Lecture 11: Confidence Set

Ma Xuejun

School of Mathematical Sciences

Soochow University

https://xuejunma.github.io

Outline

Confidence Sets

- Related to the hypothesis testing problem, an interesting topic is the confidence sets.
- In point estimation, our estimator is $T(X_1, \cdots, X_n)$
- Once we observed the data, our estimate is $T(X_1, \cdots, X_n)$. It is consistent (close to the truth), yet it does not equal to the truth.
- Moreover, in most cases, $P(T(X_1, \cdots, X_n) = \theta_0) = 0$!

Definition: Confidence Intervals

An interval estimate for θ, is any pair of function $L : X^n \to \mathbb{R}$, $U : X^n \to \mathbb{R}$, such that $L(x_{1:n}) \leq U(x_{1:n})$, any $x_{1:n} \in X^n$. The random interval $|L(X_{1:n}), U(X_{1:n})|$ is called an *interval estimator*.

- For an interval, we can claim the probability that it contains the true parameter.
- It is called the *coverage probability* of an interval estimator that

$$P(\theta \in [L(X_{1:n}), U(X_{1:n})]; \theta).$$

$\inf_{\theta \in \Theta} P(\theta \in [L(X_{1:n}), U(X_{1:n})]; \theta)$ is called the $confidence\ coefficient$.

Example

Let $X_i \overset{i.i.d}{\sim} Unif[0,\theta], i = 1, \cdots, n$. Set $Y = X_{(n)}$. We are interested in an interval estimator for θ. Consider the interval with the form$[aY, bY]$ for some $1 \leq a < b$, Then,

$$P(aY \leq \theta \leq bY; \theta) = P\Big(\frac{1}{b} \leq Y/\theta \leq \frac{1}{a}; \theta\Big).$$

The CDF of Y is

$$P(Y \leq c) = \Big(\frac{c}{\theta}\Big)^n, \quad P\Big(\frac{Y}{\theta} \leq c\Big) = P(Y \leq c\theta) = c^n.$$

Therefore, the coverage probability is

$$P(aY \leq \theta \leq bY; \theta) = (1/a)^n - (1/b)^n.$$

The confidence coeffient is the same.

Question: Is the confidence coefficient always the same with the coverage probability? Answer: No!

Example,II

Still consider the previous example. Now we consider the confidence interval with the form $[Y+c, Y+d], 0 \leq c < d$. Now the coverage probability is

$$\begin{aligned} P(Y+c \leq \theta \leq Y+d; \theta) &= P(\theta - d \leq Y \leq \theta - c; \theta) \\ &= \Big(\frac{\theta - c}{\theta}\Big)^n - \Big(\frac{\theta - d}{\theta}\Big)^n \\ &= \Big(1 - c/\theta\Big)^n - \Big(1 - d/\theta\Big)^n \end{aligned}$$

The coverage probability changes with θ. Note that

$$\lim_{\theta \to \infty} P(\theta \in [Y+c, Y+d]; \theta) = 0.$$

So the confidence coefficient is 0.

Confidence Sets

General methods to get the confidence sets:

- Probability Inequalities
- Inverting a test
- Pivots

Review of Probability Inequalities

- Markov Inequality: for non-negative random variable X,

$$P(X \geq a) \leq \frac{E[X]}{a}$$

- Chebyshev's inequality. Let $\mu = E[X]$ and $\sigma^2 = Var(X)$. Then,

$$P(|X - \mu| \geq t) \leq \sigma^2 / t^2$$

- Normal Tail Inequality. Let $X \sim N(0, 1)$, then we have

$$P(|X| > \epsilon) \leq \frac{2e^{-\epsilon^2/2}}{\epsilon}$$

Proof. Set $Y = |X| \cdot 1\{|X| > \epsilon\}$. Then $P(|X| > \epsilon) = P(Y > \epsilon)$.

$$E[Y] = 2\int_{\epsilon}^{\infty} y \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = \frac{2}{\sqrt{2\pi}} (-e^{-y^2/2})|_{\epsilon}^{\infty} \leq 2e^{-\epsilon^2/2}.$$

