Financial Time Series I

Topic 4: Discrete Data: Contingency Tables

Hung Chen
Department of Mathematics
National Taiwan University

10/30/2002

OUTLINE

1. Probability Model
   - Mean and Variance
   - Limiting Distribution
   - Mining of association rules in Basket Analysis
2. Goodness of Fit Test
   - Embedding and Nested Models
   - Test of Independence
3. Logistic Regression for Binary Response
   - Logit or Probit
   - Likelihood Equation
   - Likelihood Ratio Test
4. Tests for A Simple Null Hypothesis
   - Likelihood Ratio Test
   - Wald Test
   - Rao's Score test

Contingency Tables

We start with a probability model to describe the data summarized in terms of contingency table.

- Consider a sequence of n independent trials, with k possible outcomes for each trial. For a 2×2 table, $k = 4$ and n is the total number of observations.
- Let p_j denote the probability of occurrence of the jth outcome in any given trial ($\Sigma_1^k p_j = 1$).
- Let n_j denote the number of occurrences of the jth outcome in the series of n trials ($\Sigma_i^k n_j = n$). $(n_1, \ldots, n_k)$ is called the "cell frequency vector" associated with the n trials.
- The exact distribution of $(n_1, \ldots, n_k)$ is the multinomial distribution $MN(n, \boldsymbol{p})$ where $\boldsymbol{p} = (p_1, \ldots, p_k)$.
- $E(n_i) = np_i$, $Var(n_i) = np_i(1 - p_i)$ and $Cov(n_i, n_j) = -np_ip_j$, so that $E(n_1, \ldots, n_k) = n\boldsymbol{p}$, $Cov((n_1, \ldots, n_k)) = n(\boldsymbol{D}_p - \boldsymbol{p}^t\boldsymbol{p})$, where $\boldsymbol{D}_p = diag(\boldsymbol{p})$.

- Let $\hat{\boldsymbol{p}} = n^{-1}(n_1, \ldots, n_k)$ be the vector of sample proportions, and set $\boldsymbol{U}_n = \sqrt{n}(\hat{\boldsymbol{p}} - \boldsymbol{p})$. Then $E(\boldsymbol{U}_n) = \mathbf{0}$, $Cov(\boldsymbol{U}_n) = \boldsymbol{D}_p - \boldsymbol{p}^t\boldsymbol{p}$.

We now use "Cramer-Wold device" to prove *asymptotic multivariate normality of cell frequency vectors.*

Theorem. The random vector $\boldsymbol{U}_n$ converges in distribution to k-variate normal with mean $\mathbf{0}$ and covariance $\boldsymbol{D}_p - \boldsymbol{p}^t\boldsymbol{p}$.

- Compute the characteristic function of $E \exp(it \, \Sigma_{i=1}^n u_i)$ where $\boldsymbol{U}_n = (u_1, \ldots, u_k)$.
- Observe that

$$\begin{aligned}
&E\left(\exp\left[it \sum_{j=1}^k \lambda_j u_j\right]\right) \\
&= E\left(\exp\left[\sum_{j=1}^k it\lambda_j \left(\frac{n_j}{\sqrt{n}} - \sqrt{n}p_j\right)\right]\right) \\
&= \exp\left(-it\sqrt{n} \sum_{j=1}^k \lambda_j p_j\right) \cdot E\left(\exp\left[\frac{it}{\sqrt{n}} \sum_{j=1}^k \lambda_j n_j\right]\right) \\
&= \exp\left(-it\sqrt{n} \sum_{j=1}^k \lambda_j p_j\right) \cdot \left(\sum_{j=1}^k p_j \exp\left(\frac{it}{\sqrt{n}}\lambda_j\right)\right)^n \\
&= \left(\sum_{j=1}^k p_j \cdot \exp\left[\frac{it}{\sqrt{n}}\left(\lambda_j - \sum_{i=1}^k \lambda_i p_i\right)\right]\right)^n
\end{aligned}$$

$$
\begin{aligned}
&= \Bigg\{\sum_{j=1}^{k} p_j \Bigg[1 + \frac{it}{\sqrt{n}}(\lambda_j - \sum_{i=1}^{k} \lambda_i p_i) - \frac{t^2}{2n}(\lambda_j - \sum_{i=1}^{k} \lambda_i p_i)^2 + o \\
&= \Bigg\{1 - \frac{t^2}{2n} \sum_{j=1}^{k} p_j \left(\lambda_j - \sum_{i=1}^{k} \lambda_i p_i\right)^2 + o(n^{-1})\Bigg\}^n \\
&\longrightarrow \exp\left(-\frac{1}{2}(\lambda_1, \ldots, \lambda_k)(\mathbf{D}_p - \mathbf{p}^t\mathbf{p})(\lambda_1, \ldots, \lambda_k)^t\right).
\end{aligned}
$$

- The limit being the ch.f. of the multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{D}_p - \boldsymbol{p}^t\boldsymbol{p}$.

Assumptions:

- Every individual in the population under study can be classified as falling into one and only one of k categories, we say that the categories are mutually exclusive and exhaustive.

