Categorical Data Analysis

Chapter 1

Deyuan Li
School of Management
Fudan University

Fall 2015

Outline

1. 1.1 Categorical Response Data
2. 1.2 Distributions for Categorical Data
3. 1.3 Statistical Inference for Categorical Data
4. 1.4 Statistical Inference for Binomial Parameters
5. 1.5 Statistical Inference for Multinomial Parameters

1.1 Categorical Response Data

Categorical variable（属性变量）: a measurement consisting of a set of categories.

For example:

(1). liberal, moderate or conservative in political philosophy;

(2). normal, benign, probably benign, suspicious or malignant for diagnoses of lung cancer.

1.1.1 Response-explanatory variables

Y: *Response* variable = *dependent* variable

X: *Explanatory* variable = *independent* variable

This book focuses on methods for categorical response variables.

1.1.2 Nominal-ordinal scale distinction

Two types of *Categorical* variables:

- *Nominal* variables（名义变量）: categories without a natural ordering;
  e.g., religious affiliation (Catholic, Protestant, Jewish, Muslim, other).
  The order of listing the categories is irrelevant.
- *Ordinal* variables（有序变量）: having ordered categories;
  e.g., social class (upper, middle, lower).
  The distances between categories make no sense.

1.1.2 Nominal-ordinal scale distinction

Interval variables have numerical distances between any two values; e.g., blood pressure level, annual income.

The way that a variable is measured determines its classification.

For example, "education" is

-
-
-

1.1.2 Nominal-ordinal scale distinction

Hierarchy: interval $>$ ordinal $>$ nominal.

Statistical methods for variables of one type can also be used for variables at higher levels, but not at lower levels.

(1) Methods for nominal variables can be used with ordinal variables by ignoring the ordering of categories.

(2) Methods for ordinal variables cannot be used with nominal variables.

However, it is usually best to apply methods appropriate for the actual scale.

1.1.3 Continuous-discrete variable

Continuous variables: take lots of values.
Discrete variables: take few values.

This book deals with 4 types of discrete response:

- nominal variables;
- ordinal variables;
- discrete interval variables having relatively few values;
- continuous variables grouped into a small number of categories.

1.1.4 Quantitative-qualitative variable

Nominal variables are *qualitative*.

Interval variables are *quantitative*.

The position of ordinal variables is fuzzy:

- qualitative
  $\Rightarrow$ using methods for nominal variables;
- quantitative
  $\Rightarrow$ assigning numerical scores （赋分）to categories.

Outline

1. 1.1 Categorical Response Data
2. 1.2 Distributions for Categorical Data
3. 1.3 Statistical Inference for Categorical Data
4. 1.4 Statistical Inference for Binomial Parameters
5. 1.5 Statistical Inference for Multinomial Parameters

1.2 Distributions for Categorical Data

Binomial distribution （二项分布）

Multinomial distribution（多项分布）

Poisson distribution（泊松分布）

1.2.1 Binomial distribution

- Bernoulli trial: binary observations with 1=success and 0=failure.

$$P(Y=1)=\pi, \quad P(Y=0)=1-\pi.$$

- $Y_1, Y_2, \cdots, Y_n$ are iid Bernoulli trials. Then

$$Y=\sum_{i=1}^{n} Y_i,$$

has the binomial distribution, denoted by $Y \sim \text{Bin}(n, \pi)$.

1.2.1 Binomial distribution

- For $y = 0, 1, 2, \ldots, n$,

$$P(Y = y) = \binom{n}{y} \pi^y (1-\pi)^{n-y} = \frac{n!}{y!(n-y)!} \pi^y (1-\pi)^{n-y}.$$

- Mean $\mu = E(Y) = n\pi$, variance $\sigma^2 = \text{var}(Y) = n\pi(1-\pi)$, skewness $E(Y-\mu)^3 = (1-2\pi)/\sqrt{n\pi(1-\pi)}$.

- By central limit theorem (CLT), as $n \to \infty$,

$$\frac{Y-\mu}{\sigma} = \frac{\sqrt{n}(\frac{1}{n}\sum_{i=1}^n Y_i - \pi)}{\sqrt{\pi(1-\pi)}} \xrightarrow{d} N(0,1).$$

1.2.2 Multinomial distribution

- Consider n iid trials. Each trial has c ($c > 2$) possible outcomes. For trial i ($i = 1, \ldots, n$), let

$$Y_{ij} = \begin{cases} 1 & \text{if outcome in category } j \quad (j = 1, \ldots, c); \\ 0 & \text{otherwise.} \end{cases}$$

  Let $\pi_j = P(Y_{ij} = 1)$ for all i, then $\sum_{j=1}^{c} \pi_j = 1$.

