5. Likelihood-based Inferences

6/30/2024

Contents

1 Likelihood Review

Let θ be an unknown parameter or a vector of unknown parameters. Assume that $\{X_1, X_2, \cdots, X_n\}$ is a set of IID random variables with distribution $F(x;\theta)$. When the distribution is discrete, the probability mass function (PMF) is used in the likelihood function. If the distribution is continuous, the probability distribution density function (PDF) is used in the likelihood function.

We only focus on continuous distributions in this series of research tutorials. In the rest of this note, we assume $\{X_1, X_2, \cdots, X_n\}$ is a set of IID continuous random variables with density function $f(x;\theta)$. **Discrete distributions such as binomial and Posson distributions are used occasionally to explain some concepts.**

Recall that the likelihood of observing an IID random sample $\{x_1, x_2, \cdots, x_n\}$ from the distribution with density function $f(x;\theta)$ is given by

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i : \theta)$$

The MLE of θ, denoted by $\hat{\theta}$ is te solution to the optimization problem

$$\hat{\theta} = \arg\max_{\theta \in \Omega} L(\theta : \mathbf{x}).$$

Since the logarithm of the likelihood is easier to handle in mathematics, we define the above optimization problem using the log-likelihood, that is

$$\hat{\theta} = \arg\max_{\theta \in \Omega} l(\theta : \mathbf{x}),$$

where $l(\theta : \mathbf{x}) = \log[L(\theta : \mathbf{x})]$.

The method of moment (MM) and the maximum likelihood estimation (MLE) provide a single estimated value from a random sample to approximate the unknown parameter. This estimation is called **point estimation**. This note will focus on inferences based on MLE.

1.1 Some Basic Concepts

We briefly review some of the concepts introduced in an earlier note.

Parameter and Statistic: A parameter is a numerical characteristic of a population. A statistic is a numerical value calculated from a sample taken from the population.

Estimation, Estimator, and Estimate: An estimation is a method. An estimator is a formula or function defined based on a data set. An estimate is a value calculated using the formula (i.e., estimator) from the data set.

Point Estimate: a point estimate is a single value used to estimate the population parameter. It is a random number because the sample is random.

Interval estimate: an interval estimate is a range (interval) of values that contains the true value of the unknown parameter.

Remark - The point estimate is only a descriptive statistic based on the sample, not the population. For example, when considering the average starting salaries of recent statistics graduates at a university, we take a random sample of students and calculate the average salary, say \$55,000. We can not say the average starting salary of all statistics graduates at the university is \$55,000. A better way to describe the starting salary of the population of statistics graduates is that the starting salary is **around \$55,000** - This is a range! That is, *we need an interval estimate to generalize the information from the sample to the population - confidence interval method!*

2 Goodness Measures of Point Estimate

Several measures are commonly used to assess the goodness of a point estimate. Before introducing these measures, please keep in mind that an estimate from a random sample is also random. This means that a point estimate has a distribution. We will discuss the probability distribution in a subsequent section. For now, we only assume a distribution associated with the point estimate so that we can define these measures rigorously using the definition of statistical expectation.

Assume that $\hat{\theta}$ is a point estimate of θ based on an estimation method. The measures of goodness of point estimate are based on estimation error which is defined to be $\text{error}(\hat{\theta}) = \hat{\theta} - \theta$

Bias: the bias of point estimate $\hat{\theta}$ is defined to be $\text{bias}(\hat{\theta}) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$.

The bias of a point estimate could be positive, negative, or zero. When the bias is zero, $E(\hat{\theta}) = \theta$, the corresponding point estimate is called **unbiased point estimate**. The **unbiasedness** is a good feature for a point estimate.

Mean Square Error (MSE): The MSE of point estimate $\hat{\theta}$ is defined to be $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$.

MSE calculates the average of squared error. It is a legitimate measure of the goodness of a point estimate. **The smaller the MSE, the better the point estimate.**

Relationship among Bias, MSE, and Variance: $\text{MSE}(\hat{\theta}) = \text{var}(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2$

The derivation of the above relationship is straightforward by noting that $(\hat{\theta}-\theta)^2 = \{[(\hat{\theta}-E(\hat{\theta})]+[E(\hat{\theta})-\theta]\}^2$.. Expanding this binomial with some algebraic clean-up, you will see the relationship.

Absolute Estimation Error is defined to be $\epsilon = |\hat{\theta} - \theta|$.

We can not evaluate the above goodness measures unless the true value of the population parameter is available. *This is a dilemma: we try to measure the goodness of the point estimate of an unknown parameter, but calculating these measures requires the true value of the unknown parameter.*

In fact, we use measures in different ways:

1. In simulation studies, we assume the true value of the population parameter and generate random samples,
2. If prior information on the population parameter is available, we can approximate these measures based on a random sample.
3. We can use these measures to derive other inferential procedures such as confidence intervals.

3 Interval Estimates

One issue with a point estimate is that it has information about accuracy and precision. When we say the value of the population parameter is close to the sample statistics, but we don't know how close it is, and how close is considered as *close*. This section introduces the general framework to derive the confidence interval for a population parameter θ based on its point estimate $\hat{\theta}$ **with the assumption that $\hat{\theta}$ has a known density $f(\hat{\theta})$.**

We consider imposing a bound to the absolution estimation error $\epsilon = |\hat{\theta} - \theta| < b$, we can then calculate the probability

$$P(|\hat{\theta} - \theta| < b) = P(\theta - b < \hat{\theta} < \theta + b) = \int_{\theta-b}^{\theta+b} f(\hat{\theta}) d\hat{\theta} = p_0$$

The above p_0 is the derived probability that the absolute estimation error is bounded by given b with θ being known. Under these assumptions, the inequality $\theta - b < \hat{\theta} < \theta + b$ in the middle of the above equation tells how the point estimate is close to the parameter with probability p_0. **That is, with known θ, b, and $f_{\hat{\theta}}(\cdot)$, we can calculate the probability that $\hat{\theta}$ falls into interval $(\theta - b, \theta + b)$.**

Next, we look at the above equation under different assumptions: assume we choose the value of p_0, say $p_0 = 0.95 = 95\%$. $\hat{\theta}$ is calculated from a random sample and b is a known error bound. Note that the above equation can re-expressed into the following form

$$P(|\hat{\theta} - \theta| < b) = P(\hat{\theta} - b < \theta < \hat{\theta} + b) = p_0.$$

The above equation says that the random interval $(\hat{\theta} - b, \hat{\theta} + b)$ has $100p_0\%$ chance to include the true value of the parameter - This is the confidence interval with confidence level $100p_0\% = 95\%$ with the choice of $p_0 = 0.95$.

One piece of information that needs to be addressed is how to get b in the above discussion.

In one-sample confidence intervals of the population mean, we use either the central limit theorem of the strong assumption of normality of the population, the sample mean as the point estimate of the population

mean μ has a normal distribution, then b is the margin of error (i.e., absolution error bound of point estimation) that has the following form

$$b = Z_{1-\alpha/2}\frac{s}{\sqrt{n}}$$

where $\alpha = 1 -$ given confidence level which is called the significance level in testing hypotheses. $Z_{1-\alpha/2}$ is the quantile of the distribution of point estimate $\hat{\mu} = \bar{X}$ which is normal in the one sample confidence of population mean.

Therefore, the critical information required in deriving the confidence interval of a population parameter is the distribution of the point estimate of the parameter. It is customarily called the sampling distribution of the point estimate.

4 Sampling Distribution of MLE

Recall the central limit theorem concerning the distribution of sample means we introduced in the elementary statistics:

Central Limit Theorem

Assume that a random sample $\{x_1, x_2, \cdots, x_n\}$ *is taken from a population with mean* μ *and standard deviation* σ*. Define* $\bar{X} = \sum_{i=1}^{n} x_i/n$ *to be the point estimate of population mean* μ*. If the sample is large,* $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$

Remark: The central limit theorem (CLT) **does not assume** a specific distribution of the population but unknown mean μ and variance σ^2. The only vague condition is that the sample size is large. **Any results derived from the CLT are called large sample (asymptotic) results**.

One of the objectives of this note is to develop asymptotic results for the MLE that are similar to the above CLT so that we can make statistical inferences such as constructing confidence intervals and testing hypotheses. In the next few sections, we introduce the building blocks to be used in the MLE-based inferences.

4.1 Basic Set-up and Notations

Let $\{x_1, x_2, \cdots, x_n\} \overset{\text{i.i.d}}{\sim} f_\theta(x)$, $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$ is a vector of k parameters (could be a single parameter when $k = 1$). The likelihood of observing the data is given by

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$$

with corresponding log-likelihood function in the following

$$l(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i).$$

4.1.1 Gradient Vector and Score Equations

The system of score equations is given by

$$\begin{cases} \frac{\partial l(\theta)}{\partial \theta_1} = 0, \\ \frac{\partial l(\theta)}{\partial \theta_2} = 0, \\ \cdots\cdots\cdots \\ \frac{\partial l(\theta)}{\partial \theta_k} = 0. \end{cases}$$

The above system can be written in the following form

$$\frac{\partial l(\theta)}{\partial \theta} = \mathbf{0} \quad \text{or} \quad \nabla_\theta l(\theta) = \mathbf{0}$$

The mathematical notation ∇ (read */'na.bla/*) is a differential operator. It is commonly used in Calculus.

