Math 541: Statistical Theory II

Fisher Information and Cramér-Rao Bound

Instructor: Songfeng Zheng

In the parameter estimation problems, we obtain information about the parameter from a sample of data coming from the underlying probability distribution. A natural question is: how much information can a sample of data provide about the unknown parameter? This section introduces such a measure for information, and we can also see that this information measure can be used to find bounds on the variance of estimators, and it can be used to approximate the sampling distribution of an estimator obtained from a large sample, and further be used to obtain an approximate confidence interval in case of large sample.

In this section, we consider a random variable X for which the pdf or pmf is $f(x|\theta)$, where θ is an unknown parameter and $\theta \in \Theta$, with Θ is the parameter space.

1 Fisher Information

Motivation: Intuitively, if an event has small probability, then the occurrence of this event brings us much information. For a random variable $X \sim f(x|\theta)$, if θ were the true value of the parameter, the likelihood function should take a big value, or equivalently, the derivative log-likelihood function should be close to zero, and this is the basic principle of maximum likelihood estimation. We define $l(x|\theta) = \log f(x|\theta)$ as the log-likelihood function, and

$$l'(x|\theta) = \frac{\partial}{\partial\theta} \log f(x|\theta) = \frac{f'(x|\theta)}{f(x|\theta)}$$

How much the output changes w.r.t the parameters

where $f'(x|\theta)$ is the derivative of $f(x|\theta)$ with respect to θ. Similarly, we denote the second order derivative of $f(x|\theta)$ with respect to θ as $f''(x|\theta)$.

The output would be dictated by the input, not the parameters, so there's really no info to gain

You can also look at this as the output not rly changing with a change to the parameters

According to the above analysis, if $l'(X|\theta)$ is close to zero, then it is expected, thus the random variable does not provide much information about θ; on the other hand, if $|l'(X|\theta)|$ or $[l'(X|\theta)]^2$ is large, the random variable provides much information about θ. Thus, we can use $[l'(X|\theta)]^2$ to measure the amount of information provided by X. However, since X is a random variable, we should consider the average case. Thus, we introduce the following definition:

Fisher information (for θ) contained in the random variable X is defined as:

$$I(\theta) = E_\theta\left\{[l'(X|\theta))]^2\right\} = \int [l'(x|\theta))]^2 f(x|\theta)dx. \quad (1)$$

The expectation of the squared change of the log output w.r.t. the parameters

What is the expected squared derivative of the function log w.r.t. to the parameters?

We assume that we can exchange the order of differentiation and integration, then

$$\int f'(x|\theta)dx = \frac{\partial}{\partial\theta}\int f(x|\theta)dx = 0$$

Similarly,

$$\int f''(x|\theta)dx = \frac{\partial^2}{\partial\theta^2}\int f(x|\theta)dx = 0$$

It is easy to see that

$$E_\theta[l'(X|\theta)] = \int l'(x|\theta)f(x|\theta)dx = \int \frac{f'(x|\theta)}{f(x|\theta)}f(x|\theta)dx = \int f'(x|\theta)dx = 0$$

Therefore, the definition of Fisher information (1) can be rewritten as

$$I(\theta) = \mathrm{Var}_\theta\left[l'(X|\theta)\right] \tag{2}$$

Also, notice that

$$l''(x|\theta) = \frac{\partial}{\partial\theta}\left[\frac{f'(x|\theta)}{f(x|\theta)}\right] = \frac{f''(x|\theta)f(x|\theta) - [f'(x|\theta)]^2}{[f(x|\theta)]^2} = \frac{f''(x|\theta)}{f(x|\theta)} - [l'(x|\theta)]^2$$

Therefore,

$$E_\theta[l''(x|\theta)] = \int\left[\frac{f''(x|\theta)}{f(x|\theta)} - [l'(x|\theta)]^2\right]f(x|\theta)dx = \int f''(x|\theta)dx - E_\theta\left\{[l'(X|\theta)]^2\right\} = -I(\theta)$$

Finally, we have another formula to calculate Fisher information:

$$I(\theta) = -E_\theta[l''(x|\theta)] = -\int\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right]f(x|\theta)dx \tag{3}$$

negative expectation of the 2nd order derivative of the log of the function output w.r.t. the parameters
the negative expectation of the change's change of the log output w.r.t. changing the parameters

To summarize, we have three methods to calculate Fisher information: equations (1), (2), and (3). In many problems, using (3) is the most convenient choice.