- $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ic})$ represents a multinomial trial, with $\sum_{j=1}^{c} Y_{ij} = 1$.

- $N_j = \sum_{i=1}^{n} Y_{ij}$ is the number of trials having outcome in category j, and $\sum_{j=1}^{c} N_j = n$.

1.2.2 Multinomial distribution

- The counts $(N_1, N_2, \ldots, N_c)$ have the multinomial distribution, with $(c-1)$ dimensions.
- The multinomial probability mass function is

$$P(N_j = n_j,\ j = 1, 2, ..., c) = \frac{n!}{n_1! n_2! \cdots n_c!}\, \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c}.$$

- Mean $\mu_j = E(N_j) = n\pi_j$, variance $\mathrm{var}(N_j) = n\pi_j(1-\pi_j)$, and covariance $\mathrm{cov}(N_j, N_k) = -n\pi_j\pi_k$ (shown in the next page).
- The marginal distribution of each N_j is binomial.

1.2.2 Multinomial distribution

Proofs.

1.2.2 Multinomial distribution

So,

1.2.2 Multinomial distribution

Another proof.

1.2.3 Poisson distribution

- $Y \sim \text{Poisson}(\mu)$:

$$P(Y = y) = (e^{-\mu}\mu^{y})/y!, \quad y = 0, 1, 2, \ldots.$$

$E(Y) = \text{var}(Y) = \mu$ and skewness $E(Y - \mu)^3 = 1/\sqrt{\mu}$.

- Unimodal（单峰）with the mode $[\mu]$, i.e

$$P(Y = [\mu]) > \max_{y \neq [\mu]} P(Y = y).$$

- Normal approximation:

$$\frac{Y - \mu}{\sqrt{\mu}} \xrightarrow{d} N(0, 1), \quad \text{as } \mu \to \infty.$$

Why?

1.2.4 Overdispersion

Overdispersion（超离散、过度离散）: the phenomenon when count observations exhibit variability exceeding that expected.

Suppose Y is a random variable with variance $\text{var}(Y|\mu)$ for given μ, but μ itself varies. Let $\theta = E(\mu)$, then unconditionally,

When $Y|\mu$ is Poisson, then $E(Y) = E(\mu) = \theta$ and $\text{var}(Y) = E(\mu) + \text{var}(\mu) = \theta + \text{var}(\mu) > \theta$.

1.2.4 Overdispersion

Proof of $\text{var}(Y) = E[\text{var}(Y|\mu)] + \text{var}[E(Y|\mu)]$:

The negative binomial （负二项分布）is a related distribution for count data that permits the variance to exceed the mean (Section 4.3.4).

1.2.5 Connection between Poisson and multinomial

Assume $Y_1, Y_2, ..., Y_c$ are independent and $Y_i \sim \text{Poisson}(\mu_i)$. Then $\sum_{i=1}^{c} Y_i \sim \text{Poisson}(\mu)$ with $\mu = \sum_{i=1}^{c} \mu_i$.

Condition on $\sum_{i=1}^{c} Y_i = n$, $(Y_1, ..., Y_c)$ is the multinomial $(n, \{\pi_i\})$ distribution, since

$$\begin{aligned} & P[(Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c) | \sum Y_j = n] \\ = \; & \frac{P(Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c)}{P(\sum Y_j = n)} = \frac{\prod_i [\exp(-\mu_i)\mu_i^{n_i}/n_i!]}{\exp(-\sum \mu_j)(\sum \mu_j)^n/n!} \\ = \; & \frac{n!}{\prod_i (n_i!)} \prod_i \pi_i^{n_i}, \quad \text{where } \pi_i = \mu_i/(\sum \mu_j). \end{aligned}$$

Outline

1. 1.1 Categorical Response Data
2. 1.2 Distributions for Categorical Data
3. 1.3 Statistical Inference for Categorical Data
4. 1.4 Statistical Inference for Binomial Parameters
5. 1.5 Statistical Inference for Multinomial Parameters

1.3 Statistical Inference for Categorical Data

What is estimation?

Parameter estimation: maximum likelihood (ML) estimation （极大似然估计）.

