Date: 12/26/2011

File Name: D:\A File\Research\Likelihood 11.12.26\Likelihood_8.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_0:

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x; \theta_0) \tag{1}$$

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) - (2) the likelihood is given by

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

As follows, we discuss (3) for discrete and continuous p.d.f., respectively.

Case 1: If $f(x; \theta_0)$ in (1) is discrete, then we have $P(X = x) = f(x; \theta_0)$; in turn, equation (3) becomes

$$P\{X_1 = x_1, \ldots, X_n = x_n\} = \prod_{i=1}^{n} f(x_i; \theta_0). \tag{4}$$

Case 2: If $f(x; \theta_0)$ in (1) is continuous, then for a small constant $\delta > 0$, we have $P(X = x) \approx P(x - \delta < x < x + \delta)$; in turn, equation (3) becomes

$$\begin{aligned} & P\{X_1 = x_1, \ldots, X_n = x_n\} \\ \approx & \prod_{i=1}^{n} P(x_i - \delta < X_i < x_i + \delta) = \prod_{i=1}^{n} [F_X(x_i + \delta; \theta_0) - F_X(x_i - \delta; \theta_0)] \\ = & \prod_{i=1}^{n} [2\delta f(\xi_i; \theta_0)] = (2\delta)^n \prod_{i=1}^{n} f(\xi_i; \theta_0). \\ \approx & (2\delta)^n \prod_{i=1}^{n} f(x_i; \theta_0), \end{aligned} \tag{5}$$

where $F_X(x; \theta_0)$ is the d.f. corresponding to $f(x; \theta_0)$, ξ_i is between $(x_i - \delta)$ and $(x_i + \delta)$ and we assume $f(x; \theta)$ is continuous.

Thus, equation (5) shows that likelihood (3) is approximately proportional to $\prod_{i=1}^{n} f(x_i; \theta_0)$.

Based on (4) and (5), the *likelihood function* for θ_0 with random sample (1)-(2) is given by

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \theta), \text{ for } \theta \in \Theta, \tag{6}$$

where $\mathbf{x} = (x_1, \ldots x_n)$ and Θ is the parameter space for θ_0 in (1). Therefore, for discrete or continuous p.d.f. $f(x; \theta)$, maximizing likelihood (3) and maximizing likelihood function (6) with respective to θ are equivalent.

2 Maxium Likelihood Estimator

For random sample (1)-(2), *maximum likelihood estimator (MLE)* is given by

$$\hat{\theta} = \arg\max_{\theta\in\Theta} L(\theta; \mathbf{x}). \tag{7}$$

Mathematically, MLE is the the value that can maximize the likelihood function. Recall the relation between likelihood function $L(\theta; \mathbf{x})$ and likelihood, hence statistically, MLE selects values of θ in Θ that produce a distribution that gives the obseved data the greatest likelihood.

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

$$\hat{\theta} = \arg\max_{\theta\in\Theta} l(\theta; \mathbf{x}) = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} l(\theta; x_i), \tag{8}$$

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

$$\nabla_\theta l(\hat{\theta}; \mathbf{x}) = 0. \tag{9}$$

This is known as the *likelihood equation*. So for the random sample (1)-(2)

$$\sum_{i=1}^{n} \nabla_\theta l(\hat{\theta}; x_i) = 0. \tag{10}$$

Evidently, there may be solutions of (10) that are not maxima or only local maxima, thus we need to refer to other properties of the likelihood function.

Example 2.1. *Suppose the p.d.f. in (1) is given by* $f(x; \mu) = \exp\{-(x-\mu)^2/2\}/\sqrt{2\pi}$, *where* $\theta_0 = \mu_0$ *in this example. Find the MLE of* μ_0.

