20

Functional Delta Method

The delta method was introduced in Chapter 3 as an easy way to turn the weak convergence of a sequence of random vectors $r_n(T_n - \theta)$ into the weak convergence of transformations of the type $r_n\big(\phi(T_n) - \phi(\theta)\big)$. It is useful to apply a similar technique in combination with the more powerful convergence of stochastic processes. In this chapter we consider the delta method at two levels. The first section is of a heuristic character and limited to the case that T_n is the empirical distribution. The second section establishes the delta method rigorously and in general, completely parallel to the delta method for $\mathbb{R}^k$, for Hadamard differentiable maps between normed spaces.

20.1 von Mises Calculus

Let $\mathbb{P}_n$ be the empirical distribution of a random sample $X_1, \ldots, X_n$ from a distribution P. Many statistics can be written in the form $\phi(\mathbb{P}_n)$, where ϕ is a function that maps every distribution of interest into some space, which for simplicity is taken equal to the real line. Because the observations can be regained from $\mathbb{P}_n$ completely (unless there are ties), any statistic can be expressed in the empirical distribution. The special structure assumed here is that the statistic can be written as a fixed function ϕ of $\mathbb{P}_n$, independent of n, a strong assumption.

Because $\mathbb{P}_n$ converges to P as n tends to infinity, we may hope to find the asymptotic behavior of $\phi(\mathbb{P}_n) - \phi(P)$ through a differential analysis of ϕ in a neighborhood of P. A first-order analysis would have the form

$$\phi(\mathbb{P}_n) - \phi(P) = \phi'_P(\mathbb{P}_n - P) + \cdots,$$

where ϕ'_P is a "derivative" and the remainder is hopefully negligible. The simplest approach towards defining a derivative is to consider the function $t \mapsto \phi(P + tH)$ for a fixed perturbation H and as a function of the real-valued argument t. If ϕ takes its values in $\mathbb{R}$, then this function is just a function from the reals to the reals. Assume that the ordinary derivatives of the map $t \mapsto \phi(P + tH)$ at $t = 0$ exist for $k = 1, 2, \ldots, m$. Denoting them by $\phi_P^{(k)}(H)$, we obtain, by Taylor's theorem,

$$\phi(P + tH) - \phi(P) = t\phi'_P(H) + \cdots + \frac{1}{m!}t^m\phi_P^{(m)}(H) + o(t^m).$$

Substituting $t = 1/\sqrt{n}$ and $H = \mathbb{G}_n$, for $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ the empirical process of the observations, we obtain the *von Mises expansion*

$$\phi(\mathbb{P}_n) - \phi(P) = \frac{1}{\sqrt{n}}\phi'_P(\mathbb{G}_n) + \cdots + \frac{1}{m!}\frac{1}{n^{m/2}}\phi_P^{(m)}(\mathbb{G}_n) + \cdots.$$

Actually, because the empirical process $\mathbb{G}_n$ is dependent on n, it is not a legal choice for H under the assumed type of differentiability: There is no guarantee that the remainder is small. However, we make this our working hypothesis. This is reasonable, because the remainder has one factor $1/\sqrt{n}$ more, and the empirical process $\mathbb{G}_n$ shares at least one property with a fixed H: It is "bounded." Then the asymptotic distribution of $\phi(\mathbb{P}_n) - \phi(P)$ should be determined by the first nonzero term in the expansion, which is usually the first-order term $\phi'_P(\mathbb{G}_n)$. A method to make our wishful thinking rigorous is discussed in the next section. Even in cases in which it is hard to make the differentation operation rigorous, the von Mises expansion still has heuristic value. It may suggest the type of limiting behavior of $\phi(\mathbb{P}_n) - \phi(P)$, which can next be further investigated by ad-hoc methods.

We discuss this in more detail for the case that $m = 1$. A first derivative typically gives a *linear* approximation to the original function. If, indeed, the map $H \mapsto \phi'_P(H)$ is linear, then, writing $\mathbb{P}_n$ as the linear combination $\mathbb{P}_n = n^{-1}\sum \delta_{X_i}$ of the Dirac measures at the observations, we obtain

$$\phi(\mathbb{P}_n) - \phi(P) \approx \frac{1}{\sqrt{n}}\phi'_P(\mathbb{G}_n) = \frac{1}{n}\sum_{i=1}^{n}\phi'_P(\delta_{X_i} - P). \tag{20.1}$$

Thus, the difference $\phi(\mathbb{P}_n) - \phi(P)$ behaves as an average of the independent random variables $\phi'_P(\delta_{X_i} - P)$. If these variables have zero means and finite second moments, then a normal limit distribution of $\sqrt{n}\big(\phi(\mathbb{P}_n) - \phi(P)\big)$ may be expected. Here the zero mean ought to be automatic, because we may expect that

$$\int \phi'_P(\delta_x - P)\,dP(x) = \phi'_P\left(\int (\delta_x - P)\,dP(x)\right) = \phi'_P(0) = 0.$$

The interchange of order of integration and application of ϕ'_P is motivated by linearity (and continuity) of this derivative operator.

The function $x \mapsto \phi'_P(\delta_x - P)$ is known as the *influence function* of the function ϕ. It can be computed as the ordinary derivative

$$\phi'_P(\delta_x - P) = \frac{d}{dt}_{|t=0}\phi\big((1-t)P + t\delta_x\big).$$

The name "influence function" originated in developing robust statistics. The function measures the change in the value $\phi(P)$ if an infinitesimally small part of P is replaced by a pointmass at x. In robust statistics, functions and estimators with an unbounded influence function are suspect, because a small fraction of the observations would have too much influence on the estimator if their values were equal to an x where the influence function is large.

