M50011 Revision

Teddy Wu

February 2026

1 Estimators

2 Cramer-Rao Lower bound

Definition 2.1. For a statistical model, suppose $T = T(X)$ is an unbiased estimator for $\theta \in \Theta \subseteq \mathbb{R}$ based on the observed data $x = (x_1, ..., x_n)$, with joint pdf $f_\theta(x) = f(x_1, ..., x_n : \theta)$. The *Fisher information* is defined as

$$I_f(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \ln f_\theta(X) \right)^2 \right]$$

In defining the Fisher information, it is most natural to view θ as a variable indexing the parametric model $\{f_\theta : \theta \in \Theta\}$. The log-likelihood $\ln f_\theta(X)$ is therefore a function of the parameter θ, so we can take partial derivatives with respect to θ.
Here, $\mathbb{E}_\theta$ means when taking expectation, X follows distribution with parameter θ.
In statistical inference, the data are generated from some true but unknown parameter value θ_0. The quantity that is ultimately relevant for variance bounds and asymptotic theory is therefore the value of this function at the true parameter, $I(\theta_0)$.

Proposition 2.2. *Under regularity conditions, the Fisher information can be written as*

$$I_f(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \ln f_\theta(X) \right]$$

The log-likelihood is a function of the parameter θ, but it also depends on the observed data X. We take the second derivative of the log-likelihood with respect to θ, and then average it over all possible realizations of X under the model. The resulting quantity $I(\theta)$ can be interpreted as an average measure of the convexity of the log-likelihood at θ.
When considering a maximum likelihood estimator, the log-likelihood typically has a bell-shaped (concave) form around its maximum. At the maximizer, the second derivative is negative, which is why a minus sign appears in the definition of Fisher information. A larger

magnitude of curvature means that the log-likelihood drops more sharply when we move away from the true parameter. Consequently, small deviations of the parameter lead to large changes in the likelihood, making the parameter easier to distinguish from nearby alternatives. In this sense, higher Fisher information corresponds to greater precision of the estimator.

Proof. For simplicity, denote $\frac{\partial}{\partial\theta} f_\theta$ as f'_θ and $\frac{\partial^2}{\partial\theta^2} f_\theta$ as f''_θ. Then

$$\begin{aligned}
-I_f(\theta) &= \mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f_\theta(X)\right] \\
&= \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\frac{f'_\theta(X)}{f_\theta(X)}\right] \\
&= \mathbb{E}_\theta\left[\frac{f_\theta(X)f''_\theta(X) - f'_\theta(X)^2}{f_\theta(X)^2}\right] \\
&= \mathbb{E}_\theta\left[\frac{f''_\theta(X)}{f_\theta(X)} - \frac{f'_\theta(X)^2}{f_\theta(X)^2}\right] \\
&= \mathbb{E}_\theta\left[\frac{f''_\theta(X)}{f_\theta(X)}\right] - \mathbb{E}_\theta\left[\frac{f'_\theta(X)^2}{f_\theta(X)^2}\right] \\
&= \mathbb{E}_\theta\left[\frac{f''_\theta(X)}{f_\theta(X)}\right] - \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ln f_\theta(X)\right)^2\right]
\end{aligned}$$

and we have

$$\begin{aligned}
\mathbb{E}_\theta\left[\frac{f''_\theta(X)}{f_\theta(X)}\right] &= \int \frac{f''_\theta(x)}{f_\theta(x)}\cdot f_\theta(x)dx \\
&= \int f''_\theta(x)dx \\
&= \frac{\partial^2}{\partial\theta^2}\int f_\theta(x)dx \\
&= \frac{\partial^2}{\partial\theta^2}1 = 0
\end{aligned}$$

combining we get

$$\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f_\theta(X)\right] = -\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ln f_\theta(X)\right)^2\right]$$

□

Remark 2.3. Suppose $X_1, \cdots, X_n$ are iid. Then

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

where $x = (x_1, \cdots, x_n)$ and $f_\theta^{(1)}$ is the pmf/pdf of a single observation. Thus

$$\begin{aligned} I_f(\theta) &= -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f_\theta(X))\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\sum_{i=1}^n \ln f_\theta^{(1)}(X_i))\right] \\ &= \sum_{i=1}^n -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f_\theta^{(1)}(X_i)\right] \\ &= nI_{f^{(1)}}(\theta) \end{aligned}$$

Therefore the Fisher information is proportional to the sample size. This make sense because the more independent observations we have, the more precise the estimator.

