STAT 545 Generalized Linear Models and Categorical Data Analysis: Part I

Instructor: Yisheng Li

1 Introduction: Distributions and Inference for Categorical Data

1.1 Categorical Response Data

A *categorical variable* has a measurement scale consisting of a set of categories. It arises from many different areas:

1. Social science: political philosophy is often measured as liberal, moderate, or conservative.
2. Biomedical sciences: Response Evaluation Criteria in Solid Tumors (RECIST) (for evaluation of target lesions): complete response (CR), partial response (PR), stable disease (SD), progressive disease (PD)
3. Behavioral sciences: smoking status: smoking vs abstinent; physical activity level: physically active vs sedentary
4. Epidemiology, public health, genetics, education, marketing, engineering, and so on.

1.1.1 Response vs Explanatory Variables

Most statistical analyses distinguish between *response (or dependent) variables* and *explanatory (or independent) variables.* For instance, typical regression models describe how the mean of a response variable changes according to the values of explanatory variables.

Examples:

1) A relationship between quality of life (QOL) (response) and disease stage (explanatory) and marital status (explanatory).

2) Is social economic status (SES) (explanatory) related to the probability of success in smoking cessation (response)?

This course focuses on categorical response variables, while explanatory variables are not restricted to categorical variables (as in ordinary regression).

1.1.2 Nominal vs Ordinal Scale

Categorical variables have two primary types of scales. Variables having categories without a natural ordering are called *nominal*.

Nominal: Race/ethnicity: White, Hispanic, Black (African American), Asian, etc.; mode of transportation to work: automobile, bicycle, bus, subway, walk; choice of residence: apartment, condominium, house, other.

For nominal variables, statistical analysis does not depend on the ordering of the categories.

Categorical variables having ordered categories are called *ordinal* variables:

Judgment of things: poor, fair, good, excellent

Frequency: never, rare, sometimes, often, always

Income (per annum): < 20K, ≥ 20K and < 40K, ≥ 40K and < 100K, ≥ 100K

Although categories of an ordinal variable are ordered, distances between categories are unknown.

Statistical analysis for ordinal variables utilize the category ordering.

An *interval variable* is one that *does* have numerical distances between any two values. Examples: blood pressure, tumor size, gene expression level, actual income, number of cigarettes smoked per day, etc.

The way a variable is measured determines its classification as nominal, ordinal, or interval. For example, income is an interval variable if it is measured as actual $; it becomes an ordinal variable if it is measured as a range at each level (such as < 20K, ≥ 20K, etc., per annum).

Hierarchy of variable types: interval (highest), ordinal (next), nominal (lowest).

Statistical methods for one type of variables can be used with higher types of variables (by ignoring certain information in the higher type of variable so that it is essentially treated as a lower type variable). For example, ignoring the order in an ordinal variable, one could then apply an analysis method for nominal variable on an ordinal variable.

1.1.3 Continuous vs Discrete Variables

Variables are categorized as continuous or discrete, according to the number of values they can take. The continuous-discrete classification, in practice, distinguishes between variables that take lots of values and variables that take few values. For instance, statisticians often treat discrete interval variables having a large number of values (such as test scores) as continuous, using them in methods for continuous responses.

This course deals with: 1) nominal variables; 2) ordinal variables; 3) discrete interval variables having relatively few values; 4) continuous variables grouped into a small number of categories.

1.1.4 Quantitative vs Qualitative Variables

Nominal variables are qualitative.

Interval variables are quantitative.

Ordinal variables are unclear. However, ordinal variables possess important quantitative features: 1) Each category has a greater or smaller magnitude of characteristic than another category; 2) Although not possible to measure, usually an underlying continuous variable is present. Given the quantitative nature of ordinal variables, analysts often either assign numerical scores to categories or assume an underlying continuous distribution (in the so-called latent variable models).

1.2 Distributions for Categorical Data

Data analysis generally requires assumption about the distribution of the data. For continuous responses, the normal distribution plays the central role. For categorical responses, we have three key distributions: *binomial, multinomial*, and *Poisson.*

1.2.1 Binomial Distribution

Suppose there are n independent trials. Each (Bernoulli) trial has an outcome of either success or failure (i.e., $Y_i = 1$ for a success and $Y_i = 0$ for a failure). Assume the success probability is $P(Y_i = 1) = \pi$. Then the number of successes out of the n trials, denoted as $Y = \sum_i Y_i$, follows a binomial distribution, denoted as bin(n,π).

The probability mass function for the possible outcomes y for Y is

$$p(y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y},\ y = 0, \ldots, n.$$

The binomial distribution has mean and variance $\mu = E(Y) = n\pi$ and $\sigma^2 = \text{var}(Y) = n\pi(1-\pi)$. The skewness is described by $E(Y-\mu)^3/\sigma^3 = (1-2\pi)/\sqrt{n\pi(1-\pi)}$. The distribution converges to normality as n increases, for fixed π.

When the n Bernoulli trials are not independent, the number of successes out of the n trials may not follow a binomial distribution. One example is *hypergeometric distribution*, arising when each binary outcome is sampled from a finite population without replacement. Another example is related to overdispersion (hopefully we can cover some related models on this later in the class).

1.2.2 Multinomial Distribution

Some trials have more than two possible outcomes. Suppose that each of the n independent, identical trials can have outcomes in any of c categories. Let $Y_{ij} = 1$ if trial i has outcome in category j and $Y_{ij} = 0$ otherwise. Let $n_j = \sum_i Y_{ij}$ denote the number of trials having outcome in category j. The counts $(n_1, n_2, \ldots, n_c)$ have the *multinomial distribution.*

Let $\pi_j = P(Y_{ij} = 1)$ denote the probability of outcome in category j for each trial. The multinomial probability mass function is

$$p(n_1, n_2, \ldots, n_c) = \left(\frac{n!}{n_1! n_2! \ldots n_c!} \right) \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c}.$$

Since $\sum_j n_j = n$, this is $(c-1)$-dimensional, with $n_c = n - (n_1 + \ldots n_{c-1})$. The binomial distribution is the special case with $c = 2$.

For the multinomial distribution,

$$E(n_j) = n\pi_j,\ \text{var}(n_j) = n\pi_j(1-\pi_j),\ \text{cov}(n_j, n_k) = -n\pi_j\pi_k,\ \text{if } j \neq k.$$

The marginal distribution of each n_j is binomial.

1.2.3 Poisson Distribution

The binomial and multinomial data result from a fixed number of trials. Sometimes the number of trials may not be fixed (i.e., could be as large as you can imagine).

Example. Let $y =$ the # of deaths due to automobile accidents on motorways in Italy during this coming week. There is no fixed upper limit n for y. y must be nonnegative integers. The simplest distribution with its probability mass on this range (i.e., nonnegative integers without upper limit) is *Poisson.* Its probability mass function is

$$p(y) = \frac{e^{-\mu}\mu^y}{y!},\ y = 0, 1, 2, \ldots$$

It satisfies $E(Y) = \text{var}(Y) = \mu$. It is unimodal with mode equal to the integer part of μ. Its skewness is described by $E(Y-\mu)^3/\sigma^3 = 1/\sqrt{\mu}$. The distribution approaches normality as μ increases.

The Poisson distribution is used for counts of events that occur randomly over time or space, when outcomes in disjoint periods or regions are independent. It also applies as an approximation for the binomial when n is large and π is small, with $\mu = n\pi$.

Example. If each of the 50 million people driving in Italy next week is an independent trial with probability 0.000002 of dying in a fatal accident that week, the # of deaths Y is a bin(50000000,0.000002) variate, or approximately Poisson with $\mu = n\pi = 50000000(0.000002) = 100$.

A key feature of the Poisson distribution is that its variance equals its mean. That is, sample counts vary more when their mean is higher.

1.2.4 Overdispersion

In practice, count observations often exhibit variability exceeding that predicted by the binomial or Poisson. This phenomenon is called *overdispersion*.

We assumed above that each person has the same probability of dying in a fatal accident in the next week. More realistically, these probabilities vary, due to factors such as amount of time spent driving, whether the person wears a seat belt, and geographical location. Such variation causes fatality counts to display more variation than predicted by the Poisson model.

Suppose Y is a random variable with variance $\text{var}(Y \mid \mu)$ for given μ, but μ is random (varying based on the above described factors). Let $\theta = E(\mu)$. Then we have

$$E(Y) = E[E(Y \mid \mu)], \ \text{var}(Y) = E[\text{var}(Y \mid \mu)] + \text{var}[E(Y \mid \mu)].$$

In the case of Poisson random variable Y given μ, $E(Y) = \theta$ and $\text{var}(Y) = \theta + \text{var}(\mu) > \theta$.

Assuming a Poisson distribution for a count variable is often too simplistic, because of factors that cause overdispersion. The *negative binomial* is a related distribution for count data that permits the variance to exceed the mean.

Similar problems may arise for binomial (or multinomial) distributions. Suppose we want to estimate the response rate of certain treatment for a certain type of cancer across the country. We randomly sample 20 patients from each of 30 randomly sampled cancer hospitals in the country and record the number of responses (e.g., complete responses), denoted as y_i, at each hospital. Due to factors that may differ between hospitals (e.g., skill level of physicians, general health status of the patient population [related to race composition, socioeconomic status, etc.]) and may not be measured, y_i may exhibit a larger variability than a binomial random variable with a fixed probability does (such as being clustered near 0 and 20 more). If time permits, later in this course we will introduce models for this type of data (e.g., generalized linear mixed models).

1.2.5 Connection between Poisson and Multinomial Distributions

In Italy this next week, let $y_1 = \#$ of people who die in automobile accidents, $y_2 = \#$ who die in airplane accidents, and $y_3 = \#$ who die in railway accidents. A Poisson model for (Y_1, Y_2, Y_3) treats these as independent Poisson random variables, with parameters (μ_1, μ_2, μ_3). The joint probability mass function for $\{Y_i\}$ is the product of the three Poisson mass functions. The total also has a Poisson distribution, with parameter $\sum \mu_i$.

With Poisson sampling the total count n is random. If we start with a Poisson model and then condition on n, $\{Y_i\}$ no longer have Poisson distribution, since each Y_i cannot exceed n. Given n, Y_i are also no longer independent, since the value of one affects the possible range for the others.

For c independent Poisson variates, with $E(Y_i) = \mu_i$, let us derive their conditional distribution given that $\sum Y_i = n$. The conditional probability of a set of counts $\{n_i\}$ satisfying this condition is

$$P\left[(Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c) \mid \sum Y_j = n\right]$$
$$= \frac{P(Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c)}{P(\sum Y_j = n)}$$
$$= \frac{\prod_i \left[\exp(-\mu_i)\mu_i^{n_i}/n_i!\right]}{\exp\left(-\sum \mu_j\right)\left(\sum \mu_j\right)^n/n!} = \frac{n!}{\prod_i n_i!}\prod \pi_i^{n_i},$$

where $\{\pi_i = \mu_i/\left(\sum \mu_j\right)\}$. This is the multinomial $(n, \{\sum \mu_j\})$ distribution, characterized by the sample size n and the probabilities $\{\pi_i\}$.

Many categorical data analyses assume a multinomial distribution. Such analyses usually have the same parameter estimates as those of analyses assuming a Poisson distribution, because of the similarity in the likelihood functions. We will see examples of this later in the Chapter of Inference for Two-Way Contingency Tables.

1.3 Statistical Inference for Categorical Data

The choice of distribution for the response variable is only the first step in data analysis. In practice, that distribution has unknown parameter values. We review methods of using sample data to make inference about the parameters.

1.3.1 Likelihood Functions and Maximum Likelihood Estimation

Throughout the first half of the course, we use *maximum likelihood (ML)* for parameter estimation. Under regularity conditions, such as the parameter space having fixed dimension with true value falling in its interior, maximum likelihood estimators have desirable properties: They have large-sample normal distributions; they are asymptotically consistent, converging to the parameter as n increases; and they are asymptotically efficient, producing large-sample standard errors no greater than those from other estimation methods.

Given the data, for a chosen probability distribution the *likelihood function* is the probability of those data, treated as a function of the unknown parameter. The ML estimate is the parameter value that maximizes this function. This is the parameter value under which the data observed have the highest probability of occurrence. The parameter value that maximizes the likelihood function also maximizes the log of that function. It is simpler to maximize the log likelihood since it is a sum rather than a product of terms.

We denote a parameter for a generic problem by β and its ML estimate by $\hat{\beta}$. The likelihood function is $l(\beta)$ and the log-likelihood function is $L(\beta) = \log\left[l(\beta)\right]$. For many models, $L(\beta)$ has concave shape and $\hat{\beta}$ is the point at which the derivative equals 0. The ML estimate is then the solution of the likelihood equation, $\partial L(\beta)/\partial\beta = 0$. Often β is multidimensional, denoted by $\boldsymbol{\beta}$, and $\hat{\boldsymbol{\beta}}$ is the solution of a set of likelihood equations.

Let SE denote the standard error of $\hat{\boldsymbol{\beta}}$, and let $\text{cov}(\hat{\boldsymbol{\beta}})$ denote the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$. Under regularity conditions, $\text{cov}(\hat{\boldsymbol{\beta}})$ is the inverse of the *information matrix.* The (j,k) element of the information matrix is

$$-E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_j\beta_k}\right). \tag{1}$$

The standard errors are the square roots of the diagonal elements for the inverse information matrix. The greater the curvature of the log likelihood, the smaller the standard errors. This is reasonable, since large curvature implies that the log likelihood drops quickly as $\boldsymbol{\beta}$ moves away from $\hat{\boldsymbol{\beta}}$; hence, the data would have been much more likely to occur if $\boldsymbol{\beta}$ took a value near $\hat{\boldsymbol{\beta}}$ rather than a value far from $\hat{\boldsymbol{\beta}}$.

1.3.2 Likelihood Function and ML Estimate for Binomial Parameter

The part of the likelihood function involving the parameters is called the *kernel.* Since the maximization of the likelihood is with respect to the parameters, the rest is irrelevant.

The binomial log likelihood (ignoring the coefficient $\begin{pmatrix} n \\ y \end{pmatrix}$, which does not depend on the parameter) is then

$$L(\pi) = \log\left[\pi^y(1-\pi)^{n-y}\right] = y\log(\pi) + (n-y)\log(1-\pi). \tag{2}$$

Differentiating with respect to π yields

$$\partial L(\pi)/\partial\pi = y/\pi(n-y)/(1-\pi) - (y-n\pi)/[\pi(1-\pi)]. \tag{3}$$

Equating this to 0 gives the likelihood equation, which has solution $\hat{\pi} = y/n$, the sample proportion of successes for the n trials.

Calculating $\partial^2 L(\pi)/\partial\pi^2$, taking the expectation, and combining terms, we get

$$-E\left[\partial^2 L(\pi)/\partial\pi^2\right] = E\left[y/\pi^2 + (n-y)/(1-\pi)^2\right] = n/\left[\pi(1-\pi)\right]. \tag{4}$$

Thus, the asymptotic variance of $\hat{\pi}$ is $\pi(1-\pi)/n$. This is also expected, since $E(Y) = n\pi$ and $\text{var}(Y) = n\pi(1-\pi)$, $\hat{\pi} = Y/n$ has mean and standard error

$$E(\hat{\pi}) = \pi,\ \sigma(\hat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

1.3.3 Wald-Likelihood Ratio-Score Test Triad

Three standard ways exist to use the likelihood function to perform large-sample inference. We introduce these for a significance test of a null hypothesis $H_0 : \beta = \beta_0$ and then discuss their relation to interval estimation. They all exploit the large-sample normality of ML estimators.

Wald test. With nonnull standard error SE of $\hat{\beta}$, the test statistic

$$z = \left(\hat{\beta} - \beta_0\right)/\text{SE}$$

has an approximate standard normal distribution when $\beta = \beta_0$. This type of statistic, using the nonnull standard error, is called a *Wald statistic*.

The multivariate extension for the Wald test of $H_0: \boldsymbol{\beta} = \boldsymbol{\beta}_0$ has test statistic

$$W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \left[\text{cov}\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

(The prime denotes the transpose.) The nonnull covariance is based on the curvature (1) of the log likelihood evaluated at $\hat{\boldsymbol{\beta}}$. The asymptotic multivariate normal distribution for $\hat{\boldsymbol{\beta}}$ implies an asymptotic chi-squared distribution for W. The df equals the rank of cov($\hat{\boldsymbol{\beta}}$), which is the number of nonredundant parameters in $\boldsymbol{\beta}$.

Likelihood-ratio test. The likelihood ratio test uses the likelihood function through the ratio of two maximizations: 1) the maximum over the possible parameter values under H_0, and 2) the maximum over the larger set of parameter values under $H_0 \cup H_a$, where H_a is an alternative hypothesis.

Let l_0 and l_1 denote the maximized values of the likelihood function under H_0 and $H_0 \cup H_a$, respectively. Wilks (1935, 1938) show that $-2 \log \Lambda$, where $\Lambda = l_0/l_1$, has an asymptotic chi-squared distribution with df = difference in the dimension of the parameter space under $H_0 \cup H_a$ and under H_0, as $n \to \infty$. The *likelihood-ratio test statistic* equals

$$-2 \log \Lambda = -2 \log (l_0/l_1) = -2(L_0 - L_1),$$

where L_0 and L_1 denote the maximized log likelihood functions.

For example, suppose $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)'$ and $H_0 : \ \boldsymbol{\beta}_0 = \mathbf{0}$. Then l_1 is the likelihood function calculated at the $\boldsymbol{\beta}$ value for which the data would have been most likely; l_0 is the likelihood function calculated at the $\boldsymbol{\beta}_1$ value for which the data would have been most likely, assuming $\boldsymbol{\beta}_0 = 0$. Therefore, l_1 is always at least as large as l_0, and the likelihood ratio test statistic is always nonnegative.

