exercise 1.6.5a, it is easily seen that

$$\text{E}(n) = Np\,.$$

The variance of n is

$$\text{Var}(n) = \sum_{r=0}^{N} (r - Np)^2 \binom{N}{r} p^r (1-p)^{N-r}\,.$$

Again, by writing n as the sum of N independent Bin $(1, p)$ random variables and now using Exercise 1.6.5b, it is easily seen that

$$\mathrm{Var}(n) = Np(1-p)\,.$$

In this book, we will often need to look at both the number of successes and the number of failures at the same time. If the number of successes is n_1 and the number of failures is n_2, then

$$\begin{aligned} n_2 &= N - n_1 \\ n_1 &\sim \mathrm{Bin}(N, p) \end{aligned}$$

and

$$n_2 \sim \mathrm{Bin}(N, 1-p)\,.$$

The last result holds because, with independent identical trials, the number of outcomes that we call failures must also have a binomial distribution. If p is the probability of success, the probability of failure is $1-p$. Of course,

$$\begin{aligned} \mathrm{E}(n_2) &= N(1-p) \\ \mathrm{Var}(n_2) &= N(1-p)p\,. \end{aligned}$$

Note that $\mathrm{Var}(n_1) = \mathrm{Var}(n_2)$, regardless of the value of p. Finally,

$$\mathrm{Cov}(n_1, n_2) = -Np(1-p)$$

and

$$\mathrm{Corr}(n_1, n_2) = -1\,.$$

There is a perfect linear relationship between n_1 and n_2. If n_1 goes up one unit, n_2 goes down one unit. When we look at both successes and failures, write

$$(n_1, n_2) \sim \mathrm{Bin}\big(N, p, (1-p)\big)\,.$$

1.4 The Multinomial Distribution

The multinomial distribution is a generalization of the binomial to more than two categories. Suppose we have N independent identical trials. On each trial, we check to see which of q events occurs. In such a situation, we assume that on each trial, one of the q events must occur. Let n_i, $i = 1, \ldots, q$, be the number of times that the ith event occurs. Let p_i be the probability that the ith event occurs on any trial. Note that the p_i's must satisfy $p_1 + p_2 + \cdots + p_q = 1$. In this situation, we say that $(n_1, \ldots, n_q)$ has a multinomial distribution with parameters $N, p_1, \ldots, p_q$. Write

$$(n_1, \ldots, n_q) \sim \mathrm{Mult}(N, p_1, \ldots, p_q)\,.$$

The distribution is

$$\begin{aligned}\Pr(n_1 = r_1, \ldots, n_q = r_q) &= \frac{N!}{r_1! \cdots r_q!} p_1^{r_1} \cdots p_q^{r_q} \\ &= \frac{N!}{\prod_{i=1}^{q} r_i!} \prod_{i=1}^{q} p_i^{r_i}\end{aligned}$$

for $r_i \geq 0$ and $r_1 + \cdots + r_q = N$. Note that if $q = 2$, this is just a binomial distribution. In general, each individual component is

$$n_i \sim \text{Bin}(N, p_i)$$

so

$$\text{E}(n_i) = Np_i$$

and

$$\text{Var}(n_i) = Np_i(1 - p_i)\,.$$

Also, it can be shown that

$$\text{Cov}(n_i, n_j) = -Np_ip_j \qquad \text{for} \qquad i \neq j\,.$$

EXAMPLE 1.4.1. In Example 1.1.4, probabilities were given for the eight categories determined by combining high and low socioeconomic status, liberal and conservative political philosophy, and Democratic and Republican political affiliation. Suppose a sample of 50 individuals was taken from a population that had the probabilities associated with Example 1.1.4,

	Democrat		Republican		
	Liberal	Conservative	Liberal	Conservative	Total
High	.12	.12	.04	.12	.4
Low	.18	.18	.06	.18	.6

The number of individuals falling into each of the eight categories has a multinomial distribution with $N = 50$ and these p_i's. The expected numbers of observations for each category are given by Np_i. It is easily seen that the expected counts for the cells are

	Democrat		Republican	
	Liberal	Conservative	Liberal	Conservative
High	6	6	2	6
Low	9	9	3	9

Note that the expected counts need not be integers.

The variance for, say, the number of high liberal Republicans is $50(.04)(1-.04) = 1.92$. The variance of the number of high liberal Democrats is $50(.12)(1-.12) = 5.28$. The covariance between the number of high liberal Republicans and the number of high liberal Democrats is $-50(.04)(.12) = -.24$. The correlation between the numbers of high liberal Democrats and Republicans is $-.24/\sqrt{(1.92)(5.28)} = -0.075$.

Now, suppose that the 50 individuals fall into the categories as listed in the table below.

	Democrat		Republican	
	Liberal	Conservative	Liberal	Conservative
High	5	7	4	6
Low	8	7	3	10

The probability of getting this particular table is

$$\frac{50!}{5!7!4!6!8!7!3!10!}(.12)^5(.12)^7(.04)^4(.12)^6(.18)^8(.18)^7(.06)^3(.18)^{10} = .000007.$$

The fact that this is a very small number is not surprising. There are a lot of possible tables, so the probability of getting any particular one is small. In fact, the table that has the highest probability can be shown to have a probability of only .000142. Although this probability is also very small, it is more than 20 times larger than the probability of the table given above.

Product-Multinomial Distributions

For $i = 1, \ldots, t$, take independent multinomials where the ith has s_i possible outcomes, i.e.,

$$(n_{i1}, \ldots, n_{is_i}) \sim \text{Mult}(N_i, p_{i1}, \ldots, p_{is_i});$$

then we say that the n_{ij}'s have a product-multinomial distribution. By independence, the probability of any set of outcomes, say $\Pr(n_{ij} = r_{ij} \text{ all } i, j)$, is the product of the multinomial probabilities for each i. In other notation,

$$\Pr(n_{ij} = r_{ij} \text{ all } i, j) = \prod_{i=1}^{t} \Pr(n_{ij} = r_{ij} \text{ all } j)$$

and for $r_{ij} \geq 0$, $j = 1, \ldots, s_i$, with $\sum_{j=1}^{s_i} r_{ij} = N_i$, we have

$$\Pr(n_{ij} = r_{ij} \text{ all } j) = \left(N_i! \Big/ \prod_{j=1}^{s_i} r_{ij}! \right) \prod_{j=1}^{s_i} (p_{ij})^{r_{ij}}.$$

Thus,

$$\Pr(n_{ij} = r_{ij} \text{ all } i,j) = \prod_{i=1}^{t} \left(N_i! \Big/ \prod_{j=1}^{s_i} r_{ij}! \right) \prod_{j=1}^{s_i} (p_{ij})^{r_{ij}} ,$$

where $r_{ij} \geq 0$ all i,j and $r_{i1} + \cdots + r_{is_i} = N_i$ for all i. Expected values, variances, and covariances within a particular multinomial are obtained by ignoring the other multinomials. Covariances between counts in different multinomials are zero because such observations are independent.

EXAMPLE 1.4.2. In Example 1.4.1 we considered taking a sample of 50 people from a population with the probabilities given in Example 1.1.4. Suppose we can identify and sample two subpopulations, the high socioeconomic group and the low socioeconomic group. If we take independent random samples of 30 from the high group and 20 from the low group, the numbers of individuals in the eight categories has a product-multinomial distribution with $t = 2$, $N_1 = 30$, $s_1 = 4$, $N_2 = 20$, and $s_2 = 4$. The probabilities of the four categories associated with high socioeconomic status are the conditional probabilities given high status. For example, the probability of a liberal Republican in the high group is $.04/.4 = .1$; the probability of a liberal Democrat is $.12/.4 = .3$. Similarly, the probability of a liberal Republican in the low socioeconomic group is $.06/.6 = .1$. The table of probabilities appropriate for the product-multinomial sampling described is the table of conditional probabilities given socioeconomic status:

	Democrat		Republican		
	Liberal	Conservative	Liberal	Conservative	Total
High	.3	.3	.1	.3	1.0
Low	.3	.3	.1	.3	1.0

Although the probabilities for each category are the same for both high and low status, this is just an oddity of the particular example under consideration. Typically, the probabilities will be different in the different groups. In fact, the different groups do not even need to be divided into the same categories, although in most of our applications, the categories will be identical for all groups.

The expected counts for cells are computed separately for each multinomial. The expected count for high liberal Republicans is $30(.1) = 3$. With samples of 30 from the high group and 20 from the low group, the expected counts are

	Democrat		Republican		
	Liberal	Conservative	Liberal	Conservative	Total
High	9	9	3	9	30
Low	6	6	2	6	20

Similarly, variances and covariances are found for each multinomial separately. The variance of the number of high liberal Republicans is $30(.1)(1-.1) = 2.7$. The covariance between the numbers of low liberal Democrats and low liberal Republicans is $-20(.3)(.1) = -0.6$. The covariance between counts in different multinomials is zero because counts in different multinomials are independent, e.g., the covariance between the numbers of high liberal Democrats and low liberal Republicans is zero because all high status counts are independent of all low status counts.

To find the probability of any particular table, find the probability associated with the high group and multiply it by the probability of the low group. The probability of getting the table

	Democrat		Republican		
	Liberal	Conservative	Liberal	Conservative	Total
High	10	10	2	8	30
Low	5	8	1	6	20

is

$$\left[\frac{30!}{10!10!2!8!}(.3)^{10}(.3)^{10}(.1)^2(.3)^8\right]\left[\frac{20!}{5!8!1!6!}(.3)^5(.3)^8(.1)^1(.3)^6\right]$$
$$= (.045716)(.008117) = .000371\,.$$

EXERCISE 1.1. Find the expected counts for a sample of size 20 from the population with probabilities given in Example 1.1.3. Now, conditioning on residence, find the expected counts for a sample of size 8 from Montana and a sample of size 12 from Haiti.

1.5 The Poisson Distribution

The binomial and multinomial distributions are appropriate and useful when the number of trials are not too large (whatever that means) and the probabilities of occurrences are not too small. For phenomena that have a very small probability of occurring on any particular trial, but for which an extremely large number of trials are available, the Poisson distribution is appropriate. For example, the number of suicides in a year might have a Poisson distribution. The probability of anyone committing suicide is very small, but in a large population, a substantial number of people will do it.

One of the most famous examples of a Poisson distribution is due to Bortkiewicz (1898). He examines the yearly total of men in the Prussian army corps who were kicked by horses and died of their injuries. Again, the probability that any individual will be mortally hoofed on a given day

is very small, but for an entire army corps over an entire year, the number is fairly substantial. In particular, Fisher (1925) cites the 200 observations from 10 corps over a 20-year period as:

Deaths	0	1	2	3	4	5+
Frequencies	109	65	22	3	1	0

The idea is to view these as the results of a random sample of size 200 from a Poisson distribution. (Incidentally, the identity of the individual who introduced this example is one of the compelling mysteries in the history of statistics. It has been ascribed to at least four different people: Bortkiewicz, Bortkewicz, Bortkewitsch, and Bortkewitch.)

A third example of Poisson sampling is the number of microorganisms in a solution. One can imagine dividing the solution into a very large number of hypothetical units with very small volume (i.e., just big enough for the microorganism to be contained in the unit). If microorganisms are rare in the solution, then the probability of getting an organism in any particular unit is very small. Now, if we extract say one cubic centimeter of the solution, we have a very large number of trials. The number of organisms in the extracted solution should follow a Poisson distribution. Note that the Poisson distribution depends on having relatively few organisms in the solution. If that assumption is not true, one can dilute the solution until it is true.

Finally, and perhaps most importantly, the number of people who arrive during a 5-minute period to buy tickets for a Bruce Springsteen concert can be modeled with a Poisson distribution. Time can be divided into arbitrarily small intervals. The probability that anyone in the population will show up during any particular interval is very small. However, in 5 minutes there are a very large number of intervals.

The Poisson distribution can be arrived at as the limit of a $\text{Bin}(N, p)$ distribution where $N \to \infty$ and $p \to 0$. However, the two convergences must occur in such a way that $Np \to \lambda$. (To do this rigorously, we would let p be a function of N, say p_N.) The value λ is the parameter of the Poisson distribution. If n is a random variable with a Poisson distribution and parameter λ, write

$$n \sim \text{Pois}(\lambda)\,.$$

The distribution is defined by giving the probabilities and outcomes, i.e.,

$$\Pr(n = r) = \lambda^r e^{-\lambda}/r! \tag{1}$$

for $r = 0, 1, \ldots$.

It is not difficult to arrive at (1) by looking at binomial probabilities.

The corresponding binomial probability for $n = r$ is

$$\binom{N}{r} p^r (1-p)^{N-r} = [(Np)^r (1-p)^N / r!](1-p)^{-r} \frac{N!}{(N-r)! N^r} . \qquad (2)$$

With $N \to \infty$, $p \to 0$, and $Np \to \lambda$,

$$\begin{aligned} (Np)^r &\to \lambda^r \\ (1-p)^N &\to e^{-\lambda} \\ (1-p)^{-r} &\to 1 \\ N!/(N-r)!N^r &\to 1 \end{aligned}$$

Substituting these limits into the right-hand side of (2) gives the probability displayed in (1).

Using (1), we can compute the expected value and the variance of n. It is not difficult to show that

$$\mathrm{E}(n) = \lambda$$

and

$$\mathrm{Var}(n) = \lambda .$$

Lindgren (1993) gives a more detailed discussion of the assumptions behind the Poisson model. Fisher (1925) gives a nice discussion of the uses of the Poisson model and the analysis of Poisson data.

We close with two facts about independent Poisson random variables. If $n_1, \ldots, n_q$ are independent with $n_i \sim \mathrm{Pois}(\lambda_i)$, then the total of all the counts is

$$n_1 + n_2 + \cdots + n_q \sim \mathrm{Pois}(\lambda_1 + \cdots + \lambda_q)$$

and the counts given the total are

$$(n_1, \ldots, n_q) | N \sim \mathrm{Mult}(N, p_1, \ldots, p_q)$$

where

$$N = n_1 + \cdots + n_q$$

and

$$p_i = \lambda_i / (\lambda_1 + \cdots + \lambda_q), \quad i = 1, \ldots, q .$$

The conditional distribution is important for the analysis of log-linear models. If we have a table of counts that is comprised of independent Poisson random variables, we can always compute the grand total for the table. Looking at the conditional distribution given the total leads us to an analysis based on a multinomial distribution. The multinomial is the most commonly assumed distribution for tables of counts. Our discussion will focus almost entirely on multinomial and product-multinomial sampling.