Theorem 2.4 (**Cramer-Rao Lower Bound**). *Let T be an unbiased estimator of θ defined as above. Under mild regular conditions:*

$$\mathrm{MSE}_\theta(T) = \mathrm{Var}_\theta(T) \geq \frac{1}{I(\theta)} \text{ for all } \theta \in \Theta.$$

Proof. By Cauchy-Schwartz we know $\mathbb{E}[YZ]^2 \leq \mathbb{E}[Y^2]\mathbb{E}[Z^2]$, so

$$\begin{aligned} \mathrm{Var}_\theta(T)I_f(\theta) &= \mathbb{E}_\theta[(T-\mathbb{E}_\theta(T))^2]\cdot\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ln f_\theta(X)\right)^2\right] \\ &\geq \mathbb{E}_\theta\left[(T-\mathbb{E}_\theta(T))\cdot\left(\frac{\partial}{\partial\theta}\ln f_\theta(X)\right)\right]^2 \\ &= \mathbb{E}_\theta\left[(T-\mathbb{E}_\theta(T))\cdot\left(\frac{\frac{\partial}{\partial\theta}f_\theta(X)}{f_\theta(X)}\right)\right]^2 \\ &= \left(\int (T(x)-\mathbb{E}_\theta(T))\cdot\left(\frac{\frac{\partial}{\partial\theta}f_\theta(x)}{f_\theta(x)}\right)\cdot f_\theta(x)dx\right)^2 \\ &= \left(\int T(x)\cdot\frac{\partial}{\partial\theta}f_\theta(x)dx - \int \mathbb{E}_\theta(T)\cdot\frac{\partial}{\partial\theta}f_\theta(x)dx\right)^2 \\ &= \left(\frac{\partial}{\partial\theta}\int T(x)f_\theta(x)dx - \mathbb{E}_\theta(T)\frac{\partial}{\partial\theta}\int f_\theta(x)dx\right)^2 \\ &= \left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta(T) - E_\theta(T)\frac{\partial}{\partial\theta}1\right)^2 \\ &= \left(\frac{\partial}{\partial\theta}\theta\right)^2 = 1. \end{aligned}$$

□

In practice, the data are generated from the true parameter θ_0, so the relevant variance is $\mathrm{Var}_{\theta_0}(T)$, and the Cramer-Rao lower bound gives

$$\mathrm{Var}_{\theta_0}(T) \geq \frac{1}{I(\theta_0)}.$$

Example 2.5. Let $X_1, ..., X_n$ iid with $X_1 \sim Bern(\theta)$, $\theta \in \Theta = (0, 1)$. To compute $I_f(\theta)$, we first consider the one random variable and use $I_f(\theta) = nI_{f^{(1)}}(\theta)$.

We know $f_\theta^{(1)}(0) = \Pr(X_1 = 0) = 1 - \theta$, $f_\theta^{(1)}(1) = \Pr(X_1 = 1) = \theta$. We can make the pmf continuous by letting

$$f_\theta^{(1)}(x) = \theta^x(1 - \theta)^{1-x}.$$

Now

$$\begin{aligned} I_{f^{(1)}}(\theta) &= -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial\theta^2} \ln f_\theta^{(1)}(X_1) \right] \\ &= -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial\theta^2} (X_1 \ln(\theta) + (1 - X_1) \ln(1 - \theta)) \right] \\ &= -\mathbb{E}_\theta \left[-\frac{X_1}{\theta^2} - \frac{1 - X_1}{(1 - \theta)^2} \right] \\ &= \frac{\mathbb{E}_\theta[X_1]}{\theta^2} + \frac{1 - \mathbb{E}_\theta[X_1]}{(1 - \theta)^2} \\ &= \frac{1}{\theta} + \frac{1}{1 - \theta}. \end{aligned}$$

Therefore

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

The unbiased estimator $\bar{X}$ has variance

$$\mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X_1)}{n} = \frac{\theta(1 - \theta)}{n},$$

which is exactly the Cramer-Rao lower bound.

3 Asymptotic properties

Theorem 3.1 (**Weak Law of Large Numbers**). *If $Y_1, ..., Y_n$ are iid random variables and $\mathbb{E}[Y_i] = \mu$ for all i, then*

$$\bar{Y}_n \xrightarrow{P} \mu.$$

Sample mean converges to mean in probability.

Theorem 3.2 (**Central Limit Theorem**). *If $Y_1, ..., Y_n$ are iid random variables with $\mathbb{E}[Y_i] = \mu$ and $\mathrm{Var}(Y_i) = \sigma^2$ for all i, then*

$$\sqrt{n}(\bar{Y}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Under suitable rescaling, sample mean converges to a normal random variable in distribution.