Score test. The *score statistic* is due to R. A. Fisher and C. R. Rao. The score test is based on the slope and expected curvature of the log-likelihood function $L(\beta)$ at the null value β_0. It uses the value of the score function

$$u(\beta) = \partial L(\beta)/\partial \beta,$$

evaluated at β_0. The value $u(\beta_0)$ tends to be larger in absolute value when $\hat{\beta}$ is farther from β_0. Denote $-E\left[\partial^2 L(\beta)/\partial \beta^2\right]$ (i.e., the information) evaluated at β_0 by $i(\beta_0)$. The score statistic is the ratio of $u(\beta_0)$ to its null SE, which is $[i(\beta_0)]^{1/2}$. This has an asymptotic standard normal null distribution. The chi-squared form of the score statistic is

$$\frac{[u(\beta_0)]^2}{i(\beta_0)} = \frac{\left[\partial L(\beta)/\partial \beta \mid_{\beta=\beta_0}\right]^2}{-E\left[\partial^2 L(\beta)/\partial \beta^2 \mid_{\beta=\beta_0}\right]}.$$

In the multiparameter case, the score statistic is a quadratic form based on the vector of partial derivatives of the log likelihood with respect to $\boldsymbol{\beta}$ and the inverse information matrix, both evaluated at the ML estimates of $\boldsymbol{\beta}$ under H_0 (a special case of which is $\boldsymbol{\beta} = \boldsymbol{\beta}_0$).

Geometric illustration of the three tests in the univariate case for $H_0: \ \beta = 0$. The Wald test uses the behavior of $L(\beta)$ at the ML estimate $\hat{\beta}$, having chi-squared form $\left[\hat{\beta}/\text{SE}\right]^2$. The score test is based on the slope and curvature of $L(\beta)$ at $\beta = 0$. The likelihood-ratio test combines information about $L(\beta)$ at both $\hat{\beta}$ and $\beta = 0$. It compares the log-likelihood values L_1 at $\hat{\beta}$ and L_0 at $\beta_0 = 0$ using the chi-squared statistic $-2(L_0 - L_1)$. In a sense, this statistic uses the most information of the three types of test statistic.

As $n \to \infty$, the Wald, likelihood-ratio, and score tests have certain asymptotic equivalences. For small to moderate sample sizes, the likelihood-ratio test is usually more reliable than the Wald test.

1.3.4 Constructing Confidence Intervals

In practice, it is more informative to construct confidence intervals for parameters than to test hypotheses about their values. For any of the three test methods, a confidence interval results from inverting the test. For example, a 95% confidence interval for β is the set of β_0 for which the test of $H_0 : \beta = \beta_0$ has a p-value exceeding 0.05.

Let z_α denote the z-score from the standard normal distribution having right-tailed probability α; this is the $100(1-\alpha)$ percentile of that distribution. Let $\chi^2_{df}(\alpha)$ denote the $100(1-\alpha)$ percentile of the chi-squared distribution with degrees of freedom df.

The Wald confidence interval is the set of β_0 for which $|\hat{\beta} - \beta_0|/\text{SE} < z_{\alpha/2}$. This gives the interval $\hat{\beta} \pm z_{\alpha/2}\text{SE}$. The likelihood-ratio-based confidence interval is the set of β_0 for which $-2[L(\beta_0 - L(\hat{\beta})] < \chi^2_1(\alpha)$. [Recall that $\chi^2_1(\alpha) = z^2_{\alpha/2}$.] The score confidence interval can be constructed similarly.

Note that all three tests are based on asymptotic null distributions. In small samples or moderate or large samples when a model contains many parameters, $\hat{\beta}$ may be far from normality. In that case, inference (hypothesis test or confidence interval) made by the Wald and likelihood-ratio tests can be very different. In many cases, an exact small-sample distribution of a test statistic may exist so that we do not have to rely on the large-sample normality to make inference. In other cases, higher-order asymptotic methods may be available.

Despite that the Wald confidence interval is less reliable when the sample size is small to moderate, it is the most commonly used approach in practice, mainly because of its simplicity in construction. It appears that nowadays more and more statistical software packages produce likelihood-ratio-based confidence intervals. The LR-based interval is preferable for categorical data with small to moderate sample sizes.

For the linear regression model assuming a normal response, all three types of tests/confidence intervals are identical.

1.4 Statistical Inference for Binomial Parameters

We illustrate inference methods for categorical data by presenting tests and confidence intervals fro the binomial parameter π, based on y successes in n independent trials.

1.4.1 Tests about a Binomial Parameter

Consider $H_0 : \pi = \pi_0$. Since H_0 has a single parameter, we use the normal rather than chi-squared forms of Wald and score test statistics. This allows us to test against one-sided as well as two-sided alternatives.

The Wald statistic is

$$z_W = \frac{\hat{\pi} - \pi_0}{\text{SE}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}},$$

where $\hat{\pi} = y/n$ is the ML estimate of π.

Evaluating the binomial score (3) and information (4) at π_0 yields

$$u(\pi_0) = \frac{y}{n} - \frac{n-y}{1-\pi_0}, \; i(\pi_0) = \frac{n}{\pi_0(1-\pi_0)}.$$

The normal form of the score statistic is therefore

$$z_S = \frac{u(\pi_0)}{[i(\pi_0)]^{1/2}} = \frac{y - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}. \tag{5}$$

The Wald statistic z_W uses the standard error evaluated at $\hat{\pi}$, the score statistic z_S uses it evaluated at π_0. The score statistic is preferable, as it uses the actual null SE rather than an estimate. Its null sampling distribution is closer to standard normal than that of the Wald statistic.

The binomial log-likelihood function (2) equals $L_0 = y \log \pi_0 + (n-y)\log(1-\pi_0)$ under H_0 and $L_1 = y \log \hat{\pi} + (n-y)\log(1-\hat{\pi})$ under $H_0 \cup H_a$. The likelihood ratio statistic simplifies to

$$-2(L_0 - L_1) = 2\left(y \log \frac{\hat{\pi}}{\pi_0} + (n-y)\log\frac{1-\hat{\pi}}{1-\pi_0}\right).$$

Rewritten as

$$-2(L_0 - L_1) = 2\left(y \log \frac{y}{n\pi_0} + (n-y)\log\frac{n-y}{n-n\pi_0}\right),$$

it compares observed success and failure counts to fitted (i.e., null) counts by

$$2\sum \text{observed } \log \frac{\text{observed}}{\text{fitted}}. \tag{6}$$

We will see that this formula also holds for tests about Poisson and multinomial parameters. (6) has an asymptotic chi-squared distribution with df = 1.

1.4.2 Confidence Intervals for a Binomial Parameter

A significance test merely indicates whether a particular π value (such as $\pi = 0.5$) is plausible. We learn more by using a confidence interval to determine the range of plausible values.

Inverting the Wald test statistic gives the interval of π_0 values for which $|z_W| < z_{\alpha/2}$, or

$$\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}. \tag{7}$$

Unfortunately, the performance of this confidence interval is not satisfactory: 1) It is poor when n is not very large; 2) The coverage probability usually falls below the nominal confidence level; 3) The coverage probability is particularly poor when the true parameter is near 0 or 1. A simple adjustment that adds $1/2z_{\alpha/2}^2$ observations of each type (success and failure) to the sample before using the formula performs much better.

The score confidence interval contains π_0 values for which $|z_S| < z_{\alpha/2}$. Its endpoints are the π_0 solutions to the equations

$$(\hat{\pi} - \pi_0)\sqrt{\pi_0(1-\pi_0)/n} = \pm z_{\alpha/2}.$$

This interval is

$$\hat{\pi}\left(\frac{n}{n+z_{\alpha/2}^2}\right) + \frac{1}{2}\left(\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\right)$$

$$\pm z_{\alpha/2}\sqrt{\frac{1}{n+z_{\alpha/2}^2}\left[\hat{\pi}(1-\hat{\pi})\left(\frac{n}{n+z_{\alpha/2}^2}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\right)\right]}.$$

The midpoint $\tilde{\pi}$ of the interval is $\tilde{\pi} = (y + z^2_{\alpha/2})/(n + z_{\alpha/2})$, a weighted average of $\hat{\pi}$ and $1/2$, with the weight $n/(n + z^2_{\alpha/2})$ for $\hat{\pi}$ increases as n increases. The square of the coefficient of $z_{\alpha/2}$ is a weighted average of the variance of a sample proportion when $\pi = \hat{\pi}$ and the variance of a sample proportion when $\pi = 1/2$, using the adjusted sample size $n + z_\alpha)$ in place of n. This interval has much better performance than the Wald interval.

The likelihood-ratio-based confidence interval is more complex computationally, but simple in principle. It is the set of π_0 for which the likelihood ratio test has a p-value exceeding α. Equivalently, it is the set of π_0 for which double the log likelihood drops by less than $\chi^2_1(\alpha)$ from its value at the ML estimate $\hat{\pi} = y/n$.

In-class Exercise Example. Suppose we tossed a coin 10 times and observed 3 heads and 7 tails. We want to test $H_0 : \pi = 0.5$, where π is the probability of head. We also want to construct 95% confidence intervals for π.

1.5 Statistical Inference for Multinomial Parameters

We consider inference for multinomial parameters $\{\pi_j\}$. We assume of n observations, n_j occur in category j, $j = 1, \dots, c$.

1.5.1 Estimation of Multinomial Parameters

Recall the multinomial probability mass function

$$p(n_1, n_2, \dots, n_c) = \left(\frac{n!}{n_1!n_2! \dots n_c!}\right) \pi_1^{n_1} \pi_2^{n_2} \dots \pi_c^{n_c}. \tag{8}$$

We derive ML estimates of $\{\pi_j\}$. (8) is proportional to the kernel

$$\prod_j \pi_j^{n_j} \text{ where all } \pi_j \geq 0 \text{ and } \sum_j \pi_j = 1. \tag{9}$$

The ML estimates are the $\{\pi_j\}$ that maximize (9).

The multinomial log-likelihood function is

$$L(\boldsymbol{\pi}) = \sum_j n_j \log \pi_j.$$

Since $\sum_{j=1}^c \pi_j = 1$, L is a function of $(\pi_1, \ldots, \pi_{c-1})$. Also, $\partial\pi_c/\partial\pi_j = -1$, $j = 1, \ldots, c-1$.

Since

$$\frac{\partial \log \pi_c}{\partial \pi_j} = \frac{1}{\pi_c}\frac{\partial \pi_c}{\partial \pi_j} = -\frac{1}{\pi_c},$$

differentiating $L(\boldsymbol{\pi})$ with respect to π_j gives the likelihood equation

$$\frac{\partial L(\boldsymbol{\pi})}{\partial \pi_j} = \frac{n_j}{\pi_j} - \frac{n_c}{\pi_c} = 0.$$

The ML solution satisfies $\hat{\pi}_j/\hat{\pi}_c = n_j/n_c$. Since the ML solution should also satisfy $\sum_{j=1}^c \hat{\pi}_j = 1$, we have

$$\sum_j \hat{\pi}_j = 1 = \frac{\hat{\pi}_c \left(\sum_j n_j\right)}{n_c} = \frac{\hat{\pi}_c n}{n_c}.$$

Thus, $\hat{\pi}_c = n_c/n$ and then $\hat{\pi}_j = n_j/n$. Based on further mathematical arguments (not presented here), this solution does maximize the likelihood. Thus, the ML estimates of $\{\pi_j\}$ are the sample proportions.

1.5.2 Pearson Statistics for Testing a Specified Multinomial

Assume a multinomial distribution of $\{n_i : \ i = 1, \ldots, c\}$.

Test $H_0 : \ \pi_i = \pi_{i0}$, where π_{i0} are known, and $\sum \pi_{i0} = \sum \pi_i = 1$.

Pearson test statistic:

$$X^2 = \sum_i \frac{(n_i - m_i)^2}{m_i},$$

where $m_i = n\pi_{i0}$. A statistic of this form is called a *Pearson chi-squared statistic.*

Under H_0, $X^2 \xrightarrow{\mathcal{L}} \mathcal{X}^2_{c-1}$, as $n \to \infty$.

Example: Suppose we throw a dice 50 times, and obtain 10 1's, 8 2's, 5 3's, 6 4's 13 5's, 8 6's. Test whether this dice is fair (i.e., proportion of each of 1-6 is 1/6).

$H_0: \ \pi_1 = \ldots = \pi_6 = 1/6.$

$n_1 = 10$, $n_2 = 8$, $n_3 = 5$, $n_4 = 6$, $n_5 = 13$, $n_6 = 8$, $n = 50$. Under H_0, $m_1 = \ldots = m_6 = 25/3$.

$X^2 = \frac{(10-25/3)^2}{25/3} + \frac{(8-25/3)^2}{25/3} + \frac{(5-25/3)^2}{25/3} + \frac{(6-25/3)^2}{25/3} + \frac{(13-25/3)^2}{25/3} + \frac{(8-25/3)^2}{25/3} = 4.96$. P-value $= .42$. We cannot reject the null hypothesis of $\pi_j = 1/6$ at $\alpha = 0.05$.

1.5.3 Chi-Squared Theoretical Justification

Recall that we assumed a multinomial distribution for $\{n_i : \ i = 1, \ldots, c\}$. Let $\hat{\boldsymbol{\pi}} = (n_1/n, \ldots, n_{c-1}/n)'$ and $\boldsymbol{\pi}_0 = (\pi_{10}, \ldots, \pi_{c-1,0})'$. One can show the following:

1. $\sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) \overset{\mathcal{L}}{\to} N(0, \boldsymbol{\Sigma}_0)$, where element σ_{ij} of $\boldsymbol{\Sigma}_0$ satisfies $\sigma_{ij} = -\pi_i \pi_j$, if $i \neq j$, and $\sigma_{ij} = \pi_i(1-\pi_i)$, if $i = j$ (based on multivariate central limit theorem).

2. $\boldsymbol{\Sigma}_0^{-1}$ has the (i,j)th element equal to $1/\pi_{c0}$ if $i \neq j$, and $(1/\pi_{i0} + 1/\pi_{c0})$ if $i = j$ (verify this by showing $\boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_0^{-1} = \mathbf{I}$).

3. $n(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)' \boldsymbol{\Sigma}_0^{-1} (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)$ simplifies to X^2.

Since $\sqrt{n}\boldsymbol{\Sigma}_0^{-1/2}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)$ has an asymptotic $(c-1)$-dimensional standard multivariate normal distribution, $n(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)' \boldsymbol{\Sigma}_0^{-1} (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)$ (or X^2) has an asymptotic chi-squared distribution with df $= c - 1$.

1.5.4 Likelihood-Ratio Chi-Squared

An alternative test for multinomial parameters uses the likelihood-ratio test. Recall the kernel of the multinomial likelihood (9).

Under H_0 the likelihood is maximized when $\hat{\pi}_j = \pi_{j0}$. In the general case, it is maximized when $\hat{\pi}_j = n_j/n$. The ratio of the likelihoods equals

$$\Lambda = \frac{\prod_j (\pi_{j0})^{n_j}}{\prod_j (n_j/n)^{n_j}}.$$

Thus the likelihood-ratio statistic, denoted by G^2, is

$$G^2 = -2 \log \Lambda = 2 \sum n_j \log \left(n_j / n\pi_{j0}\right). \tag{10}$$

This statistic, which has form (6), is called the *likelihood-ratio chi-squared statistic.* The larger the value of G^2, the greater the evidence against H_0.

In the general case, the parameter space consists of $\{\pi_j\}$ subject to $\sum_j \pi_j = 1$, so the dimensionality is $c - 1$. Under H_0, the $\{\pi_j\}$ are specified completely, so the dimension is 0. The difference in these dimensions equals $c - 1$. For large n, G^2 has a chi-squared null distribution with df $= c - 1$.

When H_0 holds, the Pearson X^2 and the likelihood ratio G^2 both have asymptotic chi-squared distributions with df $= c - 1$. In fact, they are asymptotically equivalent in that case; specifically, $X^2 - G^2$ converges in probability to 0. When H_0 is false, they tend to grow proportionally to n; they need not take similar values, however, even for very large n.

For fixed c, the distribution of X^2 usually converges to chi-squared more quickly than that of G^2. The chi-squared approximation is usually poor for G^2 when $n/c < 5$. When c is large, it can be decent for X^2 for n/c as small as 1 if the table does not contain both very small and moderately large expected frequencies.

1.5.5 Testing with Estimated Expected Frequencies

What if H_0 is not a known set of parameters?

Suppose H_0 involves a small set of (say t) parameters as compared to the full set of $c - 1$ parameters. Then we can use the MLEs of m_i to replace the m_i's in X^2. The resulting test statistic then follows a $\mathcal{X}^2_{c-1-t}$ distribution under H_0.

Example: A sample of 156 dairy calves born in Okeechobee County, Florida, were classified according to whether they caught pneumonia within 60 days after birth. Calves that got a pneumonia infection were also classified according to whether they got a secondary infection within two weeks after the first infection cleared up. Note there would not be a secondary infection if there was not a primary one.

Primary and Secondary Pneumonia Infections of Calves[⋆]

	Secondary Infection	
Primary Infection	Yes	No
Yes	30 (38.1)	60 (39.0)
No	0 (–)	63 (78.9)

⋆ Values in parentheses are estimated expected frequencies.

This is a multinomial sample with positive probabilities for cells (yes, yes), (yes, no) and (no, no) only. The probability for a calf to fall in cell (no, yes) (or (2,1)) is 0. Such a cell is called a *structural zero*.

