Nonparametric Statistics

Ch.7 Bootstrap

Motivation

❖ Very often, we want to estimate the variance of a statistic. This is sometimes possible or easy to obtain, but, in many cases, it is hard to find or almost impossible.

❖ For example, to construct a confidence interval for the population mean, we need to know the variance of the sample mean.

$$\text{Confidence interval : } \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad where \quad var\,(\bar{X}) = \frac{\sigma^2}{n}$$

Replacing σ^2 by S^2 enables us to obtain a confidence interval.

❖ Suppose that we have $\hat{\theta}$, an estimator of a parameter θ and want to construct a confidence interval. Is it possible to find $var\,(\hat{\theta})$ in a general setting?

Motivation

- "Bootstrap" is a easy-to-use and powerful method to estimate the variance of a statistic. More generally, the "Bootstrap" method can provide approximations of sampling distributions of statistics.

- The "Bootstrap" method is particularly useful when no results on distributions are available. However, even if some distributional properties are known, bootstrap generally provides competitive or better approximations.

- There are some cases where the bootstrap does not work. However, we do not mention about this topic.

Bootstrap

- A basic idea is to replace the **unknown** distribution function F by its empirical version $\hat{F}_n$ which is in principle **known**.

- Suppose that we have a sample $\mathcal{X} = \{X_1, \ldots, X_n\}$ and a statistic T_n is a function of it $T_n = T(\mathcal{X})$. Basically, we want to know the distribution of T_n, $G(x) = P(T_n \leq x)$ which relies on F. The problem here is that we do not know F.

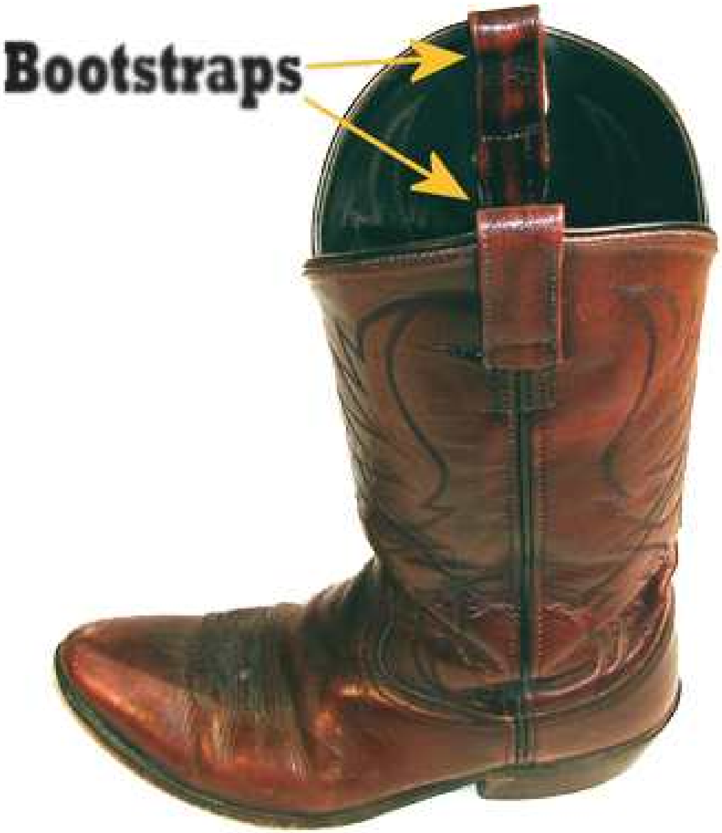
- We replace F by $\hat{F}_n$ and resample from it. That is, we have $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ from $\hat{F}_n$. Note that

$$P(X_i^* \leq x) = \hat{F}_n(x).$$

  So, $T_n^* = T(\mathcal{X}^*)$ only depends on $\hat{F}_n$ which is known. T_n^* is a surrogate statistic to T_n.

- The **known** sampling properties of T_n^* could mimic the **unknown** sampling properties of T_n.

Bootstrap : "Adventures of Baron"



Bootstrap : Monte-Carlo simulations

❖ In the simplest case, the bootstrap estimates could be calculated analytically.