- A randomly selected member of the population will fall into one of the k categories with probability $\mathbf{p}$, where $\mathbf{p}$ is the vector of cell probabilities

$$\mathbf{p} = (p_1, p_2, \ldots, p_k)$$

and $\Sigma_{i=1}^{k} p_i = 1$.

- Here the cells are strung out into a line for purposes of indexing only; their arrangement and ordering does not reflect anything about the characteristics of individuals falling into a particular cell.
- The p_i reflect the relative frequency of each category in the population.
- Mining of association rules in Basket Analysis:
  - A basket bought at the food store consists of the following four items: Apples, Bread, Coke, Milk, Tissues.
  - Data on all baskets is available (through cash registers)
  - Goal: Discover association rules of the form
Bread&Milk =¿ Coke&Tissue
  - This analysis is also called linkage analysis or item analysis.
  - Properties of association rules:
    - The support of the rule is the *Proportion of baskets with Bread&Milk&Coke&Tissue.*

  - ∗ The confidence of the rule is the Sup (Bread&Milk&Coke&Tissue)/Sup(Bread&Milk) which is simply the estimated conditional probability in statistical terms.
  - ∗ The lift of the rule is the Sup (B&M&C&T)/Sup(B&M)Sup(C&T). How do you connect it with $P(A \cap B)/P(A)P(B)$?
- – Search for rules with high confidence and support
  - ∗ Will the results be affected by randomness?
  - ∗ Add the requirement that the rule is statistically significant in the test against independence (i.e. against lift=1)
  - ∗ The number of such tests to be performed in a moderate problem reaches tens of thousands
- – You can put all of them in a *huge* contingency table.

2×2 Tables

- As an example, we might be interested in *whether hair color is related to eye color*. Conduct a study by collecting a random sample and get a count of the number of people who fall in this particular cross-classification determined by hair color and eye color.
- When the cells are defined in terms of the categories of two or more variables, a structure relating to the nature of the data is imposed. The natural structure for two variables is often a rectangular array with columns corresponding to the categories of one variable and rows to categories of the second variable; three variables creates layers of two-way tables, and so on.
- The simplest contingency table is based on four cells, and the categories depend on two variables. The four cells are arranged in a 2×2 table whose two rows correspond to the categorical variable A and whose two columns correspond to the second categorical variable B. Double subscripts refer to the position of the cells in our arrangement.

- The first subscript gives the category number of variable A, the second of variable B, and the two-dimensional array is displayed as a grid with two rows and two columns.

- The probability p_{ij} is the probability of an individual being in category i of variable A and category j of variable B. Usually, we have some theory in mind which can be checked in terms of hypothesis testing such as

$$H_0 : \boldsymbol{p} = \boldsymbol{\pi} \quad (\boldsymbol{\pi} \text{ a fixed value}).$$

- Then the problem can phrased as n observations from the k-cell multinomial distribution with cell probabilities $p_1, \ldots, p_k$. Then we encounter the problem of *proving asymptotic multivariate normality of cell frequency vectors.*

- To test H_0, it can be proceed by the Pearson chi square test, which is to reject H_0 if X^2 is too large, where

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i}.$$

This test statistic was first derived by Pearson (1900). Then we need to answer two questions. The first one is to determine what kind of the magnitude of X^2 is the so-called *too large*. The second one is whether the Pearson chi-square test is a reasonable testing procedure. These questions will be tackled by deriving the asymptotic distribution of the Pearson chi square statistic under H_0 and a local alternative of H_0.

- Using matrix notation, X^2 can be written as
$$X^2 = \boldsymbol{U}_n \mathbf{D}_{\boldsymbol{\pi}}^{-1} \boldsymbol{U}_n^t,$$
where
$$\mathbf{U}_n = \sqrt{n}(\hat{\mathbf{p}} - \boldsymbol{\pi}), \quad \hat{\boldsymbol{p}} = n^{-1}(n_1, \ldots, n_k), \;\; \text{and } \mathbf{D}_{\boldsymbol{\pi}} = diag$$
- Let $g(\mathbf{x}) = \mathbf{x}\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{x}^t$ for $\mathbf{x} = (x_1, \ldots, x_k)$. Evidently, g is a continuous function of $\mathbf{x}$. It can be shown that $\mathbf{U}_n \xrightarrow{d} \mathbf{U}$, where $\mathbf{U}$ has the multivariate normal distribution $\mathcal{N}(0, \mathbf{D}_{\boldsymbol{\pi}} - \boldsymbol{\pi}^t\boldsymbol{\pi})$. Then we have
$$\mathbf{U}_n \mathbf{D}_{\boldsymbol{\pi}}^{-1} \mathbf{U}_n^t \xrightarrow{d} \mathbf{U}\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{U}^t.$$
Thus the asymptotic distribution of X^2 under H_0, which is the distribution of $\mathbf{U}\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{U}^t$,

where $\mathbf{U}$ has the $\mathcal{N}(0, \mathbf{D}_{\boldsymbol{\pi}} - \boldsymbol{\pi}^t\boldsymbol{\pi})$ distribution. This reduces the problem to finding the distribution of a quadratic form of a multivariate normal random vector. The above process is the so-called δ method.