With Markov Inequality,

$$P(|X| > \epsilon) = P(Y > \epsilon) \leq \frac{E[Y]}{\epsilon} < \frac{2e^{-\epsilon^2/2}}{\epsilon}.$$

Probability Inequalities

- Chernoff's inequality. Let X be a random variable. For $t \geq 0$,

$$P(|X| > \epsilon) = P(e^{tX} > e^{t\epsilon}) \leq e^{-t\epsilon} E[e^{tX}] \Rightarrow P(|X| > \epsilon) \leq \inf_{t \geq 0} e^{-t\epsilon} E[e^{tX}]$$

- Hoeffding's inequality. Let $X_1, \cdots, X_n$ be i.i.d. r.v.'s with mean μ and $a \leq X_i \leq b$.

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2}, \quad \epsilon > 0.$$

Example. For i.i.d. Bernoulli(p) random sample $X_1, \cdots, X_n$, we have $E[X] = p$ and they are bounded by 0 and 1. So,

$$P(|\bar{X}_n - p| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

- Bernstein's Inequality. Let $X_1, \cdots, X_n$ be i.i.d. r.v.'s with mean μ, variance σ^2 and $a \leq X_i \leq b$. Then we have

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{\frac{n\epsilon^2}{2(\sigma^2 + (b-a)\epsilon)}} \quad \epsilon > 0.$$

For the r.v.'s that concentrate in a small interval, this bound is more helpful.

Confidence Intervals

- Let $X_1, \cdots, X_n \sim Bernoulli(p)$. By Hoeffding's inequality,

$$P(|\bar{X}_n - p| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

  So, to construct a confidence interval with confidence coefficient $1-\alpha$, we let $\alpha = 2e^{-2n\epsilon^2}$, and solve it with $\sqrt{\log(2/\alpha)/2n}$. For the interval $[\bar{X} - \epsilon, \bar{X} + \epsilon]$, we have

$$\begin{aligned} P(\bar{X} - \epsilon \leq p \leq \bar{X} + \epsilon; p) &= P(-\epsilon \leq \bar{X} - p \leq \epsilon) \\ &= P(|\bar{X} - p| \leq \epsilon) \geq 1 - 2e^{-2n\epsilon^2} = 1 - \alpha. \end{aligned}$$

- Now, consider the Poisson distribution. Let $X_1, \cdots, X_n \sim Poisson(\lambda)$. We want to construct a confidence interval for λ.
  Recall that $\sum X_i \sim Poisson(n\lambda)$, with mean $n\lambda$ and variance $n\lambda$.
  With Chebyshev's inequality, there is

$$P(|\bar{X}_n - \lambda| \geq \epsilon) \leq \lambda/n\epsilon^2.$$

  Set $\alpha = \lambda/n\epsilon^2$, which solves that $\epsilon_n = \sqrt{\lambda/n\alpha}$. The $1-\alpha$-confidence intervals is $[\bar{X} - \sqrt{\lambda/n\alpha}, \bar{X} + \sqrt{\lambda/n\alpha}]$.

Inverting a Test

- Consider the Hypothesis testing problem

$$H_0 : \ \theta = \theta_0 \quad versus \quad H_1 : \ \theta \neq \theta_0.$$

Say that we have a test statistic T and rejection region R. We consider level α test, so that $P(T \in R; \theta_0) \leq \alpha$, and so $P(T \notin R; \theta_0) \geq 1 - \alpha$.

