ST5224: Advanced Statistical Theory II

Chen Zehua

Department of Statistics & Applied Probability

Thursday, March 29, 2012
An LR test is often equivalent to a test based on a statistic $Y(X)$ whose distribution under $H_0$ can be used to determine the rejection region of the LR test with size $\alpha$.

When this technique fails, it is difficult or even impossible to find an LR test with size $\alpha$, even if the c.d.f. of $\lambda(X)$ is continuous.

In the i.i.d. case we can obtain the asymptotic distribution (under $H_0$) of the likelihood ratio $\lambda(X)$ so that an LR test having asymptotic significance level $\alpha$ can be obtained.

In many problems $\Theta_0$ is determined by $H_0 : \theta = g(\vartheta)$, where $\vartheta$ is a $(k - r)$-vector of unknown parameters and $g$ is a continuously differentiable function from $\mathbb{R}^{k-r}$ to $\mathbb{R}^k$ with a full rank $\partial g(\vartheta)/\partial \vartheta$. For example, if $\Theta = \mathbb{R}^2$ and $\Theta_0 = \{ (\theta_1, \theta_2) \in \Theta : \theta_1 = 0 \}$, then $\vartheta = \theta_2$, $g_1(\vartheta) = 0$, and $g_2(\vartheta) = \vartheta$. 
Theorem 6.5 (Asymptotic distribution of likelihood ratio)

Assume the conditions in Theorem 4.16.
Suppose that \( H_0 : \theta = g(\vartheta) \), where \( \vartheta \) is a \((k - r)\)-vector of unknown parameters and \( g \) is a continuously differentiable function from \( \mathbb{R}^{k-r} \) to \( \mathbb{R}^{k} \) with a full rank \( \partial g(\vartheta) / \partial \vartheta \).
Under \( H_0 \), \(-2 \log \lambda_n \rightarrow_d \chi^2_r\), where \( \lambda_n = \lambda(X) \) and \( \chi^2_r \) is a random variable having the chi-square distribution \( \chi^2_r \).
Consequently, the LR test with rejection region \( \lambda_n < e^{-\chi^2_{r,\alpha}/2} \) has asymptotic significance level \( \alpha \), where \( \chi^2_{r,\alpha} \) is the \((1 - \alpha)\)th quantile of the chi-square distribution \( \chi^2_r \).

Proof
Without loss of generality, we assume that there exist an MLE \( \hat{\theta} \) and an MLE \( \hat{\vartheta} \) under \( H_0 \) such that
\[
\lambda_n = \sup_{\theta \in \Theta_0} \frac{\ell(\theta)}{\sup_{\theta \in \Theta} \ell(\theta)} = \frac{\ell(g(\hat{\vartheta}))}{\ell(\hat{\theta})}.
\]
Let \( s_n(\theta) = \partial \log \ell(\theta) / \partial \theta \) and \( I_1(\theta) \) be the Fisher information about \( \theta \) contained in \( X_1 \).
Following the proof of Theorem 4.17 in §4.5.2, we can obtain that
\[ \sqrt{n}l_1(\theta)(\hat{\theta} - \theta) = n^{-1/2}s_n(\theta) + o_p(1), \]
and
\[ 2[\log \ell(\hat{\theta}) - \log \ell(\theta)] = n(\hat{\theta} - \theta)^\top l_1(\theta)(\hat{\theta} - \theta) + o_p(1). \]

Then
\[ 2[\log \ell(\hat{\theta}) - \log \ell(\theta)] = n^{-1}[s_n(\theta)]^\top [l_1(\theta)]^{-1}s_n(\theta) + o_p(1). \]

