Asymptotic Distribution of the Likelihood Ratio Test Statistic

Let $X_1, \ldots, X_n$ be a sample from density $f(x|\boldsymbol{\theta})$ where $\boldsymbol{\theta} \subset \Theta \subset \mathbb{R}^k$. The likelihood ratio test provides a general method for testing H_0: $\boldsymbol{\theta} \in \Theta_0$ versus H_1: $\boldsymbol{\theta} \in \Theta - \Theta_0$ for a given subset Θ_0 of Θ. This tests rejects H_0 when the likelihood ratio test statistic,

$$\lambda_n = \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} \prod_1^n f(x_j|\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta} \prod_1^n f(x_j|\boldsymbol{\theta})} = \frac{L_n(\boldsymbol{\theta}_n^*)}{L_n(\hat{\boldsymbol{\theta}}_n)} \tag{1}$$

is too small, where $\boldsymbol{\theta}_n^*$ is the MLE over Θ_0, and $\hat{\boldsymbol{\theta}}_n$ is the MLE over Θ. 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 Θ_0 is a $(k-r)$-dimensional subspace of Θ. Writing the components of the vector $\boldsymbol{\theta} \in \mathbb{R}^k$ as $\boldsymbol{\theta}^T = (\theta^1, \theta^2, \ldots, \theta^k)$, we assume the null hypothesis is of the form

$$H_0: \quad \theta^1 = \theta^2 = \cdots = \theta^r = 0 \tag{2}$$

where $1 \le r \le k$. More general situations, in which H_0 is of the form H_0: $g_1(\boldsymbol{\theta}) = \cdots = g_r(\boldsymbol{\theta}) = 0$ for some smooth real-valued functions $g_1, \ldots, g_r$, 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 $1 \le r \le k$. Suppose that the true value $\boldsymbol{\theta}_0$ satisfies H_0. Then*

$$-2\log\lambda_n \xrightarrow{\mathscr{L}} \chi_r^2. \tag{3}$$

Proof. $-2\log\lambda_n = 2[l_n(\hat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}_n^*)]$ where $\hat{\boldsymbol{\theta}}_n$ = MLE over Θ, and $\boldsymbol{\theta}_n^*$ = MLE over Θ_0. Expand $l_n(\boldsymbol{\theta}_n^*)$ about $\hat{\boldsymbol{\theta}}_n$:

$$l_n(\boldsymbol{\theta}_n^*) = l_n(\hat{\boldsymbol{\theta}}_n) + \dot{l}_n(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta}_n^* - \hat{\boldsymbol{\theta}}_n) - n(\boldsymbol{\theta}_n^* - \hat{\boldsymbol{\theta}}_n)^T \mathbf{I}_n(\boldsymbol{\theta}_n^*)(\boldsymbol{\theta}_n^* - \hat{\boldsymbol{\theta}}_n),$$

where

$$\mathbf{I}_n(\boldsymbol{\theta}_n^*) = -\frac{1}{n}\int_0^1\int_0^1 v\ddot{l}_n\big(\hat{\boldsymbol{\theta}}_n + uv(\boldsymbol{\theta}_n^* - \hat{\boldsymbol{\theta}}_n)\big)\,du\,dv \xrightarrow{\text{a.s.}} \frac{1}{2}\mathscr{I}(\boldsymbol{\theta}_0),$$

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

$$\begin{aligned} -2\log\lambda_n &= 2n(\boldsymbol{\theta}_n^* - \hat{\boldsymbol{\theta}}_n)^T \mathbf{I}_n(\boldsymbol{\theta}_n^*)(\boldsymbol{\theta}_n^* - \hat{\boldsymbol{\theta}}_n) \\ &\sim n(\boldsymbol{\theta}_n^* - \hat{\boldsymbol{\theta}}_n)^T \mathscr{I}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_n^* - \hat{\boldsymbol{\theta}}_n). \end{aligned} \tag{4}$$

