Topic VI:
Confidence Set

Wei You

香港科技大學
THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

Fall, 2023

Introduction

Point estimation

- Point estimation uses a statistics to estimate the unknown parameter.
- For each realized set of samples, the statistic takes only a single value.
- A point estimator for an unknown parameter θ provides no information about accuracy.
- Confidence sets (intervals) addresses this deficiency by seeking a (random) set (interval) that brackets θ with high probability.

Confidence Sets

Let $\boldsymbol{\theta} \in \Theta$ be a k-vector of unknown parameters of a unknown population $\mathbb{P} \in \mathcal{P}$.

Confidence set

Let $C(\boldsymbol{X}) \subset \Theta$ be a measurable set depending only on the sample $\boldsymbol{X}$. If

$$\inf_{\mathbb{P} \in \mathcal{P}} \mathbb{P}(\boldsymbol{\theta} \in C(\boldsymbol{X})) \geq 1 - \alpha,$$

where $\alpha \in (0, 1)$ is a fixed constant, then $C(\boldsymbol{X})$ is called a confidence set for $\boldsymbol{\theta}$ with level of significance $1 - \alpha$.

Confidence Sets

- A confidence set is a random element that covers the unknown $\boldsymbol{\theta}$ with certain probability.
- $\inf_{P \in \mathcal{P}} \mathbb{P}(\boldsymbol{\theta} \in C(X))$ is called the confidence coefficient of $C(\boldsymbol{X})$, i.e., the worst-case coverage probability.
- The concepts of level of significance and confidence coefficient are very similar to the level of significance and size in hypothesis testing. More on this later.

Confidence interval

Consider a real-valued θ. If $C(\boldsymbol{X}) = [\underline{\theta}(\boldsymbol{X}), \overline{\theta}(\boldsymbol{X})]$, then $C(\boldsymbol{X})$ is called a confidence interval for θ. If $C(\boldsymbol{X}) = (-\infty, \overline{\theta}(\boldsymbol{X})]$ or $C(\boldsymbol{X}) = [\underline{\theta}(\boldsymbol{X}), \infty)$, then $C(\boldsymbol{X})$ is called a upper (or lower) confidence bound for θ.

- A confidence set (or interval) is also called a set (or an interval) estimator of θ.

Confidence Sets

Example: Confidence interval for normal mean with known σ^2. We only need to consider $\underline{\theta}(\bar{X}), \overline{\theta}(\bar{X})$ since $\bar{X}$ is sufficient. Consider the confidence intervals of the form $[\bar{X} - c, \bar{X} + c]$, where $c > 0$ is fixed. Note that

$$\mathbb{P}(\mu \in [\bar{X} - c, \bar{X} + c]) = \mathbb{P}(|\bar{X} - \mu| \leq c) = 1 - \Phi(-\sqrt{n}c/\sigma)$$

is independent of μ.

- We can choose a confidence interval with an arbitrarily large confidence coefficient (by letting $c \to \infty$), but the chosen confidence interval may be so wide that it is practically useless.
- When σ^2 is unknown, then $[\bar{X} - c, \bar{X} + c]$ has a confidence coefficient of 0 (by letting $\sigma \to \infty$), therefore, is not a good inference procedure.

Confidence Sets

The previous example suggests that

- A reasonable approach is to choose a level of significance $1-\alpha \in (0,1)$ (just like the level of significance in hypothesis testing) and find a confidence interval (set) with this level of significance.
- For all confidence intervals satisfying a significance level, the one with the shortest interval length is preferred.
  - For $C(\boldsymbol{X}) = [\underline{\theta}(\boldsymbol{X}), \overline{\theta}(\boldsymbol{X})]$, the length is $\overline{\theta}(\boldsymbol{X}) - \underline{\theta}(\boldsymbol{X})$.
  - For $C(\boldsymbol{X}) = (-\infty, \overline{\theta}(\boldsymbol{X})]$ or $C(\boldsymbol{X}) = [\underline{\theta}(\boldsymbol{X}), \infty)$, the length is infinite, we may consider the distance $\overline{\theta}(\boldsymbol{X}) - \theta$ (or $\theta - \underline{\theta}(\boldsymbol{X})$).

Confidence Sets

Example: Normal population with $\boldsymbol{\theta} = (\mu, \sigma^2)$ for a given α. We know that $(\bar{X}, S^2)$ is sufficient. Hence, we focus on $C(\bar{X}, S^2)$. Note that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1), \quad \frac{S^2}{\sigma^2/(n-1)} \sim \chi^2_1.$$

We can find constants $\tilde{c}_\alpha$, $c_{1,\alpha}$ and $c_{2,\alpha}$, such that

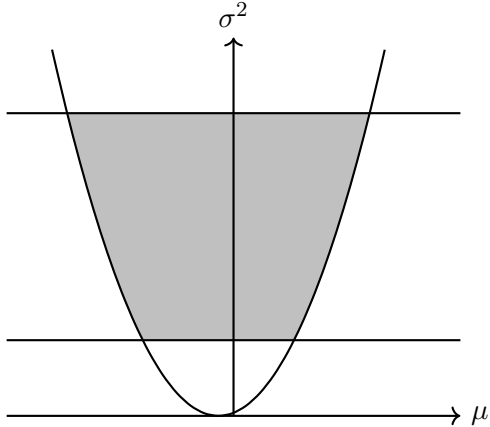
$$\mathbb{P}\left(-\tilde{c}_\alpha \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \tilde{c}_\alpha\right) = \sqrt{1-\alpha}, \quad \mathbb{P}\left(c_{1,\alpha} \leq \frac{S^2}{\sigma^2/(n-1)} \leq c_{2,\alpha}\right) = \sqrt{1-\alpha}.$$

Hence, by independence of $\bar{X}$ and S^2

$$\mathbb{P}\left(-\tilde{c}_\alpha \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \tilde{c}_\alpha, c_{1,\alpha} \leq \frac{S^2}{\sigma^2/(n-1)} \leq c_{2,\alpha}\right) = 1 - \alpha.$$

Rearrange terms, we have

$$\mathbb{P}\left(\frac{n(\bar{X}-\mu)^2}{\tilde{c}_\alpha} \le \sigma^2, \frac{(n-1)S^2}{c_{2,\alpha}} \le \sigma^2 \le \frac{(n-1)S^2}{c_{1,\alpha}}\right) = 1-\alpha, \quad \forall \boldsymbol{\theta}.$$



Asymptotic significance level

Let $\boldsymbol{\theta} \in \Theta$ be a k-vector of unknown parameters of a unknown population $\mathbb{P} \in \mathcal{P}$, and $C(\boldsymbol{X})$ be a confidence set for $\boldsymbol{\theta}$.

- If $\liminf_n \inf_{\mathbb{P}\in\mathcal{P}} \mathbb{P}(\boldsymbol{\theta} \in C(\boldsymbol{X})) \geq 1-\alpha$, then $1-\alpha$ is an asymptotic significance level of $C(\boldsymbol{X})$.
- If $\lim_n \inf_{\mathbb{P}\in\mathcal{P}} \mathbb{P}(\boldsymbol{\theta} \in C(\boldsymbol{X}))$ exists, then it is called a limiting confidence coefficient of $C(\boldsymbol{X})$.

Example: For a sample $X_1, \ldots, X_n$ from Uniform $(0, \theta)$, $Y = \max_i X_i$ is the MLE.

$$\mathbb{P}_\theta(\theta \in [aY, bY]) = \mathbb{P}_\theta(\frac{1}{a} \leq \frac{Y}{\theta} \leq \frac{1}{b}) = \frac{1}{a^n} - \frac{1}{b^n}$$

$$\mathbb{P}_\theta(\theta \in [Y+c, Y+d]) = \mathbb{P}_\theta(1 - \frac{d}{\theta} \leq \frac{Y}{\theta} \leq 1 - \frac{c}{\theta}) = \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n$$

What is the confidence coefficient and limiting confidence coefficient for each case?