Similarly, under \( H_0 \),
\[ 2[\log \ell(g(\hat{\varphi})) - \log \ell(g(\varphi))] = n^{-1}[\tilde{s}_n(\varphi)]^\top [\tilde{l}_1(\varphi)]^{-1}\tilde{s}_n(\varphi) + o_p(1), \]
where \( \tilde{s}_n(\varphi) = \partial \log \ell(g(\varphi))/\partial \varphi = D(\varphi)s_n(g(\varphi)), \)
\( D(\varphi) = \partial g^\top(\varphi)/\partial \varphi, \) and \( \tilde{l}_1(\varphi) \) is the Fisher information about \( \varphi \) (under \( H_0 \)) contained in \( X_1 \). Thus, we obtain that, under \( H_0 \),
\[ -2 \log \lambda_n = 2[\log \ell(\hat{\theta}) - \log \ell(g(\hat{\varphi}))] = n^{-1}[s_n(g(\varphi))]^\top B(\varphi)s_n(g(\varphi)) + o_p(1) \]
where \( B(\varphi) = [l_1(g(\varphi))]^{-1} - [D(\varphi)]^\top [\tilde{l}_1(\varphi)]^{-1}D(\varphi). \)
By the CLT, \( n^{-1/2}[l_1(\theta)]^{-1/2}s_n(\theta) \rightarrow_d Z \), where \( Z = N_k(0, I_k) \). Then, it follows from Theorem 1.10(iii) that, under \( H_0 \),

\[
-2 \log \lambda_n \rightarrow_d Z^\tau[l_1(g(\vartheta))]^{1/2}B(\vartheta)[l_1(g(\vartheta))]^{1/2}Z.
\]

Let \( D = D(\vartheta), B = B(\vartheta), A = l_1(g(\vartheta)), \) and \( C = \tilde{l}_1(\vartheta) \). Then

\[
(A^{1/2}BA^{1/2})^2 = A^{1/2}BABA^{1/2}
\]

\[
= A^{1/2}(A^{-1} - D^\tau C^{-1}D)A(A^{-1} - D^\tau C^{-1}D)A^{1/2}
\]

\[
= (I_k - A^{1/2}D^\tau C^{-1}DA^{1/2})(I_k - A^{1/2}D^\tau C^{-1}DA^{1/2})
\]

\[
= I_k - 2A^{1/2}D^\tau C^{-1}DA^{1/2} + A^{1/2}D^\tau C^{-1}DAD^\tau C^{-1}DA^{1/2}
\]

\[
= I_k - A^{1/2}D^\tau C^{-1}DA^{1/2}
\]

\[
= A^{1/2}BA^{1/2},
\]

where the fourth equality follows from the fact that \( C = DAD^\tau \). This shows that \( A^{1/2}BA^{1/2} \) is a projection matrix.
The rank of $A^{1/2}BA^{1/2}$ is
\[
\text{tr}(A^{1/2}BA^{1/2}) = \text{tr}(I_k - D^\tau C^{-1}DA) \\
= k - \text{tr}(C^{-1}DAD^\tau) \\
= k - \text{tr}(C^{-1}C) \\
= k - (k - r) = r.
\]

Thus, by Exercise 51 in §1.6,
\[
Z^\tau[l_1(g(\vartheta))]^{1/2}B(\vartheta)[l_1(g(\vartheta))]^{1/2}Z = \chi_r^2.
\]

**Asymptotic tests**

Tests whose rejection regions are constructed using asymptotic theory (so that these tests have asymptotic significance level $\alpha$) are called *asymptotic tests*, which are useful when a test of exact size $\alpha$ is difficult to find.

The LR test in Theorem 6.5 is one example of an asymptotic test.
Wald test and Score test

The hypothesis $H_0 : \theta = g(\vartheta)$ is equivalent to a set of $r \leq k$ equations:

$$H_0 : R(\theta) = 0,$$

where $R(\theta)$ is a continuously differentiable function from $\mathbb{R}^k$ to $\mathbb{R}^r$.

Wald (1943) introduced a test that rejects $H_0$ when the value of

$$W_n = [R(\hat{\theta})]^\tau \{[C(\hat{\theta})]^\tau [I_n(\hat{\theta})]^{-1} C(\hat{\theta})\}^{-1} R(\hat{\theta})$$

is large, where $C(\theta) = \partial R(\theta)/\partial \theta$, $I_n(\theta)$ is the Fisher information matrix based on $X_1, ..., X_n$, and $\hat{\theta}$ is an MLE or RLE of $\theta$.

