Examples and Extensions.

March 5, 2003

The asymptotic distribution of the likelihood ratio statistic, the Wald statistic and the score statistic under a sequence of contiguous alternatives of the form $\theta_n = \theta_0 + h\,n^{-1/2}$ helps us approximate the power of these tests at fixed alternatives close to the null.

Consider a p–dimensional regular parametric model from which we have i.i.d. observations $X_1, X_2, \ldots, X_n$. Suppose that we wish to test the null hypothesis $H_0 : \theta = \theta_0$. Consider a sequence of local alternatives of the form $\theta_0 + h/\sqrt{n}$; under this sequence, the likelihood ratio statistic converges in distribution to $\chi^2_p(h^T\, I(\theta_0)\, h)$. Thus, for large n, the probability that the likelihood ratio test rejects H_0 at level α is approximately the probability that a $\chi^2_p(h^T\, I(\theta_0)\, h)$ random variable is greater than $q_{1-\alpha;p}$ where $q_{1-\alpha;p}$ is the $1-\alpha$'th quantile of the χ^2_p distribution.

Now, let θ_1 be close to $\theta_0 \in \mathcal{H}_0$ and suppose we have n observations. Then we can write $\theta_1 = \theta_0 + h/\sqrt{n}$, where $h = \sqrt{n}\,(\theta_1 - \theta_0)$. Note here that h does indeed depend on n, but pretending that this is a fixed quantity, we can get an approximation to the power of the likelihood ratio test provided n is big and h is not too large. The approximate power β under the alternative θ_1 is,

$$\beta(\theta_1) = \text{Prob}\left(\chi^2_p(n\,(\theta_1 - \theta_0)^T\, I(\theta_0)\,(\theta_1 - \theta_0)) \geq q_{1-\alpha,p}\right)\,.$$

We can use the above equation to find the sample size needed to achieve a desired power (which is stipulated before). Similar approximations to the power at fixed alternatives can be made when we test a sub-parameter, rather than the full parameter. We will elaborate on these issues in a more general context. We now look at a simple example.

Example 1: Normal distributions. Let $X_1, X_2, \ldots, X_n$ be a sample from $N(\mu, \sigma^2)$. Consider testing $H_0 : \mu = 0, \sigma = 1$. Using standard asymptotic theory for regular parametric models, we conclude that the likelihood ratio statistic for this problem converges in distribution to χ^2_2. The limit distribution under a more general hypothesis of the form $H_0 : \mu = \mu_0, \sigma = \sigma_0$ is also χ^2_2. It is an instructive exercise to derive this from first principles without using the general asymptotic machinery developed.

We will derive the limit distribution of the likelihood ratio statistic for testing $H_0 : \mu = 0$ from first principles.(By standard asymptotic theory, we know that this has to be χ^2_1.)

The joint likelihood of the observations $X_1, X_2, \ldots, X_n$ is,

$$L(X_1, \ldots, X_n, \mu, \sigma^2) = \frac{1}{(\sqrt{2\,\pi}\,\sigma)^n} \exp\left(-\frac{1}{2\,\sigma^2}\,(X_i - \mu_0)^2\right).$$

The unrestricted MLEs of (μ, σ^2) are $(\overline{X}_n, \hat{\sigma}^2)$ where $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (X_i - \overline{X})^2 \equiv n^{-1}\,S^2$. The MLEs of (μ, σ^2) under $H_0 : \mu = 0$ are $(0, n^{-1} \sum_{i=1}^n X_i^2) \equiv (0, \hat{\sigma}_0^2$. Direct computation now yields,

$$\begin{aligned}
2\,\log\,\lambda_n &= 2\,\log\left(\frac{1}{(\sqrt{2\,\pi}\,\hat{\sigma})^n} \exp\,(-n/2) \div \frac{1}{(\sqrt{2\,\pi}\,\hat{\sigma}_0)^n} \exp\,(-n/2)\right) \\
&= 2\,\log\,\frac{\hat{\sigma}_0^n}{\hat{\sigma}^n} \\
&= n\,\log\,\hat{\sigma}_0^2 - n\,\log\,\hat{\sigma}^2 \\
&= n\,\log\,\hat{\sigma}^2 + \frac{n}{\hat{\sigma}_0^2}\,\left(\hat{\sigma}_0^2 - \hat{\sigma}^2\right) - \frac{n}{2\,\tilde{\sigma}_0^4}\,\left(\hat{\sigma}_0^2 - \hat{\sigma}^2\right)^2 - n\,\log\,\hat{\sigma}^2\,,
\end{aligned}$$

where $\tilde{\sigma_0}^2$ is a (random) point that lies between $\hat{\sigma}_0^2$ and $\hat{\sigma}^2$. Now,

$$n\,(\hat{\sigma_0}^2 - \hat{\sigma}^2) = \sum_{i=1}^n X_i^2 - \sum_{i=1}^n (X_i - \overline{X})^2 = n\,\overline{X}^2\,.$$

