Date: 11/11/2011

File Name: D:\A File\Research\Likelihood 11.10.20\Likelihood_5.tex

Parametric Likelihood Inference

Xuan Yao

Maximum likelihood principle is one of the milestones in statistical literature in the past century. Here we give a brief review of the parametric likelihood inference. Throughout, we consider the following random sample from a known *p.d.f.* with unknown parameter θ:

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x;\theta) \tag{1}$$

with the actual observations (realizations)

$$x_1, \ldots, x_n. \tag{2}$$

1 Likelihood Function

Likelihood is the probability of observing the data we observed. Thus, for random sample (1) and (2) the likelihood is given by

$$P\{X_1 = x_1, \ldots, X_n = x_n\} = P\{X_1 = x_1\}P\{X_2 = x_2\}\cdots P\{X_n = x_n\} = \prod_{i=1}^{n} P\{X_i = x_i\}. \tag{3}$$

To derive the definition of likelihood function as follows, we discuss (3) for discrete and continuous *p.d.f.*, respectively.

Case 1 If $X_1, \ldots, X_n$ in (1)is a discrete, then we have $P(X = x) = f(x;\theta)$; in turn, equation (3) becomes

$$P\{X_1 = x_1, \ldots, X_n = x_n\} = \prod_{i=1}^{n} f(x_i;\theta). \tag{4}$$

Case 2 If $X_1, \ldots, X_n$ in (1)is a conitnuous, then we have $P(X = x) \approx P(x - \delta < x < x + \delta)$; in turn, equation (3) becomes

$$P\{X_1 = x_1, \ldots, X_n = x_n\} \approx \prod_{i=1}^{n} P(x_i - \delta < x_i < x_i + \delta) = \prod_{i=1}^{n} (f(x_i;\theta) \cdot 2\delta) = (2\delta)^n \cdot \prod_{i=1}^{n} f(x_i;\theta). \tag{5}$$

The last equation of (5) shows that (3) is proportional to $\prod_{i=1}^{n} f(x_i;\theta)$.

Let us define $L(\theta;\mathbf{x}) = \prod_{i=1}^{n} f(x_i;\theta)$. We have shown that if we have a discrete sample, $L(\theta;\mathbf{x})$ is exactly the likelihood. If we have a continuous sample, $L(\theta;\mathbf{x})$ is proportional to the likelihood. Since later we want to maximize the likelihood with respect to the parameter θ (the reason of doing it will be discussed in section two), it is equivalent to maximize the $L(\theta;\mathbf{x})$ with respect to θ.

2 Maxium Likelihood Estimator

In statistics, maximum-likelihood estimator (MLE) is a method of estimating the parameters of a statistical model. When applied to a data set and given a statistical model, maximum-likelihood estimator provides estimates for the model's parameters.

In general, the method of maximum likelihood selects values of the model parameters that produce a distribution that gives the observed data the greatest probability (i.e., parameters that maximize the likelihood function defined in Section 1). Therefore we define the Maximum Likelihood Estimate (MLE) of parameter θ_o as

$$\hat{\theta} = \arg\max_{\theta\in\Theta} L(\theta; \mathbf{x}). \tag{6}$$

For many applications involving likelihood functions, it is more convenient to work in terms of natural logarithm of the likelihood function, called log-likelihood, than in terms of the likelihood function itself. Because the logarithm is a monotonically increasing function, the logarithm of a function achieves its maximum value at the same points as the function itself, and hence the log-likelihood can be used in place of the likelihood in maximum likelihood estimator and related techniques and we can write the MLE as

$$\hat{\theta} = \arg\max_{\theta\in\Theta} l(\theta; \mathbf{x}) = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} l(\theta; x_i), \tag{7}$$

where $l(\theta; x_i) = \ln L(\theta; x_i)$; $l(\theta; \mathbf{x}) = \ln L(\theta; \mathbf{x})$.
If Θ is open, $l(\theta; \mathbf{x})$ is differentiable in θ and $\hat{\theta}$ exists then $\hat{\theta}$ must satisfy the estimating equation

$$\nabla_\theta l(\hat{\theta}; \mathbf{x}) = 0. \tag{8}$$

This is known as the **likelihood equation**. If the X_i are independent with densities $f_i(x, \theta)$ the likelihood equation simplifies to

$$\sum_{i=1}^{n} \nabla_\theta \ln f_i(x_i, \hat{\theta}) = 0, \tag{9}$$

which again enables us to analyse the behaviour of $\hat{\theta}$ using known properties of sums of independent random variables. Evidently, there may be solutions of (9) that are not maxima or only local maxima, thus we need to refer to other properties of the likelihood function.

