Large Sample Theory

1 Basic concepts and tools

Sample size “n” tends to infinity, procedures depend on sample, in particular they depend on n. We have sequences of procedures.

Suppose $X_1, X_2, \ldots$ is a sequence of random variables sampled independently from $f(x|\theta)$. Let

$$W_n = W_n(X_1, \ldots, X_n)$$

be a sequence of estimators. For example, if $W_n = \bar{X}$, then $W_1 = X_1$, $W_2 = \frac{1}{2}(X_1 + X_2)$, $W_3 = \frac{1}{3}(X_1 + X_2 + X_3)$, etc.

**Definition 1.1** A sequence $W_n = W_n(X_1, \ldots, X_n)$ is a consistent sequence of estimators, if, for every $\epsilon > 0$ and every $\theta \in \Theta$,

$$\lim_{n \to \infty} P_{\theta}(|W_n - \theta| < \epsilon) = 1,$$

or equivalently,

$$\lim_{n \to \infty} P_{\theta}(|W_n - \theta| \geq \epsilon) = 0.$$

In other words, the sequence $W_n$ converges in probability to the “true value”, no matter what this true value is.

**Example:** Consistency of $\bar{X}$. If $X_1, X_2, \ldots \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma^2)$, then

$$P_{\theta, \sigma^2}(|\bar{X} - \theta| < \epsilon) = P_{\theta, \sigma^2}\left(\frac{\sqrt{n}(\bar{X} - \theta)}{\sigma} < \frac{\sqrt{n}\epsilon}{\sigma}\right) \to 1 \text{ as } n \to \infty.$$

In general, if $X_1, X_2, \ldots$ are i.i.d., $E X_1 = \mu$, $\text{Var}(X_1) = \sigma^2 < \infty$. Then by WLLN,

$$\bar{X}_n \xrightarrow{n \to \infty, P} \mu.$$

**Chebykov’s inequality.** For nonnegative random variable $Y \geq 0$, there holds

$$P(Y \geq r) \leq \frac{E[Y]}{r}.$$

**Theorem 1.2** If $\lim_{n \to \infty} \text{Var}_\theta(W_n) = 0$ and $\lim_{n \to \infty} \text{Bias}_\theta(W_n) = 0$ for all $\theta$, then $W_n$ is a consistent estimator of $\theta$.

**Proof** First, we have

$$E_\theta(W_n - \theta)^2 = \text{Var}_\theta(W_n) + [\text{Bias}_\theta(W_n)]^2 \xrightarrow{n \to \infty} 0.$$

Then, for any $\epsilon > 0$, then

$$P_\theta(|W_n - \theta| > \epsilon) \leq \frac{E_\theta(W_n - \theta)^2}{\epsilon^2} \xrightarrow{n \to \infty} 0.$$
This gives
\[ W_n \xrightarrow{n \to \infty} P \theta. \]

**Example**: In the case \( W_n = \bar{X} \), we have \( \text{Bias}_\theta(W_n) = 0 \) and \( \text{Var}_\theta(\bar{X}_n) = \frac{\sigma^2}{n} \to 0 \).

**Theorem 1.3 (Slutsky’s theorem)** If \( X_n \xrightarrow{d} X \) and \( Y_n \xrightarrow{P} a \), then
- \( X_n + Y_n \xrightarrow{d} X + a \);
- \( X_n \cdot Y_n \xrightarrow{d} aX \).

**Theorem 1.4 (Continuous mapping theorems)**

1. If \( X_n \xrightarrow{P} X \), then \( \phi(X_n) \xrightarrow{P} \phi(X) \) for every \( \phi \) continuous with \( X \)-probability 1;
2. If \( X_n \xrightarrow{d} X \), then \( \phi(X_n) \xrightarrow{d} \phi(X) \) for every \( \phi \) continuous with \( X \)-probability 1.

**Theorem 1.5 (The \( \delta \)-method)** Let \( g : \mathbb{R} \to \mathbb{R} \) be differentiable, and suppose that
\[ \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \]
then
\[ \sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, (g'(\theta))^2\sigma^2). \]
More generally, if \( a_n(\hat{\theta}_n - \theta) \xrightarrow{d} Y \) for some \( a_n \to \infty \) and a random variable \( Y \), then
\[ a_n(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} g'(\theta) \cdot Y. \]

**Intuition**:
\[ \sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \approx \sqrt{n}g'(\theta)(\hat{\theta}_n - \theta). \]

**Example**: Let \( X_1, \ldots, X_n \ldots \) be i.i.d. with \( \mathbb{E}[X_1] = \mu_1 \) and \( \text{Var}[X_1] = \sigma^2 \). Then by CLT, we have
\[ \sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \]
which further implies that
\[ \sqrt{n}(\bar{X}^2 - \mu^2) \xrightarrow{d} \mathcal{N}(0, (2\mu)^2\sigma^2). \]