Thus the likelihood ratio statistic reduces to,

$$2\,\log\,\lambda_n = \frac{n\,\overline{X}^2}{\hat{\sigma}_0^2} - \frac{1}{2\,\tilde{\sigma_0}^4\,n}\,\left(n\,\overline{X}^2\right)^2\,.$$

Now, $\hat{\sigma}_0^2$ and $\tilde{\sigma_0}^2$ are both consistent for σ^2 under H_0; furthermore, under H_0,

$$\frac{\sqrt{n}\,\overline{X}}{\sigma} \sim N(0,1)$$

so it follows that,

$$\frac{n\,\overline{X}^2}{\hat{\sigma}_0^2} = \frac{n\,\overline{X}^2}{\sigma^2}\,\frac{\sigma^2}{\hat{\sigma}_0^2} \to \chi_1^2\,.$$

Consequently, the quantity,

$$\frac{1}{2\,\tilde{\sigma_0}^4\,n}\,\left(n\,\overline{X}^2\right)^2 = \frac{1}{n}\,O_p(1) = o_p(1)\,.$$

It follows that 2 log λ_n is asymptotically distributed as χ_1^2. This finishes the derivation.

We will now consider the behavior of the likelihood ratio statistic under local alternatives of the form (μ_n, σ_n) where $\mu_n = \mu_0 + h_1/\sqrt{n}$ and $\sigma_n = \sigma_0 + h_2/\sqrt{n}$. Under local alternatives of this type, the likelihood ratio statistic for testing $H_0 : \mu = \mu_0, \sigma = \sigma_0$ converges in distribution

to $\chi^2_2(h^T\, I(\mu_0, \sigma_0)\, h)$. Here $h = (h_1, h_2)^T$ and $I(\mu_0, \sigma_0)$ is the information matrix based on one observation, say X_1 at parameter values (μ_0, σ_0). (Note that we use the mean and the standard deviation as parameters, rather than the mean and the variance.) Check that the information matrix $I(\mu_0, \sigma_0)$ based on one observation from a normal density with mean μ_0 and variance σ_0^2 is,

$$I(\mu_0, \sigma_0) = \begin{pmatrix} \frac{1}{\sigma_0^2} & 0 \\ 0 & \frac{2}{\sigma_0^2} \end{pmatrix}.$$

This shows that the information bound for estimating the mean does not change in the presence of knowledge of the variance. This makes sense for the normal distribution for which the MLE's of the mean and the variance are independent of one another and the variance has no functional relationship to the mean. The non-centrality parameter can now be explicitly computed as $\Delta = h_1/\sigma_0^2 + 2\, h_2^2/\sigma_0^2$. More concretely, let $\mu_0 = 0$ and $\sigma_0 = 1$. We seek to compute an approximation to the power of the likelihood ratio test at the point $\theta_1 = (\mu_1, \sigma_1) = (0.2, 1.2)$ when we have a sample of size 100. Now $(.2, 1.2) = (0, 1) + \sqrt{n}\,(.2, .2)/\sqrt{n} \equiv (0, 1) + (h_1, h_2)/\sqrt{n}$ with $h_1 = \sqrt{n}(0.2) = h_2$. The approximate distribution of the likelihood ratio statistic under $(0.2, 1.2)$ based on $n = 100$ observations is therefore a non-central χ^2_2 distribution with non-centrality parameter$\Delta = (0.2 \times 10)^2 + 2\,(0.2 \times 10)^2 = 12$, based on which the approximate power can be computed. If $q_{2,.95}$ denotes the 95'th percentile of the distribution of the (central) χ^2_2 distribution, then $\beta_{n=100}\,(.2, 1.2) =_{appx} \text{Prob}(\chi^2_2(12) \geq q_{2,.95})$.