Pivotal quantities

Perhaps the most popular method of constructing confidence sets is the use of pivotal quantities defined as follows.

Pivotal quantity

A known (Borel) function Q of $(\boldsymbol{X}, \boldsymbol{\theta})$ is called a pivotal quantity if and only if the distribution of $Q(\boldsymbol{X}, \boldsymbol{\theta})$ does not depend on $\mathbb{P} \in \mathcal{P}$.

- Note that a pivotal quantity depends on $\mathbb{P}$ through $\boldsymbol{\theta}$.
- A pivotal quantity is usually not a statistic, although its distribution is known.
- Different from ancillary statistics! Pivotal quantity need to contain and contain only the target parameter.

Pivotal quantities

With a pivotal quantity $Q(\boldsymbol{X}, \boldsymbol{\theta})$, we find cosntants a, b such that

$$\mathbb{P}(a \leq Q(\boldsymbol{X}, \boldsymbol{\theta}) \leq b) \geq 1 - \alpha.$$

Then we have a level $1 - \alpha$ confidence set

$$C(\mathbf{X}) = \{\boldsymbol{\theta} \in \Theta : a \leq Q(\boldsymbol{X}, \boldsymbol{\theta}) \leq b)\}$$

- If $Q(\boldsymbol{X}, \boldsymbol{\theta})$ has continuous pdf, then we can always choose a and b such that

  $$\mathbb{P}(a \leq Q(\boldsymbol{X}, \boldsymbol{\theta}) \leq b) = 1 - \alpha,$$

  holds for any α, so the confidence coefficient of $C(\mathbf{X})$ is $1 - \alpha$.
- If $Q(\boldsymbol{X}, \boldsymbol{\theta})$ is discrete, the confidence coefficient of $C(\mathbf{X})$ may not be $1 - \alpha$.

Computing $C(\boldsymbol{X})$

When $Q(\boldsymbol{X}, \boldsymbol{\theta})$, a and b are chosen, we need to compute the confidence set $C(\mathbf{X})$. This can be done by inverting $a \leq Q(\boldsymbol{X}, \boldsymbol{\theta}) \leq b$, i.e., finding $\boldsymbol{\theta}$ such that the inequality holds under a specific $\boldsymbol{X}$.

- If θ is real valued and $Q(\boldsymbol{X}, \theta)$ is monotone in θ when $\boldsymbol{X}$ is fixed, then

$$C(\boldsymbol{X}) = [\underline{\theta}(\boldsymbol{X}), \overline{\theta}(\boldsymbol{X})]$$

  for some $\underline{\theta}(\boldsymbol{X}), \overline{\theta}(\boldsymbol{X})$. Confidence interval!
- If $Q(\boldsymbol{X}, \theta)$ is not monotone in θ, it may be union of several intervals. Which is less preferred because it is harder to interpret.
- If $\boldsymbol{\theta}$ is multivariate, inverting $a \leq Q(\boldsymbol{X}, \boldsymbol{\theta}) \leq b$ can be complicated.
- In most cases where explicit form of $C(\boldsymbol{X})$ is not available, numerical calculation can help.

Example

Example: Location-scale families.

- μ unknown σ^2 known and $\theta = \mu$, then $\bar{X} - \mu$ is pivotal.

$$C(\boldsymbol{X}) = \{\mu : c_1 \leq \bar{X} - \mu \leq c_2\} = [\bar{X} - c_2, \bar{X} - c_1]$$

  The choice of c_1, c_2 is not unique. One popular choice is equal-tailed $c_1 = -c_2$.
- σ^2 unknown μ known and $\theta = \sigma^2$. Pivotal quantities S/σ, $(\bar{X} - \mu)/\sigma$, $\prod_i (X_i - \mu)/\sigma$...

$$C(\boldsymbol{X}) = \{\sigma : c_1 \leq S/\sigma \leq c_2\} = [S/c_2, S/c_1]$$

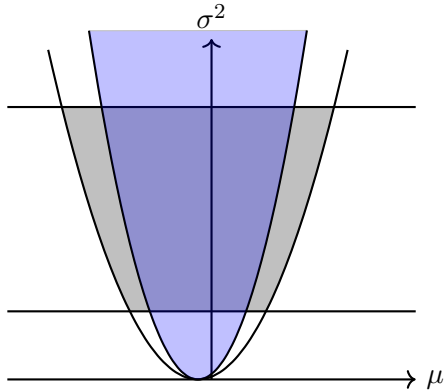
$$C(\boldsymbol{X}) = \{\sigma : c_1 \leq (\bar{X} - \mu)/\sigma \leq c_2\} = [(\bar{X} - \mu)/c_2, (\bar{X} - \mu)/c_1]$$

- σ^2 unknown μ unknown and $\theta = \sigma^2$. Pivotal quantity S/σ. Note that $(\bar{X} - \mu)/\sigma$, $\prod_i (X_i - \mu)/\sigma$ are not pivotal.

- σ unknown μ unknown and $\boldsymbol{\theta} = (\mu, \sigma^2)$. Pivotal $t(\boldsymbol{X}) = \sqrt{n}(\bar{X} - \mu)/S$, $\sqrt{n}(\bar{X} - \mu)/\sigma$...

$$C(\boldsymbol{X}) = \{(\mu, \sigma^2) : c_1 \leq (\bar{X} - \mu)/\sigma \leq c_2\}$$

If Normal, blue region is the $C(\boldsymbol{X})$ compared with our bounded confidence set.



Pivoting the CDF

The following provide a general method to find pivoral quantities.

Lemma

Let $\boldsymbol{T}(\boldsymbol{X}) = (T_1(\boldsymbol{X}), ..., T_s(\boldsymbol{X}))$ be independent statistics. Suppose that each T_i has a continuous CDF $F_{T_i,\boldsymbol{\theta}}$ indexed by $\boldsymbol{\theta}$. Then

$$Q(\boldsymbol{X}, \boldsymbol{\theta}) = \prod_{i=1}^{s} F_{T_i,\boldsymbol{\theta}}(T_i(\boldsymbol{X}))$$

is a pivotal quantity.

Proof. Because $F_{T_i,\boldsymbol{\theta}}(T_i(\boldsymbol{X}))$ are i.i.d. Uniform(0,1). □

- Naive choice is $T(\boldsymbol{X}) = X_1$ $(s = 1)$ or $\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{X}$ $(s = n)$, but that usually is not a good one. **Example:** Uniform$(0, \theta)$.

Pivoting the CDF

Corollary

Suppose θ and T in the lemma above are real-valued and let α_1 and α_2 be fixed positive constants such that $\alpha_1 + \alpha_2 = \alpha \leq 1/2$. Then the set

$$C(T) = \{\theta : \alpha_1 \leq Q(T, \theta) \leq 1 - \alpha_2\}$$

has probability $1 - \alpha$, where $Q(T, \theta) = F_{T,\theta}(T)$ is the pivotal quantity.

Proof. Because $F_{T,\theta}(T(\boldsymbol{X}))$ are i.i.d. Uniform(0,1). □

There is no guarantee that it is an interval.

We can construct confidence intervals for θ when $F_{T,\theta}(t)$ is monotone in θ.