If H_0 were simple, say H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, then $\boldsymbol{\theta}_n^* = \boldsymbol{\theta}_0$ and we would be finished, because we know $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathscr{L}} \mathscr{N}(\mathbf{0}, \mathscr{I}(\boldsymbol{\theta}_0)^{-1})$. To find the asymptotic distribution of $\sqrt{n}(\boldsymbol{\theta}_n^* - \hat{\boldsymbol{\theta}}_n)$ in general, expand $\dot{l}_n(\boldsymbol{\theta}_n^*)$ about $\hat{\boldsymbol{\theta}}_n$:

$$\begin{aligned} \frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\theta}_n^*) &= \frac{1}{\sqrt{n}}\dot{l}_n(\hat{\boldsymbol{\theta}}_n) + \frac{1}{n}\int_0^1 \ddot{l}_n\big(\hat{\boldsymbol{\theta}}_n + v(\boldsymbol{\theta}_n^* - \hat{\boldsymbol{\theta}}_n)\big)\,dv\sqrt{n}(\boldsymbol{\theta}_n^* - \hat{\boldsymbol{\theta}}_n) \\ &\sim -\mathscr{I}(\theta_0)\sqrt{n}(\boldsymbol{\theta}_n^* - \hat{\boldsymbol{\theta}}_n) \end{aligned}$$

Thus

$$\sqrt{n}(\boldsymbol{\theta}_n^* - \hat{\boldsymbol{\theta}}_n) \sim -\mathscr{I}(\theta_0)^{-1}\frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\theta}_n^*) \tag{5}$$

and

$$-2\log\lambda_n \sim \frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\theta}_n^*)^T \mathscr{I}(\theta_0)^{-1}\frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\theta}_n^*). \tag{6}$$

To find the asymptotic distribution of $\dot{l}_n(\boldsymbol{\theta}_n^*)$, expand about $\boldsymbol{\theta}_0$:

$$\frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\theta}_n^*) = \frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\theta}_0) + \frac{1}{n}\int_0^1 \ddot{l}_n(\boldsymbol{\theta}_0 + v(\boldsymbol{\theta}_n^* - \boldsymbol{\theta}_0))\,dv\,\sqrt{n}\,(\boldsymbol{\theta}_n^* - \boldsymbol{\theta}_0). \tag{7}$$

Partition $\mathscr{I}(\boldsymbol{\theta}_0)$ into four matrices,

$$\mathscr{I}(\boldsymbol{\theta}_0) = \begin{bmatrix} \underset{r\times r}{\mathbf{G}_1} & \underset{r\times(k-r)}{\mathbf{G}_2} \\ \underset{(k-r)\times r}{\mathbf{G}_2^T} & \underset{(k-r)\times(k-r)}{\mathbf{G}_3} \end{bmatrix},$$

and let

$$\mathbf{H} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_3^{-1} \end{bmatrix}.$$

Note that the last $k-r$ components of $\dot{l}_n(\boldsymbol{\theta}_n^*)$ are zero, so that $\mathbf{H}\dot{l}_n(\boldsymbol{\theta}_n^*) = \mathbf{0}$ and

$$\mathbf{H}\frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\theta}_0) \sim \mathbf{H}\mathscr{I}(\boldsymbol{\theta}_0)\sqrt{n}\,(\boldsymbol{\theta}_n^* - \boldsymbol{\theta}_0) = \sqrt{n}\,(\boldsymbol{\theta}_n^* - \boldsymbol{\theta}_0)$$

since the first r components of $\boldsymbol{\theta}_n^*$ and $\boldsymbol{\theta}_0$ are zero. Substituting into Eq. (7), we find

$$\frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\theta}_n^*) \sim [\mathbf{I} - \mathscr{I}(\boldsymbol{\theta}_0)\mathbf{H}]\frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\theta}_0).$$

From the Central Limit Theorem,

$$\frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\theta}_0) = \sqrt{n}\left(\frac{1}{n}\dot{l}_n(\boldsymbol{\theta}_0)\right) \xrightarrow{\mathscr{L}} \mathscr{N}(\mathbf{0}, \mathscr{I}(\boldsymbol{\theta}_0)).$$