Example 1: Suppose random variable X has a Bernoulli distribution for which the parameter θ is unknown ($0 < \theta < 1$). We shall determine the Fisher information $I(\theta)$ in X.

The point mass function of X is

$$f(x|\theta) = \theta^x(1-\theta)^{1-x} \quad \text{for } x = 1 \text{ or } x = 0.$$

Therefore

$$l(x|\theta) = \log f(x|\theta) = x\log\theta + (1-x)\log(1-\theta)$$

and

$$l'(x|\theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta} \quad \text{and} \quad l''(x|\theta) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}$$

Since $E(X) = \theta$, the Fisher information is

$$I(x|\theta) = -E[l''(x|\theta)] = \frac{E(X)}{\theta^2} + \frac{1-E(X)}{(1-\theta)^2} = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}$$

Example 2: Suppose that $X \sim N(\mu, \sigma^2)$, and μ is unknown, but the value of σ^2 is given. find the Fisher information $I(\mu)$ in X.

Imagine like knowing the values for all parameters in a neural net, but we need to find the fisher information for a single unknown parameter

For $-\infty < x < \infty$, we have

$$l(x|\mu) = \log f(x|\mu) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

Hence,

$$l'(x|\mu) = \frac{x-\mu}{\sigma^2} \quad \text{and} \quad l''(x|\mu) = -\frac{1}{\sigma^2}$$

It follows that the Fisher information is

$$I(\mu) = -E[l''(x|\mu)] = \frac{1}{\sigma^2}$$

When you sample a Normal distribution when you know the variance, but not the mean, the Fisher info says you should expect that the change in the change of the log output should decrease by 1/(variance)

This tells me that the change in the change of log output w.r.t. changing the mean parameter, should decrease by 1/(varience) every time you resample the function

This makes sense, as when you change the mu parameter, you should be able to understand a small amount of info about the distribution

If we make a transformation of the parameter, we will have different expressions of Fisher information with different parameterization. More specifically, let X be a random variable for which the pdf or pmf is $f(x|\theta)$, where the value of the parameter θ is unknown but must lie in a space Θ. Let $I_0(\theta)$ denote the Fisher information in X. Suppose now the parameter θ is replaced by a new parameter μ, where $\theta = \phi(\mu)$, and ϕ is a differentiable function. Let $I_1(\mu)$ denote the Fisher information in X when the parameter is regarded as μ. We will have $I_1(\mu) = [\phi'(\mu)]^2 I_0[\phi(\mu)]$.

Proof: Let $g(x|\mu)$ be the p.d.f. or p.m.f. of X when μ is regarded as the parameter. Then $g(x|\mu) = f[x|\phi(\mu)]$. Therefore,

$$\log g(x|\mu) = \log f[x|\phi(\mu)] = l[x|\phi(\mu)],$$

and

$$\frac{\partial}{\partial\mu}\log g(x|\mu) = l'[x|\phi(\mu)]\phi'(\mu).$$

It follows that

$$I_1(\mu) = E\left\{\left[\frac{\partial}{\partial\mu}\log g(X|\mu)\right]^2\right\} = [\phi'(\mu)]^2 E\left(\{l'[X|\phi(\mu)]\}^2\right) = [\phi'(\mu)]^2 I_0[\phi(\mu)]$$

This will be verified in exercise problems.

Suppose that we have a random sample $X_1, \cdots, X_n$ coming from a distribution for which the pdf or pmf is $f(x|\theta)$, where the value of the parameter θ is unknown. Let us now calculate the amount of information the random sample $X_1, \cdots, X_n$ provides for θ.