Theorem

- *If $F_{T,\theta}(t)$ is nonincreasing in θ for each t, define $\underline{\theta}$ and $\overline{\theta}$ by*

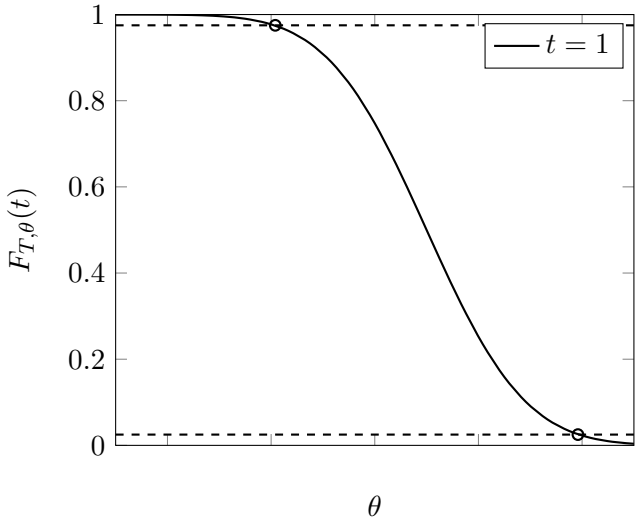
$$\overline{\theta} = \sup\{\theta : F_{T,\theta}(T) \geq \alpha_1\}, \quad \textit{and} \quad \underline{\theta} = \sup\{\theta : F_{T,\theta}(T-) \leq 1 - \alpha_2\}$$

- *If $F_{T,\theta}(t)$ is nondecreasing in θ for each t, define $\underline{\theta}$ and $\overline{\theta}$ by*

$$\underline{\theta} = \sup\{\theta : F_{T,\theta}(T) \geq \alpha_1\}, \quad \textit{and} \quad \overline{\theta} = \sup\{\theta : F_{T,\theta}(T-) \leq 1 - \alpha_2\}$$

Then $[\underline{\theta}, \overline{\theta}]$ is a $1 - \alpha$ confidence interval.

- For a special case where $F_{T,\theta}(t)$ is continuous and strictly monotone in θ, the proof follows immediately from the previous corollary.


1
0.8
0.6
0.4
0.2
0
$F_{T,\theta}(t)$
θ
$t = 1$

Remarks

- This works when the CDF $F_{T,\theta}(t)$ is not continuous, and we do not need strict monotonicity, see Theorem 7.1 of Jun Shao.
- When the parametric family has monotone likelihood ratio in T, then $F_{T,\theta}(t)$ is nonincreasing. Hence, we can find $1-\alpha$ confidence intervals.

Example: Location exponential

Example: If $X_1, \ldots, X_n$ is a sample from $f(x|\mu) = e^{-(x-\mu)}1_{\{x \geq \mu\}}$, then $Y = \min_i X_i$ is sufficient for μ with pdf $f_Y(y|\mu) = ne^{-n(y-\mu)}1_{\{y \geq \mu\}}$. Then

$$F_{Y,\mu}(y) = 1 - e^{-n(y-\mu)}, y \geq \mu,$$

is strictly decreasing in μ.

For a fixed α, find

$$F_{Y,\mu_U(y)}(y) = \int_{\mu_U(y)}^{y} ne^{-n(u-\mu_U(y))}du = \frac{\alpha}{2}, \quad 1-F_{Y,\mu_L(y)}(y) = \int_{y}^{\infty} ne^{-n(u-\mu_L(y))}du = \frac{\alpha}{2}.$$

We have

$$\mu_U(y) = y + \frac{1}{n}\log(1-\alpha/2), \quad \mu_L(y) = y + \frac{1}{n}\log(\alpha/2).$$

Thus, the $1-\alpha$ confidence interval is

$$\left[Y + \frac{1}{n}\log(\alpha/2), Y + \frac{1}{n}\log(1-\alpha/2)\right].$$

Inverting Acceptance Regions of Tests

Another popular method of constructing confidence sets is to use a close relationship between confidence sets and hypothesis tests.

For any test φ, the set $\{\boldsymbol{x} : \varphi(\boldsymbol{x}) \neq 1\}$ is called the acceptance region. Note that this terminology is not precise when φ is a randomized test.

Theorem

For each $\theta_0 \in \Theta$,let φ_{θ_0} be a test for $H_0 : \theta = \theta_0$ (versus some H_1) with significance level at most α and acceptance region $A(\theta_0)$. For each $\boldsymbol{x}$, define

$$C(\boldsymbol{x}) = \{\theta : x \in A(\theta_0)\}.$$

Then $C(\boldsymbol{X})$ is a level $1 - \alpha$ confidence set for θ. If φ_{θ_0} is nonrandomized and has size α for every θ_0, then $C(\boldsymbol{X})$ has confidence coefficient $1 - \alpha$.

Proof. The following holds for all $\theta_0 \in \Theta$,

$$\mathbb{P}(\theta_0 \in C(\boldsymbol{X})) = 1 - \mathbb{P}(\boldsymbol{X} \notin A(\theta_0)) = 1 - \mathbb{P}(\varphi_{\theta_0} = 1) \geq 1 - \mathbb{E}[\varphi_{\theta_0}] = 1 - \alpha.$$

Hence,

$$\inf_{\theta_0 \in \Theta} \mathbb{P}(\theta_0 \in C(\boldsymbol{X})) \geq 1 - \alpha.$$

If φ_{θ_0} is nonrandomized, then the inequality hold as equality.

Theorem

Let $C(\boldsymbol{X})$ be a confidence set for θ with confidence coefficient $1 - \alpha$. For any $\theta_0 \in \Theta$, define a region $A(\theta_0) = \{x : \theta_0 \in C(\boldsymbol{X})\}$, then the test $\varphi(\boldsymbol{X}) = 1 - \mathbb{1}_{A(\theta_0)}(\boldsymbol{X})$ has significance level at most α for testing $H_0 : \theta = \theta_0$ versus some H_1.

Inverting Acceptance Regions of Tests

- It is relative easy to construct a level α test.
- In general, there is no guarantee that the confidence set is an interval.
- Good tests usually result in good confidence sets.

Example: For a sample $X_1, \ldots, X_n$ from $\mathcal{N}(\mu, \sigma^2)$, consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. A size α test has rejection region $\{\mathbf{x} : |\bar{x} - \mu_0| > z_{\alpha/2}\sigma/\sqrt{n}\}$.

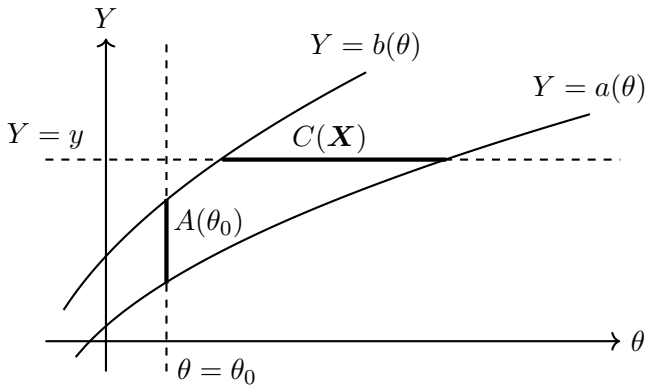
$$\begin{aligned} &\mathbb{P}\left(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu_0 \leq \bar{X} + z_{\alpha/2}\sigma/\sqrt{n} | \mu = \mu_0\right) = 1 - \alpha \quad \forall \mu_0 \\ \Rightarrow &\mathbb{P}_\mu\left(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}\right) = 1 - \alpha \end{aligned}$$

So the confidence coefficient is $1 - \alpha$.

One-sided test will give you a confidence upper bound or lower bound.

Illustration

Example: Suppose $A(\theta_0)) = \{Y : a(\theta_0) \leq Y \leq b(\theta_0)\}$ for some real valued statistic $Y(\boldsymbol{X})$ and some nondecreasing $a(\theta)$ and $b(\theta)$.



Binomial One-sided

Example: Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(p). Consider the following test, $H_0 : p = p_0$; $H_1 : p > p_0$.