Sol 2.1. *Using (6), we have*

$$\begin{aligned} l(\mu; \boldsymbol{x}) &= \ln L(\mu; \boldsymbol{x}) = \sum_{i=1}^{n} \ln f(x_i; \mu) \\ &= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2. \end{aligned} \tag{11}$$

To maximize the log-likelihood, we differentiate (11) w.r.t. μ *and set the derivative to be zero,*

$$\frac{\partial l(\mu; \boldsymbol{x})}{\partial \mu} = \sum_{i=1}^{n}(x_i - \mu) = 0. \tag{12}$$

The solution for (12) is $\mu = \sum_{i=1}^{n} x_i/n = \bar{\boldsymbol{x}}$. *Now, let us take the second derivative of (11),*

$$\frac{\partial^2 l(\mu; \boldsymbol{x})}{\partial \mu^2} = -n < 0. \tag{13}$$

By (13), we conclude that the first derivative of log-likelihood function is a decreasing. 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 log-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* $\hat{\mu} = \bar{\boldsymbol{x}}$.

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

Example 2.2. *Suppose the p.d.f. in (1) given by uniform distribution on* $(0, \theta)$. *Find* $\hat{\theta}$.

Sol 2.2. *Since* $f(x;\theta) = I_{(0,\theta)}(x)/\theta$, *we can write the log-likelihood function of the random sample as*

$$l(\theta; \boldsymbol{x}) = -n \ln \theta \cdot I_{(0,\theta)}(\boldsymbol{x}), \tag{14}$$

where I *is a indicator function, i.e., for any given set* A, $I_A(x) = 1$ *if* $x \in A$; *otherwise* $I_A(x) = 0$. *However, since*

$$\frac{\partial l(\theta; \boldsymbol{x})}{\partial \theta} = -\frac{n}{\theta} \cdot I_{(0,\theta)}(\boldsymbol{x}), \tag{15}$$

we have no solution for $\partial l(\theta; \boldsymbol{x})/\partial\theta = 0$.
Although we cannot maximize the likelihood by setting the derivative to be 0, from (15), we can see that the first derivative of log-likelihood function is negative, which indicates that the log-likelihood function is decreasing. 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* $\hat{\theta} = X_{(n)}$.

3 Properties of MLE

Let us start this section with a convenient computational property for MLE, namely, plug-in property. Then we will present the asymptotic distribution, consistency and efficiency of MLE. At the end this section, we will discuss several disadvantages of MLE.

3.1 Plug in Property

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

Theorem 3.1. *Let* $\hat{\theta}$ *denote the MLE of* θ_0 *in random sample (1)-(2). 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 p.d.f. of (1) in terms of* η *(i.e., re-parametrize the model using* η*). Then the MLE of* η_0 *is* $h(\hat{\theta})$.

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

3.2 Asymptotic Distribution

A nice asymptotic distribution will simplify computation for large sample. The following theorem will show that MLE is asymptotically normal when sample size is sufficiently large.

Theorem 3.2. *Suppose* $\hat{\theta}_n$ *is the MLE of true value* θ_0 *in random sample (1)-(2). Let* $I(\theta_0)$ *denote the Fisher Information in* X. *As* n *goes to infinity,* $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *goes to* $N(0, I^{-1}(\theta_0))$ *in distribution.*

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

$$\begin{aligned} 0 &= \frac{\partial l(\hat{\theta}_n; \mathbf{x})}{\partial \theta} = \frac{\partial l(\hat{\theta}_0; \mathbf{x})}{\partial \theta} + \frac{\partial^2 l(\hat{\theta}_0; \mathbf{x})}{\partial \theta^2}(\hat{\theta}_n - \theta_0) + o(\|\hat{\theta}_n - \theta_0\|^2) \\ &= \sum_{i=1}^{n} \frac{\partial l(\theta_0; x_i)}{\partial \theta} + \sum_{i=1}^{n} \frac{\partial^2 l(\theta_0; x_i)}{\partial \theta^2}(\hat{\theta}_n - \theta_0) + o(\|\hat{\theta}_n - \theta_0\|^2). \end{aligned} \tag{16}$$

By multiplying $1/\sqrt{n}$ on both side of (16) and ignoring the higher order reminders, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 l(\theta_0;x_i)}{\partial\theta^2}\cdot\sqrt{n}(\hat{\theta}_n-\theta_0)=-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l(\theta_0;x_i)}{\partial\theta}. \tag{17}$$