- Define the acceptance region $A(\theta_0)$, where $A(\theta_0)$ is the set in X^n.

$$A(\theta_0) = \{(x_1, \cdots, x_n) : T(x_1, \cdots, x_n) \notin R(\theta_0)\}.$$

- Define the confidence set. The confidence set is a set in the parameter space Θ, defined by the observations $(x_1, \cdots, x_n)$.

$$C_n = C_n(x_1, \cdots, x_n) = \{\theta : (x_1, \cdots, x_n) \in A(\theta)\}.$$

- Coverage Probability:

$$\begin{aligned} P(\theta \in C; \theta) &= P((X_1, \cdots, X_n) \in A(\theta); \theta) \\ &= P(T(X_1, \cdots, X_n) \notin R(\theta); \theta) \geq 1 - \alpha. \end{aligned}$$

Inverting a Test

- The procedure seems hard to understand, yet the procedure is easy
- Let $X_1, \cdots, X_n \sim N(\theta, 1)$,The LRT of $H_0:\ \theta = \theta_0 \quad versus \quad H_1:\ \theta \neq \theta_0$ has rejection region as

$$|\bar{X} - \theta_0| \geq \frac{\sigma}{\sqrt{n}} z_{\alpha/2}.$$

So, the acceptance region is ($A(\theta)$ is a set about $x_{1:n}$, which changes with θ)

$$A(\theta) = \{(x_1, \cdots, x_n); |\bar{x} - \theta| < \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\},$$

and so $\theta \in C(X^n)$ if and only if

$$|\bar{X} - \theta_0| \geq \frac{\sigma}{\sqrt{n}} z_{\alpha/2}.$$

In other words, the confidence interval is $(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2})$. This interval has confidence coefficient as $1 - \alpha$.

Relationship

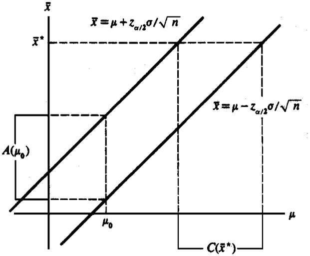


Figure 9.2.1 *Relation between confidence intervals and acceptance regions for tests*

- The hypothesis test fixed the parameter, and asks what sample values (the acceptance region) are consistent with fixed vale.
- The confidence set fixes the sample value, and asks what parameter (the confidence interval) make this sample value most plausible.

Inverting a Test

- As long as we have a test, we can find the confidence interval with it. It applies for whatever test, the Wald test, the Neyman-Pearson test, the t test and the F-test, etc.
- With this procedure, it is possible that we cannot get an interval. That' s why we call it "$confidence\ sets$" instead of confidence intervals
- With a $1-\alpha$ confidence set $C(x_1, \cdots, x_n)$, we can also figure out a test:
$$reject\ H_0:\ \theta = \theta_0\ if\ \theta_0 \notin C(x_1, \cdots, x_n).$$
It is a level α test.
- However, it is much less used. The most general direction is from hypothesis testing problems to the confidence interval estimation, i.e., the distribution is the same for every $\theta \in \Theta$.

Pivot

Definition: Pivot

A function $Q(X_1, \cdots, X_n, \theta)$ is a *pivot* if the distribution of Q does not depend on θ.

- If the distribution of Q is known, with the relationship between $X_1, \cdots, X_n$ and θ in Q, we can build a confidence interval.
- Let a and b be such that

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

  The confidence interval follows as $C(x) = \{\theta : a \leq Q(X, \theta) \leq b\}$
- Example. $N(0, 1)$ distribution. $\bar{X} - \theta \sim N(0, 1/n)$, which does not depend on θ
- Any location families has pivot as $\bar{X} - \theta$.

Example

Let $X_1, \cdots, X_n \overset{i.i.d}{\sim} Unif(0, \theta)$. Let $Q = X_{(n)}/\theta$. Then the CDF of Q is

$$P(Q \leq t) = \prod_{i=1}^{n} P(X_i \leq t\theta) = \left(\frac{t\theta}{\theta}\right)^n = t^n, \quad 0 < t \leq 1.$$

It does not depend on θ, so Q is a pivot.
To find a $1-\alpha$ confidence interval, note that

$$P(c \leq Q \leq 1) = 1 - P(Q \leq c) = 1 - c^n.$$

Let $1 - \alpha = 1 - c^n$, then $c = \alpha^{1/n}$.