1.6 Exercises

EXERCISE 1.6.1. In a *Newsweek* article on "The Wisdom of Animals" (May 23, 1988), one of the key issues considered was whether animals (other than humans) understand relationships between symbols. Some animals can associate symbols with objects; the question is whether they can tell the difference between commands such as "take the red ball to the blue ball" and "take the blue ball to the red ball." In discussing sea lions, it was indicated that out of a large pool of objects, they correctly identify symbols 95% of the time but are only correct 45% of the time on relationships. Presumably, this referred to a simple relationship between two objects; for example, a sea lion could be shown symbols for "blue ball," "take to," "red ball." It was then concluded that, "considering the number of objects present in the pool, the odds are exceedingly long of getting even that proportion [45%] right by sheer chance." Assume a simple model in which sea lions know the nature of the relationship (it is repeated in a long series of trials), e.g., take one object to another, but make independent choices for identifying each object and the order in the relationship. Assume also that they have no idea what the correct order should be in the relationship, i.e., the two possible orders are equally probable. Compute the probability a sea lion will perform the task correctly. Why is the conclusion given in the article wrong? What does the number of objects present in the pool have to do with all this?

EXERCISE 1.6.2. Consider a 2×2 table of multinomial probabilities that models how subjects respond on two separate occasions.

	First Trial	
Second Trial	A	B
A	p_{11}	p_{12}
B	p_{21}	p_{22}

Show that

$$\Pr(A \text{ Second Trial } | B \text{ First Trial }) = \Pr(B \text{ Second Trial } | A \text{ First Trial })$$

if and only if the event that a change occurs between the first and second trials is independent of the outcome on the first trial.

EXERCISE 1.6.3. Weisberg (1975) reports the following data on the number of boys among the first seven children born to a collection of 1,334 Swedish ministers.

Number of Boys	0	1	2	3	4	5	6	7
Frequency	6	57	206	362	365	256	69	13

Assume that the number of boys has a Bin(7, .5) distribution. Compute the probabilities for each of the eight categories $0, 1, \ldots, 7$. From the sample

of 1,334 families, what is the expected frequency for each category? What is the distribution of the number of families that fall into each category? Summarize the fit of the assumed binomial model by computing

$$X^2 = \sum_{i=0}^{7} \frac{(Observation_i - Expected_i)^2}{Expected_i}.$$

The statistic X^2 is known as Pearson's chi-square statistic. For large samples such as this, if the *Expected* values are correct, X^2 should be one observation from a $\chi^2(7)$ distribution. (The 7 is one less than the number of categories.) Compute X^2 and compare it to tabled values of the $\chi^2(7)$ distribution. Does X^2 seem like it could reasonably come from a $\chi^2(7)$? What does this imply about how well the binomial model fits the data? Can you distinguish which assumptions made in the binomial model may be violated?

EXERCISE 1.6.4. The data given in the previous problem may be 1,334 independent observations from a $\text{Bin}(7, p)$ distribution. If so, use the defining assumptions of the binomial distribution to show that this is the same as one observation from a $\text{Bin}(1334 \times 7, p)$ distribution. Estimate p with

$$\hat{p} = \frac{\text{Total number of boys}}{\text{Total number of trials}}.$$

Repeat the previous problem, replacing .5 with $\hat{p}$. Compare X^2 to a $\chi^2(6)$ distribution, reducing the degrees of freedom by one because the probability p is being estimated from the data.

EXERCISE 1.6.5. Let y_1, y_2, y_3, and y_4 be random variables and let a_1, a_2, a_3, and a_4 be real numbers. Show that the following relationships hold for finite discrete distributions.

(a) $\text{E}(a_1y_1 + a_2y_2 + a_3) = a_1\text{E}(y_1) + a_2\text{E}(y_2) + a_3$.

(b) $\text{Var}(a_1y_1 + a_2y_2 + a_3) = a_1^2\text{Var}(y_1) + a_2^2\text{Var}(y_2)$ for y_1 and y_2 independent.

(c) $\text{Cov}(a_1y_1 + a_2y_2, a_3y_3 + a_4y_4) = \sum_{i=1}^{2}\sum_{j=3}^{4} a_ia_j\text{Cov}(y_i, y_j)$.

EXERCISE 1.6.6. Assume that

$$(n_1, \ldots, n_q) \sim \text{Mult}(N, p_1, \ldots, p_q)$$

and let t be an integer less than q. Define $y = n_1 + \cdots + n_t$ and $\tilde{p} = p_1 + \cdots + p_t$. Show that

$$(y, n_{t+1}, \ldots, n_q) \sim \text{Mult}(N, \tilde{p}, p_{t+1}, \ldots, p_q)$$

so that

$$\mathrm{E}(y) = N\tilde{p}$$

and

$$\mathrm{Var}(y) = N\tilde{p}(1-\tilde{p}).$$

EXERCISE 1.6.7. Suppose $y \sim \mathrm{Bin}(N, p)$. Let $\hat{p} = y/N$. Show that $\mathrm{E}(\hat{p}) = p$ and that $\mathrm{Var}(\hat{p}) = p(1-p)/N$.

2

Two-Dimensional Tables and Simple Logistic Regression

At this point, it is not our primary intention to provide a rigorous account of logistic regression and log-linear model theory. Such a treatment demands extensive use of advanced calculus and asymptotic theory. On the other hand, some knowledge of the basic issues is necessary for a correct understanding of *applications of logistic regression and log-linear models*. In this chapter, we address these basic issues for the simple case of two-dimensional tables and simple logistic regression. For a more elementary discussion of two-dimensional tables and simple logistic regression including substantial data analysis, see Christensen (1996a, Chapter 8). In fact, we assume that the reader is familiar with such analyses and use the topics in this chapter primarily to introduce key theoretical ideas.

2.1 Two Independent Binomials

Consider two binomials arranged in a 2×2 table. Our interest is in examining possible differences between the two binomials.

EXAMPLE 2.1.1. A survey was conducted to examine the relative attitudes of males and females about abortion. Of 500 females, 309 supported legalized abortion. Of 600 males, 319 supported legalized abortion. The data can be summarized in tabular form:

OBSERVED VALUES

	Support	Do Not Support	Total
Female	309	191	500
Male	319	281	600
Total	628	472	1,100

Note that the totals on the right-hand side of the table (500 and 600) are fixed by the design of the study. The totals along the bottom of the table are observed random variables. It is assumed that for each sex, the numbers of supporters and nonsupporters form a binomial random vector (ordered pair). We are interested in whether these numbers indicate that a person's sex affects their attitude toward legalized abortion. Note that the categories are Support and Do Not Support legalized abortion. Not supporting legalized abortion is distinct from opposing it. Anyone who is indifferent neither supports nor opposes legalized abortion.

We now introduce the notation that will be used for tables of counts in this book. For a 2×2 table, the observed values are denoted by n_{ij}, $i = 1, 2$ and $j = 1, 2$. Marginal totals are written $n_{i\cdot} \equiv n_{i1} + n_{i2}$ and $n_{\cdot j} \equiv n_{1j} + n_{2j}$. The total of all observations is $n_{\cdot\cdot} \equiv n_{11} + n_{12} + n_{21} + n_{22}$. The probability of having an observation fall in the ith row and jth column of the table is denoted p_{ij}. The number of observations that one would expect to see in the ith row and jth column (based on some statistical model) is denoted m_{ij}. For independent binomial rows, $m_{ij} = n_{i\cdot} p_{ij}$. Marginal totals $p_{i\cdot}$, $p_{\cdot j}$, $m_{i\cdot}$, and $m_{\cdot j}$ are defined like $n_{i\cdot}$ and $n_{\cdot j}$.

All of this notation can be summarized in tabular form.

OBSERVED VALUES

		Columns 1	2	Totals
Rows	1	n_{11}	n_{12}	$n_{1\cdot}$
	2	n_{21}	n_{22}	$n_{2\cdot}$
	Totals	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

PROBABILITIES

		Columns 1	2	Totals
Rows	1	p_{11}	p_{12}	$p_{1\cdot}$
	2	p_{21}	p_{22}	$p_{2\cdot}$
	Totals	$p_{\cdot 1}$	$p_{\cdot 2}$	$p_{\cdot\cdot}$

EXPECTED VALUES

		Columns 1	2	Totals
Rows	1	m_{11}	m_{12}	$m_{1\cdot}$
	2	m_{21}	m_{22}	$m_{2\cdot}$
	Totals	$m_{\cdot 1}$	$m_{\cdot 2}$	$m_{\cdot\cdot}$

Our interest is in finding estimates of the p_{ij}'s, developing models for the p_{ij}'s, and performing tests on the p_{ij}'s. Equivalently, we can concern ourselves with estimates, models, and tests for the m_{ij}'s.

In Example 2.1.1, our interest is in whether sex is related to support for legalized abortion. Note that $p_{11}+p_{12}=1$ and $p_{21}+p_{22}=1$. (Equivalently $m_{11}+m_{12}=500$ and $m_{21}+m_{22}=600$.) If sex has no effect on opinion, the distribution of support versus nonsupport should be the same for both sexes. In particular, it is of interest to test the null hypothesis (model)

$$H_0 : p_{11} = p_{21} \text{ and } p_{12} = p_{22}\,.$$

With $p_{i1}+p_{i2}=1$, the equality $p_{11}=p_{21}$ holds if and only if $p_{12}=p_{22}$ holds. In other words, females and males have the same probability of "support" if and only if they have the same probability for "do not support." It suffices to test that the probability of support is the same for both sexes, i.e.,

$$H_0 : p_{11} = p_{21}$$

or, equivalently,

$$H_0 : p_{11} - p_{21} = 0\,.$$

To test this hypothesis, we need an estimate of $p_{11}-p_{21}$ and the standard error (SE) of the estimate. Each row is binomial with sample size $n_{i\cdot}$, so a natural estimate of p_{ij} is the proportion of observations falling in cell ij relative to the total number of observations in the ith row, i.e.,

$$\hat{p}_{ij} = n_{ij}/n_{i\cdot}\,.$$

For the abortion example, $\hat{p}_{11}=309/500$ and $\hat{p}_{21}=319/600$. The estimate of $p_{11}-p_{21}$ is

$$\hat{p}_{11}-\hat{p}_{21} = (n_{11}/n_{1\cdot}) - (n_{21}/n_{2\cdot})\,.$$

The two rows of the table were sampled independently so the variance of $\hat{p}_{11}-\hat{p}_{21}$ is

$$\begin{aligned}\mathrm{Var}(\hat{p}_{11}-\hat{p}_{21}) &= \mathrm{Var}(\hat{p}_{11}) + \mathrm{Var}(\hat{p}_{21}) \\ &= p_{11}p_{12}/n_{1\cdot} + p_{21}p_{22}/n_{2\cdot}\,,\end{aligned}$$

cf. Exercise 1.6.7. Finally,

$$\mathrm{SE}(\hat{p}_{11}-\hat{p}_{21}) = \sqrt{\hat{p}_{11}\hat{p}_{12}/n_{1\cdot} + \hat{p}_{21}\hat{p}_{22}/n_{2\cdot}}\,.$$

For the abortion example,

$$\mathrm{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{(309/500)(191/500)}{500} + \frac{(319/600)(281/600)}{600}} = .0298\,.$$

One other thing is required before we can perform a test. We need to know the distribution of $[(\hat{p}_{11} - \hat{p}_{21}) - (p_{11} - p_{21})]/\mathrm{SE}(\hat{p}_{11} - \hat{p}_{21})$. By appealing to the Central Limit Theorem and the Law of Large Numbers (cf. Lindgren, 1993), if $n_{1\cdot}$ and $n_{2\cdot}$ are large, we can use the approximate distribution

$$\frac{(\hat{p}_{11} - \hat{p}_{21}) - (p_{11} - p_{21})}{\mathrm{SE}(\hat{p}_{11} - \hat{p}_{21})} \sim N(0,1)\,.$$

To perform a test of

$$H_0 : p_{11} - p_{21} = 0$$

versus

$$H_A : p_{11} - p_{21} \neq 0,$$

assume that H_0 is true and look for evidence against H_0. If H_0 is true, the approximate distribution is

$$\frac{(\hat{p}_{11} - \hat{p}_{21}) - 0}{\mathrm{SE}(\hat{p}_{11} - \hat{p}_{21})} \sim N(0,1)\,.$$

If the alternative hypothesis is true, $\hat{p}_{11} - \hat{p}_{21}$ still estimates $p_{11} - p_{21}$ so the test statistic

$$\frac{(\hat{p}_{11} - \hat{p}_{21}) - 0}{\mathrm{SE}(\hat{p}_{11} - \hat{p}_{21})}$$

tends to be either a large positive value if $p_{11} - p_{21} > 0$ or a large negative value if $p_{11} - p_{21} < 0$. An $\alpha = .05$ level test rejects H_0 if

$$\frac{(\hat{p}_{11} - \hat{p}_{21}) - 0}{\mathrm{SE}(\hat{p}_{11} - \hat{p}_{21})} > 1.96$$

or if

$$\frac{(\hat{p}_{11} - \hat{p}_{21}) - 0}{\mathrm{SE}(\hat{p}_{11} - \hat{p}_{21})} < -1.96\,.$$

The values -1.96 and 1.96 cut off the probability .025 from the bottom and top of a $N(0,1)$ distribution, respectively. Thus, the total probability of rejecting H_0 when H_0 is true is $.025 + .025 = .05$. Recall that this test is based on a large sample approximation to the distribution of the test statistic.

For the abortion example, the test statistic is

$$\frac{(309/500) - (319/600)}{.0298} = 2.90\,.$$

Because $2.90 > 1.96$, H_0 is rejected at the $\alpha = .05$ level. There is evidence of a relationship between sex and attitudes about legalized abortion. These data indicate that females are more likely to support legalized abortion.

Before leaving this test procedure, we mention an alternative method for computing $\mathrm{SE}(\hat{p}_{11} - \hat{p}_{21})$. Recall that

$$\mathrm{Var}(\hat{p}_{11} - \hat{p}_{21}) = p_{11}p_{12}/n_{1\cdot} + p_{21}p_{22}/n_{2\cdot}\,.$$

If H_0 is true, $p_{11} = p_{21}$ and $p_{12} = p_{22}$. These facts can be used in estimating the variance of $\hat{p}_{11} - \hat{p}_{21}$. A pooled estimate of $p \equiv p_{11} = p_{21}$ is $(n_{11} + n_{21})/(n_{1\cdot} + n_{2\cdot}) = n_{\cdot 1}/n_{\cdot\cdot} = 628/1100$. A pooled estimate of $(1-p) \equiv p_{12} = p_{22}$ is $n_{\cdot 2}/n_{\cdot\cdot} = 472/1100$. This yields

$$\mathrm{Var}(\hat{p}_{11} - \hat{p}_{21}) = p(1-p)(1/n_{1\cdot} + 1/n_{2\cdot})$$

and

$$\begin{aligned}\mathrm{SE}(\hat{p}_{11} - \hat{p}_{21}) &= \sqrt{(628/1100)(472/1100)[(1/500) + (1/600)]} \\ &= .0300\,.\end{aligned}$$

The test statistic computed with the new standard error is

$$\frac{(309/500) - (319/600)}{.0300} = 2.87777\,.$$

For these data, the results are essentially the same.