By L.L.N.,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 l(\theta_0;x_i)}{\partial\theta^2}\to E\left[\frac{\partial^2 l(\theta_0;x)}{\partial\theta^2}\right]=I(\theta)\text{ in probability;} \tag{18}$$

By C.L.T.,

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l(\theta_0;x_i)}{\partial\theta}\to N\left(0,E\left(\frac{\partial l(\theta_0;x)}{\partial\theta}\right)^2\right)=N(0,I(\theta_0))\text{ in distribution.} \tag{19}$$

Plug (18) and (19) back to (16) and apply Slusky Theorem, we can get $\sqrt{n}(\hat{\theta}_n-\theta_0)$ goes to $N(0,I^{-1}(\theta_0))$ in distribution as n goes to infinity. □

3.3 Consistency

We can get the consistency of MLE by digging into Theorem 3.2:

Theorem 3.3. *Maximum likelihood estimator is consistent*

Proof: By Theorem 3.2, $\sqrt{n}(\hat{\theta}_n-\theta_0)\sim N(0,I^{-1}(\theta_0))$ asymptotically. Therefore as n goes to infinity, $\mathrm{Var}(\sqrt{n}(\hat{\theta}_n-\theta))$ will approach $I^{-1}(\theta_0)$. Therefore,

$$P(\|\hat{\theta}_n-\theta_0\|>\epsilon)=P(\sqrt{n}\|\hat{\theta}_n-\theta_0\|>\sqrt{n}\epsilon)=P((\sqrt{n}\|\hat{\theta}_n-\theta_0\|)^2>n\epsilon^2). \tag{20}$$

By Chebyshev Inequality,

$$P(\|\hat{\theta}_n-\theta_0\|>\epsilon)\le\frac{\mathrm{E}(\sqrt{n}\|\hat{\theta}_n-\theta_0\|^2)}{n\epsilon^2}=\frac{\mathrm{Var}(\sqrt{n}(\hat{\theta}_n-\theta_0))}{n\epsilon^2}\to\frac{I^{-1}(\theta_0)}{n\epsilon^2}\to 0. \tag{21}$$

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

3.4 Efficiency

Before we define efficiency, let us give the following theorem without proof.

Theorem 3.4. *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*

$$VarT(X)\ge\frac{\Psi^2(\theta)}{I(\theta)}. \tag{22}$$

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

$$VarCovT(X)\ge\frac{\partial\Psi(\theta)}{\partial\theta}\cdot I(\theta)^{-1}\cdot\left(\frac{\partial\Psi(\theta)}{\partial\theta}\right)^T \tag{23}$$

Remark 3.1. *both the left side and right side of (23) 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 (22) or (23), then it is *efficient*. We can also give the definition *efficiency*, namely,

$$e(\theta) = \frac{\Psi^2(\theta) I^{-1}(\theta)}{\mathrm{Var} X}, \theta \in \Theta \subset R. \tag{24}$$

(22) implies efficiency is always smaller than one. Since Theorem 3.4 is based on unbiased estimator, i.e., $\mathrm{E}T(X) = \theta$, $\mathrm{E}(T(X) - \theta)^2 = \mathrm{E}(T(X) - \mathrm{E}T(X))^2 = Var(T(X) - \theta) = VarT(X)$. In other words, the variance of a unbiased statistic shows the *mean square error (MSE)*. The less the variance is, the more accurate and precise the statistic is.

However, based on Theorem 3.4, we can never have an ideal statistic with zero 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 equivalently, when efficiency is one. Efficiency, in this sense, tells how accurate our statistic is.

Theorem 3.5. *Maximum likelihood estimator is asymptotically efficient.*

Proof: From Therorem 3.2 and Theorem 3.3, we can conclude that asymptotically, MLE is unbiased with variance $I^{-1}(\theta_0)$, which is the the lower bound presented in (23) or (22). So when n is large enough, MLE is efficient. □

3.5 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 the assumed p.d.f. $f(x;\theta)$. 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 does 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 be sensitive to the initial point for iteration

4 Likelihood Ratio Test

A *likelihood ratio test (LRT)* is used to compare the fitness of two models, one of which (the null model) is a special case of the other (the alternative model). Suppose the parameter θ_0 in (1) belongs to a set Θ, then LRT can be defined as follows.

