7 Likelihood Ratio and related tests

[CB8.2, CB10.3, BD4.9]

Definition 1 *The likelihood ratio test statistic for testing* $H : \theta \in \Theta_0$ *vs* $K : \theta \in \Theta_1$ *is*

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta \mid x)}{\sup_{\theta \in \Theta} L(\theta \mid x)}$$

where $\Theta = \Theta_0 \bigcup \Theta_1$.

A likelihood ratio test is of the form

$$\phi(x) = \begin{cases} 1 & \text{if} \quad \lambda(x) \le c \\ 0 & \text{if} \quad \lambda(x) > c \end{cases} \tag{1}$$

The value of c is determined from the level of the test such that $P_H(\lambda \le c) = \alpha$.
Example 1: Consider testing $H : \mu = \mu_0$ vs $K : \mu \ne \mu_0$ where $X_1, \cdots, X_n$ are iid $\mathcal{N}(\mu, 1)$.
For the numerator, $\sup_{\theta \in \Theta_0} L(\theta \mid x) = \frac{1}{(2\pi)^{n/2}} exp(-\frac{1}{2} \sum_{i=1}^n (x_i - \mu_0)^2)$
The sup in the denominator is attained at $\theta = \bar{x}$ which is mle.
Hence $\sup_{\theta \in \Theta} L(\theta \mid x) = \frac{1}{(2\pi)^{n/2}} exp(-\frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2)$

$$\lambda(x) = exp(-\frac{1}{2}(\sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x_i - \bar{x})^2))$$

$\lambda(x) < c \iff \sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x_i - \bar{x})^2 > c_1 \iff |\, \bar{x} - \mu_0 \,| > c_2$.
Thus the likelihood ration test rejects H when $\bar{x}$ differs from μ_0 by a large amount. The amount is determined from the constraint given by the level of the test, that is, $P_{\mu_0}(|\, \bar{x} - \mu_0 \,| > c_2) = \alpha$.

Exercise: Find the likelihood ration test for $H : \theta \le \theta_0$ vs $K : \theta > \theta_0$ when $X_1, \cdots, X_n$ are iid from the exponential distribution with pdf $f(x \mid \theta) = exp(-x + \theta) I(x > \theta)$.

7.1 Large Sample Distribution of LRT

Let $X_1, \cdots, X_n$ be iid with density $f(x, \theta)$. We are interested in testing $H : \theta = \theta_0$ against $K : \theta \ne \theta_0$, where θ is of dimension k, using a likelihood ratio test. To carry out the test, we need to determine the appropriate critical value c. Recall that c is determined by the requirement that $P_H(\lambda(x) < c) = \alpha$. In order to determine the critical value, we thus need to determine the distribution of $\lambda(X)$ when the null hypothesis is true. We now develop a large sample approximation to solve this problem.

Let $\hat{\theta} = \text{argmax}_\theta L(\theta)$ denote the mle, and write the maximized likelihood ratio statistic as

$$\lambda(x) = \frac{L(\theta_0)}{L(\hat{\theta})} \tag{2}$$

Define the statistic $\xi_{LR}(x) = -2\ln(\lambda(x)) = 2(l(\hat{\theta}) - l(\theta_0))$ where $l(\theta) = \ln L(\theta)$. Since ξ_{LR} is a monotonic decreasing transformation of λ, the LR test can be implemented by rejecting the null hypothesis when $\xi_{LR}(x)$ is large.

To find the approximate distribution of $\xi_{LR}(X)$ under the null hypothesis, write

$$l(\theta_0) = l(\hat{\theta}) + (\theta_0 - \hat{\theta})' \frac{\partial l(\hat{\theta})}{\partial \theta} + \frac{1}{2}(\theta_0 - \hat{\theta})' \frac{\partial^2 l(\tilde{\theta})}{\partial \theta \partial \theta}(\theta_0 - \hat{\theta}) \tag{3}$$

where $\tilde{\theta}(\omega)$ is between θ_0 and $\hat{\theta}(\omega)$. Since mle is the root of the likelihood equation, $\frac{\partial l(\hat{\theta})}{\partial \theta} = 0$. We have

$$\begin{aligned} \xi_{LR} &= -(\theta_0 - \hat{\theta})' \frac{\partial^2 l(\tilde{\theta})}{\partial \theta \partial \theta}(\theta_0 - \hat{\theta}) & (4) \\ &= \sqrt{n}(\theta_0 - \hat{\theta})' \left(-\frac{1}{n} \frac{\partial^2 l(\tilde{\theta})}{\partial \theta \partial \theta} \right) \sqrt{n}(\theta_0 - \hat{\theta}) & (5) \end{aligned}$$

Proceeding as in our derivations of the properties of the maximum likelihood estimator,