- Now we state without proof the following general result on the distribution of a quadratic form of a multivariate normal random variable. It can be found in Chapter 3b in Rao (1973) and Chapter 3.5 of Serfling (1980).
  Theorem If $\mathbf{X} = (X_1, \ldots, X_d)$ has the multivariate normal distribution $\mathcal{N}(0, \Sigma)$ and $Y = \mathbf{X}\mathbf{A}\mathbf{X}^t$ for some symmetric matrix $\mathbf{A}$, then $\mathcal{L}[Y] = \mathcal{L}[\Sigma_{i=1}^d \lambda_i Z_i^2]$, where $Z_1^2, \ldots, Z_d^2$ are independent chi square variables with one degree of freedom each and $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $\mathbf{A}^{1/2}\Sigma(\mathbf{A}^{1/2})^t$.

- Apply the above theorem to the present problem, we see that $\mathcal{L}[\mathbf{U}\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{U}^t] = \mathcal{L}[\Sigma_{i=1}^d \lambda_i Z_i^2]$, where λ_i are the eigenvalues of

$$\mathbf{B} = \mathbf{D}_{\boldsymbol{\pi}}^{-1/2}(\mathbf{D}_{\boldsymbol{\pi}} - \boldsymbol{\pi}^t\boldsymbol{\pi})\mathbf{D}_{\boldsymbol{\pi}}^{-1/2} = \mathbf{I} - \sqrt{\boldsymbol{\pi}}^t\sqrt{\boldsymbol{\pi}},$$

  where $\sqrt{\boldsymbol{\pi}} = (\sqrt{\pi}_1, \ldots, \sqrt{\pi}_k)$.

- Now it remains to find the eigenvalues of

$\mathbf{B}$. Since $\mathbf{B}^2 = \mathbf{B}$ and $\mathbf{B}$ is symmetric, the eigenvalues of $\mathbf{B}$ are all either 1 or 0. Moreover,

$$\sum_{i=1}^{k} \lambda_i = tr(\mathbf{B}) = k - 1.$$

- Therefore, we establish the result that under the simple hypothesis H_0, Pearson's chi-square statistic X^2 has an asymptotic chi square distribution with $k - 1$ degrees of freedom.

Remarks:

- We already examined the limiting distribution of the Pearson chi square statistic under H_0 by employing δ method.
- In essence, the δ method requires two ingredients:
  first, a random variable (which we denote here by $\hat{\theta}_n$) whose distribution depends on a real-valued parameter θ in such a way that

$$\mathcal{L}[\sqrt{n}(\hat{\theta}_n - \theta)] \to N(0, \sigma^2(\theta)); \qquad (1)$$

and second, a function $f(x)$ that can be differentiated at $x = \theta$ so that it possesses the following expansion about θ:

$$f(x) = f(\theta) + (x-\theta)f^{'}(\theta) + o(|x-\theta|) \text{ as } x \to \theta. \tag{2}$$

- The δ method for finding approximate means and variances (asymptotic mean and asymptotic variance) of a function of a random variable is justified by the following theorem.

  Theorem (The one-dimensional δ method.) If $\hat{\theta}_n$ is a real-valued random variable and θ is a real-valued parameter such that (1) holds, and if f is a function satisfying (2), then the asymptotic distribution of $f(\hat{\theta}_n)$ is given by

  $$\mathcal{L}[\sqrt{n}(f(\hat{\theta}_n) - f(\theta))] \to N(0, \sigma^2(\theta)[f^{'}(\theta)]^2). \tag{3}$$

  Proof. Set $\Omega_n = R$, $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n \times \cdots = \times_{n=1}^{\infty}\Omega_n$, and P_n to be the probability distribution of $\hat{\theta}_n$ on R. Note that Ω is the set of all sequences $\{t_n\}$ such that $t_n \in \Omega_n$. We define two subsets of Ω:

  $$\begin{aligned} S &= \{\{t_n\} \in \Omega : t_n - \theta = O(n^{-1/2})\}, \\ T &= \{\{t_n\} \in \Omega : f(t_n) - f(\theta) - (t_n - \theta)f^{'}(\theta) = o(n^{-1/2})\} \end{aligned}$$

Since f satisfies (2), then $S \subset T$. By (1), we have

$$n^{1/2}(\hat{\theta}_n - \theta) = O_P(1) \text{ and hence } \hat{\theta}_n - \theta = O_P(n^{-1/2}). \tag{4}$$

Note that S occurs in probability and hence T also occur in probability since $S \subset T$. Finally,

$$f(\hat{\theta}_n) - f(\theta) - (\hat{\theta}_n - \theta)f^{'}(\theta) = o_P(n^{-1/2}) \tag{5}$$

or

$$\sqrt{n}(f(\hat{\theta}_n) - f(\theta)) = \sqrt{n}(\hat{\theta}_n - \theta)f^{'}(\theta) + o_P(1). \tag{6}$$

Now let $V_n = \sqrt{n}(f(\hat{\theta}_n) - f(\theta))$, $U_n = \sqrt{n}(\hat{\theta}_n - \theta)$, and $g(x) = xf^{'}(\theta)$ for all real numbers x. Then (6) may be rewritten as

$$V_n = g(U_n) + o_P(1).$$

Goodness-of-Fit to Composite Multinomial Models

Consider a sample from a population in genetic equilibrium with respect to a single gene with two alleles. If we assume the three different genotypes are identifiable, we are led to suppose that there are three types of individuals whose frequencies are given by the so-called *Hardy-Weinberg proportions*

$$p_1 = \theta^2, p_2 = 2\theta(1-\theta), p_3 = (1-\theta)^2,$$

where $0 < \theta < 1$.

