Homework 2

Liu Huihang
SA18017026
QQ: 184050473
MAIL: huihang@mail.ustc.edu.cn

Problem 1 *Let $X_1, \dots, X_n$ iid $\sim F$, F_n be the empirical distribution function and $a < b$ be fixed real numbers, Let $\theta = T(F) = F(b) - F(a)$.*

(1) Find θ's plug-in estimator $\hat{\theta}$;

(2) Find the influence function and empirical influence function of θ;

(3) Estimate the standard error for $\hat{\theta}$;

(4) Find an expression for an approximate $1 - \alpha$ confidence interval for θ.

Solution

(1) $\theta = T(F) = F(b) - F(a) = \int I(a < x \le b) dF(x)$.

The plug-in estimator is $\hat{\theta} = \int I(a < x \le b) dF_n(x) = \frac{1}{n} \sum_{i=1}^n I(a < X_i \le b)$.

(2) θ is a linear functional. Thus the influence function is $\mathrm{IF}(x) = I(a < x \le b) - T(F)$ and the empirical influence function is $\widehat{\mathrm{IF}}(x) = I(a < x \le b) - T(F_n)$.

(3) Denote the estimated standard error of $\hat{\theta}$ by $\widehat{\mathbf{se}}$.

Note that, it cannot be obtained by doing some direct calculations. Because $\mathbb{V}\left(\sum_{i=1}^n I(a < X_i \le b)\right) = \sum_{i=1}^n \mathbb{V}(I(a < X_i \le b)) = n \times (F(b) - F(a)) \times (1 - F(b) + F(a))$ is still undetermined. So we shell use influence function to get estimation of variance.

$$\hat{\tau}^2 = \frac{1}{n} \sum_{i=1}^n \widehat{\mathrm{IF}}^2(X_i) = \frac{1}{n} \sum_{i=1}^n \left(I(a < X_i \le b) - \frac{1}{n} \sum_{j=1}^n I(a < X_j \le b) \right)^2 .$$

Then $\widehat{\mathbf{se}} = \hat{\tau}/\sqrt{n}$, according to Theorem 2.22 in Wasserman (2006).

(4) Using Theorem 2.22 in Wasserman (2006), we have

$$\frac{\sqrt{n}(T(F) - F(F_n))}{\hat{\tau}} \rightsquigarrow N(0, 1).$$

Thus, a $1 - \alpha$, pointwise asymptotic confidence interval for $\theta = T(F)$ is

$$T(F_n) \pm z_{\alpha/2} \widehat{\mathbf{se}},$$

where $\widehat{\mathbf{se}}$ is calculated in (3). ■

Problem 2 *Let $b(\epsilon) = \sup_x |T(F) - T(F_\epsilon)|$, $F_\epsilon = (1 - \epsilon)F + \epsilon \delta_x$. A breakdown point of estimator ϵ^* is definded as $\epsilon^* = \inf\{\epsilon > 0 : b(\epsilon) = \infty\}$. Find*

(1) Breakdown point of mean;

(2) Breakdown point of median.

Solution

(1) Firstly, we analyze the problem from a mathematical point of view. Let $T = T(F) = \int x dF$. Then $T(F) - T(F_\epsilon) = \int x d(F - F_\epsilon) = \int x d(\epsilon F - \epsilon \delta_x) = \epsilon T(F) - \epsilon x$. So $\sup_x |T(F) - T(F_\epsilon)| = \sup_x \epsilon |T(F) - x| = \epsilon \sup_x |T(F) - x|$.

We have that, $\forall \epsilon > 0$, $\sup_x |T(F) - T(F_\epsilon)| = \infty$. Thus $\epsilon^* = 0$.

Then, intuitively(Geyer, 2006), it is obvious from the formula from the mean

$$\frac{x_1 + \cdots + x_n}{n}$$

that if we hold $x_1, \ldots, x_{n-1}$ fixed and let x_n go to infinity, the sample mean also goes to infinity. In short even one gross outlier ruins the sample mean. The finite sample breakdown point is $1/n$. The asymptotic breakdown point is zero.

(2) Let $T = T(F) = F^{-1}(1/2) = \inf\{\mu | F(\mu) \geq 1/2\}$. We have

$$\begin{aligned} T(F_\epsilon) =& F_\epsilon^{-1}(1/2) = \inf\{\mu | (1-\epsilon)F(\mu) + \epsilon\delta_x(\mu) \geq 1/2\} \\ =& \begin{cases} F^{-1}(1/2(1-\epsilon)), & x > F^{-1}(1/2(1-\epsilon)) \\ x, & F^{-1}((1/2-\epsilon)/(1-\epsilon)) < x \leq F^{-1}(1/2(1-\epsilon)) \\ F^{-1}((1/2-\epsilon)/(1-\epsilon)), & x \leq F^{-1}((1/2-\epsilon)/(1-\epsilon)) \end{cases} \end{aligned}$$

So $\sup_x |T(F) - T(F_\epsilon)| = \infty$ when $\epsilon = 1/2$. Thus $\epsilon^* = 1/2$.