Definition 4.1. *The likelihood ratio test statistic for testing* $H_0 : \theta_0 \in \Theta_0, \Theta_0 \subset \Theta$ *vs* $H_1 : \theta_0 \in \Theta_0^c$ *is*

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta; \boldsymbol{x})}{\sup_{\theta \in \Theta} L(\theta; \boldsymbol{x})}. \tag{25}$$

A likelihood ratio test (LRT) is any test that has a rejection region of the form $\{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq c_\alpha\}$*, where* c *is any number satisfying* $0 \leq c_\alpha \leq 1$*.*

Remark 4.1. .

1. Since the numerator of (25) 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 be close to one. Otherwise, likelihood ratio will be close to zero

The constant c_α in Definition 4.1 is decided by the level of the test. For a test of level α,

$$\alpha = P(\text{reject } H_0 | H_0) = P_{\theta_0 \in \Theta_0}(\lambda(\mathbf{x}) < c_\alpha), \tag{26}$$

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 $1 - \alpha$.

A special case for (25) is testing $H_0 : \ \theta_0 = \theta_0^*$ vs $H_1 : \theta_0 \neq \theta_0^*$. Further, let us suppose the MLE exists. Since we have only one candidate under null hypothesis, $\lambda(\mathbf{x})$ becomes

$$\lambda(\mathbf{x}) = \frac{L(\theta_0^*; \mathbf{x})}{\sup_{\theta \in \Theta} L(\theta; \mathbf{x})} = \frac{L(\theta_0^*; \mathbf{x})}{L(\hat{\theta}; \mathbf{x})}. \tag{27}$$

The calculation of c_α calls for an explicit distribution of $\lambda(\mathbf{x})$. LRT has a nice χ^2 distribution when we have a large sample size. Let us present this property starting with the simple $H_0 : \ \theta_0 = \theta_0^*$ vs $H_1 : \theta_0 \neq \theta_0^*$.

Theorem 4.1. *Suppose* $\theta_0 \in \Theta \subset \mathcal{R}$*. For testing* $H_0 : \theta_0 = \theta_0^*$ *vs* $H_1 : \theta_0 \neq \theta_0^*$*, with random samples are (1)-(2). Then under* H_0*, as* $n \to \infty$*,* $-2 \ln \lambda(\boldsymbol{x}) \to \chi_1^2$ *in distribution.*

Proof: First expand $\ln L(\theta; \mathbf{x}) = l(\theta; \mathbf{x})$ in a Taylor series around the MLE $\hat{\theta}$, giving,

$$l(\theta; \mathbf{x}) = l(\hat{\theta}; \mathbf{x}) + l'(\hat{\theta}; \mathbf{x})(\theta - \hat{\theta}) + \frac{1}{2} l''(\hat{\theta}; \mathbf{x})(\theta - \hat{\theta})^2 + \ldots. \tag{28}$$

Now substitute the expansion for $l(\theta_0^*; \mathbf{x})$ in $-2 \ln \lambda(\mathbf{x}) = -2l(\theta_0^*; \mathbf{x}) + 2 \ln(\hat{\theta}; \mathbf{x})$, and get

$$-2 \ln \lambda(\mathbf{x}) \approx -l''(\hat{\theta}; \mathbf{x})(\theta_0^* - \hat{\theta}), \tag{29}$$

where we use the fact that $l'(\hat{\theta}; \mathbf{x}) = 0$. Since $l''(\hat{\theta}; \mathbf{x})$ is the observed fisher information $\hat{I}_n(\hat{\theta})$ and $\hat{I}_n(\hat{\theta})/n \to I(\theta_0^*) = I(\theta_0)$ under H_0. It follows from Theorem 3.2 and Slutsky's Theorem that $-2 \ln \lambda(\mathbf{x}) \to \chi_1^2$. □

Theorem 4.1 can be extended to the cases where the null hypothesis concerns a vector of parameters. The following generalization, which we state without proof, allows us to ensure Theorem 4.1 is true for large samples.