If we consider the likelihood ratio statistic for testing $\mu = 0$, this converges to χ^2_1 under the null hypothesis. To approximate the power of this test at $(.2, 1.2)$ based on 100 observations, once again write $(.2, 1.2) = (0 \equiv \mu_0, 1 \equiv \sigma_0) + (h_1, h_2)/10$, with $h_1 = h_2 = 2$ as in the previous paragraph. By the theory we have developed previously, the limit distribution at $(0.2, 1.2)$ is approximately $\chi^2_1(\Delta = h_1^2/\sigma_0^2 = 4)$ (recall that the non-centrality parameter is $h_1^T\, I_{11.2}\, h_1$, where $I_{11.2}$ is the efficient information for the estimation of μ; this is just $1/\sigma_0^2$ in the current case).

Example 2: Efficient estimation of a Poisson probability. Let $X_1, X_2, \ldots, X_n$ be i.i.d. Poi(θ). The common underlying density is,

$$f(x, \theta) = \frac{e^{-\theta}\, \theta^x}{x\,!}.$$

Thus,

$$l(x, \theta) = -\theta + x\, \log\, \theta - \log\, x\,!.$$

Also, the joint likelihood is,

$$L(\theta, X_1, \ldots, X_n) = \frac{e^{-n\,\theta}\, \theta^{\sum_{i=1}^n X_i}}{\Pi_{i=1}^n X_i\,!}.$$

Clearly $\sum_{i=1}^n X_i$ is complete sufficient for θ. We have,

$$\dot{l}(x, \theta) = \frac{x}{\theta} - 1\,,$$

and

$$\ddot{l}(x,\theta) = -\frac{x}{\theta^2}\,.$$

Hence,

$$I(\theta) = -E_\theta\left(\ddot{l}(X_1,\theta)\right) = \frac{1}{\theta}\,.$$

Check that in this model $\hat{\theta}_n = \overline{X}_n$. By standard asymptotic theory,

$$\sqrt{n}\,(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I(\theta)^{-1}\,\dot{l}(X_i,\theta) + o_p(1)\,.$$

In this model, this is actually a trivial identity, since

$$I(\theta)^{-1}\,\dot{l}(X_i,\theta) = X_i - \theta\,.$$

Now consider, $q(\theta) = e^{-\theta} = P_\theta(X_1 = 0)$. Then, the information bound for estimating $q(\theta)$ is,

$$IB_q(\theta) = \frac{q^{'}(\theta)^2}{I(\theta)} = e^{-2\,\theta}\,\theta\,.$$

Also,

$$\sqrt{n}\,(e^{-\overline{X}_n} - e^{-\theta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(-e^{-\theta}\,(X_i - \theta)\right) + o_p(1) \to N(0, \theta\,e^{-2\,\theta})\,.$$

Here,

$$\tilde{l}(x,\theta,q) = -e^{-\theta}\,(x-\theta)\,,$$

is the efficient influence function for estimating $q(\theta)$. Any estimator T_n for which,

$$\sqrt{n}\,(T_n - q(\theta)) \to N(0, IB_q(\theta))$$

is said to be asymptotically efficient for estimating q. We know that $q(\hat{\theta})$, the MLE is indeed asymptotically efficient. What about the UMVUE of $q(\theta)$? First let us compute the UMVUE of $q(\theta)$. Note that $1\,(X_1 = 0)$ is an unbiased estimate of $q(\theta)$. The UMVUE is then given by $E(1\,(X_1 = 0) \mid \sum_{i=1}^n X_i) = P\,(X_1 = 0 \mid \sum_{i=1}^n X_i)$. Now, given $\sum_{i=1}^n X_i$, the vector $(X_1, X_2, \ldots, X_n)$ follows a Multinomial$(\sum_{i=1}^n X_i, n^{-1}, n^{-1} \ldots, n^{-1})$. Thus, conditional on $\sum_{i=1}^n X_i$, X_1 follows $Bin(\sum_{i=1}^n X_i, n^{-1})$. Thus,

$$P(X_1 = 0 \mid \sum_{i=1}^{n} X_i) = \left(1 - \frac{1}{n}\right)^{\sum_{i=1}^n X_i} = T_n^\star\,.$$

Now, for each n,

$$\mathrm{Var}_\theta\,\sqrt{n}\,(T_n^\star - q(\theta)) > IB_q(\theta)\,,$$

since $T_n^\star$ is not linear in the score function. What can we say about the asymptotic variance of $T_n^\star$? Now,

$$\sqrt{n}\,(T_n^\star - q(\theta)) = \sqrt{n}\,(T_n^\star - q(\hat{\theta})) + \sqrt{n}\,(\hat{q}(\theta) - q(\theta)) = \sqrt{n}\,(T_n^\star - q(\hat{\theta})) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(-e^{-\theta}\,(X_i - \theta)\right) + o_p(1)\,.$$