The test procedures discussed above work nicely for two independent binomials, but, unfortunately, they do not generalize to more than two binomials or to situations in which there are more than two possible outcomes (e.g., support, oppose, no opinion). An alternative test procedure is based on what is known as the *Pearson chi-square test statistic*. This test is equivalent to the test given above using the pooled estimate of the standard error. Moreover, Pearson's chi-square is applicable in more general problems. The Pearson test statistic is based on comparing the observed table values in the 2×2 table with estimates of the expected values that are obtained assuming that H_0 is true.

In the abortion example, if H_0 is true, then $p = p_{11} = p_{21}$ and $\hat{p} = 628/1100$. Similarly, $(1-p) = p_{12} = p_{22}$ and $(1-\hat{p}) = 472/1100$. As before, the expected values are $m_{ij} = n_{i\cdot}p_{ij}$. The estimated expected values under H_0 are $\hat{m}_{ij}^{(0)} = n_{i\cdot}\hat{p}_{ij}$, where $\hat{p}_{ij}$ is $\hat{p}$ if $j = 1$ and $(1-\hat{p})$ if $j = 2$. More generally,

$$\hat{m}_{ij}^{(0)} = n_{i\cdot}(n_{\cdot j}/n_{\cdot\cdot})\,. \tag{1}$$

The Pearson chi-square statistic is defined as

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(n_{ij} - \hat{m}_{ij}^{(0)}\right)^2}{\hat{m}_{ij}^{(0)}}\,.$$

If H_0 is true, then n_{ij} and $\hat{m}_{ij}^{(0)}$ should be near each other, and the terms $(n_{ij} - \hat{m}_{ij}^{(0)})^2$ should be reasonably small. If H_0 is not true, then the $\hat{m}_{ij}^{(0)}$'s, which are estimates based on the assumption that H_0 is true, should do a poor job of predicting the n_{ij}'s. The terms $(n_{ij} - \hat{m}_{ij}^{(0)})^2$ should be larger when H_0 is not true.

Note that a prediction $\hat{m}_{ij}^{(0)}$ that is, say, three away from the observed value n_{ij}, can be either a good prediction or a bad prediction depending on how large the value in the cell should be. If $n_{ij} = 4$ and $\hat{m}_{ij}^{(0)} = 1$, the prediction is poor. If $n_{ij} = 104$ and $\hat{m}_{ij}^{(0)} = 101$, the prediction is good. The $\hat{m}_{ij}^{(0)}$ in the denominator of each term of X^2 is a scale factor that corrects for this problem.

The hypothesis $H_0 : p_{11} = p_{21}$ and $p_{12} = p_{22}$ is rejected at the $\alpha = .05$ level if

$$X^2 \geq \chi^2(.95, 1)\,.$$

The test is based on the fact that if H_0 is true, then as $n_{1\cdot}$ and $n_{2\cdot}$ get large, X^2 has approximately a $\chi^2(1)$ distribution. This is a consequence of the Central Limit Theorem and the Law of Large Numbers, cf. Exercise 2.1.

For the abortion example

$\hat{m}_{ij}^{(0)}$	Support	Do Not Support	Totals
Female	285.5	214.5	500
Male	342.5	257.5	600
Totals	628	472	1100

$$X^2 = 8.2816,$$

$$\chi^2(.95, 1) = 3.84\,.$$

Because $8.2816 > 3.84$, the $\alpha = .05$ test rejects H_0.

Note that $8.2816 = (2.8777)^2$ and that $3.84 = (1.96)^2$. For 2×2 tables, the results of Pearson chi-square tests are exactly equivalent to the results of normal theory tests using the pooled estimate in the standard error. By definition, $\chi^2(1-\alpha, 1) = [z(1-\frac{\alpha}{2})]^2$ for $\alpha \in (0, .5]$. Also,

$$X^2 = \frac{(\hat{p}_{11} - \hat{p}_{21})^2}{\hat{p}(1-\hat{p})\left(\frac{1}{n_{1\cdot}} + \frac{1}{n_{2\cdot}}\right)}\,. \tag{2}$$

EXERCISE 2.1. Prove equation (2).

By comparing the n_{ij}'s to the $\hat{m}_{ij}^{(0)}$'s, we can examine the nature of the differences in the two binomials. One simple way to do this comparison is to examine a table of *residuals*, i.e., the $(n_{ij} - \hat{m}_{ij}^{(0)})$'s. In order to make

accurate evaluations of how well $\hat{m}_{ij}^{(0)}$ is predicting n_{ij}, the residuals need to be rescaled or standardized. Define the *Pearson residuals* as

$$\tilde{r}_{ij} = \frac{n_{ij} - \hat{m}_{ij}^{(0)}}{\sqrt{\hat{m}_{ij}^{(0)}}},$$

$i = 1, 2$, $j = 1, 2$. Note that $X^2 = \sum_{ij} \tilde{r}_{ij}^2$. The Pearson residuals for the abortion data are

$\tilde{r}_{ij}$	Support	Do Not Support
Female	1.39	−1.60
Male	−1.27	1.46

The positive residual 1.39 indicates that more females support legalized abortion than would be expected under H_0. The negative residual -1.27 indicates that fewer males support abortion than would be expected under H_0. Together, the values 1.39 and -1.27 indicate that proportionately more females support legalized abortion than males. Equivalently, proportionately more males do not support legalized abortion than females.

2.1.1 The Odds Ratio

A commonly used technique in the analysis of count data is the examination of odds ratios. In the abortion example, the odds of females supporting legalized abortion is p_{11}/p_{12}. The odds of males supporting legalized abortion is p_{21}/p_{22}. The odds ratio is

$$\frac{(p_{11}/p_{12})}{(p_{21}/p_{22})} = \frac{p_{11}p_{22}}{p_{12}p_{21}}.$$

Note that if the two binomials are identical, then $p_{11} = p_{21}$ and $p_{12} = p_{22}$, so

$$\frac{p_{11}p_{22}}{p_{12}p_{21}} = 1.$$

An alternative to using Pearson's chi-square for examining whether two binomials are the same is to examine the estimated odds ratio. Using $\hat{p}_{ij} = n_{ij}/n_{i\cdot}$ gives

$$\frac{\hat{p}_{11}\hat{p}_{22}}{\hat{p}_{12}\hat{p}_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

For the abortion example, the estimate is

$$\frac{(309)(281)}{(191)(319)} = 1.425.$$

This is only an estimate of the population odds ratio, but it is fairly far from the target value of 1. In particular, we have estimated that the odds

of a female supporting legalized abortion are about one and a half times as large as the odds of a male supporting legalized abortion.

We may wish to test the hypothesis that the odds ratio equals 1. Equivalently, we can test whether the log of the odds ratio equals 0. The log odds ratio is $\log(1.425) = .354$. The asymptotic (large sample) standard error of the log odds ratio is

$$\begin{aligned}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}} &= \sqrt{\frac{1}{309}+\frac{1}{191}+\frac{1}{319}+\frac{1}{281}} \\ &= .123\,.\end{aligned}$$

The estimate minus the hypothesized value over the standard error is

$$\frac{.354-0}{.123} = 2.88\,.$$

Comparing this to a $N(0,1)$ distribution indicates that the log-odds ratio is greater than zero, thus the odds ratio is greater than 1. Note that this test is not equivalent to the other tests considered, even though the numerical value of the test statistic is similar to the other normal theory tests.

A 95% confidence interval for the log odds ratio has the end points $.354 \pm 1.96(.123)$. This gives an interval $(.113, .595)$. The log odds ratio is in the interval $(.113, .595)$ if and only if the odds ratio is in the interval $(e^{.113}, e^{.595})$. Thus, a 95% confidence interval for the odds ratio is $(e^{.113}, e^{.595})$, which simplifies to $(1.12, 1.81)$. We are 95% confident that the true odds of women supporting legalized abortion is between 1.12 and 1.81 times greater than the odds of men supporting legalized abortion.

2.2 Testing Independence in a 2 × 2 Table

In Section 1, we obtained a 2×2 table by looking at two populations, each divided into two categories. In this section, we consider only one population but divide it into two categories in each of two different ways. The two different ways of dividing the population will be referred to as factors.

In Section 1, we examined differences between the two populations. In this section, we examine the nature of the one population being sampled. In particular, we examine whether the two factors affect the population independently or whether they interact to determine the nature of the population.

EXAMPLE 2.2.1. As part of a longitudinal study, a sample of 3182 people without cardiovascular disease were cross-classified by two factors: personality type and exercise. Personality type was categorized as type A or type B. Type A persons show signs of stress, uneasiness, and hyperactivity. Type

B persons are relaxed, easygoing, and normally active. Exercise is categorized as persons who exercise regularly and those who do not. The data are given in the following table:

		Personality		
	n_{ij}	A	B	Totals
Exercise	Regular	483	477	960
	Other	1101	1121	2222
	Totals	1584	1598	3182

Although notations for observations (n_{ij}'s), probabilities (p_{ij}'s), and expected values (m_{ij}'s) are identical to those in Section 1, the meaning of these quantities has changed. In Section 1, the rows were two independent binomials, so $p_{11}+p_{12} = 1 = p_{21}+p_{22}$. In this section, there is only one population, so the constraint on the probabilities is that $p_{11}+p_{12}+p_{21}+p_{22} = 1$.

In this section, our primary interest is in determining whether the row factor is independent of the column factor and if not, how the factors deviate from independence. The probability of an observation falling in the ith row and jth column of the table is p_{ij}. The probability of an observation falling in the ith row is $p_{i\cdot}$. The probability of the jth column is $p_{\cdot j}$. Rows and columns are independent if and only if for all i and j

$$p_{ij} = p_{i\cdot}p_{\cdot j}\,. \tag{1}$$

The sample size is $n_{\cdot\cdot}$, so the expected counts in the table are

$$m_{ij} = n_{\cdot\cdot}p_{ij}\,.$$

If rows and columns are independent, this becomes

$$m_{ij} = n_{\cdot\cdot}p_{i\cdot}p_{\cdot j}\,. \tag{2}$$

It is easily seen that condition (1) for independence is equivalent to

$$m_{ij} = m_{i\cdot}m_{\cdot j}/n_{\cdot\cdot}\,. \tag{3}$$

Pearson's chi-square can be used to test independence. Pearson's statistic is again

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(n_{ij} - \hat{m}_{ij}^{(0)}\right)^2}{\hat{m}_{ij}^{(0)}}$$

where $\hat{m}_{ij}^{(0)}$ is an estimate of m_{ij} based on the assumption that rows and columns are independent. If we take $\hat{m}_{i\cdot} = n_{i\cdot}$ and $\hat{m}_{\cdot j} = n_{\cdot j}$, then equation (3) leads to

$$\hat{m}_{ij}^{(0)} = n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}\,. \tag{4}$$

Equation (4) can also be arrived at via equation (2). An obvious estimate of $p_{i\cdot}$ is

$$\hat{p}_{i\cdot} = n_{i\cdot}/n_{\cdot\cdot}.$$

Similarly,

$$\hat{p}_{\cdot j} = n_{\cdot j}/n_{\cdot\cdot}.$$

Substitution into equation (2) leads to equation (4). It is interesting to note that equation (4) is numerically identical to formula (2.1.1), which gives $\hat{m}_{ij}$ for two independent binomials. Just as in Section 1, for the purposes of testing, X^2 is compared to a $\chi^2(1)$ distribution. Pearson residuals are again defined as

$$\tilde{r}_{ij} = \frac{n_{ij} - \hat{m}_{ij}^{(0)}}{\sqrt{\hat{m}_{ij}^{(0)}}}.$$

For the personality-exercise data, we get

		Personality		
	$\hat{m}_{ij}^{(0)}$	A	B	Totals
Exercise	Regular	477.9	482.1	960
	Other	1106.1	1115.9	2222
	Totals	1584	1598	3182

$$X^2 = .156$$

The test is not significant for any reasonable α level. (The *P value* is greater than .5.) There is no significant evidence against independence of exercise and personality type. In other words, the data are consistent with the interpretation that knowledge of personality type gives no new information about exercise habits or, equivalently, knowledge of exercise habits gives no new information about personality type.

2.2.1 The Odds Ratio

Just as in examining the equality of two binomials, the odds ratio can be used to examine the independence of two factors in a multinomial sample. In the personality-exercise data, the odds that a person exercises regularly are $p_{1\cdot}/p_{2\cdot}$. In addition, the odds of exercising regularly can be examined separately for each personality type. For type A personalities, the odds are p_{11}/p_{21}, and for type B personalities, the odds are p_{12}/p_{22}. Intuitively, if exercise and personality types are independent, then the odds of regular exercise should be the same for both personality types. In particular, the ratio of the two sets of odds should be one.

Proposition 2.2.2. If rows and columns are independent, then the

odds ratio

$$\frac{(p_{11}/p_{21})}{(p_{12}/p_{22})} = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

equals one.

Proof. By equation (1)

$$\frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{p_{1\cdot}p_{\cdot 1}p_{2\cdot}p_{\cdot 2}}{p_{1\cdot}p_{\cdot 2}p_{2\cdot}p_{\cdot 1}} = 1.$$

□

If the odds ratio is estimated under the assumption of independence, $\hat{p}_{ij} = \hat{p}_{i\cdot}\hat{p}_{\cdot j} = n_{i\cdot}n_{\cdot j}/(n_{\cdot\cdot})^2$; so the estimated odds ratio is always one. A more interesting approach is to estimate the odds ratio without assuming independence and then see how close the estimated odds ratio is to one. With this approach, $\hat{p}_{ij} = n_{ij}/n_{\cdot\cdot}$ and

$$\frac{\hat{p}_{11}\hat{p}_{22}}{\hat{p}_{12}\hat{p}_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

In the personality-exercise example, the estimated odds ratio is

$$\frac{(483)(1121)}{(477)(1101)} = 1.03$$

which is very close to one. The log odds are .0305, the asymptotic standard error is $[1/483 + 1/477 + 1/1101 + 1/1121]^{1/2} = .0772$, and the test statistic for H_0 that the log odds equal 0 is $.0305/.0772 = .395$. Again, there is no evidence against independence.

EXERCISE 2.2. Give a 95% confidence interval for the odds ratio. Explain what the confidence interval means.

2.3 $I \times J$ Tables

The situation examined in Section 1 can be generalized to consider samples from I different populations, each of which is divided into J categories. We assume that the samples from different populations are independent and that each sample follows a multinomial distribution. This is *product-multinomial sampling.*

Similarly, a sample from one population that is categorized by two factors can be generalized beyond the case considered in Section 2. We allow one factor to have I categories and the other factor to have J categories.

Between the two factors, the population is divided into a total of IJ categories. The distribution of counts within the IJ categories is assumed to have a multinomial distribution. Consequently, this sampling scheme is multinomial sampling.

