Estimation of statistical functionals

(notes by Pavel Chigansky)

1. Statistical functionals

Let $X_1, \dots, X_n$ be an i.i.d. sample from an unknown distribution F on $\mathbb{R}^d$. Rather than estimating F itself, it is often required to estimate some *functional* of F, that is, a map from the space of distribution functions to $\mathbb{R}^k$. Some common examples of such statistical functionals are

$$\mu(F) = \int_{\mathbb{R}} x dF(x), \qquad \text{(mean)}$$

$$\sigma^2(F) = \int_{\mathbb{R}} (x-\mu(F))^2 dF(x), \qquad \text{(variance)}$$

$$q_p(F) = \inf\Big\{x \in \mathbb{R} : F(x) \ge p\Big\}, \qquad \text{(p-quantile[1]).}$$

Functionals of the form

$$T(F) = \int_{\mathbb{R}} \psi(x) dF(x), \tag{1.1}$$

where $\psi : \mathbb{R} \mapsto \mathbb{R}$ is a known function, are called *linear*. Thus the mean $\mu(F)$ and, more generally, higher moments are linear functionals. Nonlinear functionals can be formed by composition

$$T(F) = h\big(T_1(F), \dots, T_k(F)\big),$$

where $T_j(F)$ are linear functionals and $h : \mathbb{R}^k \mapsto \mathbb{R}$ is some nonlinearity. Variance $\sigma^2(F)$ is an example with $k = 2$, the two first moments $T_j(F) = \int x^j dF(x)$, $j = 1, 2$ and the function $h(x_1, x_2) = x_2 - x_1^2$. Many useful functionals are *implicit* being defined by solution to an equation such as

$$\int g(x,t) dF(x) = 0.$$

If F is continuous, the quantile $q_p(F)$ is such a functional with $g(x,t) = \mathbf{1}_{\{x \le t\}} - p$.

If a functional can be applied to purely discrete distributions, a reasonable estimator is obtained by replacing F with the empirical distribution $\widehat{F}_n$. It is easy

---

lecture notes for "Advanced Statistical Models B" course.

[1] $F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \ge p\}$ is a generalized inverse: when F is strictly increasing, this reduces to the usual inverse.

to see that for the examples above, the empirical mean, variance and quantiles are obtained this way (Problem 1). This chapter explores some of the properties of such *plug-in* or *substitution* estimators. For simplicity the case $d = k = 1$ will be considered.

2. Nonparametric Delta-method

2.1. A refresh on parametric Delta-method. Delta-method is a common technique of constructing consistent estimators and confidence sets for sufficiently regular statistical functionals. Let us first recall the standard parametric setup, when F belongs to a parametric family $\mathcal{F}_\Theta = (F_\theta)_{\theta\in\Theta}$ with $\Theta \subseteq \mathbb{R}^m$. In this case, any functional of F can be viewed as a function of θ,

$$q(\theta) := T(F_\theta), \quad \theta \in \Theta.$$

Typically, under mild regularity conditions, such as identifiability and smoothness in θ, standard estimation methods, e.g., maximization of likelihood or method of moments, etc. produce a sequence of estimators $\widehat{\theta}_n$ which is $\sqrt{n}$-consistent and asymptotically normal with the limit covariance matrix $\Sigma(\theta)$,

$$\sqrt{n}\big(\widehat{\theta}_n - \theta\big) \xrightarrow[n\to\infty]{d(\mathbb{P}_F)} N\big(0, \Sigma(\theta)\big). \tag{2.1}$$

If $q(\theta)$ has a continuous gradient $\nabla q(\theta)$, the plug-in estimator $q(\widehat{\theta}_n)$ is also $\sqrt{n}$-consistent and asymptotically normal for $q(\theta)$,

$$\sqrt{n}\big(q(\widehat{\theta}_n) - q(\theta)\big) \xrightarrow[n\to\infty]{d(\mathbb{P}_F)} N\Big(0, \nabla q(\theta)\Sigma(\theta)\nabla q(\theta)^\top\Big), \quad \forall F \in \mathcal{F}_\Theta. \tag{2.2}$$

If the limit covariance matrix has a continuous inverse, this limit allows to construct asymptotic confidence sets, such as ellipsoids, boxes, etc. (Problem 3). In dimension one, $d = k = 1$, the interval with the endpoints

$$q(\widehat{\theta}_n) \pm \frac{1}{\sqrt{n}}\big|q'(\widehat{\theta}_n)\big|\sqrt{\Sigma(\widehat{\theta}_n)}z_{1-\alpha/2}$$

is an asymptotic $1-\alpha$ confidence interval.