In many examples the derivative takes the form of an "expectation operator" $\phi'_P(H) = \int \tilde{\phi}_P\,dH$, for some function $\tilde{\phi}_P$ with $\int \tilde{\phi}_P\,dP = 0$, at least for a subset of H. Then the influence function is precisely the function $\tilde{\phi}_P$.

20.2 ***Example (Mean).*** The sample mean is obtained as $\phi(\mathbb{P}_n)$ from the mean function $\phi(P) = \int s\,dP(s)$. The influence function is

$$\phi'_P(\delta_x - P) = \frac{d}{dt}_{|t=0} \int s\,d\big[(1-t)P + t\delta_x\big](s) = x - \int s\,dP(s).$$

In this case, the approximation (20.1) is an identity, because the function is linear already. If the sample space is a Euclidean space, then the influence function is unbounded and hence the sample mean is not robust. □

20.3 ***Example (Wilcoxon).*** Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a bivariate distribution. Write $\mathbb{F}_n$ and $\mathbb{G}_n$ for the empirical distribution functions of the X_i and Y_j, respectively, and consider the Mann-Whitney statistic

$$T_n = \int \mathbb{F}_n\,d\mathbb{G}_n = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n 1\{X_i \le Y_j\}.$$

This statistic corresponds to the function $\phi(F, G) = \int F\,dG$, which can be viewed as a function of two distribution functions, or also as a function of a bivariate distribution function with marginals F and G. (We have assumed that the sample sizes of the two samples are equal, to fit the example into the previous discussion, which, for simplicity, is restricted to i.i.d. observations.) The influence function is

$$\begin{aligned}\phi'_{(F,G)}(\delta_{x,y} - P) &= \frac{d}{dt}_{|t=0} \int \big[(1-t)F + t\delta_x\big]\,d\big[(1-t)G + t\delta_y\big] \\ &= F(y) + 1 - G_-(x) - 2\int F\,dG.\end{aligned}$$

The last step follows on multiplying out the two terms between square brackets: The function that is to be differentiated is simply a parabola in t. For this case (20.1) reads

$$\int \mathbb{F}_n\,d\mathbb{G}_n - \int F\,dG \approx \frac{1}{n}\sum_{i=1}^n \left(F(Y_i) + 1 - G_-(X_i) - 2\int F\,dG\right).$$

From the two-sample U-statistic theorem, Theorem 12.6, it is known that the difference between the two sides of the approximation sign is actually $o_P(1/\sqrt{n})$. Thus, the heuristic calculus leads to the correct answer. In the next section an alternative proof of the asymptotic normality of the Mann-Whitney statistic is obtained by making this heuristic approach rigorous. □

20.4 ***Example (Z-functions).*** For every θ in an open subset of $\mathbb{R}^k$, let $x \mapsto \psi_\theta(x)$ be a given, measurable map into $\mathbb{R}^k$. The corresponding Z-function assigns to a probability measure P a zero $\phi(P)$ of the map $\theta \mapsto P\psi_\theta$. (Consider only P for which a unique zero exists.) If applied to the empirical distribution, this yields a Z-estimator $\phi(\mathbb{P}_n)$.

Differentiating with respect to t across the identity

$$0 = (P + t\delta_x)\psi_{\phi(P+t\delta_x)} = P\psi_{\phi(P+t\delta_x)} + t\psi_{\phi(P+t\delta_x)}(x),$$

and assuming that the derivatives exist and that $\theta \mapsto \psi_\theta$ is continuous, we find

$$0 = \left(\frac{\partial}{\partial\theta}P\psi_\theta\right)_{\theta=\phi(P)}\left[\frac{d}{dt}\phi(P + t\delta_x)\right]_{t=0} + \psi_{\phi(P)}(x).$$

The expression enclosed by squared brackets is the influence function of the Z-function. Informally, this is seen to be equal to

$$-\left(\frac{\partial}{\partial\theta}P\psi_\theta\right)^{-1}_{\theta=\phi(P)}\psi_{\phi(P)}(x).$$

In robust statistics we look for estimators with bounded influence functions. Because the influence function is, up to a constant, equal to $\psi_{\phi(P)}(x)$, this is easy to achieve with Z-estimators!

The Z-estimators are discussed at length in Chapter 5. The theorems discussed there give sufficient conditions for the asymptotic normality, and an asymptotic expansion for $\sqrt{n}\big(\phi(\mathbb{P}_n)-\phi(P)\big)$. This is of the type (20.1) with the influence function as in the preceding display. □

20.5 ***Example (Quantiles).*** The pth quantile of a distribution function F is, roughly, the number $\phi(F) = F^{-1}(p)$ such that $FF^{-1}(p) = p$. We set $F_t = (1-t)F + t\delta_x$, and differentiate with respect to t the identity

$$p = F_tF_t^{-1}(p) = (1-t)F\big(F_t^{-1}(p)\big) + t\delta_x\big(F_t^{-1}(p)\big).$$

This "identity" may actually be only an inequality for certain values of p, t, and x, but we do not worry about this. We find that

$$0 \equiv -F\big(F^{-1}(p)\big) + f\big(F^{-1}(p)\big)\left[\frac{d}{dt}F_t^{-1}(p)\right]_{|t=0} + \delta_x\big(F^{-1}(p)\big).$$

The derivative within square brackets is the influence function of the quantile function and can be solved from the equation as

$$\phi'_F(\delta_x - F) = -\frac{1_{[x,\infty)}\big(F^{-1}(p)\big) - p}{f\big(F^{-1}(p)\big)}.$$

The graph of this function is given in Figure 20.1 and has the following interpretation. Suppose the pth quantile has been computed for a large sample, but an additional observation x is obtained. If x is to the left of the pth quantile, then the pth quantile decreases; if x is to the right, then the quantile increases. In both cases the rate of change is constant, irrespective of the location of x. Addition of an observation x at the pth quantile has an unstable effect.