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta_0) &\Rightarrow \mathcal{N}(0, I(\theta_0)^{-1}) & (6) \\ -\frac{1}{n} \frac{\partial^2 l(\tilde{\theta})}{\partial \theta \partial \theta'} &\xrightarrow{P} I(\theta_0) & (7) \end{aligned}$$

so that by Slutsky and the Continuous Mapping Theorem,

$$\xi_{LR} \overset{H_0}{\Rightarrow} \chi^2_k \tag{8}$$

An asymptotically justified level $1 - \alpha$ confidence set based on the LR statistic is hence of the form

$$\theta^* \mid (\hat{\theta} - \theta^*)' \hat{V}^{-1} (\hat{\theta} - \theta^*) < c \tag{9}$$

where $\hat{V} = \left(-\frac{\partial^2 l(\tilde{\theta})}{\partial \theta \partial \theta'} \right)^{-1}$ and c solves $P(\chi^2_k > c) = \alpha$. This confidence set may be recognized as the interior of an ellipse centered at $\theta = \hat{\theta}$. In the one-dimensional case, we obtain a confidence interval $(\hat{\theta} - c^* \hat{V}^{-1/2}, \hat{\theta} + c^* \hat{V}^{-1/2})$ where c^* is the positive number that solves $P(\mathcal{N}(0,1) > c^*) = \alpha/2$.

7.2 Wald statistic

A close cousin of the LR statistic is the Wald statistic

$$\xi_W = \sqrt{n}(\hat{\theta} - \theta_0) \left(-\frac{1}{n} \frac{\partial^2 l(\hat{\theta})}{\partial \theta \partial \theta'} \right) \sqrt{n}(\hat{\theta} - \theta_0) \tag{10}$$

which differs from ξ_{LR} only because the estimated information matrix is evaluated at $\hat{\theta}$ rather than $\tilde{\theta}$. Note that we can compute the Wald statistic without doing any computations under the null hypothesis.

Since both $\hat{\theta}$ and $\tilde{\theta}$ converge in probability to θ_0 under the null hypothesis, $\xi_W - \xi_{\text{LR}} \overset{P,H_0}{\to} 0$

The motivation of the Wald statistic is that under the null hypothesis, the difference between the estimator $\hat{\theta}$ and θ_0 satisfies $\sqrt{n}(\hat{\theta}-\theta_0) \Rightarrow \mathcal{N}(0, I(\theta_0)^{-1})$and $-\frac{1}{n}\frac{\partial^2 l(\hat{\theta})}{\partial\theta\partial\theta'}$ consistently estimates $I(\theta_0)^{-1}$. Under the alternative, $\| \hat{\theta} - \theta_0 \|$ is large and we reject.

7.3 Lagrange Multiplier statistic

Another approximation to ξ_{LR} is given by the Lagrange Multiplier test statistic

$$\begin{aligned} \xi_{\text{LM}} &= \sqrt{n}S_n(\theta_0)\left(-\frac{1}{n}\frac{\partial^2 l(\theta_0)}{\partial\theta\partial\theta'}\right)^{-1}\sqrt{n}S_n(\theta_0) &(11)\\ &= \left(n^{-1/2}\sum_{i=1}^n s_i(\theta_0)\right)'\left(-\frac{1}{n}\frac{\partial^2 l(\theta_0)}{\partial\theta\partial\theta'}\right)^{-1}\left(n^{-1/2}\sum_{i=1}^n s_i(\theta_0)\right) &(12)\end{aligned}$$

with the advantage that we do not need to compute $\hat{\theta}$ in order to compute ξ_{LM}.

Since under the null hypothesis $n^{-1/2}\sum_{i=1}^n s_i(\theta_0) \Rightarrow \mathcal{N}(0, I(\theta_0))$ and $-\frac{1}{n}\frac{\partial^2 l(\theta_0)}{\partial\theta\partial\theta'} \overset{P}{\to} I(\theta_0)$ we also find $\xi_{\text{LM}} \overset{H_0}{\Rightarrow} \chi^2_k$.

7.4 Pearson's chi-square

[Lehman 5.5, Ferguson 9,10, Rao 6b]

Let $\underline{X}_1, \underline{X}_2, \cdots, \underline{X}_n$ be iid from a multinomial$_k(1, \underline{p})$ distribution, where $\underline{p}$ is a k-vector with nonnegative entries that sum to one. That is,

$$P(\underline{X}_i = e_j) = p_j \quad \text{for all} \quad 1 \le j \le k \tag{13}$$

where $\quad e_j =$ the k vector with 1 at the j-th position and 0's everywhere else.

Note that the multinomial distribution is a generalization of the binomial distribution to the case in which there are k categories of outcome instead of only 2. Also note that we ordinarily do not consider a binomial random variable to be a 2-vector, but we could easily do so if the vector contained both the number of successes and the number of failures. Equation (13) implies that the random vector $\underline{X}_i$ has expectation $\underline{p}$ and covariance matrix

$$\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -p_1p_k & -p_2p_k & \cdots & p_k(1-p_k) \end{pmatrix} \tag{14}$$

Using the Cramer-Wold device, the multivariate central limit theorem implies

$$\sqrt{n}(\bar{\underline{X}}_n - \underline{p}) \Rightarrow \mathcal{N}_k(\underline{0}, \Sigma). \tag{15}$$

Note that the sum of the j-th column of Σ is $p_j - p_j(p_1 + \cdots + p_k) = 0$, which is to say that the sum of the rows of Σ is the zero vector, so Σ is not invertible.