❖ In general, the bootstrap estimates could also be obtained by Monte-Carlo simulations as follows:

- Draw B samples of size n ($\mathcal{X}^{(b)}, b = 1, \ldots, B$) from $\hat{F}_n$
- Compute $T(\mathcal{X}^{(b)}), b = 1, \ldots, B$.
- By the strong law of large numbers, CDF of $T_n = T(\mathcal{X})$ can be approximated by $\frac{1}{B}\sum_{b=1}^{B} I(\mathcal{X}^{(b)} \leq x)$

❖ Simply speaking, $T(\mathcal{X}^{(b)}), b = 1, \ldots, B$ could be regarded as a sample from the true distribution of $T_n = T(\mathcal{X})$.

Bootstrap : Monte-Carlo simulations

❖ How do we simulate from $\hat{F}_n$?

=> $\hat{F}_n$ gives probability 1/n to each data point. Therefore, drawing n points at random from $\hat{F}_n$ is the same as drawing a sample of size n with replacement from the original data.

Bootstrap : Monte-Carlo simulations

❖ Ex] Sample mean : $\bar{X} = T(X_1, \dots, X_n)$

- $Bias\ (\bar{X}) = E(\bar{X}) - \mu = 0$
- $Var\,(\bar{X}) = \sigma^2/n$
- CDF of $\bar{X}$ = CDF of $N(\mu, \frac{\sigma^2}{n})$

❖ The bootstrap estimator : Monte-Carlo simulations

- Draw B samples from $\hat{F}_n$ and compute $\bar{X}^{(b)} = T(\mathcal{X}^{(b)}), b = 1, \dots, B$.
- $Bias\ (\bar{X}) \approx \frac{1}{B}\sum_b \bar{X}^{(b)} - \bar{X}$
- $Var\,(\bar{X}) \approx \frac{1}{B}\sum_b \left(\bar{X}^{(b)} - \frac{1}{B}\sum_b \bar{X}^{(b)}\right)^2 = \frac{1}{B}\sum_b \left(\bar{X}^{(b)}\right)^2 - \left(\frac{1}{B}\sum_b \bar{X}^{(b)}\right)^2$
- CDF of $\bar{X} \approx \frac{1}{B}\sum_{b=1}^{B} I(\bar{X}^{(b)} \le x)$

Bootstrap : Monte-Carlo simulations

❖ In general for $T_n = T(X_1, \dots, X_n)$ to estimate $\theta(F)$.

- $Bias\ (T_n) = E(T_n) - \theta(F) = 0$

- $Var\,(T_n) = E(T_n - E(T_n))^2$

❖ The bootstrap estimator : Monte-Carlo simulations

- Draw B samples from $\hat{F}_n$ and compute $T_n^{(b)} = T(\mathcal{X}^{(b)}), b = 1, \dots, B$.