But,

$$\begin{aligned} \sqrt{n}\,(T_n^\star - q(\hat{\theta})) &= \sqrt{n}\left(\left(1 - \frac{1}{n}\right)^{n\,\overline{X}} - e^{-\overline{X}}\right) \\ &= \xi_n^{\overline{X}} - e^{-1\,\overline{X}}\,, \end{aligned}$$

where $\xi_n = (1 - n^{-1})^n$. But then,

$$\sqrt{n}\,(T_n^\star - q(\hat{\theta})) = \overline{X}\,\tilde{\xi}_n^{\overline{X}-1}\,\sqrt{n}\,(\xi_n - e^{-1}) \to_p \theta\,\left(e^{-1}\right)^{(\theta-1)}\,0 = 0\,,$$

where $\tilde{\xi}_n$ lies between ξ_n and e^{-1} and therefore converges to e^{-1}, $\overline{X}$ converges in probability to θ and $\sqrt{n}(\xi_n - e^{-1})$ converges to 0. It follows that,

$$\sqrt{n}\,(T_n^\star - q(\theta)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(-e^{-\theta}\,(X_i - \theta)\right) + o_p(1)\,;$$

thus the UMVUE is asymptotically linear in the efficient influence function and hence, attains the information bound asymptotically.

Examples 3: Testing a general null hypothesis $g(\theta) = c$. Consider a p–dimensional regular parametric submodel $\{P_\theta : \theta \in \Theta\}$ with θ varying in Θ, an open subset of $\mathbb{R}^p$. The density of P_θ with respect to an appropriate underlying dominating measure is written as $f(x, \theta)$. Consider testing a null hypothesis H_0 of the form $g(\theta) = c$ where $g(\theta) = (g_1(\theta), g_2(\theta), \ldots, g_k(\theta))$ is a continuously differentiable transformation from Θ to $\mathbb{R}^k$ and $c = (c_1, c_2, \ldots, c_k)$ with $\bigtriangledown\, g(\theta)_{k\times p}$, the gradient matrix having full row rank. Now let θ_0 be a point in the null hypothesis; so $g(\theta_0) = c$. We will show that the likelihood ratio statistic for testing H_0 is asymptotically distributed as χ^2_k.

Since $\bigtriangledown\, g(\theta_0)$ is of full row rank k, we can find $p - k$ (column) vectors, $v_{k+1}, v_{k+2}, \ldots, v_p$ (in $\mathbb{R}^p$) such that the k rows of $\bigtriangledown\, g(\theta_0)$ along with these $p - k$ vectors form a basis of $\mathbb{R}^p$. Hence, the matrix

$$M = \left(\bigtriangledown\, g(\theta_0)^T, v_{k+1}, v_{k+2}, \ldots, v_p)^T\right)$$

is invertible. Now, define transformations $g_{k+1}, g_{k+2}, \ldots, g_p$ on Θ where $g_{k+i}(\theta) = v_{k+i}^T\,(\theta - \theta_0)$, for $i = 1, 2, \ldots, p - k$ and consider the transformation frm Θ to $\mathbb{R}^p$ given by,

$$g_{ext}(\theta) = (g_1(\theta), g_2(\theta), \ldots, g_p(\theta))\,.$$

Now, g_{ext} is a continuously differentiable transformation and $\bigtriangledown\, g_{ext}(\theta_0) = M$ is invertible by construction. Consequently, by the inverse function theorem, there exists an open ball $B_\epsilon(\theta_0)$

around θ_0 and an open ball $B_\delta(\xi_0)$ around $\xi_0 = g_{ext}(\theta_0)$ such that g_{ext} is a continously differentiable bijection from $B_\epsilon(\theta_0)$ to $B_\delta(\xi_0)$ with inverse function h being a continuously differentiable bijection from $B_\delta(\xi_0)$ to $B_\epsilon(\theta_0)$. Now, $\{Q_\xi \equiv P_{h(\xi)} : \xi \in B_\delta(\xi_0)\}$ is a regular parametric model with density $\tilde{f}(x,\xi) = f(x,h(\xi))$. Consider, testing the null hypothesis $\tilde{H}_0 : \xi_j = c_j, j = 1,2,\ldots,k$. This is equivalent to the null hypothesis $H_{0,rest} : \theta \in B_\epsilon(\theta_0)\ ,\ g(\theta) = c$. Also, note that

$$\lambda_{n,rest} = \frac{\sup_{\theta \in B_\epsilon(\theta_0)} \Pi_{i=1}^n f(X_i,\theta)}{\sup_{\theta \in B_\epsilon(\theta_0)\ ,\ g(\theta)=c} \Pi_{i=1}^n f(X_i,\theta)} = \frac{\sup_{\xi \in B_\delta(\xi_0)} \Pi_{i=1}^n f(X_i,h(\xi))}{\sup_{\xi \in B_\delta(\xi_0)\ ,\ (\xi_j = c_j, i=1,2,\ldots,k)} \Pi_{i=1}^n f(X_i,h(\xi))} \equiv \tilde{\lambda}_n\,.$$