$$P(c \leq Q \leq 1) = 1 - c^n = 1 - \alpha.$$

The $1-\alpha$ confidence interval is

$$\begin{aligned} C(X_{1:n}) &= \{\alpha^{1/n} \leq X_{(n)}/\theta \leq 1\} = \{X_{(n)} \leq \theta \leq X_{(n)}/\alpha^{1/n}\} \\ &= (X_{(n)}, X_{(n)}/\alpha^{1/n}) \end{aligned}$$

Confidence Sets of CDF

Let $X_1, \cdots, X_n \sim F$. The empirical CDF is

$$\hat{F}(x) = \frac{1}{n} \sum 1\{X_i \leq x\}.$$

This is an estimation. Can we find the confidence sets for $F(x)$?

- This is nonparametric estimation. Yet we can still apply the parametric approximations
- For fixed x, note that $\hat{F}(x)$ is the average of n Bernoulli$(F(x))$, we can apply the confidence interval results for the Bernoulli random variables
- We are interested in the confidence sets for the whole CDF. We want to figure out $L(x)$ and $U(x)$, so that

$$P(L(x) \leq F(x) \leq U(x) \ for\ all\ x) \geq 1 - \alpha$$

Pivot

Empirical CDF: $\hat{F}(x) = \frac{1}{n}\sum 1\{X_i \le x\}$.

- Let $K_n = \sup_x |\hat{F}(x) - F(x)|$. K_n measures the largest difference between the empirical CDF and the truth. Once K_n is properly bounded, the confidence sets for $F(x)$ among all x can be fixed.
- For continuous F, K_n is a pivot. To see this, let $U_i = F(X_i)$. Then $U_1, \cdots, U_n \overset{i.i.d}{\sim} Unif(0,1)$. So,

$$\begin{aligned} K_n &= \sup_x |\hat{F}(x) - F(x)| = \sup_x |\frac{1}{n}\sum 1\{X_i \le x\} - F(x)| \\ &= \sup_x |\frac{1}{n}\sum 1\{F(X_i) \le F(x)\} - F(x)| \\ &= \sup_x |\frac{1}{n}\sum 1\{U_i \le F(x)\} - F(x)| \\ &= \sup_{0 \le t \le 1} |\frac{1}{n}\sum 1\{U_i \le t\} - t| \end{aligned}$$

  The result does not depend on F.
- Find a number c, so that $P(\sup_{0\le t\le 1} |\frac{1}{n}\sum 1\{U_i \le t\}| > c) = \alpha$.
- The confidence set is then $C = \{F : \sup_x |F_n(x) - F(x)| < c\}$.

Credible Sets

In Bayesian statistics, what is the confidence set?

- Recall. For Bayesian statistics, the parameters are not constants. There is a prior $\pi(\theta)$ for the parameter θ
- With the observed data, we update the prior $\pi(\theta)$ to the posterior $\pi(\theta|X)$
- If we have a loss function, we summarize $\pi(\theta|X)$ into an estimator with smallest Bayes risk.
- However, for Bayesian statisticians, $\pi(\theta|X)$ is the estimation for the parameter θ
- Confidence sets: the probability that the estimated set include the true parameter θ_0
- In Bayesian, there is no $truth$. They update the prior distribution with more and more data, to get a more and more accurate posterior distribution. So, no $confidence\ interval$ thing!
- Yet, there is so-called $credible\ sets$

Credible Sets

- Assume we observe a random sample $X_1, \cdots, X_n \sim F(x;\theta)$, and the prior is $\pi(\theta)$
- With the data, we have the posterior $\pi(\theta)$
- The $1-\alpha$ credible set C is defined as

$$P(L(X_{1:n}) \leq \theta \leq U(X_{1:n}) | X) \geq 1-\alpha.$$

- We still have a set here. The set has probability $1-\alpha$
- Difference: For confidence set, θ is fixed, $L(X)$ and $U(X)$ are random. The probability is the probability that (L, U) contains θ. If we draw the samples again and again, then the probability it covers θ is $1-\alpha$. For credible sets, θ is random. With the given data, we are interested in the interval that θ concentrates on.
- To find the credible set, just figure out the posterior distribution, and draw an interval for θ with probability $1-\alpha$.