Probability Structure in the General Case

	Secondary Infection		
Primary Infection	Yes	No	Total
Yes	π_{11}	π_{12}	$\pi_{11} + \pi_{12}$
No	–	π_{22}	π_{22}

where $\pi_{11} + \pi_{12} + \pi_{22} = 1$.

We want to test whether the probability of primary infection was the same as the conditional probability of secondary infection, given that the calf got the primary infection, or test

$$H_0: \ \pi_{11} + \pi_{12} = \pi_{11}/(\pi_{11} + \pi_{12})$$

(equivalent to $\pi_{11} = (\pi_{11} + \pi_{12})^2$). Let $\pi = \pi_{11} + \pi_{12}$, the probability of the primary infection. An alternative notation for the cell probabilities under H_0 therefore is $(\pi^2, \pi(1-\pi), 1-\pi)$.

Probability Structure Under H_0

	Secondary Infection		
Primary Infection	Yes	No	Total
Yes	π^2	$\pi(1-\pi)$	π
No	–	$1-\pi$	$1-\pi$

In order to compute X^2, we need to find the ML estimates of the expected frequencies (or equivalently, the cell probabilities) under H_0.

Sample Version

	Secondary Infection		
Primary Infection	Yes	No	Total
Yes	n_{11}	n_{12}	$n_{11}+n_{12}$
No	–	n_{22}	n_{22}

Kernel of the likelihood:

$$l = (\pi^2)^{n_{11}}(\pi-\pi^2)^{n_{12}}(1-\pi)^{n_{22}}$$

Log likelihood:

$$L = n_{11}\log(\pi^2) + n_{12}\log(\pi-\pi^2) + n_{22}\log(1-\pi).$$

Solving the likelihood equation, we get

$$\hat{\pi} = (2n_{11}+n_{12})/(2n_{11}+2n_{12}+n_{22}).$$

Therefore, the expected frequencies under H_0 are (recall $n = n_{11}+n_{12}+n_{22} = 156$) $\hat{m}_{11} = n\hat{\pi}^2 = 38.1$, $\hat{m}_{12} = n(\hat{\pi}-\hat{\pi}^2)^2 = 39.0$, and $\hat{m}_{22} = n(1-\hat{\pi}) = 78.9$.

One can then compute the Pearson chi-squared statistic for testing H_0, which turns out to be $X^2 = 19.7$ with df $= c-1-t = 3-1-1 = 1$. P-value $< 1e-5$. There is strong evidence against the null hypothesis. The researchers concluded that the primary infection had an immunizing effect that reduced the likelihood of a secondary infection.

2 Describing Contingency Tables

We introduce tables that display relationships between categorical variables. We also define parameters that summarize their association.

2.1 Probability Structure of Contingency Tables

The joint distribution between two categorical variables determines their relationship. This distribution also determines the marginal and conditional distributions.

2.1.1 Joint, Marginal and Conditional Distributions

Let X & Y denote two categorical variables, X having I levels, and Y having J levels. There are IJ possible bivariate outcomes. Because X & Y are discrete, the joint distribution can be displayed in a rectangular table:

	Y			
X	1	$\dots$	J	Total
1	π_{11}	$\dots$	π_{1J}	π_{1+}
	$(\pi_{1\|1})$	$\dots$	$(\pi_{J\|1})$	(1.0)
.	.	$\dots$	.	.
.	.	$\dots$	.	.
.	.	$\dots$	.	.
I	π_{I1}	$\dots$	π_{IJ}	π_{I+}
	$(\pi_{1\|I})$	$\dots$	$(\pi_{J\|I})$	(1.0)
Total	π_{+1}	$\dots$	π_{+J}	1.0

where $\pi_{ij} = P(X = i, Y = j)$, $IJ - 1$ independent parameters.

When the cells contain frequency counts of the outcomes, the table is called a contingency table. A contingency table having I rows and J columns is called an $I \times J$ (I-by-J) table:

	Y			
X	1	...	J	Total
1	n_{11}	...	n_{1J}	n_{1+}
.	.	...	.	.
.	.	...	.	.
.	.	...	.	.
I	n_{I1}	...	n_{IJ}	n_{I+}
Total	n_{+1}	...	n_{+J}	n

Marginal distributions:

$$\pi_{i+} = \sum_{j=1}^{J} \pi_{ij}, \ \pi_{+j} = \sum_{i=1}^{I} \pi_{ij}$$

– row and column totals obtained by summing the joint probabilities. Subscript "+" denotes the sum over the index it replaces.

Conditional distributions:

Often, in a contingency table, one (e.g., Y) is considered a response variable & the other (X) is considered an explanatory variable. In that case, we can define conditional probabilities:

$$\pi_{j|i} = \pi_{ij}/\pi_{i+}, \ \sum_{j=1}^{J} \pi_{j|i} = 1.$$

2.1.2 Independence

$$\pi_{ij} = \pi_{i+}\pi_{+j}, \ i = 1, \ldots, I, j = 1, \ldots, J.$$

There are $I + J - 2 (= (I-1) + (J-1))$ parameters. Also,

$$\pi_{j|i} = \pi_{ij}/\pi_{i+} = \pi_{i+}\pi_{+j}/\pi_{i+} = \pi_{+j}, \ j = 1, \ldots, I, \tag{11}$$

meaning that each conditional distribution of Y is identical to the marginal distribution of Y, or equivalently, probability of column response j is the same in each row. When Y is response, (11) provides a more natural definition of independence.

We use similar notations for the sample distributions:

Cell frequencies (sample joint distribution): $n_{ij}(p_{ij})$

Row totals: $n_{i+}(p_{i+})$

Column totals: $n_{+j}(p_{+j})$

Total: $n = \sum_i \sum_j n_{ij}$

We have $p_{ij} = n_{ij}/n$, $p_{i+} = n_{i+}/n$, $p_{+j} = n_{+j}/n$, $p_{j|i} = n_{ij}/n_{i+}$.

2.1.3 Sensitivity and Specificity in Diagnostic Tests

A diagnostic test is often used to detect whether an individual has certain disease condition. Let X be the true disease status of an individual, with $X = 1$: diseased, $X = 2$: not diseased. Let Y be the outcome of a diagnostic test with $Y = 1$: positive, $Y = 2$: negative.

Important questions of interest include $P(X = 1 \mid Y = 1)$ & $P(X = 2 \mid Y = 2)$. Two important parameters to describe whether a diagnostic test is good or not:

$$P(Y = 1 \mid X = 1) \text{ or } \pi_{1|1} \text{ — sensitivity}$$
$$P(Y = 2 \mid X = 2) \text{ or } \pi_{2|2} \text{ — specificity.}$$

Further let ρ denote the probability that a subject has the disease.

Exercise:

a) Calculate $P(X = 1 \mid Y = 1)$ using Bayes Theorem.

Bayes Theorem: $X \sim P(X)$, $Y \mid X \sim P(Y \mid X)$, then $P(X \mid Y) = P(X)P(Y \mid X)/P(Y)$.

$P(X = 1 \mid Y = 1) = P(X = 1 \,\&\, Y = 1)/P(Y = 1) = \pi_{1|1}\rho/[\pi_{1|1}\rho + \pi_{1|2}(1 - \rho)]$, where $\pi_{1|2} = 1 - \pi_{2|2}$.

b) Suppose $\pi_{1|1} = \pi_{2|2} = 0.95$ and $\rho = 0.005$. Calculate $P(X = 1 \mid Y = 1)$.

$P(X = 1 \mid Y = 1) = .95 * .005/(.95 * .005 + (1 - .95) * (1 - .005)) = .08715596 \approx .087.$

c) Calculate π_{ij}, $i = 1, 2$, $j = 1, 2$, and interpret the result in b).

$\pi_{11} = P(X = 1)\pi_{1|1} = \rho\pi_{1|1} = 0.005 * .95 = .00475;$

$\pi_{12} = \rho - \pi_{11} = .005 - .00475 = .00025;$

$\pi_{21} = P(X = 2)\pi_{1|2} = (1 - \rho)(1 - \pi_{2|2}) = (1 - 0.005) * (1 - .95) = .04975;$

$\pi_{22} = P(X = 2) - \pi_{21} = 1 - \rho - \pi_{21} = 1 - .005 - .04975 = .94525.$

We thus obtain the following table for the joint distribution of X and Y:

	Y		
X	1	2	Total
1	.00475	.00025	.005
	(.95)	(.05)	(1.0)
2	.04975	.94525	.995
	(.05)	(.95)	(1.0)
Total	.545	.9455	1.0

A take-home message from this example: For a rare disease, even if a diagnostic test has both high sensitivity and specificity, the probability of disease for a person who is diagnosed as positive is still low.

2.2 Compare Proportions in 2×2 Tables

2.2.1 Difference of Proportions

$\pi_{1|h} - \pi_{1|i}$ $(= \pi_{2|i} - \pi_{2|h})$ – between-row comparisons. For an $I \times J$ table, independence iff all differences are 0.

If both variables are response variables, we can also compare:

1) $P(\text{row } 1 \mid \text{col } 1) - P(\text{row } 1 \mid \text{col } 2) = \pi_{11}/\pi_{+1} - \pi_{12}/\pi_{+2}$;

2) $P(\text{col } 1 \mid \text{row } 1) - P(\text{col } 1 \mid \text{row } 2) = \pi_{11}/\pi_{1+} - \pi_{21}/\pi_{2+}$;

These two differences of proportion are generally different.

2.2.2 Relative Risk

For 2×2 tables, the relative risk is the ratio $\pi_{1|1}/\pi_{1|2}$.

Independence iff $\pi_{1|1}/\pi_{1|2} = 1$.

Comparison on the second response gives a different relative risk: $\pi_{2|1}/\pi_{2|2} = (1-\pi_{1|1})/(1-\pi_{1|2})$.

2.2.3 Odds Ratio

Again consider a 2×2 table.

Odds 1: $\Omega_1 = \pi_{1|1}/\pi_{2|1}$

Odds 2: $\Omega_2 = \pi_{1|2}/\pi_{2|2}$

Interpretation of $\Omega_1 = 3$: In the first row, the odds of the response being in the first column is three times the odds of the response being in the second column.

Odds ratio: $\theta = \Omega_1/\Omega_2 = \pi_{11}\pi_{22}/\left(\pi_{12}\pi_{21}\right)$.

Assuming all $\pi_{ij} > 0$, then independence iff $\theta = 1$.

If $\theta > 1$, subjects in row 1 are more likely to make response 1 than are subjects in row 2; i.e., $\pi_{1|1} > \pi_{1|2}$ (but note $\theta \neq \pi_{1|1}/\pi_{1|2}$).

If $0 < \theta < 1$, $\pi_{1|1} < \pi_{1|2}$.

If one $\pi_{ij} = 0$, then $\theta = 0$ or ∞.

Properties:

1) Odds ratio θ does not change with the orientation of the table.

2) When the order of the rows or columns is reversed, new θ is the inverse of the original value, which represents the same level of association between X and Y, but in the opposite direction.

Sample version of θ: invariant to multiplication within rows/columns and row/column interchange.

Example: Cross-Classification of Aspirin Use and Myocardial Infarction

	Myocardial Infarction		
	Fatal Attack	Non-Fatal Attack	No Attack
Placebo	18	171	10,845
Aspirin	5	99	10,933

Collapsing fatal and non-fatal attacks:

	Myocardial Infarction	
	Attack	No Attack
Placebo	189	10,845
Aspirin	104	10,933

Sample difference of proportion of attack between aspirin usage: $(18+171)/(18+171+10845)-(5+99)/(5+99+10933) = .0171 - .0094 = .0077$.

The relative risk is $.0171/.0094 = 1.82$.

Interpretation: The proportion suffering heart attacks for those taking placebo was 1.82 times the proportion for those taking aspirin.

The sample odds ratio is

$$\frac{189 \times 10933}{10845 \times 104} = 1.83.$$

Interpretation: The odds of suffering heart attacks for those taking placebo was 1.83 times the odds for those taking aspirin.

2.2.4 Relationship between Relative Risk and Odds Ratio

Odds ratio = relative risk $\times \frac{1-\pi_{1|2}}{1-\pi_{1|1}}$.

When both $\pi_{1|1}$ and $\pi_{1|2}$ are small, odds ratio $\approx$ relative risk (e.g., for rare diseases).

Property: $|\text{odds ratio} - 1| > |\text{relative risk} - 1|$ if X and Y are not independent.

Why?

$$|\mathrm{OR} - 1| = \left|\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} - 1\right| = \left|\frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{12}\pi_{21}}\right|,$$

$$|\mathrm{RR} - 1| = \left|\frac{\pi_{11}/\left(\pi_{11}+\pi_{12}\right)}{\pi_{21}/\left(\pi_{21}+\pi_{22}\right)} - 1\right| = \left|\frac{\pi_{11}\left(\pi_{21}+\pi_{22}\right)}{\left(\pi_{11}+\pi_{12}\right)\pi_{21}} - 1\right| = \left|\frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\left(\pi_{11}+\pi_{12}\right)\pi_{21}}\right|$$

Thus, $|\mathrm{OR} - 1| > |\mathrm{RR} - 1|$ when X and Y are not independent (assuming $\pi_{11} \neq 0$).

2.2.5 Odds Ratio for $I \times J$ Tables

$$\frac{\pi_{ac}\pi_{bd}}{\pi_{ad}\pi_{bc}}$$

$I(I-1)/2 \times J(J-1)/2$ of these, redundant.

A minimal set of odds ratios:

$$\frac{\pi_{i,j}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}}$$

Another minimal set of odds ratios:

$$\frac{\pi_{ij}\pi_{IJ}}{\pi_{Ij}\pi_{iJ}},$$

$i = 1, \ldots, I-1$, $j = 1, \ldots, J-1$.

When π_{i+} and π_{+j}, $i = 1, \ldots, I$, $j = 1, \ldots, J$ are known, the minimal set of odds ratios determine the joint probabilities (and vise versa). In this sense, $(I-1)(J-1)$ parameters can describe any association in an $I \times J$ table (conditional on the marginal probabilities).

2.3 Summary Measures of Association

2×2 vs $I \times J$ tables.

Full joint distribution vs model building vs summary measures of association.

2.3.1 Measures of Ordinal Association

Interval vs ordinal vs nominal variables

Quantitative vs qualitative variables

Interval variables: Pearson correlation

Monotonicity – Does Y tend to increase as X increases?

Concordant vs discordant pairs in a contingency table:

Concordant, if the subject ranking higher on variable X also ranks higher on variable Y.

Discordant, if the subject ranking higher on variable X ranks lower on variable Y.

Tied, if the subjects have the same classification (or ranking) on X and/or Y.

Lymph node localization example: A study of lymphatic mapping and localization of lymph nodes using fluorescence sodium vs technetium Tc 99m sulfur colloid radioactivity. Concordance is the primary efficacy endpoint of the study.

Job satisfaction example

Cross-classification of job satisfaction by income

	Job Satisfaction			
	Very	Little	Moderately	Very
Income (US$)	Dissatisfied	Dissatisfied	Satisfied	Satisfied
< 6000	20	24	80	82
$6000 - 15000$	22	38	104	125
$15000 - 25000$	13	28	81	113
> 25000	7	18	54	92

Concordant, discordant pairs in this example.

Total # of concordant pairs, denoted as C, equals $20 \times (38 + 104 + 125 + 28 + \ldots + 92) + 24 \times (104 + 125 + \ldots + 92) + \ldots = 109,520$.

Total # of discordant pairs, denoted as D, equals $24 \times (22 + 13 + 7) + 80 \times (22 + 38 + 13 + \ldots + 18) + \ldots = 84,915$.

$C > D$ suggests a tendency for low income to occur with low job satisfaction and high income with high job satisfaction.

Can we calculate the probability of concordance and discordance for two independent observations?

$$\Pi_c = 2\sum_i \sum_j \pi_{ij}(\sum_{h>i} \sum_{k>j} \pi_{hk})$$
$$\Pi_d = 2\sum_i \sum_j \pi_{ij}(\sum_{h>i} \sum_{k<j} \pi_{hk})$$

Several measures of association utilize the difference $\Pi_c - \Pi_d$. If $\Pi_c > \Pi_d$, association positive; if $\Pi_c < \Pi_d$, association negative.

2.3.2 Gamma

For a pair of subjects, define

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d}.$$

What is the interpretation of γ?

P(concordance | the pair is untied) - P(discordance | the pair is untied)

Sample version of gamma: $\hat{\gamma} = \frac{C-D}{C+D}$.

For 2×2 tables, γ simplifies to

$$Q = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{11}\pi_{22} + \pi_{12}\pi_{21}} = \frac{\theta - 1}{\theta + 1},$$

where θ is the OR.

2.3.3 Measures of Nominal Association

For interval variables, R-squared, intraclass correlation coefficient, etc., describe the proportional reduction in variance from the marginal distribution to the conditional distributions of the response.

$V(Y)$: a measure of variation for the marginal distribution $\{\pi_{+1}, \ldots, \pi_{+J}\}$ of the response Y

$V(Y \mid i)$: the same measure computed for the conditional distribution $\{\pi_{+1|i}, \ldots, \pi_{+J|i}\}$ at the ith setting of an explanatory variable X. For a categorical X, $E[V(Y \mid X)] = \sum_i \pi_{i+} V(Y \mid i)$.

A measure of proportional reduction in variation:

$$\frac{V(Y) - E[V(Y \mid X)]}{V(Y)}$$

2.3.4 Concentration and Uncertainty Measures

One variation measure: $V(Y) = \sum \pi_{+j}(1 - \pi_{+j}) = 1 - \sum \pi_{+j}^2$.

Minimum: 0, when $\pi_{+j} = 1$ for some j.

Maximum: $(J-1)/J$ when $\pi_{+j} = 1/J$ for all j.