An $I \times J$ table of the observations n_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, can be written

		Factor 2 (Categories)				
	n_{ij}	1	2	...	J	Totals
	1	n_{11}	n_{12}	...	n_{1J}	$n_{1\cdot}$
Factor 1	2	n_{21}	n_{22}	...	n_{2J}	$n_{2\cdot}$
(Populations)	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	I	n_{I1}	n_{I2}	...	n_{IJ}	$n_{I\cdot}$
	Totals	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot J}$	$n_{\cdot\cdot}$

with similar tables for the probabilities p_{ij} and the expected values m_{ij}.

The analysis of product-multinomial sampling begins by testing whether all of the I multinomial populations are identical. In other words, we wish to test the model

$$H_0 : \; p_{1j} = p_{2j} = \cdots = p_{Ij} \text{ for all } j = 1, \ldots, J . \tag{1}$$

against the alternative

$$H_A : \text{ model (1) is not true} .$$

This is described as testing for *homogeneity of proportions.*

We continue to use Pearson's chi-square test statistic to evaluate the appropriateness of the null hypothesis model. Pearson's chi-square requires estimates of the expected values m_{ij}. Each sample i has a multinomial distribution with $n_{i\cdot}$ trials, so

$$m_{ij} = n_{i\cdot} p_{ij} .$$

If H_0 is true, p_{ij} is the same for all values of i. A pooled estimate of the common value of the p_{ij}'s is

$$\hat{p}_{ij}^{(0)} = n_{\cdot j}/n_{\cdot\cdot} .$$

In other words, if all the populations have the same probability for category j, an estimate of this common probability is the total number of observations in category j divided by the overall total number of observations. From this probability estimate we obtain

$$\hat{m}_{ij}^{(0)} = n_{i\cdot}(n_{\cdot j}/n_{\cdot\cdot}) .$$

In both $\hat{p}_{ij}^{(0)}$ and $\hat{m}_{ij}^{(0)}$, the superscript (0) is used to indicate that the estimate was obtained under the assumption that H_0 was true. Pearson's chi-square test statistic is

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \hat{m}_{ij}^{(0)}\right)^2}{\hat{m}_{ij}^{(0)}} .$$

For large samples, if H_0 is true, the approximation

$$X^2 \sim \chi^2((I-1)(J-1))$$

is valid. H_0 is rejected in an α level test if

$$X^2 > \chi^2(1-\alpha, (I-1)(J-1)) .$$

Note that if $I = J = 2$, these are precisely the results discussed in Section 1.

The analysis of a multinomial sample begins by testing for independence of the two factors. In particular, we wish to test the model

$$H_0 : p_{ij} = p_{i\cdot} p_{\cdot j}, \quad i = 1, \ldots, I, \; j = 1, \ldots, J . \tag{2}$$

We again use Pearson's chi-square. The marginal probabilities are estimated as

$$\hat{p}_{i\cdot} = n_{i\cdot}/n_{\cdot\cdot}$$

and

$$\hat{p}_{\cdot j} = n_{\cdot j}/n_{\cdot\cdot} .$$

Because $m_{ij} = n_{\cdot\cdot} p_{ij}$, if the model in (2) is true, we can estimate m_{ij} with

$$\begin{aligned} \hat{m}_{ij}^{(0)} &= n_{\cdot\cdot} \hat{p}_{i\cdot} \hat{p}_{\cdot j} \\ &= n_{\cdot\cdot} (n_{i\cdot}/n_{\cdot\cdot})(n_{\cdot j}/n_{\cdot\cdot}) \\ &= n_{i\cdot} n_{\cdot j}/n_{\cdot\cdot} \end{aligned}$$

where the (0) in $\hat{m}_{ij}^{(0)}$ indicates that the estimate is obtained assuming that (2) holds. The Pearson chi-square test statistic is

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \hat{m}_{ij}^{(0)}\right)^2}{\hat{m}_{ij}^{(0)}}$$

which, if (2) is true and the sample size is large, is approximately distributed as a $\chi^2\left((I-1)(J-1)\right)$. H_0 is rejected at the α level if

$$X^2 > \chi^2\left(1-\alpha, (I-1)(J-1)\right) .$$

Once again, if $I = J = 2$, we obtain the previous results given for 2×2 tables. Moreover, the test procedures for product-multinomial sampling and

for multinomial sampling are numerically identical. Only the interpretations of the tests differ.

EXAMPLE 2.3.1. Fifty-two males between the ages of 11 and 30 were operated on for knee injuries using arthroscopic surgery. The patients were classified by type of injury: twisted knee, direct blow, or both. The results of the surgery were classified as excellent (E), good (G), and fair or poor (F-P). These data can reasonably be viewed as either multinomial or product-multinomial. As a multinomial, we have 52 people cross-classified by type of injury and result of surgery. However, we can also think of the three types of injuries as defining different populations. Each person sampled from a population is given arthroscopic surgery and then the result is classified. Because our primary interest is in the result of surgery, we choose to *think* of the sampling as product-multinomial. The form of the analysis is identical for both sampling schemes. The data are

		Result			
	n_{ij}	E	G	F-P	Totals
Injury	Twist	21	11	4	36
	Direct	3	2	2	7
	Both	7	1	1	9
	Totals	31	14	7	52

The estimated expected counts under H_0 are

		Result			
	$\hat{m}_{ij}^{(0)}$	E	G	F-P	Totals
Injury	Twist	21.5	9.7	4.8	36
	Direct	4.2	1.9	.9	7
	Both	5.4	2.4	1.2	9
	Totals	31	14	7	52

with

$$X^2 = 3.229$$

and

$$df = (3-1)(3-1) = 4\,.$$

If the sample size is large, X^2 can be compared to a χ^2 distribution with four degrees of freedom. If we do this, the *P value* for the test is .52. Unfortunately, it is quite obvious that the sample size is not large. The number of observations in many of the cells of the table is small. This is a serious problem and aspects of the problem are discussed in Section 4, the subsection of Section 3.5 on Other Sampling Models, and Chapter 8 However, to the extent that this book focuses on distribution theory, it focuses on asymptotic distributions. For now, we merely state that in this

example, the n_{ij}'s and $\hat{m}_{ij}^{(0)}$'s are such that, when taken together with the very large P *value*, we feel safe in concluding that these data provide no evidence of different surgical results for the three types of injuries. (This conclusion is borne out by the fact that an exact small sample test yields a similar P *value*, cf. Section 3.5.)

2.3.1 Response Factors

In Example 2.3.1, the result of surgery can be thought of as a response, whereas the type of injury is used to explain the response. Similarly, in Example 2.1.1, opinions on abortions can be considered as a response and sex can be considered as an explanatory factor.

The existence of response factors is often closely tied to the sampling scheme. Product-multinomial sampling is commonly used with an independent multinomial sample taken for every combination of the explanatory factors and the categories of the multinomials being the categories of the response factors. This is illustrated in Example 2.1.1 where there are two independent multinomials (binomials), one for males and one for females. The categories for each multinomial are Support and Do Not Support legalized abortion. Example 3.5.2 in the next chapter involves two explanatory factors, Sex and Socioeconomic Status, and one response factor, Opinion on Legalized Abortion. Each of the four combinations obtained from the two sexes and the two statuses define an independent multinomial. In other words, there is a separate multinomial sample for each combination of sex and socioeconomic status. The categories of the response factor, Support and Do Not Support legalized abortion, are the categories of the multinomials.

More generally, the categories of a response factor can be cross-classified with other response factors or explanatory factors to yield the categories in a series of independent multinomials. This situation is of most interest when there are several factors involved. Some factors can be cross-classified to define the multinomial populations while other factors can be cross-classified with the response factors to define the categories of the multinomials. Example 2.3.1 illustrates the simplest case in which there is one explanatory factor, Injury, crossed with one response factor, Result, to define the categories of the multinomial. Both Injury and Result have three levels so the multinomial has a total of nine categories. With only two factors in the table, there can be only one multinomial sample because there are no other factors available to define various multinomial populations. Example 3.5.1 is more general in that it has two independent multinomials, one for each sex. Each multinomial has six categories. The categories are obtained by cross-classifying the explanatory factor, Political Party, having three levels, with the response factor, Abortion Opinion, having two levels.

In this more general sampling scheme, one often conditions on all factors other than the response so that the analysis is reduced to that of the original

sampling scheme in which every combination of explanatory factors defines an independent multinomial. Again, this is illustrated in Example 2.3.1. While the sampling was multinomial, we treated the sampling as product-multinomial with an independent multinomial for each level of the Injury. The justification for treating the data as product-multinomial is that we conditioned on the Injury.

While the sampling techniques described above are probably the most commonly used, there are alternatives that are also commonly used. For example, in medicine a response factor is often some disease with levels that are various states of the disease. If the disease is at all rare, it may be impractical to sample different populations and see how many people fall into the various levels of the disease. In this case, one may need to take the disease levels as populations, sample from these populations, and investigate various characteristics of the populations. Thus the "explanatory" factors discussed above would be considered descriptive factors here. This sampling scheme is often called *retrospective* for obvious reasons. The other schemes discussed above are called *prospective*. These issues are discussed in more detail in the introduction to Chapter 4 and in Sections 4.7 and 11.7.

2.3.2 Odds Ratios

The null hypotheses (1) and (2) can be rewritten in terms of odds ratios.

Proposition 2.3.2. Under product-multinomial sampling $p_{1j} = \cdots = p_{Ij} > 0$ for all $j = 1, \ldots, J$ if and only if

$$\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}} = 1$$

for all i, $i' = 1, \ldots, I$ and j, $j' = 1, \ldots, J$.

Proof. a) *Equality of probabilities across rows implies that the odds ratios equal one.* By substitution,

$$\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}} = \frac{p_{ij}p_{ij'}}{p_{ij'}p_{ij}} = 1.$$

b) *All odds ratios equal to one implies equality of probabilities across rows.* Recall that $p_{i\cdot} = 1$ for all $i = 1, \ldots, I$, so that $p_{\cdot\cdot} = I$. In addition, $p_{ij}p_{i'j'}/p_{ij'}p_{i'j} = 1$ implies $p_{ij}p_{i'j'} = p_{ij'}p_{i'j}$. Note that

$$\begin{aligned} p_{ij} = p_{ij}p_{\cdot\cdot}/I &= \frac{1}{I}\sum_{i'=1}^{I}\sum_{j'=1}^{J} p_{ij}p_{i'j'} \\ &= \frac{1}{I}\sum_{i'=1}^{I}\sum_{j'=1}^{J} p_{ij'}p_{i'j} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{I}\sum_{j'=1}^{J} p_{ij'} \sum_{i'=1}^{I} p_{i'j} \\
&= \frac{1}{I}\sum_{j'=1}^{J} p_{ij'} p_{\cdot j} \\
&= \frac{1}{I} p_{\cdot j} \sum_{j'=1}^{J} p_{ij'} \\
&= \frac{1}{I} p_{\cdot j} p_{i\cdot} \\
&= p_{\cdot j}/I\,.
\end{aligned}
$$

Because this holds for any i and j, $p_{\cdot j}/I = p_{1j} = p_{2j} = \cdots = p_{Ij}$ for $j = 1, \ldots, J$. □

Proposition 2.3.3. Under multinomial sampling, $0 < p_{ij} = p_{i\cdot}p_{\cdot j}$ for all $i = 1, \ldots, I$, $j = 1, \ldots, J$ if and only if

$$\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}} = 1$$

for all $i, i' = 1, \ldots, I$ and $j, j' = 1, \ldots, J$.

Proof. a) *Independence implies that the odds ratios equal one.*

$$\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}} = \frac{p_{i\cdot}p_{\cdot j}p_{i'\cdot}p_{\cdot j'}}{p_{i\cdot}p_{\cdot j'}p_{i'\cdot}p_{\cdot j}} = 1.$$

b) *All odds ratios equal to one implies independence.* If $p_{ij}p_{i'j'}/p_{ij'}p_{i'j} = 1$ for all i, i', j, and j', then $p_{ij}p_{i'j'} = p_{ij'}p_{i'j}$. Moreover, because $p_{\cdot\cdot} = 1$,

$$
\begin{aligned}
p_{ij} = p_{ij}p_{\cdot\cdot} = \sum_{i'=1}^{I}\sum_{j'=1}^{J} p_{ij}p_{i'j'} &= \sum_{i'=1}^{I}\sum_{j'=1}^{J} p_{ij'}p_{i'j} \\
&= \sum_{i'=1}^{I} p_{i'j} \sum_{j'=1}^{J} p_{ij'} \\
&= \sum_{i'=1}^{I} p_{i'j}p_{i\cdot} \\
&= p_{i\cdot}\sum_{i'=1}^{I} p_{i'j} \\
&= p_{i\cdot}p_{\cdot j}
\end{aligned}
$$

□

There is a great deal of redundancy in specifying that

$$\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}} = 1$$

for all i, i', j, and j'. For example, if $i = i'$, then $p_{ij}p_{i'j'}/p_{ij'}p_{i'j} = p_{ij}p_{ij'}/p_{ij'}p_{ij} = 1$ and no real restriction has been placed on the p_{ij}'s. A similar result occurs if $j = j'$. More significantly, if

$$p_{12}p_{23}/p_{13}p_{22} = 1$$

and

$$p_{12}p_{24}/p_{14}p_{22} = 1,$$

then

$$\begin{aligned} 1 &= (p_{12}p_{23}/p_{13}p_{22})(p_{14}p_{22}/p_{12}p_{24}) \\ &= p_{14}p_{23}/p_{13}p_{24}\,. \end{aligned}$$

In other words, the fact that two of the odds ratios equal one implies that a third odds ratio equals one. It turns out that the condition

$$\frac{p_{11}p_{ij}}{p_{1j}p_{i1}} = 1$$

for $i = 2, \ldots, I$ and $j = 2, \ldots, J$ is equivalent to the condition that all odds ratios equal one.

Proposition 2.3.4. $p_{ij}p_{i'j'}/p_{ij'}p_{i'j} = 1$ for all i, i', j and j' if and only if $p_{11}p_{ij}/p_{1j}p_{i1} = 1$ for all $i \neq 1$, $j \neq 1$.

