1 Likelihood ratio tests

For two competing hypotheses $H_0$ and $H_1$ about the parameter $\theta$, the likelihood ratio is often used to make a comparison. For example, for $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, the likelihood ratio is $L(\theta_0)/L(\theta_1)$, and large (resp. small) values of this ratio indicate that the data $x$ favors $H_0$ (resp. $H_1$). A more difficult and somewhat more general problem is $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$, where $\Theta_0$ is a subset of $\Theta$. In this case, one can define the likelihood ratio as

$$T_n = T_n(X, \Theta_0) = \sup_{\Theta \in \Theta_0} L(\theta) \sup_{\Theta \notin \Theta_0} L(\theta).$$

The interpretation of this likelihood ratio is the same as before, i.e., if the ratio is small, then data lends little evidence to the null hypothesis. For practical purposes, we need to know what it means for the ratio to be “small”; this means we need the null distribution of $T_n$, i.e., the distribution of $T_n$ under $P_\theta$, when $\theta \in \Theta_0$. For $\Theta \subset \mathbb{R}^d$, consider the testing problem $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$, where $\Theta_0$ is a subset of $\Theta$ that specifies the values $\theta_0, \ldots, \theta_m$ of $\theta$, where $m$ is a fixed integer between 1 and $d$. The following result, known as Wilks’s Theorem, gives conditions under which the null distribution of $W_n = -2 \log T_n$ is asymptotically of a convenient form.

**Theorem 1.** Suppose the conditions of consistency of MLE hold. Under the setup described in the previous paragraph, $W_n \rightarrow \text{ChiSq}(m)$ in distribution, under $P_\theta$ with $\theta \in \Theta_0$.

**Proof.** We focus here only on the case $d = m = 1$. That is, $\Theta_0 = \{\theta_0\}$ is a singleton, and we want to know the limiting distribution of $W_n$ under $P_{\theta_0}$. Clearly,

$$W_n = -2\ell_n(\theta_0) + 2\ell_n(\hat{\theta}_n),$$

where $\hat{\theta}_n$ is the MLE and $\ell_n$ is the log-likelihood. By the assumed continuity of the log-likelihood, do a two-term Taylor approximation of $\ell_n(\theta_0)$ around $\hat{\theta}_n$:

$$\ell_n(\theta_0) = \ell_n(\hat{\theta}_n) + \ell_n'(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n) + \frac{\ell_n''(\hat{\theta}_n)}{2}(\theta_0 - \hat{\theta}_n)^2,$$

where $\hat{\theta}_n$ is between $\theta_0$ and $\hat{\theta}_n$. Since $\ell_n'(\hat{\theta}_n) = 0$, we get

$$W_n = -\ell_n''(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n)^2 = -\frac{\ell_n''(\hat{\theta}_n)}{n}\{n^{1/2}(\hat{\theta} - \theta_0)\}^2.$$
From the conditions on consistency of MLE, we have that \( \sqrt{n}(\hat{\theta}_n - \theta_0) \to N(0, I(\theta_0)^{-1}) \) in distribution, as \( n \to \infty \). Also, in the proof of that theorem, we showed that \( n^{-1} \ell''(\theta_0) \to -I(\theta_0) \) under \( P_{\theta_0} \) for any consistent \( \hat{\theta}_n \). Indeed, we can write

\[
\ell''(\hat{\theta}) = \ell''(\theta_0) + \ell''(\hat{\theta}_n) - \ell''(\theta_0)
\]

and we have that

\[
|\ell''(\hat{\theta}_n) - \ell''(\theta_0)| \leq \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\partial^2}{\partial \theta^2} \log p\theta(X_i)\big|_{\theta = \hat{\theta}_n} - \frac{\partial^2}{\partial \theta^2} \log p\theta(X_i)\big|_{\theta = \theta_0} \right|
\]

Using Condition C2, the upper bound is bounded by \( n^{-1} \sum_{i=1}^{n} M(X_i)|\hat{\theta}_n - \theta_0| \), which goes to zero in probability under \( P_{\theta_0} \) since \( \hat{\theta}_n \) is consistent. Therefore, \( \ell''(\hat{\theta}_n) \) has the same limiting behavior as \( \ell''(\theta_0) \). Finally, by Slutsky, we get

\[
W_n \to I(\theta_0)N(0, I(\theta_0)^{-1})^2 \equiv N(0, 1)^2 \equiv \text{ChiSq}(1).
\]