2.2. Heuristics of nonparametric Delta method. Let us sketch the basic heuristics for generalization of the Delta-method to nonparametric models, introduced by R. von Mises. Denote by F the true distribution from which the sample is drawn. Then for any $t \in [0,1]$ and any other distribution G the function $F + t(G-F)$ is also a legitimate distribution. Here $t(G-F)$ can be viewed as a perturbation of F in the direction $G-F$, whose magnitude is controlled by t.

Suppose T is such that the limit

$$\dot{T}_F(G-F) := \lim_{t\to 0}\frac{1}{t}\Big(T\big(F + t(G-F)\big) - T(F)\Big) \tag{2.3}$$

exists and, moreover, has the form

$$\dot{T}_F(G-F) = \int \psi(x)d(G-F) = \int \big(\psi(x) - \mathbb{E}_F\psi\big)dG \tag{2.4}$$

for some function ψ, specific to F and T, but independent of G. Analogously to multivariate calculus, the limit (2.3) should be thought of as a directional derivative of T at the distribution F in the direction $G-F$. The integrand $L_F(x) := \psi(x) - \mathbb{E}_F\psi$ is called *the influence curve* of T and formally it is the derivative evaluated in the direction $\Delta_x - F$, where $\Delta_x(u) = \mathbf{1}_{\{u\geq x\}}$ is the distribution function of the point measure [2] δ_x.

The convergence in (2.3) suggests the following useful limit

$$\sqrt{n}\big(T(\widehat{F}_n)-T(F)\big) = \frac{1}{n^{-1/2}}\Big(T\big(F+n^{-1/2}n^{1/2}(\widehat{F}_n-F)\big)-T(F)\Big) =$$
$$\dot{T}_F\big(\sqrt{n}(\widehat{F}_n-F)\big)\big(1+o_{\mathbb{P}_F}(1)\big) = \sqrt{n}\int\big(\psi(x)-\mathbb{E}_F\psi\big)d\widehat{F}_n\big(1+o_{\mathbb{P}_F}(1)\big) =$$
$$\frac{1}{\sqrt{n}}\sum_{j=1}^n\big(\psi(X_i)-\mathbb{E}_F\psi\big)\big(1+o_{\mathbb{P}_F}(1)\big)\xrightarrow[n\to\infty]{d(\mathbb{P}_F)}N(0,V_F),$$

where the convergence is due to the standard CLT and Slutsky's lemma, and

$$V_F = \mathrm{Var}_F(\psi) = \int L_F(x)^2 dF(x). \tag{2.5}$$

If the plug-in estimator $V_{\widehat{F}_n}$ is consistent for V_F, then by Slutsky's lemma, the interval with the endpoints at

$$T(\widehat{F}_n) \pm \frac{1}{\sqrt{n}}\sqrt{V_{\widehat{F}_n}}z_{1-\alpha/2} \tag{2.6}$$

is an asymptotic $1-\alpha$ confidence interval for $T(F)$.

2.3. Linear functionals. For linear functionals of the form (1.1) this program can be implemented by elementary means. The plug-in estimator is the empirical mean of ψ,

$$T(\widehat{F}_n) = \int \psi(x)d\widehat{F}_n(x) = \frac{1}{n}\sum_{j=1}^n \psi(X_j).$$