- In the Hardy-Weinberg model, the probability model to describe the data is multinomial with parameter falling in
$$\Theta = \{\boldsymbol{\theta} : \theta_i \geq 0, 1 \leq i \leq 3, \sum_{i=1}^{3} \theta_i = 1\}.$$
- The theory we want to test can be described by a multinomial with parameter falling in
$$\Theta_0 = \{(\eta^2, 2\eta(1-\eta), (1-\eta)^2, 0 \leq \eta \leq 1\},$$
which is a one-dimensional curve in the two-dimensional parameter space Θ.
- To test the adequancy of the Hardy-Weinberg model means testing $H_0 : \boldsymbol{\theta} \in \Theta_0$ versus $H_1 : \boldsymbol{\theta} \in \Theta_1$ where $\Theta_1 = \Theta - \Theta_0$.

In general, we can describe Θ_0 parametrically as

$$\Theta_0 = \{(\theta_1(\boldsymbol{\eta}), \ldots, \theta_k(\boldsymbol{\eta})) : \theta(\boldsymbol{\eta} \in \Xi\},$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_q)^T$, is a subset of q-dimensional space, and the map $\boldsymbol{\eta} \to (\theta_1(\boldsymbol{\eta}), \ldots, \theta_k(\boldsymbol{\eta}))^T$ takes Ξ into Θ_0. To avoid trivialities we assume $q < k - 1$.

Now we consider the likelihood ratio test for H_0 versus H_1.

- Let $p(n_1, \ldots, n_k, \boldsymbol{\theta})$ denote the frequency function.
  - Maximizing $p(n_1, \ldots, n_k, \boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta_0$.
  - Denote the maximizer by $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \ldots, \hat{\eta}_q)$.
  - The logarithm of

    $$\sup_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta}; \mathbf{x})$$

    is $\Sigma_{i=1}^k n_i \log \theta_i(\hat{\boldsymbol{\eta}})$ up to a constant.
- The logarithm of

  $$\sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \mathbf{x})$$

  is $\Sigma_{i=1}^k n_i \log(n_i/n)$ up to a constant.

- Suppose that we can define $\theta_j^{'} = g_j(\boldsymbol{\theta})$, $j = 1, \ldots, r$, where g_j is chosen so that H_0 becomes equivalent to $(\theta_1^{'}, \ldots, \theta_q^{'})^T$ ranges over an open subset of R^q and $\theta_j = \theta_{0j}$, $j = q+1, \ldots, r$ for specified θ_{0j}.
  For example, to test the Hardy-Weinberg model we set $\theta_1^{'} = \theta_1$, $\theta_2^{'} = \theta_2 - 2\sqrt{\theta_1}(1 - \sqrt{\theta_1})$ and test $H_1 : \theta_2^{'} = 0$.
- Apply the standard result on likelihood ratio test, under H_0, λ_n approximately has a χ^2_{r-q} distribution for large n.

Example 1 Consider Hardy-Weinberg model.

- $\hat{\eta} = (2n_1 + n_2)/2n$
- Reject H_0 if $\lambda_n \geq \chi^2_1(1-\alpha)$ with

$$\boldsymbol{\theta}(\hat{\eta}) = \left(\left(\frac{2n_1 + n_2}{2n} \right)^2, \frac{(2n_1+n_2)(2n_3+n_2)}{2n^2}, \left(\frac{2n_3 + n_2}{2n} \right)^2 \right)^T$$

For the Wald statistic and Rao score statistic, they are approximately χ^2_{r-q} distributed for large n under H_0.

- Wald statistic:

$$W_n = \sum_{j=1}^{k} \frac{[N_j - n\theta_j(\hat{\boldsymbol{\eta}})]^2}{n\theta_j(\hat{\boldsymbol{\eta}})}.$$

- Rao score statistic:

$$S_n = \sum_{j=1}^{k} \frac{[N_j - n\theta_j(\hat{\boldsymbol{\eta}})]^2}{\theta_j(\hat{\boldsymbol{\eta}})}.$$

- They are identical to Pearson's χ^2 statistic

$$SUM\frac{(Observed - Expected)^2}{Expected}.$$

Example 2 (The Fisher Linkage Model).

- A self-crossing of maize heterozygous on two characteristic (starchy versus sugary; green base leaf versus white base leaf) leads to four possible offspring types: (1) sugary-white; (2) sugary-green; (3) starchy-white; (4) starchy-green.
- $(N_1, \ldots, N_4)$ has a $MN(n, \theta_1, \ldots, \theta_4)$ distribution.
- Fisher (1958) specifies that

  $$\theta_1 = \frac{1}{4}(2+\eta), \theta_2 = \theta_3 = \frac{1}{4}(1-\eta), \theta_4 = \frac{1}{4}\eta$$

  where η is an unknown number between 0 and 1.