Bootstrap

Let $X_1, \cdots, X_n \overset{i.i.d}{\sim} F(X;\theta)$. Let $\hat{\theta}_n = g(X_1, X_2, \cdots, X_n)$ be an estimator. Let $\sigma_n^2 = Var(\hat{\theta}_n)$

- Note: $\hat{\theta}_n = g(X_1, X_2, \cdots, X_n)$ is also a r.v., where the CDF of $\hat{\theta}_n$ can be calculated if we know F.
- σ_n^2 can be calculated if we know the CDF of $\hat{\theta}_n$. Yet, it may be quite complicated, especially for the estimators without explicit formula.
- If we know the CDF of $\hat{\theta}_n$, we can draw a sample $\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}, \cdots, \hat{\theta}_n^{(B)}$, and estimate the variance through sample variance

$$\sigma_n^2 = \frac{1}{B-1} \sum_{i=1}^{B} (\hat{\theta}_n^{(i)} - \frac{1}{B} \sum_j \hat{\theta}_n^{(j)})^2$$

- If we know F, then we do not need to calculate the CDF for $\hat{\theta}_n$, and we cal still draw a sample $\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}, \cdots, \hat{\theta}_n^{(B)}$. Since we can draw $X_1^{(i)}, \cdots, X_n^{(i)}$ to calculate $\hat{\theta}_n^{(i)}$, $i = 1, \cdots, B$.

Bootstrap

- If we know F, we can get the empirical variance for $\hat{\theta}_n$
- Now, we do not know F. However, we have the empirical CDF F_n

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

In our tutorial, we show that $F_n(x)$ is consistent with $F(x)$.

- Therefore, we can draw samples from $F_n(x)$.

$$\begin{aligned} draw \quad & X_1^*, \cdots, X_n^* \sim F_n \\ Compute \quad & \hat{\theta}_n^{(1)} = g(X_1^*, \cdots, X_n^*) \\ draw \quad & X_1^*, \cdots, X_n^* \sim F_n \\ Compute \quad & \hat{\theta}_n^{(2)} = g(X_1^*, \cdots, X_n^*) \\ \vdots \quad & \vdots \\ draw \quad & X_1^*, \cdots, X_n^* \sim F_n \\ Compute \quad & \hat{\theta}_n^{(B)} = g(X_1^*, \cdots, X_n^*) \end{aligned}$$

- The variance is: $\sigma_B^2 = \frac{1}{B-1} \sum_{i=1}^{B} (\hat{\theta}_n^{(i)} - \frac{1}{B} \sum_j \hat{\theta}_n^{(j)})^2$

Bootstrap

- The algorithm is called Bootstrap Variance Estimator
- According to the definition of F_n, it is a discrete r.v., with PMF as

$$P(X = x_i) = 1/n, \quad i = 1, \cdots, n.$$

  So, the random sample is to draw n samples from $x_1, \cdots, x_n$ with replacement.

- The intuition is that

$$\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_n^{(i)} - \frac{1}{B}\sum_j \hat{\theta}_n^{(j)})^2 \approx Var(\hat{\theta}_n^{(i)}) \approx Var(\hat{\theta}_n)$$

  where the first term is the Bootstrap estimator, the second term is the true variance of the estimator with CDF F_n, and the third term is the truth.

- The difference between the first and second item is due to the fact that B is finite. Yet we can make B as large as possible. The difference between the second the third term is due to that n is finite.

Example

Consider $X_1, \cdots, X_n \sim F$. Now we are interested in the median of F. Obviously, the median of $X_i's$ is a reasonable estimator. Yet, what's the variance of this estimator?

(1) Draw $Y_1, \cdots, Y_n$ with replacement from $\{X_1, \cdots, X_n\}$.

