

6 Statistical Functionals and Robustness

If nothing else is assumed, X_1, X_2, \dots, X_N will be i.i.d., real-valued random variables with distribution function F throughout the whole chapter. With some abuse of notation, F will sometimes also denote the distribution $\mathcal{L}(X_j)$ of X_j .

$$\hat{F}_N(x) = \frac{1}{N} \sum_{j=1}^N 1_{(-\infty, x]}(X_j) = \frac{1}{N} \#\{j \leq N; X_j \leq x\}$$

denotes the empirical distribution function of the sample X_1, X_2, \dots, X_N .

Many statistics can be written as $T_N = T(\hat{F}_N)$, where $T : \mathcal{F} \rightarrow \mathbb{R}$ is some functional on \mathcal{F} , the set of all distribution functions on \mathbb{R} , i.e.

$$\mathcal{F} = \{F : \mathbb{R} \rightarrow [0, 1]; F \text{ increasing, right-continuous, } F(x) \xrightarrow{x \rightarrow -\infty} 0, F(x) \xrightarrow{x \rightarrow +\infty} 1\}$$

By the Glivenko-Cantelli Theorem: $\sup_x |\hat{F}_N(x) - F(x)| \xrightarrow{N \rightarrow \infty} 0$.

Therefore, we expect: if T is smooth, then

$$T_N = T(\hat{F}_N) \xrightarrow{N \rightarrow \infty} T(F),$$

i.e. T_N is a consistent estimate of $T(F)$.

Examples:

a) For some measurable $h(x)$,

$$\begin{aligned} T(F) &= \int h(x) dF(x) = \mathcal{E} h(X_1) \\ T(\hat{F}_N) &= \int h(x) \hat{F}_N(x) = \frac{1}{N} \sum_{j=1}^N h(X_j). \end{aligned}$$

Here, T is a linear statistical functional, i.e. for $F, G \in \mathcal{F}$

$$T(\alpha F + \beta G) = \alpha T(F) + \beta T(G)$$

if $0 \leq \alpha, \beta \leq 1$, $\alpha + \beta = 1$, which implies $\alpha F + \beta G \in \mathcal{F}$, too.

In particular, for $h(x) = x^k$, we get $T(\hat{F}_N) = \frac{1}{N} \sum_{j=1}^N X_j^k$, the k -th sample moment.

b) sample variance $\hat{\sigma}_N^2$:

$$\begin{aligned} T(F) &= \int \left(x - \int z dF(z)\right)^2 dF(x) = \text{var } X_1, \\ T(\hat{F}_N) &= \frac{1}{N} \sum_{j=1}^N \left(X_j - \frac{1}{N} \sum_{i=1}^N X_i\right)^2 = \hat{\sigma}_N^2. \end{aligned}$$

c) Consider the parametric model

$$\mathcal{M}_\Theta : \mathcal{L}(X_j) \in \{P_\theta, \theta \in \Theta\}$$

Let F_θ = denote the cumulative distribution function of P_θ , i.e. $F_\theta(x) = P_\theta(-\infty, x]$, and let $F = \mathcal{L}(X_j)$ denote the true distribution function of the data.

If $F \neq F_\theta$ for all $\theta \in \Theta$, \mathcal{M}_Θ does not hold; in this case, we speak of a misspecified model.

Nevertheless, we can pretend that \mathcal{M}_Θ holds, and calculate the ML-estimate $\hat{\theta}_N$. If we explicitly take into account misspecification, $\hat{\theta}_N$ is called quasi- (or pseudo-) maximum likelihood (QML) estimate.

For $x \in (R)$, let $L(\theta|x), l(\theta|x)$ denote the one-point likelihood resp. log-likelihood, i.e. if P_θ has a density p_θ : $L(\theta|x) = p_\theta(x), l(\theta|x) = \log p_\theta(x)$.

Likelihood: $L_N(\theta|X) = \prod_{j=1}^N L(\theta|X_j)$ where $X = (X_1, \dots, X_N)^T$.

log-likelihood: $l_N(\theta|X) = \sum_{j=1}^N l(\theta|X_j) = N \int l(\theta|x) d\hat{F}_N(x)$

$$\hat{\theta}_N = \arg \max_{\theta \in \Theta} \int l(\theta|x) d\hat{F}_N(x) =: T(\hat{F}_N)$$

$$\theta_0 = \arg \max_{\theta \in \Theta} \int l(\theta|x) dF(x) =: T(F).$$

Under suitable identifiability assumptions, θ_0 is well-defined even if \mathcal{M}_Θ is misspecified. θ_0 is the (unknown) parameter corresponding to that P_θ for which F_θ is as similar as possible to F .

Alternatively, under appropriate assumptions, $\hat{\theta}_N$ is given as a zero of

$$\frac{\partial}{\partial \theta} l_N(\theta|X) = \psi_N(\theta|X) = \sum_{j=1}^N \psi(\theta|X_j)$$

i.e. $\hat{\theta}_N = T(\hat{F}_N)$ is implicitly defined by

$$\frac{1}{N} \sum_{j=1}^N \psi(\hat{\theta}_N|X_j) = \int \psi(\hat{\theta}_N|x) d\hat{F}_N(x) = 0.$$

Correspondingly, $\theta_0 = T(F)$ can be defined as solution of

$$\int \psi(\theta_0|x) dF(x) = 0$$

d) Chi-square goodness-of-fit statistic: Let $I_1 \cup \dots \cup I_q = \mathbb{R}$ be a partition of \mathbb{R} into disjoint intervals, and $p_k = \text{pr}(X_j \in I_k)$. Let Z_k denote the number of observations X_j in the interval $I_k, k = 1, \dots, q$. We have

$$\frac{Z_k}{N} = \frac{1}{N} \sum_{j=1}^N 1_{I_k}(X_j) = \int_{I_k} 1_{I_k}(x) d\hat{F}_N(x),$$

and, therefore, the test statistic for testing the hypothesis $H_0 : p_k = p_k^0$ is

$$\begin{aligned}\chi^2 &= \sum_{k=1}^q \frac{(Z_k - N p_k^0)^2}{N p_k^0} = N \cdot \sum_{k=1}^q \frac{1}{p_k^0} \left(\int_{I_k} d\hat{F}_N(x) - p_k^0 \right)^2 = N T(\hat{F}_N) \\ T(F) &= \sum_{k=1}^q \frac{1}{p_k^0} \left(\int_{I_k} dF(x) - p_k^0 \right)^2\end{aligned}$$

e) For testing $H_0 : F = F_0$, the Cramér - von Mises test statistics is given by

$$\begin{aligned}T(\hat{F}_N) &= \int (\hat{F}_N(x) - F_0(x))^2 dF_0(x) \\ T(F) &= \int (F(x) - F_0(x))^2 dF_0(x)\end{aligned}$$

whereas the test statistic of the Kolmogorov-Smirnov test is

$$T(\hat{F}_N) = \sup_x |\hat{F}_N(x) - F_0(x)|, \quad T(F) = \sup_x |F(x) - F_0(x)|$$

Guideline (von Mises). The type of asymptotic distribution of $T_N = T(\hat{F}_N)$ depends on which is the first non-vanishing term in the Taylor expansion of T at F :

if the linear term does not vanish, then T_N asymptotically normal, i.e.

$$\sqrt{N}(T(\hat{F}_N) - T(F)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \dots)$$

if the linear term vanishes, we get "higher" types of limit laws, i.e. if the first nonvanishing term is the m -th order term,

$$N^{\frac{m}{2}}(T(\hat{F}_N) - T(F)) \xrightarrow{\mathcal{L}} \dots$$

6.1 Asymptotics beyond LLN and CLT

We are interested in more detailed information about the asymptotic distribution of statistics for $N \rightarrow \infty$ than provided by the LLN or CLT. Let $E X_j = \mu$, $\text{var } X_j = \sigma^2 < \infty$, such that

$$Z_N = \frac{1}{\sqrt{N}} \sum_{j=1}^N \frac{X_j - \mu}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \quad \text{by CLT.}$$

Theorem 6.1 (Law of the iterated logarithm (LIL), Hartman and Wintner):

Let X_1, \dots, X_N be i.i.d., $E X_j = \mu$, $\text{var } X_j = \sigma^2 < \infty$. Then,

$$\limsup_{N \rightarrow \infty} \frac{1}{\sqrt{2 \log \log N}} \frac{1}{\sqrt{N}} \sum_{j=1}^N \frac{X_j - \mu}{\sigma} = \limsup_{N \rightarrow \infty} \frac{Z_N}{\sqrt{2 \log \log N}} = 1 \quad \text{a.s.}$$

i.e., for any $\varepsilon > 0$:

$$\begin{aligned}pr \left(\frac{Z_N}{\sqrt{2 \log \log N}} > 1 + \varepsilon \text{ for only finitely many } N \right) &= 1 \\ pr \left(\frac{Z_N}{\sqrt{2 \log \log N}} > 1 - \varepsilon \text{ infinitely often} \right) &= 1\end{aligned}$$

Let G_N denote the distribution function of Z_N . By the CLT:

$$G_N(t) \xrightarrow[N \rightarrow \infty]{} \Phi(t) \quad \text{for all } t.$$

This convergence is uniform with rate $\frac{1}{\sqrt{N}}$:

Theorem 6.2 (Berry-Esséen): Let X_1, \dots, X_N be i.i.d., $E X_j = \mu$, $\text{var} X_j = \sigma^2$, $\gamma = E|X_j - \mu|^3 < \infty$. Then,

$$\sup_t |G_N(t) - \Phi(t)| \leq c \cdot \frac{\gamma}{\sigma^3 \sqrt{N}} \quad \text{for all } N \geq 1$$

(currently best value: $c = 0.8$, van Beek (1972)).