Proof. Clearly, if the odds ratios are one for all i, i', j, and j', then $p_{11}p_{ij}/p_{1j}p_{i1} = 1$ for all i and j. Conversely,

$$\begin{aligned} 1 &= \left(\frac{p_{11}p_{ij}}{p_{1j}p_{i1}}\right)\left(\frac{p_{11}p_{i'j'}}{p_{1j'}p_{i'1}}\right) \Big/ \left(\frac{p_{11}p_{ij'}}{p_{1j'}p_{i1}}\right)\left(\frac{p_{11}p_{i'j}}{p_{1j}p_{i'1}}\right) \\ &= \frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}} \end{aligned}$$

□

Of course, in practice the p_{ij}'s are never known. We can investigate independence by examining the estimated odds ratios

$$\hat{p}_{ij}\hat{p}_{i'j'}/\hat{p}_{ij'}\hat{p}_{i'j} = n_{ij}n_{i'j'}/n_{ij'}n_{i'j}$$

or, equivalently, we can look at the log of this. For large samples, the log of the estimated odds ratio is normally distributed with standard error

$$\text{SE} = \sqrt{\frac{1}{n_{ij}} + \frac{1}{n_{ij'}} + \frac{1}{n_{i'j}} + \frac{1}{n_{i'j'}}}\,.$$

This allows the construction of asymptotic tests and confidence intervals for the log odds ratio. Of particular interest is the hypothesis

$$H_0 : p_{ij}p_{i'j'}/p_{ij'}p_{i'j} = 1\,.$$

After taking logs, this becomes

$$H_0 : \log(p_{ij}p_{i'j'}/p_{ij'}p_{i'j}) = 0\,.$$

EXAMPLE 2.3.5. We continue with the knee injury data of Example 2.3.1. From Proposition 2.3.4, it is sufficient to examine

$$\begin{aligned}
\frac{n_{11}n_{22}}{n_{12}n_{21}} = 21(2)/11(3) &= 1.27\,,\\
\frac{n_{11}n_{23}}{n_{13}n_{21}} = 21(2)/4(3) &= 3.5\,,\\
\frac{n_{11}n_{32}}{n_{12}n_{31}} = 21(1)/11(7) &= .27\,,\\
\frac{n_{11}n_{33}}{n_{13}n_{31}} = 21(1)/4(7) &= .75\,.
\end{aligned}$$

Although the X^2 test indicated no difference in the populations (populations = injury types), *at least* two of these estimated odds ratios *seem* substantially different from 1. In particular, relative to having an F-P result, the odds of an excellent result are about 3.5 times larger with twist injuries than with direct blows. Also, relative to having a good result, the odds of an excellent result from a twisted knee are only .27 of the odds of an excellent result with both types of injury. These numbers seem quite substantial, but they are difficult to evaluate without some idea of the error to which the estimates are subject. To this end, we use the large sample standard errors for the log odds ratios. Testing whether the log odds ratios are different from zero, we get

odds ratio	log (odds ratio)	SE	z
1.27	0.2412	0.9858	0.24
3.5	1.2528	1.0635	1.18
.27	−1.2993	1.1320	−1.15
.75	−0.2877	1.2002	−0.24

The large standard errors and small z values are consistent with the result of the X^2 test. None of the odds ratios appear to be substantially different from 1. Of course, it should not be overlooked that the standard errors are really only valid for large samples and we do not have large samples. Thus, all of our conclusions about the individual odds ratios must remain tentative.

2.4 Maximum Likelihood Theory for Two-Dimensional Tables

In this section, we introduce the *likelihood function*, *maximum likelihood estimates*, and *(generalized) likelihood ratio tests*. A valuable result for maximum likelihood estimation is given below without proof.

Lemma 2.4.1. Let $f(p_1, \ldots, p_r) = \sum_{i=1}^{r} n_i \log p_i$. If $n_i > 0$ for $i = 1, \ldots, r$, then, subject to the conditions $0 < p_i < 1$ and $p_\cdot = 1$, the maximum of $f(p_1, \ldots, p_r)$ is achieved at the point $(p_1, \ldots, p_r) = (\hat{p}_1, \ldots, \hat{p}_r)$ where $\hat{p}_i = n_i / n_\cdot$.

In this section, we consider product-multinomial sampling of I populations, with each population divided into the same J categories. The I populations will form the rows of an $I \times J$ table. No results will be presented for multinomial sampling in an $I \times J$ table. The derivations of such results are similar to those presented here and are left as an exercise.

The probability of obtaining the data $n_{i1}, \ldots, n_{iJ}$ from the ith multinomial sample is

$$\frac{n_{i\cdot}!}{\prod_{j=1}^{J} n_{ij}!} \prod_{j=1}^{J} p_{ij}^{n_{ij}}.$$

Because the I multinomials are independent, the probability of obtaining all of the values n_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, is

$$\prod_{i=1}^{I} \left[\frac{n_{i\cdot}!}{\prod_{j=1}^{J} n_{ij}!} \prod_{j=1}^{J} p_{ij}^{n_{ij}} \right]. \tag{1}$$

Thus, if we know the p_{ij}'s, we can find the probability of obtaining any set of n_{ij}'s. In point of fact, we are in precisely the opposite position. We do not know the p_{ij}'s, but we do know the n_{ij}'s. The n_{ij}'s have been observed. If we think of (1) as a function of the p_{ij}'s, we can write

$$L(p) = \prod_{i=1}^{I} \left[\frac{n_{i\cdot}!}{\prod_{j=1}^{J} n_{ij}!} \prod_{J=1}^{J} p_{ij}^{n_{ij}} \right] \tag{2}$$

where $p = (p_{11}, p_{12}, \ldots, p_{IJ})$. $L(p)$ is called the likelihood function for p. Some values of p give a very small probability of observing the n_{ij}'s that were actually observed. Such values of p are unlikely to be the true value of p. The true value of p is likely to be some value that gives a relatively large probability of observing what was actually observed. If we wish to estimate p, it makes sense to use the value of p that gives the largest probability of seeing what was actually observed. In other words, it makes sense to estimate p with a value $\hat{p}$ that maximizes the likelihood function $L(p)$. Such a value is called a *maximum likelihood estimate (MLE)* of p.

Rather than maximizing the likelihood function (which involves many products), it is often easier to maximize the log of the likelihood function (in which products change to sums). Because the logarithm is a strictly increasing function, the maximum of the likelihood and the maximum of the log of the likelihood occur as the same point.

For product-multinomial sampling, the log-likelihood function is

$$\log L(p) = \sum_{i=1}^{I} \left[\log(n_{i\cdot}!) - \sum_{j=1}^{J} \log(n_{ij}!) + \sum_{j=1}^{J} n_{ij} \log p_{ij} \right].$$

To maximize this as a function of p, we can ignore any terms that do not depend on p. It suffices to maximize

$$\ell(p) = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log p_{ij}.$$

The maximum is achieved when we maximize each of the terms $\sum_{j=1}^{J} n_{ij} \log p_{ij}$. By Lemma 2.4.1, the maximum is achieved at $p = \hat{p}$, where

$$\hat{p}_{ij} \equiv n_{ij}/n_{i\cdot}\,.$$

We can also obtain maximum likelihood estimates for the expected counts m_{ij}. Because $m_{ij} = n_{i\cdot} p_{ij}$, the MLE of m_{ij} is

$$\hat{m}_{ij} = n_{i\cdot} \hat{p}_{ij} = n_{ij}\,.$$

This follows from the *invariance of maximum likelihood estimates*: For any parameter θ and MLE $\hat{\theta}$, the MLE of a function of θ, say $f(\theta)$, is the corresponding function of the MLE, $f(\hat{\theta})$, cf. Cox and Hinkley (1974, p. 287).

If we change the model so that the null hypothesis

$$H_0 : p_{1j} = \ldots = p_{Ij}, \qquad j = 1, \ldots, J,$$

is true, we get different maximum likelihood estimates. Let $\pi_j = p_{1j} = \cdots = p_{Ij}$. The log-likelihood function becomes

$$\log L(p) = \sum_{i=1}^{I} \left[\log(n_{i\cdot}!) - \sum_{j=1}^{J} n_{ij}! + \sum_{j=1}^{J} n_{ij} \log \pi_j \right].$$

Ignoring terms that do not involve the p_{ij}'s, we are led to maximize

$$\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\pi_j$$

or, equivalently,

$$\sum_{j=1}^{J} n_{\cdot j}\log\pi_j\,.$$

By Lemma 2.4.1, the maximum likelihood estimates become

$$\hat{p}_{ij}^{(0)} = \hat{\pi}_j = n_{\cdot j}/n_{\cdot\cdot}$$

where the (0) in $\hat{p}_{ij}^{(0)}$ is used to indicate that the estimate was obtained assuming that H_0 was true.

Maximum likelihood estimates of the m_{ij}'s are easily obtained under the null model H_0. Because $m_{ij} = n_{i\cdot}p_{ij}$,

$$\hat{m}_{ij}^{(0)} = n_{i\cdot}\hat{p}_{ij}^{(0)} = n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}\,.$$

Note that $\hat{p}_{ij}^{(0)}$ and $\hat{m}_{ij}^{(0)}$ are precisely the estimates used in Section 3 to test H_0.

The likelihood function can also be used as the basis for a test of whether H_0 is true. The data have a certain likelihood of being observed, which can be summarized as the maximum value that the likelihood function achieves. If we place any restriction on the possible values of the p_{ij}'s, we will reduce the likelihood of observing the data. If placing a restriction on the p_{ij}'s reduces the likelihood too much, we can infer that the restriction on the p_{ij}'s is not likely to be valid. The relative reduction in the likelihood can be measured by looking at the maximum of $L(p)$ subject to the restriction divided by the overall maximum of $L(p)$. If this ratio gets too small, we will reject the assumption that the restriction on the p_{ij}'s is valid. In particular, if the restriction on the p_{ij}'s is that H_0 is true, we reject H_0 when the likelihood ratio is too small.

Again, we can simplify the mathematics by examining the log of the likelihood ratio and rejecting H_0 when the log gets too small. Of course, the log of the likelihood ratio is just the difference in the log-likelihoods. The maximum value of the log-likelihood when the reduced model H_0 is true is

$$\log L(\hat{p}^{(0)}) = \sum_{i=1}^{I}\left[\log(n_{i\cdot}!) - \sum_{j=1}^{J}\log(n_{ij}!) + \sum_{j=1}^{J} n_{ij}\log(n_{\cdot j}/n_{\cdot\cdot})\right].$$

The overall maximum of the log-likelihood is

$$\log L(\hat{p}) = \sum_{i=1}^{I}\left[\log(n_{i\cdot}!) - \sum_{j=1}^{J}\log(n_{ij}!) + \sum_{j=1}^{J} n_{ij}\log(n_{ij}/n_{i\cdot})\right].$$

The difference is

$$\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log(n_{\cdot j}/n_{\cdot\cdot}) - \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log(n_{ij}/n_{i\cdot})\,.$$

If we multiply by -2 and simplify, we get a likelihood ratio test statistic

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} \hat{m}_{ij}\log\left(\frac{\hat{m}_{ij}}{\hat{m}_{ij}^{(0)}}\right)$$

where $\hat{m}_{ij} = n_{ij}$ is the MLE of m_{ij} in the unrestricted model and $\hat{m}_{ij}^{(0)} = n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$ is the MLE of m_{ij} under the restriction that H_0 is true.

The reason for multiplying by -2 is that with this multiplication, the approximation

$$G^2 \sim \chi^2\left((I-1)(J-1)\right)$$

is valid when H_0 is true and the samples are large. Note that because H_0 was to be rejected for small values of the likelihood ratio, after taking logs and multiplying by -2, H_0 should be rejected for large values of G^2. In particular, for large samples, an α level test of H_0 is rejected if

$$G^2 > \chi^2\left(1-\alpha,(I-1)(J-1)\right).$$

EXAMPLE 2.4.2. Computing the likelihood ratio test statistic using the data and estimated expected cell counts on knee operations in Example 2.3.1 gives

$$G^2 = 3.173\,.$$

This is similar to, but distinct from, the Pearson test statistic $X^2 = 3.229$. Both are based on 4 degrees of freedom. In this example, formal tests using either statistic suffer from the fact that the sample is not large.

Larntz (1978) indicates that, for small samples, the actual size of the test that rejects H_0 if

$$X^2 > \chi^2(1-\alpha,(I-1)(J-1))$$

tends to be nearer the nominal size α than the corresponding likelihood ratio test. This is related to the fact that G^2 becomes too large when the observations are small but the estimated expected cell counts are not. Kreiner (1987) comes to similar conclusions. From the results of Larntz and others, Fienberg (1979) concludes that (a) if the minimum expected cell count is about 1, χ^2 tests often work well and (b) if the sample size is 4 to 5 times the number of cells in the table, χ^2 tests give *P values* with the

correct order of magnitude. In practice, the first of these conclusions must compare the *estimated* expected cell counts to 1. In Example 2.4.2, the test statistics are similar, so the choice of test is not important. The data also pass both of the criteria mentioned by Fienberg. The rules of thumb given in this paragraph can be applied to higher-dimensional tables. Although X^2 has the advantage alluded to above, G^2 is more convenient to use in analyzing higher-dimensional tables. The likelihood ratio test statistic will be used almost exclusively after Chapter 3.

Discussion

There are philosophical grounds for preferring the use of G^2. The likelihood principle indicates that one's conclusions should depend on the relative values of the likelihood function. The likelihood function depends only on the data that actually occurred. Because G^2 is computed from the likelihood, its use *can* be consistent with the likelihood principle. Unfortunately, the standard use of G^2 is to compute a *P value* or to perform an α level test. Both of these procedures depend on data that could have happened but did not, so these uses for G^2 violate the likelihood principle. An excellent discussion of the likelihood principle is given by Berger and Wolpert (1984).

Although formal tests are conducted throughout this book, the real emphasis is on informal evaluation of models using G^2 and other tools. The emphasis is on modeling and data analysis, not formal inferential procedures. Nevertheless, a relatively complete account is given of the standard results in formal log-linear model methodology. Bayesian methods are the primary inferential methods that satisfy the likelihood principle. Chapter 13 discusses Bayesian logistic regression — but not log-linear models. Santner and Duffy (1989) include discussion of Bayesian methods.

Exercise 2.3. For multinomial sampling, H_0 is the restriction that $p_{ij} = p_{i\cdot}p_{\cdot j}$ for all i and j. Show that

(a) the unrestricted MLE of p_{ij} is $\hat{p}_{ij} = n_{ij}/n_{\cdot\cdot}$

(b) the unrestricted MLE of m_{ij} is $\hat{m}_{ij} = n_{ij}$

(c) the MLE of p_{ij} under H_0 is $\hat{p}_{ij}^{(0)} = \hat{p}_{i\cdot}\hat{p}_{\cdot j} = (n_{i\cdot}/n_{\cdot\cdot})(n_{\cdot j}/n_{\cdot\cdot})$

(d) the MLE of m_{ij} under H_0 is $\hat{m}_{ij}^{(0)} = n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$

(e) the likelihood ratio test rejects H_0 when

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{m}_{ij}\log\left(\frac{\hat{m}_{ij}}{\hat{m}_{ij}^{(0)}}\right)$$

gets too large.

2.5 Log-Linear Models for Two-Dimensional Tables

It is our intention to exploit the similarities between analysis of variance (ANOVA) and regression on the one hand and log-linear and logistic regression models on the other. We begin by discussing two-factor analysis of variance.