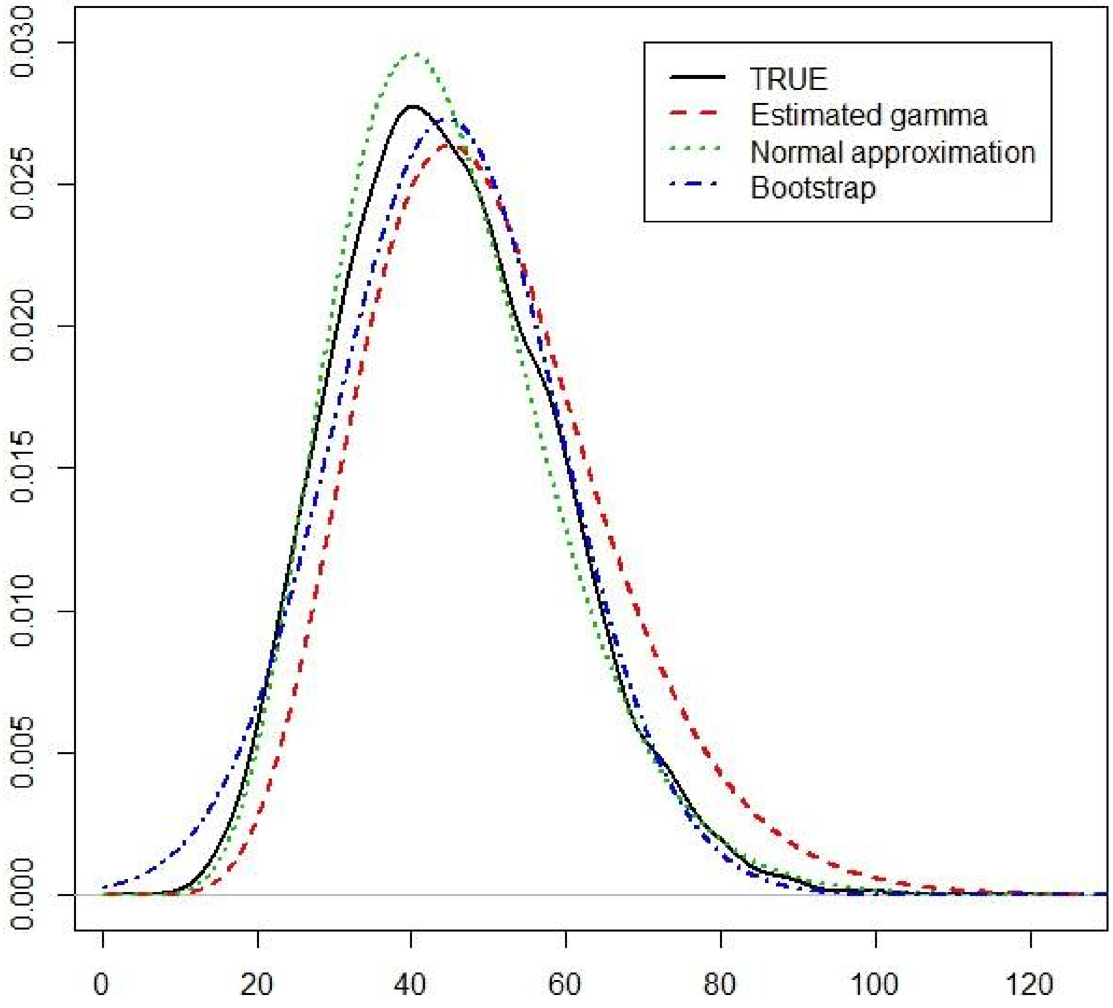
- $Bias\ (T_n) \approx \frac{1}{B}\sum_b T_n^{(b)} - \theta(\hat{F}_n)$

- $Var\,(T_n) \approx \frac{1}{B}\sum_b \left(T_n^{(b)} - \frac{1}{B}\sum_b T_n^{(b)}\right)^2 = \frac{1}{B}\sum_b \left(T_n^{(b)}\right)^2 - \left(\frac{1}{B}\sum_b T_n^{(b)}\right)^2$

- CDF of $T_n \approx \frac{1}{B}\sum_{b=1}^{B} I(\,T_n^{(b)} \leq x)$

Bootstrap : Example

- $X_i \sim Exp\,(\mu), \mu = 50, n = 10$
- Bootstrap estimation of the distribution of the sample mean with $B = 5000$.
- $\Gamma(n, \frac{\mu}{n})$ vs $\Gamma(n, \frac{\bar{X}}{n})$ vs $N(\bar{X}, \frac{s^2}{n})$ vs Bootstrap
- Bootstrap mean & var : 44.91 & 198.35 (Note that the true values are 50 & 250)



Parametric Bootstrap

❖ So far, we have estimated F nonparametrically. There is also a parametric bootstrap. Let F_θ depends on a parameter θ. Then, the parametric bootstrap is drawing a sample from $F_{\hat{\theta}}$ instead of $\hat{F}_n$.

❖ How do we simulate from $F_{\hat{\theta}}$? Use the fact that

$$X \sim F \;\; \rightarrow F^{-1}(U) \sim F \text{ where } U \sim U(0,1)$$

Bootstrap confidence interval

❖ Normal interval

$$T_n \pm z_{\alpha/2}\sqrt{Var^*(T_n)}$$

where $Var^*(T_n)$ is a Bootstrap estimator of the variance of T_n. This is not accurate unless T_n is well approximated by normal.