Example 2.1. *Suppose X_i, $i = 1, \ldots, n$ is a i.i.d. sample from normal distribution with p.d.f. $f(x, \mu) = \exp\{-(x-\mu)^2/2\}/\sqrt{2\pi}$. Find the MLE of μ.*

Sol 2.1. *Since we have a continuous i.i.d. sample, by Definition (??),*

$$\begin{aligned} l(\mu; \boldsymbol{x}) &= \ln L(\mu; \boldsymbol{x}) = \sum_{i=1}^{n} \ln f(x_i, \mu) \\ &= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2. \end{aligned} \tag{10}$$

To maximize the log-likelihood, we differentiate (10) w.r.t. μ and set it to be zero,

$$\frac{\partial l(\mu; \boldsymbol{x})}{\partial \mu} = \sum_{i=1}^{n}(x_i - \mu) = 0. \tag{11}$$

and get $\hat{\mu} = \sum_{i=1}^{n} x_i/n = \bar{x}$. Now, let us take the second derivative of (10),

$$\frac{\partial^2 l(\mu; \boldsymbol{x})}{\partial \mu^2} = -n < 0. \tag{12}$$

By (12), we conclude that the first derivative of likelihood is a decreasing function. Since it attains 0 *if and only if* $\mu = \bar{\boldsymbol{x}}$*, the first derivative will be positive on* $(-\infty, 0)$ *and negative on* $)(0, \infty)$*. This suffice to show that the likelihood function will increase on* $(-\infty, 0)$ *whereas decrease on* $(0, \infty)$*, which means the likelihood function get its maximum at* $\mu = \bar{\boldsymbol{x}}$*. Hence by definition of MLE, the maximum likelihood estimator of* μ *is* $\hat{\mu} = \bar{\boldsymbol{x}}$

In some cases, the differentiating method is not applicable. This often happens when the domain of random variable is depend on parameter.

Example 2.2. *Suppose* X_i*,* $i = 1, 2, \ldots, n$ *is a i.i.d. sample from uniform distribution on* $(0, \theta)$*. Find the MLE of* θ*.*

Sol 2.2. *Since* $f(x, \theta) = \frac{1}{\theta} I(0 \leq x \leq \theta)$*, we can write the log-likelihood function of the sample as*

$$l(\theta; \boldsymbol{x}) = -n \ln \theta \cdot I(0 \leq x_i \leq \theta)_{i=1}^{n}. \tag{13}$$

However, since $-n \ln \theta$ *is a unbounded decreasing function, we cannot maximize it by setting the derivative to be 0. However, note that* $\theta \geq x_i$*,* $i = 1, 2, \ldots, n$*, we can maximize the likelihood function by picking up the smallest possible* θ*. Hence we get the MLE* $\hat{\theta} = X_{(n)}$

MLE holds a nice "Plug-in" property, which means that MLEs are unaffected by re-parametrization, i.e., MLEs are equivariant under one-to-one transformations.

Theorem 2.1. *Let* $\boldsymbol{X} \sim P_\theta$*,* $\theta \in \Theta$ *and let* $\hat{\theta}$ *denote the MLE of* θ*. Suppose that* h *is a one-to-one function from* Θ *onto* $h(\Theta)$*. Define* $\eta = h(\theta)$ *and let* $f(\boldsymbol{x}, \eta)$ *denote the density or frequency function of* $\boldsymbol{X}$ *in terms of* η *(i.e., reparametrize the model using* η*). Then the MLE of* η *is* $h(\hat{\theta})$*.*

Proof: Since h is onto and one-to-one, it is also invertible. Define $L^\star(\eta) = L(\theta)$ where $\theta = h^{-1}(\eta)$. So for any η, $L^\star(\hat{\eta}) = L(\hat{\theta}) \geq L(\theta) = L^\star(\eta)$ and hence $\hat{\eta} = h(\hat{\theta})$ maximizes $L^\star(\hat{\eta})$. □

3 Asymptotic Property of MLE

We expect a good estimator holds several nice properties. For example, we hope that the estimator will approximate the true value of parameter as sample size grows large. We are also interested in the distribution to calculate confidence intervals. Efficiency, another important value, will describe the accuracy of our statistic. In this section, we will discuss those properties of MLE.