Let us denote the joint pdf of $X_1, \cdots, X_n$ as

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

then

$$l_n(\mathbf{x}|\theta) = \log f_n(\mathbf{x}|\theta) = \sum_{i=1}^{n} \log f(x_i|\theta) = \sum_{i=1}^{n} l(x_i|\theta).$$

and

$$l_n'(\mathbf{x}|\theta) = \frac{f_n'(\mathbf{x}|\theta)}{f_n(\mathbf{x}|\theta)} \tag{4}$$

We define the Fisher information $I_n(\theta)$ in the random sample $X_1, \cdots, X_n$ as

$$I_n(\theta) = E_\theta \left\{ [l_n'(\mathbf{X}|\theta)]^2 \right\} = \int \cdots \int [l_n'(X|\theta)]^2 f_n(\mathbf{x}|\theta) dx_1 \cdots dx_n$$

which is an n-dimensional integral. We further assume that we can exchange the order of differentiation and integration, then we have

$$\int f_n'(\mathbf{x}|\theta) dx = \frac{\partial}{\partial \theta} \int f_n(\mathbf{x}|\theta) dx = 0$$

and,

$$\int f_n''(\mathbf{x}|\theta) dx = \frac{\partial^2}{\partial \theta^2} \int f_n(\mathbf{x}|\theta) dx = 0$$

It is easy to see that

$$E_\theta[l_n'(\mathbf{X}|\theta)] = \int l_n'(\mathbf{x}|\theta) f_n(\mathbf{x}|\theta) d\mathbf{x} = \int \frac{f_n'(\mathbf{x}|\theta)}{f_n(\mathbf{x}|\theta)} f_n(\mathbf{x}|\theta) d\mathbf{x} = \int f_n'(\mathbf{x}|\theta) d\mathbf{x} = 0 \tag{5}$$

Therefore, the definition of Fisher information for the sample $X_1, \cdots, X_n$ can be rewritten as

$$I_n(\theta) = \mathrm{Var}_\theta \left[l_n'(\mathbf{X}|\theta) \right].$$

It is similar to prove that the Fisher information can also be calculated as

$$I_n(\theta) = -E_\theta \left[l_n''(\mathbf{X}|\theta) \right].$$

From the definition of $l_n(\mathbf{x}|\theta)$, it follows that

$$l_n''(\mathbf{x}|\theta) = \sum_{i=1}^{n} l''(x_i|\theta).$$

Therefore, the Fisher information

$$I_n(\theta) = -E_\theta\left[l_n''(\mathbf{X}|\theta))\right] = -E_\theta\left[\sum_{i=1}^{n} l''(X_i|\theta)\right] = -\sum_{i=1}^{n} E_\theta\left[l''(X_i|\theta)\right] = nI(\theta).$$

In other words, the Fisher information in a random sample of size n is simply n times the Fisher information in a single observation.

Example 3: Suppose $X_1, \cdots, X_n$ form a random sample from a Bernoulli distribution for which the parameter θ is unknown ($0 < \theta < 1$). Then the Fisher information $I_n(\theta)$ in this sample is

$$I_n(\theta) = nI(\theta) = \frac{n}{\theta(1-\theta)}.$$

Example 4: Let $X_1, \cdots, X_n$ be a random sample from $N(\mu, \sigma^2)$, and μ is unknown, but the value of σ^2 is given. Then the Fisher information $I_n(\theta)$ in this sample is

$$I_n(\mu) = nI(\mu) = \frac{n}{\sigma^2}.$$

2 Cramér-Rao Lower Bound and Asymptotic Distribution of Maximum Likelihood Estimators

Suppose that we have a random sample $X_1, \cdots, X_n$ coming from a distribution for which the pdf or pmf is $f(x|\theta)$, where the value of the parameter θ is unknown. We will show how to used Fisher information to determine the lower bound for the variance of an estimator of the parameter θ.