Hence,

$$\frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\theta}_n^*) \xrightarrow{\mathscr{L}} [\mathbf{I} - \mathscr{I}(\boldsymbol{\theta}_0)\mathbf{H}]\mathbf{Y}, \qquad \text{where } \mathbf{Y} \in \mathscr{N}(\mathbf{0}, \mathscr{I}(\boldsymbol{\theta}_0)),$$

so that from Eq. (6),

$$\begin{aligned} -2\log\lambda_n &\xrightarrow{\mathscr{L}} \mathbf{Y}^T[\mathbf{I} - \mathscr{I}(\boldsymbol{\theta}_0)\mathbf{H}]^T\mathscr{I}(\boldsymbol{\theta}_0)^{-1}[\mathbf{I} - \mathscr{I}(\boldsymbol{\theta}_0)\mathbf{H}]\mathbf{Y} \\ &= \mathbf{Y}^T\left[\mathscr{I}(\boldsymbol{\theta}_0)^{-1} - \mathbf{H}\right]\mathbf{Y} \qquad [\text{because } \mathbf{H}\mathscr{I}(\boldsymbol{\theta}_0)\mathbf{H} = \mathbf{H}] \\ &= \mathbf{Z}^T\mathscr{I}(\boldsymbol{\theta}_0)^{1/2}\left[\mathscr{I}(\boldsymbol{\theta}_0)^{-1} - \mathbf{H}\right]\mathscr{I}(\boldsymbol{\theta}_0)^{1/2}\mathbf{Z}, \end{aligned}$$

where $\mathbf{Z} = \mathscr{I}(\boldsymbol{\theta}_0)^{-1/2}\mathbf{Y} \in \mathscr{N}(\mathbf{0}, \mathbf{I})$). It is easily checked that the matrix $\mathbf{P} = \mathscr{I}(\boldsymbol{\theta}_0)^{1/2}[\mathscr{I}(\boldsymbol{\theta}_0)^{-1} - \mathbf{H}]\mathscr{I}(\boldsymbol{\theta}_0)^{1/2}$ is a projection and that rank($\mathbf{P}$) = trace($\mathbf{P}$) = trace($\mathscr{I}(\boldsymbol{\theta}_0)[\mathscr{I}(\boldsymbol{\theta}_0)^{-1} - \mathbf{H}]$) = trace($\mathbf{I} - \mathscr{I}(\boldsymbol{\theta}_0)\mathbf{H}$) = r. Therefore $-2\log\lambda_n \xrightarrow{\mathscr{L}} \mathbf{Z}^T\mathbf{P}\mathbf{Z} \in \chi_r^2$, as was to be shown. ■

Note: The maximum-likelihood estimates that appear in the definition of λ_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 $\mathscr{N}(\mu, \sigma^2)$. Find the likelihood ratio test of the hypothesis H_0: $\mu = 0$, $\sigma = 1$. Here $r = 2$ and

$$L_n(\mu, \sigma) = \left[\frac{1}{\sqrt{2\pi}\,\sigma}\right]^n \exp\left\{-\frac{1}{2\sigma^2}\sum_1^n (X_j - \mu)^2\right\},$$

so that

$$\lambda_n = \frac{L_n(0,1)}{L_n(\bar{X}, s)} = \frac{\exp\left\{-\frac{1}{2}\sum_1^n X_j^2\right\}}{s^{-n}\exp\{-n/2\}},$$

since the maximum-likelihood estimates of (μ, σ) under Θ are $\hat{\mu} = \bar{X}$, and $\hat{\sigma}^2 = s^2 = (1/n)\Sigma_1^n(X_i - \bar{X})^2$. Hence,

$$-2\log\lambda_n = -n\log s^2 + \sum_1^n X_j^2 - n \xrightarrow{\mathscr{L}} \chi_2^2$$

when H_0 is true. At the 5% level, we reject H_0 if

$$-2\log\lambda_n > \chi^2_{2;0.05} = 2\log 20 = 5.99\ldots.$$

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 $\Sigma_1^c\, p_i = 1$. Thus,

$$L_n(p_1, \ldots, p_c) = \binom{n}{x_1 \ldots x_c} \prod_1^c p_j^{x_j}$$

provided X_i are integers ≥ 0, and $\Sigma_1^c X_i = n$. 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 $\Sigma_1^c\, p_i = 1$. The maximum-likelihood estimates of the p_i under Θ are $\hat{p}_i = X_i/n$ for $i = 1, \ldots, c$. Hence,