The gradient vector of $l(\theta)$ is defined to be

$$\nabla_\theta l(\theta) = \left(\frac{\partial l(\theta)}{\partial \theta_1}, \frac{\partial l(\theta)}{\partial \theta_2}, \cdots, \frac{\partial l(\theta)}{\partial \theta_k} \right)$$

4.2 Expected Value of Score Functions

To derive the asymptotic normality of the MLE, we need to use the following result from advanced real analysis: the order of differentiation and integration is exchangeable. The proof of general results requires more advanced mathematical tools in the abstract measure theory (to be covered in the doctoral-level real analysis course).

Lemma 1: Suppose that $f(x, \theta)$ is differentiable in θ and there exists a function $g(x, \theta)$ such that

1. $\left| \frac{\partial f(x,\vartheta)}{\partial \vartheta} \right| \leq g(x, \theta)$ for all x and θ such that $|\vartheta - \theta| \leq \delta_0$;
2. $\int_{-\infty}^{\infty} g(x, \theta) dx < \infty$ for all θ.

Then

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x, \theta) dx = \int_{-\infty}^{\infty} \frac{\partial f(x, \theta)}{\partial \theta} dx.$$

We will not prove the lemma in this note. For the likelihood function, the regularity conditions are satisfied. So we can use the lemma in statistical inference.

Fact. The expected value of the score function is equal to zero.

Proof: We will use the definition of expectation and the above lemma in the following proof. Without loss of generality, we consider the log-likelihood of observing a single data point $f(x : \theta)$ and assume θ to be a univariate parameter.

$$E\left[\frac{\partial \log f(x, \theta)}{\partial \theta} \right] \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} \frac{\partial \log f(x, \theta)}{\partial \theta} f(x, \theta) dx$$

$$= \int_{-\infty}^{\infty} \left[\frac{\partial f(x, \theta)/\partial \theta}{f(x, \theta)} \right] f(x, \theta) dx = \int_{-\infty}^{\infty} \frac{\partial f(x, \theta)}{\partial \theta} dx$$

$$\stackrel{\text{switch}}{=} \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} f(x, \theta) dx = \frac{\partial}{\partial \theta}(1) = 0.$$

Remarks Lemma 1 introduced in this subsection is related to several *big* theorems in mathematics (advanced calculus and real analysis).

1. The two conditions in Lemma 1 are also called *regularity conditions*. Many statistical theorems (or procedures) require some regularity conditions needed in mathematical proofs and derivations. Make sure that the regularity conditions are satisfied in practical applications.
2. If the integral is definite with scalar integral limits, then above Lemma 1 is a special case of **Leibniz Rule** in Calculus.

3. More general forms of the above Lemma 1 are introduced in the real analysis textbooks under various convergence theorems such as **Lebesgue Dominant Convergence (i.e., bounded convergence theorem) and Monotone convergence theorem**.
4. Since the derivative is the limit of the rate of change (instantaneous rate), there is also a rule of exchange in the order of integral and limit.
5. The integral is also viewed as an *infinite sum*, there is also a rule of exchange of the order of summation and derivative (or summation and limit).
6. **These types of exchange order operations are frequently used in statistics deriving asymptotic statistical results.**

To conclude this subsection, we use the same idea in Lemma 1 to derive the following Lemma concerning the second-order derivative of the log-likelihood function. Lemma 2 links the two definitions of the **Fisher Information** to be introduced in the next subsection.

Lemma 2: Under some similar regularity conditions (as stated in Lemma 1), we have

$$E\left[\frac{\frac{\partial^2}{\partial\theta^2}f(x,\theta)}{f(x,\theta)}\right]=0.$$

Proof: Using the definition of expectation and the exchange of the order of the derivative and the integral, we have

$$E\left[\frac{\frac{\partial^2}{\partial\theta^2}f(x,\theta)}{f(x,\theta)}\right]=\int\left[\frac{\frac{\partial^2}{\partial\theta^2}f(x,\theta)}{f(x,\theta)}\right]f(x,\theta)dx$$

$$=\int\frac{\partial^2}{\partial\theta^2}f(x,\theta)dx\overset{\text{switch}}{=}\frac{\partial^2}{\partial\theta^2}\int f(x,\theta)dx=\frac{\partial^2}{\partial\theta^2}(1)=0.$$

4.3 Fisher Information Matrix

The information on the **variance and covariance of the MLE** is contained in the Fisher information matrix. It must be explicitly specified when making inferences about the MLE of model parameters. To better understand the concept, we start with the case with single-parameter models.

4.3.1 Fisher Information Number

Assume that θ is a univariate parameter of the population with density $f(x,\theta)$. Let $\{x_1,x_2,\cdots,x_n\}\sim f(x,\theta)$ be an IID sample. We use vector $\mathbf{x}$ to denote the set of random samples. The log-likelihood of observing the data is a function of θ that has the following form

$$l(\theta:\mathbf{x})=\log L(\theta:\mathbf{x})=\sum_{i=1}^{n}\log f(x_i,\theta).$$

The **Fisher Information Number** based on a random sample with size n is defined to be of the following form

$$I_n(\theta) \stackrel{\text{def}}{=} E_{\mathbf{x}}\left[\left(\frac{\partial}{\partial\theta}\log L(\mathbf{x},\theta)\right)^2\right].$$

Lemma 1 in the previous subsection says that

$$E_X\left[\frac{\partial}{\partial\theta}f(X_i,\theta)\right] = 0 \quad \text{for} \quad i = 1, 2, \cdots, n.$$

Therefore,

$$\left\{E_{\mathbf{x}}\left[\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log f(x_i,\theta)\right]\right\}^2 = \left\{E_{\mathbf{x}}\left[\frac{\partial}{\partial\theta}\log L(\theta:\mathbf{x})\right]\right\}^2 = 0$$

Using the formula $\text{Var}(X) = E(X^2) - [E(X)]^2$ for all random variable X, we have

$$I_n(\theta) \stackrel{\text{def}}{=} E_{\mathbf{x}}\left[\left(\frac{\partial}{\partial\theta}\log L(\theta:\mathbf{x})\right)^2\right] - \left\{E_{\mathbf{x}}\left[\frac{\partial}{\partial\theta}\log L(\theta:\mathbf{x})\right]\right\}^2 = \text{Var}_{\mathbf{x}}\left[\log L(\theta:\mathbf{x})\right].$$

This means that **the Fisher Information $I_n(\theta)$ is the variance of the score function**.

Next, we present an alternative definition of the Fisher Information. Note that

$$\frac{\partial^2}{\partial\theta^2}\left[\log L(\theta:\mathbf{x})\right] = \frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta}\left[\log L(\theta:\mathbf{x})\right]\right\} = \frac{\partial}{\partial\theta}\left\{\left[\frac{\partial L(\theta:\mathbf{x})/\partial\theta}{L(\theta:\mathbf{x})}\right]\right\}$$

Using the multiplicative rule of derivative, we have

$$= \frac{L(\theta:\mathbf{x})\frac{\partial^2}{\partial\theta^2}L(\theta:\mathbf{x}) - \left[\frac{\partial}{\partial\theta}L(\theta:\mathbf{x})\right]^2}{L^2(\theta:\mathbf{x})}$$

$$= \frac{\frac{\partial^2}{\partial\theta^2}L(\theta:\mathbf{x})}{L(\theta:\mathbf{x})} - \left[\frac{\frac{\partial}{\partial\theta}L(\theta:\mathbf{x})}{L(\theta:\mathbf{x})}\right]^2.$$

From Lemma 2, the first term of the above equation is zero. Therefore,

$$\frac{\partial^2}{\partial\theta^2}\left[\log L(\theta:\mathbf{x})\right] = -\left[\frac{\frac{\partial}{\partial\theta}L(\theta:\mathbf{x})}{L(\theta:\mathbf{x})}\right]^2.$$

This means we can also define the Fisher Information number can also be derived as

$$I_n(\theta) = -E\left\{\frac{\partial^2}{\partial\theta^2}\left[\log L(\theta:\mathbf{x})\right]\right\}.$$

This means that the Fisher Information number is the negative expectation of the second-order derivative of the log-likelihood.

