Primer on Likelihood Ratio Tests

Ariel Boyarsky

January 15, 2021

1 Likelihood Ratio Tests

Suppose that we have data,

$$X \sim P_{\theta_0}$$

where $P_{\theta_0}$is some distribution parameterized by θ_0. Given some estimate of the parameters $\hat{\theta}$ we want to test the likelihood that we see X conditional on $\theta_0 = \hat{\theta}$. That is we are interested in the test,

$$H_0 : \theta_0 \in \Theta_0 \text{ v.s. } H_1 : \theta \in \Theta$$

where Θ_0 is some restriction of our parameter space, for example $\Theta_0 = \{\theta : \theta_i = 0 \text{ for all } i\}$. Thus we test for some realization X whether or not the null hypothesis likely to admit this realization. So we can define the likelihood ratio,

$$\lambda(x) = \frac{\sup\{L(\theta; x) : \theta \in \Theta_0\}}{\sup\{L(\theta; x) : \theta \in \Theta\}}$$

Clearly since $\Theta_0 \subset \Theta$ we have that the likelihood ratio is bounded from above by 1. If the ratio approaches 1 this implies that the null hypothesis is "as likely" as the alternative which covers the full parameter space. However, if the ratio is close to 0 this implies the likelihood of the null is small relative to the alternative. This gives us evidence to reject the null. Also note that we can reverse the ratio and thus the above logic if we wish. Consider testing,

$$H_0 : \theta_0 \in \Theta_0 \text{ v.s. } H_1 : \theta \in \Theta_1$$

The critical region for some $k \in (0, 1)$ is then given by,

$$C_1 = \{x : \lambda(x) \leq k\}$$

Furthermore, for some k the test is at the significance level α if,

$$\sup\{P(\lambda(x) \leq k; \theta \in \Theta_0)\} = \alpha$$

Thus if we can determine the appropriate k for a given α we can examine if the likelihood ratio likes within the critical region for that k. To determine this we must look to the cdf of $\lambda(X)$. So for asymptotic inference we need to consider the asymptotic distribution of the likelihood ratio. In particular, consider the function $-2\log\lambda(x)$ which is decreasing and so we can equivalently write the critical region as,

$$C_1 = \{x : \Lambda(x) \geq c\} = \{x : -2\log\lambda(x) \geq c\}$$

with

$$\Lambda(x) \equiv -2\log\lambda(x) = 2\left[l(\hat{\theta}; x) - l(\theta_0; x)\right]$$

where $l(\cdot; x)$ is the log-likelihood. Then, $\Lambda(x)$ is the **likelihood ratio statistic.**

1.1 Asymptotic Behavior

To determine the asymptotic distribution begin with a Taylor expansion of the log-likelihood at θ_0 about $\hat{\theta}$,

$$\begin{aligned} l(\theta_0) &= l(\hat{\theta}) + \left(\hat{\theta} - \theta_0\right) l'(\hat{\theta}) + \frac{1}{2}\left(\hat{\theta} - \theta_0\right)^2 l''(\hat{\theta}) + o_{P_{\theta_0}}(1) \\ &= l(\hat{\theta}) + \frac{1}{2}\left(\hat{\theta} - \theta_0\right)^2 l''(\hat{\theta}) + o_{P_{\theta_0}}(1) \end{aligned}$$

Using the fact that $\hat{\theta}$ is the maximizes the likelihood and lies within the interior of Θ. So,

$$\begin{aligned} \Lambda(x) = 2\left[l(\hat{\theta}) - l(\theta_0)\right] &= -\left(\hat{\theta} - \theta_0\right)^2 l''(\hat{\theta}) \\ &= \left(\hat{\theta} - \theta_0\right)^2 \left[J(\hat{\theta})\right] \\ &= \left(\hat{\theta} - \theta_0\right)^2 \left[I(\theta_0)\frac{J(\hat{\theta})}{I(\theta_0)}\right] \end{aligned}$$

Here we use that fact that the observed information is given by,

$$J(\theta) = \frac{\partial^2 l(\theta)}{\partial \theta^2}$$

So that the Fisher information is given by,

$$I(\theta) = \mathbb{E}\left[J(\theta)\right]$$

And we know that,

$$\left(\hat{\theta} - \theta_0\right)\sqrt{I(\theta_0)} \xrightarrow{d} N(0,1) \text{ and } \frac{J(\hat{\theta})}{I(\theta_0)} \xrightarrow{p} 1$$