**Definition 1.6 (Asymptotic variance)** If \( T_n \) is an estimator, such that
\[ k_n(T_n - \tau(\theta)) \xrightarrow{d} \mathcal{N}(0, \tau^2), \]
then \( \tau^2 \) is called asymptotic variance, or variance of the limit distribution of \( T_n \).
2 Likelihood, score function, and Fisher information

Let $X_1, X_2, \ldots, \overset{i.i.d.}{\sim} f(x|\theta)$, and let

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

be the likelihood function. Correspondingly, the log-likelihood function is

$$\ell(\theta|x) = \log L(\theta|x) = \sum_{i=1}^{n} \log f(x_i|\theta).$$

We also define the score function

$$\Psi(\theta|x) = \ell'(\theta|x) = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \log f(x_i|\theta).$$

**Theorem 2.1** We have

$$\mathbb{E}_\theta \Psi(\theta|X) = 0.$$

**Theorem 2.2 (Cramér-Rao Inequality)** Let $W(X) = W(X_1, \ldots, X_n)$ be some estimator satisfying some regularity conditions. Then

$$\text{Var}_\theta(W(X)) \geq \frac{\left( \frac{d}{d\theta} \mathbb{E}_\theta W(X) \right)^2}{n \mathbb{E}_\theta \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2}.$$

**Definition 2.3 (Fisher information number)** We denote the Fisher information number for one observation as

$$I(\theta) := \mathbb{E}_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 \right).$$

If $f(x|\theta)$ satisfies some regularity condition, there usually holds

$$I(\theta) = \mathbb{E}_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 \right) = -\mathbb{E}_\theta \left( \frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right).$$

Furthermore, the Fisher information number for $n$ observations is denoted as

$$I_n(\theta) := nI(\theta).$$

The Fisher information can be represented as

$$I_n(\theta) = \mathbb{E}_\theta \left( [\ell'(\theta)]^2 \right) = -\mathbb{E}_\theta \ell''(\theta),$$

or

$$I_n(\theta) = \mathbb{E}_\theta \left( [\Psi(\theta)]^2 \right) = -\mathbb{E}_\theta \Psi'(\theta).$$

**Definition 2.4 (Observed Fisher information)** For an estimator $\hat{\theta}_n$, the observed Fisher information number is

$$\hat{I}_n(\hat{\theta}_n) = -\frac{\partial^2}{\partial \theta^2} \sum_{i=1}^{n} \log f(X_i|\theta) \bigg|_{\theta=\hat{\theta}_n}.$$

The observed Fisher information number can also be represented as

$$\hat{I}_n(\hat{\theta}_n) = -\ell''(\hat{\theta}_n) = -\Psi'(\hat{\theta}_n).$$
3 Asymptotic theory for MLE

**Theorem 3.1 (Consistency of MLE)** Let \( \hat{\theta} \) be the MLE of \( \theta \). Let \( \tau(\theta) \) be a continuous function of \( \theta \). Under regularity conditions on \( f(x|\theta) \), for every \( \epsilon > 0 \) and every \( \theta \in \Theta \),

\[
\lim_{n \to \infty} \mathbb{P}_\theta(|\tau(\hat{\theta}) - \tau(\theta)| \geq \epsilon) = 0.
\]

That is, \( \tau(\hat{\theta}) \) is a consistent estimator of \( \tau(\theta) \).

**Theorem 3.2 (Asymptotic efficiency of MLEs)** Let \( X_1, X_2, \ldots \) be i.i.d. \( f(x|\theta) \), let \( \hat{\theta} \) denote the MLE of \( \theta \), let \( \hat{\theta}_0 \) denote the MLE of \( \theta \), and let \( \tau(\theta) \) be a continuous function of \( \theta \). Under certain regularity conditions on \( f(x|\theta) \),

\[
\sqrt{n} [\tau(\hat{\theta}_n) - \tau(\theta)] \to \mathcal{N} \left(0, \frac{|\tau'(\theta)|^2}{I(\theta)} \right),
\]

which attains the Cramér-Rao lower bound. That is, \( \tau(\hat{\theta}) \) is a consistent and asymptotically efficient estimator of \( \tau(\theta) \).

Notice that in the case of \( \tau(\theta) = \theta \), the above asymptotic normality can be also represented as

\[
\sqrt{n} (\hat{\theta}_n - \theta) \to \mathcal{N} \left(0, \frac{1}{I(\theta)} \right),
\]

which can be rewritten as

\[
\sqrt{I_n(\theta)} (\hat{\theta}_n - \theta) \to \mathcal{N} (0, 1) .
\]

Usually, by Slutsky’s Theorem,

\[
\sqrt{I_n(\hat{\theta}_n)} (\hat{\theta}_n - \theta) \to \mathcal{N} (0, 1) .
\]

This asymptotic normality gives the following Wald’s test:

**MLE based Wald’s test:** Let \( \hat{\theta}_n \) be the MLE. The Wald’s test for \( H_0 : \theta \leq \theta_0 \) vs. \( H_1 : \theta > \theta_0 \) is to compare \( \sqrt{I_n(\theta)} (\hat{\theta}_n - \theta) \) to \( \mathcal{N}(0,1) \).