Since we assume an IID sample in most cases, we can derive the relationship between the Fisher information number associated with the entire observed data set and that based on the individual observation in the data set. Denote $I_0(\theta)$ as the Fisher information number of a single observation of an IID sample.

$$I_n(\theta) = -E\left\{\frac{\partial^2}{\partial\theta^2}\left[\log L(\theta:\mathbf{x})\right]\right\} = -E\left\{\frac{\partial^2}{\partial\theta^2}\left[\sum_{i=1}^{n}\log f(\theta:x_i)\right]\right\}$$

$$= \sum_{i=1}^{n}\left\{-E\left[\frac{\partial^2}{\partial\theta^2}\log f(\theta:x_i)\right]\right\} = nI_0(\theta).$$

Remarks Some comments on the Fisher Information:

1. The Fisher information is only a function of the parameter since it is defined as an expectation (with respect to $\mathbf{X}$) of the log-likelihood.
2. If we replace the parameter with an estimated one such as MLE, we obtain $\widehat{I_n(\theta)} = I_n(\hat{\theta})$. $\widehat{I_n(\theta)}$ **is called observed Fisher information number!**
3. **The Fisher information number is always positive.**
4. The Fisher information number is dependent on the sample size. This will be used in developing the asymptotic normality for the MLE in the next section.

Example: Consider an IID sample $\{x_1, x_2, \cdots, x_n\} \sim f(x,\theta) = \theta e^{-\theta x}$. Find the Fisher information number.

Note that the log-likelihood function of θ is given by

$$l(\theta:\mathbf{x}) = n\log\theta - \theta\sum_{i=1}^{n} x_i.$$

The second order derivative of $l(\theta)$ with respect to θ is

$$\frac{\partial^2}{\partial\theta^2}l(\theta) = \frac{\partial}{\partial\theta}\left(\frac{n}{\theta} - \sum_{i=1}^{n}x_i\right) = -\frac{n}{\theta^2}.$$

By definition,

$$I_n(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}l(\theta,\mathbf{x})\right) = -E(-\frac{n}{\theta^2}) = \frac{n}{\theta^2}.$$

The exponential density function has another form (i.e., reparametrization) $f(x,\beta) = \frac{1}{\beta}e^{-x/\beta}$. A natural question is whether we need to repeat the same calculation in the above example to find the Fisher information $I_n(\beta)$. The answer is unnecessary. We can *reparametrization* in the Fisher information number.

Let $\eta = \psi(\theta)$ and $\psi(\cdot)$ is invertible, that is, $\theta = \psi^{-1}(\eta)$ exists. Assume that two density forms of X are $f(x,\theta)$ and $g(x,\eta)$. Clearly, $g(x,\eta) = f(x,\psi^{-1}(\eta))$ *(why?)*. Assume further that $I_n(\theta) = I_n[\psi^{-1}(\eta)]$ is known.

Due to reparametrization, the same random variable has different forms of density function. When we work on the expectation with the random variable, we should specify the associated density function. Next, we express $I_g(\eta)$ with respect to $I_f(\theta)$.

$$I_{n,g}(\eta) = E_g\left[\left(\frac{\partial}{\partial\eta}\log g(x,\eta)\right)^2\right] = E_g\left[\left(\frac{\partial}{\partial\eta}\log f(x,\psi^{-1}(\eta))\right)^2\right]$$

$$= E_g\left[\left(\frac{\partial}{\partial\eta}\log f(x,\psi^{-1}(\eta)) \times \frac{\partial}{\partial\eta}\psi^{-1}(\eta)\right)^2\right]$$

$$= E_g\left[\left(\frac{\partial}{\partial\eta}\log f(x,\psi^{-1}(\eta))\right)^2 \times \left(\frac{\partial}{\partial\eta}\psi^{-1}(\eta)\right)^2\right]$$

$$= E_g\left[\left(\frac{\partial}{\partial\eta}\log f(x,\psi^{-1}(\eta))\right)^2\right] \times \left(\frac{\partial}{\partial\eta}\psi^{-1}(\eta)\right)^2 = I_{n,f}(\theta) \times \left(\frac{\partial}{\partial\eta}\psi^{-1}(\eta)\right)^2.$$

That is,

$$I_{n,g}(\eta) = I_{n,f}(\theta) \times \left(\frac{\partial}{\partial\eta}\psi^{-1}(\eta)\right)^2.$$

4.3.2 Fisher Information Matrix

For multi-parameter models (distributions), we need a Fisher information matrix to characterize the covariance structure of the MLE of the vector of multiple parameters. Denote $f(x;\theta)$ be the density function of X with a k-dimensional parameters $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$ *(Caution: if not specified, all vectors in this series of note are column vectors. This is also true in most mathematics and statistics books and literature)*

Under the same setup, the likelihood function of observing $\{x_1, x_2, \cdots, x_n\} \sim f(x,\theta)$ is given by

$$L(\theta : \mathbf{x}) = \prod_{i=1}^{n} f(x_i, \theta)$$

The gradient vector of the log-likelihood is

$$\frac{\partial}{\partial\theta}\log L(\theta,\mathbf{x}) = \left(\frac{\partial}{\partial\theta_1}\log L(\theta,\mathbf{x}), \frac{\partial}{\partial\theta_2}\log L(\theta,\mathbf{x}), \cdots, \frac{\partial}{\partial\theta_k}\log L(\theta,\mathbf{x}),\right).$$

Denote

$$\frac{\partial}{\partial\theta^T}\log L(\theta,\mathbf{x}) = \left(\frac{\partial}{\partial\theta_1}\log L(\theta,\mathbf{x}), \frac{\partial}{\partial\theta_2}\log L(\theta,\mathbf{x}), \cdots, \frac{\partial}{\partial\theta_k}\log L(\theta,\mathbf{x}),\right)^T$$

The **Fisher information matrix** of the k-dimensional parameter θ is defined to be

$$\mathbb{I}_n(\theta) = E\left(\frac{\partial}{\partial\theta}\log L(\theta,\mathbf{x})\frac{\partial}{\partial\theta^T}\log L(\theta,\mathbf{x})\right).$$

The explicit matrix form of the two-parameter case is

$$\mathbb{I}_n(\theta) = E\left[\left(\frac{\partial}{\partial\theta_1}\log L(\theta,\mathbf{x}), \frac{\partial}{\partial\theta_2}\log L(\theta,\mathbf{x})\right)\left(\frac{\partial}{\partial\theta_1}\log L(\theta,\mathbf{x}), \frac{\partial}{\partial\theta_2}\log L(\theta,\mathbf{x})\right)^T\right]$$

$$= E\begin{bmatrix} \left(\frac{\partial}{\partial\theta_1}\log L(\theta,\mathbf{x})\right)^2 & \frac{\partial}{\partial\theta_1}\log L(\theta,\mathbf{x})\frac{\partial}{\partial\theta_2}\log L(\theta,\mathbf{x}) \\ \frac{\partial}{\partial\theta_1}\log L(\theta,\mathbf{x})\frac{\partial}{\partial\theta_2}\log L(\theta,\mathbf{x}) & \left(\frac{\partial}{\partial\theta_2}\log L(\theta,\mathbf{x})\right)^2 \end{bmatrix}.$$

Remarks

1. The Fisher information matrix is a **square matrix**. Its dimension is dependent on the dimension of the vector of parameters. For the case of a k-dimensional parameter vector, the dimension of the corresponding Fisher information matrix is $k \times k$.
2. The individual cell element in the Fisher information matrix is expressed in the following

$$[\mathbb{I}_n(\theta)]_{i,j} = \frac{\partial}{\partial\theta_i}\log L(\theta,\mathbf{x})\frac{\partial}{\partial\theta_j}\log L(\theta,\mathbf{x})$$

3. The Fisher information matrix is the covariance matrix of the gradient vector $\partial \log L(\theta:\mathbf{x})/\partial\theta$.
4. The Fisher information at each individual observed data point x_i is still a $k \times k$ square matrix and is similarly given by

$$I_0(\theta) = E\begin{bmatrix} \left(\frac{\partial}{\partial\theta_1}\log L(\theta,x_i)\right)^2 & \frac{\partial}{\partial\theta_1}\log L(\theta,x_i)\frac{\partial}{\partial\theta_2}\log L(\theta,x_i) \\ \frac{\partial}{\partial\theta_1}\log L(\theta,x_i)\frac{\partial}{\partial\theta_2}\log L(\theta,x_i) & \left(\frac{\partial}{\partial\theta_2}\log L(\theta,x_i)\right)^2 \end{bmatrix}.$$

5. For an IID sample $\{x_1, x_2, \cdots, x_n\} \sim f(\mathbf{x}:\theta)$,

$$I_n(\theta) = nI_0(\theta).$$

Analogous to the case of single parameter distributions, we also have the following alternative definition of the Fisher information matrix for multi-parameter distributions

$$\mathbb{I}_n(\theta) = -E\left(\frac{\partial^2}{\partial\theta\partial\theta^T}\log L(\theta:\mathbf{x})\right).$$

where

$$\mathbb{H}_n(\theta,\mathbf{x}) = \frac{\partial^2}{\partial\theta\partial\theta^T}\log L(\theta:\mathbf{x}) = \begin{bmatrix} \frac{\partial^2}{\partial\theta_1^2}\log L(\theta,x_i) & \frac{\partial^2}{\partial\theta_1\partial\theta_2}\log L(\theta,x_i) & \cdots & \frac{\partial^2}{\partial\theta_1\partial\theta_k}\log L(\theta,x_i) \\ \frac{\partial^2}{\partial\theta_2\partial\theta_1}\log L(\theta,x_i)) & \frac{\partial^2}{\partial\theta_2^2}\log L(\theta,x_i) & \cdots & \frac{\partial^2}{\partial\theta_2\partial\theta_k}\log L(\theta,x_i) \\ \vdots & \vdots & \vdots & \\ \frac{\partial^2}{\partial\theta_k\partial\theta_1}\log L(\theta,x_i) & \frac{\partial^2}{\partial\theta_k\partial\theta_2}\log L(\theta,x_i) & \cdots & \frac{\partial^2}{\partial\theta_k^2}\log L(\theta,x_i) \end{bmatrix}_{k\times k}$$

is the well-known **Hessian matrix**. **Hessian Matrices** are widely used in optimization problems.