- $T = \sum_{i=1}^n X_i \sim$ Binomial(n, p) is a sufficient statistic.
- The binomial distribution has monotone likelihood ratio, so the acceptance region of a UMP test is $A = \{t : t < m(p_0)\}$, where
- For a given α, find $m(p_0)$ such that

$$\sum_{y=m(p_0)+1}^{n} \binom{n}{y} p_0^y (1-p_0)^{n-y} \leq \alpha < \sum_{y=m(p_0)}^{n} \binom{n}{y} p_0^y (1-p_0)^{n-y}$$

Then $m(p)$ is an integer-valued, nondecreasing step-function of p. Define

$$\underline{p} = \inf\{p : m(p) \geq t\}$$

then a level $1-\alpha$ confidence interval for p is $(\underline{p}, 1]$. (The confidence coefficient may not be $1-\alpha$. Need randomized CI. Not discussed here.)

Good Tests and Good Confidence Intervals

The duality between hypothesis testing and confidence intervals suggests that better tests should give better confidence intervals. This is indeed the case.

- Let $C(\boldsymbol{X})$ be a confidence set constructed from a UMP test φ^* for $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ and let $C'(\boldsymbol{X})$ be a competing confidence set.
- Define $\varphi'(\boldsymbol{X}) = 1$ if $\theta_0 \notin C'(\boldsymbol{X})$ and 0 otherwise. Then φ' is a test with level at most α.
- φ^* is UMP, so for any $\theta > \theta_0$

$$\mathbb{P}_\theta(\theta_0 \notin C(\boldsymbol{X})) = \mathbb{E}_\theta[\varphi^*] \geq \mathbb{E}_\theta[\varphi'] = \mathbb{P}_\theta(\theta_0 \notin C'(\boldsymbol{X})).$$

- This implies that $C(\boldsymbol{X})$ has a smaller chance of covering any incorrect $\theta > \theta_0$.

$$\mathbb{P}_\theta(\theta_0 \in C(\boldsymbol{X})) \leq \mathbb{P}_\theta(\theta_0 \in C'(\boldsymbol{X})).$$

- In practice, we care about the (expected) length of the confidence interval, and $\mathbb{P}_\theta(\theta_0 \in C(\boldsymbol{X}))$ seem less relevant.
- But, using Fubini's theorem, there is a relation between the two.

Consider a simple case where $\theta \in \mathbb{R}$. Let $\lambda(A)$ denote the length of the confidence set. Then Fubini's theorem implies that

$$\begin{aligned}
\mathbb{E}_\theta[\lambda(C(\boldsymbol{X}) \cap (-\infty, \theta])] &= \mathbb{E}_\theta\left[\int_{-\infty}^{\theta} \mathbb{1}_{C(\boldsymbol{X})}(\theta_0) d\theta_0\right] \\
&= \int \int_{-\infty}^{\theta} \mathbb{1}_{C(\boldsymbol{x})}(\theta_0) d\theta_0 d\mathbb{P}_\theta(\boldsymbol{x}) \\
&= \int_{-\infty}^{\theta} \mathbb{P}_\theta(\theta_0 \in C(\boldsymbol{x})) d\theta_0 \\
&\leq \int_{-\infty}^{\theta} \mathbb{P}_\theta(\theta_0 \in C'(\boldsymbol{x})) d\theta_0 = \mathbb{E}_\theta[\lambda(C'(\boldsymbol{X}) \cap (-\infty, \theta])].
\end{aligned}$$

Shortest Confidence Interval of the form $[a, b]$

Definition (Unimodal)

A function $f(\cdot)$ is unimodal if there exists x^* such that it is nondecreasing when $x \leq x^*$ and nonincreasing when $x \geq x^*$.

Theorem

Let $f(\cdot)$ be a unimodal pdf. If a, b satisfy

(i) $\int_a^b f(x)dx = 1 - \alpha$

(ii) $f(a) = f(b) > 0$

(iii) $a \leq x^* \leq b$ *where x^* is a mode of $f(\cdot)$*

then $[a, b]$ is the shortest among all that satisfies (i).

For symmetric pdf like t and normal, we should use $t_{n-1,\alpha/2}$ and $z_{\alpha/2}$.

Optimizing Expected Length

Example: For normal with unknown σ, one may use the pivot $(\bar{X}-\mu)/(s/\sqrt{n})$, as we have shown.

- The CI is
$$\bar{x} - b\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} - a\frac{s}{\sqrt{n}}$$
- The length is $(b-a)s/\sqrt{n}$, which is a function of s.
- If we want to minimize the expected length, then $(b-a)\mathbb{E}S/\sqrt{n}$. The minimum is achieved when $a = -b = t_{n-1,\alpha/2}$.

Use the MLE to Construct Confidence Interval

We have seen that MLE is usually a good estimator of a parameter θ. We now briefly introduce the procedure to construct confidence intervals based on MLEs.

Key Questions:

- Is MLE consistent? I.e., do we have $\widehat{\theta}_n \xrightarrow{p} \theta$? (c.f. LLN)
- What is the (asymptotic) distribution of MLE? I.e., can we find the limiting distribution of $\sqrt{n}(\widehat{\theta}_n - \theta)$? (c.f. CLT)

Suppose the limiting distribution of $\sqrt{n}(\widehat{\theta}_n - \theta)$ is $\mathbb{P}$ independent of θ, then we can choose c_1 and c_2 such that

$$\mathbb{P}(c_1 \leq \sqrt{n}(\widehat{\theta}_n - \theta) \leq c_2) \approx 1 - \alpha.$$

Inverting the set will give us a confidence interval for θ.

Consistency of the MLE

Technical Conditions

1. Strong identifiability: For every $\epsilon > 0$, we have

$$\inf_{\tilde{\theta}:|\tilde{\theta}-\theta|\geq\epsilon} \mathsf{KL}(\theta, \tilde{\theta}) > 0,$$

where $\mathsf{KL}(\cdot, \cdot)$ is the Kullback-Leibler divergence[a]

$$\mathsf{KL}(\theta, \tilde{\theta}) = \mathbb{E}_\theta[\log(f_\theta(X)/f_{\tilde{\theta}}(X))].$$

2. Uniform LLN: Let $R_n(\theta, \tilde{\theta}) = \frac{1}{n}\sum_{i=1}^n \log\left(f_\theta(X_i)/f_{\tilde{\theta}}(X_i)\right)$, we have

$$\sup_{\tilde{\theta}} |R_n(\theta, \tilde{\theta}) - \mathsf{KL}(\theta, \tilde{\theta})| \xrightarrow{p} 0$$

[a] It can be viewed as a measure of the information discriminating between θ and $\tilde{\theta}$ when θ is the true value of the unknown parameter. If $\mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$, then $\mathsf{KL}(\theta, \tilde{\theta}) > 0$.

Consistency of the MLE

Theorem

Suppose that the Technical Conditions hold, then MLE is consistent.

Inconsistency of the MLE

- The MLE can fail to be consistent when the model is not strongly identifiable.
- Also when the uniform LLN fails. This typically happens when the parameter space is too large.

Example: Suppose $\{Y_{i,1}, Y_{i,2} \overset{i.i.d.}{\sim} N(\mu_i, \sigma^2)\}_{i=1}^n$ We want to estimate σ^2. The log-likelihood is

$$l(\sigma^2, \mu) = -n \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [(Y_{i,1} - \mu_i)^2 + (Y_{i,2} - \mu_i)^2]$$

MLE for μ_i is $(Y_{i,1} + Y_{i,2})/2$, and MLE for σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{2n} \sum_{i=1}^{n} [(Y_{i,1} - \widehat{\mu}_i)^2 + (Y_{i,2} - \widehat{\mu}_i)^2] = \frac{1}{4n} \sum_{i=1}^{n} [(Y_{i,1} - Y_{i,2})^2] \xrightarrow{p} \frac{\sigma^2}{2}.$$

Limiting Distribution of the MLE

Example: Bernoulli(p) MLE. The MLE is $\widehat{p}_n = \bar{X}_n$. We know by CLT that

$$\sqrt{n}\frac{\widehat{p}_n - p}{\sqrt{p(1-p)}} \Rightarrow N(0,1).$$

The asymptotic distribution is normal!