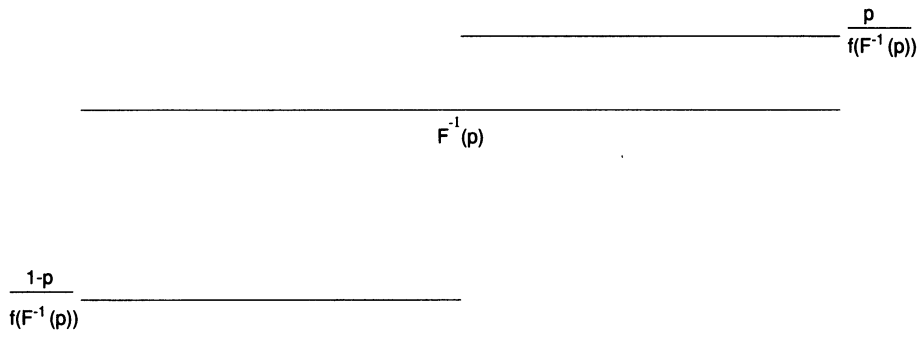

Figure 20.1. Influence function of the pth quantile.

The von Mises calculus suggests that the sequence of empirical quantiles $\sqrt{n}\big(\mathbb{F}_n^{-1}(t) - F^{-1}(t)\big)$ is asymptotically normal with variance $\operatorname{var}_F \phi_F'(\delta_{X_1}) = p(1-p)/f \circ F^{-1}(p)^2$. In Chapter 21 this is proved rigorously by the delta method of the following section. Alternatively, a pth quantile may be viewed as an M-estimator, and we can apply the results of Chapter 5. □

20.1.1 *Higher-Order Expansions*

In most examples the analysis of the first derivative suffices. This statement is roughly equivalent to the statement that most limiting distributions are normal. However, in some important examples the quadratic term dominates the von Mises expansion.

The second derivative $\phi_P''(H)$ ought to correspond to a *bilinear map*. Thus, it is better to write it as $\phi_P''(H, H)$. If the first derivative in the von Mises expansion vanishes, then we expect that

$$\phi(\mathbb{P}_n) - \phi(P) \approx \frac{1}{2}\frac{1}{n}\phi_P''(\mathbb{G}_n, \mathbb{G}_n) = \frac{1}{2}\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \phi_P''(\delta_{X_i} - P, \delta_{X_j} - P).$$

The right side is a *V-statistic* of degree 2 with kernel function equal to $h_P(x, y) = \frac{1}{2}\phi_P''(\delta_x - P, \delta_y - P)$. The kernel ought to be symmetric and degenerate in that $Ph_P(X, y) = 0$ for every y, because, by linearity and continuity,

$$\begin{aligned}\int \phi_P''(\delta_x - P, \delta_y - P)\, dP(x) &= \phi_P''\left(\int (\delta_x - P)\, dP(x), \delta_y - P\right)\\ &= \phi_P''(0, \delta_y - P) = 0.\end{aligned}$$

If we delete the diagonal, then a V-statistic turns into a U-statistic and hence we can apply Theorem 12.10 to find the limit distribution of $n\big(\phi(\mathbb{P}_n) - \phi(P)\big)$. We expect that

$$n\big(\phi(\mathbb{P}_n) - \phi(P)\big) = \frac{2}{n}\sum\sum_{i<j} h_P(X_i, X_j) + \frac{1}{n}\sum_{i=1}^n h_P(X_i, X_i) + o_P(1).$$

If the function $x \mapsto h_P(x, x)$ is P-integrable, then the second term on the right only contributes a constant to the limit distribution. If the function $(x, y) \mapsto h_P^2(x, y)$ is $(P \times P)$-integrable, then the first term on the right converges to an infinite linear combination of independent χ_1^2-variables, according to Example 12.12.

20.6 ***Example (Cramér–von Mises).*** The Cramér–von Mises statistic is the function $\phi(\mathbb{F}_n)$ for $\phi(F) = \int (F - F_0)^2\, dF_0$ and a fixed cumulative distribution function F_0. By direct calculation,

$$\phi(F + tH) = \phi(F) + 2t\int (F - F_0)H\, dF_0 + t^2\int H^2\, dF_0.$$

Consequently, the first derivative vanishes at $F = F_0$ and the second derivative is equal to $\phi_{F_0}''(H) = 2\int H^2\, dF_0$. The von Mises calculus suggests the approximation

$$\phi(\mathbb{F}_n) - \phi(F_0) \approx \frac{1}{2}\frac{1}{n}\phi_{F_0}''(\mathbb{G}_n) = \frac{1}{n}\int \mathbb{G}_n^2\, dF_0.$$

This is certainly correct, because it is just the definition of the statistic. The preceding discussion is still of some interest in that it suggests that the limit distribution is nonnormal and can be obtained using the theory of V-statistics. Indeed, by squaring the sum that is hidden in $\mathbb{G}_n^2$, we see that

$$n\phi(\mathbb{F}_n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\int \big(1_{X_i \le x} - F_0(x)\big)\big(1_{X_j \le x} - F_0(x)\big)\, dF_0(x).$$

In Example 12.13 we used this representation to find that the sequence $n\phi(\mathbb{F}_n) \rightsquigarrow (1/6) + \sum_{j=1}^{\infty} j^{-2}\pi^{-2}\big(Z_j^2 - 1\big)$ for an i.i.d. sequence of standard normal variables $Z_1, Z_2, \ldots$, if the true distribution F_0 is continuous. □

20.2 Hadamard-Differentiable Functions

Let T_n be a sequence of statistics with values in a normed space $\mathbb{D}$ such that $r_n(T_n - \theta)$ converges in distribution to a limit T, for a given, nonrandom θ, and given numbers $r_n \to \infty$. In the previous section the role of T_n was played by the empirical distribution $\mathbb{P}_n$, which might, for instance, be viewed as an element of the normed space $D[-\infty, \infty]$. We wish to prove that $r_n\big(\phi(T_n) - \phi(\theta)\big)$ converges to a limit, for every appropriately differentiable map ϕ, which we shall assume to take its values in another normed space $\mathbb{E}$.