- To test the validity of the linkage model we would take

  $$\Theta_0 = \{(\frac{1}{4}(2+\eta), \frac{1}{4}(1-\eta), \frac{1}{4}(1-\eta), \frac{1}{4}) : 0 \leq \eta \leq 1\}$$

  a "one-dimensional curve" of the three-dimensional parameter space Θ.

- The likelihood equation under H_0 becomes

  $$\frac{n_1}{2+\eta} - \frac{n_2+n_3}{1-\eta} + \frac{n_4}{\eta} = 0.$$

- We obtain critical values from χ_1^2 table.

Testing Independence of Classification in Contingency Tables

- Many important characteristics have only two categories.
- An individual either is or is not inoculated against a disease; is or is not a soker; is male or female; and so on.
- We often want to know whether such characteristics are linked or are independent. For example, do smoking and lung cancer have any relation to each other?
- Let us call the possible categories or states of the first characteristic A and $\bar{A}$ and of the second B and $\bar{B}$.
  - A randomly selected individual from the population can be one of four types AB, $A\bar{B}$, $\bar{A}B$, $\bar{A}\bar{B}$.
  - Denote the probabilities of these types by θ_{11}, θ_{12}, θ_{21}, θ_{22}, respectively.
- Independent classification means that the events (being an A) and (being a B) are independent or in terms of the θ_{ij},

$$\theta_{ij} = (\theta_{i1} + \theta_{i2})(\theta_{1j} + \theta_{2j}).$$

- The data are assembled in what is called a 2×2 *contingency table.*

- Testing independence can be put as $H_0 : \boldsymbol{\theta} \in \Theta_0$ versus $H_1 : \boldsymbol{\theta} \notin \Theta_0$ where Θ_0 is a two-dimensional subset of Θ given by

  $$\Theta_0 = \{(\eta_1\eta_2, \eta_1(1-\eta_2), \eta_2(1-\eta_1), (1-\eta_1)(1-\eta_2)) : 0 \leq \eta_1, \eta$$

  The degree of freedom of chi^2 test is 1.

- For $\boldsymbol{\theta} \in \Theta_0$, $\hat{\eta}_1 = (n_{11} + n_{12})/n$ and $\hat{\eta}_2 = (n_{11} + n_{21})/n$.

- Pearson's statistic is

  $$n \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{[N_{ij} - (N_{i1} + N_{i2})(N_{1j} + N_{2j})/n]^2}{(N_{i1} + N_{i2})(N_{1j} + N_{2j})}.$$

- Pearson's statistic can be rewritten as Z^2 where

  $$Z = \left(\frac{N_{11}}{N_{11} + N_{21}} - \frac{N_{12}}{N_{12} + N_{22}}\right) \sqrt{\frac{(N_{11} + N_{21})(N_{12} + N_{22})n}{(N_{11} + N_{12})(N_{21} + N_{22})}}$$

  Thus,

  $$Z = \sqrt{n}[\hat{P}(A|B) - \hat{P}(A|\bar{B})] \left[\frac{\hat{P}(B)}{\hat{P}(A)} \frac{\hat{P}(\bar{B})}{\hat{P}(\bar{A})}\right]^{1/2}$$

  where $\hat{P}$ is the empirical distribution.

- Z indicates what directions they deviate from independence.

$a \times b$ Contingency Tables

- Consider contingency tables for two nonnumerical characteristics having a and b states, respectively, $a, b \geq 2$.
- If we take a sample of size n from a population and classify them according to each characteristic we obtain a vector N_{ij}, $i = 1, \ldots, a$, $j = 1, \ldots, b$ where N_{ij} is the number of individuals of type i for characteristic 1 and j for characteristic 2.
- $\{N_{ij} : 1 \leq i \leq a, 1 \leq j \leq b\}$ are multinomially distributed with $\{\theta_{ij} : 1 \leq i \leq a, 1 \leq j \leq b\}$ where

  $\theta_{ij} = P($A randomly selected individual is of type i for 1 ar
- The hypothesis that the characteristics are assigned independently becomes $H_0 : \theta_{ij} = \eta_{i1}\eta_{j2}$ for $1 \leq i \leq a$, $1 \leq j \leq b$ where the η_{i1}, η_{j2} are nonnegative and $\Sigma_{i=1}^{a} \eta_{i1} = \Sigma_{j=1}^{b} \eta_{j2} = 1$.
- N_{ij} can be arranged in a $a \times b$ contingency table. Write $C_j = \Sigma_{i=1}^{a} N_{ij}$ and $R_i = \Sigma_{j=1}^{b} N_{ij}$.

- Pearson's χ^2 for the hypothesis of independence is

$$n \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(N_{ij} - R_i C_j / n)^2}{R_i C_j},$$

which has approximately a $\chi^2_{(a-1)(b-1)}$ distribution under H_0.