Bootstrap confidence interval

❖ Pivotal interval

Let $\theta(F)$ be a parameter of interest and $T_n = T(\mathcal{X})$ is its estimator. We call $T_n - \theta(F)$ "pivot" or "root". If we know the distribution of $T_n - \theta(F)$, say H. We have

$$P\left(u\left(\frac{\alpha}{2}\right) \le T_n - \theta(F) \le u\left(1-\frac{\alpha}{2}\right)\right) = 1-\alpha$$

where $u(a)$ is the a-quantile satisfying $H\big(u(a)\big) = P(T_n - \theta(F) \le u(a)) = a$. Then, $100(1-\alpha)\%$ confidence interval for θ is given by

$$(T_n - u\left(1-\frac{\alpha}{2}\right), T_n + u\left(\frac{\alpha}{2}\right))$$

A basic idea is to estimate $u\left(1-\frac{\alpha}{2}\right)$ and $u\left(\frac{\alpha}{2}\right)$ by Bootstrapping.

Bootstrap confidence interval

❖ Pivotal interval (continued)

Let $\widehat{H}^*$ be a Bootstrap estimator of H, that is, the distribution function of $T_n^* - \theta(\hat{F}_n)$. Note that

$$\widehat{H}^*(x) = P^*\left(T_n^* - \theta(\hat{F}_n) \le x\right) \approx \frac{1}{B}\sum_{b=1}^{B} I(T_n^* - \theta(\hat{F}_n) \le x)$$

Practical computation

Let $u^*(a)$ be the quantile of $\widehat{H}^*$ and $v^*(a)$ be the quantile of $\hat{G}^*$ where $\hat{G}^*(x) = P^*(T_n^* \le x) \approx \frac{1}{B}\sum_{b=1}^{B} I(T_n^* \le x)$. Then, $u^*(a) = v^*(a) - \theta(\hat{F}_n)$ and $v^*(a)$ is directly obtained by $\hat{G}^*$. Therefore, $100(1-\alpha)\%$ confidence interval is given by

$$(T_n - u^*\left(1-\frac{\alpha}{2}\right), T_n + u^*\left(\frac{\alpha}{2}\right))$$

Bootstrap confidence interval

❖ Pivotal interval (continued)

Note here that $T_n = \theta(\hat{F}_n)$ in general and therefore $100(1-\alpha)\%$ confidence interval is given by

$$\left(2T_n - v^*\left(1-\frac{\alpha}{2}\right), 2T_n - v^*\left(\frac{\alpha}{2}\right)\right)$$

Bootstrap : Example (continued)

- 2.5% and 97.5% sample quantiles of 5000 bootstrap sample means are $v^*(0.025) = 20.96, v^*(0.975) = 75.07$, respectively. Therefore, 95% pivotal confidence interval for the mean is

$$\left(2\bar{X} - v^*(0.975), 2\bar{X} - v^*(0.025)\right) = (2 \times 44.60 - 75.07, 2 \times 44.60 - 20.96)$$
$$= (14.13\,, 68.25)$$

- Normal interval is given by

$$\bar{X} \pm z_{\alpha/2}\sqrt{Var^*(\bar{X})} = 44.60 \pm 1.96 \times \sqrt{198.35} = (17.00\,, 72.21)$$

- Note that other confidence intervals are avaiable.

Bootstrap : Example (correlation)

❖ Consider the case that we are interested in estimating the distribution of the sample correlation coefficient r.

❖ Do you know anything about the sampling distribution of r? Some asymptotic results exist, but they demand advanced statistical techniques.

❖ We can automatically obtain the sampling distribution of r via bootstrapping!

Bootstrap : Example (correlation)

❖ Let

$$X \sim N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}4 & 1.5\\1.5 & 1\end{pmatrix}\right).$$

Here $\rho = 0.75$. We have a sample of size 100, and the sample correlation coefficient $r = 0.7367$.