Consider a balanced ANOVA model $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$. We can change the symbols used to denote the parameters and rewrite the model as

$$y_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} + e_{ijk}, \tag{1}$$

$i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$. The e_{ijk}'s are assumed to be independent $N(0, \sigma^2)$. We can estimate σ^2 and test for interaction. If interaction exists, we can look at contrasts in the interaction; if no interaction exists, we can test for main effects and look at contrasts in the main effects. If some factor levels correspond to quantitative values, then regression ideas can be incorporated into the ANOVA. The estimate of σ^2 is the mean squared error

$$\text{MSE} = \frac{1}{IJ(K-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (y_{ijk} - \bar{y}_{ij\cdot})^2 \ .$$

Everything in the analysis other than the estimate of σ^2 is a function of the $\bar{y}_{ij\cdot}$'s. In particular, we can form an $I \times J$ table of the $\bar{y}_{ij\cdot}$'s. The goal of the analysis is to explore the structure of this table. The ANOVA model (1) and the corresponding contrasts in interactions and main effects have proved to be very useful tools in exploring this $I \times J$ table.

Let us reconsider what the ANOVA model is really saying. Basically, the ANOVA model is saying that the y_{ijk}'s are independent and that

$$y_{ijk} \sim N(m_{ij}, \sigma^2)$$

where

$$m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \ . \tag{2}$$

Our goal is to examine the structure of the m_{ij}'s. To do that, we use the MLEs of the m_{ij}'s, which are

$$\hat{m}_{ij} = \bar{y}_{ij\cdot} \ .$$

Our statistical inferences are based on the fact that the $\hat{m}_{ij}$'s are independent with

$$\hat{m}_{ij} \sim N(m_{ij}, \sigma^2/K)$$

and that the MSE is an estimate of σ^2, which is independent of the $\hat{m}_{ij}$'s. It is of interest to note that although the MSE is not the MLE of σ^2, exactly the same tests and confidence intervals for the m_{ij}'s would be obtained if

the MLE for σ^2 was used in place of the MSE (and suitable adjustments in distributions were made).

If we impose a restriction on the m_{ij}'s – for example, the restriction of no interaction

$$m_{ij} = u + u_{1(i)} + u_{2(j)} \tag{3}$$

– the MLEs of the m_{ij}'s change. In particular,

$$\hat{m}_{ij} = \bar{y}_{\cdots} + (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}) + (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}) \tag{4}$$

and the MLE of σ^2 also changes. It can be shown that the usual F test for no interaction is just the likelihood ratio test for no interaction.

To examine an $I \times J$ table of counts, we use similar techniques. The table entries have the property that

$$\mathrm{E}(n_{ij}) = m_{ij} \, .$$

Again, we are interested in the structure of the m_{ij}'s; however, instead of considering linear models like (2) and (3), we consider log-linear models such as

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

and

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} \, .$$

Our analysis will again rely on the MLEs of the m_{ij}'s and on likelihood ratio tests; however, there are some differences. The n_{ij}'s are typically multinomial or product-multinomial. Small sample results similar to those from standard analysis of variance are not available. Traditionally, tests and confidence intervals have been based on large sample approximate distributions. On the other hand, multinomial distributions depend only on the p_{ij}'s or, equivalently, the m_{ij}'s, so there is no need to deal with a term analogous to σ^2 in normal theory. Finally, the ANOVA model (1) is balanced; it has K observations in each cell of the table. This balance leads to simplifications in the analysis. If there are different numbers of observations in the cells, the simplifications are lost. For example, the simple formula (4) for MLEs under the no-interaction model does not apply. Log-linear models are analogous to ANOVA models with unequal numbers of observations. They almost never display all the simplifications associated with balanced observations in ANOVA and they only occasionally have simple formulas for MLEs. Although most work on log-linear models has used large sample (*asymptotic*) distributions, recently there has been considerable work on exact conditional inference and Bayesian inference for small samples. See the subsection on Other Sampling Methods in Section 3.5 for further discussion of exact conditional inference and Chapter 13 for a discussion of Bayesian methods.

There are several reasons for writing ANOVA type models for the $\log(m_{ij})$'s rather than the m_{ij}'s. One is that the large sample theory can be

worked out. In other words, one reason to do it is because it can be done. Another reason is that log-linear models arise in a natural fashion from the mathematics of Poisson sampling, cf. Chapter 9. Multinomial expected cell counts are bounded between 0 and the sample size N, these bounds place awkward limits on the parameters of ANOVA type models for the m_{ij}'s. Such problems do not arise in log-linear modeling. One of the best reasons for considering log-linear models is that they often have very nice interpretations. We now examine the interpretations of log-linear models for two factors.

Consider a multinomial sample. We know that $m_{ij} = n_{..}p_{ij}$. We can write a log-linear model

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \,. \tag{5}$$

This has accomplished absolutely nothing! The terms $u_{12(ij)}$ are sufficient to explain the m_{ij}'s. The u, $u_{1(i)}$, and $u_{2(j)}$ terms are totally redundant; they can have any values at all, and yet, by choosing the $u_{12(ij)}$'s appropriately, equation (5) can be made to hold. Because model (5) has enough u terms to completely explain any set of m_{ij}'s, model (5) is referred to as a *saturated* model.

A more interesting example of a log-linear model occurs when the rows and columns of the table are independent. If

$$m_{ij} = n_{..}p_{i\cdot}p_{\cdot j} \,,$$

then

$$\log m_{ij} = \log n_{..} + \log p_{i\cdot} + \log p_{\cdot j} \,.$$

In other words, if rows and columns are independent, a log-linear model of the form

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} \tag{6}$$

holds. However, if we are to base our analysis on log-linear models, it is even more important to know that if model (6) holds, then rows and columns are independent.

Theorem 2.5.1. For multinomial sampling in an $I \times J$ table, $\log(m_{ij}) = u + u_{1(i)} + u_{2(j)}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, if and only if $p_{ij} = p_{i\cdot}p_{\cdot j}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$.

Proof. We have already shown that independence implies the log-linear model.

If the log-linear model holds, then

$$m_{ij} = e^{u + u_{1(i)} + u_{2(j)}} \,.$$

Let $a = e^u$, $a_{1(i)} = e^{u_{1(i)}}$, and $a_{2(j)} = e^{u_{2(j)}}$. Let $a_{1(\cdot)} = \sum_{i=1}^{I} a_{1(i)}$ and similarly for $a_{2(\cdot)}$. Note that

$$\begin{aligned} p_{ij} &= m_{ij}/n_{\cdot\cdot} = a\, a_{1(i)} a_{2(j)}/n_{\cdot\cdot}\,, \\ p_{i\cdot} &= a a_{1(i)} a_{2(\cdot)}/n_{\cdot\cdot}\,, \\ p_{\cdot j} &= a a_{1(\cdot)} a_{2(j)}/n_{\cdot\cdot}\,, \end{aligned}$$

and

$$1 = p_{\cdot\cdot} = a a_{1(\cdot)} a_{2(\cdot)}/n_{\cdot\cdot}\,.$$

Substitution gives

$$\begin{aligned} p_{i\cdot} p_{\cdot j} &= a a_{1(i)} a_{2(\cdot)} a a_{1(\cdot)} a_{2(j)}/n_{\cdot\cdot}^2 \\ &= (a a_{1(i)} a_{2(j)}/n_{\cdot\cdot})(a a_{1(\cdot)} a_{2(\cdot)}/n_{\cdot\cdot}) \\ &= a a_{1(i)} a_{2(j)}/n_{\cdot\cdot} \\ &= p_{ij}\,. \end{aligned}$$

Thus, the log-linear model implies independence. □

For product-multinomial sampling,

$$m_{ij} = n_{i\cdot} p_{ij} \tag{7}$$

and the log-linear model

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

holds trivially. Now, consider the model under H_0. If $\pi_j = p_{1j} = \cdots = p_{Ij}$ for all $j = 1, \ldots, J$, then

$$m_{ij} = n_{i\cdot} \pi_j\,.$$

Theorem 2.5.2. For product-multinomial sampling in an $I \times J$ table where rows are independent samples, $\log m_{ij} = u + u_{1(i)} + u_{2(j)}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, if and only if $p_{1j} = \cdots = p_{Ij}$, $j = 1, \ldots, J$.

Proof. If for each j the probabilities p_{ij} are equal, we have $m_{ij} = n_{i\cdot}\pi_j$ and $\log m_{ij} = \log n_{i\cdot} + \log \pi_j$. Taking $u = 0$, $u_{1(i)} = \log(n_{i\cdot})$, and $u_{2(j)} = \log(\pi_j)$ shows that the log-linear model holds.

Conversely, if $\log m_{ij} = u + u_{1(i)} + u_{2(j)}$, then $m_{ij} = a a_{1(i)} a_{2(j)}$, where $a = e^u$, $a_{1(i)} = e^{u_{1(i)}}$, and $a_{2(j)} = e^{u_{2(j)}}$. Note that $p_{i\cdot} = 1$, so from (7), $m_{i\cdot} = n_{i\cdot}$ and

$$n_{i\cdot} = a\, a_{1(i)} a_{2(\cdot)}\,.$$

Because $p_{ij} = m_{ij}/n_{i\cdot}$,

$$\begin{aligned} p_{ij} &= aa_{1(i)}a_{2(j)}/n_{i\cdot} \\ &= aa_{1(i)}a_{2(j)}/aa_{1(i)}a_{2(\cdot)} \\ &= a_{2(j)}/a_{2(\cdot)}\,. \end{aligned}$$

This is true for any i, so $a_{2(j)}/a_{2(\cdot)} = p_{1j} = p_{2j} = \cdots = p_{Ij}$, $j = 1, \ldots, J$. □

2.5.1 Odds Ratios

In applications with high-dimensional tables, it is rare that there are no important interactions. In order to explore the nature of the interactions, we need to look at contrasts in the interactions. To do this, we need a method of defining contrasts in the interactions. We begin by reviewing methods for examining interactions in analysis of variance.

Let q_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, be any set of numbers with the property that $q_{i\cdot} = q_{\cdot j} = 0$. In balanced analysis of variance,

$$\sum_{i=1}^{I}\sum_{j=1}^{J} q_{ij}m_{ij} \tag{8}$$

is a contrast in the interactions. Using model (2) and the fact that $q_{i\cdot} = q_{\cdot j} = 0$, the contrast (8) can also be written as

$$\sum_{i=1}^{I}\sum_{j=1}^{J} q_{ij}u_{12(ij)}$$

which involves only the interactions. The most interpretable way of obtaining a contrast in the interactions is to define the interaction contrast in terms of contrasts in the main effects. Let a_i, $i = 1, \ldots, I$, determine a contrast in the rows (thus, $a_{\cdot} = 0$) and let b_j, $j = 1, \ldots, J$, determine a contrast in the columns (so $b_{\cdot} = 0$). Then, if we take $q_{ij} = a_i b_j$, we get a contrast in the interactions. Recall that if there is no interaction, all interaction contrasts equal zero. Conversely, the interaction has $(I-1)(J-1)$ degrees of freedom, so specifying that any $(I-1)(J-1)$ linearly independent contrasts in the interaction are all zero is equivalent to specifying that there is no interaction.

A valuable data analytic technique for examining interactions in two-way analysis of variance is the *interaction plot*. It consists of plotting the I curves determined by connecting the points $(j, \hat{m}_{ij})$, $j = 1, \ldots, J$, with line segments. In this plot, $\hat{m}_{ij} = \bar{y}_{ij\cdot}$, the estimate of m_{ij} in model (2). If there is no interaction, $m_{ij} = u + u_{1(i)} + u_{2(j)}$ and the I theoretical curves (j, m_{ij}) are parallel. If interaction exists, the theoretical curves are not

parallel. The curves $(j, \hat{m}_{ij})$ estimate the theoretical curves. If the curves $(j, \hat{m}_{ij})$ are approximately parallel, it suggests that there is no interaction. If interaction exists, the estimated curves can suggest the nature of the interaction. Whether the plots are approximately parallel depends on the variability of the $\hat{m}_{ij}$'s.

Rather than plotting the I curves based on $(j, \hat{m}_{ij})$, one can plot the J curves based on $(i, \hat{m}_{ij})$, $i = 1, \ldots, I$. Again, in the absence of interactions, the curves should be approximately parallel. If the column treatments correspond to quantitative levels, say, x_j, $j = 1, \ldots, J$, then plots of $(x_j, \hat{m}_{ij})$ are appropriate. Again, one looks for parallelism. Similar plots can be constructed for row treatments with quantitative levels.

In log-linear models, the same procedures can be applied to the $\log(m_{ij})$'s. In particular, specifying that an odds ratio equals one is equivalent to specifying that an interaction contrast is zero. First, note that odds ratios can be written in terms of expected values. For product-multinomial sampling,

$$m_{ij} = n_{i\cdot} p_{ij}$$

and for multinomial sampling,

$$m_{ij} = n_{\cdot\cdot} p_{ij} \,.$$

In either case,

$$\frac{p_{ij} p_{i'j'}}{p_{ij'} p_{i'j}} = \frac{m_{ij} m_{i'j'}}{m_{ij'} m_{i'j}} \,.$$

If

$$\frac{m_{ij} m_{i'j'}}{m_{ij'} m_{i'j}} = 1 \,,$$

then taking logs gives

$$\log m_{ij} - \log m_{ij'} - \log m_{i'j} + \log m_{i'j'} = 0 \,.$$

This is precisely the statement that the interaction contrast

$$\sum_{r=1}^{I} \sum_{s=1}^{J} q_{rs} \log(m_{rs}) \tag{9}$$

equals zero, where $q_{ij} = q_{i'j'} = 1$, $q_{ij'} = q_{i'j} = -1$, and $q_{rs} = 0$ for all other pairs (r, s). In particular, the coefficients q_{rs} can be obtained by combining the contrast in the rows $a_i = 1$, $a_{i'} = -1$, and $a_r = 0$ for all other r with the contrast in the columns $b_j = 1$, $b_{j'} = -1$, and $b_s = 0$ for all other s. Observe that the contrast (9) can also be written

$$\sum_{r=1}^{I} \sum_{s=1}^{J} q_{rs} u_{12(rs)}$$

where we have used model (5) and the fact that $q_{r\cdot} = q_{\cdot s} = 0$.

If we specify that

$$\frac{m_{11}m_{ij}}{m_{1j}m_{i1}} = 1$$

for all $i = 2, \ldots, I$ and $j = 2, \ldots, J$, then we have specified that $(I-1)(J-1)$ linearly independent interaction contrasts in the $\log(m_{ij})$'s are all equal to zero; hence, there is no interaction.

As with analysis of variance, an interaction plot can be a valuable tool in the analysis of log-linear models. The I curves that connect the sets of points $(j, \log(\hat{m}_{ij}))$, $j = 1, \ldots, J$, are the basis of the interaction plot. The estimated expected counts $\hat{m}_{ij}$ are estimated using model (5), which contains interaction. Under model (5), $\hat{m}_{ij} = n_{ij}$. The I curves estimate the theoretical curves based on $(j, \log(m_{ij}))$. If there is no interaction, the theoretical curves are parallel and estimated curves should indicate this. If interaction exists, the nature of the interaction should be suggested by the estimated curves.

EXAMPLE 2.5.3. Consider the data given below on the relationship between college of enrollment and political affiliation for university students.