There are several possibilities for defining differentiability of a map $\phi : \mathbb{D} \mapsto \mathbb{E}$ between normed spaces. A map ϕ is said to be *Gateaux differentiable* at $\theta \in \mathbb{D}$ if for every fixed h there exists an element $\phi'_\theta(h) \in \mathbb{E}$ such that

$$\phi(\theta + th) - \phi(\theta) = t\phi'_\theta(h) + o(t), \qquad \text{as } t \downarrow 0.$$

For $\mathbb{E}$ the real line, this is precisely the differentiability as introduced in the preceding section. Gateaux differentiability is also called "directional differentiability," because for every possible direction h in the domain the derivative value $\phi'_\theta(h)$ measures the direction of the infinitesimal change in the value of the function ϕ. More formally, the $o(t)$ term in the previous displayed equation means that

$$\left\| \frac{\phi(\theta + th) - \phi(\theta)}{t} - \phi'_\theta(h) \right\|_{\mathbb{E}} \to 0, \qquad \text{as } t \downarrow 0. \tag{20.7}$$

The suggestive notation $\phi'_\theta(h)$ for the "tangent vectors" encourages one to think of the directional derivative as a map $\phi'_\theta : \mathbb{D} \mapsto \mathbb{E}$, which approximates the difference map $\phi(\theta + h) - \phi(\theta) : \mathbb{D} \mapsto \mathbb{E}$. It is usually included in the definition of Gateaux differentiability that this map $\phi'_\theta : \mathbb{D} \mapsto \mathbb{E}$ be linear and continuous.

However, Gateaux differentiability is too weak for the present purposes, and we need a stronger concept. A map $\phi : \mathbb{D}_\phi \mapsto \mathbb{E}$, defined on a subset $\mathbb{D}_\phi$ of a normed space $\mathbb{D}$ that contains θ, is called *Hadamard differentiable* at θ if there exists a continuous, linear map $\phi'_\theta : \mathbb{D} \mapsto \mathbb{E}$ such that

$$\left\| \frac{\phi(\theta + th_t) - \phi(\theta)}{t} - \phi'_\theta(h) \right\|_{\mathbb{E}} \to 0, \qquad \text{as } t \downarrow 0, \text{ every } h_t \to h.$$

(More precisely, for every $h_t \to h$ such that $\theta + th_t$ is contained in the domain of ϕ for all small $t > 0$.) The values $\phi'_\theta(h)$ of the derivative are the same for the two types

of differentiability. The difference is that for Hadamard-differentiability the directions h_t are allowed to change with t (although they have to settle down eventually), whereas for Gateaux differentiability they are fixed. The definition as given requires that $\phi'_\theta : \mathbb{D} \mapsto \mathbb{E}$ exists as a map on the whole of $\mathbb{D}$. If this is not the case, but ϕ'_θ exists on a subset $\mathbb{D}_0$ and the sequences $h_t \to h$ are restricted to converge to limits $h \in \mathbb{D}_0$, then ϕ is called Hadamard differentiable *tangentially* to this subset.

It can be shown that Hadamard differentiability is equivalent to the difference in (20.7) tending to zero uniformly for h in compact subsets of $\mathbb{D}$. For this reason, it is also called *compact differentiability*. Because weak convergence of random elements in metric spaces is intimately connected with compact sets, through Prohorov's theorem, Hadamard differentiability is the right type of differentiability in connection with the delta method.

The derivative map $\phi'_\theta : \mathbb{D} \mapsto \mathbb{E}$ is assumed to be linear and continuous. In the case of finite-dimensional spaces a linear map can be represented by matrix multiplication and is automatically continuous. In general, linearity does not imply continuity.

Continuity of the map $\phi'_\theta : \mathbb{D} \mapsto \mathbb{E}$ should not be confused with continuity of the dependence $\theta \mapsto \phi'_\theta$ (if ϕ has derivatives in a neighborhood of θ-values). If the latter continuity holds, then ϕ is called *continuously differentiable*. This concept requires a norm on the set of derivative maps but need not concern us here.

For completeness we discuss a third, stronger form of differentiability. The map $\phi : \mathbb{D}_\phi \mapsto \mathbb{E}$ is called *Fréchet differentiable* at θ if there exists a continuous, linear map $\phi'_\theta : \mathbb{D} \mapsto \mathbb{E}$ such that

$$\big\|\phi(\theta + h) - \phi(\theta) - \phi'_\theta(h)\big\|_{\mathbb{E}} = o\big(\|h\|\big), \qquad \text{as } \|h\| \downarrow 0.$$

Because sequences of the type th_t, as employed in the definition of Hadamard differentiability, have norms satisfying $\|th_t\| = O(t)$, Fréchet differentiability is the most restrictive of the three concepts. In statistical applications, Fréchet differentiability may not hold, whereas Hadamard differentiability does. We did not have this problem in Section 3.1, because Hadamard and Fréchet differentiability are equivalent when $\mathbb{D} = \mathbb{R}^k$.

