Asymptotic Distribution of the Likelihood Ratio Test Statistic

Let $X_1, \ldots, X_n$ be a sample from density $f(x|\theta)$ where $\Theta \subset \Theta_0 \subset \mathbb{R}^k$. The likelihood ratio test provides a general method for testing $H_0$: $\theta \in \Theta_0$ versus $H_1$: $\theta \in \Theta - \Theta_0$ for a given subset $\Theta_0$ of $\Theta$. This test rejects $H_0$ when the likelihood ratio test statistic,

$$\lambda_n = \sup_{\theta \in \Theta_0} \frac{\prod f(x_i|\theta)}{\prod f(x_i|\hat{\theta}_n)} = \frac{L_n(\theta^*_n)}{L_n(\hat{\theta}_n)}$$

is too small, where $\theta^*_n$ is the MLE over $\Theta_0$, and $\hat{\theta}_n$ is the MLE over $\Theta$. When the sample size is large, evaluation of a cutoff point can be facilitated in many important situations by the following theorem. These situations occur when $\Theta_0$ is a $(k-r)$-dimensional subspace of $\Theta$. Writing the components of the vector $\theta \in \mathbb{R}^k$ as $\theta^T = (\theta^1, \theta^2, \ldots, \theta^k)$, we assume the null hypothesis is of the form

$$H_0: \theta^1 = \theta^2 = \cdots = \theta^r = 0$$

where $1 \leq r \leq k$. More general situations, in which $H_0$ is of the form $H_0$: $g_1(\theta) = \cdots = g_s(\theta) = 0$ for some smooth real-valued functions $g_1, \ldots, g_s$, can be put into this form by a reparametrization. The integer $r$ represents the number of restrictions under the null hypothesis.

**Theorem 22** [Wilks (1938)]. Suppose the assumptions of Theorem 18 are satisfied and that $H_0$: $\theta^1 = \theta^2 = \cdots = \theta^r = 0$ where $1 \leq r \leq k$. Suppose that the true value $\theta_0$ satisfies $H_0$. Then

$$-2 \log \lambda_n \overset{d}{\to} \chi^2_r.$$

**Proof.** $-2 \log \lambda_n = 2[L_n(\hat{\theta}_n) - l_n(\theta^*_n)]$ where $\hat{\theta}_n$ is MLE over $\Theta$, and $\theta^*_n$ is MLE over $\Theta_0$. Expand $l_n(\theta^*_n)$ about $\hat{\theta}_n$:

$$l_n(\theta^*_n) = l_n(\hat{\theta}_n) + l_n'(\hat{\theta}_n)(\theta^*_n - \hat{\theta}_n) - n(\theta^*_n - \hat{\theta}_n)^T I_n(\theta^*_n)(\theta^*_n - \hat{\theta}_n),$$

where

$$I_n(\theta^*_n) = -\frac{1}{n} \int_0^1 \int_0^1 I_n(\hat{\theta}_n + uv(\theta^*_n - \hat{\theta}_n)) \, du \, dv \overset{a.s.}{\to} \frac{1}{2} \mathcal{F}(\theta_0),$$

as in the proof of Theorem 18. For sufficiently large $n$, $l_n'(\hat{\theta}_n) = 0$, so

$$-2 \log \lambda_n = 2n(\theta^*_n - \hat{\theta}_n)^T I_n(\theta^*_n)(\theta^*_n - \hat{\theta}_n) \sim n(\theta^*_n - \hat{\theta}_n)^T \mathcal{F}(\theta_0)(\theta^*_n - \hat{\theta}_n).$$