Recall that Fisher Information for Bernoulli is

$$\mathcal{I}(p) = \frac{1}{p(1-p)}.$$

We have

$$\sqrt{n}(\widehat{p} - p) \Rightarrow N(0, [\mathcal{I}(p)]^{-1}).$$

Limiting Distribution of the MLE

Example: Counter example. Uniform$(0, \theta)$, the MLE is $\widehat{\theta}_n = X_{(n)}$.

But we have shown that

$$n(\widehat{\theta}_n - \theta) \Rightarrow -\text{Exp}(1/\theta).$$

Not normal!

Remark

- Fisher information is not defined for Uniform distributions: Regularity conditions fail
  - The partial derivative of $f(X;\theta)$ with respect to θ exists almost everywhere.
  - The support of $f(X;\theta)$ does not depend on θ.

Limiting Distribution of the MLE

Sufficient Conditions

1. The dimension of the parameter space does not change with n fixed, i.e., $\boldsymbol{\theta} \in \mathbb{R}^d$. We have seen that if d grows the MLE need not even be consistent.
2. $f(x|\boldsymbol{\theta})$ is smooth (three times differentiable) in θ.
3. Conditions in Cremér-Rao holds. (The uniform example!)
4. The parameter θ is identifiable.
5. If the parameter space is restricted, i.e. $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ then θ is in the interior of the set Θ (i.e. cannot be on its boundary).

Theorem (Limiting distribution of MLE)

Under Sufficient Conditions,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \Rightarrow N(0, [\mathcal{I}(\boldsymbol{\theta})]^{-1}).$$

For a rigorous proof (under more general conditions), see Section 9.3 of Keener, Theoretical Statistics.

Asymptotic Confidence Set

Now, under the assumption that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \Rightarrow N(0, [\mathcal{I}(\boldsymbol{\theta})]^{-1}),$$

we can construct asymptotic confidence sets. Let's focus on $\theta \in \mathbb{R}$.

$$\sqrt{n\mathcal{I}(\theta)}(\widehat{\theta}_n - \theta) \Rightarrow N(0, 1),$$

This is called an approximate pivot.

$$\mathbb{P}_\theta(\sqrt{n\mathcal{I}(\theta)}|\widehat{\theta}_n - \theta| \leq z_{\alpha/2}) \to 1 - \alpha, \quad \text{as } n \to \infty.$$

Then a $1 - \alpha$ asymptotic confidence set is defined as

$$C(\boldsymbol{X}) = \{\theta \in \Theta : \sqrt{n\mathcal{I}(\theta)}|\widehat{\theta}_n - \theta| \leq z_{\alpha/2}\}.$$

Asymptotic Confidence Interval

$$C(\boldsymbol{X}) = \{\theta \in \Theta : \sqrt{n\mathcal{I}(\theta)}|\widehat{\theta}_n - \theta| \leq z_{\alpha/2}\}.$$

- Note that $I(\theta)$ depends on θ! So it is not necessarily a confidence interval.
- To avoid the trouble, if we assume that $I(\cdot)$ is continuous and bounded away from 0, then by continuous mapping theorem

$$\sqrt{\mathcal{I}(\widehat{\theta}_n)/\mathcal{I}(\theta)} \xrightarrow{p} 1.$$

- By Slutsky's theorem

$$\sqrt{n\mathcal{I}(\widehat{\theta}_n)}(\widehat{\theta}_n - \theta) \Rightarrow N(0,1).$$

- We have an asymptotic confidence interval (this is called the **Wald interval**)

$$\left(\widehat{\theta}_n - \frac{z_{\alpha/2}}{\sqrt{n\mathcal{I}(\widehat{\theta}_n)}}, \widehat{\theta}_n + \frac{z_{\alpha/2}}{\sqrt{n\mathcal{I}(\widehat{\theta}_n)}}\right)$$

Remark

Our asymptotic confidence interval requires calculation of Fisher Information. In addition, it might be argued that confidence intervals should be based solely on the shape of the likelihood function, and not on quantities that involve an expectation like Fisher Information.

- Recall that
$$\mathcal{I}(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2 l(\theta|X)}{\partial \theta^2} \right] = -\mathbb{E}_\theta \left[l''(\theta|X) \right].$$
One may expect that
$$-l''(\widehat{\theta}|\boldsymbol{X})/n \xrightarrow{p} \mathcal{I}(\theta).$$
- Hence, we should have
$$\sqrt{-l''(\widehat{\theta}|\boldsymbol{X})}(\widehat{\theta}_n - \theta) \Rightarrow N(0,1).$$
- $-l''(\theta|\boldsymbol{X})$ is called the observed Fisher Information. Hence,
$$\left(\widehat{\theta}_n - \frac{z_{\alpha/2}}{\sqrt{-l''(\widehat{\theta}|\boldsymbol{X})}}, \widehat{\theta}_n + \frac{z_{\alpha/2}}{\sqrt{-l''(\widehat{\theta}|\boldsymbol{X})}} \right)$$

One More Thing...

The previous confidence interval relies on the log-likelihood only through $\widehat{\theta}_n$ and the curvature at $\widehat{\theta}_n$. Can we use the full shape of the likelihood function?

- Taylor expansion of $l(\theta|\boldsymbol{X})$ around $\widehat{\theta}_n$ (Note that $l'(\widehat{\theta}_n|\boldsymbol{X}) = 0$.)

$$2l(\widehat{\theta}_n|\boldsymbol{X}) - 2l(\theta|\boldsymbol{X}) \approx \left[\sqrt{-l''(\theta|\boldsymbol{X})}(\widehat{\theta}_n - \theta)\right]^2$$

- We expect that

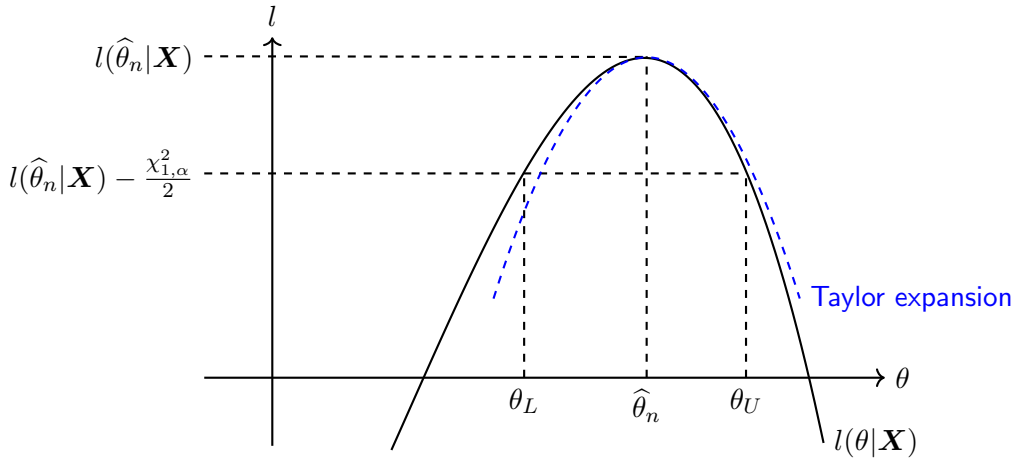
$$2l(\widehat{\theta}_n|\boldsymbol{X}) - 2l(\theta|\boldsymbol{X}) \Rightarrow Z^2 \sim \chi^2_1.$$

- Now, we define

$$C(\boldsymbol{X}) = \{\theta \in \Theta : 2l(\widehat{\theta}_n|\boldsymbol{X}) - 2l(\theta|\boldsymbol{X}) \leq \chi^2_{1,\alpha}\}.$$

- This is called a profile confidence interval.

Profile Confidence Interval



See also Figure 7.2 in Jun Shao for 2-D example.