Caution: **Hessian Matrix and Jacobian Matrix** are sometimes confusing.

1. The **Hessian matrix** is the square matrix of second-order partial derivatives of the objective function to be optimized.
2. The **Jacobian matrix** is the first-order derivatives of a vector of functions such as $(f_1(\theta), f_2(\theta), \cdots, f_k(\theta))$ where $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$. The **Jacobian matrix** associated with the vector of functions is defined by

$$J(\theta) = \begin{bmatrix} \frac{\partial}{\partial\theta_1} f_1(\theta) & \frac{\partial}{\partial\theta_2} f_1(\theta) & \cdots & \frac{\partial}{\partial\theta_k} f_1(\theta) \\ \frac{\partial}{\partial\theta_1} f_2(\theta) & \frac{\partial}{\partial\theta_2} f_2(\theta) & \cdots & \frac{\partial}{\partial\theta_k} f_2(\theta) \\ \vdots & \vdots & \vdots & \\ \frac{\partial}{\partial\theta_1} f_k(\theta) & \frac{\partial}{\partial\theta_2} f_k(\theta) & \cdots & \frac{\partial}{\partial\theta_k} f_k(\theta) \end{bmatrix}_{k\times k}$$

3. When finding the MLE of unknown parameters from the objective log-likelihood function $l(\theta)$, we first calculate gradient functions

$$\left(\frac{\partial l(\theta)}{\partial\theta_1}, \frac{\partial l(\theta)}{\partial\theta_2}, \cdots, \frac{\partial l(\theta)}{\partial\theta_k}\right) \stackrel{\text{def}}{=} [f_1(\theta), f_2(\theta), \cdots, f_k(\theta)]$$

Then the **Hessian Matrix** associated with the objective function $l(\theta)$ and the **Jacobian Matrix** associated with gradient (score) functions **are identical**.

4.4 Multiple Random Variables and Random Vectors

We first introduce the concept of covariance and properties associated with multiple random variables.

Let X and Y be two variables, the covariance that measures the linear relationship between two random variables is defined by

$$\text{cov}(X, Y) = E\{[X - E(X)][Y - E(Y)]\}.$$

Some properties of covariance:

1. $\text{cov}(X, Y) = \text{cov}(Y, X)$.
2. When $X = Y$, $\text{cov}(X, Y) = \text{var}(X)$.
3. Let a and b are scalars,

- $\text{cov}(aX, Y) = a \times \text{cov}(X, Y)$;
- $\text{cov}(aX, bY) = ab \times \text{cov}(X, Y)$;
- $\text{cov}(aX + b, Y) = a \times \text{cov}(X, Y)$

3. Let $Z = aX + bY$, $E(Z) = E[aX + bY] = aE[X] + bE[Y]$ and $\text{var}(Z) = a^2\text{var}(X) = 2ab \cdot \text{cov}(X, Y) + b^2\text{var}(Y)$. *(Prove this using the definition of the variance.)*
4. The (*linear*) correlation coefficient between X and Y is given by

$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$$

For a given sample $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, the sample correlation coefficient is given by

$$r = \frac{\left[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right]/n}{\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

The variance of the linear combination of random variables X and Y, $Z = aX + bY$, involves covariance of X and Y. In general, we can consider the linear combination of a sequence of random variables $\{X_1, X_2, \cdots, X_p\}$, say $W = \sum_{i=1}^p a_i X_i$, we can calculate the variance of W in the following

$$\text{var}(W) = \sum_{i=1}^p a_i^2 \text{var}(X_i) + \sum_{i\neq j} a_i a_j \text{cov}(X_i, X_j)$$

The above variance of W is obtained based on the definition of variance and binomial expansion. However, we can write the linear combination of the random variance in the vector form

$$W = (a_1, a_2, \cdots, a_p) \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}.$$

Therefore, W is the dot product of a scalar vector (coefficient of the linear combination) and a random vector. This motivates us to study random vectors.

Random Vectors and Properties

We have discussed the distribution of single random variables and the related characterization. If we have a vector of p random variables

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}.$$

.

Each random variable has support (i.e., domain) $\mathbb{R}_i \subseteq \mathbb{R}$ for $i = 1, 2, \cdots, p$. Let

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}.$$

.

be the realization (i.e., observed data values) of random variable $\mathbb{X}$. This means

$$\mathbf{x} \in (\mathbb{R}_1 \times \mathbb{R}_2 \times \cdots \times \mathbb{R}_p)^T.$$

The **mean of the random vector** is

$$E(\mathbf{X}) \stackrel{\text{def}}{\equiv} \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}.$$

The **variance-covariance matrix** is defined by

$$V[\mathbf{X}] \stackrel{\text{def}}{\equiv} \begin{bmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_p) \\ \text{cov}(X_2, X_1) & \text{var}(X_2) & \cdots & \text{cov}(X_2, X_p) \\ \vdots & \vdots & \vdots & \vdots \\ \text{cov}(X_k, X_1) & \text{cov}(X_p, X_2) & \cdots & \text{var}(X_p) \end{bmatrix}_{p \times p}$$

Some Properties of Random Vectors

1. A constant vector a (vector of constants) satisfies $E[a] = a$.
2. For random vectors $\mathbf{X}$ and $\mathbf{Y}$, $E[\mathbf{X} + \mathbf{Y}] = E[\mathbf{X}] + E[\mathbf{Y}]$
3. Let $\mathbf{a}$ be a scalar (column) vector and $\mathbf{X}$ be a (column) random vector. $\mathbf{a}^T\mathbf{X}$ is well defined. Then $\text{var}(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\text{cov}(\mathbf{X})\mathbf{a}$.

Example Derive the variance of $Z = aX + bY$ using the properties of random vectors. Let $\mathbf{d} = (a, b)$

$$\text{var}(Z) = \text{var}(\mathbf{d}^T\mathbf{Z}) = \mathbf{d}^T\text{cov}(\mathbf{Z})\mathbf{d}$$

$$= [a, b] \begin{bmatrix} \text{var}(X) & \text{cov}(X, Y) \\ \text{cov}(Y, X) & \text{var}(Y) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

Expand the above matrix to get the quadratic form

$$\text{var}(Z) = a^2\text{var}(X) + 2ab \times \text{cov}(X, Y) + b^2\text{var}(Y).$$

4.5 Multivariate Normal Distribution

Recall the density function of univariate random variable X with mean μ and variance σ^2 is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{for} \quad -\infty < x < \infty.$$

Example: Suppose X is the height (in inches) and Y is the weight (in pounds) of a male student in a large university. Furthermore suppose that X and Y follow normal distribution with parameters $\mu_X = 69$, $\mu_Y = 155$, $\sigma_X = 2.5$, and $\sigma_Y = 20$. Furthermore, the correlation coefficient between X and Y is $\rho = 0.55$. There may be different distributions that satisfy the given conditions. However, the given conditions uniquely determine the bivariate normal distribution that has the following joint density function

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_x}{\sigma_x} \right)^2 + \left(\frac{y-\mu_y}{\sigma_y} \right)^2 - \rho \left(\frac{x-\mu_x}{\sigma_x} \right) \left(\frac{y-\mu_y}{\sigma_y} \right) \right] \right\}.$$

Obviously, the density for the Bivariate Normal is ugly, and it only gets worse when we consider higher dimensional joint densities of more normal distributions. We can write the density in a more compact form using matrix notation,

Denote

$$\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \quad \mathbf{\Sigma} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

The matrix form of the bivariate normal distribution is given by

$$f(\mathbf{x}) = \frac{1}{2\pi}(\det\mathbf{\Sigma})^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) \right].$$

Σ is the covariance matrix of $\mathbf{x}$. The above expression can be generalized to multivariate normal distribution. The multivariate normal distribution is uniquely determined if the covariance matrix is specified.