Under weak regularity, ML estimators have good properties:

1. they have large sample normal distributions;
2. they are asymptotically consistent;
3. they converge to the parameters as n increases;
4. they are asymptotically efficient, producing large-sample standard errors no greater than those from other estimation methods.

1.3.1 Likelihood functions

Given the data, for a chosen probability distribution the *likelihood function*（似然函数） is the probability of those data, treated as a function of the unknown parameter(s).

The ML estimate is the value of a parameter that maximizes this (log-) likelihood function;

—- that means, under this parameter value the data observed have the highest probability of occurrence.

Question: how to build likelihood function for censored data?

1.3.1 Likelihood functions

Let

$\boldsymbol{\beta} =$ the vector of parameters for a model,

$\mathcal{L}(\boldsymbol{\beta}) =$ the likelihood function,

$L(\boldsymbol{\beta}) = \log[\mathcal{L}(\boldsymbol{\beta})] =$ the log-likelihood function.

If $L(\boldsymbol{\beta})$ has concave shape, then the ML estimate $\hat{\boldsymbol{\beta}}$ is the solution of the likelihood equations

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0} \Longrightarrow \hat{\boldsymbol{\beta}}.$$

1.3.1 Likelihood functions

- Cov($\hat{\beta}$) is the inverse of the *information matrix*（信息矩阵）.
- The (j, k) element of the information matrix (I) is
- The standard error (SE) of β_j is
- The estimated SE is

1.3.2 ML estimate for binomial parameter

- The part of a likelihood function involving the parameters is called the *kernel*（核）. Since the maximization is with respect to the parameters, the rest is irrelevant.
- For the binomial distribution, the binomial coefficient $\binom{n}{y}$ has no influence on the ML estimate of π. Thus, we can just use the kernel as the likelihood function:

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

- Solve the equation $\partial L(\pi)/\partial\pi = 0$ and we obtain $\hat{\pi} = y/n$, i.e., the sample proportion of successes for the n trials.

1.3.2 ML estimate for binomial parameter

- Because $-E[\partial^2 L(\pi)/\partial\pi^2] = n/[\pi(1-\pi)]$, the asymptotic variance of $\hat{\pi}$ is $\pi(1-\pi)/n$.

- Recall that, for binomial variate Y,

$$E(Y) = n\pi \;\text{ and }\; \text{var}(Y) = n\pi(1-\pi).$$

- Hence, the distribution of $\hat{\pi} = Y/n$ has mean and SE:

$$\begin{aligned} E(\hat{\pi}) &= E(Y/n) = E(Y)/n = (n\pi)/n = \pi, \\ \sigma(\hat{\pi}) &= \sigma(Y/n) = \sigma(Y)/n = [\sqrt{n\pi(1-\pi)}]/n \\ &= \sqrt{\pi(1-\pi)/n} \\ &\quad \text{(as obtained from the information matrix).} \end{aligned}$$

1.3.3 Wald - likelihood ratio - score test triad

Three standard ways to perform large-sample inference using the likelihood function.

(1) Wald test

- For $H_0 : \beta = \beta_0$, given $\sigma(\hat{\beta}) \neq 0$, the test statistic

$$z = (\hat{\beta} - \beta_0)/\sigma(\hat{\beta})$$

  has an approximate $N(0, 1)$ when $\beta = \beta_0$.
- Equivalently, the z^2 has a χ^2_{df} with $df = 1$.

1.3.3 Wald - likelihood ratio - score test triad

χ_n^2 distribution with $df = n$. Its density function:

$$f(x; n) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \quad , x \geq 0,$$

with

$$\Gamma(t) := \int_0^\infty x^{t-1} e^{-x} dx.$$

Properties:

1. if $X_1, X_2, ..., X_n$ i.i.d. from $N(0, 1)$, then $\sum_{i=1}^n X_i^2 \sim \chi_n^2$;
2. $E(X) = n, \quad V(X) = 2n$;
3. if $X_1 \sim \chi_{n_1}^2, X_2 \sim \chi_{n_2}^2$ and $X_1 \perp X_2$, then $X_1 + X_2 \sim \chi_{n_1+n_2}^2$.

1.3.3 Wald - likelihood ratio - score test triad

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

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

- The asymptotic multivariate normal distribution for $\hat{\boldsymbol{\beta}}$ implies an asymptotic chi-square distribution for W.
- The df equal the rank of $\text{cov}(\hat{\boldsymbol{\beta}})$.