By the classical LLN, this estimator is consistent, if $\mathbb{E}_F|\psi(X_1)| < \infty$. Furthermore, for all $t\in(0,1)$,

$$\frac{1}{t}\Big(T(F+t(G-F))-T(F)\Big) = \int \psi d(G-F) =: \dot{T}_F(G-F)$$

and hence, if $\mathbb{E}_F\psi(X_1)^2<\infty$,

$$\sqrt{n}(T(\widehat{F}_n)-T(F)) = \dot{T}_F\big(\sqrt{n}(\widehat{F}_n-F)\big) = \sqrt{n}\int\psi d(\widehat{F}_n-F) =$$
$$\frac{1}{\sqrt{n}}\sum_{j=1}^n(\psi(X_j)-\mathbb{E}_F\psi)\xrightarrow[n\to\infty]{d(\mathbb{P}_F)}N(0,V_F),$$

by the standard CLT, where the asymptotic variance is

$$V_F = \mathrm{Var}_F(\psi) = \int(\psi(x)-\mathbb{E}_F\psi)^2 dF(x). \tag{2.7}$$

---

[2] δ_x is the point probability measure, $\delta_x(A) = \mathbf{1}_{\{x\in A\}}$ for any measurable set A.

The plug-in estimator $V_{\widehat{F}_n}$ for V_F is consistent (Problem 2) and thus (2.6) is indeed an asymptotic $1-\alpha$ confidence interval.

2.4. Nonlinear functionals. For functionals more complicated than linear, the von Mises program is much harder to justify and requires quite sophisticated mathematical tools, see [3], [4, Ch 21]. There are several notions of functional derivatives, of which the following turns out to be the most suitable.

DEFINITION 2.1. *Let D and E be linear normed spaces and $T : D_T \mapsto E$ be a map defined on a subset $D_T \subseteq D$. T is Hadamard differentiable at $x \in D_T$ if there exists a continuous linear map $\dot{T}_x : D \mapsto E$ such that*

$$\big\|T(x+th_t)-T(x)-\dot{T}_x(th)\big\|_E = o(t), \quad \text{as } t \to 0 \tag{2.8}$$

for any $h \in D$ and any h_t such that $x+th_t \in D_T$ and $\|h_t - h\|_D \to 0$.

REMARK 2.2. A weaker notion of the *Gateaux derivative* is obtained, if (2.8) is required to hold only for $h_t := th$. Obviously, if Hadamard derivative exists, the Gateaux derivative exists as well and both derivatives coincide. It turns out however, that Gateaux differentiability is not sufficient to imply the desired convergence in distribution. A stronger *Fréchet derivative* requires that instead of (2.8)

$$\big\|T(x+h)-T(x)-\dot{T}_x(h)\big\|_E = o(\|h\|), \quad \text{as } \|h\| \to 0$$

holds. If Fréchet derivative exists, then Hadamard derivative also exists and the two coincide. However, it turns out that some functionals of interest, which are Hadamard differentiable, are not Fréchet differentiable.

In the statistical context, D is the linear subset set of functions which can be obtained as differences of distribution functions. These are right continuous functions with finite left limits, called *cadlag*. Various norms can be chosen on this space and a natural choice is the uniform norm $\|\cdot\|_\infty$. The target set is $E = \mathbb{R}$ with its usual Euclidian norm. The subset D_T consists of distribution functions.

Note that convergence (2.8) is required to hold for a functional $\dot{T}_x$, whose domain is D. If instead it holds only on some subset $D_0 \subset D$, then T is said to be Hadamard differentiable *tangentially* to D_0. In statistical context, D_0 is the subset of continuous functions. If the true distribution F is continuous, the process $\sqrt{n}(\widehat{F}_n - F)$ converges weakly to $\overline{B} \circ F$, where $\overline{B}$ is the Brownian bridge. This limit process has continuous trajectories, which is the reason for this choice of D_0.

THEOREM 2.3. *Let T be a statistical functional, Hadamard differentiable at a continuous distribution F tangentially to the subset of continuous functions. Then*

$$\sqrt{n}\big(T(\widehat{F}_n) - T(F)\big) \xrightarrow[n\to\infty]{d(\mathbb{P}_F)} \dot{T}_F(\overline{B} \circ F).$$

Often the Hadamard derivative has the form, cf. (2.4),

$$\dot{T}_F(h) = \int \psi(x)dh(x) \tag{2.9}$$

for some function ψ. Then the random variable $\dot{T}_F(\overline{B} \circ F)$ is normal with zero mean and variance given by the formula (2.7) (Problem 10).