1 Asymptotic Distribution

Theorem 3.1. X_i*,* $i = 1, 2, \ldots, n$ *is a i.i.d. sample from* $f(x, \theta)$*. Suppose* $\hat{\theta}_n$ *is the MLE of* θ *and* θ_0 *is the true value. Let* $I(\theta)$ *denote the Fisher Information in* X*. As* n *goes to infinity,* $\sqrt{n}(\hat{\theta}_n - \theta)$ *goes to* $N(0, I^{-1}(\theta_0))$ *in distribution.*

Proof: Let us make Taylor expansion of $\partial \ln f(\mathbf{x}, \hat{\theta})/\partial\theta$ at $\theta = \theta_0$,

$$\begin{aligned} 0 &= \frac{\partial \ln f(\mathbf{x}, \hat{\theta})}{\partial \theta} = \frac{\partial \ln f(\mathbf{x}, \theta_0)}{\partial \theta} + \frac{\partial^2 \ln f(\mathbf{x}, \theta_0)}{\partial \theta^2}(\hat{\theta} - \theta_0) + o(\|\hat{\theta} - \theta_0\|^2) \\ &= \sum_{i=1}^{n} \frac{\partial \ln f(x_i, \theta_0)}{\partial \theta} + \sum_{i=1}^{n} \frac{\partial^2 \ln f(x_i, \theta_0)}{\partial \theta^2}(\hat{\theta} - \theta_0) + o(\|\hat{\theta} - \theta_0\|^2). \end{aligned} \tag{14}$$

By multiplying $1/\sqrt{n}$ on both side of (14) and ignoring the higher order reminders, we obtain

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ln f(x_i, \theta_0)}{\partial \theta^2} \cdot \sqrt{n}(\hat{\theta} - \theta_0) = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \ln f(x_i, \theta_0)}{\partial \theta}. \tag{15}$$

By L.L.N.,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ln f(x_i,\theta_0)}{\partial\theta^2} \to E\left[\frac{\partial^2 \ln f(x,\theta_0)}{\partial\theta^2}\right] = I(\theta) \text{ in probability;} \tag{16}$$

By C.L.T.,

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ln f(x_i,\theta_0)}{\partial\theta} \to N\left(0, E\left(\frac{\partial \ln f(x,\theta_0)}{\partial\theta}\right)^2\right) = N(0, I(\theta_0)) \text{ in distribution.} \tag{17}$$

Plug (16) and (17) back to (14) and apply Slusky Theorem, we can get $\sqrt{n}(\hat{\theta}_n - \theta)$ goes to $N(0, I^{-1}(\theta_0))$ □

2 Consisitency

From (3.1) we can see that the MLE is asymptotically normal as sample size goes to infinity. Moreover, we can conclude that

Theorem 3.2. *Maximum likelihood estimator is consistent*

Proof: By (3.1), $\sqrt{n}(\hat{\theta}_n - \theta) \sim N(0, I^{-1}(\theta_0))$. Therefore asymptotically, $\mathrm{Var}(\sqrt{n}(\hat{\theta}_n - \theta)) = I^{-1}(\theta_0)$

$$P(\|\hat{\theta}_n - \theta\| > \epsilon) = P(\sqrt{n}\|\hat{\theta}_n - \theta\| > \sqrt{n}\epsilon) = P((\sqrt{n}\|\hat{\theta}_n - \theta\|)^2 > n\epsilon^2). \tag{18}$$

By Chebyshev Inequality,

$$P(\|\hat{\theta}_n - \theta\| > \epsilon) \le \frac{\mathrm{E}(\sqrt{n}\|\hat{\theta}_n - \theta\|^2)}{n\epsilon^2} = \frac{\mathrm{Var}(\sqrt{n}(\hat{\theta}_n - \theta))}{n\epsilon^2} \to \frac{I^{-1}(\theta_0)}{n\epsilon^2} \to 0. \tag{19}$$

Hence as n goest to infinity, MLE will goes to the true value with probability one. In other words, it's consistent. □

3 Efficiency

Before we give the definition of efficiency, we give the following theorem without proof.

Theorem 3.3. *Let $T(X)$ be a unbiased estimator of a function $\Psi(\theta)$ of the scalar parameter θ. Then lower bound of the variance of $T(X)$ is given by*

$$VarT(X) \ge \frac{\Psi^2(\theta)}{I^{-1}(\theta)}. \tag{20}$$

If θ is a $k \times 1$ column vector, the lower bound is

$$VarCovT(X) \ge \frac{\partial\Psi(\theta)}{\partial\theta} \cdot I(\theta)^{-1} \cdot \left(\frac{\partial\Psi(\theta)}{\partial\theta}\right)^T \tag{21}$$

Remark 3.1. *In (21), both the left side and right side are matrix. For matrix A and B, $A \ge B$ means that $A - B$ is positive semi-defined.*