20.8 ***Theorem (Delta method).*** *Let $\mathbb{D}$ and $\mathbb{E}$ be normed linear spaces. Let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ be Hadamard differentiable at θ tangentially to $\mathbb{D}_0$. Let $T_n : \Omega_n \mapsto \mathbb{D}_\phi$ be maps such that $r_n(T_n - \theta) \rightsquigarrow T$ for some sequence of numbers $r_n \to \infty$ and a random element T that takes its values in $\mathbb{D}_0$. Then $r_n\big(\phi(T_n) - \phi(\theta)\big) \rightsquigarrow \phi'_\theta(T)$. If ϕ'_θ is defined and continuous on the whole space $\mathbb{D}$, then we also have $r_n\big(\phi(T_n) - \phi(\theta)\big) = \phi'_\theta\big(r_n(T_n - \theta)\big) + o_P(1)$.*

Proof. To prove that $r_n\big(\phi(T_n) - \phi(\theta)\big) \rightsquigarrow \phi'_\theta(T)$, define for each n a map $g_n(h) = r_n\big(\phi\big(\theta + r_n^{-1}h\big) - \phi(\theta)\big)$ on the domain $\mathbb{D}_n = \big\{h : \theta + r_n^{-1}h \in \mathbb{D}_\phi\big\}$. By Hadamard differentiability, this sequence of maps satisfies $g_{n'}(h_{n'}) \to \phi'_\theta(h)$ for every subsequence $h_{n'} \to h \in \mathbb{D}_0$. Therefore, $g_n\big(r_n(T_n - \theta)\big) \rightsquigarrow \phi'_\theta(T)$ by the extended continuous-mapping theorem, Theorem 18.11, which is the first assertion.

The seemingly stronger last assertion of the theorem actually follows from this, if applied to the function $\psi = (\phi, \phi'_\theta) : \mathbb{D} \mapsto \mathbb{E} \times \mathbb{E}$. This is Hadamard-differentiable at (θ, θ) with derivative $\psi'_\theta = (\phi'_\theta, \phi'_\theta)$. Thus, by the preceding paragraph, $r_n\big(\psi(T_n) - \psi(\theta)\big)$ converges weakly to $\big(\phi'_\theta(T), \phi'_\theta(T)\big)$ in $\mathbb{E} \times \mathbb{E}$. By the continuous-mapping theorem, the difference $r_n\big(\phi(T_n) - \phi(\theta)\big) - \phi'_\theta\big(r_n(T_n - \theta)\big)$ converges weakly to $\phi'_\theta(T) - \phi'_\theta(T) = 0$. Weak convergence to a constant is equivalent to convergence in probability. ■

Without the *chain rule*, Hadamard differentiability would not be as interesting. Consider maps $\phi : \mathbb{D} \mapsto \mathbb{E}$ and $\psi : \mathbb{E} \mapsto \mathbb{F}$ that are Hadamard-differentiable at θ and $\phi(\theta)$, respectively. Then the composed map $\psi \circ \phi : \mathbb{D} \mapsto \mathbb{F}$ is Hadamard-differentiable at θ, and the derivative is the map obtained by composing the two derivative maps. (For Euclidean spaces this means that the derivative can be found through matrix multiplication of the two derivative matrices.) The attraction of the chain rule is that it allows a calculus of Hadamard-differentiable maps, in which differentiability of a complicated map can be established by decomposing this into a sequence of basic maps, of which Hadamard differentiability is known or can be proven easily. This is analogous to the chain rule for real functions, which allows, for instance, to see the differentiability of the map $x \mapsto \exp\cos\log(1 + x^2)$ in a glance.

20.9 *Theorem (Chain rule).* *Let $\phi : \mathbb{D}_\phi \mapsto \mathbb{E}_\psi$ and $\psi : \mathbb{E}_\psi \mapsto \mathbb{F}$ be maps defined on subsets $\mathbb{D}_\phi$ and $\mathbb{E}_\psi$ of normed spaces $\mathbb{D}$ and $\mathbb{E}$, respectively. Let ϕ be Hadamard-differentiable at θ tangentially to $\mathbb{D}_0$ and let ψ be Hadamard-differentiable at $\phi(\theta)$ tangentially to $\phi'_\theta(\mathbb{D}_0)$. Then $\psi \circ \phi : \mathbb{D}_\phi \mapsto \mathbb{F}$ is Hadamard-differentiable at θ tangentially to $\mathbb{D}_0$ with derivative $\psi'_{\phi(\theta)} \circ \phi'_\theta$.*

Proof. Take an arbitrary converging path $h_t \to h$ in $\mathbb{D}$. With the notation $g_t = t^{-1}\big(\phi(\theta + th_t) - \phi(\theta)\big)$, we have

$$\frac{\psi \circ \phi(\theta + th_t) - \psi \circ \phi(\theta)}{t} = \frac{\psi\big(\phi(\theta) + tg_t\big) - \psi\big(\phi(\theta)\big)}{t}.$$

By Hadamard differentiability of ϕ, $g_t \to \phi'_\theta(h)$. Thus, by Hadamard differentiability of ψ, the whole expression goes to $\psi'_{\phi(\theta)}\big(\phi'_\theta(h)\big)$. ■

20.3 Some Examples

In this section we give examples of Hadamard-differentiable functions and applications of the delta method. Further examples, such as quantiles and trimmed means, are discussed in separate chapters.

The Mann-Whitney statistic can be obtained by substituting the empirical distribution functions of two samples of observations into the function $(F, G) \mapsto \int F\,dG$. This function also plays a role in the construction of other estimators. The following lemma shows that it is Hadamard-differentiable. The set $BV_M[a, b]$ is the set of all cadlag functions $z : [a, b] \mapsto [-M, M] \subset \mathbb{R}$ of variation bounded by M (the set of differences of $z_1 - z_2$ of two monotonely increasing functions that together increase no more than M).