		Political Affiliation			
		Rep.	Dem.	Ind.	Total
College	Letters	34	61	16	111
	Engineering	31	19	17	67
	Agriculture	19	23	16	58
	Education	23	39	12	74
	Totals	107	142	61	310

The Pearson and likelihood ratio test statistics for independence (no interaction) are

$$X^2 = 16.16$$

and

$$G^2 = 16.39\,.$$

The test has $(4-1)(3-1) = 6$ degrees of freedom. The 99th percentile of a $\chi^2(6)$ is 16.81, so the *P value* for either statistic is a little above .01. An interaction plot uses the values $\log(n_{ij})$ given below.

	Political Affiliation		
	Rep.	Dem.	Ind.
Letters	3.5	4.1	2.8
Engineering	3.4	2.9	2.8
Agriculture	2.9	3.1	2.8
Education	3.1	3.7	2.5

The interaction plot is given in Figure 2.1. The curves for Letters and Education are almost parallel. The curve for Agriculture is similar but not nearly as concave. In fact, the Agriculture curve is nearly horizontal. The

FIGURE 2.1. Interaction Plot

Engineering curve is clearly the main source of interaction. It does not behave at all like the other three. It is clearly not parallel to the others. If Engineering is dropped from the table and the resulting 3×3 table is fit, one gets $X^2 = 5.770$ and $G^2 = 5.536$ on 4 degrees of freedom. The *P value* is a bit larger than .2. Without Engineering, there is no evidence for lack of independence. This confirms that the main source of interaction is in Engineering.

2.6 Simple Logistic Regression

In this section, we deal with simple logistic regression in which we use a predictor variable to estimate probabilities. Simple logistic regression, in fact, all of logistic regression, can be viewed as an extension of standard regression analysis. It can also be viewed as modeling the interactions in two-dimensional tables.

EXAMPLE 2.6.1. *O-ring Data.*
Table 2.1 presents data from Dalal, Fowlkes, and Hoadley (1989) on field O-ring failures in the 23 pre-*Challenger* space shuttle launches. See also Lavine (1991) and Martz and Zimmer (1992). *Challenger* was the shuttle that blew up on take off. Temperature is the predictor variable. The *Challenger* explosion occurred during a takeoff at 31 degrees Fahrenheit. Each flight is viewed as an independent trial. The result of a trial is 1 if any field O-rings failed on the flight and 0 if all the O-rings functioned properly. A

simple logistic regression uses temperature to model the probability that any O-ring failed. Such a model allows us to predict O-ring failure from temperature.

TABLE 2.1. O-ring Failure Data

Case	Flight	Failure	Success	Temperature
1	14	1	0	53
2	9	1	0	57
3	23	1	0	58
4	10	1	0	63
5	1	0	1	66
6	5	0	1	67
7	13	0	1	67
8	15	0	1	67
9	4	0	1	68
10	3	0	1	69
11	8	0	1	70
12	17	0	1	70
13	2	1	0	70
14	11	1	0	70
15	6	0	1	72
16	7	0	1	73
17	16	0	1	75
18	21	1	0	75
19	19	0	1	76
20	22	0	1	76
21	12	0	1	78
22	20	0	1	79
23	18	0	1	81

Let p_i be the probability that any O-ring fails in case i. A simple linear logistic regression model for these data is

$$\text{logit}(p_i) \equiv \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \tau_i,$$

where τ_i is the known temperature and β_0 and β_1 are unknown intercept and slope parameters (coefficients). The logistic regression model presents the log odds of O-ring failure as a linear function of temperature.

We again use maximum likelihood estimates. The likelihood function for logistic regression is discussed later in this section. The procedure for finding maximum likelihood estimates is discussed later in the book. For now, we merely present results and use analogies to standard regression.

The coefficient estimates, standard errors, and z values are

Variable	Estimate	Std. Error	z
Intercept	15.04	7.316	2.06
Temperature	−0.2321	0.1073	−2.16

The z values are simply the estimate divided by the standard error. They are test statistics for testing whether a coefficient equals zero. In particular, $z = -2.16$ yields a P *value* for $H_0 : \beta_1 = 0$ that is approximately .03. An alternative and preferred test is presented later.

To predict the probability of any O-ring failures for a flight at a temperature of τ,

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1\tau$$

which can be rearranged into

$$p = \frac{\exp(\beta_0 + \beta_1\tau)}{1 + \exp(\beta_0 + \beta_1\tau)}.$$

The estimated probability is

$$\hat{p} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1\tau)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1\tau)}.$$

Figure 2.2 gives a plot of the estimated probabilities as a function of temperature. The *Challenger* was launched at $\tau = 31$ degrees Fahrenheit, so the predicted log odds are $15.04 - (.2321)31 = 7.8449$ and the predicted probability of an O-ring failure is $e^{7.8449}/(1 + e^{7.8449}) = .9996$. Actually, there are problems with this prediction because we are predicting very far from the observed data. The lowest temperature at which a shuttle had previously been launched was 53 degrees, very far from 31 degrees. According to the fitted model, a launch at 53 degrees has probability .939 of O- ring failure, so even with the caveat about predicting beyond the range of the data, the model indicates an overwhelming probability of failure.

Before discussing logistic regression in general, we review standard one-way ANOVA and simple linear regression with normal errors. Suppose we have independent observations y_{ij} on I populations. The one-way ANOVA model is

$$y_{ij} = m_i + \varepsilon_{ij} \tag{1}$$

ε_{ij}'s independent $N(0, \sigma^2)$, $i = 1, \ldots, I$, $j = 1, \ldots, N_i$. Here, $\mathrm{E}(y_{ij}) \equiv m_i$. Alternatively, when a predictor variable x_i is available for each population, a simple linear regression model for the y_{ij}'s is

$$y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij} \ . \tag{2}$$

Model (2) is specifying a linear structure for the m_i's defined in model (1), i.e., for $i = 1, \ldots, I$

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

In general, we construct similar models for binomial data, except that the models are for the log odds rather than for the expected values. In a simple

FIGURE 2.2. O-ring Failure Probabilities

logistic regression, we have independent observations from I populations; each is $y_i \sim \text{Bin}(N_i, p_i)$. The N_i trials in the binomial play the same role as the N_i replicate observations in ANOVA. Recall that $\text{E}(y_i) = m_i = N_i p_i$ and that the odds are $p_i/(1-p_i)$. Logistic regression specifies a linear structure for the log odds

$$\text{logit}(p_i) \equiv \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i, \tag{3}$$

$i = 1, \ldots, I$. Note that by varying x_i, the term on the right of equation (3) can take on any real value. Although the probabilities p_i are restricted to be between 0 and 1, the log odds can also take on any real value.

Alternatively, the logistic regression model for the log odds can be viewed as a log-linear model. For example, we can think of the O-ring data in Table 2.1 as providing a two-way table in which there are independent samples from 23 populations and a binomial response of failure or success. Associated with the 23 populations in this 23×2 table are temperature values that we can use to model the interaction.

In general, we rewrite a logistic regression with I independent binomials as an $I \times 2$ two-way table. This involves substantially changing the notation we have used. The sampling is product-multinomial (actually, product-binomial). $y_i \sim \text{Bin}(N_i, p_i)$, so $(N_i - y_i) \sim \text{Bin}(N_i, 1-p_i)$. In terms of a two-way table, write $y_i \equiv n_{i1}$ and $N_i - y_i \equiv n_{i2}$. Note that $N_i = n_{i\cdot}$ for all i. Also, $p_i \equiv p_{i1}$ and $1 - p_i \equiv p_{i2}$ with similar definitions for the expected values. In particular, $m_i = N_i p_i = n_{i\cdot} p_{i1} = m_{i1}$ and $m_{i2} = N_i(1-p_i)$, so $p_i/(1-p_i) = m_{i1}/m_{i2}$.

The log-linear version of model (3) is

$$\log(m_{ij}) = u_{1(i)} + u_{2(j)} + \eta_j x_i \tag{4}$$

where the usual interaction term $u_{12(ij)}$ from (2.5.5) is being replaced in the model by a more specific interaction term, $\eta_j x_i$. Of course, x_i is the known predictor variable, but η_j is an unknown parameter. This is an interaction term because it involves both the i and j subscripts, just like $u_{12(ij)}$. The relationship between the logistic model (3) and the log-linear model (4) is that

$$\begin{aligned}
\log\left(\frac{p_i}{1-p_i}\right) &= \log\left(\frac{m_{i1}}{m_{i2}}\right) \\
&= \log(m_{i1}) - \log(m_{i2}) \\
&= \left[u_{1(i)} + u_{2(1)} + \eta_1 x_i\right] - \left[u_{1(i)} + u_{2(2)} + \eta_2 x_i\right] \\
&= \left[u_{2(1)} - u_{2(2)}\right] + \left[\eta_1 x_i - \eta_2 x_i\right] \\
&\equiv \beta_0 + \beta_1 x_i
\end{aligned}$$

where $\beta_0 \equiv \left[u_{2(1)} - u_{2(2)}\right]$ and $\beta_1 \equiv [\eta_1 - \eta_2]$.

As in Section 4, we can use maximum likelihood to estimate the parameters and to generate tests. The likelihood function $L(p)$ for a two-dimensional table was given in (2.4.2). Equation (3) can be rearranged to give

$$\begin{aligned}
p_i &= \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}, \\
1 - p_i &= \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)}.
\end{aligned}$$

Recalling that $p_i \equiv p_{i1}$, $(1 - p_i) \equiv p_{i2}$, $y_i \equiv n_{i1}$, and $N_i - y_i \equiv n_{i2}$, substitution into (2.4.2) gives the likelihood function

$$L(\beta_0, \beta_1) = \prod_{i=1}^{I} \left[\frac{n_{i\cdot}!}{\prod_{j=1}^{2} n_{ij}!} \left\{ \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right\}^{n_{i1}} \left\{ \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \right\}^{n_{i2}} \right].$$

It is by no means obvious what values of β_0 and β_1 will maximize this function. In Chapters 10 and 11, we discuss the *Newton–Raphson* method for obtaining such maxima. For now, we rely on a computer program to give us the maximizing values. (See Subsections 2.6.1 and 4.4.2 for SAS, BMDP, and GLIM computer commands.)

As in the example, if p is the probability for a predictor x,

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \quad \text{and} \quad p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

Given the MLEs $\hat{\beta}_0$ and $\hat{\beta}_1$, we get the estimated probability associated with x:

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

In particular, this formula provides the $\hat{p}_i$'s when doing predictions at the x_i's. It also provides $\hat{m}_{ij}$'s through $\hat{m}_{ij} = n_{i\cdot}\hat{p}_{ij}$. We can try to test model (4) against the more general saturated model (2.5.5). Recall that the MLEs for the expected cell counts under model (2.5.5) are just the n_{ij}'s, so

$$\begin{aligned} G^2 &= 2\sum_{i=1}^{I}\sum_{j=1}^{2} n_{ij}\log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right) \\ &= 2\sum_{i=1}^{I}[n_{i1}\log(n_{i1}/\hat{m}_{i1}) + n_{i2}\log(n_{i2}/\hat{m}_{i2})] \\ &= 2\sum_{i=1}^{I}[y_i\log(y_i/N_i\hat{p}_i) + (N_i - y_i)\log((N_i - y_i)/N_i(1-\hat{p}_i))]\,. \end{aligned}$$

In this formula, if $y_i = 0$, then $y_i\log(y_i)$ is taken as zero.

The Pearson test statistic is

$$\begin{aligned} X^2 &= \sum_{i=1}^{I}\sum_{j=1}^{2}\frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \\ &= \sum_{i=1}^{I}\left[\frac{(y_i - N_i\hat{p}_i)^2}{N_i\hat{p}_i} + \frac{[(N_i - y_i) - N_i(1-\hat{p}_i)]^2}{N_i(1-\hat{p}_i)}\right] \\ &= \sum_{i=1}^{I}\left[\frac{(y_i - N_i\hat{p}_i)^2}{N_i\hat{p}_i} + \frac{(y_i - N_i\hat{p}_i)^2}{N_i(1-\hat{p}_i)}\right] \\ &= \sum_{i=1}^{I}\frac{(y_i - N_i\hat{p}_i)^2}{N_i\hat{p}_i(1-\hat{p}_i)}\,. \end{aligned}$$

The degrees of freedom for the tests are $23 - 2 = 21$, i.e., the number of cases minus one for the intercept and one for temperature. This computation is based on model (3). Alternatively, based on model (4), the degrees of freedom are the number of cells in the two-way table, 23×2, minus 23 for fitting row effects and the grand mean, minus 1 for column effects, and minus 1 for fitting the interaction term based on temperature, i.e., $46 - 23 - 1 - 1 = 21$.

G^2 and X^2 are appropriate test statistics, but, unfortunately, for them to have large sample χ^2 distributions, we need the n_{ij}'s to get large. In this example, the n_{ij}'s are 0 or 1, so a χ^2 test is inappropriate for this example. In general, *a χ^2 test of a logistic regression model against the*

saturated model (2.5.5) is appropriate **only** *when the sample sizes N_i for the I populations are all large.*

We can also use model (4) as a full model and test it against a reduced model. Since models (4) and (3) are equivalent, we specify a reduced model for model (3), say

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0. \tag{5}$$

Testing model (5) against model (3) is equivalent to testing $H_0 : \beta_1 = 0$. Given an estimate $\hat{\beta}_0$ for model (5), we get $\hat{p}_i = e^{\hat{\beta}_0}/(1+e^{\hat{\beta}_0})$, estimates $\hat{p}_{ij}$, and estimates, say, $\hat{m}_{ij}^{(0)} = n_{i\cdot}\hat{p}_{ij}$, where the (0) indicates that the expected cell count is estimated under $H_0 : \beta_1 = 0$. The test statistic is

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{2}\hat{m}_{ij}\log\left(\frac{\hat{m}_{ij}}{\hat{m}_{ij}^{(0)}}\right).$$

Unlike standard regression analysis where the t test for $H_0 : \beta_1 = 0$ is equivalent to the F test, in logistic regression the z test described earlier can give different results than the G^2 test described here. Both tests of H_0 will generally be valid whenever I is large.

EXERCISE 2.4 Show that the independence model (2.5.6) implies model (5). Hint: Use the same method as was used to show that model (4) implies model (3).

Crude standardized residuals can be defined as

$$\tilde{r}_i = \frac{y_i - N_i\hat{p}_i}{\sqrt{N_i\hat{p}_i(1-\hat{p}_i)}}, \tag{6}$$

so that Pearson's chi-squared is $X^2 = \sum_{i=1}^{s}\tilde{r}_i^2$. Note that $\mathrm{Var}(y_i) = N_i p_i(1-p_i)$, making this definition of crude standardized residuals an estimate of $y_i - \mathrm{E}(y_i)/\sqrt{\mathrm{Var}(y_i)}$. (These are "crude" in that they ignore the variability of $\hat{p}_i$.) When $N_i = 1$, the residuals will not have an asymptotic normal distribution, which is a major reason why these residuals do not behave like residuals in normal theory models.

