EXERCISE 2

WEIYU LI

1. Suppose $X_1, \ldots, X_n \sim F$, $F_n(x)$ is the EDF. For fixed real numbers $a < b$, let $\theta = T(F) = F(b) - F(a)$.

(1) Solve the plug-in estimation of θ, namely $\hat{\theta}$.

Solve. The plug-in estimation of θ is $\hat{\theta} = T(F_n) = F_n(b) - F_n(a)$. □

(2) Solve the influence function and empirical influence function of θ.

Solve. By definition, the influence function is

$$\begin{aligned} IF(x; T, F) &= \lim_{\epsilon \to 0} \frac{T\left((1-\epsilon)F + \epsilon\delta_x\right) - T(F)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\left[(1-\epsilon)\left(F(b) - F(a)\right) + \epsilon\left(\delta_x(b) - \delta_x(a)\right)\right] - \left[F(b) - F(a)\right]}{\epsilon} \\ &= \left(\delta_x(b) - \delta_x(a)\right) - \left(F(b) - F(a)\right) \\ &= \begin{cases} F(a) - F(b), & a, b < x \text{ or } a, b \geq x \\ F(a) - F(b) + 1, & a < x \leq b \end{cases}. \end{aligned}$$

Analogously, the empirical influence function is

$$\begin{aligned} IF(x; T, F_n) &= \lim_{\epsilon \to 0} \frac{T\left((1-\epsilon)F_n + \epsilon\delta_x\right) - T(F_n)}{\epsilon} \\ &= \begin{cases} F_n(a) - F_n(b), & a, b < x \text{ or } a, b \geq x \\ F_n(a) - F_n(b) + 1, & a < x \leq b \end{cases}. \end{aligned}$$

□

Remark 1. *We can regard T as a linear functional $T(F) = \int I_{(a,b]}(x) dF(x)$ and use the conclusions we already know (e.g., Page 16 in Lec2).*

(3) Estimate the standard deviation of $\hat{\theta}$.

Solve. An estimation of $se = \sqrt{Var(\hat{\theta})}$ can be $\hat{se} = \hat{\tau}/\sqrt{n}$, where $\hat{\tau}^2 = \frac{1}{n}\sum_{i=1}^n IF^2(X_i; T, F_n)$. In specific,

$$\hat{se} = \frac{\sqrt{\sum_{i=1}^n \left[I_{(a,b]}(X_i) + F_n(a) - F_n(b)\right]^2}}{n}.$$

□

Date: 2019/09/23.

liweiyu@mail.ustc.edu.cn.

(4) Give an asymptotic $1-\alpha$ confidence interval of θ.

Solve. For the linear functional T, we have

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

Then asymptotically an $1-\alpha$ confidence interval is

$$T(F_n) \pm z_{\alpha/2}\hat{se} = F_n(b) - F_n(a) \pm z_{\alpha/2}\frac{\sqrt{\sum_{i=1}^n \left[I_{(a,b]}(X_i) + F_n(a) - F_n(b)\right]^2}}{n}.$$

□

2. Denote $b(\epsilon) = \sup_x |T(F) - T(F_\epsilon)|$, $F_\epsilon = (1-\epsilon)F + \epsilon\delta_x$. Define a breakdown point of an estimator as $\epsilon^* = \inf\{\epsilon : b(\epsilon) = \infty\}$.

(1) Solve the breakdown point of the mean.

Solve. The mean of a distribution F is

$$T(F) = \int x dF.$$

Thus,

$$T(F_\epsilon) = \int xd\left((1-\epsilon)F + \epsilon\delta_x\right) = (1-\epsilon)T(F) + \epsilon x,$$

and

$$b(\epsilon) = \sup_x |\epsilon T(F) - \epsilon x| = \infty,\ \forall \epsilon > 0.$$

Consequently, $\epsilon^* = 0$. □

(2) Solve the breakdown point of the median.