Wilks theorem facilitates construction of an approximate size-\( \alpha \) test of \( H_0 \) when \( n \) is large, i.e., by rejecting \( H_0 \) iff \( W_n \) is more than \( \chi^2_{m,1-\alpha} \), the 100(1 - \( \alpha \)) percentile of the ChiSq(\( m \)) distribution. The advantage of Wilks theorem appears in cases where the exact sampling distribution of \( W_n \) is intractable, so that an exact (analytical) size-\( \alpha \) test is not available. Monte Carlo can often be used to find a test, but Wilks theorem gives a good answer and only requires use of a simple chi-square table. One can also use the Wilks theorem result to obtain approximation confidence regions. Let \( W_n(\theta_0) = -2\log T_n(X; \theta_0) \), where \( \theta_0 \) is a fixed generic value of the full \( d \)-dimensional parameter \( \theta \), i.e., \( H_0 : \theta = \theta_0 \). Then an approximate 100(1 - \( \alpha \))% confidence region for \( \theta \) is \( \{\theta_0 : W_n(\theta_0) \leq \chi^2_{m,1-\alpha}\} \). An interesting and often overlooked aspect of Wilks theorem is that the asymptotic null distribution does not depend on the true values of those parameters unspecified under the null. For example, in a gamma distribution problem with the goal of testing if the shape is equal to some specified value, the null distribution of \( W_n \) does not depend on the true value of the scale.

### 1.1 Cautions concerning the first-order theory

One might be tempted to conclude that the desirable properties of the likelihood-based methods presented in the previous section are universal, i.e., that maximum likelihood estimators will “always work.” Moreover, based on the form of the asymptotic variance of the MLE and its similarity to the Cramer-Rao lower bound, it is tempting to conclude that the MLE is asymptotically efficient. However, both of these conclusions are technically false in general. Indeed, there are examples where
1. the MLE is not unique or even does not exist

2. the MLE “works” (in the sense of consistency), but the conditions of the theory are not met so asymptotic normality fails; and

3. the MLE is not even consistent!

Non-uniqueness or non-existence of the MLE are roadblocks to practical implementation but, for some reason, aren’t viewed as much of a concern from a theoretical point of view. The case where the MLE works but is not asymptotically normal is also not really a problem, provided that one recognizes the non-regular nature of the problem and makes the necessary adjustments. The most concerning of these points is inconsistency of the MLE. Since consistency is a rather weak property, inconsistency of the MLE means that its performance is poor and can give very misleading results. The most famous example of inconsistency of the MLE, due to Neyman and Scott (1948), is given next.

Example: (Neyman and Scott 1948). Let \( X_{ij} \) be independent normal random variables, \( X_{ij} \sim N(\mu_i, \sigma^2) \), \( i = 1, \ldots, n \) and \( j = 1, 2 \); the case of two \( j \) levels is the simplest, but the result holds for any fixed number of levels. The point here is that \( X_{i1} \) and \( X_{i2} \) have the same mean \( \mu_i \), but there are possibly \( n \) different means. The full parameter is \( \theta = (\mu_1, \ldots, \mu_n, \sigma^2) \), which is of dimension \( n + 1 \). It is easy to check that the MLEs are given by

\[
\hat{\mu}_i = \frac{1}{2}(X_{i1} + X_{i2}), \quad i = 1, \ldots, n
\]

\[
\hat{\sigma}^2 = \frac{1}{4n} \sum_{i=1}^{n} (X_{i1} - X_{i2})^2
\]

It is easy to see that as \( n \to \infty \), \( \hat{\sigma}^2 \to \sigma^2/2 \) in probability, so that the MLE of \( \sigma^2 \) is inconsistent!

The issue here that is causing inconsistency is that the dimension of the nuisance parameter, the means \( \mu_1, \ldots, \mu_n \), is increasing with \( n \). In general, when the dimension of the parameter depends on \( n \), consistency of the MLE will be a concern so one should be careful.

2 Wald and score tests

There are common large sample alternatives to likelihood ratio testing. These are the “Wald tests” and the (Rao) “score tests.” Consider \( \Theta \subset \mathbb{R}^d \) and \( m \) genuinely different or independent restrictions on \( \theta \) given by

\[
g_1(\theta) = g_2(\theta) = \cdots g_m(\theta) = 0,
\]
for instance $\theta_1 - \theta_1^0 = \theta_2 - \theta_2^0 = \cdots = \theta_m - \theta_m^0 = 0$. Define

$$g(\theta) = (g_1(\theta), g_2(\theta), \ldots, g_m(\theta))' = (0, \ldots, 0)'$$

that is $H_0 : g(\theta) = 0$. The obvious likelihood ratio statistic for this hypothesis is

$$\lambda_n = \frac{\sup_{\theta} f(X \mid \theta)}{\sup_{\theta: g(\theta) = 0} f(X \mid \theta)^{1/2}}$$