Theorem 6.3 (Edgeworth expansion) Let X_1, \dots, X_N be i.i.d. with $E|X_j|^m < \infty$ for some $m > 3$. Let $\mu_1 = E X_j$, $\mu_k = E (X_j - E X_j)^k$, $k = 2, \dots, m$. Then, for $N \rightarrow \infty$,

$$\sup_t |G_N(t) - \Phi(t) - \varphi(t) \sum_{k=3}^m N^{-\frac{k}{2}+1} R_k(t)| = o(N^{-\frac{m}{2}+1}),$$

where $R_k(t)$ is a polynomial in t depending only on μ_2, \dots, μ_m , but not on N, m or otherwise on F .

In particular, for the first non-trivial case ($m = 3$), $R_3(t) = \frac{\mu_3}{6\sigma^3}(1 - t^2)$ and

$$G_N(t) = \Phi(t) + \frac{\mu_3}{6\sigma^3 \sqrt{N}}(1 - t^2)\varphi(t) + o\left(\frac{1}{\sqrt{N}}\right)$$

6.2 Asymptotics based on differentiation of functionals

In the context of ML-estimates, asymptotic normality is proven by linearization (delta method). We transfer this idea to general estimates based on functionals.

Idea: Assume Taylor expansion

$$T(\hat{F}_N) = T(F) + T'(F; \hat{F}_N - F) + \frac{1}{2}T''(F; \hat{F}_N - F) + \dots,$$

or, more precisely, for some m

$$\begin{aligned} T(\hat{F}_N) - T(F) &= \sum_{k=1}^m \frac{1}{k!} T^{(k)}(F; \hat{F}_N - F) + R_{m,N} \\ &= V_{m,N} + R_{m,N} \end{aligned}$$

$V_{m,N}$ multilinear ($m = 1$: linear, $m = 2$: bilinear, ...)

Proving asymptotic normality of $T(\hat{F}_N)$: Consider $m = 1$.

a) Show $\sqrt{N}R_{1,N} \xrightarrow{\text{p}} 0$

b) Check: $V_{1,N} = T'(F; \hat{F}_N - F)$ is sample mean of i.i.d. mean 0-variables. Then, by CLT,

$$\sqrt{N}V_{1,N} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(T, F))$$

c) a)+b) imply $\sqrt{N}(T(\hat{F}_N) - T(F)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(T, F))$.

Proving LIL: Consider $m = 1$. Analogously as asymptotic normality, but instead of a) show

$$\begin{aligned} \sqrt{N}R_{1,N} &= o(\sqrt{\log \log N}) \quad \text{a.s.} \\ \implies \limsup_{N \rightarrow \infty} \frac{\sqrt{N}(T(\hat{F}_N) - T(F))}{\sigma(T, F) \sqrt{2 \log \log N}} &= 1 \quad \text{a.s.} \end{aligned}$$

Proving Berry-Esséen rate of convergence: Consider $m = 2$.

Show $\text{pr}(|R_{2,N}| > \frac{c}{\sqrt{N}}) = O(\frac{1}{\sqrt{N}})$ for some $c > 0$, which implies

$$\sup_t \left| \text{pr} \left(\frac{\sqrt{N}(T(\hat{F}_N) - T(F))}{\sigma(T, F)} \leq t \right) - \Phi(t) \right| = O\left(\frac{1}{\sqrt{N}}\right).$$

If $\text{pr}(V_{1,N} = c) = 1$, i.e. $T'(F, \hat{F}_N - F)$ is a degenerate random variable, then find smallest m s.th. $V_{m,N}$ not degenerate, show $N^{m/2}R_{m,N} \xrightarrow{p} 0$ which implies

$$N^{\frac{m}{2}}(T(\hat{F}_N) - T(F)) \xrightarrow{\mathcal{L}} \dots$$

Typically, $m = 2$ in that case (e.g. χ^2 -test statistic, Cramér-von Mises test).

In functional analysis, there are various concepts of differentiation. The simplest one is:

Definition: Let \mathcal{V} be a linear space, $T : \mathcal{V} \rightarrow \mathbb{R}$ a functional on \mathcal{V} . T is Gateaux differentiable at u in direction v ($u, v \in \mathcal{V}$) if there is a linear functional $L_u : \mathcal{V} \rightarrow \mathbb{R}$ such that

$$\lim_{\lambda \rightarrow 0} \frac{T(u + \lambda v) - T(u)}{\lambda} = L_u(v)$$

$T'(u; v) := L_u(v)$ is called the Gateaux derivative of T at u in direction v . If L_u is defined for all $v \in \mathcal{V}$, then $T'(u; \cdot) = L_u$ is called the Gateaux derivative of T at u (i.e. the derivative of a functional T is a linear functional).

Of course, setting $g(\lambda) = T(u + \lambda v)$, we just have $T'(u; v) = g'(0)$ (derivative in \mathbb{R}).

We now consider \mathcal{F} (= set of distribution functions). \mathcal{F} is not a linear space, however a convex set.

Definition: Let $T : \mathcal{F} \rightarrow \mathbb{R}$ be a functional on \mathcal{F} . For $F, G \in \mathcal{F}$, let $F_\lambda = (1 - \lambda)F + \lambda G = F + \lambda(G - F)$, $0 \leq \lambda \leq 1$ ($F_\lambda \in \mathcal{F}$ too, by convexity).

$$\lim_{\lambda \rightarrow 0+} \frac{T(F + \lambda(G - F)) - T(F)}{\lambda} = T'(F; G - F)$$

is called Gateaux derivative of T at F in direction $G - F$. If this exists for all $G \in \mathcal{F}$ and if $T'(F; \cdot)$ linear, then T is called Gateaux differentiable at F .

In some statistics textbooks, $T'(F; G)$ is written instead of $T'(F; G - F)$, but that does not conform with the use of Gateaux derivatives in functional analysis.

T is not necessarily defined for all distributions $F \in \mathcal{F}$; e.g. if T represents the variance.

Examples:

a) $T(F) = \int h(x) dF(x)$ is linear functional itself. In this case

$$\begin{aligned} T(F + \lambda(G - F)) &= T((1 - \lambda)F + \lambda G) = (1 - \lambda) \int h(x) dF(x) + \lambda \int h(x) dG(x) \\ \implies \frac{T(F + \lambda(G - F)) - T(F)}{\lambda} &= \int h(x) dG(x) - \int h(x) dF(x) \quad (\text{for all } \lambda) \\ \implies T'(F; G - F) &= \int h(x) d(G - F)(x) = T(G - F) = T(G) - T(F) \\ \implies T'(F; \cdot) &= T \quad \text{for all } F. \end{aligned}$$

b) Let X_1, X_2 be independent with $\mathcal{L}(X_1) = \mathcal{L}(X_2) = F$

$$T(F) = \iint h(x_1, x_2) dF(x_1) dF(x_2) = \mathbb{E} h(X_1, X_2)$$

Then,

$$\begin{aligned} T(F + \lambda(G - F)) &= \iint h(x_1, x_2) d(F + \lambda(G - F))(x_1) d(F + \lambda(G - F))(x_2) \\ &= \underbrace{\iint h(x_1, x_2) dF(x_1) dF(x_2)}_{=T(F)} + \lambda \iint [h(x_1, x_2) + h(x_2, x_1)] dF(x_1) d(G - F)(x_2) + \lambda^2 \dots \end{aligned}$$

Subtracting $T(F)$, dividing by λ and letting $\lambda \rightarrow 0$, we get

$$T'(F; G - F) = \int \left\{ \int [h(x_1, x_2) + h(x_2, x_1)] dF(x_1) \right\} d(G - F)(x_2)$$

linear in $G - F$.