For testing $H_0 : \theta = \theta_0$ with a known $\theta_0$, $R(\theta) = \theta - \theta_0$ and

$$W_n = (\hat{\theta} - \theta_0)^\tau I_n(\hat{\theta})(\hat{\theta} - \theta_0).$$

Rao (1947) introduced a score test that rejects $H_0$ when the value of

$$R_n = [s_n(\tilde{\theta})]^\tau [I_n(\tilde{\theta})]^{-1} s_n(\tilde{\theta})$$

is large, where $s_n(\theta) = \partial \log \ell(\theta)/\partial \theta$ is the score function and $\tilde{\theta}$ is an MLE or RLE of $\theta$ under $H_0 : R(\theta) = 0$. 
Theorem 6.6
Assume the conditions in Theorem 4.16.
(i) Under \( H_0 : R(\theta) = 0 \), where \( R(\theta) \) is a continuously differentiable function from \( \mathcal{R}^k \) to \( \mathcal{R}^r \), \( W_n \xrightarrow{d} \chi^2_r \) and, therefore, the test rejects \( H_0 \) if and only if \( W_n > \chi^2_{r,\alpha} \) has asymptotic significance level \( \alpha \), where \( \chi^2_{r,\alpha} \) is the \((1 - \alpha)\)th quantile of the chi-square distribution \( \chi^2_r \).
(ii) The result in (i) still holds if \( W_n \) is replaced by \( R_n \).

Remarks

▶ Wald’s test, Rao’s score test, and the LR test are asymptotically equivalent.
▶ Wald’s test requires computing \( \hat{\theta} \), not \( \tilde{\theta} = g(\hat{\theta}) \).
▶ Rao’s score test requires computing \( \tilde{\theta} \), not \( \hat{\theta} \).
▶ The LR test requires computing both \( \hat{\theta} \) and \( \tilde{\theta} \) (or solving two maximization problems), but it may be more efficient.
▶ Hence, one may choose one of these tests in terms of computation and efficiency in a particular application.
Proof

(i) Using Theorems 1.12 and 4.17,

\[ \sqrt{n}[R(\hat{\theta}) - R(\theta)] \rightarrow_d N_r \left(0, [C(\theta)]^\tau [I_1(\theta)]^{-1} C(\theta)\right) , \]

where \( I_1(\theta) \) is the Fisher information about \( \theta \) contained in \( X_1 \).

Under \( H_0 \), \( R(\theta) = 0 \) and, therefore (by Theorem 1.10),

\[ n[R(\hat{\theta})]^\tau \{[C(\theta)]^\tau [I_1(\theta)]^{-1} C(\theta)\}^{-1} R(\hat{\theta}) \rightarrow_d \chi^2_r \]

Then the result follows from Slutsky’s theorem (Theorem 1.11) and the fact that \( \hat{\theta} \rightarrow_p \theta \) and \( I_1(\theta) \) and \( C(\theta) \) are continuous at \( \theta \).

(ii) From the Lagrange multiplier, \( \tilde{\theta} \) satisfies

\[ s_n(\tilde{\theta}) + C(\tilde{\theta})\lambda_n = 0 \quad \text{and} \quad R(\tilde{\theta}) = 0. \]

Using Taylor’s expansion, one can show (exercise) that under \( H_0 \),

\[ [C(\tilde{\theta})]^\tau (\tilde{\theta} - \theta) = o_p(n^{-1/2}) \quad (1) \]

and

\[ s_n(\theta) - I_n(\theta)(\tilde{\theta} - \theta) + C(\tilde{\theta})\lambda_n = O_p(1), \quad (2) \]

where \( I_n(\theta) = nI_1(\theta) \).
Multiplying \([C(\tilde{\theta})]^{\tau}[I_n(\theta)]^{-1}\) to the left-hand side of (2) and using (1), we obtain that

\[
[C(\tilde{\theta})]^{\tau}[I_n(\theta)]^{-1} C(\tilde{\theta}) \lambda_n = -[C(\tilde{\theta})]^{\tau}[I_n(\theta)]^{-1} s_n(\theta) + o_p(n^{-1/2}),
\]

which implies

\[
\lambda_n^{\tau}[C(\tilde{\theta})]^{\tau}[I_n(\theta)]^{-1} C(\tilde{\theta}) \lambda_n \rightarrow_d \chi^2_r
\]

(exercise).