1.3.3 Wald - likelihood ratio - score test triad

(2) Likelihood-ratio test (似然比检验, LRT)

- $\mathcal{L}_0 =$ the maximum of the likelihood function under H_0,
  $\mathcal{L}_1 =$ the maximum of the likelihood function under $H_0 \cup H_a$.
- Then $\mathcal{L}_1 \geq \mathcal{L}_0$ and $\Lambda = \mathcal{L}_0/\mathcal{L}_1 \leq 1$.
- The LRT statistic equals

  It has a limiting χ^2_{df} as $n \to \infty$.
- $df = dim(H_0 \cup H_a) - dim(H_0)$: the difference in the dimensions of the parameter spaces under $H_0 \cup H_a$ and under H_0.

1.3.3 Wald - likelihood ratio - score test triad

(3) Score test（得分检验）

- $u(\beta_0) =$ the score function $\partial L(\beta)/\partial\beta$ evaluated at β_0,
  $\iota(\beta_0) =$ the information $-E[\partial^2 L(\beta)/\partial\beta^2]$ evaluated at β_0.
- The score test statistic is

  approximating $N(0, 1)$ or χ^2_{df} with $df = 1$.
- The multivariate extension has test statistic

$$[u(\boldsymbol{\beta})]' \left\{-E[\partial^2 L(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^2]\right\}^{-1} [u(\boldsymbol{\beta})], \quad \text{evaluated at } \boldsymbol{\beta} = \boldsymbol{\beta}_0.$$

1.3.3 Wald - likelihood ratio - score test triad

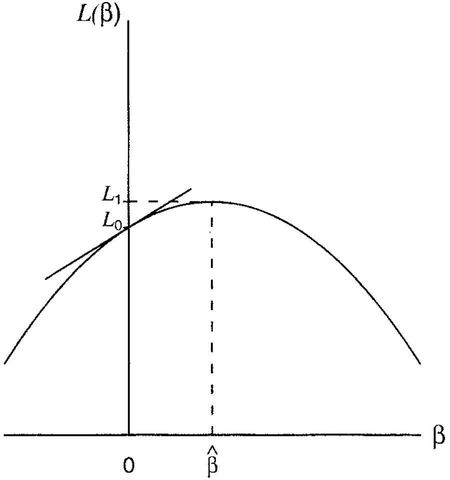

FIGURE 1.1 Log-likelihood function and information used in three tests of H_0: $\beta = 0$.

1.3.3 Wald - likelihood ratio - score test triad

Comparison:

1. To test $H_0 : \beta = 0$, the LRT statistic uses the most information (not only at $H_0 : \beta = 0$), and is the most useful.
2. As $n \to \infty$, the Wald, likelihood-ratio and score tests have certain asymptotic equivalences.
3. For small to moderate sample sizes, the LRT is usually more reliable than the Wald test.
4. Wald 检验直接比较$\hat{\beta} - \beta_0$，而得分检验比较$g(\hat{\beta}) - g(\beta_0)$，并且取$g = u$ (so that $u(\hat{\beta}) = 0$)。

1.3.4 Constructing confidence intervals

Let z_a be the $100(1-a)$ percentile of $N(0,1)$, i.e. $P(N(0,1) > z_a) = a$, and $\chi^2_{df}(a)$ be the $100(1-a)$ percentile of χ^2_{df}.

To construct a $100(1-\alpha)\%$ CI:

- Wald CI: contains the set of β_0 for which $|\hat{\beta}-\beta_0|/\mathrm{SE}(\hat{\beta}) < z_{\alpha/2} \Rightarrow \hat{\beta} \pm z_{\alpha/2}\,\mathrm{SE}(\hat{\beta})$.
- Likelihood-ratio-based CI: contains the set of β_0 for which $-2[L(\beta_0) - L(\hat{\beta})] < \chi^2_1(\alpha) = z^2_{\alpha/2}$.

The likelihood-ratio-based CI is preferred over the Wald CI for categorical data with small to moderate n.

Outline

1. 1.1 Categorical Response Data
2. 1.2 Distributions for Categorical Data
3. 1.3 Statistical Inference for Categorical Data
4. 1.4 Statistical Inference for Binomial Parameters
5. 1.5 Statistical Inference for Multinomial Parameters

1.4 Statistical Inference for Binomial Parameters

Recall:

1. The binomial distribution has one parameter π.
2. With y successes in n independent trials, the ML estimator of π is

$$\hat{\pi} = y/n.$$

1.4.1 Tests about a binomial parameter

Consider $H_0 : \pi = \pi_0$.