We get higher order derivatives analogously:

$$T^{(k)}(F; G - F) = \frac{d^k}{d\lambda^k} g(0+), \quad \text{where } g(\lambda) = T(F + \lambda(G - F))$$

i.e. as the common right-sided k -th derivative of $g(\lambda)$ at 0.

If $g(\lambda)$ is regular enough, we get the Taylor-expansion

$$g(1) - g(0) = \sum_{k=1}^n \frac{1}{k!} g^{(k)}(0) 1^k + \frac{1}{(m+1)!} g^{(m+1)}(\lambda^*) 1^{m+1} \quad \text{for some } 0 \leq \lambda^* \leq 1. \quad (1)$$

In terms of T :

$$T(G) - T(F) = \sum_{k=1}^m \frac{1}{k!} T^{(k)}(F; G - F) + R_m(G).$$

For $G = \hat{F}_N$, we get as above with $R_{m,N} = R_m(\hat{F}_N)$:

$$T(\hat{F}_N) - T(F) = V_{m,N} + R_{m,N}. \quad (2)$$

$V_{m,N}$ typically is of a nice form easy to handle asymptotically. To get asymptotic properties of $T(\hat{F}_N) - T(F)$, we have to show that the remainder is negligible, i.e.

$$N^{\frac{m}{2}} R_{m,N} \xrightarrow{p} 0.$$

If the Taylor expansion in the form (1) holds, it is typically shown that

$$N^{\frac{m}{2}} \sup_{0 \leq \lambda \leq 1} |g_N^{(m+1)}(\lambda)| \xrightarrow{p} 0 \quad \text{with } g_N(\lambda) = T(F + \lambda(\hat{F}_N - F)).$$

Considering the supremum solves the otherwise complicated problem that λ^* depends on the random \hat{F}_N and the unknown F .

In this approach, we require one degree more of differentiability ($m+1$) than necessary for just defining $V_{m,N}$. Often, the remainder $R_{m,N}$ of (2) can be handled directly without requiring that it is of the form $\frac{1}{(m+1)!} g^{(m+1)}(\lambda^*)$.

Theorem 6.4 *Let X_1, \dots, X_N be i.i.d. with distribution function F ; let T be a functional with Taylor expansion of order 1*

$$T(\hat{F}_N) - T(F) = V_{1,N} + R_{1,N},$$

where $\sqrt{N}R_{1,N} \xrightarrow{p} O$ and

$$V_{1,N} = \frac{1}{N} \sum_{j=1}^N h(F; X_j).$$

Let $\mu(T; F) = E h(F; X_1) = \int h(F; x) dF(x)$ and

$$0 < \sigma^2(T; F) = \text{var } h(F; X_1) = \int (h(F; x) - \mu(T; F))^2 dF(x) < \infty.$$

Then,

$$\sqrt{N}(T(\hat{F}_N) - T(F) - \mu(T; F)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(T; F)).$$

Proof: $h(F; X_j), j = 1, \dots, N$, are i.i.d. with mean $\mu(T; F)$, variance $\sigma^2(T; F)$. By CLT for i.i.d. data,

$$\sqrt{N} \frac{V_{1,N} - \mu(T; F)}{\sigma(T; F)} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Apply the assumption on $R_{1,N}$ and Slutsky's Lemma to finish the proof. \square

Another important case are statistics where we need a Taylor expansion up to order 2 as the first term is degenerate.

Theorem 6.5 Let X_1, \dots, X_N be i.i.d. with distribution function F ; let T be a functional with Taylor expansion of order 2.

$$T(\hat{F}_N) - T(F) = V_{1,N} + \frac{1}{2}V_{2,N} + R_{2,N}.$$

Assume $NR_{2,N} \xrightarrow{p} 0$ and

$$\begin{aligned} V_{1,N} &= \frac{1}{N} \sum_{j=1}^N h_1(F; X_j) \quad \text{with} \quad \text{var } h_1(F; X_1) = 0 \\ V_{2,N} &= \frac{1}{N^2} \sum_{i,j=1}^N h(F; X_i, X_j) \end{aligned}$$

where $h(F; x, y)$ satisfies

$$h(F; x, y) = h(F; y, x) \quad \text{for all } x, y$$

$$E h^2(F; X_1, X_2) < \infty, \quad E |h(F; X_1, X_1)| < \infty$$

$$E h(F; x, X_1) = c \quad \text{for all } x, \text{ some constant } c.$$

Let $\mu(T; F) = E h(F; X_1, X_2)$, and let $\lambda_1, \lambda_2, \dots$ denote the eigenvalues of the operator $A : L^2(F) \rightarrow L^2(F) = \{\gamma; \int \gamma^2(x) dF(x) < \infty\}$ given by

$$A\gamma(x) = \int [h(F; x, y) - \mu(T; F)] \gamma(y) dF(y), \quad x \in \mathbb{R}.$$

Then,

$$N(T(\hat{F}_N) - T(F) - \mu(T; F)) \xrightarrow{\mathcal{L}} \sum_{k=1}^{\infty} \lambda_k Z_k^2$$

where Z_1, Z_2, \dots i.i.d. $\mathcal{N}(0, 1)$.

Proof: As Theorem 6.4, but instead of CLT, asymptotic results for so-called U-statistics are used. □

We also get higher-order asymptotics from the Taylor expansion of $T(F)$:

Theorem 6.6 Under the assumptions of Theorem 6.4 and, additionally,

$R_{1,N} = O\left(\sqrt{\frac{\log \log N}{N}}\right)$ a.s., we have

$$\limsup_{N \rightarrow \infty} \frac{\sqrt{N}(T(\hat{F}_N) - T(F) - \mu(T; F))}{\sigma(T; F)\sqrt{2 \log \log N}} = 1 \quad \text{a.s.}$$

In a similar manner, one also gets Berry-Esséen rates.

Applications

A: Sample moments

The k -th central moment of $\mathcal{L}(X_j)$ is given as

$$\mu_k = \mathbb{E} (X_j - \mathbb{E} X_j)^k = \int [x - \int y dF(y)]^k dF(x) = T(F).$$

The corresponding sample moment is

$$T(\hat{F}_N) = \frac{1}{N} \sum_{j=1}^N (X_j - \bar{X}_N)^k = \hat{\mu}_k.$$

Set $\mu_F = \int x dF(x)$, $F_\lambda = F + \lambda(G - F)$. Then,

$$\mu_{F_\lambda} = \mu_F + \lambda(\mu_G - \mu_F)$$

$$T(F_\lambda) = \int (x - \mu_{F_\lambda})^k dF(x) + \lambda \int (x - \mu_{F_\lambda})^k d(G - F)(x)$$

$$\begin{aligned} \frac{d}{d\lambda} T(F_\lambda) &= \int \frac{d}{d\lambda} (x - \mu_{F_\lambda})^k dF(x) + \int (x - \mu_{F_\lambda})^k d(G - F)(x) \\ &\quad + \lambda \cdot \int \frac{d}{d\lambda} (x - \mu_{F_\lambda})^k d(G - F)(x) \\ &= \int \frac{d}{d\lambda} (x - \mu_F - \lambda(\mu_G - \mu_F))^k dF_\lambda(x) + \int (x - \mu_{F_\lambda})^k d(G - F)(x) \end{aligned}$$

For $\lambda = 0$, we get

$$\begin{aligned} T'(F; G - F) &= -k(\mu_G - \mu_F) \int (x - \mu_F)^{k-1} dF(x) + \int (x - \mu_F)^k d(G - F)(x) \\ &= \int [(x - \mu_F)^k - k \mu_{k-1} x] d(G - F)(x) \\ &= \int [(x - \mu_F)^k - k \mu_{k-1} x] dG(x) - (\mu_k - k \mu_{k-1} \mu_F). \end{aligned}$$

Setting $h(F; x) = (x - \mu_F)^k - k \mu_{k-1} x - (\mu_k - k \mu_{k-1} \mu_F)$ and using $\mu(T; F) = \int h(F; x) dF(x) = 0$, we get from Theorem 6.4

$$\sqrt{N}(T(\hat{F}_N) - T(F)) = \sqrt{N}(\hat{\mu}_k - \mu_k) \xrightarrow[\mathcal{L}]{\mathcal{L}} \mathcal{N}(0, \sigma^2(T, F)) \quad \text{with}$$

$$\sigma^2(T, F) = \int h^2(F; x) dF(x) = \mu_{2k} - \mu_k^2 - 2k \mu_{k+1} \mu_{k-1} + 2k \mu_k \mu_{k-1} \mu_F + k^2 \mu_{k-1}^2 \mu_2.$$

Of course, we have to check that the remainder term

$$R_{1,N} = \hat{\mu}_k - \frac{1}{N} \sum_{j=1}^N (X_j - \mu_F)^k + k \mu_{k-1} (\bar{X}_N - \mu_F)$$

satisfies $\sqrt{N}R_{1,N} \xrightarrow{\mathbb{P}} 0$.