Logistic Regression for Binary Response

- Consider Bernoulli responses Y that can only take on the values 0 and 1. Examples are
  - medical trials where at the end of the trial the patient has either recovered ($Y = 1$) or has not recovered ($Y = 0$),
  - election polls where a voter either supports a proposition ($Y = 1$) or does not ($Y = 0$),
  - market research where a potential customer either desires a new product ($Y = 1$) or does not ($Y = 0$)
  - multiple-choice test where an examiner either gets a correct answer ordoes not
- Assume the distribution of Y depends on the known covariate vector $\mathbf{z}$ in R^p.
- Assume that the data are grouped or replicated so that for each fixed i, we observe the number of successes $X_i = \Sigma_{j=1}^{m_i} Y_{ij}$ where Y_{ij} is the response on the jth of the m_i trials in block i, $1 \leq i \leq k$.
  Thus, we observe independent $X_1, \ldots, X_k$ with X_i binomial $Bin(m_i, \pi)$, where $\pi =$

$\pi(\mathbf{z})$ is the probability of success for a case with covariate vector $\mathbf{z}_i$.

- Consider the *logistic transform* $g(\pi)$, usually called the *logit*, which is

$$\eta = g(\pi) = \log[\pi/(1-\pi)].$$

- We choose the following parametric model for $\pi(\mathbf{z})$

$$logit(\pi(\mathbf{z})) = \mathbf{z}^T\boldsymbol{\beta}.$$

This model will allow that each component of $\mathbf{z}$ takes values on R.

- The above model is called the *logistic linear regression model.*
In practice, the *probit* $g_1(\pi) = \Phi^{-1}(\pi)$ where Φ is the $N(0,1)$ cdf and the *log-log transform* $g_2(\pi) = \log[-\log(1-\pi)]$are also being used.

- The log likelihood $\ell(\pi(\boldsymbol{\beta})) \equiv \ell_N(\boldsymbol{\beta})$ of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is, if $N = \Sigma_{i=1}^k m_i$,

$$\ell_N(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j T_j - \sum_{i=1}^{k} m_i \log(1+\exp(\mathbf{z}_i\boldsymbol{\beta})) + \sum_{i=1}^{k} \log \begin{pmatrix} m_i \\ X_i \end{pmatrix}$$

where $T_j = \Sigma_{i=1}^k z_{ij} X_i$.

- The likelihood equations are

$$\mathbf{Z}^T(\mathbf{X} - \boldsymbol{\mu}) = 0,$$

where $\mathbf{Z} = (z_{ij})_{m\times p}$ by observing

$$\begin{aligned} \mu_i &= E(X_i) = m_i\pi_i \\ E(T_j) &= \sum_{i=1}^{k} z_{ij}\mu_i \end{aligned}$$

- The MLE $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ solves $E_{\boldsymbol{\beta}}(T_j) = T_j$, $j = 1, \ldots, p$.
- To solve the above nonlinear equations, we use the Newton-Raphson algorithm
- The Fisher information matrix is $\mathbf{Z}^T\mathbf{W}\mathbf{Z}$ where $\mathbf{W} = diag\{m_i\pi_i(1-\pi_i)\}_{k\times k}$.

Testing

- Let $\omega = \{\boldsymbol{\eta} : \eta_i = \mathbf{z}_i^T\boldsymbol{\beta}, \boldsymbol{\beta} \in R^p\}$. Consider two different kinds of tests.
  - Let $\Omega = R^k$. Test $H_0 : \boldsymbol{\eta} \in \omega$ versus $H_1 : \boldsymbol{\eta} \in \Omega \setminus \omega$.
  - Let ω_0 be a q-dimensional linear subspace of ω with $q < r$.. Test $H_0 : \boldsymbol{\eta} \in \omega_0$ versus $H_1 : \boldsymbol{\eta} \in \omega \setminus \omega_0$.

- For the first set of hypotheses, the MLEs of π_i and μ_i are X_i/m_i and X_i. The log-likelihood ratio test statistics is

  $$2 \sum_{i=1}^{k} [X_i \log(X_i/\hat{\mu}_i) + X_i^{'} \log(X_i^{'}/\hat{\mu}_i^{'})]$$

  where $X_i^{'} = m_i - X_i$ and $\mu_i^{'} = m_i - \hat{\mu}_i$.

  - Note that it just measure the distance between the fit $\hat{\mu}$ based on the model ω and the data.
  - By the multivariate delta method, it has asymptotically a χ^2_{k-p} distribution for $\boldsymbol{\eta} \in \omega$ as $m_i \to \infty$, $i = 1, \ldots, k < \infty$.

- For the second set of hypotheses, the log-likelihood ratio test statistics is

  $$2 \sum_{i=1}^{k} \left[X_i \log\left(\frac{\hat{\mu}_i}{\hat{\mu}_{0i}}\right) + X_i^{'} \log\left(\frac{\hat{\mu}_i^{'}}{\hat{\mu}_{0i}^{'}}\right)\right]$$

  where $\hat{\mu}_0$ is the MLE of μ under H_0 and $\mu_{0i}^{'} = m_i - \hat{\mu}_{0i}$.
  It has an asymptotical χ^2_{p-q} distribution as $m_i \to \infty$, $i = 1, \ldots, k < \infty$.

Tests for A Simple Null Hypothesis

- Let $X_1, \ldots, X_n$ be iid with $X_1 \sim N(\theta, 1)$.
  - Test $H_0 : \theta = 0$ versus $H_1 : \theta = \theta_0 > 0$.
  - How do we find a good test for the above simple hypothesis?
- Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0 \in R^s$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}^0$.
- We consider three large sample tests.

