Lection 5. Robust Statistics

Let

$$\widehat{L}(x) = \widehat{L}(x; f) = \lim_{\varepsilon \to 0} \frac{f((1-\varepsilon)\widehat{F}_n + \varepsilon\delta_x) - f(\widehat{F}_n)}{\varepsilon}$$

be a Gateuax derivative of f at the point $\widehat{F}_n$. If $L_F(x)$ is a continuous functional of F under the uniform norm, then

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{L}^2(X_i)$$

is a consistent estimator for $\sigma^2(F)$. Therefore,

$$P_F\left(f(F) \in \left(f(\widehat{F}_n) - \frac{z_{1-\alpha/2}\widehat{\sigma}}{\sqrt{n}}, f(\widehat{F}_n) + \frac{z_{1-\alpha/2}\widehat{\sigma}}{\sqrt{n}}\right)\right) \to 1-\alpha.$$

This interval is called the *infinitesmall jackknife* interval.

Example 1. Let $f(F) = \int_{\mathbb{R}} a(u)dF(u)$, where a is bounded. Then

$$L_F(x) = \lim_{\varepsilon \to 0} \frac{f((1-\varepsilon)F + \varepsilon\delta_x) - f(F)}{\varepsilon} = a(x) - \int_{\mathbb{R}} a(u)dF(u).$$

Therefore, L_F is a continuous functional of F and

$$\overline{\widehat{L}^2(X)} = \frac{1}{n}\sum_{i=1}^{n}\left(a(X_i) - \overline{a(X)}\right)^2 = \overline{a(X)^2} - \overline{a(X)}^2$$

is a consistent estimator of $\sigma^2(F)$. So,

$$\left(\overline{a(X)} - \frac{z_{1-\alpha/2}\sqrt{\overline{a(X)^2} - \overline{a(X)}^2}}{\sqrt{n}}, \overline{a(X)} + \frac{z_{1-\alpha/2}\sqrt{\overline{a(X)^2} - \overline{a(X)}^2}}{\sqrt{n}}\right)$$

is an asymptotical $1-\alpha$ confidence interval for $\mathbf{E}a(X)$.

Example 2. Consider $f(F) = F^{-1}(1/2)$. Then

$$\sqrt{n}\frac{f(\widehat{F}_n) - f(F)}{\frac{1}{2p(x_{1/2})}} \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

Therefore, we need to estimate $p(x_{1/2})$. We can't estimate it by $\overline{\widehat{L}(X)}$ since f is not Hadamard differentiable at $\widehat{F}$.

The Jackknife Method

Let's consider another estimator for $\sigma^2(F)$. Let

$$\widehat{F}_{n-1,i}(x) = \widehat{F}_{n-i}(x; x_1, ...x_{i-1}, x_{i+1}, ..., x_n) = \frac{1}{n-1}\sum_{j \neq i} I_{x_j \leq x}.$$

Then

$$L_{\widehat{F}_n}(x_i) = \lim_{\varepsilon \to 0} \frac{f\left((1-\varepsilon)\widehat{F}_n + \varepsilon\delta_{x_i}\right) - f(\widehat{F}_n)}{\varepsilon} \approx \frac{f\left(\left(1+\frac{1}{n-1}\right)\widehat{F}_n - \frac{1}{n-1}\delta_{x_i}\right) - f(\widehat{F}_n)}{-\frac{1}{n-1}} = (n-1)(f(\widehat{F}_n) - f(\widehat{F}_{n-1,i})).$$

Therefore, it's natural to estimate $\sigma^2(F) = \int_{\mathbb{R}} L^2(x)dF(x) = D_F L^2(X)$ by

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(L_{\widehat{F}_n}(x_i) - \overline{L_{\widehat{F}_n}(x)}\right)^2 \approx (n-1)\sum_{i=1}^{n}\left(f(\widehat{F}_{n-1,i}) - \overline{f(\widehat{F}_{n-1,\cdot})}\right)^2 =: nS^2_{jack}.$$

So, it's natural to use an interval

$$f(F) \in \left(f(\widehat{F}_n) - z_{1-\alpha/2}S_{jack}, f(\widehat{F}_n) + z_{1-\alpha/2}S_{jack}\right).$$

This interval is called the *jackknife* interval. It's often used to estimate the variance $D_F\widehat{\theta}(X_1, ..., X_n)$ of an estimator $\widehat{\theta}$. The jackknife estimator is

$$S^2_{jack} = \frac{n-1}{n}\sum_{i=1}^{n}(\widehat{\theta}(-i) - \overline{\widehat{\theta}})^2,\ \widehat{\theta}(-i) := \widehat{\theta}(X_1, ..., X_{i-1}, X_{i+1}, ..., X_n),\ \overline{\widehat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}(-i).$$

Example 3. Let $\widehat{\theta}(X_1, ..., X_n) = \overline{X}$. Then

$$\widehat{\theta}(-i) = \frac{1}{n-1}(\overline{X}n - X_i),\ \overline{\widehat{\theta}} = \overline{X}.$$

So

$$S^2_{jack} = \frac{S^2_0}{n}.$$

It's a natural estimator for the $\mathbf{D}\overline{X}$.