(2) Let $\theta_i = median(Y_1, \cdots, Y_n)$

(3) Repeat $1-2$ for $B = 10000$ times. So that we have $\theta_1, \cdots, \theta_B$.

(4) Estimate the variance as

$$\sigma_B^2 = \frac{1}{B-1}\sum_{i=1}^{B}(\theta_i - \bar{\theta})^2$$

Note. If F is normal distribution with variance 1, according to our analysis about the median for normal distribution, the asymptotic variance is $\frac{1}{4\phi(0)^2}$

R code

```
rm(list=ls())
x <-  rnorm(200, 5, 1)
m1 <- median(x)
#Bootstrap Algorithm
B <- 100000;
theta <- rep(0, B);
for(i in 1:B){
  y <-  sample(x, 200, replace = TRUE)
  theta[i] <-  median(y)
}
mvar <-  var(theta)
1 / 4 / dnorm(0) ^ 2
mvar * 200


```

Bootstrap Confidence Interval

- If the estimator is asymptotic normal distributed, then the variance is enough for a confidence interval (and that's one way to achieve CI with Bootstrap)
- More accurate way is to find the distribution for $\sqrt{n}(\hat{\theta} - \theta)$
- If F is known, the empirical distribution for $\hat{\theta}$ can be estimated through

$$\widetilde{F}_n(t) = \frac{1}{B}\sum_{i=1}^{B} 1\{\sqrt{n}(\hat{\theta}_i - \theta) \leq t\},$$

  where $\hat{\theta}_i,\ i = 1, \cdots, B$ are independent observations drawn from the distribution for $\hat{\theta}_i$
- Again, we do not know F, but we know the empirical distribution for F.
- For the empirical distribution, the truth is $\hat{\theta}$
- The random draws are $\hat{\theta}_1^*, \cdots, \hat{\theta}_B^*$, We can have $\bar{F}_n(t)$ as empirical CDF of $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$. Hopefully, $\bar{F}_n(t)$ is close to $\widetilde{F}_n(t)$

Bootstrap Confidence Interval: Procedure

Bootstrap Confidence Interval:

(1) Draw a bootstrap sample $X_1^* \cdots, X_n^* \sim F_n$. Compute $\hat{\theta}^* = g(X_1^* \cdots, X_n^*)$

(2) Repeat Step 1 for B times, yielding estimators $\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_B^*$

(3) Define

$$\bar{F}_n(t) = \frac{1}{B}\sum_{i=1}^{B} 1\{\sqrt{n}(\hat{\theta}_j^* - \hat{\theta}_n) \leq t\},\ \hat{\theta}_n = g(X_1, \cdots, X_n).$$

(4) The confidence interval is

$$C_n = [\hat{\theta}_n - \frac{t_{1-\alpha/2}}{\sqrt{n}}, \hat{\theta}_n - \frac{t_{\alpha/2}}{\sqrt{n}}]$$

where $t_{\alpha/2} = \bar{F}^{-1}(\alpha/2), t_{1-\alpha/2} = \bar{F}^{-1}(1-\alpha/2)$

Example

Consider the polynomial regression model $Y = g(X) + \epsilon$, where $X, Y \in R$ and $g(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$. Therefore, the function is

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

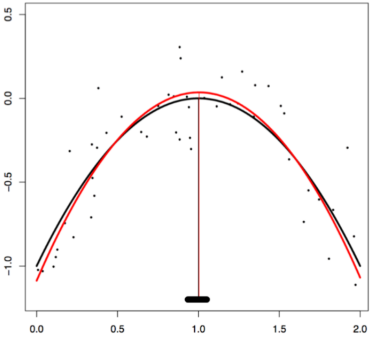
Given data $(X_1, Y_1), \cdots, (X_n, Y_n)$ we can estimate $\beta = (\beta_0, \beta_1, \beta_2)$ with the least squares estimator $\hat{\beta} =$. We are interested in the location at which $g(x)$ is maximized. It is easy to see that the maximum occurs at $x = -(1/2)\beta_1/\beta_2 = \theta$. A point estimate of θ is $\hat{\theta} = -(1/2)\hat{\beta}_1/\hat{\beta}_2$. Now we want to find a Bootstrap confidence interval for θ.