Then the result follows from (3) and the fact that

\[C(\tilde{\theta}) \lambda_n = -s_n(\tilde{\theta}), \quad I_n(\theta) = nI_1(\theta),\]

and \(I_1(\theta)\) is continuous at \(\theta\).
Testing in multinomial distributions

Consider \( n \) independent trials with \( k \) possible outcomes for each trial. Let \( p_j > 0 \) be the probability that the \( j \)th outcome occurs in a given trial and \( X_j \) be the number of occurrences of the \( j \)th outcome in \( n \) trials. Then \( X = (X_1, ..., X_k) \) has the multinomial distribution (Example 2.7) with the parameter \( p = (p_1, ..., p_k) \).

Let \( \xi_i = (0, ..., 0, 1, 0, ..., 0) \), where the single nonzero component 1 is located in the \( j \)th position if the \( i \)th trial yields the \( j \)th outcome. Then \( \xi_1, ..., \xi_n \) are i.i.d. and \( X/n = \bar{\xi} = \sum_{i=1}^{n} \xi_i/n \).

\( X/n \) is an unbiased estimator of \( p \) and, by the CLT,

\[
Z_n(p) = \sqrt{n} \left( \frac{X}{n} - p \right) = \sqrt{n}(\bar{\xi} - p) \rightarrow_d N_k(0, \Sigma),
\]

where \( \Sigma = \text{Var}(X/\sqrt{n}) \) is a symmetric \( k \times k \) matrix whose \( i \)th diagonal element is \( p_i(1 - p_i) \) and \((i, j)\)th off-diagonal element is \(-p_ip_j\).

We first consider the problem of testing

\[
H_0 : p = p_0 \quad \text{versus} \quad H_1 : p \neq p_0,
\]

where \( p_0 = (p_{01}, ..., p_{0k}) \) is a known vector of cell probabilities.
\( \chi^2 \) tests

For testing \( H : \mathbf{p} = \mathbf{p}_0 \) vs \( H_1 : \mathbf{p} \neq \mathbf{p}_0 \), a class of tests related to the asymptotic tests described in §6.4.2 is the class of \( \chi^2 \)-tests. A popular test is based on the following \( \chi^2 \)-statistic:

\[
\chi^2 = \sum_{j=1}^{k} \frac{(X_j - np_{0j})^2}{np_{0j}} = \| D(\mathbf{p}_0) Z_n(\mathbf{p}_0) \|^2,
\]

where \( D(c) \) with \( c = (c_1, \ldots, c_k) \) is the \( k \times k \) diagonal matrix whose \( j \)th diagonal element is \( c_j^{-1/2} \).

Another one is based on the following modified \( \chi^2 \)-statistic:

\[
\tilde{\chi}^2 = \sum_{j=1}^{k} \frac{(X_j - np_{0j})^2}{X_j} = \| D(X/n) Z_n(\mathbf{p}_0) \|^2.
\]

The next result shows that a test of asymptotic significance level \( \alpha \) rejects \( H_0 : \mathbf{p} = \mathbf{p}_0 \) when \( \chi^2 > \chi_{k-1,\alpha}^2 \) (or \( \tilde{\chi}^2 > \chi_{k-1,\alpha}^2 \)), where \( \chi_{k-1,\alpha}^2 \) is the \( (1 - \alpha) \)th quantile of \( \chi_{k-1}^2 \).

Thus, these tests are called (asymptotic) \( \chi^2 \)-tests.
**Theorem 6.8**

Let \( \phi = (\sqrt{p_1}, \ldots, \sqrt{p_k}) \) and \( \Lambda \) be a \( k \times k \) projection matrix.  

(i) If \( \Lambda \phi = a \phi \), then  

\[
[Z_n(p)]^T D(p) \Lambda D(p) Z_n(p) \to_d \chi_r^2,
\]

where \( \chi_r^2 \) has the chi-square distribution \( \chi_r^2 \) with \( r = \text{tr}(\Lambda) - a \).

(ii) The same result holds if \( D(p) \) in (i) is replaced by \( D(X/n) \).

**Remark**

The \( \chi^2 \)-statistic and the modified \( \chi^2 \)-statistic are special cases of the statistics in Theorem 6.8(i) and (ii), respectively, with \( \Lambda = I_k \) satisfying \( \Lambda \phi = \phi \).