and standard theory suggests (in iid cases) that $2 \log \lambda_n \to \chi_m^2$. Suppose that $\hat{\theta}_n$ is an MLE of $\theta$. Then if $H_0 : g(\theta) = 0$ is true, $g(\hat{\theta}_n)$ ought to be near 0, and one can think about rejecting $H_0$ if it is not. The questions are how to measure "nearness? and how to set a critical value in order to have a test with size approximately $\alpha$. The Wald approach to doing this is as follows. We expect (under suitable conditions) that under $P_\theta$ the $(d$-dimensional) estimator $\hat{\theta}_n$ has

$$\sqrt{n}(\hat{\theta}_n - \theta) \to N_d(0, I(\theta)^{-1})$$

Then, if

$$G(\theta)_{m \times d} = \left( \frac{\partial g_i(\theta)}{\partial \theta_j} \right)$$

the delta method suggests that

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \to N_m(0, G(\theta)I(\theta)^{-1}G(\theta)')$$

Abbreviate $G(\theta)I(\theta)^{-1}G(\theta)'$ as $B(\theta)$. Since $g(\theta) = 0$ if $H_0$ is true, the above suggests that under $H_0$

$$ng(\hat{\theta}_n)'B(\theta)^{-1}g(\hat{\theta}_n) \to \chi_m^2.$$ 

Now $ng(\hat{\theta}_n)'B(\theta)^{-1}g(\hat{\theta}_n)$ can not serve as a test statistic, since it involves $\theta$, which is not completely specified by $H_0$. But it is plausible to consider the statistic

$$W_n = ng(\hat{\theta}_n)'B(\theta)^{-1}g(\hat{\theta}_n)$$

and hope that under suitable conditions $W_n \to \chi_m^2$ under $H_0$. If this can be shown to hold up, one may reject for $W_n > (1 - \alpha)$ quantile of the $\chi_m^2$ distribution. This is the use of “expected Fisher information” in defining a Wald statistic. With $H_n(\theta)$ the matrix of second partials of the log-likelihood evaluated at $\theta$, an “observed Fisher information” version of the above is to let

$$B_n^*(\theta) = G(\theta)\left(-\frac{1}{n}H_n(\theta)\right)^{-1}G(\theta)'.$$
where $H_n$ is the Hessian matrix of $\ell(\theta)$ and use the test statistic

$$W_n^* = n g(\hat{\theta}_n)' B_n^*(\hat{\theta}_n)^{-1} g(\hat{\theta}_n)$$

The “(Rao) score test” or “$\chi^2$” test is an alternative to the LR and Wald tests. The motivation for it is that on occasion it can be easier to maximize the log likelihood subject to $g(\theta) = 0$ than to simply maximize it without constraints. Let $\hat{\theta}_n$ be a “restricted” MLE (i.e. a maximizer of $\ell(\theta)$ subject to $g(\theta) = 0$). One might expect that if $H_0 : g(\theta) = 0$ is true, then $\hat{\theta}_n$ ought to be nearly an unrestricted maximizer of $\ell(\theta)$ and the partial derivatives of $\ell(\theta)$ should be nearly 0 at $\hat{\theta}_n$. On the other hand, if $H_0$ is not true, there is little reason to expect the partials to be nearly 0.

So again with $\ell(\theta)$, one might consider the statistic

$$R_n = \frac{1}{n} \left( \frac{\partial}{\partial \theta_i} \ell(\theta) \bigg|_{\theta = \hat{\theta}_n} \right)' I(\hat{\theta}_n)^{-1} \left( \frac{\partial}{\partial \theta_i} \ell(\theta) \bigg|_{\theta = \hat{\theta}_n} \right)$$

or using the observed Fisher information $-H_n(\hat{\theta}_n)$ in place of $I(\hat{\theta}_n)$.

What is not obvious is that $R_n$ can typically be shown to differ from $2 \log \lambda_n$ by a quantity tending to 0 in $P_\theta$ probability. That is, this statistic can be calibrated using $\chi^2$ critical points and can form the basis of a test of $H_0$ asymptotically equivalent to the likelihood ratio test. In fact, all of the test statistics mentioned here are asymptotically equivalent to the LRT (Wald and Rao score alike). Observe that

1. an LRT requires computation of an MLE, $\hat{\theta}_n$, and a restricted MLE, $\tilde{\theta}_n$,
2. a Wald test requires only computation of the MLE, $\hat{\theta}_n$, and
3. a Rao score test requires only computation of the restricted MLE, $\tilde{\theta}_n$. 