20.10 *Lemma.* *Let $\phi : [0, 1] \mapsto \mathbb{R}$ be twice continuously differentiable. Then the function $(F_1, F_2) \mapsto \int \phi(F_1)\,dF_2$ is Hadamard-differentiable from the domain $D[-\infty, \infty] \times BV_1[-\infty, \infty] \subset D[-\infty, \infty] \times D[-\infty, \infty]$ into $\mathbb{R}$ at every pair of functions of bounded variation (F_1, F_2). The derivative is given by*[†]

$$(h_1, h_2) \mapsto h_2\,\phi \circ F_1|_{-\infty}^{\infty} - \int h_{2-}\,d\phi \circ F_1 + \int \phi'(F_1)h_1\,dF_2.$$

[†] We denote by h_- the left-continuous version of a cadlag function h and abbreviate $h|_a^b = h(b) - h(a)$.

Furthermore, the function $(F_1, F_2) \mapsto \int_{(-\infty,\cdot]} \phi(F_1)\, dF_2$ *is Hamamard-differentiable as a map into* $D[-\infty, \infty]$.

Proof. Let $h_{1t} \to h_1$ and $h_{2t} \to h_2$ in $D[-\infty, \infty]$ be such that $F_{2t} = F_2 + th_{2t}$ is a function of variation bounded by 1 for each t. Because F_2 is of bounded variation, it follows that h_{2t} is of bounded variation for every t. Now, with $F_{1t} = F_1 + th_{1t}$,

$$\frac{1}{t}\left(\int \phi(F_{1t})\, dF_{2t} - \int \phi(F_1)\, dF_2\right)$$
$$= \int \left(\frac{\phi(F_{1t}) - \phi(F_1)}{t} - \phi'(F_1)h_1\right) dF_{2t} + \int \phi(F_1)\, dh_{2t} + \int \phi'(F_1)h_1\, dF_{2t}.$$

By partial integration, the second term on the right can be rewritten as $\phi \circ F_1 h_{2t}|_{-\infty}^{\infty} - \int h_{2t-}\, d\phi \circ F_1$. Under the assumption on h_{2t}, this converges to the first part of the derivative as given in the lemma. The first term is bounded above by $\left(\|\phi''\|_\infty t\|h_{1t}\|_\infty + \|\phi'\|_\infty \|h_{1t} - h_1\|_\infty\right) \int d|F_{2t}|$. Because the measures F_{2t} are of total variation at most 1 by assumption, this expression converges to zero. To analyze the third term on the right, take a grid $u_0 = -\infty < u_1 < \cdots < u_m = \infty$ such that the function $\phi' \circ F_1\, h_1$ varies less than a prescribed value $\varepsilon > 0$ on each interval $[u_{i-1}, u_i)$. Such a grid exists for every element of $D[-\infty, \infty]$ (problem 18.6). Then

$$\left|\int \phi'(F_1)h_1\, d(F_{2t} - F_2)\right| \le \varepsilon\left(\int d|F_{2t}| + d|F_2|\right)$$
$$+ \sum_{i=1}^{m+1} \left|(\phi' \circ F_1\, h_1)(u_{i-1})\right| \left|F_{2t}[u_{i-1}, u_i) - F_2[u_{i-1}, u_i)\right|.$$

The first term is bounded by $\varepsilon O(1)$, in which the ε can be made arbitrarily small by the choice of the partition. For each fixed partition, the second term converges to zero as $t \downarrow 0$. Hence the left side converges to zero as $t \downarrow 0$.

This proves the first assertion. The second assertion follows similarly. ■

20.11 ***Example (Wilcoxon).*** Let $\mathbb{F}_m$ and $\mathbb{G}_n$ be the empirical distribution functions of two independent random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from distribution functions F and G, respectively. As usual, consider both m and n as indexed by a parameter ν, let $N = m + n$, and assume that $m/N \to \lambda \in (0, 1)$ as $\nu \to \infty$. By Donsker's theorem and Slutsky's lemma,

$$\sqrt{N}(\mathbb{F}_m - F, \mathbb{G}_n - G) \rightsquigarrow \left(\frac{\mathbb{G}_F}{\sqrt{\lambda}}, \frac{\mathbb{G}_G}{\sqrt{1-\lambda}}\right),$$

in the space $D[-\infty, \infty] \times D[-\infty, \infty]$, for a pair of independent Brownian bridges $\mathbb{G}_F$ and $\mathbb{G}_G$. The preceding lemma together with the delta method imply that

$$\sqrt{N}\left(\int \mathbb{F}_m\, d\mathbb{G}_n - \int F\, dG\right) \rightsquigarrow -\int \frac{\mathbb{G}_{G-}}{\sqrt{1-\lambda}}\, dF + \int \frac{\mathbb{G}_F}{\sqrt{\lambda}}\, dG.$$

The random variable on the right is a continuous, linear function applied to Gaussian processes. In analogy to the theorem that a linear transformation of a multivariate Gaussian vector has a Gaussian distribution, it can be shown that a continuous, linear transformation of a tight Gaussian process is normally distributed. That the present variable is normally

distributed can be more easily seen by applying the delta method in its stronger form, which implies that the limit variable is the limit in distribution of the sequence

$$-\int \sqrt{N}(\mathbb{G}_n - G)_-\, dF + \int \sqrt{N}(\mathbb{F}_m - F)\, dG.$$

This can be rewritten as the difference of two sums of independent random variables, and next we can apply the central limit theorem for real variables. □

20.12 ***Example (Two-sample rank statistics).*** Let $\mathbb{H}_N$ be the empirical distribution function of a sample $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ obtained by "pooling" two independent random samples from distributions F and G, respectively. Let $R_{N1}, \ldots, R_{NN}$ be the ranks of the pooled sample and let $\mathbb{G}_n$ be the empirical distribution function of the second sample. If no observations are tied, then $N\mathbb{H}_N(Y_j)$ is the rank of Y_j in the pooled sample. Thus,

$$\int \phi(\mathbb{H}_N)\, d\mathbb{G}_n = \frac{1}{n} \sum_{j=m+1}^{N} \phi\left(\frac{R_{Nj}}{N}\right)$$

is a two-sample rank statistic. This can be shown to be asymptotically normal by the preceding lemma. Because $N\mathbb{H}_N = m\mathbb{F}_m + n\mathbb{G}_n$, the asymptotic normality of the pair $(\mathbb{H}_N, \mathbb{G}_n)$ can be obtained from the asymptotic normality of the pair $(\mathbb{F}_m, \mathbb{G}_n)$, which is discussed in the preceding example. □