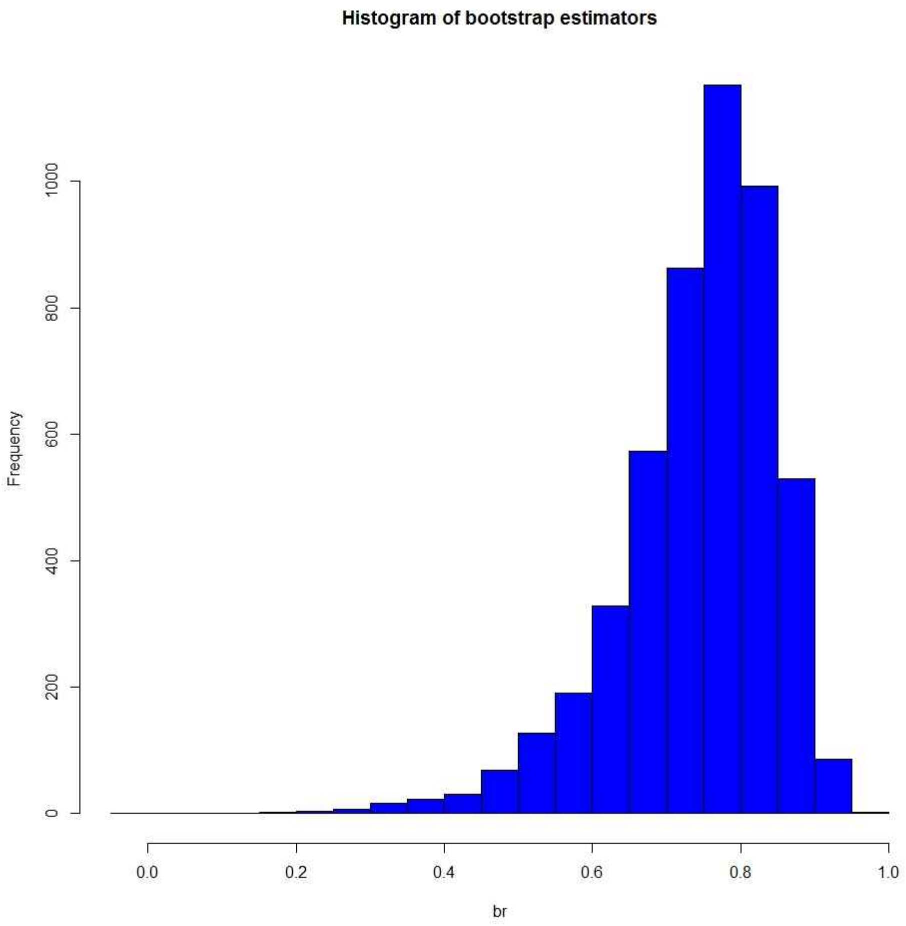
❖ We resample $B = 5000$ bootstrap samples and calculate the bootstrap estimators $r^{*,b}$ $b = 1, \ldots, 5000$. Then, the bootstrap approximations of the mean and the variance of the distribution of r are given by

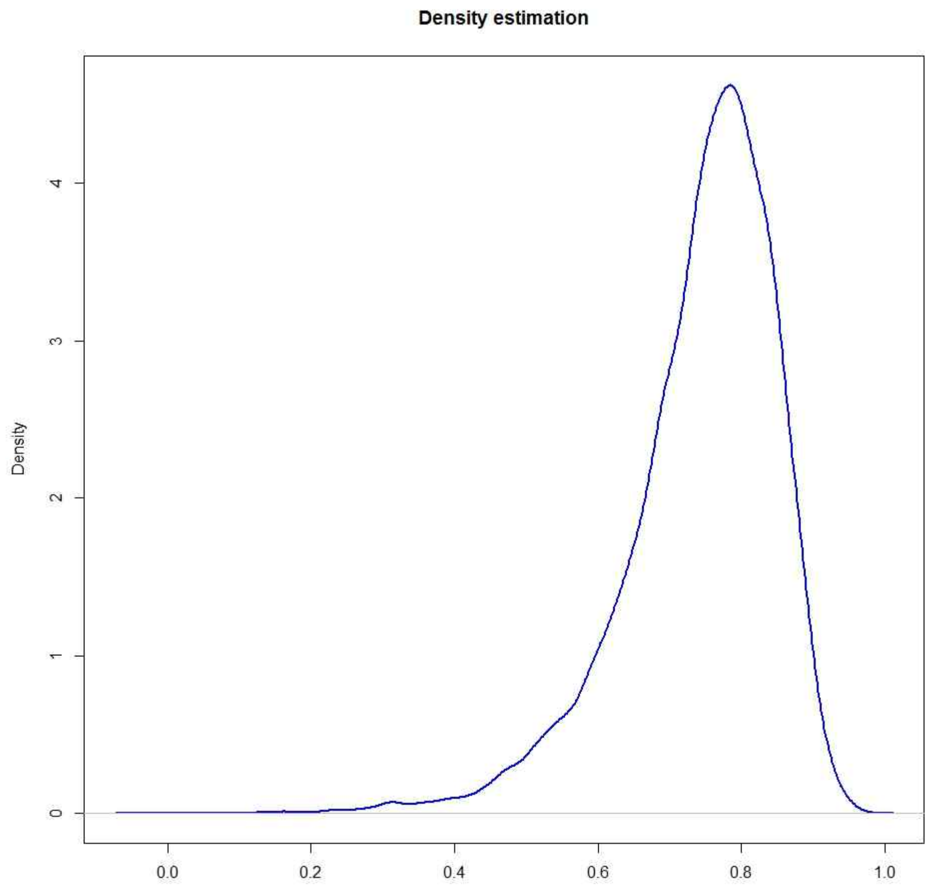
$$E^*(r) = \frac{1}{5000}\sum_b r^{*,b} = 0.7432$$