B: Maximum Likelihood

Consider the model where $\mathcal{L}(X_j) \in \{P_\theta, \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}$. The ML-estimate $\hat{\theta}_N = T(\hat{F}_N)$ w.r.t. this model solves

$$\int \psi(\hat{\theta}_N|y) d\hat{F}_N(y) = \frac{1}{N} \sum_{j=1}^N \psi(\hat{\theta}_N|X_j) = 0$$

with score function $\psi(\theta|y) = l'(\theta|y)$.

Correspondingly, $T(F) = \theta_F$ can be defined as solution of

$$\int \psi(\theta_F|y) dF(y) = 0.$$

We calculate $T'(F; G - F) = \frac{d}{d\lambda}$, where $g(\lambda)|_{\lambda=0}$, $g(\lambda) = T(F_\lambda) = \theta_{F_\lambda}$, $F_\lambda = F + \lambda(G - F)$, by implicit differentiation of the equation

$$H(\theta_{F_\lambda}, \lambda) = 0 \quad \text{with } H(\theta, \lambda) = \int \psi(\theta|y) dF_\lambda(y).$$

We get

$$0 = \frac{d}{d\lambda} H(\theta_{F_\lambda}, \lambda) \Big|_{\lambda=0} = \left(\frac{\partial}{\partial \theta} H \right) (\theta_F, 0) \frac{d}{d\lambda} \theta_{F_\lambda} \Big|_{\lambda=0} + \left(\frac{\partial}{\partial \lambda} H \right) (\theta_F, 0)$$

and therefore

$$\begin{aligned} \frac{d}{d\lambda} g(\lambda) \Big|_{\lambda=0} &= \frac{d\theta_{F_\lambda}}{d\lambda} \Big|_{\lambda=0} = - \frac{\partial H}{\partial \lambda} \Big|_{\lambda=0} / \frac{\partial H}{\partial \theta} \Big|_{\lambda=0} \\ \frac{\partial H}{\partial \lambda} \Big|_{\lambda=0} &= \int \psi(\theta_F|y) d(G - F)(y), \quad - \frac{\partial H}{\partial \theta} \Big|_{\lambda=0} = - \int \psi'(\theta_F|y) dF(y) = I(P_{\theta_F}). \end{aligned}$$

where the latter identity holds if the model is correct.

Remark that, by Lemma 3.1, $\int \psi(\theta|y) dF(y) = \mathbb{E} \psi(\theta|X_1) = 0$.

We get

$$T'(F; \hat{F}_N - F) = \frac{1}{I(P_{\theta_F})} \frac{1}{N} \sum_{j=1}^N \psi(\theta_F|X_j) = V_{1,N} \quad \text{with } h(F, x) = \frac{\psi(\theta_F|x)}{I(P_{\theta_F})}.$$

Checking (as in chapter 3), that $\sqrt{N} R_{1,N} = \sqrt{N}(T(\hat{F}_N) - T(F) - V_{1,N}) \xrightarrow{p} 0$, we get from

Theorem 6.4

$$\sqrt{N} (T(\hat{F}_N) - T(F)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(T, F)) \quad \text{where}$$

$$\sigma^2(T, F) = \mathbb{E} h^2(F; X_1) = \frac{1}{I^2(P_{\theta_F})} \mathbb{E} \psi^2(\theta_F|X_j) = \frac{1}{I(P_{\theta_F})}.$$

If the model is misspecified, i.e. F is not the distribution function of P_{θ_F} , we get the same result, but

$$\sigma^2(T, F) = \frac{\int \psi^2(\theta_F|y) dF(y)}{(\int \psi'(\theta_F|y) dF(y))^2}.$$

Numerator and denominator cannot be reduced to the Fisher information, as, then, ψ is the score function w.r.t. P_{θ_F} , not w.r.t. F .

C: Sample quantiles

We are mainly interested in $T'(F; \hat{F}_N - F)$. We can write

$$\hat{F}_N(x) = \frac{1}{N} \sum_{j=1}^N 1_{(-\infty, x]}(X_j) = \frac{1}{N} \sum_{j=1}^N 1_{[X_j, \infty)}(x).$$

As T' is linear in $\hat{F}_N - F = \frac{1}{N} \sum_{j=1}^N (1_{[X_j, \infty)}(x) - F)$, we can restrict the calculation of $T'(F, \cdot)$ to

$$F_\lambda = F + \lambda(\Delta_{x_0} - F)$$

where $\Delta_{x_0}(x) = 1_{[x_0, \infty)}(x)$ is the distribution function of a point mass in x_0 .

For $0 < \alpha < 1$, let $q_\alpha = T(F) = F^{-1}(\alpha)$ be the α -quantile of $\mathcal{L}(X_j)$, i.e., if F has a density,

$$F(q_\alpha) = \text{pr}(X_j \leq q_\alpha) = \alpha.$$

Otherwise, define F^{-1} more generally by

$$q_\alpha = F^{-1}(\alpha) = \inf\{x; F(x) \geq \alpha\}$$

which implies

$$\text{pr}(X_j < q_\alpha) \leq \alpha \leq \text{pr}(X_j \leq q_\alpha).$$

We have

$$\begin{aligned} T(F_\lambda) &= \inf\{x; F(x) + \lambda(\Delta_{x_0}(x) - F(x)) \geq \alpha\} \\ &= \inf\{x; F(x) \geq \frac{\alpha - \lambda 1_{[x_0, \infty)}(x)}{1 - \lambda}\} \end{aligned}$$

The **sample quantile** $\hat{q}_{\alpha N}$ is defined by (using right-continuity of \hat{F}_N)

$$\begin{aligned} \hat{q}_{\alpha N} &= \hat{F}_N^{-1}(\alpha) = \inf\{X_j; \hat{F}_N(X_j) \geq \alpha\} = \inf\{X_{(k)}; \hat{F}_N(X_{(k)}) \geq \alpha\} \\ &\stackrel{(*)}{=} \inf\{X_{(k)}; \frac{k}{N} \geq \alpha\} = X_{(\{N\alpha\})} \end{aligned}$$

where $X_{(1)} \leq \dots \leq X_{(N)}$ order statistics and $\{x\} = \min\{k \leq N, k \geq x\}$. Special case: $\alpha = 0.5 \implies \hat{q}_{\alpha, N} = \hat{X}_N$ sample median.

Ass: There is a neighbourhood $(q_\alpha - \delta, q_\alpha + \delta)$, $\delta > 0$, of q_α such that:

$$F'(x) = p(x) \text{ exists and is positive in } (q_\alpha - \delta, q_\alpha + \delta).$$

For λ small enough, $T(F_\lambda) \in (q_\alpha - \delta, q_\alpha + \delta)$ as $T(F) = q_\alpha$.

Case I: $x_0 > q_\alpha + \delta$, λ small enough

$$\begin{aligned} g(\lambda) = T(F_\lambda) &= \inf\left\{x \in (q_\alpha - \delta, q_\alpha + \delta); F(x) \geq \frac{\alpha - \lambda 1_{[x_0, \infty)}(x)}{1 - \lambda}\right\} \\ &= F^{-1}\left(\frac{\alpha}{1 - \lambda}\right) \\ \implies g'(\lambda) &= \frac{1}{p(F^{-1}(\frac{\alpha}{1 - \lambda}))} \cdot \frac{\alpha}{(1 - \lambda)^2} \implies g'(0) = \frac{\alpha}{p(q_\alpha)} \end{aligned}$$

Case II: $x_0 < q_\alpha + \delta$, λ small enough

$$\begin{aligned} g(\lambda) &= T(F_\lambda) = F^{-1}\left(\frac{\alpha - \lambda}{1 - \lambda}\right) \\ g'(\lambda) &= \frac{1}{p(F^{-1}(\frac{\alpha - \lambda}{1 - \lambda}))} \cdot \frac{\alpha - 1}{(1 - \lambda)^2} \implies g'(0) = \frac{\alpha - 1}{p(q_\alpha)}. \end{aligned}$$