The *cumulative hazard function* corresponding to a cumulative distribution function F on $[0, \infty]$ is defined as

$$\Lambda_F(t) = \int_{[0,t]} \frac{dF}{1 - F_-}.$$

In particular, if F has a density f, then Λ_F has a density $\lambda_F = f/(1-F)$. If $F(t)$ gives the probability of "survival" of a person or object until time t, then $d\Lambda_F(t)$ can be interpreted as the probability of "instant death at time t given survival until t." The hazard function is an important modeling tool in survival analysis.

The correspondence between distribution functions and hazard functions is one-to-one. The cumulative distribution function can be explicitly recovered from the cumulative hazard function as the *product integral* of $-\Lambda$ (see the proof of Lemma 25.74),

$$1 - F_\Lambda(t) = \prod_{0<s\le t} \big(1 - \Lambda\{s\}\big) e^{-\Lambda^c(t)}. \tag{20.13}$$

Here $\Lambda\{s\}$ is the jump of Λ at s and $\Lambda^c(s)$ is the continuous part of Λ.

Under some restrictions the maps $F \leftrightarrow \Lambda_F$ are Hadamard differentiable. Thus, from an asymptotic-statistical point of view, estimating a distribution function and estimating a cumulative hazard function are the same problem.

20.14 ***Lemma.*** *Let $\mathbb{D}_\phi$ be the set of all nondecreasing cadlag functions $F : [0, \tau] \mapsto \mathbb{R}$ with $F(0) = 0$ and $1 - F(\tau) \ge \varepsilon > 0$ for some $\varepsilon > 0$, and let $\mathbb{E}_\psi$ be the set of all nondecreasing cadlag functions $\Lambda : [0, \tau] \mapsto \mathbb{R}$ with $\Lambda(0) = 0$ and $\Lambda(\tau) \le M$ for some $M \in \mathbb{R}$.*

(i) *The map* $\phi : \mathbb{D}_\phi \subset D[0,\tau] \mapsto D[0,\tau]$ *defined by* $\phi(F) = \Lambda_F$ *is Hadamard differentiable.*

(ii) *The map* $\psi : \mathbb{E}_\psi \subset D[0,\tau] \mapsto D[0,\tau]$ *defined by* $\psi(\Lambda) = F_\Lambda$ *is Hadamard differentiable.*

Proof. Part (i) follows from the chain rule and the Hadamard differentiability of each of the three maps in the decomposition

$$F \mapsto (F, 1 - F_-) \mapsto \left(F, \frac{1}{1-F_-}\right) \mapsto \int_{[0,t]} \frac{dF}{1-F_-}.$$

The differentiability of the first two maps is easy to see. The differentiability of the last one follows from Lemma 20.10. The proof of (ii) is longer; see, for example, [54] or [55]. ■

20.15 ***Example (Nelson-Aalen estimator).*** Consider estimating a distribution function based on *right-censored data*. We wish to estimate the distribution function F (or the corresponding cumulative hazard function Λ) of a random sample of "failure times" $T_1, \ldots, T_n$. Unfortunately, instead of T_i we only observe the pair (X_i, Δ_i), in which $X_i = T_i \wedge C_i$ is the minimum of T_i and a "censoring time" C_i, and $\Delta_i = 1\{T_i \leq C_i\}$ records whether T_i is censored ($\Delta_i = 0$) or not ($\Delta_i = 1$). The censoring time could be the closing date of the study or a time that a patient is lost for further observation. The cumulative hazard function of interest can be written

$$\Lambda(t) = \int_{[0,t]} \frac{1}{1-F_-}\, dF = \int_{[0,t]} \frac{1}{1-H_-}\, dH_1,$$

for $1 - H = (1 - F)(1 - G)$ and $dH_1 = (1 - G_-)dF$, and every choice of distribution function G. If we assume that the censoring times $C_1, \ldots, C_n$ are a random sample from G and are independent of the failure times T_i, then H is precisely the distribution function of X_i and H_1 is a "subdistribution function,"

$$1 - H(x) = \mathrm{P}(X_i > x), \qquad H_1(x) = \mathrm{P}(X_i \leq x, \Delta_i = 1).$$

An estimator for Λ is obtained by estimating these functions by the empirical distributions of the data, given by $\mathbb{H}_n(x) = n^{-1}\sum_{i=1}^n 1\{X_i \leq x\}$ and $\mathbb{H}_{1n}(x) = n^{-1}\sum_{i=1}^n 1\{X_i \leq x, \Delta_i = 1\}$, and next substituting these estimators in the formula for Λ. This yields the *Nelson-Aalen estimator*

$$\hat{\Lambda}_n(t) = \int_{[0,t]} \frac{1}{1-\mathbb{H}_{n-}}\, d\mathbb{H}_{1n}.$$

Because they are empirical distribution functions, the pair $(\mathbb{H}_n, \mathbb{H}_{1n})$ is asymptotically normal in the space $D[-\infty,\infty] \times D[-\infty,\infty]$. The easiest way to see this is to consider them as continuous transformations of the (bivariate) empirical distribution function of the pairs (X_i, Δ_i). The Nelson-Aalen estimator is constructed through the maps

$$(A, B) \mapsto (1 - A, B) \mapsto \left(\frac{1}{1-A}, B\right) \mapsto \int_{[0,t]} \frac{1}{1-A_-}\, dB.$$

These are Hadamard differentiable on appropriate domains, the main restrictions being that $1 - A$ should be bounded away from zero and B of uniformly bounded variation. The

asymptotic normality of the Nelson-Aalen estimator $\hat{\Lambda}_n(t)$ follows for every t such that $H(t) < 1$, and even as a process in $D[0, \tau]$ for every τ such that $H(\tau) < 1$.

