Lection 6. Robustness

The Introduction

Another important property we consider is the robustness. Roughly, robust statistic is not affected by outliers. For example, a sample mean is not robust because one outlier in the sample can change it significally. On the other hand, a sample median is very robust statistic.

The robustness is also very important in parametric statistics, because robust statistic shows good performance in case of small departues from the parametric model.

Let $\mathcal{H}$ be a set of all distribution functions.

Definition 1. Consider an estimator $\hat{\theta} = f(\hat{F}_n)$. The maximum bias of $\hat{\theta}$ is

$$b(\varepsilon) = b(\varepsilon; F) = \sup_{F \in U_\varepsilon(F_0)} |f(F) - f(F_0)|,$$

where $U_\varepsilon(F_0)$ is a neighborhood of F_0 in $\mathcal{H}$ in some sense.

We'll consider two cases:
1) The Levy neighborhood:

$$U_\varepsilon(F_0) = U_\varepsilon^{(L)}(F_0) = \{F |\ \forall t\ F_0(t-\varepsilon) - \varepsilon \le F(t) \le F_0(t+\varepsilon) + \varepsilon\},$$

2) The contamination neighborhood

$$U_\varepsilon(F_0) = U_\varepsilon^{(C)}(F_0) = \{F |\ \exists H \in \mathcal{H} :\ F = (1-\varepsilon)F_0 + \varepsilon H.\}$$

Example 1. Let $F_0(x) = I_{x \ge 0}$. Then $U_\varepsilon^{(C)}(F_0)$ is the set of all distributions with $F(0) - F(0-) \ge 1 - \varepsilon$, $U_\varepsilon^{(L)}(F_0)$ is the set of all distributions with $F_0(\varepsilon) - F_0(-\varepsilon) \ge 1 - \varepsilon$. For example, the c.d.f. of $R[-1/4, 1/4]$ belongs to $U_{1/4}^{(L)}(F_0)$ (in fact, it belongs to $U_{1/6}^{(L)}(F_0)$), but doesn't belong even to $U_{0.99}^{(C)}(F_0)$.

Problem 1. In both cases 1) and 2) find the smallest ε such that $F_{\max(X,Y)}$, $X, Y \sim R[0,1]$ belongs to the ε-neighborhood of F_X

Example 2. Consider $f(F) = \int_{\mathbb{R}} x dF(x)$. Then,

$$b^{(C)}(\varepsilon) = \sup_{H \in \mathcal{H}} \left| \int_{\mathbb{R}} x d((1-\varepsilon)F(x) + \varepsilon H(x)) - \int_{\mathbb{R}} x dF(x) \right| = \varepsilon \sup_{H \in \mathcal{H}} \left| \int_{\mathbb{R}} x dH(x) - \int_{\mathbb{R}} x dF(x) \right| = \infty$$

for every F. Obviously, $b^{(C)}(\varepsilon) \le b^{(\varepsilon)}(F) = \infty$.

Example 3. Let $F_{\varepsilon,H}(x) = (1-\varepsilon)F(x) + \varepsilon H(x)$. Consider $f(F) = F^{-1}(1/2)$ be a median, $F^{-1}(x) = \inf\{u : F(u) \ge x\}$. Then,

$$b^{(C)}(F) = \sup_{H \in \mathcal{H}} \left| F_{\varepsilon,H}^{-1}(1/2) - F^{-1}(1/2) \right|$$

The smallest possible value for the $F_{\varepsilon,H}^{-1}(1/2)$ is $F^{-1}\left(\frac{1-2\varepsilon}{2(1-\varepsilon)}\right)$, the largest is $F^{-1}\left(\frac{1}{2(1-\varepsilon)}\right)$. Therefore,

$$b^{(C)}(F) = \max\left(F^{-1}(1/2) - F^{-1}\left(\frac{1-2\varepsilon}{2(1-\varepsilon)}\right), F^{-1}\left(\frac{1}{2(1-\varepsilon)}\right) - F^{-1}(1/2) \right).$$

Similarly,

$$b^{(C)}(F) = \max\left(F^{-1}(1/2) - F^{-1}\left(\frac{1-2\varepsilon}{2(1-\varepsilon)}\right), F^{-1}\left(\frac{1}{2(1-\varepsilon)}\right) - F^{-1}(1/2) \right) + \varepsilon.$$

The Asymptotic Breakdown Point and Gross Sensivity

Definition 2. The *asymptotic breakdown point* of f at F_0 is

$$\varepsilon^* = \varepsilon^*(F_0, f) = \varepsilon^*(F_0, \hat{\theta}) = \sup\{\varepsilon : b(\varepsilon) < b(1)\}.$$

The $b(1)$ is the worst value of $f(F) - f(F_0)$, so, roughly speaking, the breakdown point give as the limiting fraction of outliers the estimator can cope with.

Example 4. For the mean $\hat{\theta} = \overline{X}$ we have $b(\varepsilon) = \infty$ for every $\varepsilon > 0$. Therefore, $\varepsilon^*(F_0, \overline{X}) = 0$.
For the median $\hat{\theta} = MED(X)$ we have $b(1/2) = \infty$, $b(\varepsilon) < \infty$ for every $\varepsilon < 1/2$. Therefore, $\varepsilon^*(F_0, MED) = 1/2$.

Definition 3. The estimator is said to be *robust* if its asymptotic breakdown point is greater then zero.

Definition 4. The *gross sensitivity* of an estimator is $\sup_x |L_F(x)|$.

A gross sensivity is related with the reaction on small changes of c.d.f. F. An asymptotic breakdown point means maximal possible fraction of outliers the estimator can cope with.

Example 5. The median $F^{-1}(1/2)$ has no Gateuax derivative at

$$F = \begin{cases} 0, & x < 0, \\ \frac{1}{2}, & x \in [0, 1), \\ 1, & x \geq 1. \end{cases}$$

So, it's gross sensivity is equal to ∞. However, it's robust estimator with asymptotic breakdown point $1/2$ for any F.

Symmetrical Distributions

Let

$$\mathcal{F}_{symm} = \{F : \exists \theta = \theta(F) : \ F(\theta - x - 0) = 1 - F(\theta + x), F'(x) = p(x) > 0 \text{ as } x \in (\theta - c, \theta + c) \text{ for some } c \leq \infty\}$$

be a class of absolutely continuous symmetrical distributions. If there exists $\mathbf{E}X$, then $\mathbf{E}X = x_{1/2} = \theta$ and therefore MED and $\overline{X}$ are estimators for the same parameter.

As $\mathbf{D}X < \infty$ the estimator $\overline{X}$ is asymptotically normal with the asymptotical variance $\mathbf{D}X$. However, it's not robust.
The estimator MED is asymptotically normal with the asymptotical variance $1/(4p^2(x_{1/2}))$. It's robust with asymptotic breakdown point $1/2$.

Example 6. Suppose that $X_i \sim \mathcal{N}(\theta, \sigma^2)$. Then $\overline{X}$ has asymptotical variance σ^2, MED has asymptotical variance

$$\left(\frac{1}{\frac{2}{\sqrt{2\pi}\sigma}} \right)^2 = \frac{\pi \ \sigma^2}{2}.$$

Therefore, MED has greater variance.

Example 7. Suppose that $X_i \sim Cauchy(\theta)$. Then $\overline{X}$ is a bad estimator, because $\overline{X} \stackrel{d}{=} X_1$. However, MED is asymptotically normal estimator with asymptotical variance

$$\left(\frac{1}{\frac{2}{\pi}} \right)^2 = \frac{\pi^2}{4}.$$