$$\lambda_n = \frac{\binom{n}{x_1 \cdots x_c} \prod_1^c (1/c)^{x_j}}{\binom{n}{x_1 \cdots x_c} \prod_1^c (x_j/n)^{x_j}} = \prod_1^c \left(\frac{n}{cx_j} \right)^{x_j}$$

and

$$-2 \log \lambda_n = 2 \sum_1^c x_j \log\left(\frac{cx_j}{n} \right) \xrightarrow{\mathscr{L}} \chi^2_{c-1}$$

under H_0. The usual test of H_0 in this situation is of course Pearson's χ^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 $\boldsymbol{\theta}$ is the true value and that $\boldsymbol{\theta}_0$ is the parameter point in H_0 that is closest to $\boldsymbol{\theta}$. Define $\boldsymbol{\delta} = \sqrt{n}\,(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$. As in the discussion of the power of Pearson's χ^2 test, we take $\boldsymbol{\theta}$ to be converging to $\boldsymbol{\theta}_0$ in such a way that $\boldsymbol{\delta}$ is fixed. In the proof of Theorem 22, this changes the limiting distribution of $(1/\sqrt{n})\dot{l}_n(\boldsymbol{\theta}_0)$. It may be found by the expansion,

$$\frac{1}{\sqrt{n}} \dot{l}_n(\boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}} \dot{l}_n(\boldsymbol{\theta}) + \frac{1}{n} \ddot{l}_n(\boldsymbol{\theta}) \sqrt{n}\,(\boldsymbol{\theta}_0 - \boldsymbol{\theta})$$

$$\xrightarrow{\mathscr{L}} \mathbf{Y} = \mathscr{N}(\mathbf{0}, \mathscr{I}(\boldsymbol{\theta}_0)) + \mathscr{I}(\boldsymbol{\theta}_0)\boldsymbol{\delta} = \mathscr{N}(\mathscr{I}(\boldsymbol{\theta}_0)\boldsymbol{\delta}, \mathscr{I}(\boldsymbol{\theta}_0)).$$

As before, if we let $\mathbf{Z} = \mathscr{I}(\boldsymbol{\theta}_0)^{-1/2}\,\mathbf{Y}$, then $-2 \log \lambda_n \xrightarrow{\mathscr{L}} \mathbf{Z}^T \mathbf{P} \mathbf{Z}$, where $\mathbf{P} = \mathscr{I}(\boldsymbol{\theta}_0)^{1/2} [\mathscr{I}(\boldsymbol{\theta}_0)^{-1} - \mathbf{H} [\mathscr{I}(\boldsymbol{\theta}_0)^{1/2}$ is a projection of rank r, but this time $\mathbf{Z} \in \mathscr{N}(\mathscr{I}(\boldsymbol{\theta}_0)^{1/2}\boldsymbol{\delta}, \mathbf{I}))$ so that (see Exercise 4),

$$-2 \log \lambda_n \xrightarrow{\mathscr{L}} \mathbf{Z}^T \mathbf{P} \mathbf{Z} \in \chi_r^2(\varphi),$$

where the noncentrality parameter φ is

$$\varphi = \boldsymbol{\delta}^T \mathscr{I}(\boldsymbol{\theta}_0)^{1/2} \mathbf{P} \mathscr{I}(\boldsymbol{\theta}_0)^{1/2} \boldsymbol{\delta} = \boldsymbol{\delta}^T \mathscr{I}(\boldsymbol{\theta}_0) \left[\mathscr{I}(\boldsymbol{\theta}_0)^{-1} - \mathbf{H} \right] \mathscr{I}(\boldsymbol{\theta}_0) \boldsymbol{\delta}.$$

If we use the form of $\mathscr{I}(\boldsymbol{\theta})$ in terms of the matrices $\mathbf{G}_1$, $\mathbf{G}_2$, and $\mathbf{G}_3$, the noncentrality parameter φ reduces to the simpler form,

$$\varphi = \boldsymbol{\delta}_r^T (\mathbf{G}_1 - \mathbf{G}_2 \mathbf{G}_3^{-1} \mathbf{G}_2^T) \boldsymbol{\delta}_r,$$

where $\boldsymbol{\delta}_r$ is the vector of the first r components of $\boldsymbol{\delta}$. Note the effect of nuisance parameters. If $\theta_{r+1}, \ldots, \theta_k$ were known, the noncentrality parameter would be $\boldsymbol{\delta}_r^T \mathbf{G}_1 \boldsymbol{\delta}_r$.