$$Var^*(r) = \frac{1}{5000}\sum_b \left(r^{*,b}\right)^2 - \left(\frac{1}{5000}\sum_b r^{*,b}\right)^2 = 0.0113$$

Bootstrap : Example (correlation)

❖ This figure depicts the Bootstrap distribution of r^*.





Bootstrap : Example (correlation)

❖ Confidence interval :

- Normal interval

$$\left(r - 1.96\sqrt{Var^*(r)}, r + 1.96\sqrt{Var^*(r)}\right) = (0.5280, 0.9454)$$

- Pivotal interval

$$\left(2r - v^*(0.975), 2r - v^*(0.025)\right) = (0.5799, 0.9934)$$

where $v^*(a)$ is the lower a -quantile of the Bootstrap distribution.

❖ Can you give an answer for the hypotheses $H_0 : \rho = 0 \quad vs \quad H_1 : \rho \neq 0$?

More on confidence interval : percentile

❖ Percentile interval :

$$\left(v^*\left(1-\frac{\alpha}{2}\right), v^*\left(\frac{\alpha}{2}\right)\right)$$

where $v^*(a)$ is the lower a -quantile of the Bootstrap distribution.

❖ Justification : Suppose that there exists a monotone transformation function m such that $U_n = m(T_n)$ and $U_n \sim N(\phi, c^2)$ where $\phi = m\big(\theta(F)\big)$. Then, the pivotal confidence interval for ϕ is

$$\left(U_n - u_U^*\left(1-\frac{\alpha}{2}\right), U_n - u_U^*\left(\frac{\alpha}{2}\right)\right)$$

where $u_U^*(a)$ is the lower a -quantile of the Bootstrap distribution of $U_n^* - m(\theta(\hat{F}_n))$.

More on confidence interval : percentile

- Note that $U_n - m(\theta(F))$ is symmetric around 0. Therefore, $u_U^*\left(1-\frac{\alpha}{2}\right) \approx -u_U^*\left(\frac{\alpha}{2}\right)$ and $u_U^*\left(\frac{\alpha}{2}\right) \approx -u_U^*\left(1-\frac{\alpha}{2}\right)$. The interval can be rewritten as

$$\left(U_n + u_U^*\left(\frac{\alpha}{2}\right), U_n + u_U^*\left(1-\frac{\alpha}{2}\right)\right)$$

where $v_U^*(a)$ is the lower a -quantile of the Bootstrap distribution of U_n^*.

- Recall that $u_U^*(a) = v_U^*(a) - m(\theta(\hat{F}_n)) = v_U^*(a) - U_n$. Therefore, the interval is

$$\left(v_U^*\left(\frac{\alpha}{2}\right), v_U^*\left(1-\frac{\alpha}{2}\right)\right)$$

Note that this is the confidence interval for $\phi = m(\theta(F))$.

More on confidence interval : percentile

❖ Due to the monotonicity of m , the percentile confidence interval for $\theta(F)$ is given by

$$\left(v^*\left(\frac{\alpha}{2}\right), v^*\left(1-\frac{\alpha}{2}\right)\right)$$

$\because\ v_U^*(a) = m(v^*(a))$.