(1) The Wald statistic is

1.4.1 Tests about a binomial parameter

(2) The score test

- Evaluating the binomial score and information at π_0:

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

- The score statistic is

- The score statistic is preferable over the Wald statistic, as it uses the true π_0 in SE rather than the estimated $\hat{\pi}$.
- Its sampling distribution is closer to standard normal than that of the Wald statistic.

1.4.1 Tests about a binomial parameter

(3) The LRT

- $$L_0 = y \log(\pi_0) + (n - y) \log(1 - \pi_0), \text{ under } H_0,$$
  $$L_1 = y \log(\hat{\pi}) + (n - y) \log(1 - \hat{\pi}), \text{ under } H_0 \cup H_a.$$
- The LRT statistic simplifies to
- This statistic has an asymptotic χ^2_{df} with $df = 1$.

1.4.2 Confidence intervals for a binomial parameter

(1) Wald CI

$|z_W| < z_{\alpha/2}$ implies

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

This CI performs poorly unless n is very large.

1.4.2 Confidence intervals for a binomial parameter

(2) Score CI

- The score CI contains π_0 values for which $|z_S| < z_{\alpha/2}$, i.e., its endpoints are the π_0 solutions to the equations

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

- First discussed by Wilson (1927), the CI is

$$\hat{\pi}\Big(\frac{n}{n + z^2_{\alpha/2}}\Big) + \Big(\frac{1}{2}\Big)\Big(\frac{z^2_{\alpha/2}}{n + z^2_{\alpha/2}}\Big)$$

$$\pm\, z_{\alpha/2}\sqrt{\frac{1}{n + z^2_{\alpha/2}}\Big[\hat{\pi}(1 - \hat{\pi})\Big(\frac{n}{n + z^2_{\alpha/2}}\Big) + \Big(\frac{1}{2}\Big)\Big(\frac{1}{2}\Big)\Big(\frac{z^2_{\alpha/2}}{n + z^2_{\alpha/2}}\Big)\Big]}.$$

1.4.2 Confidence intervals for a binomial parameter

(3) Likelihood-ratio-based CI

This CI is more complex computationally, but simple in principle.

It is the set of π_0 such that

$$-2(L_0 - L_1) \leq \chi^2_{df}(\alpha).$$

1.4.3 Proportion of vegetarians example

A survey was conducted in 25 students. $\Rightarrow n = 25$

One question asked each student if he or she was a vegetarian.
$\Rightarrow$ Binary outcome.

None of the students answered "yes". $\Rightarrow y = 0$
$\Rightarrow \hat{\pi} = 0/25 = 0$

(1) Wald 95% CI: $0 \pm 1.96\sqrt{[0 \times (1-0)]/25} = (0, 0)$.

When the observation falls at the boundary of the sample space, often Wald methods do not provide sensible answers.

1.4.3 Proportion of vegetarians example

(2) Score 95% CI

$$\left(\frac{z^2_{\alpha/2}}{n+z^2_{\alpha/2}}\right) \pm \left(\frac{1}{2}\right)\left(\frac{z^2_{\alpha/2}}{n+z^2_{\alpha/2}}\right) = (0,\ \frac{1.96^2}{25+1.96^2}) = (0,\ 0.133).$$

This is a more believable inference than the Wald CI.

For $H_0 : \pi = 0.5$, the score statistic

$$z_S = (0 - 0.5)/\sqrt{(0.5 \times 0.5)/25} = -5.0.$$

Since $|z_S| > 1.96$, $\pi_0 = 0.5$ does not fall in the 95% CI.

For $H_0 : \pi = 0.1$, the score statistic

$$z_S = (0 - 0.1)/\sqrt{(0.1 \times 0.9)/25} = -1.67.$$

Since $|z_S| < 1.96$, $\pi_0 = 0.1$ falls in the 95% CI.

1.4.3 Proportion of vegetarians example

(3) Likelihood-ratio-based 95% CI

The log likelihood is $L(\pi) = 25 \log(1 - \pi)$.