For $\delta \rightarrow 0$, we get

$$T'(F; \Delta_{x_0} - F) = g'(0) = \frac{\alpha - 1_{[x_0, \infty)}(q_\alpha)}{p(q_\alpha)} \quad (3)$$

and, then,

$$T'(F; \hat{F}_N - F) = \frac{\alpha - \hat{F}_N(q_\alpha)}{p(q_\alpha)} = \frac{1}{N} \sum_{j=1}^N h(F; X_j) \quad \text{with } h(F; x) = \frac{\alpha - 1_{[x, \infty)}(q_\alpha)}{p(q_\alpha)}$$

We have, as $1_{[X_1, \infty)}(q_\alpha) = 1_{(-\infty, q_\alpha]}(X_1)$ is a Bernoulli variable,

$$\begin{aligned} \mathbb{E} h(F; X_1) &= \frac{\alpha - \text{pr}(X_1 \leq q_\alpha)}{p(q_\alpha)} = 0 \\ \text{var } h(F; X_1) &= \frac{1}{p^2(q_\alpha)} \text{var}(1_{(-\infty, q_\alpha]}(X_1)) = \frac{q_\alpha(1 - q_\alpha)}{p^2(q_\alpha)} = \sigma^2(T, F). \end{aligned}$$

By Theorem 6.4, we get

$$\sqrt{N}(\hat{q}_{\alpha N} - q_\alpha) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{q_\alpha(1 - q_\alpha)}{p^2(q_\alpha)}\right),$$

provided that we can show $\sqrt{N}R_{1N} \xrightarrow{p} 0$ with

$$R_{1N} = \hat{q}_{\alpha N} - q_\alpha - \frac{\alpha - \hat{F}_N(q_\alpha)}{p(q_\alpha)}.$$

For that purpose, we may use the Bahadur representation for sample quantiles.

Theorem 6.7 (Bahadur, 1966). *If F is twice continuously differentiable at q_α with $F'(q_\alpha) = p(q_\alpha) > 0$, then*

$$\hat{q}_{\alpha N} = q_\alpha + \frac{\alpha - \hat{F}_N(q_\alpha)}{p(q_\alpha)} + R_N$$

with $R_N = O\left(\left(\frac{\log N}{N}\right)^{3/4}\right)$ a.s..

The following weaker result under weaker assumptions suffices already.

Theorem 6.8 (Ghosh, 1971): *For $\sqrt{N}R_N \xrightarrow{p} 0$, it suffices that F is one times continuously differentiable at q_α with $F'(q_\alpha) > 0$.*

D) α -trimmed mean

Let F be symmetric around $\mu = \int x dF(x)$ with density p . Let $0 < \alpha < \frac{1}{2}$, and let $X_{(1)} \leq \dots \leq X_{(N)}$ be the order statistics, i.e. the ordered data X_1, \dots, X_N . To estimate μ , we eliminate a fraction α of the smallest and largest data and average over the rest, i.e.

$$\bar{X}_N^\alpha = \frac{1}{N - 2[\alpha N]} \sum_{j=[\alpha N]+1}^{N-[\alpha N]} X_{(j)} \quad [x] = \max\{k \in \mathbb{Z}, k \leq x\}$$

limiting cases: $\alpha \rightarrow 0$: \bar{X}_N , and $\alpha \rightarrow \frac{1}{2}$: \dot{X}_N (sample median)

Goal: Show that $\sqrt{N}(\bar{X}_N^\alpha - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \dots)$

Consider

$$T(F) = \frac{1}{1 - 2\alpha} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x dF(x) = \frac{1}{1 - 2\alpha} \int_{\alpha}^{1-\alpha} F^{-1}(u) du \quad \text{by substitution.}$$

Recall sample quantiles of previous example:

$$\hat{F}_N^{-1}(\alpha) = X_{(\{\alpha N\})}$$

If αN integer $\implies (1-\alpha)N$ integer and $\hat{F}_N^{-1}(\alpha) = X_{[\alpha N]}$, $\hat{F}_N^{-1}(1-\alpha) = X_{[(1-\alpha)N]} = X_{N-[\alpha N]}$.
If $\alpha N \notin \mathbb{Z} \implies [\alpha N] < \alpha N < \{\alpha N\} = [\alpha N] + 1$, $\{(1-\alpha)N\} = N - [\alpha N]$.

Therefore, if $\alpha N \notin \mathbb{Z}$

$$\bar{X}_N^\alpha = \frac{N - 2\alpha N}{N - 2[\alpha N]} \frac{1}{1 - 2\alpha} \int_{\hat{F}_N^{-1}(\alpha)}^{\hat{F}_N^{-1}(1-\alpha)} x d\hat{F}_N = \frac{1 - 2\alpha}{1 - 2\frac{[\alpha N]}{N}} T(\hat{F}_N)$$

if $\int_a^b \equiv \int_{[a,b]}$ includes the boundary points.

Otherwise, if $\alpha N \in \mathbb{Z}$

$$\begin{aligned} \bar{X}_N^\alpha &= \frac{N - 2\alpha N}{N - 2[\alpha N]} \frac{1}{1 - 2\alpha} \frac{1}{N} \left\{ \sum_{j=[\alpha N]}^{N-[\alpha N]} X_{(j)} - X_{(\alpha N)} \right\} \\ &= \frac{1 - 2\alpha}{1 - 2\frac{[\alpha N]}{N}} \left(T(\hat{F}_N) - \frac{1}{1 - 2\alpha} \frac{X_{(\alpha N)}}{N} \right) \end{aligned}$$

We conclude in both cases:

$$\bar{X}_N^\alpha - T(\hat{F}_N) = O_p\left(\frac{1}{N}\right),$$

i.e. it suffices to show

$$\sqrt{N}(T(\hat{F}_N) - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \dots)$$

As F has a density p , symmetric around $\mu \implies F^{-1}(\alpha) = \mu - \delta$, $F^{-1}(1-\alpha) = \mu + \delta$.

Define $p_\alpha(x) = \frac{1}{1-2\alpha} p(x) \cdot 1_{[\mu-\delta, \mu+\delta]}(x)$; this is again a probability density symmetric around μ which implies

$$\mu = \int x p_\alpha(x) dx = \frac{1}{1 - 2\alpha} \int_{\mu-\delta}^{\mu+\delta} x p(x) dx = T(F).$$

So, we can try to apply Theorem 6.4. As in the previous example, consider $F_\lambda = F + \lambda(\Delta_{x_0} - F)$. As

$$T(F_\lambda) = \frac{1}{1-2\alpha} \int_\alpha^{1-\alpha} F_\lambda^{-1}(u) du$$

we get, using (3)

$$\begin{aligned} \frac{d}{d\lambda} T(F_\lambda)|_{\lambda=0} &= \frac{1}{1-2\alpha} \int_\alpha^{1-\alpha} \frac{d}{d\lambda} F_\lambda^{-1}(u)|_{\lambda=0} du \\ &= \frac{1}{1-2\alpha} \int_\alpha^{1-\alpha} \frac{u - \mathbf{1}_{[x_0, \infty)}(F^{-1}(u))}{p(F^{-1}(u))} du. \end{aligned}$$

Using $\frac{d}{du} F^{-1}(u) = \frac{1}{p(F^{-1}(u))}$, integration by parts and distinguishing between the cases $F(x_0) < \alpha$, $\alpha \leq F(x_0) \leq 1 - \alpha$ and $F(x_0) > 1 - \alpha$, we get

$$T'(F; \Delta_{x_0} - F) = h(F, x_0) = \frac{1}{1-2\alpha} \begin{cases} F^{-1}(\alpha) - \mu & x_0 < F^{-1}(\alpha) \\ x_0 - \mu & \text{else} \\ F^{-1}(1-\alpha) - \mu & x_0 > F^{-1}(1-\alpha) \end{cases}$$

and from Theorem 6.4

$$\sqrt{N}(\bar{X}_N^\alpha - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(T, F)),$$

$$\sigma^2(T, F) = \mathbb{E} h^2(F; X_j) = \frac{1}{(1-2\alpha)^2} \left[\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} (x - \mu)^2 dF(x) + 2\alpha(F^{-1}(\alpha) - \mu)^2 \right]$$

6.3 Robustness

Consider again i.i.d. data X_1, \dots, X_N , the parametric model

$$\mathcal{M}_\Theta : \mathcal{L}(X_j) \in \mathcal{P}_\Theta = \{P_\theta, \theta \in \Theta\}.$$

We have derived estimates for θ and tests about θ which have certain optimality properties, at least asymptotically for $N \rightarrow \infty$. But: Models are only an approximation to reality. So what happens, e.g., if we assume \mathcal{M}_Θ to hold and use the ML-estimate for some function $b(\theta)$, e.g.

$$b(\theta) = \int x dP_\theta(x) = \mathbb{E}_\theta X_j,$$

in a situation where $\mathcal{L}(X_j) \notin \mathcal{P}_\Theta$? We have seen that such estimates may still be asymptotically normal (as QML-estimates), but no longer asymptotically efficient. How bad can they be, i.e. how large can the asymptotic variance of a QML-estimate be?

