

## 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  $\theta$ 's plug-in estimator  $\hat{\theta}$ ;
- (2) Find the influence function and empirical influence function of  $\theta$ ;
- (3) Estimate the standard error for  $\hat{\theta}$ ;
- (4) Find an expression for an approximate  $1 - \alpha$  confidence interval for  $\theta$ .

**Solution**

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

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

(2)  $\theta$  is a linear functional. Thus the influence function is  $L_F(x) = I(a < x \leq b) - T(F)$  and the empirical influence function is  $\hat{L}_F(x) = I(a < x \leq b) - T(F_n)$ .

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

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

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

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

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

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

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

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

where  $\hat{\mathbf{se}}$  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  $\epsilon^*$  is defined 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  $x_1, \dots, x_{n-1}$  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  $\text{floor}((n-1)/2)$  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  $\text{floor}((n-1)/2n)$ . The asymptotic breakdown point is one-half. ■

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

- (1) Find the influence function and empirical influence function of  $\theta, \lambda$ ;
- (2) Do  $\hat{\theta}, \hat{\lambda}$  have the same limiting distribution?
- (3) Who is more robust to outliers in  $\hat{\theta}$  and  $\hat{\lambda}$ ?

**Solution**

(1)  $\theta = T_1(F) = \int \log(x)dF(x)$  is a linear functional, so the influence function is  $\text{IF}_\theta = \log(x) - T_1(F)$ .

$\lambda = T_2(F) = \log(\int x dF)$  is not a linear functional. So

$$\begin{aligned} \text{IF}_\lambda &= \lim_{\epsilon \rightarrow 0} \frac{T_2((1-\epsilon)F + \epsilon\delta_x) - T_2(F)}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} \frac{\log(\int x d((1-\epsilon)F + \epsilon\delta_x)) - \log(\int x dF)}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} \frac{\log(\epsilon x / \int x dF + 1 - \epsilon)}{\epsilon} \end{aligned}$$

■

## 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.