Example

Truth:

$$\beta_0 = -1, , \beta_1 = 2, \beta_2 = -1$$
$$X \sim Unif(0,2),\ \epsilon \sim N(0, 0.04),$$
$$\theta = (-1/2)\beta_1/\beta_2 = 1$$

Sample: 50 points (black)
Curves: True (black) and estimated (red)



Intuition

We have the following terms:

- $F_n(t)$: the true distribution
$$F_n(t) = P(\sqrt{n}(\hat{\theta}_n - \theta) \le t).$$
If we know $F_n(t)$, we can apply it to construct a confidence interval, which is
$$C_n = [\hat{\theta} - F_n^{-1}(1-\alpha/2)/\sqrt{n}, \hat{\theta} - F_n^{-1}(\alpha/2)/\sqrt{n}]$$
- $\hat{F}_n(t)$: the true CDF of the Bootstrap estimator.
$$\hat{F}_n(t) = P(\sqrt{n}(\hat{\theta}^* - \hat{\theta}_n) \le t | X_1, \cdots, X_n)$$
When $X_1, \cdots, X_n$ is given, it does not depend on θ.
- $\bar{F}(t)$: the empirical version of $\hat{F}_n(t)$.
$$\bar{F}(t) = \frac{1}{B}\sum_{j=1}^{B} I\{\sqrt{n}(\hat{\theta}_j^* - \hat{\theta}_n) \le t\}.$$
When$B \to \infty, \bar{F}(t) \to \hat{F}_n(t)$. We assume that B is very large.

Intuition

- If $\bar{F}(t)$ is close to $F_n(t)$, then the estimation $C_n = [\hat{\theta} - F_n^{-1}(1-\alpha/2)/\sqrt{n}, \hat{\theta} - F_n^{-1}(\alpha/2)/\sqrt{n}]$ is a good estimator.
- Assumptions: $F_n(t) \to L,\ \hat{F}_n(t) \to \hat{L}$

$$
\begin{tikzcd}
F_n \arrow[r, "O(1/\sqrt{n})"] \arrow[d, dashed] & L \arrow[d, "O_P(1/\sqrt{n})"] \\
\widehat{F}_n \arrow[r, "O_P(1/\sqrt{n})"'] \arrow[d, "O(1/\sqrt{B})"] & \widehat{L} \\
\overline{F} &
\end{tikzcd}
$$

Proof for a simple case

Suppose that $X_1, \cdots, X_n \sim F$ where $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$. Suppose we want to construct a confidence interval for μ. Let $\hat{\mu}_n = \bar{X}$, we define that

$$F_n(t) = P(\sqrt{n}(\hat{\mu}_n - \mu) \leq t),\ \hat{F}_n(t) = P(\sqrt{n}(\hat{\mu}_n^* - \mu) \leq t | X_1, \cdots, X_n)$$

- According to the analysis in previous slide, we need to show $\sup_t |F_n(t) - \hat{F}_n(t)|$ is small. To prove it, we need that the distribution converges.
- The convergence can be proved through Berry-Esseen Theorem.

Berry-Esseen Theorem

Let $X_1, \cdots, X_n$ be i.i.d with mean μ and variance σ^2. Let $\mu_3 = E[|X_i - \mu|^3] < \infty$ and $\Phi(\cdot)$ be the CDF of $N(0,1)$. Then we have

$$\sup_z |P(\sqrt{n}(\bar{X}_n - \mu) \leq \sigma_z) - \Phi(z)| < \frac{33}{4} \frac{\mu_3}{\sqrt{n}\sigma^3}$$