If $H_0$ were simple, say $H_0$: $\theta = \theta_0$, then $\theta^*_n = \theta_0$ and we would be finished, because we know $\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\to} \mathcal{N}(0, \mathcal{F}(\theta_0)^{-1})$. To find the asymptotic distribution of $\sqrt{n}(\theta^*_n - \hat{\theta}_n)$ in general, expand $l_n(\theta^*_n)$ about $\hat{\theta}_n$:

$$\frac{1}{\sqrt{n}} l_n(\theta^*_n) = \frac{1}{\sqrt{n}} l_n(\hat{\theta}_n) + \frac{1}{n} \int_0^1 I_n(\hat{\theta}_n + uv(\theta^*_n - \hat{\theta}_n)) \, du \, \sqrt{n} \, (\theta^*_n - \hat{\theta}_n)$$

$$\sim -\mathcal{F}(\theta_0) \sqrt{n}(\theta^*_n - \hat{\theta}_n).$$

Thus

$$\sqrt{n}(\theta^*_n - \hat{\theta}_n) \sim -\mathcal{F}(\theta_0)^{-1} \frac{1}{\sqrt{n}} l_n(\theta^*_n)$$

and

$$-2 \log \lambda_n \sim \frac{1}{\sqrt{n}} l_n(\theta^*_n)^T \mathcal{F}(\theta_0)^{-1} \frac{1}{\sqrt{n}} l_n(\theta^*_n).$$
To find the asymptotic distribution of \( \hat{I}_n(\theta^*_n) \), expand about \( \theta_0 \):
\[
\frac{1}{\sqrt{n}} \hat{I}_n(\theta^*_n) = \frac{1}{\sqrt{n}} \hat{I}_n(\theta_0) + \frac{1}{n} \int_0^1 \hat{I}_n(\theta_0 + u(\theta^*_n - \theta_0)) \, du \sqrt{n} (\theta^*_n - \theta_0).
\]

(7)

Partition \( \mathcal{A}(\theta_0) \) into four matrices,
\[
\mathcal{A}(\theta_0) = \begin{bmatrix}
\begin{array}{ccc}
 r \times r & r \times (k-r) \\
 G_1 & G_2 \\
 (k-r) \times r & (k-r) \times (k-r) \\
 G_2' & G_3
\end{array}
\end{bmatrix},
\]

and let
\[
H = \begin{bmatrix}
0 & 0 \\
0 & G_3^{-1}
\end{bmatrix}.
\]

Note that the last \( k-r \) components of \( \hat{I}_n(\theta^*_n) \) are zero, so that \( H \hat{I}_n(\theta^*_n) = 0 \) and
\[
H \frac{1}{\sqrt{n}} \hat{I}_n(\theta_0) \sim H \mathcal{A}(\theta_0) \sqrt{n} (\theta^*_n - \theta_0) = \sqrt{n} (\theta^*_n - \theta_0)
\]
since the first \( r \) components of \( \theta^*_n \) and \( \theta_0 \) are zero. Substituting into Eq. (7), we find
\[
\frac{1}{\sqrt{n}} \hat{I}_n(\theta^*_n) \sim [I - \mathcal{A}(\theta_0)H] \frac{1}{\sqrt{n}} \hat{I}_n(\theta_0).
\]

From the Central Limit Theorem,
\[
\frac{1}{\sqrt{n}} \hat{I}_n(\theta_0) \sim \sqrt{n} \left( \frac{1}{n} \hat{I}_n(\theta_0) \right) \overset{d}{\sim} \mathcal{N}(0, \mathcal{A}(\theta_0)).
\]

Hence,
\[
\frac{1}{\sqrt{n}} \hat{I}_n(\theta^*_n) \overset{d}{\sim} [I - \mathcal{A}(\theta_0)H]Y, \quad \text{where } Y \in \mathcal{N}(0, \mathcal{A}(\theta_0)),
\]

so that from Eq. (6),
\[
-2 \log \lambda_n \overset{d}{\sim} Y^T [I - \mathcal{A}(\theta_0)H]^T \mathcal{A}(\theta_0)^{-1} [I - \mathcal{A}(\theta_0)H] Y
\]
\[
= Y^T \mathcal{A}(\theta_0)^{-1} - \mathcal{H} Y \quad \text{[because } \mathcal{H} \mathcal{A}(\theta_0)H = H]\]
\[
= Z^T \mathcal{A}(\theta_0)^{-1/2} [\mathcal{A}(\theta_0)^{-1} - \mathcal{H}] \mathcal{A}(\theta_0)^{1/2} Z,
\]