Lemma 3.3 (**Slutsky**). *If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} c$ for a constant c, then*

(i) $X_n + Y_n \xrightarrow{d} X + c$;

(ii) $X_n Y_n \xrightarrow{d} cX$;

(iii) $\frac{X_n}{Y_n} = \frac{X}{c}$ *if* $c \neq 0$.

Theorem 3.4 (**Continuous mapping theorem**). *If g is continuous, then*

(i) $X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X)$;

(ii) $X_n \xrightarrow{P} X \Rightarrow g(X_n) \xrightarrow{P} g(X)$;

(iii) $X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X)$.

Theorem 3.5. *Suppose $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$ and $X_n Y_n \xrightarrow{P} XY$.*

Theorem 3.6. *Suppose $(X_n) \xrightarrow{d} c$ for some $c \in \mathbb{R}$, then $X_n \xrightarrow{P} c$.*

Definition 3.7. A sequence of estimators $(T_n)_{n\in\mathbb{N}}$ for $g(\theta)$ is called *consistent* if for all $\theta \in \Theta$:

$$T_n \xrightarrow{P} g(\theta).$$

Example 3.8. If $X_1, ..., X_n$ are iid and $\mathbb{E}[X_i] = \mu$, then the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ as an estimator for μ is consistent by Law of Large Numbers. The sample variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ as an estimator for σ^2 is also consistent:

$$\begin{aligned} S_n^2 &= \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^n (X_i^2 - 2\bar{X}X_i + \bar{X}^2) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right) \\ &= \frac{1}{n-1}\sum_{i=1}^n X_i^2 - \frac{n}{n-1}\bar{X}^2 \end{aligned}$$

Since $\mathbb{E}[X_i^2] = \mathrm{Var}(X_i) + \mathbb{E}[X_i]^2 = \sigma^2 + \mu^2$, by Law of Large Numbers we have

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{P} \sigma^2 + \mu^2.$$

We know $\bar{X} \xrightarrow{P} \mu$, by continuous mapping theorem

$$\bar{X}^2 \xrightarrow{P} \mu^2.$$

Since $\frac{n}{n-1} \xrightarrow{a.s.} 1$, we can apply Slutsky to obtain

$$\frac{1}{n-1}\sum_{i=1}^n X_i^2 = \frac{n}{n-1} \cdot \frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{P} \sigma^2 + \mu^2,$$

$$\frac{n}{n-1}\bar{X} \xrightarrow{P} \mu^2.$$

Finally we use 3.5 to conclude

$$\frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{n}{n-1}\bar{X}^2 \xrightarrow{P} \sigma^2.$$

So S_n^2 is consistent.

Definition 3.9. A sequence of estimators $(T_n)_{n\in\mathbb{N}}$ for $g(\theta)$ is called *asymptotic unbiased* if

$$\mathbb{E}_\theta[T_n] \to g(\theta).$$

Theorem 3.10 (Criterion for consistency). *A sequence of estimators $(T_n)_{n\in\mathbb{N}}$ for $g(\theta)$ is consistent if it is asymptotic unbiased and*

$$\mathrm{Var}_\theta(T_n) \to 0.$$

Proof. For any $\epsilon > 0$:

$$\begin{aligned}
\Pr(|T_n - g(\theta)| \geq \epsilon) &= \Pr(|T_n - g(\theta)|^2 \geq \epsilon^2) \\
&\leq \frac{\mathbb{E}_\theta[(T_n - g(\theta))^2]}{\epsilon^2} \quad \text{(by Markov inequality)} \\
&= \frac{\mathrm{MSE}_\theta(T_n)}{\epsilon^2} \\
&= \frac{1}{\epsilon^2}(\mathrm{Var}_\theta(T_n) + (\underbrace{\mathbb{E}_\theta[T_n] - g(\theta)}_{\text{bias}})^2)
\end{aligned}$$

which tends to 0, since variance and bias tends to 0. □

Definition 3.11. A sequence of estimators $(T_n)_{n\in\mathbb{N}}$ for $\theta \in \Theta$ is called *asymptotic normal* if

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta))$$

for some function σ^2 of θ.