With the above example, we can specify the vector and covariance matrix required in the bivariate normal distribution.

$$\mu = \begin{bmatrix} 69 \\ 155 \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma} = \begin{bmatrix} 2.5^2 & 0.55 \times 2.5 \times 20 \\ 0.55 \times 2.5 \times 200 & 20^2 \end{bmatrix} = \begin{bmatrix} 6.25 & 27.5 \\ 27.5 & 400 \end{bmatrix}$$

$$\det \mathbf{\Sigma} = 6.25 \times 400 - 27.5^2 = 1743.75.$$

and

$$\mathbf{\Sigma}^{-1} = \frac{1}{6.25 \times 400 - 27.5^2} \begin{bmatrix} 400 & -27.5 \\ -27.5 & 6.25 \end{bmatrix} = \begin{bmatrix} 0.2293907 & -0.01577061 \\ -0.01577061 & 0.003584229 \end{bmatrix}.$$

$$f(\mathbf{x}) = 0.0038 \exp\left[-\frac{1}{2}(x - 69, y - 155) \begin{bmatrix} 0.2294 & -0.0158 \\ -0.0158 & 0.0036 \end{bmatrix} \begin{bmatrix} x - 69 \\ y - 155 \end{bmatrix} \right].$$

Notation of General Multivariate Normal Distribution

Let

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix}, \quad \text{and} \quad \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \sigma_{k3} & \cdots & \sigma_{kk} \end{pmatrix}.$$

We use the following notation to denote the k-dimensional normal distribution

$$\mathbf{X} \sim \mathcal{N}_k(\mu, \mathbf{\Sigma}).$$

Properties of Multivariate Normal Distributions

Let $\mathbf{X} = [X_1, X_2, \cdots, X_k]^T$ and $\mu = E[\mathbf{X}]$, and $\text{cov}(\mathbf{X}) = \mathbf{\Sigma} = (\sigma_{ij})$. The multivariate normal distribution is given by

$$\mathbf{X} \to_p \mathcal{N}_k(\mu, \mathbf{\Sigma}).$$

The following are properties of multivariate normal distributions that will be used in the subsequent note.

1. All **marginal distributions** *(distribution of individual component of multivariate normal)* of a multivariate normal distribution are **normal distributions**.
2. All **conditional distributions** of a multivariate normal distribution are **normal distribution**.
3. All **linear combinations of the components** of multivariate normal distribution are also **normal distributions**.

4.6 Asymptotic Normality of MLE

Recall the MLE of θ based on IID $\{x_1, x_2, \cdots, x_n\} \overset{\text{iid}}{\sim} f(x:\theta)$, denoted by $\hat{\theta}$, is the solution to the following optimization problem

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \log L(\theta : \mathbf{x})$$

where $L(\theta)$ is the likelihood of θ that given explicitly in the following

$$L(\theta : \mathbf{x}) = \prod_{i=1}^{n} f(x_i : \theta).$$

The **Fisher Information** of θ is given by

$$\mathbb{I}_n(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log L(\theta:\mathbf{x})\right)\left(\frac{\partial}{\partial\theta}\log L(\theta:\mathbf{x})\right)^T\right]$$

$$= -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log L(\theta:\mathbf{x})\right].$$

Denote the **Fisher Information Matrix** based on individual observed data points by

$$\mathbb{I}_0(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log L(\theta:x)\right] = \frac{\mathbb{I}_n(\theta)}{n}.$$

With the above discussions and notations, we have the following sampling distribution of MLE of θ.

Theorem: (*Asymptotic normality of MLE*) Under some regularity conditions, the MLE of θ based on $\{x_1, x_2, \cdots, x_n\} \overset{\text{iid}}{\sim} f(x:\theta)$ has the following asymptotic normality

$$\sqrt{n}(\hat{\theta} - \theta) \overset{\text{approx}}{\to_p} \mathcal{N}\left[\mathbf{0}, \mathbb{I}_0^{-1}(\theta)\right].$$

Since θ in the covariance matrix $\mathbb{I}_0^{-1}$ is unknown. When making inferences, we use $\widehat{\mathbb{I}_0(\theta)} = \mathbb{I}_0(\hat{\theta})$.

Remarks:

1. $\sqrt{n}(\hat{\theta}-\theta)$ is approximately normally distributed when the sample size is large. The variance of $\sqrt{n}(\hat{\theta}-\theta)$ is $\mathbb{I}_0^{-1}(\theta)$ which is independent of the sample size and is defined based on the individual random data point.
2. θ is the unknown parameter and $\hat{\theta}$ is the MLE of θ.
3. $\sqrt{n}(\hat{\theta}-\theta)$ can be considered as a sequence of random variables (indexed by the sample size n). $\to_p$ stands for **converges in distribution**.

Example: The following data represent active repair times (in hours) for an airborne communication transceiver.

0.2, 0.3, 0.5, 0.5, 0.5, 0.5, 0.6, 0.6, 0.7, 0.7, 0.7, 0.8, 0.8, 1.0, 1.0, 1.0, 1.0, 1.1, 1.3, 1.5, 1.5, 1.5, 1.5, 2.0, 2.0, 2.2, 2.5, 3.0, 3.0, 3.3, 3.3, 4.0, 4.0, 4.5, 4.7, 5.0, 5.4, 5.4, 7.0, 7.5, 8.8, 9.0, 10.3, 22.0, 24.5.

Assuming the above data set was taken from a Weibull distribution with density function $f(x, \alpha, \beta) = \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}}$.

As we derived in the previous note, the log-likelihood of observing the data is a function of α and β

$$l(\alpha, \beta) = n[\log(\alpha) + \log(\beta)] + (\beta - 1)\sum_{i=1}^{n} \log(x_i) - \alpha \sum_{i=1}^{n} x_i^{\beta}$$

The score equations are given by

$$\begin{cases} \frac{\partial l(\alpha,\beta)}{\partial \alpha} = \frac{n}{\alpha} - \sum_{i=1}^{n} x_i^{\beta} = 0, \\ \frac{\partial l(\alpha,\beta)}{\partial \beta} = \frac{n}{\beta} + \sum_{i=1}^{n} \log(x_i) - \alpha \sum_{i=1}^{n} x_i^{\beta} \log(x_i) = 0. \end{cases}$$

We next write R code to find the MLE and the Hessian matrix (the negative observed Fisher Information Matrix).

```
# Data set
x = c(0.2, 0.3, 0.5, 0.5, 0.5, 0.5, 0.6, 0.6, 0.7, 0.7, 0.7, 0.8, 0.8, 1.0, 1.0, 1.0, 1.0,
      1.1, 1.3, 1.5, 1.5, 1.5, 1.5, 2.0, 2.0, 2.2, 2.5, 3.0, 3.0, 3.3, 3.3, 4.0, 4.0, 4.5,
      4.7, 5.0, 5.4, 5.4, 7.0, 7.5, 8.8, 9.0, 10.3, 22.0, 24.5)
n = length(x)
## log-likelihood
negLogLik = function(A){
  a = A[1]
  b = A[2]
  n*log(a) + n*log(b) + (b-1)*sum(log(x)) - a*sum(x^b)
}
## gradient function
grFun = function(A){
  a = A[1]
  b = A[2]
  ga = n/a - sum(x^b)
  gb= n/b +sum(log(x)) -a*sum(x^b*log(x))
  c(ga,gb)
}
## calling R function optim()
results = optim(par = c(.5,.5), fn = negLogLik, gr = grFun, method = "BFGS",
      control = list(maxit = 20000,fnscale = -1), hessian = TRUE)
MLE = results$par
Counts = results$counts
Convergence = results$convergence
Hessian = results$hessian
I0.inv = -solve(Hessian/n)
out = list(MLE = MLE, Counts = Counts, Convergence = Convergence, Hessian = Hessian, I0.inv = I0.inv)
out
```

```
$MLE
[1] 0.3377259 0.8896178

$Counts
function gradient
      33        9

$Convergence
```

```
[1] 0

$Hessian
          [,1]      [,2]
[1,] -394.5369 -236.5124
[2,] -236.5124 -250.2269

$I0.inv
           [,1]       [,2]
[1,]  0.2631775 -0.2487532
[2,] -0.2487532  0.4149563
```

Therefore, the sampling distribution of $(\hat{\alpha}, \hat{\beta})$ is given by

$$\sqrt{n}\begin{pmatrix}\hat{\alpha}-\alpha\\ \hat{\beta}-\beta\end{pmatrix} \to_p \mathcal{N}\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}0.2631785 & -0.2487531\\ -0.2487531 & 0.4149552\end{pmatrix}\right]$$

Remarks

1. **Caution:** *The Hessian matrix $\mathbb{H}$ reported in `optim()` is negative observed Fisher information: $-\widehat{\mathbb{I}_n(\theta)} = -\mathbb{I}_n(\hat{\theta})$!*
2. The covariance matrix in the asymptotic normality of MLE requires $\mathbb{I}_0(\theta)$. The observed Fisher information $\hat{\theta}$ is equal to the Hessian matrix, $\mathbb{H}_n$, divided by sample size n. That is, $\mathbb{I}_0(\hat{\theta}) = -\mathbb{H}(\theta : \mathbf{x})/n$.
3. $\text{var}[\sqrt{n}(\hat{\alpha}) - \alpha] = n\text{var}(\hat{\alpha}) = 0.2631785$ which implies that $\text{var}(\hat{\alpha}) = 0.2631785/45 \approx 0.005848411$; $\text{var}[\sqrt{n}(\hat{\beta}) - \beta] = n\text{var}(\hat{\beta}) = 0.4149552$ which implies that $\text{var}(\hat{\beta}) = 0.4149552/45 \approx 0.009221227$; and $\text{cov}(\hat{\alpha}, \hat{\beta}) = -0.2487531/45 = -0.005527847$.
4. Let $\hat{\omega} = c\hat{\alpha} + d\hat{\beta}$. Then $\text{var}(\hat{\omega}) = \text{var}[c\hat{\alpha} + d\hat{\beta}] = c^2\text{var}(\hat{\alpha}) + d^2\text{var}(\hat{\beta}) + 2cd\text{cov}(\hat{\alpha}, \hat{\beta}) = 0.005848411c^2 + 0.009221227d^2 - 2 \times 0.005527847cd$.