Theorem 4.2. ***(Wilk's Theorem)*** *For testing* $H_0 : \theta_0 \in \Theta_0$ *vs* $H_1 : \theta_0 \in \Theta_0^c$*, suppose random samples are (1)-(2) and* $\theta_0 \in \Theta$*. Then under* H_0*, as* $n \to \infty$*,* $-2 \ln \lambda(\boldsymbol{x}) \to \chi^2$ *in distribution. The degrees of freedom of the limiting distribution is the difference between the number of free parameters specified by* $\theta \in \Theta_0$ *and the mumber of free parameters specified by* $\theta \in \Theta$*.*

The computation of ν is usually straight forward. Most often, Θ can be represented as a subset of q-dimensional Euclidean space that contains and open subset in $\mathcal{R}^q$, and Θ_0 can be represented as a subset of p-dimensional Euclidean space that contains an open subset in $\mathcal{R}^p$, where $p < q$. Then $\nu = q - p$ is the

degrees of freedom for the test statistic.

Rejection of H_0 for small values of $\lambda(\mathbf{x})$ is equivalent to rejection for large values of $-2\ln\lambda(\mathbf{x})$. Thus,

$$H_0 \text{ is rejected if and only if } -2\ln\lambda(\mathbf{x}) \geq \chi_{\nu,\alpha}, \tag{30}$$

and the asymptotic rejection region is $(\chi_{\nu,\alpha}, \infty)$.

Next let us present and example using Wilk's Theorem.

Example 4.1. *Let $\theta = (p_1, p_2, p_3, p_4, p_5)$, where p_j's are non-negative and sum to 1. For (1)-(2), $f(j;\theta) = p_j, j = 1, \ldots, 5$. Find the LRT test statistic for testing $H_0 : p_1 = p_2 = p_3$ and $p_4 = p_5$ vs $H_1 : H_0$ is not true, and the asymptotic rejection region.*

Sol 4.1. *The likelihood function under Θ is*

$$L(\theta;\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} p_i^{y_i}, \text{where } y_j = \text{number of } x_1, \ldots, x_n \text{ equal to } j. \tag{31}$$

The full parameter space, Θ, with four free parameters, is really a four-dimensional set since $p_5 = 1 - p_1 - p_2 - p_3 - p_4$. The parameter set is defined by

$$\sum_{j=1}^{4} p_j \leq 1 \text{ and } p_j \geq 0, \ j = 1, \ldots, 4, \tag{32}$$

a subset of $\mathcal{R}^4$ containing an open subset of $\mathcal{R}^4$. Thus $q = 4$. There is only one free parameter in the set specified by H_0 because once p_1 is fixed, $p_2 = p_3$ must equal to p_1 and $p_4 = p_5$ must equal $(1 - 3p_1)/2$. Thus $p = 1$ and the degrees of freedom is $\nu = 4 - 1 = 3$.
To calculate $\lambda(\boldsymbol{x})$, the MLE of θ under both Θ_0 and Θ must be determined. By setting

$$\frac{\partial}{\partial p_j} l(\theta;\boldsymbol{x}) = 0, \text{ for each of } j = 1, \ldots, 4 \tag{33}$$

and using the facts that $p_5 = 1 - p_1 - p_2 - p_3 - p_4$ and $y_5 = n - y_1 - y_2 - y_3 - y_4$, we can verify that the MLE of p_j under Θ is $\hat{p}_j = y_j/n$. Under H_0, the likelihood function reduces to

$$L(\theta;\boldsymbol{x}) = \prod_{i=1}^{n} f(\theta;x_i) = p_1^{y_1+y_2+y_3} \left(\frac{1-3p_1}{2}\right)^{y_4+y_5}. \tag{34}$$

Using the same method as (33), the MLE's under H_0 are $p_{10} = p_{20} = p_{30} = (y_1 + y_2 + y_3)/(3n)$ and $p_{40} = p_{50} = (1 - 3\hat{p}_{10})/2$. Substituting these values and the $\hat{p}_j$ values into $L(\theta;\boldsymbol{x})$ and combining terms with the same exponent yield

$$\lambda(\boldsymbol{x}) = \left(\frac{y_1+y_2+y_3}{3y_1}\right)^{y_1} \left(\frac{y_1+y_2+y_3}{3y_2}\right)^{y_2} \left(\frac{y_1+y_2+y_3}{3y_3}\right)^{y_3} \left(\frac{y_4+y_5}{2y_4}\right)^{y_4} \left(\frac{y_4+y_5}{2y_5}\right)^{y_5}. \tag{35}$$