In particular from the proof of the asymptotic normality of the MLE we have,

$$\begin{aligned} \sum_i \frac{\partial}{\partial \theta} \log f(x_i;\theta) - J(\theta)(\hat{\theta} - \theta) &\approx 0 \\ \sqrt{I(\theta)}(\hat{\theta} - \theta) &\approx \left[\sum_i \frac{\partial}{\partial \theta} \log f(x_i;\theta)\right] \frac{\sqrt{I(\theta)}}{J(\theta)} \\ &= \frac{\left[\sum_i \frac{\partial}{\partial \theta} \log f(x_i;\theta)\right]}{\sqrt{I(\theta)}} \left(\frac{J(\theta)}{I(\theta)}\right)^{-1} \end{aligned}$$

And we have that $\frac{\left[\sum_i \frac{\partial}{\partial \theta} \log f(x_i;\theta)\right]}{\sqrt{I(\theta)}} \xrightarrow{d} N(0,1)$ and $\frac{J(\theta)}{I(\theta)} \xrightarrow{p} 1$ and so the first result follows by Slutsky. Note that the first occurs because $\left[\sum_i \frac{\partial}{\partial \theta} \log f(x_i;\theta)\right]$ is simply the score function (derivative of the log-likelihood) and so it admits asymptotic normality in the MLE. Also,

$$\frac{J(\theta)}{I(\theta)} = \frac{-\sum_i \frac{\partial^2}{\partial \theta^2} \log f(x_i;\theta)}{ni(\theta)} \xrightarrow{p} \frac{\mathbb{E}\left[-\frac{\partial^2}{\partial \theta^2} \log f(x_i;\theta)\right]}{\mathbb{E}\left[-\frac{\partial^2}{\partial \theta^2} \log f(x_i;\theta)\right]} = 1$$

Using the fact the observations are a random sample such that $I(\theta) = ni(\theta)$.
So we have that by Slutsky's lemma we have,

$$\Lambda(x) = \left(\hat{\theta} - \theta_0\right)^2 \left[I(\theta_0)\frac{J(\hat{\theta})}{I(\theta_0)}\right] \xrightarrow{d} \chi^2_1$$

If instead we have that H_0 is a composite null, that is Θ_0 is a linear subspace with dimension greater than 1 so that the null and alternate hypothesis fully specify the distribution. Then it turns out that,

$$\Lambda(x) \xrightarrow{d} \chi^2_{k-l}$$

where $k = \dim \Theta$ and $l = \dim \Theta_0$. This happens because then the quadratic approximation yields a sum of $k - l$ square normals.

Example 1 (One-sample t-test)**.** Suppose we want to test,

$$H_0 : \theta = \theta_0$$

where θ is the mean of a normal distribution with unknown variance.

$$\Theta = \left\{ \left(\theta, \sigma^2\right) : \theta \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+ \right\}$$
$$\Theta_0 = \left\{ \left(\theta, \sigma^2\right) : \theta = \theta_0, \sigma^2 \in \mathbb{R}^+ \right\}$$

Then the density is given by,

$$f(x; \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \left(x - \theta\right)^2 \right)$$

Which gives a likelihood function,

$$L(\theta, \sigma^2; x) = \prod_i f(x_i; \theta, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_i \left(x_i - \theta\right)^2 \right)$$

The log-likelihood for the null is then,

$$l(\theta_0, \sigma^2; x) = -\frac{n}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_i \left(x_i - \theta_0\right)^2$$

The FOC is,

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_i \left(x_i - \theta_0\right)^2$$
$$\implies \sigma^2 = \frac{1}{n} \sum_{i=1}^n \left(x_i - \theta_0\right)^2$$

Plugging this back in we get that,

$$\begin{aligned} \sup L(\theta_0, \sigma^2; x) &= \left(2\pi \frac{1}{n} \sum_{i=1}^n \left(x_i - \theta_0\right)^2 \right)^{-n/2} \exp\left(-\frac{1}{2\frac{1}{n}\sum_{i=1}^n \left(x_i - \theta_0\right)^2} \sum_i \left(x_i - \theta\right)^2 \right) \\ &= \left(\frac{2\pi}{n} \sum_{i=1}^n \left(x_i - \theta_0\right)^2 \right)^{-n/2} \exp\left(-\frac{n}{2}\right) \end{aligned}$$

Then using the fact that $\bar{x}$ is the MLE in this setting, we know,

$$\sup L(\theta, \sigma^2; x) = \left(\frac{2\pi}{n} \sum_{i=1}^n \left(x_i - \bar{x}\right)^2 \right)^{-n/2} \exp\left(-\frac{n}{2}\right)$$

So,

$$\lambda(x) = \left(\frac{\frac{2\pi}{n}\sum_{i=1}^n \left(x_i - \theta_0\right)^2}{\frac{2\pi}{n}\sum_{i=1}^n \left(x_i - \bar{x}\right)^2} \right)^{-n/2}$$