Robust statistics is dealing with that issue of model misspecification. An estimate or test is considered to be robust if it works reasonably well even if the model is not 100% correct.

A special case is robustness against outliers, i.e. in the simple case of i.i.d. data against a few observations which are considerably apart from the majority of data.

Example: ε -contamination neighbourhood $\mathcal{P}_{\Theta, \varepsilon}$ or **gross error model**

Let $\mathcal{P}_\Theta = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}\}$ $\theta = \mu, \sigma^2$ given, be the Gaussian model with fixed variance. Let \mathcal{F}_{ac} be the class of all distribution functions on (R) having a density. Then,

$$\mathcal{P}_{\Theta, \varepsilon} = \{F, F = (1 - \varepsilon)\Phi_{\mu, \sigma^2} + \varepsilon\tilde{F}; \int x dF = \mu, \tilde{F} \in \mathcal{F}_{ac}\}$$

is the ε -contamination neighbourhood of \mathcal{P}_{Θ} . The real data are not all normal, but only the good $(1 - \varepsilon)$ fraction of them. The rest may be outliers, i.e. lying much farther away from μ than $\mathcal{N}(\mu, \sigma^2)$ -data, depending on the choice of \tilde{F} . A typical example would be $\tilde{F} = \Phi_{\mu, 9\sigma^2}$, i.e. a fraction ε of the data is much more unreliable having a much larger variance.

So, we are looking for robustness concepts allowing to characterize estimates or tests which are robust against certain forms of model misspecification. Robustness has something to do with continuity of functionals. Let, e.g., T_N be a sequence of estimates, $\mathcal{L}_F(T_N)$ be the law of T_N provided that F is the distribution function of the data X_j . Then, if $F \approx G$ we want to have $\mathcal{L}_F(T_N) \approx \mathcal{L}_G(T_N)$.

We have to specify what "≈" means. For that, we introduce a metric on the set of distribution functions or, more generally, probability measures.

Definition: a) F, G distribution functions on \mathbb{R} .

Lévy distance:

$$d_L(F, G) = \inf \{ \varepsilon > 0, F(x - \varepsilon) - \varepsilon \leq G(x) \leq F(x + \varepsilon) + \varepsilon \text{ for all } x \}$$

Kolmogorov distance:

$$d_K(F, G) = \sup_x |F(x) - G(x)|$$

b) P, Q probability measures on \mathbb{R}^d .

Prohorov distance:

$$d_P(P, Q) = \inf \{ \varepsilon > 0, P(A) \leq Q(A^\varepsilon) + \varepsilon \text{ for all } A \in \mathcal{L} \},$$

where $A^\varepsilon = \{x; \inf_{y \in A} \|x - y\| \leq \varepsilon\}$ denotes the ε -neighbourhood of the Borel set A .

total variation distance:

$$d_{TV}(P, Q) = \sup_A |P(A) - Q(A)|$$

Lemma 6.1 a) d_L, d_P are metrics

b) d_L, d_P metrize the weak (w^*) topology, i.e. the weakest topology on the set of distributions \mathcal{F} for which all the following mappings are continuous

$$F \longmapsto \int_{-\infty}^{\infty} \psi(x) dF(x) \text{ resp. } P \longmapsto \int_{\mathbb{R}^d} \psi(x) P(dx)$$

with $\psi : \mathbb{R} \longrightarrow \mathbb{R}$ resp. $\psi : \mathbb{R}^d \longrightarrow \mathbb{R}$ bounded and continuous.

Lemma 6.2 a) $d_L \leq d_P \leq d_{TV}$, $d_L \leq d_K \leq d_{TV}$

b) d_K, d_{TV} do not metrize the weak topology.

Definition: A sequence T_N , $N \geq 1$, of estimates or test statistics is called qualitatively robust at $F = F_0$ if the sequence of maps

$$F \longmapsto \mathcal{L}_F(T_N), \quad N \geq 1,$$

is equicontinuous at F_0 , i.e. for some metric d_* metrizing the weak topology (e.g. $* = L$ or P) we have: for all $\varepsilon > 0$ there are $\delta_\varepsilon, N_\varepsilon$ such that for all F and $N \geq N_\varepsilon$

$$d_*(F_0, F) \leq \delta_\varepsilon \implies d_*(\mathcal{L}_{F_0}(T_N), \mathcal{L}_F(T_N)) \leq \varepsilon. \quad (4)$$

A rather deep result of Hampel states that for statistics of the form $T_N = T(\hat{F}_N)$ robustness is essentially equivalent to continuity of T .

Let $X_1, \dots, X_N \in \mathbb{R}$ i.i.d. with distribution function F , $T_N = T(\hat{F}_N) \in \mathbb{R}^d$. In (4), we choose $d_L(F_0, F)$ and $d_P(\mathcal{L}_{F_0}(T_N), \mathcal{L}_F(T_N))$

Proposition 6.1 *If T is weakly continuous at F (i.e. $d_L(F_N, F) \longrightarrow 0 \implies T(F_N) \xrightarrow{p} T(F)$*

for $N \rightarrow \infty$), then $T_N = T(\hat{F}_N)$ is consistent at F , i.e. $T_N = T(\hat{F}_N) \xrightarrow{p} T(F)$ for $N \rightarrow \infty$.

Proof: By Lemma 6.2 and the Glivenko-Cantelli theorem

$$d_L(\hat{F}_N, F) \leq d_K(\hat{F}_N, F) = \sup_x |\hat{F}_N(x) - F(x)| \longrightarrow 0$$

which implies $T(\hat{F}_N) \xrightarrow{p} T(F)$ by Lemma 6.1. ■

Theorem 6.9 (*Hampel*): *Assume that $T_N = T(\hat{F}_N)$ is consistent at F for all F in a Lévy neighbourhood $\{F; d_L(F, F_0) < \eta\}$ of F_0 . Then, T is continuous at F_0 iff T_N is qualitatively robust at F_0 . (compare Theorem 2.6.2 of Huber)*

As consistency in basic estimation problems, qualitative robustness is a requirement which is satisfied by many estimators and which is not suitable for comparing the robustness properties of estimators directly. We need concepts of quantitative robustness.

What do we want from good robust statistics? Let us consider data X_1, \dots, X_N i.i.d. with distribution function F . We also write $F = \mathcal{L}(X_j)$ for the distribution itself. We have an ideal model

$$\mathcal{M}_\Theta : F \in \{P_\theta; \theta \in \Theta\} = \mathcal{P}_\Theta, \text{ e.g. } F = \mathcal{N}(\mu, \sigma^2) \text{ for some } \mu \in \mathbb{R}, \sigma^2 > 0.$$

However, we do not expect the model to hold perfectly, e.g. we assume that the data have a distribution function of the form

$$F(x) = (1 - \varepsilon)F_0(x) + \varepsilon G(x), \quad F_0 \in \mathcal{P}_\Theta, \quad G \in \mathcal{F}$$

for some fixed $\varepsilon > 0$. A good robust statistic should

- (1) have a good (but not necessarily optimal) efficiency if the model \mathcal{M}_Θ really holds, i.e. if $G = F_0$ and, then, $F = F_0 \in \mathcal{P}_\Theta$
- (2) have a performance which is only changing slightly due to small deviations from the model, e.g. for $F \approx F_0$ for some $F_0 \in \mathcal{P}_\Theta$, the *mse* of an estimate or the power of a test should not change too much. Small deviations can be caused by a few outliers or by many tiny errors (e.g. due to rounding).

(3) be only moderately sensitive to large deviations from the model, i.e. should not lead to completely wrong conclusions.

(1) is quantified by asymptotic relative efficiency (ARE) relative to the asymptotically optimal statistic, e.g. the ML-estimate, in model \mathcal{M}_Θ .