Moreover, recall that we have the asymptotic normality

\[
\sqrt{\frac{I_n(\theta)}{|\tau'(\theta)|}} [\tau(\hat{\theta}_n) - \tau(\theta)] \to \mathcal{N} (0, 1). 
\]

Replacing the unobserved values with observed estimates, by Slutsky’s theorem, we have

\[
\sqrt{\frac{\hat{I}_n(\hat{\theta}_n)}{|\tau'(\hat{\theta}_n)|}} [\tau(\hat{\theta}_n) - \tau(\theta)] \to \mathcal{N} (0, 1) .
\]

This gives the following approximate confidence interval:

**MLE based approximate confidence intervals:**

\[
h(\hat{\theta}_n) - z_{\frac{\alpha}{2}} \frac{|h'(\hat{\theta}_n)|}{\sqrt{\hat{I}_n(\hat{\theta}_n)}} \leq h(\theta) \leq h(\hat{\theta}_n) + z_{\frac{\alpha}{2}} \frac{|h'(\hat{\theta}_n)|}{\sqrt{\hat{I}_n(\hat{\theta}_n)}}.
\]
4 Asymptotic theory for LRT and score tests

Recall that the likelihood ratio test statistic is
\[ \lambda(X) = \frac{\sup_{\theta \in \Theta_0} L(\theta|X)}{\sup_{\theta \in \Theta} L(\theta|X)}. \]

For the simple test
\[ H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta \neq \theta_0, \]
the LRT statistic is
\[ \lambda(X) = \frac{L(\theta_0|X)}{L(\hat{\theta}_n|X)}, \]
where \( \hat{\theta}_n \) is the MLE.

**Theorem 4.1 (Asymptotic distribution of the LRT)** For testing
\[ H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta \neq \theta_0, \]
suppose \( \hat{\theta}_n \) is the MLE, and \( f(x|\theta) \) satisfies certain regularity conditions. Then under \( H_0 \), as \( n \to \infty \),
\[ -2 \log \lambda(X) \xrightarrow{d} \chi^2_1. \]

**Intuition:** Consider the Taylor expansion
\[ \ell(\theta) - \ell(\hat{\theta}) \approx \ell'(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2} \ell''(\hat{\theta})(\theta - \hat{\theta})^2. \]
Given \( \hat{\theta} \) is MLE, we have \( \ell'(\hat{\theta}) = 0 \), so
\[ 2\ell(\hat{\theta}) - 2\ell(\theta) \approx -\ell''(\hat{\theta})(\theta - \hat{\theta})^2. \]
Recall that
\[ \hat{I}_n(\hat{\theta}_n) = -\ell''(\hat{\theta}_n) \]
and
\[ \sqrt{\hat{I}_n(\hat{\theta}_n)(\hat{\theta}_n - \theta)} \xrightarrow{d} \mathcal{N}(0,1). \]
Then, under the null hypothesis,
\[ -2 \log \lambda(X) = 2\ell(\hat{\theta}) - 2\ell(\theta_0) \xrightarrow{d} \chi^2_1. \]

**Theorem 4.2 (Asymptotic distribution of the score test statistic)** For testing
\[ H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta \neq \theta_0, \]
under \( H_0 \), as \( n \to \infty \),
\[ \frac{\Psi(\theta_0)}{\sqrt{I_n(\theta_0)}} \xrightarrow{d} \mathcal{N}(0,1). \]

**Proof** Notice that
\[ \Psi(\theta_0) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i|\theta_0). \]
Since
\[ \mathbb{E}_{\theta_0} \left[ \frac{\partial}{\partial \theta} \log f(X_i|\theta_0) \right] = 0 \]
and
\[ \text{Var}_{\theta_0} \left[ \frac{\partial}{\partial \theta} \log f(X_i|\theta_0) \right] = \mathbb{E}_{\theta_0} \left( \left( \frac{\partial}{\partial \theta} \log f(X|\theta_0) \right)^2 \right) = I(\theta_0). \]
By CLT, the asymptotic normality is obtained.
5 Example: Bernoulli MLE

Let $X_1, \ldots, X_n \sim \text{Ber}(p)$. The likelihood function is

$$L(p|\mathbf{x}) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p^{\bar{x}} (1-p)^{n-\bar{x}}.$$  