The conditional variation in row i is $V(Y \mid i) = 1 - \sum \pi_{+j|i}^2$.

For an $I \times J$ table,

$$E[V(Y \mid X)] = 1 - \sum_i \pi_{i+} \sum_j \pi_{j|i}^2 = 1 - \sum \sum \pi_{ij}^2 \pi_{i+}.$$

The proportional reduction in variation is *Goodman and Kruskal's tau*

$$\tau = \frac{\sum_i \sum_j \pi_{ij}^2/\pi_{i+} - \sum_j \pi_{+j}^2}{1 - \sum_j \pi_{+j}^2},$$

also called the *concentration coefficient*.

Interpretation of τ:

Proportional prediction rule: with probability π_{+j} guess response to be in category j. Then

$$V(Y) = \sum \pi_{+j}(1 - \pi_{+j}) = P(\text{guess } j) \times P(\text{wrong guess} \mid \text{guess } j)$$

is the probability of an incorrect guess.

$V(Y \mid i)$: the probability of an incorrect guess given we know that a subject is in category i of X.

$E[V(Y \mid X)]$: averaged over the distribution of X

A large τ represents a strong association, in the sense that we can guess Y much better when we know X than when we do not.

An alternative variation measure (Theil, 1970): $V(Y) = \sum \pi_{+j} \log \pi_{+j}$, called the *entropy*. For contingency tables, this results in the proportional reduction in variation index

$$U = -\frac{\sum_i \sum_j \pi_{ij} \log(\pi_{ij}/\pi_{i+}\pi_{+j})}{\sum_j \pi_{+j} \log \pi_{+j}},$$

called the *uncertainty coefficient*.

Properties:

1) Both τ and U are well defined as long as some $\pi_{+j} > 0$;

2) $0 \le \tau \le 1$, $0 \le U \le 1$

3) $\tau = U = 0$ is equivalent to independence of X and Y;

4) $\tau = U = 1$ is equivalent to no conditional variation, in the sense that for each i, $\pi_{j|i} = 1$ for some j.

Variation measure used in τ: *Gini concentration*

Variation measure used in U: *entropy*

Note: τ and U tend to decrease with an increase of the number of response categories. But in general, a large value constitutes a "strong" association.

Religious identification example

Religious Identification Now and at Age 16

Religious Identification at Age 16	Current Religious Identification				
	Protestant	Catholic	Jewish	None or other	Total
Protestant	918	27	1	70	1016
Catholic	30	351	0	37	418
Jewish	1	1	28	1	31
None or other	29	5	0	25	59
Total	978	384	29	133	1524

The sample version of Goodman and Kruskal's tau equals 0.57.

The sample version of the uncertainty coefficient is 0.51.

There seems to be relatively strong association between religious identification now and at age 16.

3 Inference for Two-Way Contingency Tables

Most inferential methods for categorical data assume multinomial or Poisson sampling models.

3.1 Sampling Distributions

Suppose $\{n_i,\ i = 1, \ldots, c\}$ are observed counts in the c cells of a contingency table (for an $I \times J$ table, $c = IJ$). Denote by $m_i = E(n_i)$ the corresponding expected frequencies.

3.1.1 Poisson Sampling

The probability mass function

$$\frac{\exp(-m_i)m_i^{n_i}}{n_i!},\ n_i = 0, 1, 2, \ldots$$

$\mathrm{Var}(n_i) = E(n_i) = m_i$.

The counts observed in different cells are assumed independent.

For example, $n_1 = \#$ of spontaneous abortions, $n_2 = \#$ of induced abortions, $n_3 = \#$ of live births, measured in November, 1990 in London, England.

3.1.2 Multinomial Sampling

Poisson sampling assumes the total number $n = \sum_i n_i$ is random. If we condition on the total number n, then $[n_i \mid n]$ is not a Poisson distribution any more.

$$P(n_i \text{ observations in cell } i,\ i = 1, \ldots, c \mid \sum n_j = n)$$

$$= \frac{n_i \text{ observations in cell } i,\ i = 1, \ldots, c}{P(\sum n_j = n)}$$

$$= \frac{\Pi_i[\exp(-m_i)m_i^{n_i}/n_i!]}{\exp(-\sum m_j)(\sum m_j)^n/n!} = \frac{n!}{\Pi_i n_i!}\Pi_i \pi_i^{n_i}, \tag{12}$$

where $\pi_i = m_i/(\sum m_j)$, $i = 1, \ldots, c$. This is the multinomial $(n, \{\pi_i\})$ distribution.

Alternatively, if the n observations are independent and each has a probability of π_i falling in category i of the c categories, then $\{n_i\}$ follow the same distribution (12).

3.1.3 Independent Multinomial Sampling

Suppose we take observations on a categorical response variable Y, separately at various settings of an explanatory variable X. Let n_{ij} denote the # of observations in the jth response category, at the ith setting of X. Suppose the n_{i+} observations on Y at the ith setting of X are independent, each having probability distribution $\{\pi_{1|i}, \ldots, \pi_{J|i}\}$. Then the counts $\{n_{ij},\ j = 1, \ldots, J\}$ have the multinomial distribution

$$\frac{n_{i+}!}{\Pi_j n_{ij}!} \Pi_j \pi_{j|i}^{n_{ij}}. \tag{13}$$

When samples at different settings of X are independent, the joint probability function for the entire dataset is the product of (13) from the various settings. This sampling scheme is called independent multinomial sampling, sometimes also called product multinomial sampling.

Another scenario where independent multinomial sampling arises: Suppose $\{n_{ij}\}$ follow either independent Poisson sampling with means $\{m_{ij}\}$, or multinomial sampling with probabilities $\{\pi_{ij} = m_{ij}/n\}$. When X is an explanatory variable, it is sensible to perform statistical inference conditional on the totals $\{n_{i+}\}$, even when their values are not fixed by the sampling design. When we condition on $\{n_{i+}\}$, the cell counts $\{n_{ij},\ j = 1, \ldots, J\}$ have the multinomial distribution (13) with response probabilities $\{\pi_{j|i} = m_{ij}/m_{i+}\}$, and the cell counts from different rows are independent.

In *prospective* studies, $\{n_{i+}\}$ for X are often fixed, and we regard each row of J counts as an independent multinomial sample on Y.

In *retrospective* studies, the totals $\{n_{+j}\}$ for Y are often fixed, and we regard each column of I counts as an independent multinomial sample on X.

In *cross-sectional* studies, the total sample size is fixed, but not the row or column totals, and we regard the IJ cell counts as a multinomial sample.

Physical activity example

Physical Activity and Quality of Life (QOL) for Cancer Survivors

and People at Increased Risk

	Distress status at first-time visit		
Physical activity	Distressed	Non-distressed	Total
Physically active	15	20	35
Sedentary	45	20	65
Total	60	40	100

Scenario 1: Survey all visitors (seeking counseling or treatment) at the QOL Shared Resources at MD Anderson Cancer Center in the next year. Classify them into four categories, i.e., physically active vs sedentary + distressed vs non-distressed, at the first time of their visit to the Shared Resources. In this scenario, the total # of people visiting in the next year will be a random variable. Therefore, we can treat it as a Poisson random variable. The sampling is Poisson sampling with unknown expected frequencies $\{m_{11}, m_{12}, m_{21}, m_{22}\}$.

Scenario 2: Survey a random sample of 200 people who visited the Shared Resources in the past year. Classify each subject into one of the above four categories. This is a multinomial sampling with a total sample size of 200 and unknown cell probabilities $\{\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\}$.

Scenario 3: Suppose the data for the distressed and non-distressed people visiting the Shared Resources have been kept separately in some way so that we can randomly sample, e.g., 100 distressed people, and separately randomly sample 100 non-distressed people, and classify them according to their physical activity status. Now the column totals are fixed. This is independent binomial sampling within each column.

Scenario 4: Treat both row and column totals as fixed. In the 2 × 2 table case, only one cell count will be random. Hypergeometric sampling distribution – Fisher's exact test, which we will talk about later.

3.1.4 Likelihood Functions and MLEs

Assume multinomial sampling, recall:

$$l \propto \Pi_{i=1}^{c} \pi_i^{n_i}$$

$$L = \sum_{i=1}^{c} n_i \log \pi_i$$

Solving the likelihood equation (letting the first derivative of L equal 0) results in: $\hat{\pi}_i = n_i/n$, $i = 1, \ldots, c$. It can be shown that the sample counts are minimal sufficient statistics. Birch (1963) showed that such estimates are MLEs.

Conclusions:

1) For contingency tables, the MLEs of cell probabilities are the sample cell proportions;

2) The MLEs of marginal probabilities are the sample marginal proportions;

3) If two categorical variables are independent, i.e., $\pi_{ij} = \pi_{i+}\pi_{+j}$, then the MLE of π_{ij} is $\hat{\pi}_{ij} = p_{i+}p_{+j} = n_{i+}n_{+j}/n^2$.

For multinomial sampling of size n over IJ cells of a two-way table, $n_{ij} \sim B(n, \pi_{ij})$. The MLE of m_{ij}, or $\hat{m}_{ij} = n\hat{\pi}_{ij}$. Under independence, $\hat{m}_{ij} = np_{i+}p_{+j} = n_{i+}n_{+j}/n$. These are called estimated frequencies, and will be used in the test of independence later.

Many analyses yield the same estimation results for Poisson, multinomial, or independent multinomial sampling schemes, because of the similarity in the likelihood functions.

3.2 Inference for Odds Ratios, Difference of Proportions and Relative Risks

3.2.1 Delta Method

1) *Random variable case*:

Suppose $\sqrt{n}(T_n - \theta) \overset{\mathcal{L}}{\rightarrow} N(0, \sigma^2)$, g twice differentiable at θ.

Taylor expansion:

$$g(t) = g(\theta) + (t-\theta)g'(\theta) + (t-\theta)^2 g''(\theta^\star)/2,\ \ ((t-\theta)^2 g''(\theta^\star)/2 = \mathrm{O}(|t-\theta|^2)).$$

Therefore,

$$\sqrt{n}[g(T_n) - g(\theta)] = \sqrt{n}(T_n - \theta)g'(\theta) + \sqrt{n}\mathrm{O}(|T_n - \theta|^2) \text{ (note here } \sqrt{n}\mathrm{O}(|T_n-\theta|^2) = \mathrm{O}_p(n^{-1/2})),$$

and $\sqrt{n}[g(T_n) - g(\theta)] \overset{\mathcal{L}}{\to} N(0, \sigma^2[g'(\theta)]^2)$. Here $\mathrm{O}_p(z_n)$ denotes a random variable such that for every $\epsilon > 0$, there is a constant K and an integer n_0 such that $P\left[\mathrm{O}_p(z_n)/z_n < K\right] > 1 - \epsilon$ for all $n > n_0$.

2) *Random vector case*:

Let $\boldsymbol{T}_n = (T_{n1}, \ldots, T_{nc})'$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_c)'$. Suppose $\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}) \overset{\mathcal{L}}{\to} N(0, \boldsymbol{\Sigma})$. Suppose $g(t_1, \ldots, t_c)$ has a nonzero differential $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_c)'$ at $\boldsymbol{\theta}$, where $\phi_i = \partial g/\partial t_i \mid_{\boldsymbol{t}=\boldsymbol{\theta}}$. Then

$$\sqrt{n}[g(\boldsymbol{T}_n) - g(\boldsymbol{\theta})] \overset{\mathcal{L}}{\to} N(0, \boldsymbol{\phi}'\boldsymbol{\Sigma}\boldsymbol{\phi}).$$

3.2.2 Odds Ratio

In 2 × 2 table,

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

A continuity correction:

$$\tilde{\theta} = \frac{(n_{11}+0.5)(n_{22}+0.5)}{(n_{12}+0.5)(n_{21}+0.5)}.$$

$\hat{\theta}$ and $\tilde{\theta}$ have the same asymptotic normal distribution around θ.

$$\hat{\sigma}(\log\hat{\theta}) = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)^{1/2} \tag{14}$$

Average $100(1-\alpha)\%$ CI for $\log\theta$:

$$\log\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}(\log\hat{\theta}).$$

Similarly, replacing $\{n_{ij}\}$ by $\{n_{ij}+0.5\}$ improves the asymptotic standard error (ASE) accordingly.

Derivation of (14):

Suppose sample counts $\{n_i,\ i=1,\ldots,c\}$ follow a multinomial $(n,\{\pi_i\})$ distribution. Denote the sample proportion as p_i. We have $Ep_i=\pi_i$, $E(p_i-\pi_i)^2=\pi_i(1-\pi_i)/n$. $(p_1,\ldots,p_{c-1})$ have a large-sample multivariate normal distribution, or $\boldsymbol{p}=(p_1,\ldots,p_c)$ have a multivariate normal distribution with a singular covariance matrix. In fact,

$$\sqrt{n}(\boldsymbol{p}-\boldsymbol{\pi})\overset{\mathcal{L}}{\to}N(0,Diag(\boldsymbol{\pi})-\boldsymbol{\pi}\boldsymbol{\pi}'),$$

where $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_c)'$.

Let ξ be a differentiable function of $\{\pi_i\}$ and $\hat{\xi}$ be the corresponding sample value. Let $\phi_i=\partial\xi/\partial\pi_i$, $i=1,\ldots,c$, then

$$\sqrt{n}(\hat{\xi}-\xi)/\sigma\overset{\mathcal{L}}{\to}N(0,1),$$

where $\sigma^2=\sum\pi_i\phi_i^2-(\sum\pi_i\phi_i)^2$.

A large sample CI for ξ takes the form $\hat{\xi}\pm z_{\alpha/2}\hat{\sigma}/\sqrt{n}$.

Now we have $\xi=\log\theta=\log\pi_{11}+\log\pi_{22}-\log\pi_{12}-\log\pi_{21}$. We then have $\phi_{11}=\partial\log\theta/\partial\pi_{11}=1/\pi_{11}$, similarly, $\phi_{12}=-1/\pi_{12}$, $\phi_{21}=-1/\pi_{21}$, and $\phi_{22}=1/\pi_{22}$.

Since $\sum\sum\pi_{ij}\phi_{ij}=0$, $\sigma^2=\sum\sum\pi_{ij}\phi_{ij}^2=\sum\sum 1/\pi_{ij}$. Therefore, the ASE for $\log\hat{\theta}$ is $(\sum\sum 1/n_{ij})^{1/2}$.

3.2.3 Difference of Proportions and Relative Risk

Again consider a 2×2 table. We want to calculate CIs for the difference in proportions and relative risk, e.g., with a difference in proportion defined as $\pi_{1|1}-\pi_{1|2}$, and a relative risk defined as $\pi_{1|1}/\pi_{1|2}$. For these measures, we treat the rows as independent binomial samples.

Difference of proportions:

To estimate $\pi_{1|1}-\pi_{1|2}$, we use $p_{1|1}-p_{1|2}=n_{11}/n_{1+}-n_{21}/n_{2+}$.

$E(p_{1|1} - p_{1|2}) = \pi_{1|1} - \pi_{1|2}$

SD:

$$\sigma(p_{1|1} - p_{1|2}) = \left[\frac{\pi_{1|1}(1-\pi_{1|1})}{n_{1+}} + \frac{\pi_{1|2}(1-\pi_{1|2})}{n_{2+}}\right]^{1/2}$$

Estimate $\sigma(p_{1|1} - p_{1|2})$ by

$$\left[\frac{p_{1|1}(1-p_{1|1})}{n_{1+}} + \frac{p_{1|2}(1-p_{1|2})}{n_{2+}}\right]^{1/2}$$

Then a $100(1-\alpha)\%$ CI can be constructed by $(p_{1|1} - p_{1|2}) \pm z_{\alpha/2}\hat{\sigma}(p_{1|1} - p_{1|2})$.

No continuity correction is necessary assuming row totals are greater than 0.

Relative risk:

Sample relative risk is $r = p_{1|1}/p_{1|2}$. ASD of $\log r$ is

$$\sigma(\log r) = \left(\frac{1-\pi_{1|1}}{\pi_{1|1}n_{1+}} + \frac{1-\pi_{1|2}}{\pi_{1|2}n_{2+}}\right)^{1/2}.$$

(keep in mind that $\sqrt{n}(p_{1|1} - \pi_{1|1}) \xrightarrow{\mathcal{L}} N(0, \pi_{1|1}(1-\pi_{1|1}))$ We have $\sigma(\log p_{1|1}) = (1-\pi_{1|1})/(\pi_{1|1}n_{1+})$, similarly, $\sigma(\log p_{1|2}) = (1-\pi_{1|2})/(\pi_{1|2}n_{2+})$. The ASE of $\log r$ is

$$\hat{\sigma}(\log r) = \left(\frac{1-p_{1|1}}{p_{1|1}n_{1+}} + \frac{1-p_{1|2}}{p_{1|2}n_{2+}}\right)^{1/2}.$$

When $p_{1|1}$ and/or $p_{1|2}$ equal 0, both $\log r$ and the sample version of the SE are undefined.

A less biased estimator of the log relative risk is

$$\log \tilde{r} = \log\left(\frac{n_{11}+1/2}{n_{1+}+1/2}\right) - \log\left(\frac{n_{21}+1/2}{n_{2+}+1/2}\right),$$

and a corresponding CI is

$$\log \tilde{r} \pm z_{\alpha/2}\left[\frac{1}{n_{11}+1/2} - \frac{1}{n_{1+}+1/2} + \frac{1}{n_{21}+1/2} - \frac{1}{n_{2+}+1/2}\right]^{1/2}$$

.

Exponentiating endpoints gives CI for the relative risk r.