Example 3.12. Let $X \sim Bin(n, p)$, then X/n is an asymptotic normal estimator of p, since we can view X as a sum of Bernoulli random variables

$$X = \sum_{i=1}^{n} Y_i, \quad Y_i \sim Bern(p).$$

thus $X/n = \frac{1}{n}\sum_{i=1}^{n} Y_i$, by CLT be have

$$\sqrt{n}(X/n - p) \xrightarrow{d} N(0, p(1-p)).$$

Theorem 3.13 (**Delta method**). *Suppose T_n is an asymptotic normal estimator of θ with*

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta)).$$

Let $g : \Theta \to \mathbb{R}$ be a continuous function with $g'(\theta) \neq 0$. Then

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{d} N(0, g'(\theta)^2\sigma^2(\theta))$$

Proof. □

Example 3.14. Let $Y_1, ..., Y_n$ be i.i.d. $Bern(p)$ random variables, $S_n = \bar{Y_n}$. We know that by CLT

$$\sqrt{n}(S_n - p) \xrightarrow{d} N(0, p(1-p)).$$

We can use the delta method to calculate the distribution of sample odds. Let $g(x) = \frac{x}{1-x}$, so $g'(x) = (1-x)^{-2}$. By the delta method we have that

$$\sqrt{n}(g(S_n) - p/(1-p)) \xrightarrow{d} N\left(0, ((1-p)^{-2})^2 \cdot p(1-p)\right).$$

Let $T_n = \frac{S_n}{1-S_n}$ denote the sample odds, then

$$\sqrt{n}(T_n - p/(1-p)) \xrightarrow{d} N\left(0, p(1-p)^{-3}\right).$$

Theorem 3.15. *Suppose T_n is an asymptotic normal estimator of θ, then it is a consistent estimator of θ.*

Proof. The assumption says that

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta)).$$

Let $Z_n = \sqrt{n}(T_n - \theta)$, $Z = N(0, \sigma^2(\theta))$, we have $Z_n \xrightarrow{d} Z$. Since $\frac{1}{\sqrt{n}} \to 0$, by Slutsky we obtain

$$\frac{1}{\sqrt{n}} Z_n \xrightarrow{d} 0 \cdot Z = 0.$$

Therefore

$$T_n - \theta \xrightarrow{d} 0.$$

By 3.6 we also have convergence in probability. Thus

$$T_n \xrightarrow{P} \theta,$$

proving consistency. □

4 Maximum likelihood estimator

Definition 4.1. Let the pmf/pdf of X be $f(x:\theta)$. Having observed the data x (write the independent observations $x_1, ..., x_n$ as a vector), the *likelihood function* $L(\theta : x) : \Theta \to \mathbb{R}$ is

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

Setting the derivative with respect to θ to be 0 and checking the second derivative is negative, we can find the $\hat{\theta}$ that maximizes $L(\theta|x)$. This $\hat{\theta}$ is called the *maximum likelihood estimator (MLE)*. It is also important to check the endpoints of Θ to ensure the local maximum is indeed a global maximum.

Definition 4.2. Taking the natural log of the likelihood function yields the *log-likelihood*, commonly denoted as

$$l(\theta) = \ln(L(\theta : x)).$$

Example 4.3. MLEs are not necessarily unbiased. Consider iid $X_1, ..., X_n$ where $X_1 \sim N(\mu, \sigma^2)$, with μ, σ^2 unknown. Our likelihood function is

$$\begin{aligned} L(\mu, \sigma^2 : x) = f(x : \mu, \sigma^2) &= \prod_{i=1}^{n} f(x_i : \mu, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right) \end{aligned}$$

The log-likelihood is

$$l(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

We first minimise the log-likelihood with respect to μ. Taking partial derivatives:

$$\frac{\partial l}{\partial \mu} = \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(x_i - \mu) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$$

$$\sum_{i=1}^{n} x_i - n\mu = 0$$

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

So the MLE for μ is

$$\hat{\mu} = \bar{X}.$$

Plug this back to the log-likelihood:

$$l(\sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Taking partial derivatives:

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = n\sigma^2$$

So the MLE for σ^2 is

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

This is biased because

$$\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n}\mathbb{E}[S_n^2] = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

Example 4.4 (**The German Tank Problem**)**.** Suppose that German tanks are labelled with consecutive integers $1, 2, ..., N$, where the unknown parameter N denotes the total number of tanks produced. Assume that the Allies capture k tanks whose observed serial numbers are $x_1, ..., x_k$, and that each subset of k tanks is equally likely to be captured.