❖ Surprisingly, we do not need to know the transformation function m. The existence of such transformation is enough.

❖ Actually, this interval is valid whenever $U_n - \phi$ follows a symmetric distribution around 0, which means that this may fail for biased estimators.

Bootstrap : Example (correlation) revisited

❖ Percentile interval :

$$(v^*(0.025), v^*(0.975)) = (0.4800, 0.8934)$$

where $v^*(a)$ is the lower a -quantile of the Bootstrap distribution.

❖ Compare it to the aforementioned intervals.

- Normal interval

$$(0.5280, 0.9454)$$

- Pivotal interval

$$(0.5799, 0.9934)$$

Bootstrap : Real data example

❖ Duration data : 10 observations of duration times.

$$\mathcal{X} = (1,5,12,15,20,26,78,145,158,358)$$

❖ For duration times, the exponential modelling ($X \sim \exp(\mu)$) is reasonable in many cases. Note that in the exponential model $\mu = \sigma$ where $\mu = E(X)$ and $\sigma = \sqrt{Var\,(X)}$.

❖ Suppose that we are interested in estimating the sampling distribution of the sample mean $\bar{X}$.

❖ Note that $\bar{x} = 81.8\,, s = 112.94$.

Bootstrap : Real data example

❖ Several approaches are possible based on parametric (①~③) and nonparametric (④~⑤) methods.

① Exact result based on the exponential model

$$\bar{X} \sim \Gamma(n, \mu/n), \hat{\mu} = \bar{x}$$

② CLT based on the exponential model ($\mu = \sigma$)

$$\bar{X} \sim \mathrm{N}(\mu, \mu^2/n), \hat{\mu} = \bar{x}$$

③ Parametric Bootstrap

Generate Bootstrap samples from $Exp\,(\bar{x})$.