**Proof**

The result in (ii) follows from the result in (i) and \( X/n \to_p p \).

To prove (i), let \( D = D(p) \), \( Z_n = Z_n(p) \), and \( Z = N_k(0, I_k) \). From the asymptotic normality of \( Z_n \) and Theorem 1.10,

\[
Z_n^T D \Lambda D Z_n \to_d Z^T A Z \text{ \quad with \quad } A = \Sigma^{1/2} D \Lambda D \Sigma^{1/2}.
\]
From Exercise 51 in §1.6, the result in (i) follows if we can show that $A^2 = A$ (i.e., $A$ is a projection matrix) and $\text{tr}(A) = \text{tr}(\Lambda) - a$. Since $\Lambda$ is a projection matrix and $\Lambda \phi = a \phi$, $a$ must be either 0 or 1. Note that $D \Sigma D = I_k - \phi \phi^T$.

Then

$$A^3 = \Sigma^{1/2} D \Lambda D \Sigma D \Lambda D \Sigma D \Lambda D \Sigma^{1/2}$$

$$= \Sigma^{1/2} D (\Lambda - a \phi \phi^T)(\Lambda - a \phi \phi^T) \Lambda D \Sigma^{1/2}$$

$$= \Sigma^{1/2} D (\Lambda - 2a \phi \phi^T + a^2 \phi \phi^T) \Lambda D \Sigma^{1/2}$$

$$= \Sigma^{1/2} D (\Lambda - a \phi \phi^T) \Lambda D \Sigma^{1/2}$$

$$= \Sigma^{1/2} D \Lambda D \Sigma D \Lambda D \Sigma^{1/2} = A^2,$$

which implies that the eigenvalues of $A$ must be 0 or 1. Therefore, $A^2 = A$. Also,

$$\text{tr}(A) = \text{tr}[\Lambda(D \Sigma D)] = \text{tr}(\Lambda - a \phi \phi^T) = \text{tr}(\Lambda) - a.$$
Goodness of fit tests

Let $Y_1, ..., Y_n$ be i.i.d. from $F$. Consider the problem of testing

$$H_0 : F = F_0 \quad \text{versus} \quad H_1 : F \neq F_0,$$

where $F_0$ is a known c.d.f. (For instance, $F_0 = N(0, 1)$.)

One way to test $H_0 : F = F_0$ is to partition the range of $Y_1$ into $k$ disjoint events $A_1, ..., A_k$ and test $H_0 : p = p_0$ with $p_j = P_F(A_j)$ and $p_{0j} = P_{F_0}(A_j)$, $j = 1, ..., k$.

Let $X_j$ be the number of $Y_i$'s in $A_j$, $j = 1, ..., k$.

Based on $X_j$'s, the $\chi^2$-tests discussed previously can be applied. They are called goodness of fit tests.

In the goodness of fit tests above, $F_0$ in $H_0$ is known so that $p_{0j}$'s can be computed.

In some cases, we need to test the following hypotheses:

$$H_0 : F = F_\theta \quad \text{versus} \quad H_1 : F \neq F_\theta,$$

where $\theta$ is an unknown parameter in $\Theta \subset \mathbb{R}^s$.

For example, $F_\theta = N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$. 
If we still try to test $H_0 : \mathbf{p} = \mathbf{p}_0$ with $p_j = P_{F_\theta}(A_j), \ j = 1, \ldots, k$, the result discussed above is not applicable since $\mathbf{p}$ is unknown under $H_0$. A generalized $\chi^2$-test can be obtained using the following result. Let $\mathbf{p}(\theta) = (p_1(\theta), \ldots, p_k(\theta))$ be a $k$-vector of known functions of $\theta \in \Theta \subset \mathbb{R}^s$, where $s < k$. Consider the testing problem

$$H_0 : \mathbf{p} = \mathbf{p}(\theta) \quad \text{versus} \quad H_1 : \mathbf{p} \neq \mathbf{p}(\theta).$$

Note that $H_0 : \mathbf{p} = \mathbf{p}_0$ is the special case of $H_0 : \mathbf{p} = \mathbf{p}(\theta)$ with $s = 0$. Let $\hat{\theta}$ be an MLE of $\theta$ under $H_0$.