(2) Here, the main quantitative tool is the influence curve. It is motivated by the following consideration. We have a statistic which can be written as $T_N = T(\hat{F}_N)$. We have a sample X_1, \dots, X_N of "good" data, and we add an additional observation $X_{N+1} = x$. How strongly can this one observation change the estimate if x is chosen particularly badly, i.e. we consider

$$T_{N+1} - T_N = T\left(\frac{N}{N+1}\hat{F}_N + \frac{1}{N+1}\Delta_x\right) - T(\hat{F}_N), \quad \Delta_x = 1_{[x, \infty)}.$$

We let $N \rightarrow \infty$. Then, $\hat{F}_N \rightarrow F$. To get a nontrivial limit, we have to divide by $\frac{1}{N+1}$, and we have

$$\begin{aligned} \lim_{N \rightarrow \infty} (N+1)(T_{N+1} - T_N) &= \lim_{N \rightarrow \infty} \frac{T(\hat{F}_N + \frac{1}{N+1}(\Delta_x - \hat{F}_N)) - T(\hat{F}_N)}{\frac{1}{N+1}} \\ &= \lim_{\lambda \rightarrow 0} \frac{T(F + \lambda(\Delta_x - F)) - T(F)}{\lambda} \\ &= T'(F; \Delta_x - F) =: h(F, x). \end{aligned}$$

Definition: If T is Gateaux-differentiable at F in direction $\Delta_x - F$ for all x , the influence curve of the estimator $T(\hat{F}_N)$ is given as

$$IC(x; T, F) = h(F, x) = T'(F; \Delta_x - F)$$

Definition: a) gross error sensitivity $\gamma^* = \sup_x |IC(x; T, F)|$

b) local shift sensitivity $\lambda^* = \sup_{x \neq y} \left| \frac{IC(x; T, F) - IC(y; T, F)}{x - y} \right|$

a) measures influence of outliers, b) measures influence of local changes to the data (e.g. rounding). A statistic which should be robust against outliers should, e.g., have $\gamma^* < \infty$, i.e. a bounded influence curve.

Examples

a) sample mean $T_N = \bar{X}_N$, $T(F) = \int y dF(y)$

$$IC(x; T, F) = T'(F, \Delta_x - F) = T(\Delta_x) - T(F) = x - \mu.$$

The influence curve is unbounded, $\gamma^* = +\infty$, and \bar{X}_N is not robust.

b) α -trimmed mean $T_N = \bar{X}_N^\alpha$, $T(F) = \frac{1}{1-2\alpha} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x dF(x)$

$$IC(x; T, F) = \frac{1}{1-2\alpha} \begin{cases} F^{-1}(\alpha) - \mu & x < F^{-1}(\alpha) \\ x - \mu & \text{else} \\ (F^{-1}(1-\alpha) - \mu) & x > (F^{-1}(1-\alpha)) \end{cases}.$$

For symmetric F , we have $|F^{-1}(\alpha) - \mu| = |F^{-1}(1-\alpha) - \mu|$ and, therefore, $\gamma^* = F^{-1}(1-\alpha) - \mu < \infty$, so \bar{X}_N^α is (infinitesimally) robust.

The robustness concept based on the influence curve is infinitesimal in the sense that the fraction $\frac{1}{N+1}$ of bad data $\rightarrow 0$ for $N \rightarrow \infty$. What about requirement (3)? Assume that we have a model distribution F_0 , and the distribution F of the data lies in some neighbourhood \mathcal{F}_ε of F_0 , e.g.

$$\begin{aligned} \text{Lévy neighbourhood} & \quad \mathcal{F}_\varepsilon = \{F, d_L(F, F_0) < \varepsilon\} \\ \text{gross error neighbourhood} & \quad \mathcal{F}_\varepsilon = \{F, F = (1 - \varepsilon)F_0 + \varepsilon G, G \in \mathcal{F}\}. \end{aligned}$$

We consider maximum bias and variance

$$\begin{aligned} b_1(\varepsilon) &= \sup_{F \in \mathcal{F}_\varepsilon} |T(F) - T(F_0)| \\ v_1(\varepsilon) &= \sup_{F \in \mathcal{F}_\varepsilon} \sigma^2(T, F) \end{aligned}$$

provided $\sqrt{N}(T(\hat{F}_N) - T(F)) \rightarrow \mathcal{N}(0, \sigma^2(T, F))$.

We would like to discuss stronger robustness measures, e.g.

$$b(\varepsilon) = \lim_{N \rightarrow \infty} \sup_{F \in \mathcal{F}_\varepsilon} |\text{median}\{\mathcal{L}_F(T(\hat{F}_N) - T(F_0))\}|,$$

but that is difficult to handle. Fortunately, for many situations, we have $b(\varepsilon) = b_1(\varepsilon)$ (and similar for the variance measure) but this has to be checked in each case (compare section 1.4 of Huber).

Let us first consider the bias. If $\varepsilon = 1$, we have $\mathcal{F}_1 = \mathcal{F}$, so $b(1)$ is the worst what can happen.

Definition: The asymptotic breakdown point of T at F_0 is

$$\varepsilon^* = \varepsilon^*(T, F_0) = \sup\{\varepsilon; b(\varepsilon) < b(1)\}$$

ε^* is the largest fraction of outliers which an estimate can handle before giving completely false results. It is the limit of the finite sample breakdown point ε_N^* defined in the following manner:

Let T be an estimate of a location parameter $\mu = T(F_0)$. Given X_1, \dots, X_N , we add k data at arbitrary locations y_1, \dots, y_k . Let $\Delta^{(k)}(x) = \frac{1}{k} \sum_{i=1}^k 1_{[y_i; \infty)}(x)$ be the corresponding empirical distribution function. Then, $\frac{N}{N+k} \hat{F}_N(x) + \frac{k}{N+k} \Delta^{(k)}(x)$ is the empirical distribution function of the augmented sample $X_1, \dots, X_N, y_1, \dots, y_k$ of size $n + k$. Consider

$$\sup_{y_1, \dots, y_k} \left| T\left(\frac{N}{N+k} \hat{F}_N + \frac{k}{N+k} \Delta^{(k)}\right) - \mu \right| = \tau_k,$$

and define

$$k^* = \sup\{k; \tau_k < \infty\}, \quad \varepsilon_N^* = \frac{k^*}{N + k^*}$$

As $\hat{F}_N \rightarrow F$, we have $\varepsilon_N^* \rightarrow \varepsilon^*$ ($N \rightarrow \infty$).

Under regularity assumptions, we have for $F = (1 - \varepsilon)F_0 + \varepsilon G$:

$$T(F) - T(F_0) \approx \varepsilon \int IC(x, F_0, T) dG(x)$$

which implies $b_1(\varepsilon) \approx \varepsilon \gamma^*$ (ε small), usually. However, there are counterexamples (Huber).

Examples:

- a) sample mean \bar{X}_N , $\varepsilon^* = 0$.
- b) α -trimmed mean \bar{X}_N^α , $\varepsilon^* = \alpha$.
- c) sample median \dot{X}_N , $\varepsilon^* = \frac{1}{2}$.

Minimax theory for location estimates

In this section we consider the following location parameter problem:

Model: X_1, \dots, X_N i.i.d. with distribution function $F(x - \theta)$, $\theta \in \mathbb{R}$, $F \in \mathcal{F}_\varepsilon =$ gross error model around G , i.e. $F = (1 - \varepsilon)G + \varepsilon\tilde{F}$, $\tilde{F} \in \mathcal{F}$, G given .

Notation: $F_\theta, G_\theta =$ distributions corresponding to the distribution functions $F(x - \theta)$ resp. $G(x - \theta)$.

Example: $G = \mathcal{N}(0, 1) \implies G_\theta = \mathcal{N}(\theta, 1)$

Goal: Estimate location parameter θ !

The model is semiparametric, as $\mathcal{L}(X_j) = (1 - \varepsilon)G_\theta + \varepsilon\tilde{F}_\theta = (1 - \varepsilon)G_\theta + \varepsilon\bar{F}$, $\bar{F} \in \mathcal{F}$, i.e. a major model component (G_θ) is parametric, but another one (\bar{F}) is nonparametric.

Assumptions: Let $T_N = T_N(X_1, \dots, X_N) = T(\hat{F}_N)$ be an estimate of θ .

- a) T_N is translation equivariant, i.e. $T_N(X_1 + \mu, \dots, X_N + \mu) = T_N(X_1, \dots, X_N) + \mu$ for all X_j .
- b) T_N is asymptotically unbiased at G , i.e. for $\mathcal{L}(X_j) = G_\theta$

$$\lim_{N \rightarrow \infty} \mathbb{E} T(\hat{F}_N) = T(G_\theta) = \theta \quad \text{for all } \theta.$$

Together with a), it suffices for b) that $T(G) = 0$.

Denote by \mathcal{T} the class of all estimates T satisfying these assumptions.