Let $\hat{\theta} = r(X_1, \cdots, X_n) = r(\mathbf{X})$ be an arbitrary estimator of θ. Assume $E_\theta(\hat{\theta}) = m(\theta)$, and the variance of $\hat{\theta}$ is finite. Let us consider the random variable $l_n'(\mathbf{X}|\theta)$ defined in (4), it was shown in (5) that $E_\theta[l_n'(\mathbf{X}|\theta)] = 0$. Therefore, the covariance between $\hat{\theta}$ and $l_n'(\mathbf{X}|\theta)$ is

$$\begin{aligned}
\mathrm{Cov}_\theta[\hat{\theta}, l_n'(\mathbf{X}|\theta)] &= E_\theta\left\{[\hat{\theta} - E_\theta(\hat{\theta})][l_n'(\mathbf{X}|\theta) - E_\theta(l_n'(\mathbf{X}|\theta))]\right\} = E_\theta\left\{[r(\mathbf{X}) - m(\theta)]l_n'(\mathbf{X}|\theta)\right\} \\
&= E_\theta[r(\mathbf{X})l_n'(\mathbf{X}|\theta)] - m(\theta)E_\theta[l_n'(\mathbf{X}|\theta)] = E_\theta[r(\mathbf{X})l_n'(\mathbf{X}|\theta)] \\
&= \int\cdots\int r(\mathbf{x})l_n'(\mathbf{x}|\theta)f_n(\mathbf{x}|\theta)dx_1\cdots dx_n \\
&= \int\cdots\int r(\mathbf{x})f_n'(\mathbf{x}|\theta)dx_1\cdots dx_n \quad \text{(Use Equation 4)} \\
&= \frac{\partial}{\partial\theta}\int\cdots\int r(\mathbf{x})f_n(\mathbf{x}|\theta)dx_1\cdots dx_n \\
&= \frac{\partial}{\partial\theta}E_\theta[\hat{\theta}] = m'(\theta) \qquad (6)
\end{aligned}$$

By Cauchy-Schwartz inequality and the definition of $I_n(\theta)$,

$$\left\{\mathrm{Cov}_\theta[\hat{\theta}, l_n'(\mathbf{X}|\theta)]\right\}^2 \leq \mathrm{Var}_\theta[\hat{\theta}]\mathrm{Var}_\theta[l_n'(\mathbf{X}|\theta)] = \mathrm{Var}_\theta[\hat{\theta}]I_n(\theta)$$

i.e.,

$$[m'(\theta)]^2 \leq \mathrm{Var}_\theta[\hat{\theta}]I_n(\theta) = nI(\theta)\mathrm{Var}_\theta[\hat{\theta}]$$

Finally, we get the lower bound of variance of an arbitrary estimator $\hat{\theta}$ as

$$\mathrm{Var}_\theta[\hat{\theta}] \geq \frac{[m'(\theta)]^2}{nI(\theta)} \tag{7}$$

The inequality (7) is called the *information inequality*, and also known as the *Cramér-Rao inequality* in honor of the Sweden statistician H. Cramér and Indian statistician C. R. Rao who independently developed this inequality during the 1940s. The information inequality shows that as $I(\theta)$ increases, the variance of the estimator decreases, therefore, the quality of the estimator increases, that is why the quantity is called "information".

If $\hat{\theta}$ is an unbiased estimator, then $m(\theta) = E_\theta(\hat{\theta}) = \theta$, $m'(\theta) = 1$. Hence, by the information inequality, for unbiased estimator $\hat{\theta}$,

$$\mathrm{Var}_\theta[\hat{\theta}] \geq \frac{1}{nI(\theta)}$$

What it is when we have the optimal parameters

The right hand side is always called the Cramér-Rao lower bound (CRLB): under certain conditions, no other unbiased estimator of the parameter θ based on an i.i.d. sample of size n can have a variance smaller than CRLB.

Example 5: Suppose a random sample $X_1, \cdots, X_n$ from a normal distribution $N(\mu, \theta)$, with μ given and the variance θ unknown. Calculate the lower bound of variance for any estimator, and compare to that of the sample variance S^2.