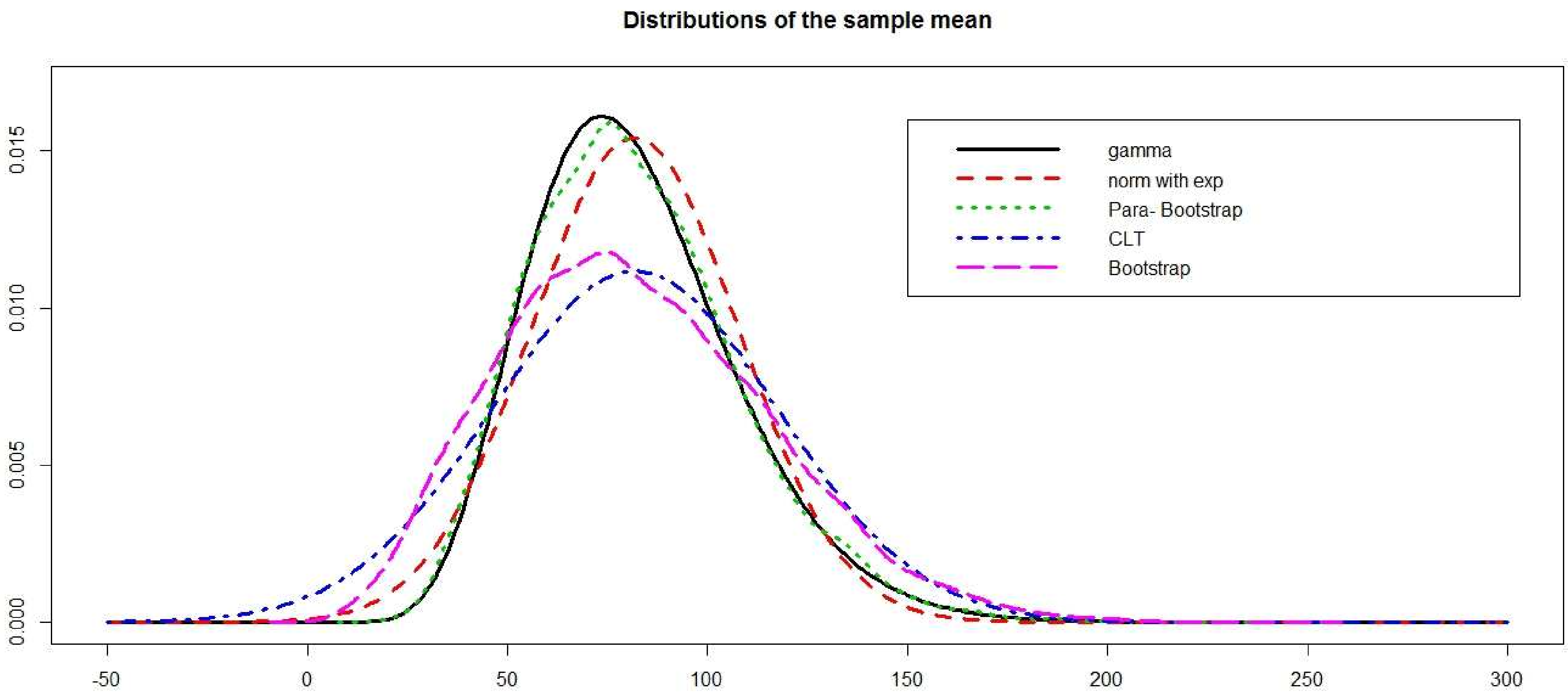
④ CLT

$$\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n), \hat{\mu} = \bar{x}\,, \widehat{\sigma^2} = s^2$$

⑤ (Nonparametric) Bootstrap

Generate Bootstrap samples from $\hat{F}_n$.

Bootstrap : Real data example



Bootstrap : Real data example

❖ Estimation of the mean and standard deviation of $\bar{X}$.

① $\hat{E}(\bar{X}) = \bar{x} = 81.8\,, \widehat{Std}\,(\bar{X}) = \frac{\bar{x}}{\sqrt{n}} = 25.87$

② $\hat{E}(\bar{X}) = \bar{x} = 81.8\,, \widehat{Std}\,(\bar{X}) = \frac{\bar{x}}{\sqrt{n}} = 25.87$

③ $\hat{E}(\bar{X}) \approx 81.89\,, \widehat{Std}\,(\bar{X}) \approx 26.10$

④ $\hat{E}(\bar{X}) = \bar{x} = 81.8\,, \widehat{Std}\,(\bar{X}) = \frac{s}{\sqrt{n}} = 35.71$

⑤ $\hat{E}(\bar{X}) \approx 82.09\,, \widehat{Std}\,(\bar{X}) \approx 33.46$

Bootstrap : Real data example

❖ 95% Confidence interval for the population mean.

① $\bar{X} \sim Gamma\left(n, \frac{\mu}{n}\right) \Rightarrow \frac{\bar{X}}{\mu} \sim Gamma\left(n, \frac{1}{n}\right)$

$$\Rightarrow P\left(G_{0.025}\left(n, \frac{1}{n}\right) \le \frac{\bar{X}}{\mu} \le G_{0.975}\left(n, \frac{1}{n}\right)\right) = 1 - \alpha$$

$$\Rightarrow C.I.\, for\ \ \mu\ : \left(\frac{\bar{x}}{G_{0.975}\left(n, \frac{1}{n}\right)}, \frac{\bar{x}}{G_{0.025}\left(n, \frac{1}{n}\right)}\right) = (47.88{,}170.58)$$

② $\bar{X} \sim N\left(\mu, \frac{\mu^2}{n}\right) \Rightarrow C.I.\, for\ \ \mu : \left(\bar{x} \pm z_{0.975} \frac{\bar{x}}{\sqrt{n}}\right) = (31.10{,}132.50)$

③ Pivotal : (24.07,124.66) , Percentile : (38.94,139.53)

④ $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow C.I.\, for\ \ \mu : \left(\bar{x} \pm z_{0.975} \frac{s}{\sqrt{n}}\right) = (11.80{,}151.80)$

⑤ Pivotal : (8.79,137.30) , Percentile : (26.30,154.81)

Bootstrap : Real data example

❖ Note that ①~③ are based on the exponential model and $\bar{x} = 81.8$, $s = 112.94$, which suggests that the parametric approximation to the exponential distribution is questionable.

❖ In this example, the sample size is too small to apply CLT, and therefore ⑤ is only possibility remained.

❖ There are more advanced Bootstrap confidence intervals to improve the accuracy of interval estimation.