Physical activity example (revisited)

Suppose we randomly sampled 100 visitors and classified them in the 2 × 2 table below:

Physical Activity and Quality of Life (QOL) for Cancer Survivors

and People at Increase Risk

	Distress status at first-time visit		
Physical activity	Distressed	Non-distressed	Total
Physically active	15	20	35
Sedentary	45	20	65
Total	60	40	100

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{15 \times 20}{45 \times 20} = .33$$

$$\tilde{\theta} = \frac{(n_{11}+.5)(n_{22}+.5)}{(n_{12}+.5)(n_{21}+.5)} = \frac{(15+.5)\times(20+.5)}{(45+.5)\times(20+.5)} = .34$$

$\hat{\theta}$ and $\tilde{\theta}$ are close because no cell count is small.

$\log\hat{\theta} = -1.099$. ASE $= (1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22})^{1/2}$. Therefore, a 95% CI for $\log\theta$ is $-1.099 \pm 1.96 \times .435 = (-1.952, -.246)$. A 95% CI for the odds ratio θ is the $(e^{-1.952}, e^{-.246})$ $= (.142, .782)$.

Now suppose the sampling was done by independent binomial sampling for the two rows. That is, we randomly sampled 35 physically active and 65 sedentary people, and classified them based on their distress status.

The difference of proportions between two rows is $p_{1|1} - p_{1|2} = 15/(15+20) - 45/(45+20) = -.264$. The SE is

$$\hat{\sigma}(p_{1|1}-p_{1|2}) = \left[\frac{p_{1|1}(1-p_{1|1})}{n_{1+}} + \frac{p_{1|2}(1-p_{1|2})}{n_{2+}}\right]^{1/2} = \left(\frac{15/35 \times 20/35}{35} + \frac{45/65 \times 20/65}{65}\right)^{1/2} = .101.$$

Therefore, a 95% CI for $\pi_{1|1} - \pi_{1|2}$ is $-.264 \pm 1.96 \times .101 = (-.462, -.066)$.

For estimating relative risk,

$$\log r = \log(p_{1|1}/p_{1|2}) = \log \frac{15/35}{45/65} = -.480$$

ASE:

$$\hat{\sigma}(\log r) = \left(\frac{1 - p_{1|1}}{p_{1|1} n_{1+}} + \frac{1 - p_{1|2}}{p_{1|2} n_{2+}} \right)^{1/2}$$

$$= \left(\frac{1}{n_{11}} - \frac{1}{n_{1+}} + \frac{1}{n_{21}} - \frac{1}{n_{2+}} \right)^{1/2} = \left(\frac{1}{15} - \frac{1}{35} + \frac{1}{45} - \frac{1}{65} \right)^{1/2} = .212.$$

Therefore, a 95% CI for $\pi_{1|1}/\pi_{1|2}$ is $(.408, .938)$.

3.3 Testing Independence

$H_0 : \ \pi_{ij} = \pi_{i+}\pi_{+j}$

We could use the Pearson X^2 statistic with n_{ij} in place of n_i and $m_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$ in place of m_i. $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ are usually unknown.

3.3.1 Pearson Chi-Squared Test

In $I \times J$ two-way contingency tables,

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}},$$

where $\hat{m}_{ij} = np_{i+}p_{+j}$. Under H_0, $X^2 \xrightarrow{\mathcal{L}} \mathcal{X}^2_{(I-1)(J-1)}$, as $n \to \infty$.

Why is $df = (I-1)(J-1)$?

$c = IJ$, $t = (I-1) + (J-1)$. So $c - 1 - t = IJ - 1 - (I-1) - (J-1) = (I-1)(J-1)$.

3.3.2 Likelihood-Ratio Chi-Squared

Let Λ be the ratio of the maximized likelihood under H_0 and under $H_0 \cup H_a$. Then

$$-2\log\Lambda \sim \mathcal{X}^2_{df}, \text{ as } n \to \infty,$$

where *df* = difference in # of parameters between H_0 and $H_0 \cup H_a$.

Assume multinomial sampling, kernel of likelihood:

$$\prod_i \prod_j \pi_{ij}^{n_{ij}}, \ \pi_{ij} \geq 0, \ \sum_i \sum_j \pi_{ij} = 1.$$

Under H_0: MLEs are $\hat{\pi}_{i+} = n_{i+}/n$, $\hat{\pi}_{+j} = n_{+j}/n$, so that $\hat{\pi}_{ij} = n_{i+}n_{+j}/n^2$.

Under the general case (i.e., $H_0 \cup H_a$), $\hat{\pi}_{ij} = n_{ij}/n$.

Then we have

$$\Lambda = \frac{\prod_i \prod_j (n_{i+}n_{+j})^{n_{ij}}}{n^n \prod_i \prod_j n_{ij}^{n_{ij}}},$$

and

$$G^2 = -2 \log \Lambda = 2 \sum_i \sum_j n_{ij} \log(n_{ij}/\hat{m}_{ij}).$$

Under H_0, as $n \to \infty$, $G^2 \overset{\mathcal{L}}{\to} \mathcal{X}^2_{df}$, where *df* = difference in # of parameters between H_0 and $H_0 \cup H_a = (I-1)(J-1)$.

Comparison between X^2 and G^2: They test the same hypothesis, same *df*, perform similarly when n is large. However, X^2 seems to be more robust in small samples.

Practical conditions: G^2 is poor if $n/(IJ) < 5$, while in certain cases of these X^2 may still perform ok.

Physical activity example (revisited) (assuming multinomial sampling):

Physical Activity and Quality of Life (QOL) for Cancer Survivors

and People at Increase Risk

	Distress status at first-time visit		
Physical activity	Distressed	Non-distressed	Total
Physically active	15 (21)	20 (14)	35
Sedentary	45 (39)	20 (26)	65
Total	60	40	100

$X^2 = (15-21)^2/21 + (20-14)^2/14 + (45-39)^2/39 + (20-26)^2/26 = 6.59$. With a null distribution of $\mathcal{X}_1^2$ ($df = (I-1)(J-1) = 1$), p-value = 0.01.

$G^2 = 2\sum_i \sum_j n_{ij} \log(n_{ij}/\hat{m}_{ij}) = 2[15\log(15/21)+20\log(20/14)+45\log(45/39)+20\log(20/26)] = 6.56$. With a null distribution of $\mathcal{X}_1^2$, p-value = 0.01.

3.3.3 Invariance of Chi-Squared to Category Orderings

The $\{\hat{m}_{ij} = n_{i+}n_{+j}/n\}$ used in X^2 and G^2 depend on the row and column marginal totals, but not on the order in which the rows and columns are listed. Thus, X^2 and G^2 do not change under permutations of rows or columns. Both row and column variables are treated as nominal scales in the tests. We essentially ignore some information when we use X^2 and G^2 to test independence between ordinal scales.

When at least one variable is ordinal, it is usually possible to construct more powerful tests of independence, which will be discussed later in this course.

3.3.4 Partitioning Chi-Squared

Some known facts: 1) A chi-squared random variable with $df = \nu$ has representation $Z_1^2+\ldots+Z_\nu^2$, where $Z_1, \ldots, Z_\nu$ are independent $N(0,1)$ random variables; 2) If X_1^2 and X_2^2 are independent random variables having chi-squared distributions with *df* ν_1 and ν_2, respectively, then $X^2 = X_1^2 + X_2^2$ has a chi-squared distribution with *df* $= \nu_1 + \nu_2$; 3) Conversely, a chi-squared statistic having *df* $= \nu$ has partitionings into independent chi-squared components. For example, it can be partitioned into ν components each having *df* $= 1$.

Partitioning chi-squared statistics for testing independence may help reveal certain aspects of the association between two variables. For example, it may help show that an association primarily reflects differences between certain categories or groupings of categories.

Consider a $2 \times J$ table:

	Y		
X	1	...	J
1	n_{11}	...	n_{1J}
2	n_{21}	...	n_{2J}

A simple partitioning of G^2 with $J-1$ components: The jth component is G^2 for testing independence in a 2×2 table, where the first column combines columns 1 through j of the original table, and the second column is column $j+1$, $j = 1, \ldots, J-1$. Each statistic has a single df.

A natural alternative seems to be partitioning G^2 based on $(J-1)$ 2×2 tables obtained by pairing each column with a particular one, say the last. However, these statistics are not asymptotically independent, and their sum does not equal G^2 for testing independence in the full table.

Now consider an $I \times J$ table and two potential partitions that result in asymptotically independent chi-squared components.

1: Compare columns 1 and 2, then combine columns 1 and 2 and compare them to column 3, and so on. Each of the $J-1$ statistics has *df* $= I - 1$.

2: A more refined partition. One example:

$$\begin{array}{cc} \sum_{a<i} \sum_{b<j} n_{ab} & \sum_{a<i} n_{aj} \\ \sum_{b<j} n_{ib} & n_{ij} \end{array}$$

for $i = 2, \ldots, I$ and $j = 2, \ldots, J$. This generates $(I-1)(J-1)$ chi-squared statistics, each having $df = 1$.

An origin of schizophrenia example:

The data classify a sample of psychiatrists by their school of psychiatric thought and by their opinion on the origin of schizophrenia.

Most Influential School of Psychiatric Thought and Ascribed Origin of Schizophrenia

	Origin of Schizophrenia		
School of Psychiatric Thought	Biogenic	Environmental	Combination
Eclectic	90	12	78
Medical	13	1	6
Psychoanalytic	19	13	50

An overall test of independence gives $G^2 = 23.04$ with $df = 4$.

We can better understand this association by partitioning G^2 into four independent components.

Subtables Used in Partitioning Chi-Squared in the Origin of Schizophrenia Example

	Bio	Env		Bio + Env	Com		Bio	Env		Bio + Env	Com
Ecl	90	12	Ecl	102	78	Ecl + Med	103	13	Ecl + Med	116	84
Med	13	1	Med	14	6	Psy	19	13	Psy	32	50

Subtable 1: Comparing eclectic and medical schools of psychiatric thought on whether the origin of schizophrenia is biogenic or environmental (assuming the opinion is either biogenic or environmental). $G^2 = 0.29$ with $df = 1$.

Subtable 2: Compare ecl and med schools in terms of proportion of bio and env combo or either. $G^2 = 1.36$ with $df = 1$.

Sum of these two G^2 equals the G^2 for the independence model applied to the first two rows of the full table, i.e., $G^2 = 1.65$ with $df = 2$. There is little evidence of a difference in thought on the ascribed origin of schizophrenia between eclectic and medical schools.

Subtable 3: Combine eclectic and medical schools and compare them to the psychoanalytic school (bio vs env). $G^2 = 12.95$ with $df = 1$.

Subtable 4: Same as subtable 3, except to compare proportions of ascribing the origin to bio + env vs com. $G^2 = 8.43$ with $df = 1$.

The psychoanalytic school seems more likely than the other schools to ascribe the origins of schizophrenia as being a combination. Among those who chose either the biogenic or environmental origin, members from the psychoanalytic school were more likely than the other schools to choose the environmental origin.

Sum of the four G^2 equals 23.04, the value for testing the overall independence in the original table.

3.3.5 Rules for Partitioning

Necessary conditions for independent partitioning:

1. Dfs for subtables sum to the df for the original table.
2. Each cell count in the original table must be a cell count in one and only one subtable.
3. Each marginal total of the original table must be a marginal total for one and only one subtable.

To check empirically, see if the sum of G^2 for subtables equals that for the original table.

X^2 does not equal the sum of X^2 for separate tables. However, asymptotically X^2 is equivalent to G^2. When there are small sample counts, safer to use X^2 to study the component tables.

3.4 Exact Test of Independence for Small Samples

We have talked about inferences (i.e., calculating confidence intervals) for odds ratios, difference of proportions, and relative risk, and Pearson's and likelihood ratio chi-squared tests. All these inferences are based on large samples. As $n \to \infty$, the multinomial distribution for $\{n_i\}$ is better approximated by a multivariate normal distribution, and X^2 and G^2 have nearly chi-squared distributions.

What if the sample size is small? For example, in a two-way table, say for certain (i, j), $n_{ij} < 5$? The large-sample Pearson's chi-squared and likelihood ratio tests may not be trustworthy in these situations.

We can try to find the exact distribution of the cell counts under the null hypothesis of independence, instead of resorting to a large sample approximation by a multivariate normal distribution.

3.4.1 Fisher's Exact Test

Consider a 2×2 table:

	Y	
X	1	2
1	n_{11}	n_{12}
2	n_{21}	n_{22}

While the unconditional joint distribution of the cell counts is a function of two unknown parameters (i.e., one marginal row probability and one marginal column probability) under the null hypothesis of independence, the joint distribution will be a distribution free of any unknown parameters conditioning on *both* the row and column totals.

Assuming Poisson, multinomial, or independent multinomial sampling, conditioning on the observed marginal totals, the distribution of n_{11}, which determines all three other cell counts, is a hypergeometric distribution as follows:

$$\frac{\begin{pmatrix} n_{1+} \\ n_{11} \end{pmatrix} \begin{pmatrix} n_{2+} \\ n_{+1} - n_{11} \end{pmatrix}}{\begin{pmatrix} n \\ n_{+1} \end{pmatrix}}.$$

Range of n_{11}: $m_- \leq n_{11} \leq m_+$, where $m_- = \max(0, n_{1+} + n_{+1} - n)$ and $m_+ = \min(n_{1+}, n_{+1})$.

Consider H_a: $\theta > 1$. Given both marginal totals, θ increases with n_{11}. Therefore, to test independence the p-value is the sum of hypergeometric probabilities for tables having n_{11} at least as large as the observed value.

Fisher's tea drinker example

An experiment to test a British woman's claim that she could distinguish whether milk or tea was added to the cup first.

	Guess poured first		
Poured first	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	8

H_0: $\theta = 1$, H_a: $\theta > 1$.

Completely natural to use the hypergeometric distribution as the null distribution of n_{11} because both row and column marginals are fixed.

Recall, p-value is the null probability of the outcomes of the observed table and those would have given even more evidence in favor of her claim.

The null probability for the observed table is

$$\frac{\begin{pmatrix} 4 \\ 3 \end{pmatrix}\begin{pmatrix} 4 \\ 1 \end{pmatrix}}{\begin{pmatrix} 8 \\ 4 \end{pmatrix}} = 0.229$$

There is only one more extreme table, with four correct guesses. The corresponding null probability is

$$\frac{\begin{pmatrix} 4 \\ 4 \end{pmatrix}\begin{pmatrix} 4 \\ 0 \end{pmatrix}}{\begin{pmatrix} 8 \\ 4 \end{pmatrix}} = 0.014$$

Therefore, the p-value is .229 + .014 = .243. The null hypothesis of no association between the truths and guesses was not rejected. (This could be due to the small sample size, though, i.e., lack of power to detect an association.)

A two-sided p-value can be defined as the sum of the probabilities of tables no more likely to occur than the observed table.

Due to the discreteness, it is usually difficult to achieve the exact significance level (or Type I error). They are usually smaller than the nominal significance level, and therefore is conservative.

Randomization on the boundary of the critical region can help achieve the exact significance level. For example, if we reject the null hypothesis with 0.157 probability (a tuned value), then the significance level equals

$$P(\text{reject } H_0) = E\ P(\text{reject } H_0 \mid n_{11}) = 1.0(0.014) + 0.157(0.229) = 0.05.$$

However, randomized tests are difficult to justify in practice. Instead, it may be recommended to simply report the p-value.

3.4.2 Derivation of Exact Conditional Distribution

We start with assuming fixed row totals, or independent multinomial sampling across rows.

Under H_0, $\pi_{j|1} = \ldots \pi_{j|I} = \pi_{+j}$, $j = 1, \ldots, J$.

The joint probabilities of $\{n_{ij}\}$ is

$$\frac{\left(\prod_i n_{i+}!\right)\left(\prod_j \pi_{+j}^{n_{+j}}\right)}{\prod_i \prod_j n_{ij}!}$$

Nuisance parameters: π_{+j}.

$\{n_{+j}\}$ are sufficient statistics for $\{\pi_{+j}\}$.

We want to condition on $\{n_{+j}\}$ to eliminate the nuisance parameters from the null distribution for n_{11}.

The distribution of $\{n_{+j}\}$ is multinomial $(n, \{\pi_{+j}\})$.

The joint distribution function of $\{n_{ij}\}$ and $\{n_{+j}\}$ is identical to the probability function of $\{n_{ij}\}$ (since $\{n_{+j}\}$ are determined by $\{n_{ij}\}$).

The conditional distribution of $\{n_{ij}\}$ given $\{n_{+j}\}$ is therefore

$$\frac{\left(\prod_i n_{i+}!\right)\left(\prod_j n_{+j}!\right)}{n! \prod_i \prod_j n_{ij}!}. \tag{15}$$

If we instead start with a multinomial sample (but not an independent multinomial sample), then we can condition on both the row and column totals to still obtain the null distribution (15).

Distribution (15) is called the *multiple hypergeometric distribution.*

3.4.3 Other Exact Tests of Independence

Exact tests of independence for an $I \times J$ table use the multiple hypergeometric distribution.

Exact test for ordinal data: A p-value could be $P[C-D \geq (C-D)_o]$, where C and D denote the # of concordant and discordant pairs and $(C-D)_o$ denotes the corresponding observed difference.

An example for exact conditional test:

	Smoking level Cigarettes/Day		
	0	$1-24$	>25
Control	25	25	12
Myocardial infarction	0	1	3

For this table, $C = 175$ and $D = 12$, so $(C-D)_o = 163$. Given the marginal totals, the only other table having $(C-D)$ at least this large has counts $(25, 26, 11)$ for row 1 and $(0, 0, 4)$ in row 2.

$P(C - D \geq 163) = 0.0183$.