Note that $L(\hat{\pi}) = L(0) = 0$.

$$\begin{aligned} -2(L_0 - L_1) &= -2[L(\pi_0) - L(\hat{\pi})] \\ &= -50 \log(1 - \pi_0) \leq \chi_1^2(0.05) = 3.84\,. \end{aligned}$$

$\Rightarrow \pi_0 \leq 1 - \exp(-3.84/50) = 0.074, \quad \Rightarrow \text{CI} = (0,\, 0.074)$.

Comments:

1. The 3 large-sample methods yield quite different results.
2. When π is near 0, the sampling distribution of $\hat{\pi}$ is highly skewed to the right for small n.

Outline

1 1.1 Categorical Response Data

2 1.2 Distributions for Categorical Data

3 1.3 Statistical Inference for Categorical Data

4 1.4 Statistical Inference for Binomial Parameters

5 1.5 Statistical Inference for Multinomial Parameters

1.5 Statistical Inference for Multinomial Parameters

Consider c categories.

- n observations in total,
- n_j observations occur in category $j = 1, \ldots, c$.
- Then $\sum_{j=1}^{c} n_j = n$.

We now present inference for multinomial parameters $\{\pi_j\}$.

1.5.1 Estimation of multinomial parameters

The ML estimates of $\{\pi_j\}$ are those that maximize the kernel

$$\prod_{j=1}^{c} \pi_j^{n_j} \quad \text{where all } \pi_j \geq 0 \text{ and } \sum_{j=1}^{c} \pi_j = 1,$$

or the log-likelihood function

$$L(\boldsymbol{\pi}) = \sum_{j=1}^{c} n_j \log \pi_j.$$

1.5.1 Estimation of multinomial parameters

Since $\pi_c = 1 - (\pi_1 + \cdots + \pi_{c-1})$, for $j = 1, \ldots, c-1$

$$\frac{\partial \pi_c}{\partial \pi_j} = -1 \quad \Rightarrow \quad \frac{\partial \log \pi_c}{\partial \pi_j} = \frac{1}{\pi_c} \frac{\partial \pi_c}{\partial \pi_j} = -\frac{1}{\pi_c}.$$

Differentiating $L(\boldsymbol{\pi})$ with respect to π_j gives

$$\frac{\partial L(\boldsymbol{\pi})}{\partial \pi_j} = \frac{n_j}{\pi_j} - \frac{n_c}{\pi_c} = 0 \quad \Rightarrow \quad \frac{\hat{\pi}_j}{\hat{\pi}_c} = \frac{n_j}{n_c} \quad \Rightarrow \quad \hat{\pi}_j = \frac{\hat{\pi}_c \, n_j}{n_c}.$$

Due to

$$\sum_{j=1}^{c} \hat{\pi}_j = 1 = \frac{\hat{\pi}_c}{n_c} \sum_{j=1}^{c} n_j = \frac{\hat{\pi}_c \, n}{n_c} \quad \Rightarrow \quad \hat{\pi}_c = \frac{n_c}{n} \quad \Rightarrow \quad \hat{\pi}_j = \frac{n_j}{n}.$$

Thus, the ML estimates of $\{\pi_j\}$ are the sample proportions.

1.5.2 Pearson statistic for testing a specified multinomial

Consider $H_0 : \pi_j = \pi_{j0},\ j = 1, \ldots, c$, where $\sum_j \pi_{j0} = 1$.

- When H_0 is true, the expected values of $\{n_j\}$, called *expected frequencies*, are $\mu_j = n\pi_{j0}$.
- The test statistic is

  For fixed n, greater differences $\{n_j - \mu_j\}$ imply greater X^2.

P-value

P-value: *The P-value of a test is the (asymptotic) probability of the observing a test statistic at least as extreme as the one computed given that the null hypothesis is true.*

One-side, two-sides P-value.

For example, Wald test statistic $z = (\hat{\beta} - \beta_0)/\sigma(\hat{\beta})$ for testing $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$. Let $X \sim N(0, 1)$ and z_0 is the value of z.

One-side P-value is

$P(X \geq z_0)$ (may be used for $H_1 : \beta > \beta_0$) or
$P(X \leq z_0)$ (may be used for $H_1 : \beta < \beta_0$);

Two-sides P-value is $P(X \geq |z_0|)$ (may be used for $H_1 : \beta \neq \beta_0$).

Since standard normal is not exactly the distribution of z, it is asymptotic P-value.

Summary of a test of hypothesis

Summary of a test of hypothesis

Judgement:

- if P-value $< \alpha$, reject H_0;
- if P-value $> \alpha$, not reject H_0;
- If P-value is very close to α, be much careful to make a conclusion;
- never say "accept H_0", "accept H_1" or "reject H_1".