Solution: We know

$$f(x|\theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left\{-\frac{(x-\mu)^2}{2\theta}\right\}$$

then

$$l(x|\theta) = -\frac{(x-\mu)^2}{2\theta} - \frac{1}{2}\log 2\pi - \frac{1}{2}\log\theta.$$

Hence

$$l'(x|\theta) = \frac{(x-\mu)^2}{2\theta^2} - \frac{1}{2\theta},$$

and

$$l''(x|\theta) = -\frac{(x-\mu)^2}{\theta^3} + \frac{1}{2\theta^2}.$$

Therefore,

$$I(\theta) = -E[l''(X|\theta)] = -E\left[-\frac{(X-\mu)^2}{\theta^3} + \frac{1}{2\theta^2}\right] = \frac{1}{2\theta^2},$$

and

$$I_n(\theta) = nI(\theta) = \frac{n}{2\theta^2}.$$

Finally, we have the Cramér-Rao lower bound $\frac{2\theta^2}{n}$.

The sample variance is defined as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

and it is known that

$$\frac{(n-1)S^2}{\theta} \sim \chi^2_{n-1}$$

then,

$$\text{Var}\left(\frac{n-1}{\theta}S^2\right) = \frac{(n-1)^2}{\theta^2}\text{Var}\left(S^2\right) = 2(n-1).$$

Therefore,

$$\text{Var}\left(S^2\right) = \frac{2\theta^2}{n-1} > \frac{2\theta^2}{n}$$

i.e., the variance of the estimator S^2 is bigger than the Cramér-Rao lower bound.

Now, let us consider the MLE $\hat{\theta}$ of θ, to make notations clear, let us assume the true value of θ is θ_0. We shall prove that as the sample size n is very big, the distribution of MLE estimator $\hat{\theta}$ is approximately normal with mean θ_0 and variance $1/[nI(\theta_0)]$. Since this is merely a limiting result, which holds as the sample size tends to infinity, we say that the MLE is **asymptotically unbiased** and refer to the variance of the limiting normal distribution as the **asymptotic variance of the MLE**. More specifically, we have the following theorem:

Theorem (The asymptotic distribution of MLE): Let $X_1, \cdots, X_n$ be a sample of size n from a distribution for which the pdf or pmf is $f(x|\theta)$, with θ the unknown parameter. Assume that the true value of θ is θ_0, and the MLE of θ is $\hat{\theta}$. Then the probability distribution of $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ tends to a standard normal distribution. In other words, the asymptotic distribution of $\hat{\theta}$ is

$$N\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$$

Proof: we shall prove that

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \sim N(0,1)$$

asymptotically. We will only give a sketch of the proof; the details of the argument are beyond the scope of this course.

Remember the log-likelihood function is

$$l(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

and $\hat{\theta}$ is the solution to $l'(\theta) = 0$. We apply Tylor expansion of $l'(\hat{\theta})$ at the point θ_0, yielding

$$0 = l'(\hat{\theta}) \approx l'(\theta_0) + (\hat{\theta} - \theta_0) l''(\theta_0)$$

Therefore,

$$\hat{\theta} - \theta_0 \approx \frac{-l'(\theta_0)}{l''(\theta_0)}$$

and

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx \frac{-n^{-1/2} l'(\theta_0)}{n^{-1} l''(\theta_0)}$$

First, let us consider the numerator of the last expression above. Its expectation is

$$E[-n^{-1/2} l'(\theta_0)] = n^{-1/2} \sum_{i=1}^{n} E\left[\frac{\partial}{\partial \theta} \log f(X_i|\theta_0)\right] = n^{-1/2} \sum_{i=1}^{n} E\left[l'(X_i|\theta_0)\right] = 0,$$

and its variance is

$$\mathrm{Var}[-n^{-1/2} l'(\theta_0)] = \frac{1}{n} \sum_{i=1}^{n} E\left[\frac{\partial}{\partial \theta} \log f(X_i|\theta_0)\right]^2 = \frac{1}{n} \sum_{i=1}^{n} E\left[l'(X_i|\theta_0)\right]^2 = I(\theta_0).$$

Next, we consider the denominator:

$$\frac{1}{n} l''(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f(X_i|\theta_0)$$

By the law of large number, this expression converges to

$$E\left[\frac{\partial^2}{\partial \theta^2} \log f(X_i|\theta_0)\right] = -I(\theta_0)$$