If a statistic attains the lower bound denoted in (20) or (21), then it is **efficient**. We can also give the definition **efficiency**, namely,

$$e(\theta) = \frac{\Psi^2(\theta)I^{-1}(\theta)}{\mathrm{Var}X}, \theta \in R \tag{22}$$

(20) implies efficiency is always smaller than one. Since we in Theorem 3.3 is based on unbiased estimator $\mathrm{E}(T(X) - \theta)^2 = Var(T(X) - \theta) = VarT(X)$. In other words, the variance of a unbiased statistic shows the mean square error. The less the variance is, the more accurate and precise the statistic is. (3.3) shows that we can never have an ideal statistic with 0 variance. An statistic with larger Fisher Information offers a lower bound closer to 0, which implies a better chance to attain preciseness. On the other hand, the best statistic in terms of MSE can be obtained when variance reaches the lower bound, or in other words, when efficiency is one. Efficiency, in this sense, tells how accurate our statistic is.

Theorem 3.4. *Maximum likelihood estimator is asymptotically efficient.*

Proof: From (3.1) and (3.2), we can conclude that asymptotically, MLE is unbiased with variance $I^{-1}(\theta)$, which is the the lower bound presented in (21) (or (20). So when n is large enough, MLE is efficient. □

4 Disadvantages of MLE

Although MLE does hold some convenient mathematical properties (plug-in) and good asymptotic behaviour (asymptotic normal, consistency and efficiency), it also has some disadvantages.

1. All the good statistical behaviour are based on sufficiently large sample size. Actually, for small sample, MLE may be significantly biased. We may also lose efficiency when sample size is small.

2. We need to assume the distribution of random sample according to prior experience or knowledge. All the calculation, no matter for large sample or small sample, is based on that $f(x, \theta)$. However, in practice, it is quite possible that the $f(x, \theta)$ we propose is not close to the real distribution, which will cause a vital damage to the whole process.

3. To derive a convenient way to calculate MLE, we assumed independence among $X_1, \ldots, X_n$. This assumption may also be violated in practise.

4. In some cases, maximum likelihood estimator is not necessary exist. Even it does exist and can be calculated by differentiating the likelihood function, the calculation might be very complex and will not lead to a explicit answer.

5. Sometimes we apply Newton-Raphson, EM and etc. to give a numerical solution to MLE. This calls for more regulation on parameter space and $p.d.f.$. These methods may also sensitive to the initial point for iteration

4 Likelihood Ratio Test

In statistics, a likelihood ratio test is a statistical test used to compare the fit of two models, one of which (the null model) is a special case of the other (the alternative model). The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than the other. This likelihood ratio, or equivalently its logarithm, can then be used to compute a p-value, or compared to a critical value to decide whether to reject the null model in favour of the alternative model. When the logarithm of the likelihood ratio is used, the statistic is known as a log-likelihood ratio statistic, and the probability distribution of this test statistic, assuming that the null model is true, can be approximated using Wilks' theorem.

1 Two-Sided Tests

First let us concentrate on the most simple case. Suppose parameter $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$. Likelihood ratio can be defined as follows,

Definition 4.1. *Suppose we wish to test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$. Then the likelihood ratio, denoted as $\Lambda(\boldsymbol{x})$ is definded as*

$$\Lambda(\boldsymbol{x}) = \frac{\sup_{\theta=\theta_0} f(\boldsymbol{x}, \theta)}{\sup_{\theta\in\Theta} f(\boldsymbol{x}, \theta)} = \frac{f(\boldsymbol{x}, \theta_0)}{\sup_{\theta\in\Theta} f(\boldsymbol{x}, \theta)} \tag{23}$$

Furthermore, if the MLE of θ exists, we can write likelihood ratio as

$$\Lambda(\boldsymbol{x}) = \frac{f(\boldsymbol{x}, \theta_0)}{f(\boldsymbol{x}, \hat{\theta})}. \tag{24}$$

Remark 4.1. .

1. *Since the numerator of (23) is maximized over a smaller region compared to the denominator, we can conclude that likelihood ratio is always smaller than one.*
2. *An optimized case is when null hypothesis is true. Recall that if we have a large sample the MLE is approximately equal to the true value. Hence the likelihood ratio will approach one.*
3. *If θ_0 is far away from the true value of θ, then the difference between numerator and denominator in (23) will also be large, which will make $\Lambda(\boldsymbol{x})$ close to 0.*

Consequently, we should reject null hypothesis if likelihood ratio is significantly small. For a test of level α,

$$\alpha = P(\text{reject } H_0|H_0) = P_{\theta_0}(\Lambda(\theta) < c_\alpha), \tag{25}$$

where c_α is a constant decided by the distribution of the likelihood ratio and level α; and the rejection region is $(0, c_\alpha)$, which means that if the likelihood ratio is smaller than c_α, we should reject the null hypothesis with probability $1 - \alpha$.