We then have the log-likelihood

$$\ell(p|\mathbf{x}) = \log L(p|\mathbf{x}) = n\bar{x} \log p + n(1-\bar{x}) \log (1-p).$$

Then the score function is

$$\psi(p|\mathbf{x}) = \ell'(p|\mathbf{x}) = \frac{n\bar{x}}{p} - \frac{n(1-\bar{x})}{1-p} = \frac{n(\bar{x} - p)}{p(1-p)}.$$  

It is straightforward to verify that

$$\mathbb{E}_p[\psi(p|\mathbf{X})] = \mathbb{E}_p \left[ \frac{n(\bar{X} - p)}{p(1-p)} \right] = 0.$$  

The MLE can be obtained by solving $\psi(p|\mathbf{x}) = p$, which gives $\hat{p}_n = \bar{x}$.

The Fisher information can be obtained through two methods:

Method 1: By

$$I_n(p) = \mathbb{E}_p \psi^2(p|\mathbf{X}) = \mathbb{E}_p \left[ \frac{n^2 (\bar{X} - p)^2}{p^2(1-p)^2} \right] = \frac{n^2 (p(1-p))}{n} = \frac{n}{p(1-p)}.$$  

Method 2: By

$$\psi'(p|\mathbf{x}) = \frac{\partial}{\partial p} \left( \frac{n(\bar{x} - p)}{p(1-p)} \right) = -\frac{np(1-p) - n(\bar{x} - p)(1-2p)}{p^2(1-p)^2},$$

we have

$$I_n(p) = -\mathbb{E} [\psi'(p|\mathbf{X})] = \frac{n}{p(1-p)}.$$  

For the observed Fisher information, we have

$$\hat{I}_n(\hat{p}) = -\psi'(\hat{p}|\mathbf{x}) = \frac{n}{\hat{p}(1-\hat{p})}.$$  

Notice that here Method 1 is not good for estimating the Fisher information in that

$$\psi(\hat{p}|\mathbf{x}) = 0 \implies \psi^2(\hat{p}|\mathbf{x}) = 0.$$  

By the asymptotic normality of MLE, we have

$$\sqrt{I_n(p)} (\hat{p} - p) = \sqrt{\frac{n}{p(1-p)}} (\hat{p} - p) \overset{d}{\to} \mathcal{N}(0,1).$$  

This can also be obtained straightforwardly by CLT. By Slutsky’s theorem, we have

$$\sqrt{\hat{I}_n(\hat{p})} (\hat{p} - p) = \sqrt{\frac{n}{\hat{p}(1-\hat{p})}} (\hat{p} - p) \overset{d}{\to} \mathcal{N}(0,1).$$
5.1 Confidence Intervals

The MLE based Wald’s test for

\[ H_0 : p \leq p_0 \quad \text{vs.} \quad H_1 : p > p_0 \]

amounts to comparing

\[ \sqrt{\frac{n}{\hat{p}(1-\hat{p})}} (\hat{p} - p_0) \]

with \( \mathcal{N}(0, 1) \). This also gives the Wald’s confidence interval

\[
\left[ \hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \ \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right].
\]

As to the LRT based confidence intervals, by the asymptotic distribution of LRT,

\[-2(\ell(p|X) - \ell(\hat{p}|X)) \xrightarrow{d} \chi^2_1.\]

Recall that

\[ \ell(p|x) = n\hat{p} \log p + n(1-\hat{p}) \log (1-p), \]

which implies that

\[ \ell(p|x) - \ell(\hat{p}|x) = n \left( \hat{p} \log \frac{\hat{p}}{p} + (1-\hat{p}) \log \frac{1-p}{1-\hat{p}} \right). \]

Then the LRT based confidence interval is

\[
\left\{ p : -2n \left( \hat{p} \log \frac{\hat{p}}{p} + (1-\hat{p}) \log \frac{1-p}{1-\hat{p}} \right) \leq \chi^2_{1, \alpha} \right\}.
\]

Recall that by the asymptotic distribution of the score statistic, we have

\[ \frac{\psi(p)}{\sqrt{I_n(p)}} \xrightarrow{d} \mathcal{N}(0, 1). \]

By \( \psi(p) = \frac{n\hat{p} - np}{p(1-p)} \) and \( I_n(p) = \frac{n}{p(1-p)} \), we have

\[ \frac{\psi(p)}{\sqrt{I_n(p)}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}. \]

Then the score test based confidence interval is

\[
\left\{ \left| \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \right| < z_{\alpha/2} \right\},
\]

which is

\[
2\hat{p} + z_{\alpha/2} \pm \sqrt{\left(2\hat{p} + z_{\alpha/2}^2/n \right)^2 - 4\hat{p}^2(1+z_{\alpha/2}^2/n)} \div 2(1+z_{\alpha/2}^2/n).
\]