We thus have

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx \frac{n^{-1/2} l'(\theta_0)}{I(\theta_0)}$$

Therefore,

$$E\left[\sqrt{n}(\hat{\theta} - \theta_0)\right] \approx \frac{E[n^{-1/2} l'(\theta_0)]}{I(\theta_0)} = 0,$$

and

$$\mathrm{Var}\left[\sqrt{n}(\hat{\theta} - \theta_0)\right] \approx \frac{\mathrm{Var}[n^{-1/2} l'(\theta_0)]}{I^2(\theta_0)} = \frac{I(\theta_0)}{I^2(\theta_0)} = \frac{1}{I(\theta_0)}$$

As $n \to \infty$, applying central limit theorem, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \sim N\left(0, \frac{1}{I(\theta_0)}\right)$$

i.e.,

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \sim N(0, 1).$$

This completes the proof.

This theorem indicates the asymptotic optimality of maximum likelihood estimator since the asymptotic variance of MLE can achieve the CRLB. For this reason, MLE is frequently used especially with large samples.

Example 6: Suppose that $X_1, X_2, \cdots, X_n$ are i.i.d. random variables on the interval $[0, 1]$ with the density function

$$f(x|\alpha) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2}[x(1-x)]^{\alpha-1}$$

where $\alpha > 0$ is a parameter to be estimated from the sample. It can be shown that

$$E(X) = \frac{1}{2}$$

$$Var(X) = \frac{1}{4(2\alpha+1)}$$

What is the asymptotic variance of the MLE?

Solution. Let's calculate $I(\alpha)$: Firstly,

$$\log f(x|\alpha) = \log\Gamma(2\alpha) - 2\log\Gamma(\alpha) + (\alpha-1)\log[x(1-x)]$$

Then,

$$\frac{\partial \log f(x|\alpha)}{\partial\alpha} = \frac{2\Gamma'(2\alpha)}{\Gamma(2\alpha)} - \frac{2\Gamma'(\alpha)}{\Gamma(\alpha)} + \log[x(1-x)]$$

and

$$\begin{aligned}\frac{\partial^2 \log f(x|\alpha)}{\partial\alpha^2} &= \frac{2\Gamma''(2\alpha)2\Gamma(2\alpha) - 2\Gamma'(2\alpha)2\Gamma'(2\alpha)}{\Gamma(2\alpha)^2} - \frac{2\Gamma''(\alpha)\Gamma(\alpha) - 2\Gamma'(\alpha)\Gamma'(\alpha)}{\Gamma(\alpha)^2} \\ &= \frac{4\Gamma''(2\alpha)\Gamma(2\alpha) - (2\Gamma'(2\alpha))^2}{\Gamma(2\alpha)^2} - \frac{2\Gamma''(\alpha)\Gamma(\alpha) - 2(\Gamma'(\alpha))^2}{\Gamma(\alpha)^2}\end{aligned}$$

Therefore,

$$I(\alpha) = -E\left(\frac{\partial^2 \log f(x|\alpha)}{\partial\alpha^2}\right) = \frac{2\Gamma''(\alpha)\Gamma(\alpha) - 2(\Gamma'(\alpha))^2}{\Gamma^2(\alpha)} - \frac{4\Gamma''(2\alpha)\Gamma(2\alpha) - (2\Gamma'(2\alpha))^2}{\Gamma^2(2\alpha)}$$

The asymptotic variance of the MLE is $\frac{1}{nI(\alpha)}$.

Example 7: The Pareto distribution has been used in economics as a model for a density function with a slowly decaying tail:

$$f(x|x_0,\theta) = \theta x_0^{\theta} x^{-\theta-1}, \quad x \geq x_0, \quad \theta > 1$$

Assume that $x_0 > 0$ is given and that $X_1, X_2, \cdots, X_n$ is an i.i.d. sample. Find the asymptotic distribution of the mle.

Solution: The asymptotic distribution of $\hat{\theta}_{MLE}$ is $N\left(\theta, \frac{1}{nI(\theta)}\right)$. Let's calculate $I(\theta)$.