Let the random variables $X_1, \ldots, X_k$ denote the observed serial numbers of the captured tanks. Under this model, every subset of size k from $\{1, \ldots, N\}$ is equally likely to be observed. Therefore, for a given observed set $\{x_1, \ldots, x_k\}$, the joint probability mass function is

$$P_N\big(\{X_1, \ldots, X_k\} = \{x_1, \ldots, x_k\}\big) = \begin{cases} \dfrac{1}{\binom{N}{k}}, & \text{if } \max\{x_1, \ldots, x_k\} \leq N, \\ 0, & \text{otherwise.} \end{cases}$$

This defines the likelihood function for N. Let

$$M = \max(X_1, \ldots, X_k)$$

denote the sample maximum. If $N < M$, the likelihood is zero, since it is impossible to observe a tank with serial number $> N$. For $N \geq M$, the likelihood $1/\binom{N}{k}$ is a decreasing function of N. Consequently, the likelihood is maximised at the smallest possible value of N, and the maximum likelihood estimator is given by

$$\hat{N} := M = \max(X_1, \ldots, X_k).$$

The expectation of this MLE is

$$\begin{aligned}\mathbb{E}[\hat{N}] &= \sum_{n=k}^{N} n \Pr(N=n) \\ &= \sum_{n=k}^{N} n \frac{\binom{n-1}{k-1}}{\binom{N}{k}} \\ &\cdots \\ &= \frac{k}{k+1}(N+1),\end{aligned}$$

We see again this is a biased MLE.

Example 4.5 (**MLE for common distributions**). Suppose $X_1, ..., X_n$ are iid random variables, then the MLE for estimator the parameter θ is

(i) $\bar{X}$ for $X_1 \sim Bern(\theta)$;

(ii) $\bar{X}$ for $X_1 \sim Poisson(\theta)$;

(iii) $1/\bar{X}$ for $X_1 \sim Exp(\theta)$.

Proof. □

Theorem 4.6 (**Invariance of MLE**). *If g is bijective and if $\hat{\theta}$ is an MLE of θ, then $\hat{\phi} = g(\hat{\theta})$ is an MLE of $g(\theta)$.*

Proof. □

Theorem 4.7 (**Asymptotic normality of MLE**). *Let $X_1, X_2, ...$ be iid observations with pdf/pmf $f_\theta(x)$, where $\theta \in \Theta$. Let $\theta_0 \in \Theta$ denote the true parameter. Under regularity conditions:*

(i) *There exists a consistent sequence $(\hat{\theta}_n)_{n\in\mathbb{N}}$ of MLEs, where $\hat{\theta}_n$ is an estimator based on $X_1, ..., X_n$. If MLE is unique for every n (there's only one point where the likelihood function attains its maximum), then that MLE is consistent;*

(ii) *Suppose $(\hat{\theta}_n)_{n\in\mathbb{N}}$ is a consistent sequence of MLEs. Then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, (I_{f^{(1)}}(\theta_0))^{-1})$$

where $I_{f^{(1)}}(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} \ln f_\theta^{(1)}(X)\right]$ is the Fisher information of a sample size 1.

After we obtain a consistent sequence of MLE $\hat{\theta}_n$, we can assume $\sqrt{n}(\hat{\theta}_n - \theta_0) \sim N(0, (I_{f^{(1)}}(\theta_0))^{-1})$ by approximation, therefore we may calculate confidence intervals.

Remark 4.8. Apply this theorem on example 2.5 we see that

$$\sqrt{n}(\bar{X} - \theta_0) \xrightarrow{d} N(0, \theta_0(1 - \theta_0)).$$

If we use CLT instead, we will get the same result.

5 Confidence interval

Definition 5.1. A $1 - \alpha$ confidence interval I for θ is a random interval I such that

$$\Pr(\theta_0 \in I) \geq 1 - \alpha,$$

where θ_0 is the true parameter.

Definition 5.2. Let θ be the unknown parameter and Y be the data we observe. A *pivotal quantity* for θ is a function $t(Y, \theta)$, only consisting the data and θ, such that the distribution of $t(Y, \theta)$ is known.

Remark 5.3. Suppose $t(Y, \theta)$ is a pivotal quantity for θ, then we can find constants a_1, a_2 such that

$$\Pr(a_1 \leq t(Y, \theta) \leq a_2) \geq 1 - \alpha.$$

If we can rearrange terms to obtain

$$\Pr(h_1(Y)) \leq \theta \leq h_2(Y)) \geq 1 - \alpha,$$

then we get $[h_1(Y), h_2(Y)]$ which is a random interval. The observed interval $[h_1(y), h_2(y)]$ is a $1 - \alpha$ confidence interval.