EXAMPLE 2.4. Let us see how this theory applies to the linear functionals of the form (1.1). To this end, let G_t be any family of distributions, such that $\|G_t - G\|_\infty \to 0$ as $t \to 0$. Then

$$T(F + t(G_t - F)) - T(F) =$$
$$\int \psi d\big(F + t(G_t - F)\big) - \int \psi dF = t \int \psi d(G_t - F).$$

Hence T has Hadamard derivative

$$\dot{T}_F(G - F) = \int \psi d(G - F),$$

if ψ is such that

$$\int \psi d(G_t - G) \to 0.$$

This is true, e.g., for bounded ψ. Indeed, for any $r > 0$, there exists a simple function ψ_r such that $\|\psi_r - \psi\|_\infty \le r$. Then

$$\left| \int \psi d(G_t - G) \right| = \int \psi_r d(G_t - G) + 2r \xrightarrow[t \to 0]{} 2r,$$

where the convergence holds since ψ_r is simple. The claim follows by arbitrariness of $r > 0$.

If ψ is unbounded T may not be Hadamard differentiable. In particular, the empirical mean is not. Indeed, take $G_t = (1 - t)G + t\Delta_{a(t)}$ for some function $a(t)$. Then $\|G_t - G\|_\infty = t \to 0$, but for $\psi(x) = x$,

$$\int \psi d(G_t - G) = -t \int x dG(x) + t a(t) \to \infty,$$

if $t a(t) \to \infty$. This of course does not contradict the result obtained by elementary calculations in Section 2.3, since the conditions of Theorem 2.3 are only sufficient. ■

EXAMPLE 2.5. Assume that F has a continuous positive density function f. Fix a number $p \in (0, 1)$ and consider the quantile functional

$$T(F) = \min\{x \in \mathbb{R} : F(x) \ge p\} = F^{-1}(p).$$

Note that this functional is nonlinear and cannot be put in the form (1.1). Let G_t be a family of distribution functions, such that $\|G_t - G\|_\infty \to 0$ for some distribution G, which is continuous at $F^{-1}(p)$. Thus, unlike in the previous example, we will be able to argue that this functional is Hadamard differentiable only tangentially to the subset of continuous functions, which, however, suffices for Theorem 2.3.

For brevity, let $q = T(F)$ and $q_t = T\big((1 - t)F + tG_t\big)$. We aim at finding the limit $\lim_{t \to 0}(q_t - q)/t$. Since both F and G are continuous at q,

$$(1 - t)F(q_t) + tG_t(q_t) = p \tag{2.10}$$

and $F(q) = p$. Subtracting these equations gives

$$F(q_t) - F(q) = tF(q_t) - tG_t(q_t),$$

or, equivalently,

$$\int_q^{q_t} f(t)dt = tF(q_t) - tG_t(q_t). \tag{2.11}$$

This implies $\int_q^{q_t} f(t)dt \xrightarrow[t\to 0]{} 0$ and, since f is positive, $q_t \to q$. Moreover, (2.11) can be rearranged as

$$\frac{q_t - q}{t}\frac{1}{q_t - q}\int_q^{q_t} f(t)dt = \big(F(q_t) - G(q_t)\big) + \big(G(q_t) - G_t(q_t)\big).$$

Taking the limit $t \to 0$ we obtain the following expression for the Hadamard derivative

$$\dot{T}_F(G-F) = \lim_{t\to 0}\frac{q_t - q}{t} = \frac{F(q) - G(q)}{f(q)} = -\int \frac{\mathbf{1}_{\{u\le q\}}}{f(q)} d(G-F),$$

which has the form (2.9) with $\psi(u) = -\mathbf{1}_{\{u\le q\}}/f(q)$. This can be also written as

$$\dot{T}_F(G-F) = \int \frac{F(q) - \mathbf{1}_{\{u\le q\}}}{f(q)} dG(u).$$

The influence function is obtained by evaluating this expression at the distribution $G(u) := \Delta_x(u) = \mathbf{1}_{\{u\ge x\}}$,

$$L_F(x) = \frac{p - \mathbf{1}_{\{x\le q\}}}{f(q)}, \quad x \in \mathbb{R}.$$

The corresponding variance is

$$V_F = \int_{\mathbb{R}} L_F(x)^2 dF(x) = \frac{1}{f(q)^2}\int_{\mathbb{R}} (p - \mathbf{1}_{\{q\ge x\}})^2 dF(x) = \frac{p(1-p)}{f(q)^2} = \frac{p(1-p)}{f(F^{-1}(p))^2}.$$