With a little algebra we can show this is equivalent to the usual one-sample t-statistic. Begin by noticing,

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\theta_0)^2 &= \sum_{i=1}^{n}\left((x_i-\bar{x})+(\bar{x}-\theta_0)\right)^2 \\
&= \sum_{i=1}^{n}(x_i-\bar{x})^2 + 2(\bar{x}-\theta_0)(x_i-\bar{x}) + (\bar{x}-\theta_0)^2 \\
&= n(\bar{x}-\theta_0)^2 + \sum_{i=1}^{n}(x_i-\bar{x})\left((x_i-\bar{x})+2(\bar{x}-\theta_0)\right) \\
&= n(\bar{x}-\theta_0)^2 + \sum_{i=1}^{n}(x_i-\bar{x})(x_i+\bar{x}-2\theta_0) \\
&= n(\bar{x}-\theta_0)^2 + \sum_{i=1}^{n}(x_i-\bar{x})^2
\end{aligned}$$

Then,

$$\begin{aligned}
\lambda(x) &= \left(\frac{\frac{2\pi}{n}\left(n(\bar{x}-\theta_0)^2+\sum_{i=1}^{n}(x_i-\bar{x})^2\right)}{\frac{2\pi}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^{-n/2} \\
&= \left(\frac{n(\bar{x}-\theta_0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}+1\right)^{-n/2}
\end{aligned}$$

Thus, if the critical region is given by,

$$C_1 = \{x : \lambda(x) \leq k\}$$

So it is clear that the above is a function of,

$$\frac{|\bar{x}-\theta_0|}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

Note that this is very close to the t-statistic,

$$\frac{\bar{x}-\theta}{S/\sqrt{n}} \sim t(n-1)$$

So we can also write the critical region as,

$$C_1 = \left\{x : \frac{|\bar{x}-\theta|}{S/\sqrt{n}} \geq c\right\}$$

That is there is a correspondence between the rejection regions of the likelihood ratio test and the one-sample t statistic.

In particular define,

$$t = \frac{\sqrt{n}\,|\bar{x}-\theta|}{\sqrt{\frac{1}{n}\sum_i(x_i-\bar{x})^2}}$$

$$\begin{aligned}
\lambda(x) &= \left(\frac{n(\bar{x}-\theta_0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}+1\right)^{-n/2} \\
&= \left(\sqrt{n}t^2+1\right)^{-n/2}
\end{aligned}$$

Hence establishing an equivalence between the two tests.

1.2 Asymptotic Power

Consider the power function for a set critical value $\chi^2_{k-l,\alpha}$,

$$\pi_n(\theta + \frac{h}{\sqrt{n}}) = P_{\theta+h/\sqrt{n}}\left(\Lambda(x) > \chi^2_{k-l,\alpha}\right)$$

2 Neyman-Pearson Lemma

Neyman and Pearson developed likelihood over a series of papers to unify various hypothesis tests. In particular, Neyman and Pearson (1933) shows that the likelihood ratio test is most powerful. This result is the Neyman-Pearson lemma.

Theorem 1 (Neyman-Pearson Lemma)**.** *Suppose $\Omega_0 = \{\theta_0\}$ and $\Omega_1 = \{\theta_1\}$ are both simple and P_{θ_0} and P_{θ_1} have densities p_{θ_0} and p_{θ_1} with respect to μ. Then for any $\alpha \in (0,1)$,*

1. *(Existence) there exists a possible randomized test,*

$$\phi : \mathcal{X} \to [0,1]$$

   (a) $\mathbb{E}_{\theta_0}[\phi(x)] = \alpha$

   (b) $\phi(x) = \begin{cases} 1 & p_1(x) > k \cdot p_0(x) \\ 0 & p_1(x) < k \cdot p_0(x) \end{cases}$ *for some* $k \equiv k(\alpha)$

2. *(Optimality) Any test ϕ that satisfies both conditions of (1) is* ***most powerful:*** *If ϕ' is any other test with $\mathbb{E}_{\theta_0}[\phi'(x)] \leq \alpha$ then $\mathbb{E}_{\theta_1}[\phi'(x)] \leq \mathbb{E}_{\theta_1}[\phi(x)]$*

3. *(Uniqueness) Any test that is most powerful must satisfy (1b) for some $k = k(\alpha)$ almost everywhere for μ.*

Remark 1. The theorem states that the likelihood ratio test which rejects the null whenever the ratio is greater than some k such that rejection occurs with probability α then this test is most powerful.