where \( Z = \mathcal{A}(\theta_0)^{-1/2} Y \in \mathcal{N}(0, I) \). It is easily checked that the matrix \( P = \mathcal{A}(\theta_0)^{1/2} [\mathcal{A}(\theta_0)^{-1} - \mathcal{H}] \mathcal{A}(\theta_0)^{1/2} \) is a projection and that \( \text{rank}(P) = \text{trace}(P) = \text{trace}(\mathcal{A}(\theta_0)[\mathcal{A}(\theta_0)^{-1} - \mathcal{H}]) = \text{trace}(I - \mathcal{A}(\theta_0)H) = r \). Therefore,
\[
-2 \log \lambda_n \overset{d}{\sim} Z^T PZ \in \chi^2_r, \quad \text{as was to be shown.} \]

Note: The maximum-likelihood estimates that appear in the definition of \( \lambda_n \) may be replaced by any of the efficient estimates, such as those of Sections 18 and 19, without disturbing the asymptotic distribution of \(-2 \log \lambda_n\).

**Example 1.** Let \( X_1, \ldots, X_n \) be a sample from \( \mathcal{N} (\mu, \sigma^2) \). Find the likelihood ratio test of the hypothesis \( H_0: \mu = 0, \sigma = 1 \). Here \( r = 2 \) and
\[
L_n(\mu, \sigma) = \left[ \frac{1}{\sqrt{2\pi}\sigma} \right]^n \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \right\},
\]

so that
\[
\lambda_n = \frac{L_n(0, 1)}{L_n(\bar{X}, s)} = \exp \left\{ -\frac{1}{2} \sum_{i=1}^n X_i^2 \right\} \frac{1}{s^{n/2} \exp(-n/2)},
\]

since the maximum-likelihood estimates of \( (\mu, \sigma) \) under \( \Theta \) are \( \hat{\theta} = \bar{X} \), and \( \hat{\sigma}^2 = s^2 = (1/n) \sum_i (X_i - \bar{X})^2 \). Hence,
\[
-2 \log \lambda_n = -n \log s^2 + \sum_{i=1}^n \frac{X_i^2}{2} - n \overset{d}{\sim} \chi^2_2
\]

when \( H_0 \) is true. At the 5% level, we reject \( H_0 \) if
\[
-2 \log \lambda_n > \chi^2_{2,0.05} = 2 \log 20 = 5.99 \ldots.
\]
A Course in Large Sample Theory

Example 2. Let $X_1, \ldots, X_c$ have a multinomial distribution based on $n$ trials, each resulting in one of $c$ outcomes (cells) with respective probabilities $p_1, \ldots, p_c$, where $p_i > 0$ for all $i$, and $\sum_i p_i = 1$. Thus,

$$L_n(p_1, \ldots, p_c) = \prod_{i=1}^{c} p_i^{x_i}$$

provided $X_i$ are integers $\geq 0$, and $\sum_i X_i = n$. Consider testing the hypothesis $H_0$: $p_1 = \cdots = p_c = 1/c$. Even though it appears that there are $c$ restrictions, we have $r = c - 1$ because of the original constraint $\sum_i p_i = 1$. The maximum-likelihood estimates of the $p_i$ under $\Theta$ are

$$\hat{p}_i = X_i/n$$

for $i = 1, \ldots, c$. Hence,

$$\lambda_n = \left[ n \prod_{i=1}^{c} (1/c)^{x_i/n} \right] \prod_{i=1}^{c} \left( x_i/n \right)^{x_i/n} = \prod_{i=1}^{c} \left( x_i/n \right)^{x_i/n}$$

and

$$-2 \log \lambda_n = 2 \sum_{i=1}^{c} x_i \log \left( x_i/n \right) \sim \chi^2_{c-1}$$

under $H_0$. The usual test of $H_0$ in this situation is of course Pearson's $\chi^2$.