Conclusions of a test of hypothesis:

- if we reject H_0, we conclude that there is enough statistical evidence to infer that the alternative hypothesis is true;
- if we do not reject H_0, we conclude that there is not enough statistical evidence to infer that the alternative hypothesis is true.

Mostly take what we are interested as alternative hypothesis.

1.5.2 Pearson statistic for testing a specified multinomial

- Let X_o^2 denote the observed value of X^2. Define

$$P\text{-value } = P(X^2 \geq X_o^2 | \pi_j = \pi_{j0}, j = 1, \ldots, c).$$

- For large sample, X^2 approximates χ^2_{df} with $df = c - 1$.
- The P-value is approximated by $P(\chi^2_{c-1} \geq X_o^2)$.
- If P-value$< \alpha = 0.05$, reject H_0; If P-value$> \alpha = 0.05$, do not reject H_0.
- This test statistic is called Pearson chi-squared statistic.

1.5.3 Example: Testing Mendel's theories

Mendel crossed pea plants of pure yellow strain (dominant，显性) with plants of pure green strain (recessive，隐性).

He predicted that second-generation hybrid seeds would be 75% yellow and 25% green.

One experiment produced $n = 8023$ seeds, of which $n_1 = 6022$ were yellow and $n_2 = 2001$ were green.

1.5.3 Example: Testing Mendel's theories

$c = 2$ and test $H_0 : \pi_{10} = 0.75$, $\pi_{20} = 0.25$.

Solution.

- The expected frequencies are
  $\mu_1 = 8023 \times 0.75 = 6017.25$ and
  $\mu_2 = 8023 \times 0.25 = 2005.75$.
- The Pearson statistic

$$X^2 = \frac{(6022 - 6017.25)^2}{6017.25} + \frac{(2001 - 2005.75)^2}{2005.75} = 0.015$$

  with df$=1$ has a P-value of 0.90.

- $\Rightarrow$ Mendel's hypothesis is not rejected.

1.5.3 Example: Testing Mendel's theories

Theorem. *If $X_1^2, \ldots, X_k^2$ are independent chi-squared statistics with degrees of freedom $\nu_1, \ldots, \nu_k$, then $\sum_i X_i^2$ has a chi-squared distribution with df$= \sum_i \nu_i$.*

- Mendel performed 84 experiments of this type.
- Based on Mendel's data, R. A. Fisher obtained a summary chi-squared statistic equal to 42, with df$=$ 84 and the P-value was 0.99996. That means

$$\sum_{i=1}^{84} X_i^2 = 42, \quad P(\chi_{84}^2 \geq 42) \approx 0.99996.$$

- Fisher thought that the fit seemed too good.

  $\Rightarrow$ Was Mendel deceived by a gardening assistant?

1.5.5 Likelihood-ratio chi-squared

Recall: the kernel of the multinomial likelihood is $\prod_j \pi_j^{n_j}$.

Under H_0 the likelihood is maximized with $\pi_j = \pi_{j0}$. Since the $\{\pi_{j0}\}$ are specified completely, the dimension (df) is 0.

In the general case, it is maximized when $\hat{\pi}_j = n_j/n$. The $\{\pi_j\}$ are subject to $\sum_j \pi_j = 1$, so the dimension (df) is $c - 1$.

The ratio of the likelihood equals $\Lambda = [\prod_j \pi_{j0}^{n_j}]/[\prod_j (n_j/n)^{n_j}]$.

Thus the likelihood-ratio statistic is

$$G^2 = -2\log\Lambda = -2\log\prod_j (n\pi_{j0}/n_j)^{n_j} = 2\sum_j n_j \log(n_j/n\pi_{j0}).$$

1.5.5 Likelihood-ratio chi-squared

As for the binomial test in Section 1.4.1, G^2 has form

$$2\sum(\text{observed}) \log\left(\frac{\text{observed}}{\text{fitted}}\right)$$

and is called the *likelihood-ratio chi-squared statistic*.

The larger the value of G^2, the greater the evidence against H_0.

For large n, G^2 has a chi-squared null distribution with df$= (c-1) - 0 = c - 1$.

1.5.5 Likelihood-ratio chi-squared

Comparison between the Pearson X^2 and the likelihood-ratio statistics G^2:

1. When H_0 holds, the Pearson X^2 and the likelihood ratio G^2
   - both have asymptotic χ^2_{df} with $df = c - 1$;
   - both are asymptotically equivalent, i.e., $X^2 - G^2$ converges in probability to zero (Section 14.3.4).