Likelihood Ratio Test

- A *likelihood ratio* statistic,

$$\Lambda_n = \frac{L(\boldsymbol{\theta}^0; \mathbf{x})}{\sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \mathbf{x})}$$

  was introduced by Neyman and Pearson (1928).
- Λ_n takes values in the interval $[0, 1]$ and H_0 is to be rejected for sufficiently small values of Λ_n.
- The rationale behind LR tests is that when H_0 is true, Λ_n tends to be close to 1, whereas when H_1 is true, Λ_n tends to be close to 0,

- The test may be carried out in terms of the statistic

$$\lambda_n = -2 \log \Lambda_n.$$

- For finite n, the null distribution of λ_n will generally depend on n and on the form of pdf of X.

- LR tests are closely related to MLE's.

- Denote MLE by $\hat{\boldsymbol{\theta}}$. For asymptotic analysis, expanding λ_n at $\hat{\boldsymbol{\theta}}$ in a Taylor series, we get

$$\begin{aligned}\lambda_n &= -2\left\{-\sum_{i=1}^{n} \log f(X_i, \hat{\boldsymbol{\theta}}) + \sum_{i=1}^{n} \log f(X_i, \boldsymbol{\theta}^0)\right\} \\ &= 2\left\{\frac{1}{2}(\boldsymbol{\theta}^0 - \hat{\boldsymbol{\theta}})^T \left(-\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log f(x;\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right) (\boldsymbol{\theta}^0 -\right.\end{aligned}$$

where $\hat{\boldsymbol{\theta}}$ lies between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^0$.

- Since $\boldsymbol{\theta}^*$ is consistent,

$$\lambda_n = n(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)^T \left(-\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta_j \partial\theta_k} L(\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right) (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)+o_P(1).$$

  By the asymptotic normality of $\hat{\boldsymbol{\theta}}$ and

$$-n^{-1} \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta_j \partial\theta_k} L(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} \xrightarrow{P} \mathbf{I}(\boldsymbol{\theta}^0),$$

  λ_n has, under H_0, a limiting chi-squared distribution on s degrees of freedom.

Example 3 Consider the testing problem $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 \neq \sigma_0^2$ based on iid $X_1, \ldots, X_n$ from the normal distribution $N(\mu_0, \sigma^2)$.

- $L(\theta^0; \mathbf{x}) = (2\pi\sigma_0^2)^{-n/2} \exp\left[- \Sigma_i (x_i - \mu_0)^2 / 2\sigma_0^2\right]$
- $\hat{\sigma}^2 = n^{-1} \Sigma_i (x_i - \mu_0)^2$ (MLE) and
$$\sup_{\theta \in \Theta} L(\theta; \mathbf{x}) = (2\pi\hat{\sigma}^2)^{-n/2} \exp(-n/2).$$
- We have
$$\Lambda_n = \left(\frac{\hat{\sigma}^2}{\sigma_0^2}\right)^{n/2} \exp\left(\frac{n}{2} - \frac{\Sigma_i (x_i - \mu_0)^2}{2\sigma_0^2}\right)$$
or under H_0
$$\lambda_n = -n \left\{ \ln\left(\frac{1}{n} \sum_{i=1}^n Z_i^2\right) - \left[1 - \left(\frac{1}{n} \sum_{i=1}^n Z_i^2\right)\right]\right\},$$
where $Z_1, \ldots, Z_n$ are iid $N(0,1)$.
- Fact: Using CLT, we have
$$\frac{n^{-1} \Sigma_{i=1}^n Z_i^2 - 1}{\sqrt{2/n}} \xrightarrow{d} N(0,1)$$
or
$$\frac{n}{2} \left(\frac{1}{n} \sum_{i=1}^n Z_i^2 - 1\right)^2 \xrightarrow{d} \chi_1^2.$$

- Note that $\ln u \approx -(1-u) - (1-u)^2/2$ when u is near 1 and $n^{-1}\Sigma_{i=1}^n Z_i^2 \rightarrow 1$ in probability by LLN.
- A common question to be asked in Taylor's series approximation is that how many terms we should consider. In this example, it refers to the use of approximation $\ln u \approx -(1-u)$ as a contrast to the second order approximation we use. If we do use the first order approximation, we will end up the difficulty of finding $\lim_n a_n b_n$ when $\lim_n a_n = \infty$ and $\lim_n b_n = 0$.
- We conclude that λ_n has a limiting chi-squared distribution with 1 degree of freedom.

The Wald Test

- Let $\hat{\boldsymbol{\theta}}_n$ denote a consistent, asymptotically normal, and asymptotically efficient sequence of solutions of the likelihood equations.

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\theta}))$$

as $n \longrightarrow \infty$.

- Because $\mathbf{I}(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$, we have

$$\mathbf{I}(\hat{\boldsymbol{\theta}}_n) \xrightarrow{P} \mathbf{I}(\boldsymbol{\theta})$$

as $n \to \infty$.

- Replace the matrix $\left(-\frac{1}{n}\Sigma_{i=1}^{n}\frac{\partial^2}{\partial\theta_j\partial\theta_k}L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}\right)$ by $\mathbf{I}(\hat{\boldsymbol{\theta}}_n)$ in large sample approximation of λ_n, we get a second statistic,

$$W_n = n(\hat{\boldsymbol{\theta}}_n - \theta^0)^T\mathbf{I}(\hat{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0),$$

which was introduced by Wald (1943).