5 Inference of MLE

Three types of inferences will be discussed in this section: confidence intervals, and significance tests.

5.1 Confidence Intervals

Because MLEs are asymptotically normally distributed, we can use the same procedure as we used in introductory statistics. Assuming the confidence level in the subsequent discussion is 95%.

For convenience, let $(\hat{\theta}_1, \hat{\theta}_2)$ be the MLE of (θ_1, θ_2) based on an IID sample $\{x_1, x_2, \cdots, x_n\} \overset{\text{i.i.d}}{\sim} f(x, \theta_1, \theta_2)$. The asymptotic sampling distribution of the MLE is assumed to be

$$\sqrt{n}\begin{pmatrix}\hat{\theta}_1-\theta_1\\ \hat{\theta}_2-\theta_2\end{pmatrix} \overset{\text{approx}}{\to_p} \mathcal{N}_2\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}\right],$$

which can be re-expressed in the following form

$$\begin{pmatrix}\hat{\theta}_1\\ \hat{\theta}_2\end{pmatrix} \overset{\text{approx}}{\to_p} \mathcal{N}_2\left[\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}, \frac{1}{n}\begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}\right],$$

where

$$\text{var}(\hat{\theta}_1) = \frac{\sigma_1^2}{n}, \quad \text{var}(\hat{\theta}_2) = \frac{\sigma_2^2}{n}, \quad \text{and} \quad \text{cov}(\hat{\theta}_1, \hat{\theta}_2) = \frac{\sigma_{12}}{n}.$$

Example (continued): We can re-express the asymptotic normality of the MLE in the example in the previous section as

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \overset{\text{approx}}{\to_p} \mathcal{N}_2 \left[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \frac{1}{45} \begin{pmatrix} 0.2631785 & -0.2487531 \\ -0.2487531 & 0.4149552 \end{pmatrix} \right],$$

this implies

$$\text{var}(\hat{\alpha}) = \frac{0.2631785}{45} = 0.005848411 \quad \text{and} \quad \text{var}(\hat{\beta}) = \frac{0.4149552}{45} = 0.009221227,$$

and

$$\text{cov}(\hat{\alpha}, \hat{\beta}) = -\frac{0.2487531}{45} = 0.005527847.$$

Recall that the MLE of (α, β) in the example are $\hat{\alpha} = 0.3377271$ and $\hat{\beta} = 0.8896159$.

For convenience, we only construct 95% confidence interval for θ_1.

- **Two-sided confidence interval**

$$\hat{\theta_1} \pm z_{0.975} \frac{\hat{\sigma_1}}{\sqrt{n}}$$

In the above **example**, the 95% confidence interval for α is

$$\hat{\alpha}_1 \pm z_{0.975} \sqrt{\frac{\hat{\sigma}_1^2}{45}} = 0.3377271 \pm 1.96 \times \sqrt{\frac{0.2631785}{45}} = (0.1878363, 0.4876179).$$

- **Lower and Upper Confidence Intervals**

The 95% lower and upper confidence intervals for α are given respectively by

$$\left(-\infty, \hat{\theta}_1 + z_{0.95} \frac{\hat{\sigma}_1}{\sqrt{n}} \right) \quad \text{and} \quad \left(\hat{\theta}_1 - z_{0.95} \frac{\hat{\sigma}_1}{\sqrt{n}}, \infty \right)$$

- **Linear Confidence Intervals**

The linear combination of θ_1 and θ_2, denoted by $\theta_0 = a\theta_1 + b\theta_2$ can be estimated by $\hat{\theta}_0 = a\hat{\theta}_1 + b\hat{\theta}_2$. By the **plugin Principle of MLE**, $\hat{\theta}_0$ is also a normal distribution. The $E[\hat{\theta}_0] = aE[\hat{\theta}_1] + bE[\hat{\theta}_2] = a\mu_1 + b\mu_2$, and $\text{var}(\hat{\theta}_0) = a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\sigma_{12}$. Then the two-sided 95% confidence interval is given by

$$\hat{\theta}_0 \pm z_{0.975} \sqrt{\text{var}(\hat{\theta}_0)}.$$

We can also construct one-sided linear confidence intervals similarly.

5.2 Significance Tests

Significance tests are common in practice. For example, after we fit Weibull distribution $f(x, \alpha, \beta) = \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}$ to a data set, we need to assess whether the model was overfitting. If $\beta = 1$, the Weibull distribution reduces to a one-parameter exponential distribution. According to the **Principle of Parsimony**, we should stay with the exponential distribution of the data set. Testing the following hypothesis addresses the above potential overfit/underfit problem.

$$\mathbb{H}_0: \quad \beta = 1 \quad \text{v.s.} \quad \mathbb{H}_a: \quad \beta \neq 1.$$

The above test is a typical significance test. To perform the above hypothesis test, we need to know the sampling distribution of the MLE of (α, β) which is assumed to be

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \overset{\text{approx}}{\to_p} \mathcal{N}_2 \left[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} \sigma_1^2/n & \sigma_{12}/n \\ \sigma_{12}/n & \sigma_2^2/n \end{pmatrix} \right],$$

Based on the above asymptotic normality, the test statistic for the above hypothesis is defined to be

$$TS = \frac{\hat{\beta} - 1}{\hat{\sigma}_2/\sqrt{n}} \sim N(0, 1).$$

The p-value of the above two-tail test is

$$\text{p-value} = P[Z > |TS|], \quad \text{where} \quad Z \quad \text{is the standard normal random variable.}$$

If the p-value is less than a threshold (for example 0.05), the null hypothesis is rejected. The one-parameter exponential distribution should be used.

Example (continued):

In the above numerical example assuming Weibull distribution, we want to see whether an exponential has a better fit. This is equivalent to testing

$$\mathbb{H}_0: \quad \beta = 1 \quad \text{v.s.} \quad \mathbb{H}_a: \quad \beta \neq 1.$$

The test statistic is

$$TS = \frac{0.8896178 - 1}{\sqrt{0.4149563/45}} = -1.149487$$

The p-value is

$$\text{p-value} = P(Z > |-1.149487|) = 2 \times P(Z > -1.149487) \approx 0.2503552.$$

The null hypothesis is **rejected**. This implies that Weibull distribution is more appropriate than exponential.

Remarks:

1. The above test is based on the assumption of a large sample.
2. Since TS is a standard normal distribution. TS^2 is a χ_1^2 distribution. That is,

$$W = \left[\frac{\hat{\beta} - 1}{\hat{\sigma}_2/\sqrt{n}}\right]^2 \sim \chi^2_1.$$

W is called **Wald** statistic. The test is called the **Wald Test**.

Example (continued):

$$W = \left[\frac{\hat{\beta} - 1}{\sqrt{\hat{\sigma}^2_2/n}}\right]^2 = \left[\frac{0.8896178 - 1}{\sqrt{0.4149563/45}}\right]^2 = (-1.149487)^2 = 1.32132.$$

The p-value based on the above statistic is

$$\text{p-value} = P(\chi^2_1 > 1.32132) \approx 0.2503553.$$

Remark: All χ^2 tests are right-tailed!

5.3 Score Test

Consider the null hypothesis

$$H_0 : \quad \theta = \theta_0.$$

which defines a **restricted parameter space**. We also need the *restricted MLE* before defining the test statistic. Let

$$\theta_{\text{rMLE}} = \arg\max_{\theta \in \Theta_R} \log L(\theta : \mathbf{x}).$$

We have shown at the beginning of the section that

$$E\left[\frac{\partial}{\partial \theta} \log L(\theta : \mathbf{x})\right] = \mathbf{0},$$

and

$$\mathbb{I}_n(\theta) = -E\left[\left(\frac{\partial}{\partial \theta} \log L(\theta, \mathbf{x})\right)\left(\frac{\partial}{\partial \theta} \log L(\theta, \mathbf{x})\right)^T\right]$$

We also showed that

$$\text{var}\left[\frac{\partial}{\partial \theta} \log L(\theta : \mathbf{x})\right] = \mathbb{I}_n(\theta)$$

and

$$\frac{\partial}{\partial \theta} \log L(\theta : \mathbf{x}) \to_p \mathcal{N}_k\left[\mathbf{0}, \mathbb{I}_n(\theta)\right]$$

where $k = \dim(\Theta) - \dim(\Theta_R)$.