Minimax-approach (now semiparametric): Choose T such that maximal risk (mse) over all possible F is minimized over \mathcal{T} .

We consider the mse-components bias and variance separately.

$$\text{maximal bias } b_1(\varepsilon) = \sup_{F \in \mathcal{F}_\varepsilon} |T(F) - \underbrace{T(G)}_{=0}|.$$

By a) we do not have to take the supremum over θ , too, as the bias does not depend on θ : $T(F_\theta) - T(G_\theta) = T(F) - T(G)$.

Let $M(F) = \text{med}(F)$ be the median of F .

Theorem 6.10 *If G has a symmetric, unimodal density (i.e. only 1 local maximum at 0), then the median M is the minimax bias estimate, i.e.*

$$\sup_{F \in \mathcal{F}_\varepsilon} |M(F)| = \min_{T \in \mathcal{T}} \sup_{F \in \mathcal{F}_\varepsilon} |T(F)|.$$

For the median, we always have $b_1(\varepsilon) = b(\varepsilon)$. Also the result holds for other neighbourhoods \mathcal{F}_ε , e.g. for Lévy neighbourhoods.

However, usually we have ε, N not too large, and in that case $\text{var} \gg \text{bias}$! So, the theorem covers extreme situations where ε large ($\gtrsim 0.25$) or N very large ($\gtrsim 1000 - 10000$). The median is the estimate of choice in such situations, but otherwise we have to look at minimax variance estimates.

Assume now $\sqrt{N}(T(\hat{F}_N) - T(F)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(T, F))$ for $F \in \mathcal{F}_\varepsilon$.

An estimate $R \in \mathcal{T}$ is called minimax robust (w.r.t. asymptotic variance) if

$$\sup_{F \in \mathcal{F}_\varepsilon} \sigma^2(R, F) = \min_{T \in \mathcal{T}} \sup_{F \in \mathcal{F}_\varepsilon} \sigma^2(T, F).$$

Heuristics (does not work in every respect, but it provides some intuition): For sake of simplicity, consider only unbiased estimates. Then, if T is any estimate and T^F is the ML-estimate w.r.t. F , we have from the Cramér-Rao inequality (Cor. 3.1) and the asymptotics of ML-estimates (Th. 3.7) with I denoting Fisher information:

$$\sigma^2(T, F) \geq \frac{1}{I(F)} = \sigma^2(T^F, F)$$

Step 1: Determine the most unfavourable distribution F_u in \mathcal{F}_ε , i.e.

$$\frac{1}{I(F_u)} = \max_{F \in \mathcal{F}_\varepsilon} \frac{1}{I(F)} = \max_{F \in \mathcal{F}_\varepsilon} \min_{T \in \mathcal{T}} \sigma^2(T, F)$$

Step 2: Choose as an estimate the ML-estimate $T^u \equiv T^{F_u}$ w.r.t. F_u .

Step 3: Prove

$$\max_{F \in \mathcal{F}_\varepsilon} \min_{T \in \mathcal{T}} \sigma^2(T, F) \geq \min_{T \in \mathcal{T}} \max_{F \in \mathcal{F}_\varepsilon} \sigma^2(T, F) \quad (5)$$

As the other inequality, i.e. "≤" in (5), is always fulfilled, we get from step 3

$$\max_{F \in \mathcal{F}_\varepsilon} \min_{T \in \mathcal{T}} \sigma^2(T, F) = \min_{T \in \mathcal{T}} \max_{F \in \mathcal{F}_\varepsilon} \sigma^2(T, F)$$

and, therefore,

$$\sigma^2(T^u, F) \leq \sigma^2(T^u, F_u) \leq \sigma^2(T, F_u) \quad \text{for all } T \in \mathcal{T}, F \in \mathcal{F}_\varepsilon.$$

Therefore, (T^u, F_u) is a saddlepoint of $\sigma^2(T, F)$, and T^u is the desired minimax robust estimate.

The main task is step 1, i.e. we have to solve

$$I(F_u) = \min_F I(F) \quad \text{under the constraint } F \in \mathcal{F}_\varepsilon.$$

This can be solved by a Lagrange multiplier technique. Afterwards, step 3 has to be checked, and some technical details have to be looked at.

Example:

Let G have a twice continuously differentiable symmetric density g with a convex support for which $-\log g(x)$ is convex (e.g. $G = \mathcal{N}(0, 1)$). Then, it can be shown that the most unfavourable distribution F_u has a density p_u given by

$$p_u(x) = \begin{cases} (1 - \varepsilon)g(x) & |x| \leq b \\ ce^{-\lambda|x|} & |x| > b \end{cases}$$

for some constants $b, c, \lambda > 0$ chosen such that p_u is continuously differentiable in $x = \pm b$ and $\int p_u(x)dx = 1$.

The ML-estimate w.r.t the most unfavourable distribution is a M-estimate for the location parameter θ

$$\hat{\theta}_N = \arg_{\theta} \max Q_N(X_1, \dots, X_N; \theta)$$

with

$$Q_N(x_1, \dots, x_N; \theta) = \sum_{j=1}^N q(x_j - \theta)$$

and

$$q(y) = \log p_u(y).$$

Q_N is just the log likelihood w.r.t. F_u . Equivalently, we get $\hat{\theta}_N$ as solution of

$$\sum_{j=1}^N \psi(X_j - \theta) = 0$$

where $\psi(y - \theta) = \frac{\partial}{\partial \theta} q(y - \theta) = -q'(y - \theta)$ is the one-dimensional score function.

Now, let, in particular, $g(x) = \varphi(x) = \mathcal{N}(0, 1)$ -density, i.e. we consider a gross error neighbourhood of the Gaussian model. Then,

$$q(y) = \begin{cases} -\frac{1-\varepsilon}{2}x^2 + \text{const.} & |x| \leq b \\ -\lambda x + \text{const.} & x > b \\ \lambda x + \text{const.} & x < -b \end{cases}$$

or, redefining ψ by dividing it by $1 - \varepsilon$ which does not change the estimate,

$$\psi(y) = \begin{cases} x & |x| \leq b \\ \pm \frac{\lambda}{1-\varepsilon} & x \begin{cases} > b \\ < -b \end{cases} \end{cases} .$$

As the constants have been chosen such that p'_u and, therefore, ψ is continuous, we finally get

$$\psi(y) = \psi_H(y) = \begin{cases} x & |x| \leq b \\ \text{sign}(x)b & |x| > b \end{cases} .$$

The resulting location estimate $\hat{\theta}_N$ which solves

$$\sum_{j=1}^N \psi_H(y - \theta) = 0$$

or

$$\sum_{j=1}^N \rho_H(y - \theta) = \min_{\theta}$$

is called Huber's M -estimate. Here,

$$\rho_H(y) = \begin{cases} y^2 & |y| \leq b \\ b|y| & |y| > b \end{cases}$$

which shows the relationship to least-squares estimates (here just \bar{X}_N) which are the limiting case for $b \rightarrow \infty$. Huber's estimate reduces the influence of extreme outliers by considering not the squared but the absolute deviation for large $|x|$. Mark that the other limiting case $b \rightarrow 0$ leads to the sample median \hat{X}_N .

Theorem 6.11 *Let $G = \Phi$, and let T_H be Huber's M -estimate with b determined by*

$$\frac{2\varphi(b)}{b} - 2\Phi(-b) = \frac{\varepsilon}{1 - \varepsilon}.$$

Then, T_H is minimax robust (w.r.t asymptotic variance) for the neighbourhood

$$\mathcal{F}_\varepsilon^0 = \{F = (1 - \varepsilon)\Phi + \varepsilon \tilde{F}, \tilde{F} \in \mathcal{F}, T_H(\tilde{F}) = 0\}.$$

Restricting the neighbourhood from \mathcal{F}_ε to $\mathcal{F}_\varepsilon^0$ is somewhat against the spirit of robustness, but has to be done to avoid particular technical problem. Frequently, it is even assumed that \tilde{F} is symmetric around 0.

Proposition 6.2 *Let T_H denote Huber's estimate, $G = \Phi$.*

a) *The influence function of T_H is bounded:*

$$IC(x, T_H, \Phi) = \frac{\psi(x)}{\int \psi'(y) d\Phi(y)}$$

where $\int \psi'(y) d\Phi(y) = \int_{-b}^b d\Phi(y) = \Phi(b) - \Phi(-b) = 2\Phi(b) - 1$

b) *The asymptotic breakdown point of T_H is $\varepsilon^* = \frac{1}{2}$, i.e. the best possible value in our situation.*

Literature:

R. J. Serfling: *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

P. J. Huber: *Robust Statistics*. Wiley, New York, 1981.