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 $\boldsymbol{\delta}^T = \sqrt{n}\,(0.2, 0.2)$. To compute φ, recall that Fisher Information for the normal distribution is

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

In this problem the matrix $\mathbf{H}$ is empty, so that $\varphi = \boldsymbol{\delta}^T \mathscr{I}(0, 1) \boldsymbol{\delta} = 6$. From the Fix Tables (Table 3) of the power of χ_2^2, we find a power of approximately $\beta = 0.58$. To get a power of 0.9 at this alternative, we need φ to be about 12.655, so we must increase n to about 106.

Note that in the calculation of the information matrix in φ 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 φ, $\varphi = 4.167$, 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 $\mathscr{N}(\mu_x, \sigma_x^2)$ and $Y_1, \ldots, Y_n$ be an independent sample from $\mathscr{N}(\mu_y, \sigma_y^2)$. Find the likelihood ratio test for testing H_0: $\mu_x = \mu_y$ and $\sigma_x^2 = \sigma_y^2$ and state its asymptotic distribution.
2. Let $X_1, \ldots, X_n$ be a sample from the exponential distribution with density $f(x|\theta) = \theta \exp\{-\theta x\} I\ (x > 0)$ and $Y_1, \ldots, Y_n$ be an independent sample from $f(y|\mu) = \mu \exp\{-\mu y\} I\ (y > 0)$. 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, $\mathscr{P}(\theta_i)$, respectively. Find the likelihood ratio test and its asymptotic distribution, for testing $H_0\colon \theta_1 = \theta_2 = \cdots = \theta_k$.
4. Show that if $\mathbf{Z} \in \mathscr{N}(\boldsymbol{\delta}, \mathbf{I}))$ and if $\mathbf{P}$ is a symmetric projection of rank r, then $\mathbf{Z}^T\mathbf{P}\mathbf{Z} \in \chi_r^2(\boldsymbol{\delta}^T\mathbf{P}\boldsymbol{\delta})$.
5. (a) Consider the likelihood ratio test of $H_0\colon \mu = 0$ against all alternatives based on a sample of size $n = 1000$ from a normal distribution with mean μ and unknown standard deviation σ. 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 σ_0?
   (b) Suppose instead the distribution is $\mathscr{G}(\alpha, \beta)$ and $H_0\colon \alpha = 1$ with β 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 β_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 $H_0\colon \boldsymbol{\theta} = \boldsymbol{\theta}_0$ when $\boldsymbol{\theta}$ is two-dimensional. Make the same assumptions as in Theorem 22, with $k = r = 2$ and take $\boldsymbol{\theta}_0 = \mathbf{0}$.
   (a) Let λ_n denote the likelihood ratio test statistic for testing H_0: $\boldsymbol{\theta} = \mathbf{0}$ against $H_1\colon \theta_1 > 0$, θ_2 unrestricted. Show that under the null hypothesis, $-2\log\lambda_n \xrightarrow{\mathscr{L}} 0.5\chi_1^2 + 0.5\chi_2^2$ (the mixture of a χ_1^2 and a χ_2^2 with probability 0.5 each).
   (b) In testing $H_0\colon \boldsymbol{\theta} = \mathbf{0}$ against $H_1\colon \theta_1 \geq 0$, $\theta_2 \geq 0$, $\boldsymbol{\theta} \neq \mathbf{0}$, show that $-2\log\lambda_n \xrightarrow{\mathscr{L}} p\delta_0 + 0.5\chi_1^2 + (0.5 - p)\chi_2^2$ under H_0, where δ_0 is the distribution degenerate at 0, and $p = \arccos(\rho)/2\pi$, where ρ is the correlation coefficient of the variables whose covariance matrix is $\mathscr{I}(\boldsymbol{\theta}_0)$. Thus the limiting distribution of $-2\log\lambda_n$ depends on the correlation of the underlying distribution.