Intuitively(Geyer, 2006), if we have n data points and we let a minority of them floor($(n-1)/2$) go to infinity leaving the rest fixed, the "floor" operation means largest integer less than or equal to, then the median stays with the majority. The median changes, but does not become arbitrarily bad. The finite sample breakdown point is floor($(n-1)/2n$). The asymptotic breakdown point is one-half. ■

Problem 3 *Let X be positive random variable with distribution function F and let $\theta = \int log(x)dF(x)$, $\lambda = log(\mu), \mu = EX$.*

(1) Find the influence function and empirical influence function of θ, λ;

(2) Do $\hat{\theta}, \hat{\lambda}$ have the same limit?

(3) Who is more robust to outliers in $\hat{\theta}$ and $\hat{\lambda}$?

Solution

(1) $\theta = T_\theta(F) = \int log(x)dF(x)$ is a linear functional, so the influence function is $\mathrm{IF}_\theta = log(x) - T_\theta(F)$ and the empirical influence function is $\widehat{\mathrm{IF}}_\theta = log(x) - T_\theta(F_n)$.

$\lambda = T_\lambda(F) = log(\int xdF)$ is not a linear functional. So

$$\begin{aligned} \mathrm{IF}_\lambda &= \lim_{\epsilon \to 0} \frac{T_\lambda((1-\epsilon)F + \epsilon\delta_x) - T_\lambda(F)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{log(\int xd((1-\epsilon)F + \epsilon\delta_x) - log(\int xdF)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{log\left(\epsilon x / \int xdF + 1 - \epsilon\right)}{\epsilon} \\ &= \frac{x}{\int xdF} - 1. \end{aligned}$$

And the empirical influence function is $\widehat{\mathrm{IF}}_\lambda = \frac{x}{\int xdF_n} - 1$.

(2) $\hat{\theta} = T_\theta(\hat{F}_n(x))$ and $\hat{\lambda} = T_\lambda(\hat{F}_n(x))$ are plug-in estimators of θ and λ respectively. First, we consider $\hat{\theta}$. We have

$$\sqrt{n}(T_\theta - T_\theta(\hat{F_n})) \rightsquigarrow N(0, \tau_\theta^2)$$

where $\tau_\theta^2 = \int \mathrm{IF}_\theta(x)dF(x)$. The limiting distribution of $\hat{\theta}$ is $N(T_\theta, \tau_\theta^2/n)$

Then, for $\hat{\lambda}$,

$$\sqrt{n}(T_\lambda - T_\lambda(\hat{F_n})) \rightsquigarrow N(0, \tau_\lambda^2)$$

where $\tau_\lambda^2 = \int \mathrm{IF}_\lambda(x)dF(x)$. The limiting distribution of $\hat{\lambda}$ is $N(T_\lambda, \tau_\lambda^2/n)$

Their limiting distributions are different.

What's more, $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n log(X_i) \to E(log(X))$ and $\hat{\lambda} = log\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \to log(E(X))$ almost sure.

$E(log(X)) = log(E(X))$ only when X is constants, otherwise $E(log(X)) < log(E(X))$ by Jensen's inequality.

(3) Technically, we use breakdown points to measure the robustness of a statistic to outliers. Thus we consider the breakdown points of those two estimators.

For $\hat{\theta}$,

$$T_\theta(F) - T_\theta(F_\epsilon) = \int log(x)d\epsilon(F - \delta_x) = \epsilon(T_\theta - log(x)).$$

Then

$$\sup_x |T_\theta(F) - T_\theta(F_\epsilon)| = \epsilon|T_\theta - log(x)|$$

Thus, breakdown point $\epsilon^* = 0$.

For $\hat{\lambda}$,

$$T_\lambda(F) - T_\lambda(F_\epsilon) = log\left(\int xdF\right) - log\left(\int xd((1-\epsilon)F + \epsilon\delta_x)\right) = log\left(\frac{\mu}{(1-\epsilon)\mu + \epsilon x}\right)$$

Then

$$\sup_x |T_\lambda(F) - T_\lambda(F_\epsilon)| = \infty, \quad \text{when } x = \frac{(\epsilon - 1)\mu}{\epsilon} \text{ or } \infty.$$

Thus, breakdown point $\epsilon^* = 0$.

So, from the breakdown point characterization of robustness, they are both sensitive to outliers. But, intuitively, log transformation can reduce the influence of outliers before taking a average of samples.

Thus I suppose $\theta = \int log(x)dF(x)$ is more robust to outliers than $\lambda = log\left(\int xdF(x)\right)$.

■

References

Charles J. Geyer. Breakdown Point Theory Notes, 2006. URL http://www.stat.umn.edu/geyer/5601/notes/break.pdf.

Larry Wasserman. *All of Nonparametric Statistics*. 2006.