Example: Poisson population. Suppose $X_1, \ldots, X_n$ are i.i.d. from a Poisson distribution with mean θ. Then

$$l(\theta|\boldsymbol{X}) = n\bar{X}\log\theta - n\theta - \log\left(\prod_{i=1}^{n} X_i!\right).$$

The MLE is $\widehat{\theta}_n = \bar{X}$ and $\mathcal{I}(\theta) = 1/\theta$. We have the following confidence regions.

- $C_1(\boldsymbol{X}) = \{\theta > 0 : \sqrt{n/\theta}|\widehat{\theta}_n - \theta| < z_{\alpha/2}\} = \{\theta > 0 : \widehat{\theta}_n^2 - 2\widehat{\theta}_n\theta + \theta^2 < z_{\alpha/2}^2\theta/n\}$. End points always positive.
- $C_2(\boldsymbol{X}) = \left(\widehat{\theta}_n - \frac{z_{\alpha/2}}{\sqrt{n\mathcal{I}(\widehat{\theta}_n)}}, \widehat{\theta}_n + \frac{z_{\alpha/2}}{\sqrt{n\mathcal{I}(\widehat{\theta}_n)}}\right) = \left(\bar{X} - \frac{z_{\alpha/2}}{\sqrt{n/\bar{X}}}, \bar{X} + \frac{z_{\alpha/2}}{\sqrt{n/\bar{X}}}\right)$, note that lower endpoint for this confidence interval will be negative if $\bar{X} \approx 0$.
- $C_3(\boldsymbol{X}) = \{\theta > 0 : 2l(\widehat{\theta}_n|\boldsymbol{X}) - 2l(\theta|\boldsymbol{X}) \leq \chi^2_{1,\alpha}\} = \{\theta \in \Theta : \theta - \bar{X} - \bar{X}\log(\theta/\bar{X}) < \chi^2_{1,\alpha}/(2n)\}$. This is an interval, but must be calculated numerically. End points always positive (because the formulation uses log-likelihood!).

Higher Dimension

Recall that under certain conditions, the MLE is asymptotically normal

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \Rightarrow N(0, [\mathcal{I}(\boldsymbol{\theta})]^{-1}).$$

Sometimes we may be interested in a function $g(\boldsymbol{\theta})$.

Multivariate delta method

If $g : \Theta \to \mathbb{R}$ is differentiable at $\boldsymbol{\theta}$, $\mathcal{I}(\theta)$ is positive definite, and $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \Rightarrow N(0, [\mathcal{I}(\boldsymbol{\theta})]^{-1})$, then

$$\sqrt{n}\left(g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})\right) \Rightarrow N(0, \nu^2(\theta)),$$

where

$$\nu^2(\boldsymbol{\theta}) = (\nabla g(\boldsymbol{\theta}))^T \mathcal{I}(\theta)^{-1} \nabla g(\boldsymbol{\theta})$$

Still proved by Taylor expansion.

A simple confidence set can be obtained if ν_n^2 is a consistent estimator of $\nu(\boldsymbol{\theta})$

$$\left(g(\widehat{\boldsymbol{\theta}}_n) - \frac{z_{\alpha/2}\nu_n^2}{\sqrt{n}}, g(\widehat{\boldsymbol{\theta}}_n) + \frac{z_{\alpha/2}\nu_n^2}{\sqrt{n}} \right)$$

What if we want a confidence set for $\boldsymbol{\theta}$?

- $(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^T \mathcal{I}(\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ will follow some χ_p^2 distribution for p-dimensional $\boldsymbol{\theta}$, then we may be able to construct a confidence set based on chi-squared distribution. We shall see this when we discuss regression models.

Theorem

For a p-dimensional normal random vector $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \Sigma)$ with a positive definite Σ, we have

$$(\boldsymbol{Y} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) \sim \chi_p^2.$$

Proof. If $\Sigma = U\Lambda U^T = U\Lambda^{1/2}(U\Lambda^{1/2})^T$ is an eigendecomposition where the columns of U are the eigenvectors and Λ is a diagonal matrix of the eigenvalues, then

$$\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \Sigma) \Leftrightarrow \boldsymbol{Y} \sim \boldsymbol{\mu} + U\Lambda^{1/2}\boldsymbol{Z}, \quad \text{where} \quad \boldsymbol{Z} \sim N(\boldsymbol{0}, I).$$

Hence,

$$(\boldsymbol{Y} - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) = \boldsymbol{Z}^T\Lambda^{1/2}U^T(U\Lambda U^T)^{-1}U\Lambda^{1/2}\boldsymbol{Z} = \boldsymbol{Z}^T\boldsymbol{Z} \sim \chi^2_p.$$

Remarks

Let us caution that in the construction of these asymptotic confidence intervals, a number of different approximations are being made:

- MLE is asymptotically normally distributed.
- $(\mathcal{I}(\widehat{\theta}))^{-1}$ approximates $(\mathcal{I}(\theta))^{-1}$.
- When we are interested in a function $g(\boldsymbol{\theta})$, $g(\widehat{\boldsymbol{\theta}})$ is approximated by its Taylor expansion.

Given the right condition, these asymptotic confidence sets are all valid in the limit $n \to \infty$, but their accuracy is not guaranteed for the finite sample size n for any given problem.

- Coverage of asymptotic confidence intervals should be checked by simulation. They might be severely overconfident for small n.

Example: Poisson(θ). We have $\sqrt{n}(\widehat{\theta} - \theta) \to N(0, \theta)$. Use Wald interval

$$\left\{ -z_{\alpha/2} \leq \frac{\sqrt{n}(\widehat{\theta} - \theta)}{\sqrt{\widehat{\theta}}} \leq z_{\alpha/2} \right\}$$

For various values of θ and n, the table below shows the simulated true probabilities that the 90% and 95% confidence intervals constructed cover θ:

	Desired coverage 90%			Desired coverage 95%		
	$\theta = 0.1$	$\theta = 1$	$\theta = 5$	$\theta = 0.1$	$\theta = 1$	$\theta = 5$
$n = 10$	0.63	0.91	0.90	0.63	0.93	0.95
$n = 30$	0.79	0.89	0.90	0.80	0.93	0.95
$n = 100$	0.91	0.90	0.90	0.93	0.94	0.95

- The table is obtained by simulating n i.i.d. Poisson, calculate the CI, check if CI cover θ, repeat for 10^6 times, and report the fraction of simulations for which θ is covered.

Statistical Functional

We have been focusing on parametric models. Now we turn to nonparametric inference.

Given any distribution, we may want to estimate the following quantity

Statistical functional

A statistical functional is a map that maps a distribution $\mathbb{P}$ to a real number (or vector).

- The mean: $\psi(\mathbb{P}) = \int x dF(x)$.
- The variance, the median, quantiles...
- For parametric models, statistical functional is simply a function $\psi(\theta)$.

Plug-In Estimator

Given data $\boldsymbol{X}$, we can find its empirical CDF

$$F_n(t) = \frac{1}{n}\sum_i \mathbb{1}_{X_i \le t}.$$

We then have an estimator for the expectation of any function g

$$\widehat{g} = \int g(x) dF_n(x) = \frac{1}{n}\sum_i g(X_i).$$

Applying this idea, we have

Plug-in estimator

Let $\mathbb{P}_n$ be the empirical distribution given $\boldsymbol{X}$,

$$\widehat{\psi}_n = \psi(\mathbb{P}_n).$$

Plug-In Estimator

Example:

- Mean: $\bar{X}$.
- Variance: $\frac{1}{n}\sum(X_i - \bar{X})^2$.
- Covariance: $\frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})$.
- Expectation of $g(x)$: $\frac{1}{n}\sum_i g(X_i)$.

Question: How do we get confidence intervals for statistical functional based on plug-in estimators?