2. When H_0 is false, X^2 and G^2 tend to grow proportionally to n; they need not to take similar values even for very large n.

1.5.5 Likelihood-ratio chi-squared

3. For fixed c, as n increases the distribution of X^2 usually converges to chi-squared more quickly than that of G^2.
4. The chi-squared approximation is usually poor for G^2 when $n/c < 5$.
5. When c is large, X^2 still works for n/c as small as 1 if there are no very small and moderately large expected frequencies.

1.5.6 Testing with estimated expected frequencies

Pearson's X^2 compares a sample distribution with a hypothetical (**known**) one $\{\pi_{j0}\}$.

In some applications, $\{\pi_{j0}\}$ are functions of a smaller set of **unknown** parameters $\boldsymbol{\theta}$, i.e., $\{\pi_{j0}\} = \{\pi_{j0}(\boldsymbol{\theta})\}$.

ML estimates $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$
$\Rightarrow$ determine ML estimates $\{\pi_{j0}(\hat{\boldsymbol{\theta}})\}$ of $\{\pi_{j0}\}$
$\Rightarrow$ determine ML estimates $\{\hat{\mu}_j = n\pi_{j0}(\hat{\boldsymbol{\theta}})\}$ of $\{\mu_j\}$ in X^2.

Replacing $\{\mu_j\}$ by estimates $\{\hat{\mu}_j\}$ affects the distribution of X^2.

When $\dim(\boldsymbol{\theta}) = p$, the true $df = (c-1) - p$.

1.5.6 Testing with estimated expected frequencies

Example.

- $n = 156$ dairy calves were classified according to whether they caught pneumonia（肺炎）within 60 days of birth (primary infection).
- Calves that got a pneumonia infection were also classified according to whether they got a secondary infection within 2 weeks after the first infection cleared up.

TABLE 1.1 Primary and Secondary Pneumonia Infections in Calves

	Secondary Infection	
Primary Infection	Yes	No
Yes	$n_{11} = 30$ (38.1)	$n_{12} = 63$ (39.0)
No	$n_{21} = 0$ (–)	$n_{22} = 63$ (78.9)

Values in parenthesis are estimated expected frequencies.

1.5.6 Testing with estimated expected frequencies

- Calves that did not get a primary infection could not get a secondary infection, so no observations can fall in the category for "no" primary infection and "yes" secondary infection.

  That combination is called a *structural zero*.

- One goal of the study was to test whether the prob. of primary infection was the same as the conditional prob. of secondary infection, given that the calf got the primary infection.

1.5.6 Testing with estimated expected frequencies

- Let π_{ab} denote the probability that a calf is classified in row a and column b of Table 1.1.
- Test
- Let $\pi = \pi_{11} + \pi_{12}$ denote the probability of primary infection.
- Equivalently test

1.5.6 Testing with estimated expected frequencies

Solution.

- $c = 3$, three categories: yes-yes, yes-no, no-no.

TABLE 1.2 Probability Structure for Hypothesis

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

- Kernel of the multinomial likelihood

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

- The log likelihood is

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

and

$$\frac{dL(\pi)}{d\pi} = \frac{2n_{11}}{\pi} + \frac{n_{12}}{\pi} - \frac{n_{12}}{1-\pi} - \frac{n_{22}}{1-\pi} = 0.$$

1.5.6 Testing with estimated expected frequencies

- The solution is $\hat{\pi} = (2n_{11} + n_{12})/(2n_{11} + 2n_{12} + n_{22})$.
- For Table 1.1, $\hat{\pi} = 0.494$. Then
  $\hat{\mu}_{11} = n\hat{\pi}^2 = 38.1, \quad \hat{\mu}_{12} = n(\hat{\pi} - \hat{\pi}^2) = 39.0,$
  $\hat{\mu}_{22} = n(1 - \hat{\pi}) = 78.9.$
- The Pearson's statistic $X^2 = 19.7$ with
  df$= (c-1) - p = (3-1) - 1 = 1$ and P-value = 0.00001.
  Strong evidence against H_0
- Many more calves got a primary infection but not a secondary infection than H_0 predicts.

  $\Rightarrow$ The primary infection had an immunizing effect that reduced the likelihood of a secondary infection.