- By Slutsky's theorem, W_n converges in distribution to χ_s^2.

- For the construction of confidence region, one generates $\{\boldsymbol{\theta}^0 : W_n \le \chi^2_{s,\alpha}\}$ which is an ellipsoid in R^s.

- As a remark, for the construction of confidence region based on λ_n, one generates $\{\boldsymbol{\theta}^0 : \lambda_n \le \chi^2_{s,\alpha}\}$ which is not necessary an ellipsoid in R^s.

The Rao Score Tests

- Both the Wald and likelihood ratio tests requires evaluation of $\hat{\boldsymbol{\theta}}_n$. Now we consider a test for which this is not necessary.

- Denote the likelihood score vector

$$q(\mathbf{x};\boldsymbol{\theta}) = (q_1(\mathbf{x};\boldsymbol{\theta}), \ldots, q_s(\mathbf{x};\boldsymbol{\theta}))^T$$

where

$$q_j(\mathbf{x};\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_j} \log f(\mathbf{x};\boldsymbol{\theta}).$$

- Write $Q(\boldsymbol{\theta}) = \Sigma_{i=1}^{n} q(X_i;\boldsymbol{\theta})$. By the central limit theorem,

$$n^{-1/2} Q(\boldsymbol{\theta}^0) \xrightarrow{d} N(0, \mathbf{I}(\boldsymbol{\theta}^0)).$$

- A third statistic,

$$\begin{aligned} V_n &= [n^{-1/2} Q(\boldsymbol{\theta}^0)]^T \mathbf{I}^{-1}(\boldsymbol{\theta}^0)[n^{-1/2} Q(\boldsymbol{\theta}^0)] \\ &= n^{-1} Q(\boldsymbol{\theta}^0)^T \mathbf{I}^{-1}(\boldsymbol{\theta}^0) Q(\boldsymbol{\theta}^0), \end{aligned}$$

  was introduced by Rao (1948).
  Again, it has a limiting χ^2_s distribution.

Example 4. Consider a sample $X_1, \ldots, X_n$ from the logistic distribution with density

$$f_\theta(x) = \frac{e^{x-\theta}}{(1+e^{x-\theta})^2}.$$

- $q(x;\theta) = -1 + 2e^{x-\theta}/(1+e^{x-\theta})$ and

$$Q(\theta^0) = -n + 2 \sum_{i=1}^{n} \frac{e^{x_i-\theta^0}}{1+e^{x_i-\theta^0}}.$$

- $I(\theta) = 1/3$ for all θ.

- The Rao scores test therefore rejects H_0 with test statistic
$$\sqrt{\frac{3}{n}}\sum_{i=1}^{n}\frac{e^{x_i-\theta^0}-1}{1+e^{x_i-\theta^0}}.$$
- In this case, the MLE does not have an explicit expression and therefore the Wald and likelihood ratio tests are less convenient.

Example 5. Consider a sequence of n independent trials, with s possible outcomes for each trials.

- Let θ_j denote the probability of occurrence of the jth outcome in any given trial.
- Let N_j denote the number of occurrences of the jth outcome in the series of n trials.
- The MLE of θ_j's are N_j/n.
- The three test statistics λ_n, W_n and V_n for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0$ against $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}^0$ are easily seen to be
$$\begin{aligned}\lambda_n &= 2\sum_{j=1}^{s} N_j \log(\frac{N_j}{n\theta_j^0}),\\ W_n &= \sum_{j=1}^{s}\frac{(N_j - n\theta_j^0)^2}{N_j},\end{aligned}$$

$$V_n \;=\; \sum_{j=1}^{s} \frac{(N_j - n\theta_j^0)^2}{n\theta_j^0}.$$

- Both W_n and V_n are referred to as chi-squared goodness of fit statistics; the latter often called the Pearson chi-squared distribution. The large sample properties was first derived by Pearson (1900).
  Pearson's chi-square statistic is easily remembered as
  $$\chi^2 = sum\frac{(Observed - Expected)^2}{Expected}.$$

Example 6. (Testing a Genetic Theory)

- In experiments on pea breading, Mendel observed the different kinds of seeds obtained by crosses from peas with round yellow seeds and peas with wrinkled green seeds.
- Possible types of progeny were: (1) round yellow; (2) wrinkled yellow; (3) round green; and (4) wrinkled green.
- Assume the seeds are produced independently. We can think of each seed as being the outcome of a multinomial trial with possible

outcomes numbered 1, 2, 3, 4 as above and associated probabilities of occurrence θ_1, θ_2, θ_3, θ_4.

- Mendel's theory predicted that $\theta_1 = 9/16$, $\theta_2 = \theta_3 = 3/16$, $\theta_4 = 1/16$.
- Data: $n = 556$, $n_1 = 315$, $n_2 = 101$, $n_3 = 108$, $n_4 = 32$.
- Pearson's chi-square statistic is

$$\frac{(315 - 556 \times 9/16)^2}{312.75} + \frac{(3.25)^2}{104.25} + \frac{(3.75)^2}{104.25} + \frac{(2.75)^2}{34.75} = 0.47,$$

which has a p-