If we treat the variables as nominal and use an exact distribution for X^2, we get a p-value (defined as $P(X^2 \geq X_o^2)$ where $X_o^2 = 6.96$ is observed X^2) of 0.052.

Both exact tests, but assuming ordinal (as it is the case) yields larger power.

Other tests/p-values: 1) Pearson's chi-squared test (assuming large sample) has a p-value of 0.031; 2) Freeman-Halton p-value is 0.034.

Usually computation is intensive for exact tests for $I \times J$ tables with $I > 2$ and/or $J > 2$.

4 The Generalized Linear Models (GLMs)

4.1 Ordinary Linear Regression

Conditional on $\{x_i\}$, $y \sim \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. Using usual matrix notation,

$$Y = X\boldsymbol{\beta} + \epsilon$$

with $\epsilon \sim N(0, \Sigma = \sigma^2 I)$.

The above equation can also be decomposed into the following parts:

1) $y \sim N(Ey \equiv \mu, \sigma^2)$;

2) $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$;

3) $\mu = \eta$.

4.2 Components of GLMs

1. Random component: Random component identifies the response variable y and assumes a probability distribution for it (e.g., success/failure and Bernoulli/binomial distribution; counts and Poisson distribution, negative binomial distribution, etc.)

Denote observations on y by $(y_1, \ldots, y_n)$. Standard GLMs treat $y_1, \ldots, y_n$ as independent.

Example: dose-response (toxicity) curve. y is the toxicity outcome (yes or no). $y \sim Bernoulli(\mu)$.

2. Systematic component: Systematic component specifies the explanatory variables. They enter linearly as predictors on the right-hand side (RHS) of the model equation (to be given). It is of the form: $\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$.

Example: dose-response (toxicity) curve: $\beta_0 + \beta_1 x$, where x is dose level.

3. Link function: Denote $\mu = Ey$, the expected value of y. The link function specifies a function $g(\cdot)$ that relates μ to the linear predictor as

$$g(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

The function $g(\cdot)$ connects the random and systematic components. The simplest link function is $g(\mu) = \mu$, known as the identity link function. This is used in the ordinary linear regression equation.

Example: dose-response (toxicity) curve: $g(\mu) = \log[\mu/(1-\mu)]$ – logit link (to be discussed later).

Other examples: If y is a cell count, $y \sim Poisson(\mu)$, $g(\mu) = \log \mu$ links with a systematic linear component – known as a log linear model.

4.3 Likelihood Functions for GLMs

$Y = (y_1, \ldots, y_n)'$. An important class of distributions of y_i's: exponential family.

Let $f_y(y; \theta, \phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y, \phi)\}$ for some functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. If ϕ is known, $f_y(y; \theta, \phi)$ is an exponential family model with canonical (natural) parameter θ.

Example: Normal distribution (treating μ as a parameter of interest)

$$\begin{aligned} &\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(y-\mu)^2\}/2\sigma^2 \\ &= \exp\left\{(y\mu - \mu^2/2)/\sigma^2 - y^2/(2\sigma^2) - \log(2\pi\sigma^2)/2\right\} \end{aligned}$$

so that $\theta = \mu$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \theta^2/2$, and $c(y, \phi) = -\{y^2/\sigma^2 + \log(2\pi\sigma^2)\}/2 = -\{y^2/\phi + \log(2\pi\phi)\}/2$.

For a general exponential family distribution, let $L(\theta, \phi; y) = \log f_y(y; \theta, \phi)$ (function of θ and ϕ). Can we derive a general formula for Ey and $Var(y)$ based on $f_y(y; \theta, \phi)$?

When no confusion is caused, we use notation $f_y(y; \theta, \phi)$ and $l(\theta)$ interchangeably.

$$E\frac{\partial L}{\partial \theta} = E\frac{\partial l(\theta)/\partial\theta}{l(\theta)} = \int \partial l(\theta)/\partial\theta dy = \partial \int l(\theta) dy/\partial\theta = 1' = 0$$

On the other hand, $E\{\partial L/\partial\theta\} = E((y - b'(\theta))/a(\phi))$. We therefore have $Ey = b'(\theta)$.

Let us now calculate $E\partial^2 L/\partial\theta^2$. We write

$$\frac{\partial^2 L}{\partial\theta^2} = \frac{\partial^2 l(\theta)/\partial\theta^2}{l(\theta)} - \left\{\frac{\partial l(\theta)/\partial\theta}{l(\theta)}\right\}^2.$$

Taking expectation of both sides, we get

$$E\left(\frac{\partial^2 L}{\partial\theta^2}\right) + E\left(\frac{\partial L}{\partial\theta}\right)^2 = 0$$

(because $\int \partial l(\theta)/\partial\theta dy = 0 \Rightarrow \int \partial^2 l(\theta)/\partial\theta^2 dy = 0$).

This implies

$$-\frac{b''(\theta)}{a(\phi)} + E\left\{\frac{y - b'(\theta)}{a(\phi)}\right\}^2 = 0,$$

or $Var(y) = b''(\theta)a(\phi)$. Here $b''(\theta)$ depends on the canonical parameter (and hence on the mean) only and will be called the *variance function* (denoted as $V(\mu)$). $a(\phi)$ is independent of θ and depends only on ϕ.

The function $a(\phi)$ is commonly of form $a(\phi) = \phi/w$, where ϕ, called the *dispersion parameter*, is constant over observations, and w is a known *prior weight* that varies from observation to observation.

Example of $a(\phi)$: For a normal model in which each observation is the mean of m independent readings we have $a(\phi) = \sigma^2/m$.

The most important exponential family distributions:

Normal:

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-(y-\mu)^2\}/2\sigma^2$$
$$= \exp\left\{(y\mu - \mu^2/2)/\sigma^2 - y^2/(2\sigma^2) - \log(2\pi\sigma^2)/2\right\}$$

Poisson:

$$\frac{\mu^y \exp(-\mu)}{y!} = \exp\{y\theta - \exp(\theta) - \log y!\}.$$

$\theta = \log\mu$, $a(\phi) = \phi = 1$, $b(\theta) = \exp(\theta)$, $c(y,\phi) = \log y!$.

Binomial (divided by m):

$$\begin{pmatrix} m \\ my \end{pmatrix} \pi^{my}(1-\pi)^{m-my} = \exp\left\{ my\log\{\pi/(1-\pi)\} + m\log(1-\pi) + \log\begin{pmatrix} m \\ my \end{pmatrix} \right\},$$

$\theta = \log\{\pi/(1-\pi)\}$, $a(\phi) = 1/m$, $\phi = 1$, $b(\theta) = -\log(1-\pi) = \log\{1+\exp(\theta)\}$, $c(y,\phi) = \log\begin{pmatrix} m \\ my \end{pmatrix}$.

Gamma:

$$\frac{y^{\nu-1}\exp(-y/\mu)}{\Gamma(\nu)\mu^{\nu}} = \exp\left\{\nu y\{-1/(\mu\nu)\} - \nu\log(\mu\nu) + \nu\log(\nu y) - \log y - \log\Gamma(\nu)\right\},$$

$a(\phi) = \phi = 1/\nu$, $\theta = -1/(\mu\nu)$, $b(\theta) = -\log(-\theta)$, $c(y,\phi) = \nu\log(\nu y) - \log y - \log\Gamma(\nu) = \log(y/\phi)/\phi - \log y - \log\Gamma(1/\phi)$.

Inverse Gaussian (assume $\mu > 0$):

$$\sqrt{\frac{\sigma^2}{2\pi y^3}}\exp\left\{-\frac{\sigma^2(y-\mu)^2}{2\mu^2 y}\right\} = \exp\left\{\log\sigma^2/2 - \log(2\pi)/2 - 3\log y/2 - \sigma^2(y-\mu)^2/(2\mu^2 y)\right\}$$

$$= \exp\left\{\log\sigma^2/2 - \log(2\pi)/2 - 3\log y/2 - \sigma^2(y - 2\mu + \mu^2/y)/(2\mu^2)\right\}$$

$a(\phi) = \phi = 1/\sigma^2$, $\theta = -1/(2\mu^2)$, $b(\theta) = -\sqrt{-2\theta}$, $c(y,\phi) = -\left\{\log(2\pi\phi y^3) + 1/(\phi y)\right\}/2$.

4.4 Processes in Model Fitting

Three distinct processes in model fitting: 1) model selection; 2) parameter estimation; 3) prediction.

4.4.1 Model Selection

Start with a particular class of models.

Issues to consider in model selection:

1. Model assumptions.

The GLM example. Assumptions: 1) Independence across observations. For example, this excludes autoregressive correlation structure of time series and spatial processes; 2) A single error term. In classical linear models, when both within and between factors are present, there are two error terms. (However, these cases also correspond to the dependent cases in my opinion.)

While these assumptions are in some sense restrictive, they may not be as restrictive as they appear.

Examples:

1) Fitting of autoregressive models using programs for fitting ordinary linear models;

2) A grouping factor (treated as nuisance) is present in a categorical data analysis. If we can somehow eliminate the effects of the nuisance factor and perform a within-group analysis, then observations can be treated as if they were independent (e.g., conditional logistic regression for matched case-control data).

2. Choice of the scale for analysis (e.g., y, $\sqrt{y}$, $\log y$, etc.)

Choice of analysis scale depends on both observations and the purpose the scale is to be used.

In classical linear regression, a good scale may imply: 1) constancy of variance; 2) normality; 3) additivity of systematic effects.

For a Poisson random variable, systematic effects are often multiplicative. $y^{1/2}$ stabilized variance, $y^{2/3}$ approximates symmetry or normality better, and $\log y$ produces additivity of the systematic effects. However, often no scale achieves all purposes for a non-normally distributed random variable.

The GLM is a good tool to reduce the scaling problems, because: 1) distribution can be explicitly specified and inference can proceed without the normality assumption; 2) additivity of effects can be specified on a transformed scale (so we can still assume additivity of effects in spite of a clearly indicated non-linear relationship between the mean of the outcome and the covariates).

Note that in the GLMs, additivity is correctly postulated as a property of the expected responses (instead of on the data themselves!)

3. Covariate selection.

Criterion needs to be defined. What is considered an optimal model in certain context? Examples: AIC, BIC, DIC.

It is unlikely that a clear winner is indicated among a large number of competing models. So usually a set of 'alternatives' are almost as good and statistically indistinguishable with the 'winner'.

4.4.2 Estimation

Once a model is chosen, parameters need to be estimated along with their precision.

Criterion of estimation. In the case of GLMs, it is goodness of fit between the observed data and the fitted values generated by the model. That is, the parameter estimates minimize the goodness-of-fit criterion. We will mainly be concerned about maximizing the likelihood or log likelihood of the parameters.

Suppose $f(y;\theta)$ is the density function or probability distribution for y. Let $L(\mu;y)$ be the log likelihood as a function of $\mu = Ey$. With n independent observations, $L(\boldsymbol{\mu};\mathbf{y}) = \sum_i \log f(y_i;\theta_i)$, where $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_n)$.

Define

$$D^\star(\mathbf{y};\boldsymbol{\mu}) = 2L(\mathbf{y};\mathbf{y}) - 2L(\boldsymbol{\mu};\mathbf{y}),$$

which we call the *scaled deviance.*

For the exponential family models considered in this class, $L(\mathbf{y};\mathbf{y})$ is the maximum likelihood achievable for an exact fit in which the fitted values are equal to the observed data. Therefore, maximizing $L(\boldsymbol{\mu};\mathbf{y})$ is equivalent to minimizing $D^\star(\mathbf{y};\boldsymbol{\mu})$, subject to appropriate model constraints for $\boldsymbol{\mu}$.

The ordinary linear regression example (assuming normality and known variance σ^2):

For a single observation y,

$$f(y;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}.$$

Log likelihood is

$$L(\mu; y) = -\log(2\pi\sigma^2)/2 - (y-\mu)^2/(2\sigma^2).$$

The maximum likelihood achievable is obtained by replacing μ by y:

$$L(y; y) = -\log(2\pi\sigma^2)/2.$$

Therefore, the scaled deviance is

$$D^\star(y;\mu) = 2\{L(y;y) - L(\mu;y)\} = (y-\mu)^2/\sigma^2.$$

Therefore, the deviance is synonymous with least squares for the normal-theory linear regression model.

4.4.3 Prediction

Generally, prediction is concerned with statements about the likely values of unobserved events (not necessarily those in the future, as is typically the case in the time series analysis).

For example, predict random effects (unobserved) in linear mixed-effects models.

4.4.4 Important Link Functions, Canonical Links and Sufficient Statistics

Appropriate link functions match ranges of μ and η.

Some principal link functions considered in this course:

For a Poisson distribution, a log link:

$$\eta = \log \mu$$

This results in multiplicative effects.

For a binomial distribution:

1. logit link

$$\eta = \log\{\mu/(1-\mu)\}$$

2. probit link

$$\eta = \Phi^{-1}(\mu),$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution.

3. complementary log-log link

$$\eta = \log\{-\log(1-\mu)\}.$$

For observations with $\mu > 0$, the power family of links is important.

$$\eta = (\mu^{\lambda} - 1)/\lambda,\ \ \lambda \neq 0;\ \ \eta = \log \mu,\ \ \lambda = 0.$$

Another alternative in the power link family:

$$\eta = \mu^{\lambda},\ \ \lambda \neq 0;\ \ \eta = \log \mu,\ \ \lambda = 0.$$

For both links, special action needs to be taken in any computation with $\lambda = 0$.

Canonical links occur when

$$\theta = \eta,$$

where θ is the canonical parameter.

The canonical links for the five distributions we talked about earlier are as follows:

$$\text{Normal: } \eta = \mu,$$

$$\text{Poisson: } \eta = \log \mu,$$

$$\text{binomial: } \eta = \log\{\pi/(1-\pi)\}$$

$$\text{gamma: } \eta = \mu^{-1}$$

$$\text{inverse Gaussian: } \eta = \mu^{-2}$$

Sufficient statistic under the canonical link: $\mathbf{X}^T\mathbf{Y}$ (a $p \times 1$ vector), with the jth component being

$$\sum_i x_{ij} y_i.$$

Derive this.

There are nice properties of the model when the link is the canonical link.

Nice statistical properties alone should not replace quality of fit as a model selection criterion. We will talk about non-canonical link functions later (e.g., probit link for binary data). However, we shall find that the canonical links are often eminently sensible on scientific grounds.

4.5 Measuring the Goodness of Fit

4.5.1 The Discrepancy of a Fit

Full model: n observations, n parameters

Null model: n observations, 1 parameter (overall mean)

Maximum log likelihood achievable under full model: $L(\mathbf{y}, \phi; \mathbf{y})$

Maximized (over $\boldsymbol{\beta}$) log likelihood under the GLM: $L(\hat{\boldsymbol{\mu}}, \phi; \mathbf{y})$

Let $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\boldsymbol{\mu}})$ and $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}(\mathbf{y})$ be the canonical parameters under the full model and the GLM, respectively. Assuming $a_i(\phi) = \phi/w_i$, the discrepancy is

$$\sum 2w_i \left\{ y_i \left(\tilde{\theta}_i - \hat{\theta}_i - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \right) \right\} / \phi = D(\mathbf{y}; \hat{\boldsymbol{\mu}})/\phi,$$

where $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ is known as the *deviance* for the current model. The deviance depends on the data only (because the maximum likelihood achievable under the full model does not depend on model).

Relationship between scaled deviance and deviance:

$$D^{\star}(\mathbf{y}; \hat{\boldsymbol{\mu}}) = D(\mathbf{y}; \hat{\boldsymbol{\mu}})/\phi.$$

Deviances for the following five distributions:

Normal: $\sum (y - \hat{\mu})^2$,

Poisson: $2 \sum \{ y \log (y/\hat{\mu}) - (y - \hat{\mu}) \}$,

binomial: $2 \sum \{ y \log (y/\hat{\mu}) + (m - y) \log [(m - y)/(m - \hat{\mu})] \}$,

gamma: $2 \sum \{ - \log (y/\hat{\mu}) + (y - \hat{\mu}) / \hat{\mu} \}$,

inverse Gaussian: $\sum (y - \hat{\mu})^2 / \hat{\mu}^2$.

Deviance for the normal distribution is the residual sum of squares.

If summed over n observations and the GLM has an intercept term, then the deviances for the Poisson and binomial distributions are G^2.

The other important measure of discrepancy: generalized Pearson X^2 statistic

$$X^2 = \sum (y - \hat{\mu})^2 / V(\hat{\mu}),$$

where $V(\hat{\mu})$ is the estimated variance function for the distribution concerned.

For the normal distribution, X^2 is again the residual sum of squares.

For the Poisson or binomial distribution it is the original Pearson X^2 statistic.

Verify this for the binomial distribution.

For normal-theory linear models, both the deviance and generalized X^2 have exact chi-squared distributions.

For other distributions, asymptotic results are available. In small samples, either of these may prove to be superior in its distributional properties. But in general, deviance has the advantage of being additive for nested models when ML estimates are used. X^2 sometimes is preferred because of its more direct interpretation.

4.5.2 The Analysis of Deviance

A table for first differences in the deviance between nested models can be used as a screening device for picking out obviously important terms.

Example: models involving two factors as predictors: 1, A, B, A+B.

4.6 Residuals

In normal linear models, $r = y - \hat{\mu}$.

We consider the theoretical form of the definition of residuals for GLMs below.

4.6.1 Pearson Residual

$$r_P = \frac{y - \mu}{\sqrt{V(\mu)}},$$

i.e., the raw residual divided by the estimated standard deviation of y.

The name Pearson comes from the fact that for a Poisson distribution, the Pearson residual is just the signed square root of the component of the Pearson X^2 goodness-of-fit statistic, so that

$$\sum r_P^2 = X^2.$$

4.6.2 Anscombe Residual

Disadvantage of the Pearson residual: markedly skewed distribution of r_P.