The plug-in estimator is the empirical quantile $T(\widehat{F}_n) = X_{(\lceil np\rceil)}$ (Problem 1). To be able to construct an asymptotic confidence interval, we need a density estimator $\widehat{f}_n$, consistent with respect to the supremum norm. In the following chapters, we will see how such estimators can be constructed for smooth densities. Then

$$X_{(\lceil np\rceil)} \pm \frac{z_{1-\alpha/2}}{\sqrt{n}}\frac{\sqrt{p(1-p)}}{\widehat{f}_n(X_{(\lceil np\rceil)})}$$

is an asymptotic confidence interval with coverage probability $1-\alpha$. ■

Computer experiment. Implement the asymptotic confidence interval from Problem 8. Run $M = 10{,}000$ MC trials to approximate the actual coverage probability and the mean interval length for the sample of size $n = 100$ from some bivariate distribution. Compare to the target coverage probability. Repeat for the sample size of $n = 1000$. Explain the obtained results.

Exercises.

PROBLEM 1. Find the plug-in estimators for the mean, variance and quantile.

PROBLEM 2. Prove that the plug-in estimator for the functional V_F defined in (2.7) is strongly consistent,

$$V_{\widehat{F}_n} \xrightarrow[n\to\infty]{\mathbb{P}_F-a.s.} V_F,$$

for any ψ and F such that $\int \psi^2 dF < \infty$.

PROBLEM 3. Let $X_1,...,X_n \sim N(\mu,\sigma^2)$ where $\theta = (\mu,\sigma^2) \in \mathbb{R}\times\mathbb{R}_+$ is the unknown parameter. Apply parametric Delta method to the m.l.e. $\widehat{\theta}_n$ for θ in order to construct asymptotic $1-\alpha$ confidence intervals for the following functionals.

(1) The fourth moment $\mathbb{E}_\theta X_1^4$.

(2) The p-th quantile.

(3) The probability $\mathbb{P}_\theta(X_1 \in [a,b])$ where $a<b$ are fixed real numbers.

PROBLEM 4. Find the plug-in estimator for the functional

$$T(F) = \mathbb{P}_F(X_1 \in (a,b]) = F(b) - F(a),$$

where $a<b$ are fixed numbers. Argue that this functional is Hadamard differentiable, find the influence function and the limit variance. Construct $1-\alpha$ asymptotic confidence interval for $T(F)$.

PROBLEM 5. Let $\mathcal{F}$ be the subset of distribution functions with continuous densities. For $F \in \mathcal{F}$ let D be the functional which maps F to its derivative at a fixed point x_0,

$$D(F) = \frac{d}{dx}F(x)\Big|_{x=x_0}$$

Argue that D is not Hadamard differentiable on $\mathcal{F}$.

Hint: Show that the limit in the definition fails for the family

$$G_t(x) := (1-\sqrt{t})\Phi(x) + \sqrt{t}\Phi((x-x_0)/t), \quad t\in(0,1),$$

where Φ is the standard normal distribution.

PROBLEM 6 (Chain rule). Consider the functional

$$H(F) = h\big(T_1(F),...,T_k(F)\big) = h(T(F)),$$

where $h:\mathbb{R}^k \mapsto \mathbb{R}$ is continuously differentiable and $T_j(F)$ are Hadamard differentiable functionals with influence functions $L_F^j(x)$, $x\in\mathbb{R}$. Show that H is Hadamard differentiable with the influence function

$$L_F(x) = \nabla h(T(F))\mathbf{L}_F(x) \tag{2.12}$$

where $\mathbf{L}_F(x)$ is the vector with entries $L_F^j(x)$.

PROBLEM 7. Use the non-parametric Delta method to construct asymptotic $1-\alpha$ confidence interval for variance

$$\sigma^2(F) = \int x^2 dF(x) - \Big(\int x dF(x)\Big)^2$$

if F is a continuous distribution on a bounded interval of $\mathbb{R}$.