Firstly,

$$\log f(x|\theta) = \log\theta + \theta\log x_0 - (\theta+1)\log x$$

Then,

$$\frac{\partial \log f(x|\theta)}{\partial\theta} = \frac{1}{\theta} + \log x_0 - \log x$$

and

$$\frac{\partial^2 \log f(x|\theta)}{\partial\theta^2} = -\frac{1}{\theta^2}$$

So,

$$I(\theta) = -E\left[\frac{\partial^2 \log f(x|\theta)}{\partial\theta^2}\right] = \frac{1}{\theta^2}$$

Therefore, the asymptotic distribution of MLE is

$$N\left(\theta, \frac{1}{nI(\theta)}\right) = N\left(\theta, \frac{\theta^2}{n}\right)$$

3 Approximate Confidence Intervals

In previous lectures, we discussed the exact confidence intervals. However, to construct an exact confidence interval requires detailed knowledge of the sampling distribution as well as some cleverness. An alternative method of constructing confidence intervals is based on the large sample theory of the previous section.

According to the large sample theory result, the distribution of $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ is approximately the standard normal distribution. Since the true value of θ, θ_0, is unknown, we will use the estimated value $\hat{\theta}$ to estimate $I(\theta_0)$. It can be further argued that the distribution of $\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0)$ is also approximately standard normal. Since the standard normal distribution is symmetric about 0,

$$P\left(-z(1-\alpha/2) \leq \sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0) \leq z(1-\alpha/2)\right) \approx 1 - \alpha$$

Manipulation of the inequalities yields

$$\hat{\theta} - z(1-\alpha/2)\frac{1}{\sqrt{nI(\hat{\theta})}} \leq \theta_0 \leq \hat{\theta} + z(1-\alpha/2)\frac{1}{\sqrt{nI(\hat{\theta})}}$$

as an approximate $100(1-\alpha)\%$ confidence interval.

Example 8: Let $X_1, \cdots, X_n$ denote a random sample from a Poisson distribution that has mean $\lambda > 0$.

It is easy to see that the MLE of λ is $\hat{\lambda} = \bar{X}$. Since the sum of independent Poisson random variables follows a Poisson distribution, the parameter of which is the sum of the parameters of the individual summands, $n\hat{\lambda} = \sum_{i=1}^{n} X_i$ follows a Poisson distribution with mean $n\lambda$. Therefore the sampling distribution of $\hat{\lambda}$ is know, which depends on the true value of λ. Exact confidence intervals for λ may be obtained by using this fact, and special tables are available.

For large samples, confidence intervals may be derived as follows. First, we need to calculate $I(\lambda)$. The probability mass function of a Poisson random variable with parameter λ is

$$f(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!} \quad \text{for } x = 0, 1, 2, \cdots$$

then

$$\log f(x|\lambda) = x \log \lambda - \lambda - \log x!$$

It is easy to verify that

$$\frac{\partial^2}{\partial \lambda^2} \log f(x|\lambda) = -\frac{x}{\lambda^2}$$

therefore

$$I(\lambda) = -E\left[-\frac{X}{\lambda^2}\right] = \frac{1}{\lambda}$$

Thus, an approximate $100(1-\alpha)\%$ confidence interval for λ is

$$\left[\bar{X} - z(1-\alpha/2)\sqrt{\frac{\bar{X}}{n}}, \bar{X} + z(1-\alpha/2)\sqrt{\frac{\bar{X}}{n}}\right]$$

Note that in this case, the asymptotic variance is in fact the exact variance, as we can verify. The confidence interval, however, is only approximate, since the sampling distribution is only approximately normal.

4 Multiple Parameter Case

Suppose now there are more than one parameters in the distribution model, that is, the random variable $X \sim f(x|\boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)^T$. We denote the log-likelihood function

as

$$l(\boldsymbol{\theta}) = \log f(x|\boldsymbol{\theta}),$$

and its first order derivative with respect to $\boldsymbol{\theta}$ is a k-dimensional vector, which is

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_1}, \cdots, \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_k} \right)^T,$$

The second order derivative of $l(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is a $k \times k$ matrix, which is

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = \left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right]_{i=1,\cdots,k; j=1,\cdots,k}$$

We define the Fisher information matrix as

$$\mathbf{I}(\boldsymbol{\theta}) = E\left[\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T \right] = Cov\left[\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right] = -E\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \right]$$

Since the covariance matrix is symmetric and semi-positive definite, these properties hold for the Fisher information matrix as well.