Proof for a simple case

- According to Berry-Esseen Theorem, $F_n(t) \to N(0, \sigma^2)$ and $\hat{F}_n(t) \to N(0, \hat{\sigma}^2)$, where $\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$. What's more ,we have the control on the convergence rate.
- By the triangle inequality,

$$
\begin{aligned}
\sup_t |F_n(t) - \hat{F}_n(t)| &= \sup_t |F_n(t) - \Phi(\frac{t}{\sigma}) + \Phi(\frac{t}{\sigma}) - \Phi(\frac{t}{\hat{\sigma}}) + \Phi(\frac{t}{\hat{\sigma}}) - \hat{F}_n(t)| \\
&\leq \sup_t |F_n(t) - \Phi(\frac{t}{\sigma})| + \sup_t |\Phi(\frac{t}{\sigma}) - \Phi(\frac{t}{\hat{\sigma}})| \\
&+ \sup_t |\Phi(\frac{t}{\hat{\sigma}}) - \hat{F}_n(t)| \\
&\leq \frac{33}{4}\frac{\mu_3}{\sqrt{n}\sigma^3} + \sup_t |\Phi(\frac{t}{\sigma}) - \Phi(\frac{t}{\hat{\sigma}})| + \frac{33}{4}\frac{\hat{\mu}_3}{\sqrt{n}\hat{\sigma}^3}
\end{aligned}
$$

where μ_3 is the third moment for empirical CDF.

Proof for a simple case

- According to Taylor expansion,

$$\Phi(t/\hat{\sigma}) = \Phi(t/\sigma) - (\sigma \hat{-} \sigma)\frac{t}{\sigma^2}\phi(t/\sigma) + \cdots$$

  Note that $t/\sigma^2\phi(t/\sigma)$ is bounded for any σ and t, and $\sigma \hat{-} \sigma = O_p(1/\sqrt{n})$, we have that $\sup_t |\Phi(\frac{t}{\sigma}) - \Phi(\frac{t}{\hat{\sigma}})| = O_p(1/\sqrt{n})$.
- Therefore, $\sup_t |F_n(t) - \hat{F}_n(t)| = O_p(1/\sqrt{n})$. Therefore, the Bootstrap CI has coverage probability as $1 - \alpha - O_p(1/\sqrt{n})$.

Parametric Bootstrap

- The procedure is totally $non-parametric$. We do not need any information from F. Therefore, it can be used to estimate any function of F, say, $E[X_1 X_2]$.
- If we know the family of distribution, say, $F = N(\mu, \sigma^2)$, then the information helps us in the Bootstrap problem.
  The Parametric Bootstrap Variance Estimator:
- Therefore, we can draw samples from $F_n(x)$.

$$
\begin{aligned}
draw \quad & X_1^*, \cdots, X_n^* \sim F_n \\
Compute \quad & \hat{\theta}_n^{(1)} = g(X_1^*, \cdots, X_n^*) \\
draw \quad & X_1^*, \cdots, X_n^* \sim F_n \\
Compute \quad & \hat{\theta}_n^{(2)} = g(X_1^*, \cdots, X_n^*) \\
\vdots \quad & \vdots \\
draw \quad & X_1^*, \cdots, X_n^* \sim F_n \\
Compute \quad & \hat{\theta}_n^{(B)} = g(X_1^*, \cdots, X_n^*)
\end{aligned}
$$

- The variance is: $\sigma_B^2 = \frac{1}{B-1} \sum_{i=1}^{B} (\hat{\theta}_n^{(i)} - \frac{1}{B} \sum_j \hat{\theta}_n^{(j)})^2$

Remarks

- The Bootstrap is a general procedure. However, it requires some assumptions. We have shown the condition for the mean estimation. The general condition is $= Hadamard\ Differentiability$. You may check it after class.
- Bootstrap highly rely on the observed data. The rate is controlled as $1/\sqrt{n}$, where n is the sample size.
- There are many modifications for the Bootstrap confidence interval, for which you can check the textbook (if interested):
  - Bootstrap percentile method (no strict proof)
  - Bootstrap bias-corrected percentile
  - Hybrid bootstrap
  - more
- Related method: jackknife. The jackknife method is to estimate the standard error by leaving out one observation at a time. A generalization of the jackknife method is cross-validation.