For the more complex null hypothesis, namely $H_0 : \ \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ with $\Theta_0 \cup \Theta_1 = \Theta$, we can define the likelihood ratio in the same way,

$$\Lambda(\mathbf{x}) = \frac{sup_{\theta\in\Theta_0} f(\mathbf{x}, \theta)}{sup_{\theta\in\Theta} f(\mathbf{x}, \theta)}, \tag{26}$$

and the rejection. With the same reason for the simple case, we reject null hypothesis if $\Lambda(\mathbf{x})$ is significantly small. The level probability of making type one error can be calculated using and the rejection region $(0, c_\alpha)$. As for the power of the test, we can obtain it by

$$p = P(\text{reject } H_0|H_1) = P_{\theta\in\Theta_1}\{\Lambda(\mathbf{x}) > c_\alpha\}. \tag{27}$$

Example 4.1. *Suppose $X_1, \ldots, X_n$ are samples from $N(\mu, \sigma^2)$. Find the test statistic for H_0: $\mu = \mu_0$. vs H_1: $\mu \neq \mu_0$.*

Sol 4.1. *By Definition 4.1, we can write*

$$\Lambda(\boldsymbol{x}) = \frac{\sup_{\sigma^2}(2\pi\sigma^2)^{-n/2}\exp\{-\sum_{i=1}^n (x_i - \mu_0)^2/(2\sigma^2)\}}{\sup_{\mu,\sigma^2}(2\pi\sigma^2)^{-n/2}\exp\{-\sum_{i=1}^n (x_i - \mu)^2/(2\sigma^2)\}}. \tag{28}$$

Note that the MLE for the numerator are

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \mu_0)^2; \tag{29}$$

and the MLE for the denominator are

$$\hat{\mu} = \bar{x}, \ \hat{\hat{\sigma}}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2. \tag{30}$$

Therefore we can calculate $\Lambda(\boldsymbol{x})$ by plugging (29) and (30) back to (28)

$$\ln\frac{1}{\Lambda(\boldsymbol{x})} \propto \frac{\hat{\sigma}^2}{\hat{\hat{\sigma}}^2} = \frac{\sum_{i=1}^n (x_i - \mu_0)^2/n}{\sum_{i=1}^n (x_i - \bar{x})^2/(n-1)} \tag{31}$$

To simplify our test rule further we use the following equation, which can be established by expanding both sides.

$$\hat{\sigma}^2 = \hat{\hat{\sigma}}^2 + (\bar{x} - \mu_0)^2 \tag{32}$$

Therefore,

$$\ln \frac{1}{\Lambda(\boldsymbol{x})} \propto 1 + (\bar{x} - \mu_0)^2 / \hat{\hat{\sigma}}^2 \tag{33}$$

Because $s^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = n\hat{\hat{\sigma}}^2$, $\hat{\sigma}^2/\hat{\hat{\sigma}}^2$ *is a monotone increasing function of* $|T_n|$ *where*

$$T_n = \frac{\sqrt{n}(x - \mu_0)}{s}. \tag{34}$$

Therefore the likelihood ratio tests reject for small values of $\Lambda(\boldsymbol{x})$*, or equivalently, large values of* $|T_n|$*. Because* T_n *has a* T *distribution under* H_0*, the size* α *critical value is* $t_{n-1,1-\alpha/2}$*. We should reject null hypothesis if* $|T_n| \geq t_{n-1,1-\alpha/2}$*.*

To discuss the power of these tests, we need to introduce the non-central t distribution with k degrees of freedom and non-centrality parameter δ*. This distribution, denoted by* $T_{k,\delta}$ *is by definition the distribution of* $Z/\sqrt{V/K}$ *where* Z *and* V *are independent and have* $N(\delta, 1)$ *and* χ^2_k *distribution respectively. Note that* $\sqrt{n}(\bar{X} - \mu)/\sigma$ *and* $(n-1)s^2/\sigma^2$ *are independent and that* $(n-1)s^2/\sigma^2$ *has a* χ^2_{n-1} *distribution. Because* $E[(\sqrt{n}(\bar{X} - \mu_0))/\sigma] = \sqrt{n}(\mu - \mu_0)/\sigma$ *and* $Var(\sqrt{n}(\bar{X} - \mu_0))/\sigma)$*=1,* $\sqrt{n}(\bar{X} - \mu_0))/\sigma$ *has* $N(\delta, 1)$ *distribution, with* $\delta = \sqrt{n}(\mu - \mu_0))/\sigma$*. Thus the ratio*

$$T_n = \frac{\sqrt{n}(\bar{X} - \mu_0))/\sigma}{\sqrt{(n-1)s^2/[(n-1)\sigma^2]}} \tag{35}$$

has a $T_{n-1,\delta}$ *distribution, and the power can be obtained from non-central t distribution. Note that the distribution of* T_n *depends on* $\theta = (\mu, \sigma^2)$ *only through* δ*.*