By Theorem 6.5, the LR test that rejects $H_0$ when $-2 \log \lambda_n > \chi^2_{k-s-1,\alpha}$ has asymptotic significance level $\alpha$, where $\chi^2_{k-s-1,\alpha}$ is the $(1-\alpha)$th quantile of $\chi^2_{k-s-1}$ and

$$\lambda_n = \prod_{j=1}^{k} [p_j(\hat{\theta})]^{X_j} / (X_j/n)^{X_j}.$$

Using the fact that $p_j(\hat{\theta})/(X_j/n) \to_p 1$ under $H_0$ and

$$\log(1 + x) = x - x^2/2 + o(|x|^2) \quad \text{as} \ |x| \to 0,$$
we obtain that
\[-2 \log \lambda_n = -2 \sum_{j=1}^{k} X_j \log \left(1 + \frac{p_j(\hat{\theta})}{X_j/n} - 1\right)\]

\[= -2 \sum_{j=1}^{k} X_j \left(\frac{p_j(\hat{\theta})}{X_j/n} - 1\right) + \sum_{j=1}^{k} X_j \left(\frac{p_j(\hat{\theta})}{X_j/n} - 1\right)^2 + o_p(1)\]

\[= \sum_{j=1}^{k} \frac{[X_j - np_j(\hat{\theta})]^2}{X_j} + o_p(1) = \sum_{j=1}^{k} \frac{[X_j - np_j(\hat{\theta})]^2}{np_j(\hat{\theta})} + o_p(1),\]

where the third equality follows from
\[\sum_{j=1}^{k} p_j(\hat{\theta}) = \sum_{j=1}^{k} X_j/n = 1.\]

**Generalized \(\chi^2\)-statistics**

The generalized \(\chi^2\)-statistics \(\chi^2\) and \(\tilde{\chi}^2\) are defined to be the previously defined \(\chi^2\)-statistics with \(p_{0j}'s\) replaced by \(p_j(\hat{\theta})'s\).
Theorem 6.9
Under $H_0 : p = p(\theta)$, the generalized $\chi^2$-statistics converge in distribution to $\chi^2_{k-s-1}$.
The $\chi^2$-test with rejection region $\chi^2 > \chi^2_{k-s-1,\alpha}$ (or $\tilde{\chi}^2 > \chi^2_{k-s-1,\alpha}$) has asymptotic significance level $\alpha$, where $\chi^2_{k-s-1,\alpha}$ is the $(1 - \alpha)$th quantile of $\chi^2_{k-s-1}$.

Discussion
Theorem 6.9 can be applied to derive a goodness of fit test for $H_0 : p = p(\theta)$ vs $H_1 : p \neq p(\theta)$.
However, one has to compute an MLE of $\theta$ under $H_0 : p = p(\theta)$, which is different from an MLE under $H_0 : F = F_\theta$ unless $F = F_\theta$ and $p = p(\theta)$ are the same; see Moore and Spruill (1975).
Many elementary textbooks, however, use an MLE under $H_0 : F = F_\theta$, which is wrong.
MLE under $p = p(\theta)$

From the multinomial distribution, the MLE $\hat{\theta}$ in the generalized $\chi^2$ test should maximize the likelihood

$$\ell(\theta) = \frac{n!}{x_1! \cdots x_k!} [p_1(\theta)]^{x_1} \cdots [p_k(\theta)]^{x_k} l_{x_1+\cdots+x_k=1}$$

This MLE $\hat{\theta}$ is different from the MLE maximizing the likelihood based on the family $\{F_\theta\}$

For testing $H_0: F = N(\mu, \sigma^2)$, for example,

$$p_j(\theta) = \Phi \left( \frac{a_j + 1 - \mu}{\sigma} \right) - \Phi \left( \frac{a_j - \mu}{\sigma} \right), \quad j = 1, \ldots, k$$

where $-\infty = a_1 < a_2 < \cdots < a_k < a_{k+1} = \infty$ and $a_j$'s are fixed constants.

This MLE $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$ is certainly different from $\hat{\mu} =$ the sample mean and $\hat{\sigma}^2 = (n - 1)/n$ times the sample variance, which is the MLE under the normal model $N(\mu, \sigma^2)$. 