We wish to derive the asymptotic distribution of Pearson's chi-square statistic

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j}, \tag{16}$$

where n_j is the random variable that is the j-th component if $n\underline{\bar{X}}_n$, the number of successes in the j-th category for trials $1, \cdots, n$. We will discuss two different ways to do this. One way avoids dealing with the singular matrix Σ, whereas the other does not.

In the first approach, define for each i, $\underline{Y}_i = (\underline{X}_{i1}, \cdots, \underline{X}_{ik-1})$. That is, let $\underline{Y}_i$ be the $k-1$-vector consisting of the first $k-1$ components of $\underline{X}_i$. Then the covariance matrix of $\underline{Y}_i$ is the upper-left $(k-1) \times (k-1)$ submatrix of Σ, which we denote by Σ^*. Similarly, let $\underline{p}^*$ denote the vector $(p_1, \cdots, p_{k-1})$. First, verify that Σ^* is invertible and that

$$\Sigma^{*-1} = \begin{pmatrix} \frac{1}{p_1} + \frac{1}{p_k} & \frac{1}{p_k} & \cdots & \frac{1}{p_k} \\ \frac{1}{p_k} & \frac{1}{p_2} + \frac{1}{p_k} & \cdots & \frac{1}{p_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{p_k} & \frac{1}{p_k} & \cdots & \frac{1}{p_{k-1}} + \frac{1}{p_k} \end{pmatrix} \tag{17}$$

Second, verify that

$$\chi^2 = n(\underline{\bar{Y}}_n - \underline{p}^*)^t (\Sigma^*)^{-1} (\underline{\bar{Y}}_n - \underline{p}^*) \tag{18}$$

The facts in equations (17) and (18) are checked in exercise 1. If we now define

$$\underline{Z}_n = \sqrt{n}(\Sigma^*)^{-1/2}(\underline{\bar{Y}}_n - \underline{p}^*), \tag{19}$$

then clearly the central limit theorem implies $\underline{Z}_n \Rightarrow \mathcal{N}_{k-1}(\underline{0}, I)$. By definition, the χ^2_{k-1} distribution is the distribution of the sum of the squares of $k-1$ independent standard normal random variables. Therefore,

$$\chi^2 = (\underline{Z}_n)^t \underline{Z}_n \Rightarrow \chi^2_{k-1}, \tag{20}$$

which is the result that leads to the familiar chi-square test.

In a second approach to deriving the limiting distribution (20), we use some properties of projection matrices.

Definition 2 *A matrix P is called a projection matrix if it is idempotent; that is, if $P^2 = P$.*

The following lemmas, to be proven in exercise 2, give some basic facts about projection matrices.

Lemma 1 *Suppose P is a projection matrix. Then every eigenvalue of P equals 0 or 1. Suppose that r denotes the number of eigenvalues of P equal to 1. Then if $Z \sim \mathcal{N}_k(\underline{0}, P)$, then, $Z^t Z \sim \chi^2_r$.*

This can be derived from the Fisher-Cochran Theorem.

Lemma 2 *The trace of a square matrix equals the sum of its eigenvalues. For matrices A and B whose sizes allow them to be multiplied in either order,* $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$*.*

Define $\Gamma = \mathrm{diag}(\underline{p})$. Clearly, equation (15) implies

$$\sqrt{n}\Gamma^{-1/2}(\underline{\bar{X}}_n - \underline{p}) \Rightarrow \mathcal{N}_k(\underline{0}, \Gamma^{-1/2}\Sigma\Gamma^{-1/2}). \tag{21}$$

Since Σ may be written in the form $\Gamma - \underline{pp}^t$,

$$\Gamma^{-1/2}\Sigma\Gamma^{-1/2} = I - \Gamma^{-1/2}\underline{pp}^t\Gamma^{-1/2} = I - \sqrt{\underline{p}}\sqrt{\underline{p}}^t \tag{22}$$

clearly has trace $k-1$; furthermore, $(I - \sqrt{\underline{p}}\sqrt{\underline{p}}^t)(I - \sqrt{\underline{p}}\sqrt{\underline{p}}^t) = I - 2\sqrt{\underline{p}}\sqrt{\underline{p}}^t + \sqrt{\underline{p}}\sqrt{\underline{p}}^t\sqrt{\underline{p}}\sqrt{\underline{p}}^t = I - \sqrt{\underline{p}}\sqrt{\underline{p}}^t$ because $\sqrt{\underline{p}}^t\sqrt{\underline{p}} = 1$, so the covariance matrix (22) is a projection matrix.

Define $\Delta_n = \sqrt{n}\Gamma^{-1/2}(\underline{\bar{X}} - \underline{p})$. Then we may check (exercise 2) that

$$\chi^2 = (\Delta_n)^t\Delta_n \tag{23}$$

Therefore, since the covariance matrix (22) is a projection with trace $k-1$, Lemma 1 and Lemma 2 prove that $\chi^2 \Rightarrow \chi^2_{k-1}$ as desired.