Example 4. Consider a sample $X_1, ..., X_{2n+1}$ such that $X_{(n-1)} = X_{(n)} = X_{(n+1)}$ and let $\widehat{\theta}(X_1, ..., X_n) = MED$. Then $\widehat{\theta}(-i) = \widehat{\theta}(X_1, ..., X_n) = MED$ and $S^2_{jack} = 0$. In this situation the jackknife estimator is nonapplicable since $L_F(x; MED)$ is not continuous as functional of F.

The Bootstrap

Let's start with the following theorem (without proof):

Theorem 1. *Let F be a distribution function, $\hat{F}_n$ be an empirical distribution function and $\hat{F}^*_n$ be an empirical distribution function for c.d.f. $\hat{F}_n$. Suppose, that f is Hadamard differentiable functional, $\sigma(F) > 0$*

$$\sqrt{n}\frac{f(\hat{F}_n) - f(F)}{\sigma(F)} \xrightarrow{d} Z \sim \mathcal{N}(0, 1).$$

Then

$$\mathbf{P}_{\hat{F}_n}\left(\sqrt{n}\frac{f(\hat{F}^*_n) - f(\hat{F}_n)}{\sigma(F)} \le x \middle| X_1, ..., X_n\right) \to \Phi(x)$$

for a.s. $X_1, ..., X_n, ...$.

Due to this theorem, we can estimate $\sigma(F)$ by the use of $f(\hat{F}^*_n)$, $f(\hat{F}_n)$ in the following way:
1) Draw $X^*_1, ..., X^*_n$ from $\hat{F}_n$.
2) Compute $\hat{F}^*_n$ and $f(\hat{F}^*_n)$.
3) Repeat steps 1), 2) k times.
4) Compute

$$\hat{\sigma}^2(F) = \frac{n}{k}\sum_{i=1}^{k}\left(f(\hat{F}^*_{n,i}) - f(\hat{F}_n)\right)^2$$

Someone can notice that 1) means "to draw a sample $X^*_1, ..., X^*_n$ with replacement from $X_1, ..., X_n$".

The idea of the bootstrap is very simple — we can generate samples from $\hat{F}_n$ since it's known distribution.

Therefore, we can estimate any parameters of $\hat{F}_n$ using consistent estimators. Due to Theorem (1) it's enough to estimate the asymptotic variance of $f(\hat{F}_n)$.

Example 5. Now we can construct an estimator of $\sigma^2_{MED}(F)$. The bootstrap give us the solution:

$$\hat{\sigma}^2(F) = \frac{n}{k}\sum_{i=1}^{k}\left(f(\hat{F}^*_{n,i}) - f(\hat{F}_n)\right)^2,$$

where $f(F) = F^{-1}(1/2)$.

The method is very popular in practice because of its simplicity. However, it fails in several situations, particularly when the rate of convergence of f is not $n^{-1/2}$:

Example 6. Consider $f(F) = \left(\int_{\mathbb{R}} x dF(x)\right)^2$, $\mathbf{E}X = 0$, $\mathbf{D}X = 1$. Then

$$L_F(f) = 2\mathbf{E}X(x - \mathbf{E}X) = 0.$$

Therefore, we can't apply Theorem 1. Let's show that the bootstrap concept fails in this situation. Really,

$$nf(\hat{F}_n) \xrightarrow{d} Z \sim \chi^2_1.$$

By theorem 1

$$\mathbf{P}\left(\sqrt{n}\frac{\overline{X}^* - \overline{X}}{DX} \le x \middle| X_1, ..., X_n\right) \to \Phi(x),$$

so

$$(\sqrt{n}(\overline{X}^* - \overline{X}), \sqrt{n}\overline{X}) \xrightarrow{d} (Z_1, Z_2) \sim \mathcal{N}(0, E).$$

Therefore,

$$n(f(\hat{F}^*_n) - f(\hat{F}_n)) = n((\overline{X}^* - \overline{X})^2 + 2\overline{X}(\overline{X}^* - \overline{X}) \to Z_1^2 + 2Z_1Z_2,$$

where Z_1, Z_2 are independent $\mathcal{N}(0,1)$. It's not χ^2_1 distribution. So, asymptotical distributions of $f(\hat{F}_n)$ and $f(\hat{F}^*_n)$ can be significally different if $\sigma(F) = 0$.

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 even one large 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:

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

2) The contamination neighborhood

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

Example 7. Let $F_0(x) = I_{x\geq 0}$. Then $U_\varepsilon^{(C)}(F_0)$ is the set of all distributions with $F(0) - F(0-) \geq 1-\varepsilon$, $U_\varepsilon^{(L)}(F_0)$ is the set of all distributions with $F_0(\varepsilon) - F_0(-\varepsilon) \geq 1-\varepsilon$. For example, the c.d.f. of $R[-1/4, 1/4]$ belongs to $U_{1/4}^{(L)}(F_0)$ 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 8. 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, $b^{(C)}(\varepsilon) \leq b^{(\varepsilon)}(F) = \infty$.

Example 9. Let $F_{\varepsilon,H}(x) = (1-\varepsilon)F(x) + \varepsilon H(x)$. Consider $f(F) = F^{-1}(1/2)$ be a median, $F^{-1}(x) = \inf\{u : F(u) \geq x\}$. 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

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