Anscombe proposed defining a residual using a transformation of y, say $A(y)$, instead of the original y, to "normalize" the residual distribution.

Wedderburn showed for GLMs, $A(\cdot)$ is as follows:

$$A(\cdot) = \int \frac{d\mu}{V^{1/3}(\mu)}.$$

For the Poisson distribution, we have

$$\int \frac{d\mu}{\mu^{1/3}} = 3\mu^{2/3}/2.$$

So we base our residual on $y^{2/3} - \mu^{2/3}$. We further do scaling, i.e., dividing by the standard deviation of $A(y)$, for which the first order approximation is $A'(\mu)\sqrt{V(\mu)}$. Therefore, the Anscombe residual for the Poisson distribution is

$$r_A = \frac{3(y^{2/3} - \mu^{2/3})/2}{\mu^{1/6}}.$$

For the gamma distribution, the Anscombe residual takes the form

$$r_A = \frac{3(y^{1/3} - \mu^{1/3})}{\mu^{1/3}}.$$

For the inverse Gaussian distribution,

$$r_A = (\log y - \log \mu)/\mu^{1/2}.$$

4.6.3 Deviance Residual

$$r_D = \text{sign}(y - \mu)\sqrt{d_i}.$$

$\sum r_D^2 = D.$

For the Poisson distribution,

$$r_D = \text{sign}(y - \mu)\left\{2\left(y\log(y/\mu) - y - \mu\right)\right\}^2.$$

Although the forms of the Anscombe and deviance residuals are very different, they turn out to be numerically very similar for a range of observed values (say $y = c\mu$ with $c \in [0, 10]$).

4.7 An Algorithm for Fitting the GLMs

We introduce an iterative procedure.

Assign an initial value $\hat{\boldsymbol{\beta}}_0$ (hence $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\eta}}_0$ for the expected values and linear predictors of y).

Let $\hat{\boldsymbol{\beta}}_k$ (hence $\hat{\boldsymbol{\mu}}_k$ and $\hat{\boldsymbol{\eta}}_k$) be the estimates in the current iteration.

Construct an adjusted dependent variate for each single observation (suppressing the subscript i for unit) $z_k = \hat{\eta}_k + (y - \hat{\mu}_k)\left(\frac{d\eta}{d\mu}|_{\hat{\mu}_k}\right)$. Further define $W_k^{-1} = \left(\frac{d\eta}{d\mu}\right)^2|_{\hat{\mu}_k}V(\hat{\mu}_k)$.

Now regress z_k on the covariates $x_1, \ldots, x_p$ with weight W_k to give new estimates $\hat{\boldsymbol{\beta}}_{k+1}$. Compute $\hat{\boldsymbol{\mu}}_{k+1}$ and $\hat{\boldsymbol{\eta}}_{k+1}$ using $\hat{\boldsymbol{\beta}}_{k+1}$. Repeat until the changes between successive $\hat{\boldsymbol{\beta}}$'s are small.

4.7.1 Justification of the Fitting Procedure

We first show that the ML equations for β_j are given by

$$\sum W(y - \mu)\frac{d\eta}{d\mu}x_j = 0,$$

where the summation is over the units.

Note

$$L = \{y\theta - b(\theta)\}/a(\phi) + c(y, \phi)$$

and

$$\frac{\partial L}{\partial \beta_j} = \frac{\partial L}{\partial \theta}\frac{d\theta}{d\mu}\frac{d\mu}{d\eta}\frac{\partial \eta}{\partial \beta_j}.$$

We have the following: $b'(\theta) = \mu$, $b''(\theta) = V \Rightarrow d\mu/d\theta = V$; $\partial\eta/\partial\beta_j = x_j$. Therefore

$$\frac{\partial L}{\partial \beta_j} = \frac{y-\mu}{a(\phi)}\frac{1}{V}\frac{d\mu}{d\eta}x_j = \frac{W}{a(\phi)}(y-\mu)\frac{d\eta}{d\mu}x_j.$$

This results in the ML equation assuming $a(\phi) = \phi$.

Fisher's scoring algorithm:

Denote $\mathbf{u} = \partial L/\partial \boldsymbol{\beta}$ and $\mathbf{A} = -E\left(\frac{\partial^2 L}{\partial \beta_r \partial \beta_s}\right)$, the Hessian matrix.

The Fisher scoring algorithm computes an adjustment of the estimate $\mathbf{b}$ of $\boldsymbol{\beta}$, denoted as $\Delta\mathbf{b}$ by solving

$$\mathbf{A}\Delta\mathbf{b} = \mathbf{u}.$$

Note

$$u_r = \sum W(y-\mu)\frac{d\eta}{d\mu}x_r,$$

so that

$$A_{rs} = -E\frac{\partial u_r}{\partial \beta_s} = -E\sum\left[(y-\mu)\frac{\partial}{\partial \beta_s}\left\{W\frac{d\eta}{d\mu}x_r\right\} + W\frac{d\eta}{d\mu}x_r\frac{\partial}{\partial \beta_s}(y-\mu)\right].$$

The first term vanishes on taking expectations while the second term reduces to

$$\sum_i W\frac{d\eta}{d\mu}x_r\frac{\partial \mu}{\partial \beta_s} = \sum_i W x_r x_s.$$

Therefore $(\mathbf{A}\mathbf{b})_r = \sum_s A_{rs}b_s = \sum W x_r \eta$, and

$$(\mathbf{A}\mathbf{b}^\star)_r = (\mathbf{A}\mathbf{b} + \mathbf{A}\Delta\mathbf{b})_r = (\mathbf{A}\mathbf{b} + \mathbf{u})_r = \sum_i W x_r\{\eta + (y-\mu)d\eta/d\mu\}.$$

These equations have the form of linear weighted least-squares equations with weight

$$W = V^{-1}\left(\frac{d\mu}{d\eta}\right)^2$$

and dependent variate

$$z = \eta + (y-\mu)\frac{d\eta}{d\mu}.$$

The fitting algorithm for the GLMs is thus justified.

4.8 GLMs for Binary Data

4.8.1 Linear Probability Model

In ordinary regression, $\mu = Ey$ is a linear function of x. For a binary response, an analogous model is

$$\pi(x) = \beta_0 + \beta_1 x.$$

This is called a linear probability model.

The model has a *structural defect*.

Example: Snoring and Heart Disease

Relationship between snoring and heart disease

	Heart disease	
Snoring	Yes	No
Never	24	1355
Occasional	35	603
Nearly every night	21	192
Every night	30	224

Two fitting methods. One assumes the binomial random component and an identity link function. No closed form for ML estimates. Iterative algorithms are required. Second: treating the outcome as normal, and use least squares estimates.

Result from the first method:

$x = (0, 2, 4, 5)$.

$\hat{\pi} = 0.0172 + 0.0198x$.

Coefficient $\beta = 0.0198$ is significant with a $SE = 0.0028$.

Results change somewhat with the score assignment for snoring.

4.8.2 Logistic Regression Model

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)},$$

or equivalently,

$$\log[\pi(x)/\{1 + \pi(x)\}] = \beta_0 + \beta_1 x.$$

The link function $\log[\pi/(1-\pi)]$ is called the logit function. Logistic regression models are often called *logit* models.

For the snoring and heart disease example, the ML fit is

$$\text{logit}[\hat{\pi}(x)] = -3.87 + 0.40x.$$

4.8.3 Probit Regression Model

$$\text{probit}[\pi(x)] = \beta_0 + \beta_1 x.$$

Snoring and heart disease example:

$$\text{probit}[\hat{\pi}(x)] = -2.061 + 0.188x.$$

4.8.4 Binary Regression and Cumulative Distribution Function

In general, let $F(\cdot)$ be a cdf function.

$$F^{-1}[\pi(x)] = \beta_0 + \beta_1 x.$$

4.9 Asymptotic Inference for Model Parameters

Recall the likelihood equations

$$\sum_{i=1}^{N} \frac{(Y_i - \mu_i)x_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \; j = 1, \ldots, p.$$

The Newton-Raphson or Fisher scoring method helps solve the likelihood equations and obtain the ML estimators. For making inference on the parameters, we also need to calculate the asymptotic standard errors of the estimators and confidence intervals of the parameters. To this end, we need to calculate the asymptotic covariance matrix of model parameter estimators, which is the inverse of the information matrix $\mathcal{I}$ and has elements $E\left[-\partial^2 L(\boldsymbol{\beta})/\partial\beta_h\partial\beta_j\right]$. To find this, for the contribution L_i to the log likelihood we use the helpful result

$$E\left(\frac{\partial^2 L_i}{\partial\beta_h\partial\beta_j}\right) = -E\left(\frac{\partial L_i}{\partial\beta_h}\right)\left(\frac{\partial L_i}{\partial\beta_j}\right),$$

which holds for exponential families (Cox and Hinkley, 1974, Sec. 4.8). Since

$$\frac{\partial L_i}{\partial\beta_j} = \frac{y_i-\mu_i}{a(\phi)}\frac{a(\phi)}{var(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}x_{ij} = \frac{(Y_i-\mu_i)x_{ij}}{var(Y_i)}\frac{\partial\mu_i}{\partial\eta_i},$$

we have

$$E\left(\frac{\partial^2 L_i}{\partial\beta_h\partial\beta_j}\right) = -E\left[\frac{(Y_i-\mu_i)x_{ih}}{var(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}\frac{(Y_i-\mu_i)x_{ij}}{var(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}\right].$$

Since $L(\boldsymbol{\beta}) = \sum_i L_i$,

$$E\left(-\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_h\partial\beta_j}\right) = \sum_{i=1}^{N}\frac{x_{ih}x_{ij}}{var(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2.$$

Thus, the information matrix has the form

$$\mathcal{I} = \mathbf{X}'\mathbf{W}\mathbf{X},$$

where $\mathbf{W}$ is the diagonal matrix with main-diagonal elements

$$w_i = \left(\partial\mu_i/\partial\eta_i\right)^2/var(Y_i).$$

The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is estimated by

$$c\hat{o}v\left(\hat{\boldsymbol{\beta}}\right) = \hat{\mathcal{I}^{-1}} = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1},$$

where $\hat{\mathbf{W}}$ is $\mathbf{W}$ evaluated at $\hat{\boldsymbol{\beta}}$. Note that the form of $\mathbf{W}$ also depends on the link function.

4.10 Inference for Generalized Linear Models

The Wald, score, and likelihood-ratio methods introduced earlier for significance testing and interval estimation apply to any GLM. Here we concentrate on likelihood-ratio inference, through the deviance of the GLM.

4.10.1 Deviance and Goodness of Fit

For some GLMs the scaled deviance has an asymptotic chi-squared distribution.

4.10.2 Deviance for Poisson Models

Assuming Poisson counts in two-way contingency tables, the deviance of the Poisson GLM that uses a log link function, contains an intercept term, and treats the X variable as an explanatory variable, reduces to the G^2 statistic we have discussed earlier. For a Poisson or multinomial model applied to a contingency table with a fixed number of cells c, the deviance has an approximate chi-squared distribution for large $\{\mu_i\}$.

4.10.3 Deviance for Binomial Models: Grouped and Ungrouped Data

With binomial responses, it is possible to construct the data file with the counts of successes and failures at each setting for the predictors, or with the individual Bernoulli 0-1 observations at the subject level. The deviance differs in the two cases. In the first case the saturated model has a parameter at each setting for the predictors, whereas in the second case it has a parameter for each subject. We refer to these as *grouped data* and *ungrouped data* cases. The approximate chi-squared distribution for the deviance occurs for the grouped data but not for ungrouped data. With grouped data, the sample size increases for a fixed number of settings of the predictors and hence a fixed number of parameters for the saturated model.

A fact: Both the deviances for a Poisson loglinear model with an intercept term and a binomial GLM with logit link have a deviance of the form

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2\sum \text{observed } \times \log\left(\text{observed}/\text{fitted}\right).$$

4.10.4 Likelihood-Ratio Model Comparison Using the Deviance

For a Poisson or binomial model M, $\phi = 1$, so the deviance (= scaled deviance) equals

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = -2\left[L(\hat{\boldsymbol{\mu}};\mathbf{y}) - L(\mathbf{y};\mathbf{y})\right].$$

Consider two models, M_0 with fitted values $\hat{\boldsymbol{\mu}}_0$ and M_1 with fitted values $\hat{\boldsymbol{\mu}_1}$, with M_0 a special case of M_1. Model M_0 is said to be *nested* within M_1.

Since M_0 is simpler than M_1, a smaller set of parameter values satisfies M_0 than satisfies M_1. Thus, $L(\hat{\boldsymbol{\mu}}_0;\mathbf{y}) \leq L(\hat{\boldsymbol{\mu}}_1;\mathbf{y})$, and it follows that

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}_1) \leq D(\mathbf{y};\hat{\boldsymbol{\mu}}_0).$$

Assuming that model M_1 holds, the likelihood-ratio test of the hypothesis that M_0 holds uses the test statistic

$$-2\left[L(\hat{\boldsymbol{\mu}}_0;\mathbf{y}) - L(\hat{\boldsymbol{\mu}}_1;\mathbf{y})\right] = D(\mathbf{y};\hat{\boldsymbol{\mu}}_0) - D(\mathbf{y};\hat{\boldsymbol{\mu}}_1).$$

The likelihood-ratio statistic comparing the two models is simply the difference between the deviances. This statistic is large when M_0 fits poorly compared to M_1. The difference between deviances also has the form of the deviance. Under regularity conditions, this difference has approximately a chi-squared null distribution with df equal to the difference between the numbers of parameters in the two models.

For binomial GLMs and Poisson loglinear GLMs with intercept, the difference in deviance uses the observed counts and the two sets of fitted values in the form

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}_0) - D(\mathbf{y};\hat{\boldsymbol{\mu}}_1) = 2\sum \text{observed } \times \log\left(\text{fitted}_1/\text{fitted}_0\right).$$

5 Logistic Regression

5.1 Interpreting Parameters in Logistic Regression

For a binary response variable Y and an explanatory variable X, let $\pi(x) = P(Y = 1 \mid X = x) = 1 - P(Y = 0 \mid X = x)$. The logistic regression model is

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}. \tag{16}$$

Equivalently, the log odds, called the *logit*, has the linear relationship

$$\operatorname{logit}[\pi(x)] = \log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x. \tag{17}$$

This equates the logit link function to the linear predictor.

5.1.1 Interpreting β: Odds Ratios

How can we interpret β in (17)? Its sign determines whether $\pi(x)$ is increasing or decreasing as x increases. The rate of climb or descent increases as $|\beta|$ increases; as $\beta \to 0$ the curve flattens to a horizontal straight line. When $\beta = 0$, Y is independent of X. For quantitative x with $\beta > 0$, the curve for $\pi(x)$ has the shape of the cdf of the logistic distribution. Since the logistic density is symmetric, $\pi(x)$ approaches 1 at the same rate as it approaches 0.

Exponentiating both sides of (17) shows that the odds are an exponential function of x. This provides a basic interpretation for the magnitude of β: The odds increase multiplicatively by e^β for every 1-unit increase in x. In other words, e^β is an odds ratio, the odds at $X = x + 1$ divided by the odds at $X = x$.

The intercept parameter α is not usually of particular interest. However, by centering the predictor about 0 [i.e., replacing x by $(x - \bar{x})$, α becomes the logit at that mean, and thus $e^\alpha/(1 + e^\alpha) = \pi(\bar{x})$. (As in ordinary regression, centering is also helpful in complex models containing quadratic or interaction terms to reduce correlations among explanatory variables, also known as collinearity.)

5.1.2 Logistic Regression with Retrospective Studies

Another property of logistic regression relates to situations in which the explanatory variable X rather than the response variable Y is random. This occurs with retrospective sampling designs, such as case-control biomedical studies. For samples of subjects having $Y = 1$ (cases) and having $Y = 0$ (controls), the value of X is observed. Evidence of an association exists if the distribution of X values differs between cases and controls. In retrospective studies, one can estimate odds ratios. Effects in the logistic regression model refer to odds ratios. Thus, one can estimate such models and estimate effects in case-control studies.

Here is a justification for this. Let Z indicate whether a subject is sampled (1 = yes, 0 = no). Let $\rho_1 = P(Z = 1 \mid y = 1)$ denote the probability of sampling a case, and let $\rho_0 = P(Z = 1 \mid y = 0)$ denote the probability of sampling a control. Even though the conditional distribution of Y given $X = x$ is not sampled, we need a model for $P(Y = 1 \mid z = 1, x)$, assuming that $P(Y = 1 \mid x)$ follows the logistic model. By Bayes' theorem,

$$P(Y = 1 \mid z = 1, x) = \frac{P(Z = 1 \mid y = 1, x)P(Y = 1 \mid x)}{\sum_{j=0}^{1} P(Z = 1 \mid y = j, x)P(Y = j \mid x)}. \tag{18}$$

Now, suppose that $P(Z = 1 \mid y, x) = P(Z = 1 \mid y)$ for $y = 0$ and 1; that is, for each y, the sampling probabilities do not depend on x. For instance, often x refers to exposure of some type, such as whether someone has been a smoker. Then, for cases and for controls, the probability of being sampled is the same for smokers and nonsmokers. Under this assumption, substituting ρ_1 and ρ_0 in (18) and dividing numerator and denominator by $P(Y = 0 \mid x)$, (18) simplifies to

$$P(Y = 1 \mid z = 1, x) = \frac{\rho_1 \exp(\alpha + \beta x)}{\rho_0 + \rho_1 \exp(\alpha + \beta x)}.$$

Then, dividing numerator and denominator by ρ_0 and using $\rho_1/\rho_0 = \exp[\log(\rho_1/\rho_0)]$ yields

$$\text{logit}\,[P(Y = 1 \mid z = 1, x)] = \alpha^\star + \beta x$$

with $\alpha^\star = \alpha + \log(\rho_1/\rho_0)$.