Solve. The median of a distribution F is

$$T(F) = Med(F),$$

where $Med(F)$ is a point such that $\mathbb{P}\big(X \le Med(F)\big) = F\big(Med(F)\big) \ge 0.5$, and $\mathbb{P}\big(X \ge Med(F)\big) = 1 - F\big(Med(F) - 0\big) \ge 0.5$ (which may not be unique in some cases). Then $T(F_\epsilon) = Med\left((1-\epsilon)F + \epsilon\delta_x\right)$.

Since $T(F)$ is a constant for all x, $b(\epsilon) = \infty$ if and only if $\sup_x |T(F_\epsilon)| = \infty$.

For any $\epsilon > 0.5$, we have $T(F_\epsilon) \ge x$, since $F_\epsilon(y) = (1-\epsilon)F(y) < 0.5, \forall y < x$. Thus letting $x \to +\infty$ gives that $|T(F_\epsilon)| \to \infty$, and then $b(\epsilon) = \infty$. On the other hand, for any $\epsilon < 0.5$,

$$0.5 \le F_\epsilon\big(Med(F_\epsilon)\big) \le (1-\epsilon)F\big(Med(F_\epsilon)\big) + \epsilon \quad \Rightarrow \quad F\big(Med(F_\epsilon)\big) \ge \frac{0.5-\epsilon}{1-\epsilon},$$

$$0.5 \ge F_\epsilon\big(Med(F_\epsilon) - 0\big) \ge (1-\epsilon)F\big(Med(F_\epsilon) - 0\big) \quad \Rightarrow \quad F\big(Med(F_\epsilon - 0)\big) \ge \frac{0.5}{1-\epsilon},$$

where the above two equations give upper- and lower-bounds for $Med(F_\epsilon)$ independent of x. Then $b(\epsilon)$ cannot go to infinity in this case. In summary, $\epsilon^* = 0.5$. □

3. Suppose a positive random variable X, whose distribution is F. Denote $\theta = \int \log(x)dF(x), \lambda = \log(\mu), \mu = EX$.

(1) Solve the influence function and empirical influence function of θ and λ.

Solve. Since θ is a linear functional, its influence function and empirical influence function are

$$IF(x;\theta,F) = \log(x) - \theta, \quad IF(x;\theta,F_n) = \log(x) - \hat{\theta},$$

where $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n \log(X_i)$.

Notice that λ is a log-transform of a linear functional $\mu = \int x dF$, then using the chain rule, we have

$$IF(x;\lambda,F) = \frac{1}{\mu} IF(x;\mu,F) = \frac{1}{\mu}(x-\mu), \quad IF(x;\lambda,F_n) = \frac{1}{\hat{\mu}}(x-\hat{\mu}),$$

where $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$. $\square$

(2) Are the limit of $\hat{\theta}$ and $\hat{\lambda}$ the same?

Answer. These two are the plug-in estimators, *i.e.*,

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^n \log(X_i), \quad \hat{\lambda} = \log\left(\frac{1}{n}\sum_{i=1}^n X_i\right).$$

As $n \to \infty$,

$$\hat{\theta} \to \theta, \quad \hat{\lambda} \to \lambda.$$

From the Jensen's inequality, $\theta \le \lambda$, where $\theta = \lambda$ if and only if $X \overset{a.s.}{=} const$. That is, the limit of $\hat{\theta}$ and $\hat{\lambda}$ are not the same, unless X is almost surely a constant. $\square$

(3) Which one is more robust to the outliers, $\hat{\theta}$ or $\hat{\lambda}$?

Answer. For both the two estimators, the gross error sensitivities are $\gamma^*_\theta = \gamma^*_\lambda = \infty$ and the breakdown points are $\epsilon^*_\theta = \epsilon^*_\lambda = 0$. The local shift sensitivities are

$$\lambda^*_\theta = \sup_{0<x<y} \frac{\log(y) - \log(x)}{y - x} = \infty, \quad \lambda^*_\lambda = \frac{1}{\mu},$$

which implies that $\hat{\lambda}$ is more robust in the sense of local shift sensitivity. $\square$