PROBLEM 8. Consider the problem of estimating the correlation coefficient

$$\rho_F = \frac{\mathrm{Cov}_F(X_1, Y_1)}{\sqrt{\mathrm{Var}_F(X_1)\,\mathrm{Var}_F(Y_1)}}$$

from the random sample $(X_1, Y_1), \dots, (X_n, Y_n) \sim F$, where F is a distribution on a bounded subset of $\mathbb{R}^2$.

(1) Specify the plug-in estimator.
(2) Extending the result from the previous problem to distributions on the plane $\mathbb{R}^2$, prove that the influence function is

$$L_F(x,y) = \widetilde{x}\widetilde{y} - \frac{1}{2}\rho_F(\widetilde{x}^2 + \widetilde{y}^2),$$

where

$$\widetilde{x} = \frac{x - \int s dF(s,t)}{\sqrt{\int s^2 dF(s,t) - (\int s dF(s,t))^2}}, \quad \widetilde{y} = \frac{y - \int t dF(s,t)}{\sqrt{\int t^2 dF(s,t) - (\int t dF(s,t))^2}}.$$

Hint: the calculations are somewhat cumbersome

(3) Detail the plug-in estimator for the limit variance.

PROBLEM 9. Let F_0 be a fixed known continuous distribution and consider the functional

$$T(F) = \int (F - F_0)^2 dF_0.$$

The Cramer-von Mises statistic $T(\widehat{F}_n)$ is used for testing the simple hypothesis $H_0 : F = F_0$ against nonparametric alternatives. This problem explores some of its properties.

(1) Detail the plug-in estimator.

(2) Show that $T(F)$ is Hadamard differentiable and find the corresponding derivative.

(3) Prove that $\dot{T}_F(\overline{B} \circ F) \sim N(0, V_F)$ where $V_F = \int L_F(x)^2 dF(x)$ and $L_F(x)$ is the influence function of T.

Hint: verify and use the formula

$$\mathrm{Cov}\big((\overline{B} \circ F)(u), (\overline{B} \circ F)(v)\big) = \int \big(\mathbf{1}_{\{u \ge x\}} - F(u)\big)\big(\mathbf{1}_{\{v \ge x\}} - F(v)\big) dF(x),$$

where $\overline{B}$ is the Brownian bridge.

(4) Construct $1-\alpha$ confidence interval for $T(F)$ assuming $F \neq F_0$. Why is this assumption necessary?

PROBLEM 10. Let F be a continuous distribution and $\overline{B}$ the Brownain bridge process, i.e., the Gaussian process with zero mean and covariance function

$$\mathrm{Cov}(\overline{B}(s), \overline{B}(t)) = s \wedge t - st, \quad s,t \in [0,1].$$

Let ψ be a sufficiently regular function[3] with bounded support so that Riemann–Stieltjes integral

$$\int (\overline{B} \circ F)(t) d\psi(t)$$

is well defined. Using the integration by parts formula, define

$$\int \psi(t) d(\overline{B} \circ F)(t) := -\int (\overline{B} \circ F)(t) d\psi(t).$$

Obviously, this is a normal random variable with zero mean. Show that it's variance is given by the formula (2.7).

[3] ψ must have finite variation, e.g., be piecewise continuously differentiable except for countably many points at which it may have summable jumps.

References

[1] Wasserman, Larry. All of nonparametric statistics. Springer Science & Business Media, 2006.
[2] I.A. Ibragimov, R. Z. Hasminskii, Statistical estimation, asymptotic theory, Springer Verlag 1981
[3] Fernholz, Luisa Turrin. Von Mises calculus for statistical functionals. Vol. 19. Springer Science & Business Media, 2012.
[4] Van der Vaart, Aad W. Asymptotic statistics. Vol. 3. Cambridge university press, 2000.

DEPARTMENT OF STATISTICS, THE HEBREW UNIVERSITY, MOUNT SCOPUS, JERUSALEM 91905, ISRAEL
E-mail address: Pavel.Chigansky@mail.huji.ac.il