Thus, the logistic regression model holds with the same effect parameter β as in the model for $P(Y = 1 \mid x)$. If the sampling rate for cases is 10 times that for controls, the intercept estimated is $\log(10) = 2.3$ larger than the one estimated with a prospective study.

With case-control studies, one cannot estimate β in other binary-response models (such as the probit models). This is an important advantage of the logit link and is a major reason why logit models have surpassed other models in popularity in biomedical studies.

Many case-control studies employ matching. Each case is matched with one or more control subjects. The controls are like the case on key characteristics such as age. The model and subsequent analysis should take the matching into account. Specifically, a conditional logistic regression approach can be used to analyze such matched case-control data.

5.2 Inference for Logistic Regression

5.2.1 Types of Inference

For a model with a single predictor,

$$\text{logit}\,[\pi(x)] = \alpha + \beta x,$$

significance tests focus on $H_0 : \beta = 0$, the hypothesis of independence. The Wald test uses the log likelihood at $\hat{\beta}$, with test statistic $z = \hat{\beta}/SE$ or its square; under H_0, z^2 is asymptotically $\mathcal{X}_1^2$. The likelihood-ratio test uses twice the difference between the maximized log likelihood at $\hat{\beta}$ and at $\beta = 0$ and also has an asymptotic $\mathcal{X}_1^2$ null distribution. The score test uses the log likelihood at $\beta = 0$ through the derivative of the log likelihood (i.e., the score function) at that point. The test statistic compares the sufficient statistic for β to its null expected value, suitably standardized $[N(0,1)$ or $\mathcal{X}_1^2]$.

For large samples, the three tests usually give similar results. The likelihood-ratio test is preferred over the Wald. It uses more information, since it incorporates the log likelihood at H_0 as well as at $\hat{\beta}$. When $|\beta|$ is relatively large, the Wald test is not as powerful as the likelihood-ratio test and can even show aberrant behavior.

A confidence interval for β results from inverting a test of $H_0: \ \beta = \beta_0$. The interval is the set of β_0 for which the chi-squared test statistic is no greater than $\mathcal{X}_1^2(\alpha) = z_{\alpha/2}^2$. For the Wald approach, this means $\left[\left(\hat{\beta} - \beta_0\right)/SE\right]^2 \leq z_{\alpha/2}^2$; the interval is $\hat{\beta} \pm z_{\alpha/2}(SE)$.

For summarizing the relationships, other characteristics may have greater importance than β, such as $\pi(x)$ at various x values. For fixed $x = x_0$, $\text{logit}\,[\hat{\pi}(x_0)] = \hat{\alpha} + \hat{\beta}x_0$ has a large-sample SE given by the estimated square root of

$$var\left(\hat{\alpha} + \hat{\beta}x_0\right) = var\left(\hat{\alpha}\right) + x_0^2 var\left(\hat{\beta}\right) + 2x_0 cov\left(\hat{\alpha}, \hat{\beta}\right).$$

A 95% confidence interval for $\text{logit}\,[\pi(x_0)]$ is $\left(\hat{\alpha} + \hat{\beta}x_0\right) \pm 1.96SE$. Substituting each endpoint into the inverse transformation $\pi(x_0) = \exp\left(\text{logit}\right)/\left[1 + \exp\left(\text{logit}\right)\right]$ gives a corresponding interval for $\pi(x_0)$.

5.2.2 Checking Goodness of Fit

In practice, there is no guarantee that a certain logistic regression model fits the data well. For any type of binary data, one way to detect lack of fit uses a likelihood-ratio test to compare the model to more complex ones. If more complex models do not fit better, this provides some assurance that the model chosen (i.e., the simpler model) is reasonable.

When the explanatory variables are solely categorical, other approaches to detecting lack of fit include Pearson X^2 (or likelihood-ratio G^2) statistic (based on an asymptotic null chi-squared distribution).

5.3 Logit Models with Categorical Predictors

Dummy variable approach.

Cochran-Armitage trend test for $I \times 2$ table with ordered rows. Assumes a linear probability model. However, this test turns out to be the score test under the linear logit model.

5.4 Multiple Logistic Regression

Multiple predictors.

5.4.1 Goodness of Fit as a Likelihood-Ratio Test and Model Comparison

Same approaches based on likelihood-ratio test statistics following a null chi-squared distribution. Note here “null” means that the reduced model fits equally well as the more complex model.

6 Models for Matched Pairs

We introduce methods for comparing categorical responses for two samples when each observation in one sample pairs with an observation in the other. Such *matched-pairs* data commonly occur in studies with repeated measurement of subjects, such as *longitudinal studies* that observe subjects over time. Because of the matching, the responses in the two samples are statistically dependent.

The following table illustrates matched-pairs data. For a poll of a random sample of 1600 voting-age British citizens, 944 indicated approval of the Prime Minister's performance in office. Six months later, of these same 1600 people, 880 indicated approval. The two cells with identical row and column response form the main diagonal of the table. These subjects had the same opinion at both surveys. They compose most of the sample, since relatively few people changed opinion. A strong association exists between opinions six months apart, the sample odds ratio being $(794 \times 570)/(150 \times 86) = 35.1$.

For matched pairs with a categorical response, a two-way contingency table with the same row and column categories summarizes the data. The table is *square*.

Example of Rating of Performance of Prime Minister

	Second Survey		
First Survey	Approve	Disapprove	Total
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

6.1 Comparing Dependent Proportions

For each of n matched pairs, let π_{ab} denote the probability of outcome a for the first observation and outcome b for the second. Let n_{ab} count the number of such pairs, with $p_{ab} = n_{ab}/n$ the sample proportion. We treat $\{n_{ab}\}$ as a sample from a multinomial $(n; \{\pi_{ab}\})$ distribution. Then p_{a+} is the proportion in category a for observation 1, and p_{+a} is the corresponding proportion for observation 2. We compare samples by comparing marginal proportions $\{p_{a+}\}$ and $\{p_{+a}\}$. With matched samples, these proportions are correlated, and methods for independent samples are inappropriate.

We focus on cases where the outcome is binary. When $\pi_{1+} = \pi_{+1}$, then $\pi_{2+} = \pi_{+2}$ also, and there is *marginal homogeneity.* Since

$$\pi_{1+} - \pi_{+1} = \pi_{12} - \pi_{21},$$

marginal homogeneity in 2×2 tables is equivalent to $\pi_{12} = \pi_{21}$. The table then shows symmetry across the main diagonal.

6.1.1 Inference for Dependent Proportions

One comparison of the marginal distributions uses $\delta = \pi_{+1} - \pi_{1+}$. Let

$$d = p_{+1} - p_{1+} = p_{2+} - p_{+2}.$$

From our earlier results for multinomial covariances, $cov(p_{+1}, p_{1+}) = cov(p_{11} + p_{21}, p_{11} + p_{12}$ simplifies to $(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})/n$. Thus,

$$var(\sqrt{n}d) = \pi_{1+}(1 - \pi_{1+}) + \pi_{+1}(1 - \pi_{+1}) - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21}). \tag{19}$$

For large samples, d has approximately a normal sampling distribution. A confidence interval for $\delta = \pi_{+1} - \pi_{1+}$ is then

$$d \pm z_{\alpha/2}\hat{\sigma}(d),$$

where

$$\hat{\sigma}^2(d) = \left[p_{1+}(1 - p_{1+}) + p_{+1}(1 - p_{+1}) - 2(p_{11}p_{22} - p_{12}p_{21})\right]/n$$

$$= \left[(p_{12} + p_{21}) - (p_{12} - p_{21})^2\right]/n, \tag{20}$$

The hypothesis of marginal homogeneity is $H_0 : \pi_{1+} = \pi_{+1}$ ($i.e.,\ \delta = 0$). The ratio $z = d/\hat{\sigma}(d)$ or its square is a Wald test statistic. Under H_0, an alternative estimated variance is

$$\hat{\sigma}_0^2 = \frac{p_{12} + p_{21}}{n} = \frac{n_{12} + n_{21}}{n^2}.$$

The score test statistic $z_0 = d/\hat{\sigma}_0(d)$ simplifies to

$$z_0 = \frac{n_{21} - n_{12}}{(n_{21} + n_{12})^{1/2}}. \tag{21}$$

The square of z_0 is a chi-squared statistic with df $= 1$. The test using it is called *McNemar's test* (McNemar, 1947).

The McNemar statistic depends only on cases classified in *different* categories for the two observations. The $n_{11} + n_{22}$ on the main diagonal are irrelevant to inference about whether π_{11} and π_{22} differ. This may seem surprising, but *all* cases contribute to inference about *how much* π_{1+} and π_{+1} differ: for instance, to estimating δ and the standard error.

Prime Minister Approval Rating Example:

The sample proportions of approval of the prime minister's performance are $p_{1+} = 944/1600 = 0.59$ for the first survey and $p_{+1} = 880/1600 = 0.55$ for the second. Using (20), a 95% confidence interval for $\pi_{+1} - \pi_{1+}$ is $(0.55 - 0.59) \pm 1.96(0.0095)$, or $(-0.06, -0.02)$. The approval rating appears to have dropped between 2 and 6%.

For testing marginal homogeneity the test statistic (21) using the null variance is

$$z_0 = \frac{86 - 150}{(86 + 150)^{1/2}} = -4.17.$$

It shows strong evidence of a drop in the approval rating.

6.1.2 Increase Precision with Dependent Samples

For *independent* samples of size n each to estimate binomial probabilities π_1 and π_2, the covariance for the sample is zero, and

$$var\left[\sqrt{n}(\text{difference of sample proportions}) = \pi_1(1-\pi_1) + \pi_2(1-\pi_2)\right].$$

Compare this variance with (19). Dependent samples usually exhibit a positive dependence, in which case $\log\theta = \log\left[\pi_{11}\pi_{22}/\left(\pi_{12}\pi_{21}\right)\right]$; that is, $\pi_{11}\pi_{22} > \pi_{12}\pi_{21}$. Thus based on (19), positive dependence implies that $var(d)$ is smaller than when the samples are independent.

An implication of the above comparison: A study design using dependent samples can help improve the precision of statistical inferences for within-subject effects.

6.2 Conditional Logistic Regression for Binary Matched Pairs

6.2.1 A Logit Model with Subject-Specific Probabilities

Like in the Prime Minister approval rating data, let (Y_{i1}, Y_{i2}) denote the ith pair of observations, $i = 1, \ldots, n$. Due to subject heterogeneity, one may believe that different subjects can have different probabilities of approving Prime Minister's performance (e.g., depending on unmeasured subject characteristics such as race, social economic status, etc.) A generalized linear model that allows this possibility could be the following logit model for the probability of $Y_{it} = 1$, $t = 1, 2$:

$$\text{logit}\left[P(Y_{it} = 1)\right] = \alpha_i + \beta x_i, \tag{22}$$

where $x_1 = 0$ and $x_2 = 1$. Although permitting subject-specific distributions, it assumes a common effect β. For subject i,

$$P(Y_{i1} = 1) = \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)},\ P(Y_{i2} = 1) = \frac{\exp(\alpha_i + \beta)}{1 + \exp(\alpha_i + \beta)}.$$

The parameter β compares the response distributions. For each subject, the odds of success for observation 2 are $\exp \beta$ times the odds for observation 1.

Given the parameters, with model (22) one normally assumes independence of responses for different subjects and for the two observations on the same subject. However, averaged over all subjects, the responses are nonnegatively associated. Suppose that $|\beta|$ is small compared to $|\alpha_i|$. A subject with a large positive α_i has high $P(Y_{ij} = 1)$ for each j and is likely to have success each time; a subject with a large negative α_i has low $P(Y_{ij} = 1)$ for each j and is likely to have a failure each time. The greater the variability in $\{\alpha_i\}$, the greater the overall positive association between responses (across j), successes (failures) for observation 1 tending to occur with successes (failures) for observation 2. This is true for any β. The positive association reflects the shared value of α_i for each observation in a pair. No association occurs only when $\{\alpha_i\}$ are identical. Thus, the model does account for the dependence in matched pairs. Fitting it takes into account nonnegative association through the structure of the model.

For this momdel, the large number of $\{\alpha_i\}$ causes difficulties with the fitting process and with the properties of ordinary ML estimators (inconsistency of the ML estimator for β). One approach of remedy (i.e., the conditional ML approach) is to treat $\{\alpha_i\}$ as nuisance parameters and maximize the likelihood function for a conditional distribution that eliminates them. This is done by conditioning the distribution of the responses on sufficient statistics for nuisance parameters.

6.2.2 Conditional Maximum Likelihood Inference for Binary Matched Pairs

For model (22), assuming independence of responses for different subjects and for the two observations on the same subject (given parameters), the joint mass function for $\{(y_{11}, y_{12}), \ldots, (y_{n1}, y_{n2})\}$ is

$$\prod_{i=1}^{n} \left(\frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)} \right)^{y_{i1}} \left(\frac{1}{1 + \exp(\alpha_i)} \right)^{1-y_{i1}}$$

$$\times \left(\frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)}\right)^{y_{i2}} \left(\frac{1}{1+\exp(\alpha_i+\beta)}\right)^{1-y_{i2}}.$$

In terms of the data, this is proportional to

$$\exp\left[\sum_i \alpha_i(y_{i1}+y_{i2}) + \beta\left(\sum_i y_{i2}\right)\right].$$

To eliminate $\{\alpha_i\}$, we condition on their sufficient statistics, the pairwise success totals $\{S_i = y_{i1}+y_{i2}\}$. Given $S_i = 0$, $P(Y_{i1} = Y_{i2} = 0) = 1$, and given $S_i = 2$, $P(Y_{i1} = Y_{i2} = 1) =$. The distribution of (Y_{i1}, Y_{i2}) depends on β only when $S_i = 1$; that is, only when outcomes differ for the two responses. Given $y_{i1}+y_{i2} = 1$, the conditional distribution is

$$P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2} \mid S_i = 1)$$

$$= P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}) / \left[P(Y_{i1} = 1, Y_{i2} = 0) + P(Y_{i1} = 0, Y_{i2} = 1)\right]$$

$$\frac{\left(\frac{\exp(\alpha_i)}{1+\exp(\alpha_i)}\right)^{y_{i1}}\left(\frac{1}{1+\exp(\alpha_i)}\right)^{1-y_{i1}}\left(\frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)}\right)^{y_{i2}}\left(\frac{1}{1+\exp(\alpha_i+\beta)}\right)^{1-y_{i2}}}{\frac{\exp(\alpha_i)}{1+\exp(\alpha_i)}\frac{1}{1+\exp(\alpha_i)} + \frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)}\frac{1}{1+\exp(\alpha_i+\beta)}}$$

$$= \exp(\beta)/\left[1+\exp(\beta)\right], \text{ if } y_{i1} = 0, y_{i2} = 1,$$

$$= 1/\left[1+\exp(\beta)\right], \text{ if } y_{i1} = 1, y_{i2} = 0.$$

Again, let $\{n_{ab}\}$ denote the count for the four possible sequences. For subjects having $S_i = 1$, $\sum_i y_{i1} = n_{12}$, the number of subjects having success for observation 1 and failure for observation 2. Similarly, for those subjects, $\sum_i y_{i2} = n_{21}$ and $\sum_i S_i = n^\star = n_{12} + n_{21}$. Since n_{21} is the sum of $n^\star$ independent, identical Bernoulli variates, its conditional distribution is binomial with parameter $\exp(\beta)/\left[1+\exp(\beta)\right]$. For testing marginal homogeneity ($\beta = 0$), the parameter equals 1/2. In summary, the conditional analysis for the logit model implies that pairs in which $y_{i1} = y_{i2}$ are irrelevant to inference about β. When this model is realistic, it provides justification for comparing marginal distributions using only the $n_{21} + n_{12}$ pairings having outcomes in different categories at the two observations.

Conditional on $S_i = 1$, the joint distribution of the marched pairs is

$$\prod_{S_i=1} \left(\frac{1}{1+\exp(\beta)} \right)^{y_{i1}} \left(\frac{\exp(\beta)}{1+\exp(\beta)} \right)^{y_{i2}}$$

where the product refers to all pairs having $S_i = 1$. Differentiating the log of this conditional likelihood and equating to 0 and solving yields the conditional ML estimator of β in model (22). You can verify that $\hat{\beta} = \log(n_{21}/n_{12}),\ SE = \sqrt{1/n_{21} + 1/n_{22}}$.

6.2.3 Random Effects in Binary Matched-Pairs Model

An alternative remedy to handling the huge number of nuisance parameters in logit model (22) treats $\{\alpha_i\}$ as *random effects*. This regards $\{\alpha_i\}$ as an unobserved random sample from a probability distribution, usually assumed to be $N(\mu, \sigma^2)$ with unknown μ and σ. It eliminates $\{\alpha_i\}$ by averaging with respect to their distribution, yielding a marginal distribution. The likelihood function then depends on β as well as the $N(\mu, \sigma^2)$ parameters (i.e., μ and σ^2). It has only three parameters and is more manageable. This model is an example of a *generalized linear mixed model* (GLMM), containing both random effects (parameters) $\{\alpha_i\}$ and the fixed effect β.

Unless the outcome variable is normal, the likelihood function for a GLMM involves intractable integration. Therefore, numerical methods, such as Gaussian quadrature, Monte Carlo methods, penalized quasi-likelihood approximation, and Bayesian methods, need to be used for fitting the model and make inference on the model parameters and/or make predictions using random effects.