Thus the test statistic is

$$-2\ln\lambda(\boldsymbol{x}) = 2\sum_{i=1}^{5} y_i \ln\left(\frac{y_i}{m_i}\right), \tag{36}$$

where $m_1 = m_2 = m_3 = (y_1 + y_2 + y_3)/3$ and $m_4 = m_5 = (y_4 + y_5)/2$. The asymptotic size α test rejects H_0 if $-2\ln\lambda(\boldsymbol{x}) \geq \chi^2_{3,\alpha}$.

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.

Theorem 4.3. *The likelihood ratio test is consistent.*

Proof: We need to show that if true value $\theta_0 \neq \theta_0^*$, 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

$$-\ln \lambda(\mathbf{x}) = \sum_{i=1}^{n} l(\hat{\theta}_n; x_i) - \sum_{i=1}^{n} l(\theta_0^*; x_i) > c. \tag{37}$$

Expand the first term in (37) at true value θ_0, we can re-write it as

$$\begin{aligned} -\ln \lambda(\mathbf{x}) &= \sum_{i=1}^{n} l(\theta_0; x_i) + \sum_{i=1}^{n}\sum_{r=1}^{s} \frac{\partial l(\theta_0; x_i)}{\partial \theta_{0,r}}(\hat{\theta}_{n,r} - \theta_{0,r}) + n o_p(\|\hat{\theta}_n - \theta_0\|) - \sum_{i=1}^{n} l(\theta_0^*; x_i) \\ &= \sum_{i=1}^{n} \ln \frac{f(x_i;\theta_0)}{f(x_i;\theta_0^*)} + \sum_{i=1}^{n} \mathbf{J}(\theta_0; x_i)^T(\hat{\theta}_n - \theta_0) + n o_p(\|\hat{\theta}_n - \theta_0\|), \end{aligned} \tag{38}$$

where $\mathbf{J}$ is Fisher Score. To use L.L.N. and C.L.T., we manipulate (38) into a more convenient form and split it into three parts, namely, nA, $\sqrt{n} \cdot B \cdot C$, and $\sqrt{o}_p(\|B\|)$,

$$\begin{aligned} -\ln \lambda(\mathbf{x}) &= n \cdot \frac{1}{n}\sum_{i=1}^{n} \ln \frac{f(x_i;\theta_0)}{f(x_i;\theta_0^*)} + \sqrt{n} \cdot \frac{1}{n}\sum_{i=1}^{n} \mathbf{J}(x_i;\theta_0)^T \sqrt{n}(\hat{\theta}_n - \theta_0) + \sqrt{n} o_p(\sqrt{n}\|\hat{\theta}_n - \theta_0\|) \\ &= nA + \sqrt{n} \cdot B \cdot C + \sqrt{o}_p(\|B\|). \end{aligned} \tag{39}$$

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

$$E_{\theta_0} \ln \frac{f(x_i;\theta_0)}{f(x_i;\theta_0^*)} = E_{\theta_0}\left(-\ln \frac{f(x_i;\theta_0^*)}{f(x_i;\theta_0)}\right)$$

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

$$A \to E_{\theta_0}\left(-\ln \frac{f(x_i;\theta_0^*)}{f(x_i;\theta_0)}\right) > -\ln E_{\theta_0} \frac{f(x_i;\theta_0^*)}{f(x_i;\theta_0)} = -\ln \int \frac{f(x_i;\theta_0^*)}{f(x_i;\theta_0)} f(x_i;\theta_0) dx = -\ln 1 = 0. \tag{40}$$

Hence we have proved that $A \to constant > 0$ with probability one. Consequently, $A \to n \cdot constant = \infty$ with probability one.
Since B is asymptotically normal by Theorem 3.1 and C approaches $E\mathbf{J}(\mathbf{x};\theta_0) = 0$, we conclude that the second and third term in (39) 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. □

In the end of this section, we present an example of small sample size. In this case, we can deduce the exact distribution of $\lambda(\mathbf{x})$ without requiring $n \to \infty$ or applying Wilk's Theorem.