Remark 5.4. Often, we only know

$$\sqrt{n}\frac{(T_n - \theta)}{\sigma(\theta)} \xrightarrow{d} N(0, 1).$$

Thus approximately

$$\sqrt{n}\frac{(T_n - \theta)}{\sigma(\theta)} \sim N(0, 1),$$

so we can use LHS as a pivotal quantity.

Example 5.5. Suppose $X_1, ..., X_n$ follow normal distribution with unknown mean θ, known variance σ^2. By CLT we have the (approximate) pivotal quantity

$$Z = \frac{\bar{X} - \theta}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Define z_α satisfying $\alpha = P(Z \leq z_\alpha)$. Thus

$$P(z_{\alpha/2} < Z < z_{1-\alpha/2}) = 1 - \alpha.$$

Plugging in we have

$$P\left(-z_{1-\alpha/2} < \frac{\bar{X} - \theta}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) = 1 - \alpha$$

where we used the fact that $z_{\alpha/2} = -z_{1-\alpha/2}$ by symmetry. Rearranging yields

$$P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \theta < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Thus

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

is a $100(1-\alpha)\%$ confidence interval of θ.

Example 5.6. Suppose independent random variables $X_1, ..., X_n$ follow normal distribution with unknown mean θ, unknown variance σ^2. Since σ^2 is unknown we use s^2 to estimate variance. Now our pivotal quantity follows t-distribution with dof $v = n - 1$:

$$T = \frac{\bar{X} - \theta}{s/\sqrt{n}} \sim t_{n-1}$$

Similarly we can find $t_{v,1-\alpha/2}$ such that

$$P\left(-t_{v,1-\alpha/2} < \frac{\bar{X} - \theta}{s/\sqrt{n}} < t_{v,1-\alpha/2}\right) = 1 - \alpha$$

Rearranging:

$$P\left(\bar{X} - t_{v,1-\alpha/2}\frac{s}{\sqrt{n}} < \theta < \bar{X} + t_{v,1-\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Thus

$$\left(\bar{X} - t_{v,1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + t_{v,1-\alpha/2}\frac{s}{\sqrt{n}}\right)$$

is a $100(1-\alpha)\%$ confidence interval of θ.

Example 5.7. Suppose independent random variables $X_1, ..., X_n$ follow normal distribution with unknown mean θ, unknown variance σ^2. To obtain a confidence interval for σ^2 we use the pivotal quantity

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

$$P\left(\chi^2_{n-1,\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1,1-\alpha/2}\right) = 1 - \alpha$$

Note that since χ^2 is not a symmetric distribution we can't write $\chi^2_{n-1,\alpha} = -\chi^2_{n-1,1-\alpha/2}$.

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right) = 1-\alpha$$

Thus

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right)$$

is a $100(1-\alpha)\%$ confidence interval of σ^2.

Remark 5.8. Construct a confidence interval from $\sqrt{n}\frac{(T_n-\theta)}{\sigma(\theta)}$ is usually not easy, since σ may dependent on θ so it is difficult to solve the inequalities for θ. However if we have a consistent estimator $\hat{\sigma}_n$ for θ, then $\hat{\sigma}_n \xrightarrow{P} \sigma(\theta)$. Since $\sqrt{n}(T_n-\theta) \xrightarrow{d} N(0,\sigma^2(\theta))$, we can apply Slutsky to obtain

$$\sqrt{n}\frac{T_n-\theta}{\hat{\sigma}_n} \xrightarrow{d} N(0,1).$$

(Here we are considering the convergence of random variables, so θ and $\sigma(\theta)$ are viewed as constants.)

Example 5.9. Let $X \sim Bin(n,\theta)$ where n is known, θ is the parameter we are trying to estimate. By CLT we have

$$\sqrt{n}(X/n-\theta) \xrightarrow{d} N(0,\theta(1-\theta)).$$

Thus approximately

$$\sqrt{n}\frac{X/n-\theta}{\sqrt{\theta(1-\theta)}} \sim N(0,1).$$

But the LHS is difficult to solve for θ. To find a consistent estimator, note that since $X/n \xrightarrow{P} \theta$, by Law of Large Numbers:

$$\frac{X}{n}\left(1-\frac{X}{n}\right) \xrightarrow{P} \theta(1-\theta),$$

so by the above remark we have

$$\sqrt{n}\frac{X/n-\theta}{\frac{X}{n}\left(1-\frac{X}{n}\right)} \sim N(0,1)$$

which can be readily solved for θ.

6 Hypothesis testing