Example 9: Fisher information for normal distribution $N(\mu, \sigma^2)$. We have

$$\boldsymbol{\theta} = (\mu, \sigma^2)^T,$$

and

$$l(\boldsymbol{\theta}) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}.$$

Thus,

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial l(\boldsymbol{\theta})}{\partial \mu}, \frac{\partial l(\boldsymbol{\theta})}{\partial \sigma^2} \right)^T = \left(\frac{x-\mu}{\sigma^2}, -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2(\sigma^2)^2} \right)^T,$$

and

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = \begin{bmatrix} -\frac{1}{\sigma^2} & -\frac{x-\mu}{(\sigma^2)^2} \\ -\frac{x-\mu}{(\sigma^2)^2} & \frac{1}{2(\sigma^2)^2} - \frac{(x-\mu)^2}{(\sigma^2)^3} \end{bmatrix}$$

For $X \sim N(\mu, \sigma^2)$, since $E(X-\mu) = 0$ and $E((X-\mu)^2) = \sigma^2$, we can easily get the Fisher information matrix as

$$\mathbf{I}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \right] = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

Similar to the one parameter case, the asymptotic distribution of MLE $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is approximately multi variate normal distribution with the true value of $\boldsymbol{\theta}$ as the mean and $[n\mathbf{I}(\boldsymbol{\theta})]^{-1}$ as the covariance matrix.

5 Exercises

Problem 1: Suppose that a random variable X has a Poisson distribution for which the mean θ is unknown ($\theta > 0$). Find the Fisher information $I(\theta)$ in X.

Problem 2: Suppose that a random variable X has a normal distribution for which the mean is 0 and the standard deviation σ is unknown ($\sigma > 0$). Find the Fisher information $I(\sigma)$ in X.

Problem 3: Suppose that a random variable X has a normal distribution for which the mean is 0 and the standard deviation σ is unknown ($\sigma > 0$). Find the Fisher information $I(\sigma^2)$ in X. Note that in this problem, the variance σ^2 is regarded as the parameter, whereas in Problem 2 the standard deviation σ is regarded as the parameter.

Problem 4: The Rayleigh distribution his defined as:

$$f(x|\theta) = \frac{x}{\theta^2} e^{-x^2/(2\theta^2)}, \quad x \geq 0, \quad \theta > 0$$

Assume that $X_1, X_2, \cdots, X_n$ is an i.i.d. sample from the Rayleigh distribution. Find the asymptotic variance of the mle.

Problem 5: Suppose that $X_1, \cdots, X_n$ form a random sample from a gamma distribution for which the value of the parameter α is unknown and the value of parameter β is known. Show that if n is large, the distribution of the MLE of α will be approximately a normal distribution with mean α and variance

$$\frac{[\Gamma(\alpha)]^2}{n\{\Gamma(\alpha)\Gamma''(\alpha) - [\Gamma'(\alpha)]^2\}}$$

Problem 6: Let $X_1, X_2, \cdots, X_n$ be an i.i.d. sample from an exponential distribution with the density function

$$f(x|\tau) = \frac{1}{\tau} e^{-x/\tau}, \quad x \geq 0, \quad \tau > 0$$

a. Find the MLE of τ.

b. What is the exact sampling distribution of the MLE.

c. Use the central limit theorem to find a normal approximation to the sampling distribution.

d. Show that the MLE is unbiased, and find its exact variance.

e. Is there any other unbiased estimate with smaller variance?

f. Using the large sample property of MLE, find the asymptotic distribution of the MLE. Is it the same as in **c.**?

g. Find the form of an approximate confidence interval for τ.

h. Find the form of an exact confidence interval for τ.