If we apply the product integral given in (20.13) to the Nelson-Aalen estimator, then we obtain an estimator $1 - \hat{F}_n$ for the distribution function, known as the *product limit estimator* or *Kaplan-Meier estimator*. For a discrete hazard function the product integral is an ordinary product over the jumps, by definition, and it can be seen that

$$1 - \hat{F}_n(t) = \prod_{i : X_i \le t} \frac{\#(j : X_j \ge X_i) - \Delta_i}{\#(j : X_j \ge X_i)} = \prod_{i : X_{(i)} \le t} \left(\frac{n-i}{n-i+1} \right)^{\Delta_{(i)}}.$$

This estimator sequence is asymptotically normal by the Hadamard differentiability of the product integral. □

Notes

A calculus of "differentiable statistical functions" was proposed by von Mises [104]. Von Mises considered functions $\phi(\mathbb{F}_n)$ of the empirical distribution function (which he calls the "repartition of the real quantities $x_1, \ldots, x_n$") as in the first section of this chapter. Following Volterra he calls ϕ m times differentiable at F if the first m derivatives of the map $t \mapsto \phi(F + tH)$ at $t = 0$ exist and have representations of the form

$$\phi_F^{(k)}(H) = \int \cdots \int \psi(x_1, \ldots, x_k)\, dH(x_1) \cdots dH(x_k).$$

This representation is motivated in analogy with the finite-dimensional case, in which H would be a vector and the integrals sums. From the perspective of our section on Hadamard-differentiable functions, the representation is somewhat arbitrary, because it is required that a derivative be continuous, whence its general form depends on the norm that we use on the domain of ϕ. Furthermore, the Volterra representation cannot be directly applied to, for instance, a limiting Brownian bridge, which is not of bounded variation.

Von Mises' treatment is not at all informal, as is the first section of this chapter. After developing moment bounds on the derivatives, he shows that $n^{m/2}\big(\phi(\mathbb{F}_n) - \phi(F)\big)$ is asymptotically equivalent to $\phi_F^{(m)}(\mathbb{G}_n)$ if the first $m-1$ derivatives vanish at F and the $(m+1)$th derivative is sufficiently regular. He refers to the approximating variables $\phi_F^{(m)}(\mathbb{G}_n)$, degenerate V-statistics, as "quantics" and derives the asymptotic distribution of quantics of degree 2, first for discrete observations and next in general by discrete approximation. Hoeffding's work on U-statistics, which was published one year later, had a similar aim of approximating complicated statistics by simpler ones but did not consider degenerate U-statistics.

The systematic application of Hamadard differentiability in statistics appears to have first been put forward in the (unpublished) thesis [125] of J Reeds and had a main focus on robust functions. It was revived by Gill [53] with applications in survival analysis in mind. With a growing number of functional estimators available (beyond the empirical distribution and product-limit estimator), the delta method is a simple but useful tool to standardize asymptotic normality proofs.

Our treatment allows the domain $\mathbb{D}_\phi$ of the map ϕ to be arbitrary. In particular, we do not assume that it is open, as we did, for simplicity, when discussing the Delta method for

Euclidean spaces. This is convenient, because many functions of statistical interest, such as zeros, inverses or integrals, are defined only on irregularly shaped subsets of a normed space, which, besides a linear space, should be chosen big enough to support the limit distribution of T_n.

PROBLEMS

1. Let $\phi(P) = \int\int h(u, v)\, dP(u)\, dP(v)$ for a fixed given function h. The corresponding estimator $\phi(\mathbb{P}_n)$ is known as a *V-statistic*. Find the influence function.

2. Find the influence function of the function $\phi(F) = \int a(F_1 + F_2)\, dF_2$ if F_1 and F_2 are the marginals of the bivariate distribution function F, and a is a fixed, smooth function. Write out $\phi(\mathbb{F}_n)$. What asymptotic variance do you expect?

3. Find the influence function of the map $F \mapsto \int_{[0,t]} (1 - F_-)^{-1}\, dF$ (the cumulative hazard function).

4. Show that a map $\phi : \mathbb{D} \mapsto \mathbb{E}$ is Hadamard differentiable at a point θ if and only if for every compact set $K \subset \mathbb{D}$ the expression in (20.7) converges to zero uniformly in $h \in K$ as $t \to 0$.

5. Show that the symmetrization map $(\theta, F) \mapsto \frac{1}{2}\big(F(t) + 1 - F(2\theta - t)\big)$ is (tangentially) Hadamard differentiable under appropriate conditions.

6. Let $g : [a, b] \mapsto \mathbb{R}$ be a continuously differentiable function. Show that the map $z \mapsto g \circ z$ with domain the functions $z : T \mapsto [a, b]$ contained in $\ell^\infty(T)$ is Hadamard differentiable. What does this imply for the function $z \mapsto 1/z$?

7. Show that the map $F \mapsto \int_{[a,b]} s\, dF(s)$ is Hadamard differentiable from the domain of all distribution functions to $\mathbb{R}$, for each pair of finite numbers a and b. View the distribution functions as a subset of $D[-\infty, \infty]$ equipped with supremum norm. What if a or b are infinite?

8. Find the first- and second-order derivative of the function $\phi(F) = \int (F - F_0)^2\, dF$ at $F = F_0$. What limit distribution do you expect for $\phi(\mathbb{F}_n)$?