Theorem: with the above notations and some regularity conditions, the following asymptotic normality holds.

$$S_n = \left[\frac{\partial}{\partial\theta}\log L(\theta_{\text{rMLE}}, \mathbf{x})\right]^T \mathbb{I}_n^{-1}(\theta_{\text{rMLE}}) \left[\frac{\partial}{\partial\theta}\log L(\theta_{\text{rMLE}}, \mathbf{x})\right] \to_p \chi_k^2,$$

where θ_{rMLE} is the restricted MLE under the null hypothesis.

Example (continued): The Weibull example revisited. We still consider

$$H_0: \quad \beta = 1 \quad \text{vs} \quad H_a: \beta \neq 1.$$

Under $H_0: \quad \beta = 1$, the Weibull distribution is reduced to the exponential distribution with density $f(x) = \alpha e^{-\alpha x}$. Therefore, the restricted MLE is $\theta_{\text{rMLE}} = (\hat{\alpha}, \beta = 1) = (1/\bar{x}, 1)$.

To define the score test statistic, we need to find the gradient vector and the Fisher information matrix based on the unrestricted parameter space.

The log-likelihood function is

$$l(\alpha, \beta) = n[\log(\alpha) + \log(\beta)] + (\beta - 1)\sum_{i=1}^n \log(x_i) - \alpha\sum_{i=1}^n x_i^\beta$$

The score functions are given by

$$\begin{cases} \frac{\partial l(\alpha,\beta)}{\partial\alpha} = \frac{n}{\alpha} - \sum_{i=1}^n x_i^\beta, \\ \frac{\partial l(\alpha,\beta)}{\partial\beta} = \frac{n}{\beta} + \sum_{i=1}^n \log(x_i) - \alpha\sum_{i=1}^n x_i^\beta \log(x_i). \end{cases}$$

```
x= c(0.2, 0.3, 0.5, 0.5, 0.5, 0.5, 0.6, 0.6, 0.7, 0.7, 0.7, 0.8, 0.8, 1.0, 1.0, 1.0, 1.0,
     1.1, 1.3, 1.5, 1.5, 1.5, 1.5, 2.0, 2.0, 2.2, 2.5, 3.0, 3.0, 3.3, 3.3, 4.0, 4.0, 4.5,
     4.7, 5.0, 5.4, 5.4, 7.0, 7.5, 8.8, 9.0, 10.3, 22.0, 24.5)
n = length(x)
##
lb = n + sum(log(x)) - (1/mean(x))*sum(x*log(x))
lb
```

```
[1] -11.12719
```

The Score vector is

$$U(\theta_{\text{rMLE}}) = [0, -11.12719]$$

and

$$\frac{\partial^2}{\partial\alpha^2}\log L(\alpha : \mathbf{x}) = -\frac{n}{\alpha^2}, \quad \text{and} \quad \frac{\partial^2}{\partial\alpha\partial\beta}\log L(\alpha : \mathbf{x}) = -\sum_{i=1}^n x_i^\beta \log(x_i)$$

$$\frac{\partial^2}{\partial\beta^2}\log L(\alpha : \mathbf{x}) = -\frac{n}{\beta^2} - \sum_{i=1}^n x_i^\beta \log(x_i)$$

```r
## Inverse observed Fisher Information matrix
x= c(0.2, 0.3, 0.5, 0.5, 0.5, 0.5, 0.6, 0.6, 0.7, 0.7, 0.7, 0.8, 0.8, 1.0, 1.0, 1.0, 1.0,
     1.1, 1.3, 1.5, 1.5, 1.5, 1.5, 2.0, 2.0, 2.2, 2.5, 3.0, 3.0, 3.3, 3.3, 4.0, 4.0, 4.5,
     4.7, 5.0, 5.4, 5.4, 7.0, 7.5, 8.8, 9.0, 10.3, 22.0, 24.5)
n = length(x)
##
laa = -n*mean(x)
lab = -sum(x*log(x))
lbb = -n-sum(x*log(x))
IIn = matrix(c(laa, lab, lab, lbb), nrow=2)
IIn.inv = solve(IIn)
IIn.inv
```

```
             [,1]         [,2]
[1,]  0.009319889 -0.008137795
[2,] -0.008137795  0.004287062
```

The observed Fisher information matrix based on the restricted MLE is given by

$$\mathbb{I}_n^{-1}(\theta_{\text{rMLE}}) = \begin{bmatrix} 0.009319889 & -0.008137795 \\ -0.008137795 & 0.004287062 \end{bmatrix}$$

```r
## Score test statistic
Sn=c(0, -11.12719)%*% IIn.inv %*% c( 0, -11.12719)
Sn
```

```
          [,1]
[1,] 0.5307998
```

The score test statistic is

$$S_n = [0, -11.12719] \begin{bmatrix} 0.009319889 & -0.008137795 \\ -0.008137795 & 0.004287062 \end{bmatrix} \begin{bmatrix} 0 \\ -11.12719 \end{bmatrix} \approx 0.5307998.$$

Since $S_n \to_p \chi_1^2$. The p-value of the score test is given by

$$\text{p-value} = P[\chi_1^2 > 0.5307998] = 0.4662708.$$

The null hypothesis $H_0 : \beta = 1$ is not rejected. This is consistent with the Wald χ^2 test.

5.4 Likelihood Ratio Test (LRT)

The likelihood ratio (LR) test

The likelihood ratio test is one of the most commonly used in likelihood-based statistical inferences. It is a test of hypothesis in which two different maximum likelihood estimates of a parameter are compared in order to decide whether to reject or not to reject a **restriction on the parameter** (such as setting $\beta = 1$ in the previous Weibull distribution).

This note focuses on parametric likelihood inferences. The LRT is used to compare two nested models. The basic setup is outlined in the following.

In essence, the LRT is based on two MLEs from two parameter spaces: **full parameter space** (Θ) and **restricted parameter space** Θ_R.

- **Full (Unrestricted) Parameter Space**: The parameter space containing all values that the parameter vector can take. For example, for normal distribution with density $N(\mu, \sigma)$, $\mu \in \mathbb{R}$ and standard deviation $\sigma \in \mathbb{R}^+$, then $\Theta \overset{\text{def}}{\equiv} \mathbb{R} \times \mathbb{R}^+$ is the parameter space of $\theta = c(\mu, \sigma)$.
- **Restricted Parameter Space**: A subspace of the unrestricted space. For example, in the 2-parameter Weibull distribution, the parameter space is $\Theta = \mathbb{R}^+ \times \mathbb{R}^+$ because both parameters (α, β) are positive. That is, Θ is spanned by (α, β). If we set $\beta = 1$, $(\alpha, 1)$ spans the restricted parameter space $\Theta_R = \mathbb{R}^+ \times \{1\}$.

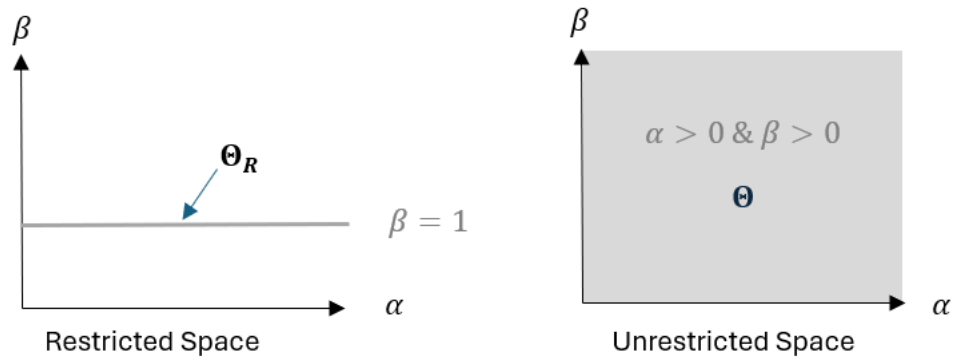

Figure 1: Parameter spaces of 2-parameter Weibull distribution

- **Unrestricted and Restricted Maximum Likelihood**: The unrestricted maximum likelihood is based on the unrestricted parameter space (Θ) and the restricted maximum likelihood is based on the restricted parameter space (Θ_R). The unrestricted and restricted MLEs are given by

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \log L(\theta : \mathbf{x})$$

and

$$\hat{\theta}_{rMLE} = \arg\max_{\theta \in \Theta_R} \log L(\theta : \mathbf{x})$$

With the above notations, we state the following theorem. The proof is out of the scope of this series of tutorials.