What Do People Do in Practice

In practice,

- find an estimator $\widehat{\psi}$ for a statistical functional ψ
- estimate the standard deviation $\widehat{\sigma}$ of $\widehat{\psi}$ (this is called the standard error of $\widehat{\psi}$)
- assume $(\widehat{\psi} - \psi)/\widehat{\sigma} \to N(0,1)$; (Note that we are no longer dealing with paramteric family, so Fisher Information is not defined.)
- construct a $1-\alpha$ confidence interval (for $\alpha = 0.05$, $z_{0.025} = -1.96$)

$$[\widehat{\psi} - 1.96\widehat{\sigma}, \widehat{\psi} + 1.96\widehat{\sigma}]$$

- This is why we say something is significant if it is 2σ away from the mean.
- This simple method performs quite well in practice, especially when n is large.
- Need a good way to estimate the standard error of ANY estimator.

Monte-Carlo – A Naive Approach

Note that an estimator $\widehat{\psi}_n$ is a function of $\boldsymbol{X} = (X_1, \ldots, X_n)$. So it's variance is a statistical functional for a n-dimensional random vector.

- Draw n i.i.d. samples and calculate $\widehat{\psi}_n$, then we have one observation $\widehat{\psi}_{n,1}$.
- Draw i.i.d. samples in groups of n for B times, then we have B observations $\widehat{\psi}_{n,i}$.
- Use plug-in estimator

$$\frac{1}{B}\sum_i (\widehat{\psi}_{n,i} - \overline{\widehat{\psi}_n})^2$$

The problem is that we then need $n \times B$ number of random samples. This means that we need to be able to simulate more data.

What if we are given only n samples?

Bootstrap

Bootstrap sample

Suppose we have a sample $X_1, \cdots, X_n$ from PDF F. A bootstrap sample is $(\tilde{X}_1, \cdots, \tilde{X}_n)$, where $\tilde{X}_i$ are i.i.d. from the empirical distribution:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}.$$

- We can always draw samples from the empirical distribution. Think of it as drawing samples uniformly from $X_1, \cdots, X_n$ with replacement.

Consider B bootstrap samples

$$
\begin{aligned}
(X_1^{(1)}, \cdots, X_n^{(1)}) &\to \widehat{\psi}_n^{(1)} \\
(X_1^{(2)}, \cdots, X_n^{(2)}) &\to \widehat{\psi}_n^{(2)} \\
&\vdots \\
(X_1^{(B)}, \cdots, X_n^{(B)}) &\to \widehat{\psi}_n^{(B)}
\end{aligned}
$$

Then one can obtain B samples of $\widehat{\psi}_n$: $\widehat{\psi}_n^{(1)}, \ldots, \widehat{\psi}_n^{(B)}$. The standard error estimated by the sample standard deviation.

Failure of Bootstrap

Example: As usual, when we need a counterexample, we try the Uniform$(0, \theta)$. Consider estimating the distribution of

$$n\left(\theta - X_{(n)}\right).$$

Note that it is a collection of statistical functionals. We know that this is an Exp(θ). We use the bootstrap estimation, i.e., the distribution of

$$n\left(X_{(n)} - X^B_{(n)}\right)$$

The MLE is $X_{(n)}$. For the bootstrap sample, it will contain $X_{(n)}$ with probability

$$1 - (1 - 1/n)^n \approx 0.63.$$

So the bootstrap distribution puts a mass $.63$ at 0. If θ large, the density of Exp(θ) at 0 can be arbitrarily small (e.g. $< .63$).

- Better use parametric bootstrap, if you know it is uniform.

Parametric Bootstrap

The bootstrap method can be applied to parametric model, $\mathbb{P}_{\boldsymbol{\theta}}$.

- It is useful if we know the parametric family, but not allowed to draw more samples from the true distribution.

Suppose we have a sample $X_1, \cdots, X_n$. Estimate $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}_n$.

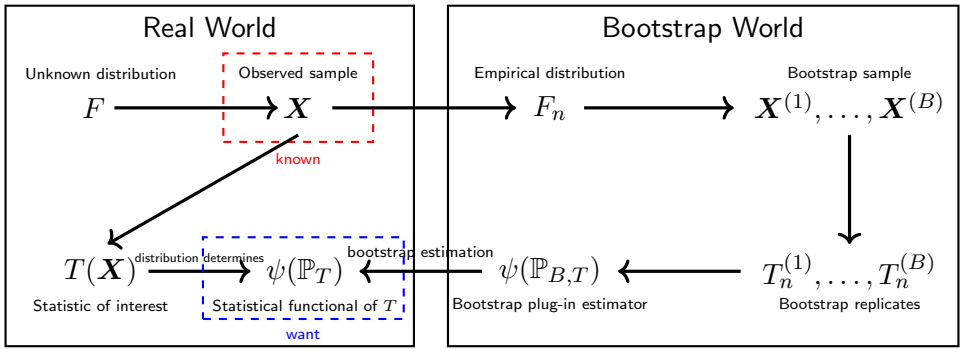
Bootstrap sample (parametric)

Suppose we have a sample $X_1, \cdots, X_n$ from PDF F. A parametric bootstrap sample is $(\tilde{X}_1, \cdots, \tilde{X}_m)$, where $\tilde{X}_i$ are i.i.d. from the $\mathbb{P}_{\widehat{\boldsymbol{\theta}}_n}$.

The rest is exactly the same.

- Nonparametric bootstrap is particularly useful if we are uncertain about the parametric family (model misspecification).

Does Bootstrap Work?



- Two asymptotic regimes: $B \to \infty$, $n \to \infty$.
- It is quite involved to prove, Read Efron's paper and follow-ups:
- B. Efron, Better Bootstrap Confidence Intervals, JASA, 1984.

Bootstrap Confidence Interval

- **Method 1.** If $\widehat{\theta}_n$ is asymptotically normal, then a $1-\alpha$ confidence interval is

$$\widehat{\theta}_n \pm z_{\alpha/2}\widehat{\sigma}_n$$

where $\widehat{\sigma}_n$ is the bootstrap estimate of the standard error.

- **Method 2: The Percentile Interval** is

$$\left(\widehat{\theta}_{(\alpha/2)}, \widehat{\theta}_{(1-\alpha/2)}\right)$$

where $\widehat{\theta}_{(\tau)}$ is the sample τ-quantile of the B bootstrap values of $\widehat{\theta}_n$.

Bayes Estimation: Introduction

- So far, we have treated the parameter θ as unknown but deterministic.
- The sample $\boldsymbol{X}$ is random and the distribution depends on θ.
- Statistical inference is about learning the unknown parameters.
- This is the frequentists' view of statistics, which is the popular view.
- In Bayesian view, statistical inference is about belief revision: I have some belief about the parameter, and I may revise it when I observe the data.

Bayes Estimators

Mathematically speaking, a belief is described by a probability distribution $\pi(\theta)$ on Θ, called the *the prior distribution*, or simply the prior.

- The probability distribution assigns more weight to the parameters which, according to the statistician's belief, are more likely to be true compared to other values.
- If the statistician has no particular preference for any value, then a prior to use may be the uniform distribution.

A Toy Example

Example: Suppose that a newborn Bayesian baby has never seen a sunrise before. Let $X = 1$ if the sun rises tomorrow, and 0 otherwise. We assume that X follows a Bernoulli(p) distribution, where p is unknown.

The Bayesian baby wants to estimate the probability $p \in \Theta = [0, 1]$ that the sun will rise tomorrow.

- With no prior knowledge of how sun behaves, he chooses Uniform$(0, 1)$ as the prior distribution $\pi(p)$ for p.
- Conditional on the value of p, the pdf of X is denoted by $f(x|p)$, this is called the *likelihood*.

A Toy Example – Updating Belief

Now on the second day of his life, he saw the sunrise for the first time, so that $X = 1$. How does he revise his prior belief?