2 One-Sided Tests

Example 4.2. (***continue***) *Suppose we are interested in testing* $H_0 : \mu \leq \mu_0$ *vs* $H_1 : \mu > \mu_0$*. Note that* $\mu = \bar{x}$ *if* $\bar{x} \leq \mu_0$ *and* $= \mu_0$ *otherwise. Thus* $\ln \Lambda(\boldsymbol{x}) = 0$ *if* $T_n \leq 0$ *and* $= (n/2)\ln(1 + T_n^2/(n-1))$ *for* $T_n > 0$*. We also argue that* $P_\delta[T_n > t]$ *is increasing in* δ*. Therefore the test that rejects* H_0 *for*

$$T_n \geq t_{n-1,1-\alpha}, \tag{36}$$

is of size α *for* $H_0 : \mu \leq \mu_0$*. Similarly, the size* α *likelihood ratio test for* $H_0 : \mu \geq \mu_0$ *vs* $H_1 : \mu < \mu_0$ *rejects null hypothesis if and only if*

$$T_n \leq t_{n-1,1-\alpha}. \tag{37}$$

The power function is $\Phi(z_\alpha + \mu\sqrt{n}/\sigma)$ *and is mnontone in* $\sqrt{n}\mu/\sigma$*.*

We can control both probabilities of error by selecting the sample size n large provided we consider alternatives of the form $|\delta| \geq \delta_1 > 0$ in the two-sided case and $\delta \leq \delta_1$ or $\delta \geq \delta_1$ in the one-sided cases.

3 Asymptotic Distribution of $\Lambda(\mathbf{x})$

Unfortunately, this significance is not always easy to measure because it is difficult to calculate the exact distribution of the likelihood ratio, which lead to the curiosity in its asymptotic distribution.

Theorem 4.1. ***(Wilk's)*** *Let* $\theta \in \Theta \subset R^k$ *and* H_0*:* $\theta_i = a_i, i = 1, 2, \ldots, s$*,* $s < k$ *vs the general alternative. As* n *goes to infinity,* $-2\ln\Lambda(\boldsymbol{x})$ *goes to* χ^2_m *in distribution under* H_0*.*

Proof: Set $\mathbf{a} = (a_1, \ldots, a_s)$ and $\hat{\theta}$ to be the MLE under H_1. By Definition 4.1, we can write $\ln \Lambda(\mathbf{x})$ as

$$\ln \Lambda(\mathbf{x}) = \sum_{i=1}^{n} \ln f(x_i, \mathbf{a}) - \sum_{i=1}^{n} \ln f(x_i, \hat{\theta}_n). \tag{38}$$

Now, take Taylor expansion for the first term at $\hat{\theta}_n$ in (38)

$$\begin{aligned} \ln \Lambda(\mathbf{x}) = \sum_{i=1}^{n} \ln f(x_i, \hat{\theta}_n) \quad &+ \quad \sum_{i=1}^{n} \sum_{r=1}^{s} \frac{\partial}{\partial \hat{\theta}_{n,r}} \ln f(x_i, \hat{\theta}_n)(a_r - \hat{\theta}_{n,r}) \\ &- \quad \frac{1}{2} \sum_{i=1}^{n} \sum_{r,q}^{s} \frac{\partial^2}{\partial \hat{\theta}_{n,r} \partial \hat{\theta}_{n,q}} \ln f(x_i, \hat{\theta}_n)(a_r - \hat{\theta}_{n,r})(a_r - \hat{\theta}_{n,q}) \\ &+ \quad o(\|\hat{\theta}_n - \mathbf{a}\|^2) - \sum_{i=1}^{n} \ln f(x_i, \hat{\theta}_n). \end{aligned} \tag{39}$$