Theorem: Let $\hat{\theta}_{\text{MLE}}$ and $\hat{\theta}_{\text{rMLE}}$ be the unrestricted MLE and restricted MLE estimated based on Θ and Θ_R respectively. Denote $\text{df} = \dim(\Theta) - \dim(\Theta_R)$, then the test statistic for testing the null hypothesis

$$H_0 : \theta_0 \in \Theta_R$$

is defined to be

$$LRT = -2 \ln \frac{L(\hat{\theta}_{rMLE} : \mathbf{x})}{L(\hat{\theta}_{MLE} : \mathbf{x})} \rightarrow_p \chi^2_{\text{df.}}$$

The p-value of the above likelihood ratio χ^2 test is determined by

$$\text{p-value} = P(\chi^2_{\text{df}} > LRT).$$

Recap of LRT

The likelihood ratio χ^2 test involves three technical steps:

1. Find the MLE of the parameters in the restricted parameter space ($\not\leq_R$) which is defined based on the null hypothesis H_0.
2. Find the MLE of the parameters in the unrestricted parameter space ($\not\leq$).
3. Evaluate the likelihood ratio statistic and find the p-value based on the resulting χ^2_{df} with df $=$ $\dim(\not\leq) - \dim(\not\leq_R)$.

Example (Weibull example revisited): We will follow the above three technical steps for testing the null hypothesis

$$H_0 : \beta = 1 \quad \text{vs} \quad H_a : \beta \neq 1.$$

Note that H_0 defines the restricted space $\not\leq_R = \mathbb{R}^+ \times 1$. We need to find a value from all possible values of α given $\beta = 1$ that maximizes the log-likelihood of observing the data.

Step 1: Find the restricted MLE is the MLE of exponential distribution with density $f(x, \alpha) = \alpha e^{-\alpha x}$.

We know the solution to the optimization problem has a closed form $\hat{\alpha} = n/\sum_{i=1}^{n} x_i = 45/163.2 \approx 0.2757353$. That is, restricted MLE $\theta_{rMLE} = 0.2757353$. The log-likelihood evaluated at the restricted MLE is given by

$$\log L(\hat{\theta}_{rMLE}, \mathbf{x}) = n \log(a) - a \sum_{i=1}^{45} x_i$$

$$= 45 \log(0.2757353) - 0.2757353 \times 163.2 = -102.9741.$$

The log-likelihood curve and its critical point is given in the figure below.

```
# Data set
x = c(0.2, 0.3, 0.5, 0.5, 0.5, 0.5, 0.6, 0.6, 0.7, 0.7, 0.7, 0.8, 0.8, 1.0, 1.0, 1.0, 1.0,
      1.1, 1.3, 1.5, 1.5, 1.5, 1.5, 2.0, 2.0, 2.2, 2.5, 3.0, 3.0, 3.3, 3.3, 4.0, 4.0, 4.5,
      4.7, 5.0, 5.4, 5.4, 7.0, 7.5, 8.8, 9.0, 10.3, 22.0, 24.5)
n = length(x)
## log-likelihood
negLogLik = function(a){
  b = 1
  n*log(a) - a*sum(x^b)
}
## gradient function
## We first plot the log-likelihood function
alpha = seq(0.01, 1, length=100)
lglik = negLogLik(alpha)
plot(alpha, lglik, type="l",  ylab ="Loglikelihood")
abline(v=0.2757353, col="red")
points(0.2757353, negLogLik(0.2757353), pch=19, col = "red", cex = 1.5)
text(0.5, -150, "(0.2757, -103.0)", col="blue")
arrows(0.2757, negLogLik(0.2757353), 0.5, -145, length = 0.05, angle = 30,
       code = 2)
```

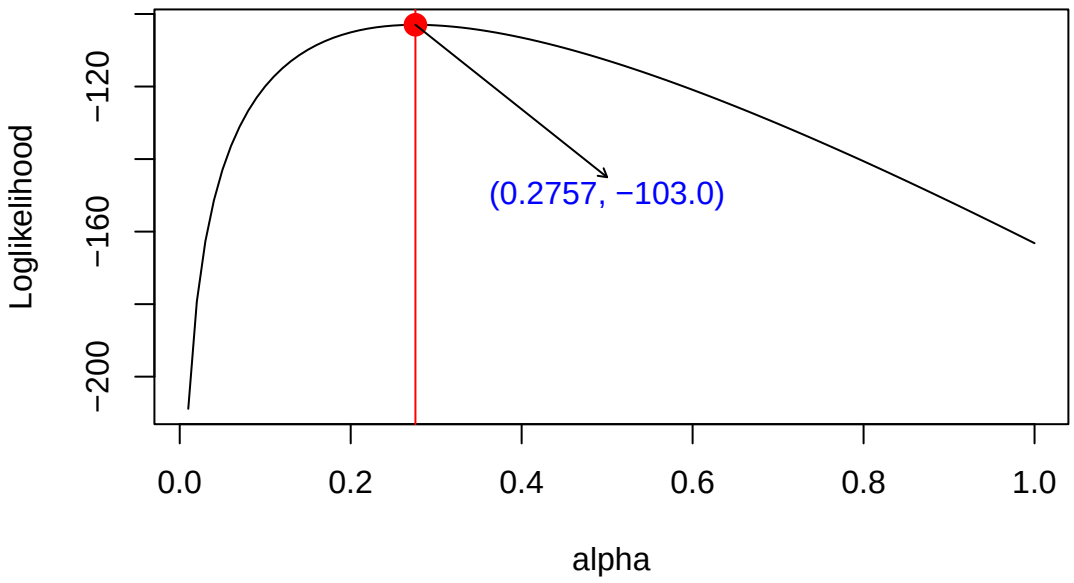

Figure 2: log-likelihood curve with critical point labeled in red.

Step 2: log-likelihood evaluated at unrestricted MLE.

The unrestricted MLE has found in the previous example $\hat{\theta}_{MLE} = (\hat{\alpha}, \hat{\beta}) \approx (0.3377271, 0.8896159)$. The log-likelihood evaluated at $(\hat{\alpha}, \hat{\beta}) \approx (0.3377271, 0.8896159)$ is

$$\log L(\hat{\theta} : \mathbf{x}) = n \log(\alpha) + n \log(\beta) + (\beta - 1) \sum_{i=1}^{n} \log(x_i) - \alpha \sum_{i=1}^{n} (x_i^{\beta})$$

```
# Data set
x = c(0.2, 0.3, 0.5, 0.5, 0.5, 0.5, 0.6, 0.6, 0.7, 0.7, 0.7, 0.8, 0.8, 1.0, 1.0, 1.0, 1.0,
      1.1, 1.3, 1.5, 1.5, 1.5, 1.5, 2.0, 2.0, 2.2, 2.5, 3.0, 3.0, 3.3, 3.3, 4.0, 4.0, 4.5,
      4.7, 5.0, 5.4, 5.4, 7.0, 7.5, 8.8, 9.0, 10.3, 22.0, 24.5)
n = length(x)
## log-likelihood
negLogLik = function(A){
  a = A[1]
  b = A[2]
  n*log(a) + n*log(b) + (b-1)*sum(log(x)) - a*sum(x^b)
}
# 0.3377259 0.8896178
A = c(0.3377259, 0.8896178)
lglik = negLogLik(A)
lglik
```

```
[1] -102.3452
```

Step 3: Evaluating the LRT and finding the p-value.

$$LRT = -2\log\frac{L(\hat{\theta}_{rMLE} : \mathbf{x})}{L(\hat{\theta}_{MLE} : \mathbf{x})} = -2\left[\log L(\hat{\theta}_{rMLE} : \mathbf{x}) - \log L(\hat{\theta}_{MLE} : \mathbf{x})\right] \to_p \chi^2_1.$$

The value of the above test statistic is

$$LRT = -2[-102.9741 - (-102.3452)] = 1.2578.$$

The p-value of the LRT test is calculated in the following

$$\text{p-value} = P[\chi^2_1 > 1.2578] = 0.7379321.$$

This implies that there is an overfit issue if we choose the two-parameter Weibull distribution to fit the given data. The one-parameter exponential is a better choice for this application.

5.5 Concluding Remarks

In this section, we introduced three major large sample tests associated with likelihood estimators: Wald, Score, and likelihood ratio tests.

Three chi-square tests are asymptotically equivalent. However, this potentially gives the wrong impression that all three approaches are equally accurate regarding approximation inference.

Many studies in theory and simulation have shown that the likelihood ratio approach is the most accurate among the three large-sample tests.

The Score and Wald approaches depend on derivatives, and thus, can change substantially if we reparameterize the model. The likelihood ratio test is recommended in practice.

The likelihood ratio test requires restricted and unrestricted MLEs that defined based on the restricted and unrestricted parameter spaces. This means that the LRT compares only two nested models or distributions.