Power. We may also find an approximation to the power of the likelihood ratio test at an alternative close to the null hypothesis. Suppose that $\theta$ is the true value and that $\theta_0$ is the parameter point in $H_0$ that is closest to $\theta$. Define $\delta = \sqrt{n} (\theta - \theta_0)$. As in the discussion of the power of Pearson's $\chi^2$ test, we take $\theta$ to be converging to $\theta_0$ in such a way that $\delta$ is fixed. In the proof of Theorem 22, this changes the limiting distribution of $\frac{1}{\sqrt{n}} \lambda_n(\theta_0)$. It may be found by the expansion,

$$\frac{1}{\sqrt{n}} \lambda_n(\theta_0) = \frac{1}{\sqrt{n}} \lambda_n(\theta) + \frac{1}{n} \lambda_n(\theta) \delta \mathcal{A}(\theta_0) \delta$$

where the noncentrality parameter $\varphi$ is

$$\varphi = \delta^T \mathcal{A}(\theta_0)^{1/2} \mathcal{A}(\theta_0)^{1/2} \delta = \delta^T \mathcal{A}(\theta_0)^{-1} \mathcal{A}(\theta_0) \delta.$$

If we use the form of $\mathcal{A}(\theta)$ in terms of the matrices $G_1$, $G_2$, and $G_3$, the noncentrality parameter $\varphi$ reduces to the simpler form,

$$\varphi = \delta_i^T (G_1 - G_2 G_3^{-1} G_2^T) \delta,$$

where $\delta_i$ is the vector of the first $r$ components of $\delta$. Note the effect of nuisance parameters. If $\theta_{r+1}, \ldots, \theta_c$ were known, the noncentrality parameter would be $\delta^T G_1 \delta$.

Example 1 (continued). Let us find the approximate power at the alternative $\mu = 0.2$, $\sigma = 1.2$, when $n = 50$ and the test is conducted at the 5% level. First we compute $\delta^T = \sqrt{n} (0.2, 0.2)$. To compute $\varphi$, recall that Fisher Information for the normal distribution is

$$\mathcal{I}(\mu, \sigma) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{bmatrix}.$$

In this problem the matrix $H$ is empty, so that $\varphi = \delta^T \mathcal{I}(0,1) \delta = 6$. From the Fix Tables (Table 3) of the power of $\chi^2$, we find a power of approximately $\beta = 0.58$. To get a power of 0.9 at this alternative, we need $\varphi$ to be about 12.655, so we must increase $n$ to about 106.

Note that in the calculation of the information matrix in $\varphi$ we used the null hypothesis value, $\sigma = 1$, but from the point of view of the asymptotic theory, the true value, $\sigma = 1.2$, should serve as well. However, this would give a smaller value of $\varphi$, $\varphi = 4.167$, and a power of about $\beta = 0.43$. The sample size is not yet large enough to smooth out this difference. Perhaps a better approximation to the power would be given using the compromise value, $\sigma = 1.1$ ($\beta = 0.50$).

Exercises

1. Let $X_1, \ldots, X_n$ be a sample from $\mathcal{A}(\mu_x, \sigma^2_x)$ and $Y_1, \ldots, Y_n$ be an independent sample from $\mathcal{A}(\mu_y, \sigma^2_y)$. Find the likelihood ratio test for testing $H_0$: $\mu_x = \mu_y$ and $\sigma^2_x = \sigma^2_y$ and state its asymptotic distribution.

2. Let $X_1, \ldots, X_n$ be a sample from the exponential distribution with density $f(x|\theta) = \theta \exp(-\theta x) (x > 0)$ and $Y_1, \ldots, Y_n$ be an independent sample from $f(y|\mu) = \mu \exp(-\mu y) (y > 0)$. Find the likelihood ratio test and its asymptotic distribution for testing $H_0$: $\mu = 2\theta$. 
3. For \( i = 1, \ldots, k \), let \( X_{i1}, X_{i2}, \ldots, X_{in} \) be independent samples from Poisson distributions, \( \mathcal{P}(\theta_i) \), respectively. Find the likelihood ratio test and its asymptotic distribution, for testing \( H_0: \theta_1 = \theta_2 = \cdots = \theta_k \).

4. Show that if \( Z \in \mathcal{N}(\bar{\theta}, I) \) and if \( P \) is a symmetric projection of rank \( r \), then \( Z^T P Z \in \chi^2_r(\bar{\theta}^T P \bar{\theta}) \).