EXAMPLE 2.6.1 CONTINUED. For the simple linear logistic regression model, $G^2 = 20.315$ with 21 degrees of freedom. For the intercept-only model, $G^2 = 28.267$ with 22 degrees of freedom. Since $N_i = 1$ for all i, neither of these G^2's is compared directly to a chi-squared distribution.

The model based test for $H_0 : \beta_1 = 0$ has $G^2 = 28.267 - 20.315 = 7.952$ on $df = 22-21 = 1$. Comparing this to a $\chi^2(1)$ distribution, the *P value* for the test is approximately .005. It is considerably smaller than the *P value* for the z test of H_0. This test is generally preferred to the z test.

Since $N_i = 1$ for all i, we delay consideration of residuals until Chapter 4.

All of the methods presented in this section carry over to multiple logistic regression in which there is more than one predictor variable. Such models are discussed in Chapter 4.

2.6.1 Computer Commands

The data are in a file 'oring.dat' that looks like Table 2.1 except it has an extra column at the right (which contains the actual number of O-rings that failed on each flight). Perhaps the simplest way to fit the logistic regression model in SAS is to use PROC GENMOD.

```
options ps=60 ls=72 nodate;
data oring;
   infile 'oring.dat';
   input ID flt f s temp junk;
   n = 1;
proc genmod data = oring;
   model f/n = temp  / link=logit
                       dist=binomial;
run;
```

The first line controls printing of the output. The next four lines define the data. The variable "n" is used to specify that there is only one trial in each of the 23 binomials. PROC GENMOD needs the data specified: "data = oring". GENMOD also needs information on the model. "link = logit" and "dist = binomial" are both needed to specify that a logistic regression is being fitted. "model f/n = temp" indicates that we are modeling the number of failures in "f" out of "n" trials using the predictor "temp" (and implicitly an intercept).

A more powerful SAS program for logistic regression is PROC LOGISTIC. Commands for this, BMDP-LR, and GLIM are given in Subsection 4.4.2.

2.7 Exercises

Exercise 2.7.1. The data in Table 2.2 are on graduate admissions by sex at the University of California, Berkeley, and are given by Bickel et al. (1975) and Freedman et al. (1978). Test for independence, examine the Pearson residuals, and evaluate the odds ratio. What conclusions do you reach? (Do not put too much credence in your analysis; the data will be reanalyzed in Exercise 3.6.4.)

Exercise 2.7.2. Cramér (1946) presents data on the distribution of birth dates for males and females born in Sweden in 1935. The data given in Table 2.3 presume a natural ordering for the months of the year that

TABLE 2.2. Graduate Admissions at Berkeley

	Male	Female
Admitted	1198	557
Rejected	1493	1278

Cramér does not specify. Analyze the data. Is it better to think of this as one multinomial sample or as two independent multinomial samples?

TABLE 2.3. Swedish Birth Dates

Month	Female	Male
January	3537	3743
February	3407	3550
March	3866	4017
April	3711	4173
May	3775	4117
June	3665	3944
July	3621	3964
August	3596	3797
September	3491	3712
October	3391	3512
November	3160	3392
December	3371	3761

EXERCISE 2.7.3. Gilby (1911) presents data on the relationships among instructor's evaluation of general intelligence, quality of clothing, and school standard. General intelligence was classified using a system of Karl Pearson's that was reported in Waite (1911). Briefly, the Intelligence classifications are A – Mentally Defective, B – Dull, C – Slow, E – Fairly Intelligent, F – Capable, and G – Very Able. Clothing was classified as I – Very Well Clad, II – Well Clad, III – Poor but Passable, IV – Insufficient, V – Worse than Insufficient. Throughout, intelligence category A was combined with B and clothing category V was combined with IV. This was done because of small numbers of observations. The third variable, Standard, seems to be similar to the American idea of a school grade. For example, roughly half of 10-year-olds were in Standard III with most of the others in II or IV. For $10\frac{1}{2}$-year-olds, about two-thirds were in standards III or IV with most of the rest in II or V. Data were collected from 36 instructors spread over eight different primary schools. Tables 2.4, 2.5, and 2.6 summarize some of the data; use the methods of Chapter 2 to analyze these data.

EXERCISE 2.7.4. *Partitioning Tables.*

TABLE 2.4. Intelligence versus Clothing

	Intelligence					
Clothing	B	C	D	E	F	G
I	33	48	113	209	194	39
II	41	100	202	255	138	15
III	39	58	70	61	33	4
IV,V	17	13	22	10	10	1

TABLE 2.5. Intelligence versus Standard

	Intelligence					
Standard	B	C	D	E	F	G
I	17	27	45	50	27	1
II	23	34	61	66	36	1
III	42	42	69	117	72	10
IV	16	25	41	75	53	11
V	18	38	66	77	45	6
VI	10	32	73	80	98	18
VII	4	19	39	52	35	11
VIII	0	2	13	18	9	1

The examination of odds ratios and residuals provide two ways to investigate lack of independence in a two-way table. The partitioning methods of Irwin (1949) and Lancaster (1949) provide another. Christensen (1996a, Section 8.6) gives extensive examples of the application of these methods. Table 2.7 gives data on the occupation of family heads for families of various religious groups. The occupations are A – Professions, B – Owners, Managers, and Officials, C – Clerical and Sales, D – Skilled, E – Semiskilled, F – Unskilled, G – Farmers, H – No Occupation. The data were extracted from Lazerwitz (1961). Although the data were collected using a complex sampling design (cf. Section 3.5), ignore this fact in your analysis. To establish

TABLE 2.6. Clothing versus Standard

	Clothing			
Standard	I	II	III	IV,V
I	20	87	56	4
II	71	88	42	20
III	157	134	41	20
IV	82	77	45	17
V	101	117	29	3
VI	127	145	32	7
VII	59	81	18	2
VIII	19	22	2	0

the effect of the Protestant groups on the lack of independence, we can isolate the Protestant groups in a separate reduced table. We can also pool the Protestants together in a collapsed table that includes the non-Protestant groups. These are both given in Table 2.8. Test each of the three tables for independence. Note that G^2 for the full table equals the sum of the G^2's for the reduced table and the collapsed table. Continue the analysis of these data by using the partitioning procedure on the reduced and collapsed tables and on subsequent generations of reduced and collapsed tables. Note that tables can also be partitioned on their columns. At its logical extreme, this leads to a collection of 2×2 tables, each with one degree of freedom for testing independence. The Lancaster-Irwin partitioning provides a method of breaking the interaction (lack of independence) G^2 for the full table into one degree of freedom components that add up to the original G^2. This is similar to using orthogonal contrasts to break up the interaction sum of squares in a balanced analysis of variance. For a theoretical justification of the Lancaster-Irwin procedure, see Exercise 8.4.3.

TABLE 2.7. 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
Roman Catholic	102	140	127	279	254	127	51	190
Jewish	36	60	30	17	17	2	0	26
No Religion	19	12	6	12	25	9	14	28

EXERCISE 2.7.5. *Fisher's Exact Test.*
Consider the problem of testing whether the probability of success is the same for two independent binomials. Let $y_i \sim \text{Bin}(N_i, p_i)$, $i = 1, 2$. Write the 2×2 table as

y_1	$N_1 - y_1$	N_1
y_2	$N_2 - y_2$	N_2
t	$N_1 + N_2 - t$	$N_1 + N_2$

(a) Find $\Pr(y_1 = r_1 \text{ and } t = t_0)$ for arbitrary r_1 and t_0.

(b) Assuming $p_1 = p_2$, find $\Pr(y_1 = r_1 | t = t_0)$.

(c) Consider the following subset of the knee injury data of Example 2.3.1

TABLE 2.8. Partitioned Tables

Reduced Table

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

Collapsed Table

Religion	A	B	C	D	E	F	G	H
Protestant	210	277	254	394	356	245	232	415
Roman Catholic	102	140	127	279	254	127	51	190
Jewish	36	60	30	17	17	2	0	26
No Religion	19	12	6	12	25	9	14	28

	Result	
Injury	E	G
Direct	3	2
Twist	7	1

Using the conditional distribution of (b), find the probability of getting the observed value 3. The *P value* for Fisher's exact test is the sum of the $\Pr(y_1 = r_1 | t = 10)$'s for every r_1 value that satisfies

$$\Pr(y_1 = r_1 | t = 10) \leq \Pr(y_1 = 3 | t = 10).$$

Find the *P value* for the data given above. Note that this test does not depend on any large sample approximations, so it is exact even for small samples. On the other hand, the computations become difficult with large samples.

EXERCISE 2.7.6. *Yule's Q.*
For 2×2 tables, a measure of association similar to a correlation coefficient is Yule's Q, which is defined as

$$Q = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}}.$$

Find Q in terms of the odds ratio. Show that Q lies between -1 and 1.

EXERCISE 2.7.7. *Freeman-Tukey Residuals.*
Freeman and Tukey (1950) suggest a variance stabilizing transformation for Poisson data that leads to using the quantities

$$\sqrt{n_{ij}} + \sqrt{n_{ij} + 1} - \sqrt{4\hat{m}_{ij}^{(0)} + 1}$$

as residuals, cf. Bishop, Fienberg, and Holland (1975, Section 4.4). Reexamine the data of Example 2.3.1 using the Freeman-Tukey residuals.

EXERCISE 2.7.8. *Power Divergence Statistics.*
Cressie and Read (1984) and Read and Cressie (1988) have introduced the *power divergence* family of test statistics

$$2I^{\lambda} = \frac{2}{\lambda(\lambda+1)} \sum_{ij} n_{ij} \left[\left(\frac{n_{ij}}{\hat{m}_{ij}^{(0)}} \right)^{\lambda} - 1 \right],$$

where for $\lambda = -1, 0$ the statistics are defined by taking limits. They establish that for any λ, the large sample distribution under H_0 is χ^2 with the usual degrees of freedom. Show that $X^2 = 2I^1$ and $G^2 = 2I^0$. Find the relationship between $2I^{-1/2}$ and the Freeman-Tukey residuals discussed in Exercise 2.7.7.

EXERCISE 2.7.9. Compute the power divergence test statistics $2I^{-1/2}$ and $2I^{1/2}$ for the knee injury data of Example 2.3.1. Compare the results to G^2 and X^2. What conclusions can be reached about knee injuries?

EXERCISE 2.7.10. *Testing for Symmetry.*
Consider a multinomial sample arranged in an $I \times I$ table. In square tables with similar categories for the two factors, it is sometimes of interest to test

$$H_0 : \ p_{ij} = p_{ji}$$

for all i and j.

(a) Give a procedure for testing this hypothesis based on testing equality of probabilities (homogeneity of proportions) in a $2 \times I(I-1)/2$ table. If you think of the $I \times I$ table as a matrix, the rows indicate whether a cell is above or below the diagonal. The columns are corresponding off diagonal pairs. Illustrate the test for a 4×4 table.

(b) Give a justification for the procedure in terms of a (conditional) sampling model.

(c) The data in Table 2.9 were given by Fienberg (1980), Yule (1900), and earlier by Galton. They report the relative heights of 205 married couples. Test for symmetry and do any other appropriate analysis for these data. Do the data display symmetry?

EXERCISE 2.7.11. *Correlated Data.*
There are actually 410 observations in Exercise 2.7.10 and Table 2.9. There are 205 men and 205 women. Why was Table 2.9 set up as a 3×3 table with only 205 observations rather than as Table 2.10, a 2×3, sex versus height table with 410 observations?

TABLE 2.9. Heights of Married Couples

	Wife		
Husband	Tall	Medium	Short
Tall	18	28	14
Medium	20	51	28
Short	12	25	9

TABLE 2.10. Heights of Married Couples

	Height		
Sex	Tall	Medium	Short
Wife	50	104	51
Husband	60	99	46

EXERCISE 2.7.12. *McNemar's Test.*
McNemar (1947) proposes a method of testing for homogeneity of proportions among two binary populations when the data are correlated. (A binary population is one in which all members fall into one of two categories. Homogeneity means that the proportions in each category are the same for both groups.) If we restrict attention in Exercise 2.7.10 and Table 2.9 to the subpopulation of Tall and Medium people, we get an example of such data. The data on a husband and wife pair cannot be considered as independent, but this problem is avoided by treating each pair as a single response. The data from the subpopulation are given below.

	Wife	
Husband	Tall	Medium
Tall	18	28
Medium	20	51

Conditionally, these data are a multinomial sample of 117. The probability of a tall woman is $p_{11} + p_{21}$ and the probability of a medium woman is one minus that. The probability of a tall man is $p_{11} + p_{12}$ and again the probability of a medium man can be found by subtraction. It follows that the probability of a tall woman is the same as the probability of a tall man if and only if $p_{21} = p_{12}$. Thus, for 2×2 tables, the problem of homogeneity of proportions is equivalent to testing for symmetry. McNemar's test is just the test for symmetry in Exercise 2.7.10 applied to 2×2 tables. Check for homogeneity of proportions in the subpopulation. For square tables that are larger than 2×2, the problem of testing for *marginal homogeneity* is more difficult and cannot, as yet, be addressed using log-linear models. Nonetheless, a test can be obtained from basic asymptotic results,

cf. Exercise 10.8.6.

EXERCISE 2.7.13. Suppose the random variables n_{ij}, $i = 1, 2$, $j = 1, \ldots, N_i$, are independent Poisson(μ_i) random variables. Find the maximum likelihood estimates for μ_1 and μ_2 and find the generalized likelihood ratio test statistic for $H_0 : \mu_1 = \mu_2$.

EXERCISE 2.7.14. Yule's Q (cf. Exercise 2.7.6.) is one of many measures of association that have been proposed for 2×2 tables. Agresti (1984, Chapter 9) has a substantial discussion of measures of association. It has been suggested that measures of association for 2×2 multinomial tables should depend solely on the conditional probabilities of being in the first column given the row, i.e., $p_{11}/p_{1\cdot}$ and $p_{21}/p_{2\cdot}$, or, alternatively, on the conditional probabilities of being in the first row given the column, i.e., $p_{11}/p_{\cdot 1}$ and $p_{12}/p_{\cdot 2}$. Moreover, it has been suggested that the measure of association should not depend on which set of conditional probabilities are used. Show that any measure of association

$$f\left(\frac{p_{11}}{p_{1\cdot}}, \frac{p_{21}}{p_{2\cdot}}\right)$$

can be written as some function of the odds

$$g\left(\frac{p_{11}}{p_{12}}, \frac{p_{21}}{p_{22}}\right).$$

Show that if

$$f\left(\frac{p_{11}}{p_{1\cdot}}, \frac{p_{21}}{p_{2\cdot}}\right) = f\left(\frac{p_{11}}{p_{\cdot 1}}, \frac{p_{12}}{p_{\cdot 2}}\right)$$

for any sets of probabilities, then $g(x, y) = g(ax, ay)$ for any x, y, and a. Use this to conclude that any such measure of association is a function of the odds ratio.