

Principle of Statistics (continued)

November 22, 2017

Contents

In the previous lecture, we saw that for a bootstrap sample (X_1^b, \dots, X_n^b) drawn from $\mathbb{P}_n(\cdot | X_1, \dots, X_n)$, we have $\sup_{t \in \mathbb{R}} |\mathbb{P}_n(\sqrt{n}(\bar{X}_n^b - \bar{X}_n) \leq t(X_1, \dots, X_n)) - \Phi(t)| \xrightarrow{a.s.} 0$ as $n \rightarrow \infty$, where $\Phi(t) = \mathbb{P}(Z \leq t)$, $Z \sim N(0, \sigma^2)$.

Remark. As in the B.vM theorem, this theorem can be used to show that $\mathbb{P}(\mu \in \mathcal{C}_n) \rightarrow 1 - \alpha$ when $n \rightarrow \infty$.

Idea: For fixed $(X_i)_{i \geq 1}$, $\mathbb{P}_n(\cdot | X_1, \dots, X_n)$ is a sequence of distributions. Considering the X_i as independent random variables drawn from P allows us to make statements "with randomness". The idea is to fix a sequence X_i (equivalent to fix ω in the original probability space) if we can show that $\sup_{t \in \mathbb{R}} |\mathbb{P}_n(\dots | X_1, \dots, X_n) - F(t)| \rightarrow 0$ "for almost all ω ", then we have almost sure convergence.

Lemma. If $A_n \sim f_n \xrightarrow{d} A \sim f$, and F is continuous (c.d.f of f), then $\sup |F_n(t) - F(t)| \rightarrow 0$ as $n \rightarrow \infty$.

Proof. By continuity of F , there exists points $-\alpha_0 = x_0 < x_1 < \dots < x_k = +\infty$ such that $F(x_i) = \frac{i}{k}$. Then for every $x \in [x_{i-1}, x_i]$, $F_n(x) - F(x) \leq F_n(x_i) - F(x_{i-1}) = F_n(x_i) - F(x_i) + \frac{1}{k}$, and $F_n(x) - F(x) \geq F_n(x_{i-1}) - F(x_i) = F_n(x_{i-1}) - F(x_{i-1}) - \frac{1}{k}$. For k large enough, $\frac{1}{k} < \frac{\varepsilon}{2}$, for n large enough (depends on k), we have $\max_{0 \leq i \leq k} |F_n(x_i) - F(x_i)| < \varepsilon/2$ (pointwise convergence) (depends on k), we have

$$\max_{0 \leq i \leq k} |F_n(x_i) - F(x_i)| < \varepsilon/2$$

(pointwise convergence). As a consequence,

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \leq \max_{0 \leq i \leq k} |F_n(x_i) - F(x_i)| + \frac{1}{k} < \varepsilon$$

□

Definition. The sequence $(Z_{n,i}, i = 1, \dots, n)_{n \geq 1}$ is a triangular array of i.i.d. random variables if:

- For all $n \geq 1$, $(Z_{n,1}, \dots, Z_{n,i}, \dots, Z_{n,n})$ is a sequence of i.i.d. random variables; For example, $Z_{11} = (Z_{1,i})$,
 $Z_{21}, Z_{22} = (Z_{2,i})$,
 \dots ,
 $Z_{n,1}, \dots, Z_{n,n} = (Z_{n,i})$.

We need independence on each line, but not across the lines. We don't need even need the same distribution at each line.

Proposition. (CLT for triangular arrays)

Let $(Z_{n,i}; i = 1, \dots, n)$ be a triangular array of iid random variables, each with finite variance. We have $\text{Var}_{Q_n}(Z_{n,i}) = \sigma_n^2 \rightarrow \sigma^2$ as $n \rightarrow \infty$, each line consists of n independent draws from Q_n . Then, under the following hypotheses (1-3), we have

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n Z_{n,i} - \mathbb{E}_{Q_n}[Z_{n,i}] \right) \xrightarrow{d} N(0, \sigma^2)$$

- (1) $\forall \delta > 0$, $nQ_n(|Z_{n,1}| > \sqrt{n}\delta) \rightarrow 0$ as $n \rightarrow \infty$;
- (2) $\text{Var}(Z_{n,1}1_{\{|Z_{n,1}| \leq \sqrt{n}\}}) \rightarrow \sigma^2$ as $n \rightarrow \infty$;

(3) $\sqrt{n}\mathbb{E}[Z_{n,1}1\{|Z_{n,1}| > \sqrt{n}\}] \rightarrow 0$ as $n \rightarrow \infty$.

(The statement of these assumptions is not examinable.)

Proof. (of the main theorem)

Fix $(X_i)_{i \geq 1}$ (equivalent to fix ω in the original probability space). Under $Q_n = \mathbb{P}_n(\cdot \mid X_1, \dots, X_n)$, $Z_{n,i} = X_i^{b(n)}$ (bootstrap on n observations), $\mathbb{E}_n[Z_{n,i}] = \mathbb{E}_{\mathbb{P}_n}[X_i^{b(n)}] = \bar{X}_n$. Then, the $(X_i^{b(n)}; i = 1, \dots, n)_{n \geq 1}$ are a triangular array of i.i.d. variables. We have that

$$\text{Var}_{\mathbb{P}_n}(X_i^{b(n)}) = \mathbb{E}_{\mathbb{P}_n}[X_i^{b(n)^2}] - (\mathbb{E}_{\mathbb{P}_n}[X_i^{b(n)}])^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i\right)^2 = \sigma_n^2$$

by definition of $\mathbb{P}_n(\mid X_1, \dots, X_n)$. For almost all the ω , or almost all infinite sequences, $\sigma_n^2 \rightarrow \sigma^2$, and hypotheses (1-3).

• By the CLT for triangular arrays, we have that

$$\sqrt{n}(\bar{X}_n^{b(n)} - \bar{X}_n) \xrightarrow{d} N(0, \sigma^2)$$

as $n \rightarrow \infty$ 'for almost all ω '.

• By lemma,

$$\sup_{t \in \mathbb{R}} |\mathbb{P}_n(\sqrt{n}(\bar{X}_n^{b(n)} - \bar{X}_n) \leq t) - \Phi(t)| \geq 0$$

as $n \rightarrow \infty$ 'for almost all ω ', meaning that it $\xrightarrow{a.s.} 0$. □

Remark. • This shows the validity of the bootstrap confidence interval for the mean.

• In genral, this can be extended to estimation of θ : Sampling from \mathbb{P}_n as for the mean, and compute the bootstrap MLE $\hat{\theta}_n^b = \hat{\theta}(X_1^b, \dots, X_n^b)$ then using $\sqrt{n}(\hat{\theta}_n^b - \hat{\theta}_n)$ as a proxy for $\sqrt{n}(\hat{\theta}_n - \theta_0)$, we will have similar results that show that taking R_n such that

$$\mathbb{P}_n(|\hat{\theta}_n^b - \hat{\theta}_n| \leq \frac{R_n}{\sqrt{n}} \mid X_1, \dots, X_n) = 1 - \alpha$$

can be used to construct a valid confidence region.

This approach is known as the non-parametric bootstrap.

Another approach is to do the same thing with $X_1^b, \dots, X_n^b \sim \mathbb{P}_{\hat{\theta}_n}$, same types of results will hold.