From previous derivations in class, we know that under parameter value ξ_0 (which in the θ parametrization corresponds to parameter value θ_0), $2\log\tilde{\lambda}_n$ converges in distribution to a χ^2_k random variable. Thus, $2\log\lambda_{n,rest}$ converges to a χ^2_k. Now the actual likelihood ratio statistic $2\log\lambda_n$ is given by,

$$2\log\lambda_n = 2\log\frac{\sup_{\theta\in\Theta}\Pi_{i=1}^n f(X_i,\theta)}{\sup_{\theta\in\Theta\,,\,g(\theta)=c}\Pi_{i=1}^n f(X_i,\theta)}\,.$$

But by consistency of the MLE's $\hat{\theta}_n$ (the unrestricted MLE) and $\hat{\theta}_n^0$ (the MLE under H_0) for the true parameter, it follows that when θ_0 is the true value, with probability increasing to 1, $\hat{\theta}_n$ and $\hat{\theta}_n^0$ both lie in $B_\epsilon(\theta_0)$ eventually. Hence,

$$2\log\lambda_n - 2\log\lambda_{n,rest} \to_p 0\,.$$

Hence $2\log\lambda_n \to_d \chi^2_k$.

How do we find the Wald statistic for testing $g(\theta) = c$? Denote the first k components of ξ by $\xi_{(1)}$ and let $I_\xi(\xi_0)$ denote the information matrix at the point ξ_0. Now, the Wald statistic for testing $H_0 : \xi_j = c_j, j = 1,2,\ldots,k$ (which is at least locally the same as testing $g(\theta) = c$) is simply given by

$$n\,\left(\hat{\xi}_{(1)} - c\right)^T\,\widehat{I_{11.2,\xi}}\,\left(\hat{\xi}_{(1)} - c\right)$$

where $\widehat{I_{11.2,\xi}}$ is an estimate of $I_{11.2,\xi} = I_{11,\xi} - I_{12,\xi}\,I_{22,\xi}^{-1}\,I_{21,\xi}$. Here $I_{11,\xi}$ is the dispersion matrix of the score for $\xi_{(1)}$ under parameter value ξ_0, $I_{12,\xi}$ is the covariance matrix between the score for $\xi_{(1)}$ and the score for $\xi_{(2)}$ under ξ_0 and so on. We need to translate the statistic in the above display in terms of the θ's. Note that $\hat{\xi}_{(1)} = g(\hat{\theta})$. Also, $I_{11.2,\xi}^{-1}$ is the information bound for estimating $\xi_{(1)} = g(\theta)$ at ξ_0. The information bound for estimating $g(\theta)$ at θ_0 is given by $\bigtriangledown g(\theta_0)\,I(\theta_0)^{-1}\,\bigtriangledown\,g(\theta_0)^T$. By the invariance of the information bound under reparametrization,

$$I_{11.2,\xi}^{-1} = \bigtriangledown g(\theta_0)\,I(\theta_0)^{-1}\,\bigtriangledown\,g(\theta_0)^T$$

so that

$$I_{11.2,\xi} = \left(\bigtriangledown g(\theta_0)\,I(\theta_0)^{-1}\,\bigtriangledown\,g(\theta_0)^T\right)^{-1}\,.$$

So,

$$\widehat{I_{11.2,\xi}} = \left(\bigtriangledown g(\hat{\theta}_n)\,I(\hat{\theta}_n)^{-1}\,\bigtriangledown\,g(\hat{\theta}_n)^T\right)^{-1}\,.$$

Thus, the Wald statistic, W_n for testing H_0 is simply,

$$W_n = n \left(g(\hat{\theta}_n) - c\right)^T \left(\bigtriangledown g(\hat{\theta}_n)\, I(\hat{\theta}_n)^{-1} \bigtriangledown g(\hat{\theta}_n)^T\right)^{-1} \left(g(\hat{\theta}_n) - c\right) .$$