Recall that $\hat{\theta}_{n,r}$ is the MLE which minimizes the likelihood function, so the second term in (39) is 0. Next, take ln on both side of (39) and multiply by -2, we obtain

$$\begin{aligned} -2 \ln \Lambda(\mathbf{x}) \quad &= \quad \sum_{i=1}^{n} \sum_{r,q}^{s} \frac{\partial^2}{\partial \hat{\theta}_{n,r} \partial \hat{\theta}_{n,q}} \ln f(x_i, \hat{\theta}_n)(a_r - \hat{\theta}_{n,r})(a_r - \hat{\theta}_{n,q}) + o(\|\hat{\theta}_n - \mathbf{a}\|^2) \\ &= \quad (\hat{\theta}_n - \mathbf{a})^T \mathbf{M} (\hat{\theta}_n - \mathbf{a}) = \sqrt{n}(\hat{\theta}_n - \mathbf{a})^T \cdot \frac{1}{n} \mathbf{M} \cdot \sqrt{n}(\hat{\theta}_n - \mathbf{a}), \end{aligned} \tag{40}$$

where $\mathbf{M}$ is an $s \times s$ matrix, with entries $m_{r,q} = \sum_{i=1}^{n} \frac{\partial^2}{\partial \hat{\theta}_{n,r} \partial \hat{\theta}_{n,q}} \ln f(x_i, \hat{\theta}_n)$, $i, j = 1, \ldots n$
By L.L.N, $\mathbf{M}/n$ goes to Fisher Information $I(\theta)$ in probability; by C.L.T.$\sqrt{n}(\hat{\theta}_n - \mathbf{a})$ goes to $N(0, I^{-1}(\theta))$ in distribution. Using Slusky again, we finish our proof. □
A natural question is when to reject null hypothesis using the statistic in Wilk's theorem. We have shown that we should reject H_0 when the likelihood ratio is significantly small, *i.e.* $\Lambda(\mathbf{x}) < c$. This is equivalent to $-2 \ln \Lambda(\mathbf{x}) > c$. Using Wilk's theorem, the level α test is

$$\alpha = P(-2 \ln \Lambda(\mathbf{x}) > \chi^2_{k,\alpha/2}) \tag{41}$$

with rejection region $(\chi^2_{k,\alpha/2}, \infty)$. .

Example 4.3. *Continue with Example (4.1). If we assume n is large, for the same $\Lambda(x)$ shown in (31) and level α, we should reject H_0 when $\Lambda(x) > c_\alpha$ where c_α is decided by $\alpha = P(-2 \ln \Lambda(x) > \chi^2_{1,\alpha})$; and the rejection region is $(\chi^2_{1,\alpha}, \infty)$.*

Although likelihood ratio test is not necessarily unbiased, we can approach the unbiasness by increasing sample size. In other words, likelihood ratio test is consistent. We give the proof of a very special case,H_0: $\theta = \theta_0$.

Theorem 4.2. *The likelihood ratio test in Theorem 4.1 is consistent.*

Proof: We need to show that if true value $\theta \neq \theta_0$, we reject H_0 with probability one as n goes to infinity. We reject the null hypothesis if $\Lambda(\mathbf{x}) < c$, or equivalently, if

$$-\ln \Lambda(\mathbf{x}) = \sum_{i=1}^{n} \ln f(x_i, \hat{\theta}_n) - \sum_{i=1}^{n} \ln f(x_i, \theta_0) > c. \tag{42}$$

Expand the first term in (42) at true value θ, we can re-write it as

$$\begin{aligned} -\ln\Lambda(\mathbf{x}) &= \sum_{i=1}^{n}\ln f(x_i,\theta)+\sum_{i=1}^{n}\sum_{r=1}^{s}\frac{\partial \ln f(x_i,\theta)}{\partial\theta_r}(\hat{\theta}_{n,r}-\theta_r)+no_p(|\hat{\theta}_n-\theta|)-\sum_{i=1}^{n}\ln f(x_i,\theta_0) \\ &= \sum_{i=1}^{n}\ln\frac{f(x_i,\theta)}{f(x_i,\theta_0)}+\sum_{i=1}^{n}\mathbf{J}(x_i,\theta)^T(\hat{\theta}_n-\theta)+no_p(|\hat{\theta}_n-\theta|), \end{aligned} \tag{43}$$

where $\mathbf{J}$ is Fisher Score, which is a $s\times 1$ vector. To use L.L.N. and C.L.T., we manipulate (43) into a more convenient form and split it into three parts, namely, nA, $\sqrt{n}\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbf{J}(x_i,\theta)^T\cdot B$, and $\sqrt{o}_p(|B|)$,

$$\begin{aligned} -\ln\Lambda(\mathbf{x}) &= n\cdot\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f(x_i,\theta)}{f(x_i,\theta_0)}+\sqrt{n}\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbf{J}(x_i,\theta)^T\sqrt{n}(\hat{\theta}_n-\theta)+\sqrt{n}o_p(\sqrt{n}|\hat{\theta}_n-\theta|) \\ &= nA+\sqrt{n}\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbf{J}(x_i,\theta)^T\cdot B+\sqrt{o}_p(|B|). \end{aligned} \tag{44}$$