- With $\pi(p)$ and $f(x|p)$, we can write down the joint distribution of p and X

$$f(x, p) = \pi(p) f(x|p).$$

- We ask the probability distribution of p, conditional on the observation $X = x$

$$f(p|x) = \frac{f(x, p)}{f(x)} = \frac{\pi(p) f(x|p)}{f(x)}.$$

This distribution is called *the posterior distribution*. It describes the updated belief of the Bayesian baby, *after* observing the sunrise.

A Toy Example – Estimation of the Parameter

To calculate the posterior distribution

$$f(p|x) \propto f(x,p) = \pi(p)f(x|p) = 1 \times p^x(1-p)^{(1-x)} \sim \text{Beta}(1+x, 2-x).$$

The posterior is a Beta distribution!

- Once the distribution of $f(p|\boldsymbol{x})$ has been obtained, one can use $\mathbb{E}[p|\boldsymbol{x} = \boldsymbol{X}]$ to estimate the parameter p.
- Since $x = 1$, the posterior is $\text{Beta}(2, 1)$ and the estimation is $2/3$.
- Compare with the estimation from the prior uniform belief with mean $1/2$, the Bayesian baby is now more confident that the sun will rise tomorrow.

Bayes Estimator – Binomial

Now that the Bayesian baby have updated his belief to a Beta(α, β). He wants to update his belief once again after observing multiple days.

- The parameter is p, with prior $\pi(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$.
- Given p, we have i.i.d. samples $X_1, X_2, \dots X_n \sim$ Bernoulli(p). Let $Y = \sum_{i=1}^n X_i \sim$Binomial$(n, p)$.
- Let $y = \sum_{i=1}^n x_i$. The posterior is

$$f(p|\mathbf{x}) \propto f(\mathbf{x}, p) = \binom{n}{y} p^y (1-p)^{n-y} \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \propto p^{\alpha+y-1}(1-p)^{n-y+\beta-1}$$

- The posterior is Beta$(y+\alpha, n-y+\beta)$. Its mean is

$$\hat{p} = \frac{y+\alpha}{\alpha+\beta+n} = \frac{n}{\alpha+\beta+n}\frac{y}{n} + \frac{\alpha+\beta}{\alpha+\beta+n}\frac{\alpha}{\alpha+\beta}$$

Bayes Estimator – Normal

- Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$. σ is known and the prior $\mu \sim \mathcal{N}(a, b^2)$.
- The posterior

$$\begin{aligned} f(\mu|\mathbf{x}) &\propto \exp\left(-\frac{(\mu-a)^2}{2b^2}\right)\exp\left(-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\left(\frac{1}{b^2}+\frac{n}{\sigma^2}\right)\mu^2 - 2\left(\frac{a}{b^2}+\frac{\sum_{i=1}^n x_i}{\sigma^2}\right)\mu\right)\right) \end{aligned}$$

- The posterior is still normal, with

$$\mathbb{E}[\mu|x] = \frac{nb^2}{nb^2+\sigma^2}\bar{X} + \frac{\sigma^2}{nb^2+\sigma^2}a \qquad \mathrm{Var}(\mu|x) = \frac{\sigma^2 b^2}{\sigma^2+nb^2}$$

Conjugate Family

Previous examples shows that, *given Bernoulli (normal) samples*, Beta (normal) prior is a conjugate family for Bernoulli (normal) distributions.

Definition (Conjugate Family)

Let $\mathcal{F} = \{f(\cdot|\theta) : \theta \in \Theta\}$ denote a class of likelihoods. A class Π of prior distributions is a conjugate family for $\mathcal{F}$ if the posterior distribution for any $f \in \mathcal{F}$ is in the class Π.

It greatly simplifies the computation (need only to update the parameters).

- Pros for Bayesian:
  - Having a distribution rather than a point. No need for large sample.
  - A learning procedure
- Cons for Bayesian:
  - No systematic way to set a prior;
  - Computationally intensive if conjugate family is not available.

Conjugate Family – Exponential Family

For exponential family, finding conjugate family is usually easy.

- Prior:

$$f_0(\boldsymbol{\eta}|\boldsymbol{\tau}, n_0) = H(\boldsymbol{\tau}, n_0) \exp(\boldsymbol{\eta}^T \boldsymbol{\tau} - n_0 A(\boldsymbol{\eta})).$$

- Likelihood:

$$f(\boldsymbol{x}|\boldsymbol{\eta}) = \left(\prod_{n=1}^{N} h(x_n) \right) exp \left(\boldsymbol{\eta}^T \left(\sum_{n=1}^{N} T(x_n) \right) - NA(\boldsymbol{\eta}) \right).$$

- Posterior:

$$f(\boldsymbol{\eta}|\boldsymbol{x}, \boldsymbol{\tau}, n_0) \propto exp \left(\boldsymbol{\eta}^T \left(\boldsymbol{\tau} + \sum_{n=1}^{N} T(x_n) \right) - (n_0 + N) A(\boldsymbol{\eta}) \right).$$

Bayesian Intervals

- So far, we have adopted the classical view of CI, that is, a random interval covers the parameter.
- In Bayesian view, the parameter is random, and we can say its probability to fall inside an interval.
- We use credible set instead of confidence set to make the distinction.
- If $\pi(\theta|\mathbf{x})$ is the posterior distribution of θ when $\mathbf{X} = \mathbf{x}$, then we want to find a credible set A

$$\mathbb{P}(\theta \in A|\mathbf{x}) = \int_A \pi(\theta|\mathbf{x})d\theta$$

Poisson Credible Set

Example: Suppose $X_i \sim Poisson(\lambda)$. Assume λ has a prior, Gamma distribution $\Gamma(a, b)$ where a is integer. The posterior after observing $\mathbf{X} = \mathbf{x}$ where $\sum x_i = y$ is

$$\pi\left(\lambda \middle| \sum x_i = y\right) = \Gamma\left(a + y, \frac{b}{nb+1}\right)$$

We can find the $\alpha/2$ and $1-\alpha/2$ quantile of the Gamma distribution. Alternatively, since a is integer[1], $\frac{2(nb+1)}{b}\lambda \sim \chi^2_{2(a+y)}$. Splitting the probability equally,

$$\lambda_L = \frac{b}{2(nb+1)}\chi^2_{2(a+y),1-\alpha/2}, \quad \lambda_U = \frac{b}{2(nb+1)}\chi^2_{2(a+y),\alpha/2}$$

[1] $X \sim \chi^2$ is a special case of gamma distribution, $X \sim \Gamma(k/2, 1/2)$.

Bayesian Optimality

Let $\pi(\theta|\mathbf{x})$ be the posterior of θ when $\mathbf{X} = \mathbf{x}$, we want to find $C(\mathbf{x})$ such that

i $\int_{C(\mathbf{x})} \pi(\theta|\mathbf{x})d\mathbf{x} = 1 - \alpha$

ii length$(C(\mathbf{x})) \leq$ length$(C'(\mathbf{x}))$ for any other set satisfying $\int_{C'(\mathbf{x})} \pi(\theta|\mathbf{x})d\mathbf{x} \geq 1 - \alpha$

We can apply a previous theorem to obtain

Corollary

If the the posterior pdf $\pi(\theta|\mathbf{x})$ is unimodal, the shortest credible interval for θ is

$$\{\theta : \pi(\theta|\mathbf{x}) \geq k\} \quad \textit{where} \quad \int_{\{\theta:\pi(\theta|\mathbf{x})\geq k\}} \pi(\theta|\mathbf{x})d\theta = 1 - \alpha.$$

The credible set is called highest posterior density (HPD) region.

Reading Materials

- Casella and Berger, Statistical Inference, Chapter 9.
- Keener, Theoretical Statistics, Chapter 9.1-9.5, 12.4.
- Jun Shao, Mathematical Statistics, Section 2.4.3, 7.1-7.2.2,7.3.1-7.3.3.