5. (a) Consider the likelihood ratio test of \( H_0: \mu = 0 \) against all alternatives based on a sample of size \( n = 1000 \) from a normal distribution with mean \( \mu \) and unknown standard deviation \( \sigma \). What is the approximate distribution of \( -2 \log \lambda_n \) if the true values of the parameters are \( \mu = 0.1 \) and \( \sigma = \sigma_0 \) for some fixed \( \sigma_0 \)?

(b) Suppose instead the distribution is \( \mathcal{N}(\alpha, \beta) \) and \( H_0: \alpha = 1 \) with \( \beta \) unknown. What is the approximate distribution of \( -2 \log \lambda_n \) if the true values of the parameters are \( \alpha = 1.1 \) and \( \beta = \beta_0 \)? (Note that this distribution is independent of \( \beta_0 \).)

6. One-Sided Likelihood Ratio Tests. The likelihood ratio test against one-sided alternatives is more complex and is no longer asymptotically distribution-free under the null hypothesis. This may be illustrated in testing \( H_0: \theta = \theta_0 \) when \( \theta \) is two-dimensional. Make the same assumptions as in Theorem 22, with \( k = r = 2 \) and take \( \theta_0 = \theta \).

(a) Let \( \lambda_n \) denote the likelihood ratio test statistic for testing \( H_0: \theta = \theta_0 \) against \( H_1: \theta_1 > 0, \theta_2 \) unrestricted. Show that under the null hypothesis, \( -2 \log \lambda_n \overset{d}{\to} 0.5 \chi_1^2 + 0.5 \chi_2^2 \) (the mixture of a \( \chi_1^2 \) and a \( \chi_2^2 \) with probability 0.5 each).

(b) In testing \( H_0: \theta = 0 \) against \( H_1: \theta_1 \geq 0, \theta_2 \geq 0, \theta_0 = 0 \), show that \( -2 \log \lambda_n \overset{d}{\to} p \delta_0 + 0.5 \chi_1^2 + (0.5 - p) \chi_2^2 \) under \( H_0 \), where \( \delta_0 \) is the distribution degenerate at 0, and \( p = \arccos(\rho) / 2\pi \), where \( \rho \) is the correlation coefficient of the variables whose covariance matrix is \( \mathcal{N}(\theta_0) \).

Thus the limiting distribution of \( -2 \log \lambda_n \) depends on the correlation of the underlying distribution.

---

Minimum Chi-Square Estimates

In this section we treat estimation problems by minimum distance methods, using a general theory of quadratic forms in asymptotically normal variables. This theory contains minimum \( \chi^2 \) methods as a particular case.

We observe a sequence of \( d \)-dimensional random vectors \( Z_n \) whose distribution depends upon a \( k \)-dimensional parameter \( \theta \) lying in a parameter space \( \Theta \) assumed to be a nonempty open subset of \( \mathbb{R}^k \) where \( k \leq d \).

It is given that the \( Z_n \) are asymptotically normal:

\[
\sqrt{n} (Z_n - \Lambda(\theta)) \overset{d}{\to} \mathcal{N}(0, C(\theta)),
\]

where \( \Lambda(\theta) \) is a \( d \) vector and \( C(\theta) \) is a \( d \times d \) covariance matrix for all \( \theta \in \Theta \). We make two assumptions on \( \Lambda(\theta) \):

- \( \Lambda(\theta) \) is bicontinuous (that is, \( \theta_n \to \theta \Rightarrow \Lambda(\theta_n) \to \Lambda(\theta) \)),
- \( \Lambda(\theta) \) has a continuous first partial derivative, \( \Lambda(\theta) \), of full rank \( k \).

We measure the distance of \( Z_n \) to \( \Lambda(\theta) \) through a quadratic form of the type

\[
Q_n(\theta) = (Z_n - \Lambda(\theta))^T M(\theta)(Z_n - \Lambda(\theta)),
\]

where \( M(\theta) \) is a \( d \times d \) covariance matrix. We assume

- \( M(\theta) \) is continuous in \( \theta \) and uniformly bounded below in the sense that for some constant \( \alpha > 0 \) we have \( M(\theta) > \alpha I \) for all \( \theta \in \Theta \).