Example 4.2. *Suppose the p.d.f. in (1) is given by $N(\mu, \sigma^2)$. Find the test statistic for H_0: $\mu_0 = \mu_0^*$. vs H_1: $\mu_0 \neq \mu_0^*$.*

Sol 4.2. *By Definition 4.1, we can write*

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\sigma^2}(2\pi\sigma^2)^{-n/2} \exp\{-\sum_{i=1}^{n}(x_i - \mu_0^*)^2/(2\sigma^2)\}}{\sup_{\mu,\sigma^2}(2\pi\sigma^2)^{-n/2} \exp\{-\sum_{i=1}^{n}(x_i - \mu)^2/(2\sigma^2)\}}. \tag{41}$$

Note that the MLE for the numerator is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0^*)^2; \tag{42}$$

and the MLE for the denominator are

$$\hat{\mu} = \bar{x},\ \hat{\hat{\sigma}}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{43}$$

Therefore we can calculate $\lambda(\boldsymbol{x})$ by plugging (42) and (43) back to (41)

$$\ln \frac{1}{\lambda(\boldsymbol{x})} \propto \frac{\hat{\sigma}^2}{\hat{\hat{\sigma}}^2} = \frac{\sum_{i=1}^{n}(x_i - \mu_0^*)^2/n}{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)} \tag{44}$$

To simplify our test rule further we use the following equation, which can be established by expanding $\hat{\sigma}^2$.

$$\hat{\sigma}^2 = \hat{\hat{\sigma}}^2 + (\bar{x} - \mu_0^*)^2 \tag{45}$$

Therefore,

$$\ln \frac{1}{\lambda(\boldsymbol{x})} \propto 1 + (\bar{x} - \mu_0^*)^2/\hat{\hat{\sigma}}^2 \tag{46}$$

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

$$T_n = \frac{\sqrt{n}(x - \mu_0^*)}{s}. \tag{47}$$

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 $|T_n| \geq t_{n-1,1-\alpha/2}$.

5 Likelihood Ratio Confidence Interval

In the previous section, we derived the fact that $-2\ln\lambda(\mathbf{x})$ has an asymptotic chi squared distribution. For fixed θ_0^* in $H_0 : \theta_0 = \theta_0^*$, the acceptance region is given by

$$\{\lambda(\mathbf{x}) : -2\ln\lambda(\mathbf{x}) \leq \chi_{1,\alpha}\} \tag{48}$$

Then by inverting the LRT, we can conclude that for (1)-(2), the set

$$\left\{\theta : -2\ln\left(\frac{L(\theta;\mathbf{x})}{L(\hat{\theta};\mathbf{x})}\right) \leq \chi^2_{1,\alpha}\right\} \tag{49}$$

is an approximate $1-\alpha$ confidence interval.

Example 5.1. *The p.d.f. in (1) is given by Bernoulli(p) and $Y = \sum_{i=1}^{n} X_i$. We have the approximate $1-\alpha$ confidence set*

$$\left\{p : -2\ln\left(\frac{p^y(1-p)^{n-y}}{\hat{p}^y(1-\hat{p})^{n-y}}\right) \leq \chi^2_{1,\alpha}\right\}. \tag{50}$$

For some special distributions, we can find the exact distribution of $\lambda(\mathbf{x})$. In this case, we can get a more accurate confidence interval by inverting the LRT of $H_0 : \theta_0 = \theta_0^*$ vs $H_1 : \theta_0 \neq \theta_0^*$. The confidence interval is of form

$$\text{accept } H_0 \text{ if } \frac{L(\theta_0^*;\mathbf{x})}{L(\hat{\theta};\mathbf{x})} \leq k(\theta_0), \tag{51}$$

with the resulting confidence region

$$\{\theta : L(\theta;\mathbf{x}) \geq k'(\mathbf{x},\theta)\}, \tag{52}$$

for some function k' that gives $1-\alpha$ confidnece.