By L.L.N, as n tends to infinity, A tends to

$$E_\theta\ln\frac{f(x_i,\theta)}{f(x_i,\theta_0)}=E_\theta\left(-\ln\frac{f(x_i,\theta_0)}{f(x_i,\theta)}\right)$$

with probablility one. Observe that $-\ln(\bullet)$ is a convex function, we can apply Jensen's Inequality to the limit of A,

$$A\to E_\theta\left(-\ln\frac{f(x_i,\theta_0)}{f(x_i,\theta)}\right)>-\ln E_\theta\frac{f(x_i,\theta_0)}{f(x_i,\theta)}=-\int\frac{f(x_i,\theta_0)}{f(x_i,\theta)}f(x_i,\theta)dx=-\ln 1=0. \tag{45}$$

Hence we have proved that $A\to constant>0$ with probability one. Consequently, $A\to n\cdot constant=\infty$ with probability one.

As for B and C, under C.L.T., both of them will tend to ininity. Since B is asymptotically normal by Theorem 3.1 and $\sum_{i=1}^{n}\mathbf{J}(x_i,\theta)/n$ approaches $E\mathbf{J}(\mathbf{x},\theta)=0$, we conclude that the second and third term in (44) is bounded with probability one. This suffice to show that $-\ln\Lambda(\mathbf{x})$ will be greater than any given constant as n goes to infinity with probability one. In other words, we reject null hypothesis with probability one. □

5 Likelihood Ration Confidence Interval

To compute the confidence interval, we need to find x_u and x_l such that $\theta\in(x_l,x_u)$ with probability $1-\alpha$ under H_0. Assume that we know the distribution of $\Lambda(\mathbf{x})$, then

$$1-\alpha=P_{\theta_0}(c_{1-\alpha/2}<\Lambda(\mathbf{x})<c_{\alpha/2}), \tag{46}$$

where under H_0, both $c_{1-\alpha/2}$ and $c_{\alpha/2}$ are in terms of θ_0.

Since

$$1-\alpha=P\left(F_{n,n-1,1-\alpha/2}\le\frac{\sum_{i=1}^{n}(x_i-\mu)^2/n}{\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)}\le F_{n,n-1,\alpha/2}\right), \tag{47}$$

the level $1-\alpha$ confidence interval of the $\Lambda(\mathbf{x})$ is $(F_{n,n-1,1-\alpha/2},F_{n,n-1,\alpha/2})$, meaning that under likelihood ratio test we believe that the statistic $\Lambda(\mathbf{x})$ will fall between $F_{n,n-1,1-\alpha/2}$ and $F_{n,n-1-2,\alpha/2}$ with probability $1-\alpha$.

If we rewrite (47) as

$$\begin{aligned} 1-\alpha &= P\left(F_{n,n-1,1-\alpha/2}\le\frac{\sum_{i=1}^{n}(x_i^2-2\mu+\mu^2)/n}{\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)}\le F_{n,n-1,\alpha/2}\right) \\ &= P(x_l<\mu^2-2\mu<x_u), \end{aligned} \tag{48}$$

where

$$x_l = \frac{F_{n,n-1,1-\alpha/2} \sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} - \sum_{i=1}^{n} x_i^2, \tag{49}$$

$$x_u = \frac{F_{n,n-1,\alpha/2} \sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} - \sum_{i=1}^{n} x_i^2. \tag{50}$$

From (48), we can solve for μ and get its confidence interval.is $\mu_l < \mu < \mu_u$, where

$$\mu_l = \max\{1 - \sqrt{1 + x_u^2}, 1 + \sqrt{1 + x_l^2}\},\ \mu_u = \min\{1 + \sqrt{1 + x_u^2}, 1 - \sqrt{1 + x_l^2}\}$$

The double sided confidence interval is $(\chi^2_{k,1-\alpha/2}, \chi^2_{k,\alpha/2})$

For the double sided confidence interval for the likelihood ratio is $(\chi^2_{1,1-\alpha/2}, \chi^2_{1,\alpha/2})$. As for the confidence interval for μ, we have similar result as shown in (48);only this time we find the critical value according to the χ^2_1 table.