Example 5.2. *Suppose that the p.d.f. of (1) is exponential(λ). Find the confidence interval for λ by inverting a level α test of $H_0 : \lambda = \lambda_0^*$ vs $H_1 : \lambda \neq \lambda_0^*$.*

Sol 5.1. *For random sample (1)-(2), the LRT statistic is given by*

$$\frac{\exp(-\sum x_i/\lambda_0^*)/\lambda_0^{*n}}{\sup_{\lambda>0}\exp(-\sum x_i/\lambda)/\lambda^n} = \frac{\exp(-\sum x_i/\lambda_0^*)/\lambda_0^{*n}}{e^{-n}/(\sum x_i/n)} = \left(\frac{\sum x_i}{n\lambda_0^*}\right)^n e^n e^{-\sum x_i/\lambda_0^*}. \tag{53}$$

For fixed λ_0^*, the acceptance region is given by

$$A(\lambda_0^*) = \left\{\boldsymbol{x} : \left(\frac{\sum x_i}{\lambda_0^*}\right)^n e^{-\sum x_i/\lambda_0^*} \geq k^*\right\}, \tag{54}$$

where k^* is a constant chosen to satisfy $P_{\lambda_0^*}(\boldsymbol{X} \in A(\lambda_0^*)) = 1-\alpha$ and the constant e^n/n has been absorbed into k^*. Inverting this acceptance region gives the $1-\alpha$ confidence set

$$C(\boldsymbol{x}) = \left\{\lambda : \left(\frac{\sum x_i}{\lambda}\right)^n e^{-\sum x_i/\lambda} \geq k^*\right\}, \tag{55}$$

which is an interval in the parameter space Θ.
The expression defining $C(\boldsymbol{x})$ depends on $\boldsymbol{x}$ only through $\sum x_i$. So the confidence interval can be expressed in the form

$$C\left(\sum x_i\right) = \left\{\lambda : L\left(\sum x_i\right) \leq \lambda \leq U\left(\sum x_i\right)\right\} \tag{56}$$

where L and U are functions determined by the constraints that the set (54) has probability $1-\alpha$ and

$$\left(\frac{\sum x_i}{L\left(\sum x_i\right)}\right)^n e^{-\sum x_i/L(\sum x_i)} = \left(\frac{\sum x_i}{U\left(\sum x_i\right)}\right)^n e^{-\sum x_i/U(\sum x_i)}. \tag{57}$$

If we set

$$\frac{\sum x_i}{L\left(\sum x_i\right)} = a \text{ and } \frac{\sum x_i}{U\left(\sum x_i\right)} = b, \text{ where } a > b \text{ are constants.} \tag{58}$$

Then (57) becomes $a^n e^{-1} = b^n e^{-b}$, which yields easily to numerical solution. To work out some details, let $n = 2$ and note that $\sum X_i \sim \Gamma(2,\lambda)$ and $\sum X_i/\lambda \sim \Gamma(2,1)$. Hence from (58), the confidence interval becomes

$$\left\{\lambda : \frac{1}{a}\sum x_i \leq \lambda \leq \frac{1}{b}\sum x_i\right\},$$

where a and b satisfy

$$P_\lambda\left(\frac{1}{a}\sum X_i \leq \lambda \leq \frac{1}{b}\sum X_i\right) = P\left(b \leq \frac{\sum X_i}{\lambda} \leq a\right) = 1-\alpha$$

Then

$$P\left(b \leq \frac{\sum X_i}{\lambda} \leq a\right) = \int_b^a te^{-t}dt = e^{-b}(b+1) - e^{-a}(a+1). \tag{59}$$

To get, for example, a 90% confidence interval, we must simultaneously satisfy the probability condition and the constraint. To Three decimal places, we got $a = 5.480$, $b = 0.441$, with a confidence doefficient of 0.90006. Thus,

$$P_\lambda\left(\frac{1}{5.480}\sum X_i \leq \lambda \leq \frac{1}{0.441}\sum X_i\right) = 0.90006.$$