Chapter 7
Statistical Functionals and the Delta Method

1. **Estimators as Functionals of $\mathbb{F}_n$ or $\mathbb{P}_n$**
2. **Continuity of Functionals of F or P**
3. **Metrics for Distribution Functions F and Probability Distributions P**
4. **Differentiability of Functionals of F or P: Gateaux, Hadamard, and Frechet Derivatives**
5. **Higher Order Derivatives**

Chapter 7

Statistical Functionals and the Delta Method

1 Estimates as Functionals of $\mathbb{F}_n$ or $\mathbb{P}_n$

Often the quantity we want to estimate can be viewed as a functional $T(F)$ or $T(P)$ of the underlying distribution function F or P generating the data. Then a simple nonparametric estimator is simply $T(\mathbb{F}_n)$ or $T(\mathbb{P}_n)$ where $\mathbb{F}_n$ and $\mathbb{P}_n$ denote the empirical distribution function and empirical measure of the data.

Notation. Suppose that $X_1, \ldots, X_n$ are i.i.d. P on $(\mathcal{X}, \mathcal{A})$. We let

$$\mathbb{P}_n \equiv \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \equiv \text{ the empirical measure of the sample,}$$

where $\delta_x \equiv$ the measure with mass one at x (so $\delta_x(A) = 1_A(x)$ for $A \in \mathcal{A}$. When $\mathcal{X} = \mathbb{R}^k$, especially when $k = 1$, we will write

$$\mathbb{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i) = \mathbb{P}_n(-\infty, x], \qquad F(x) = P(-\infty, x].$$

Here is a list of examples.

Example 1.1 The mean $T(F) = \int x dF(x)$. $T(\mathbb{F}_n) = \int x d\mathbb{F}_n(x)$.

Example 1.2 The r-th moment: for r an integer, $T(F) = \int x^r dF(x)$, and $T(\mathbb{F}_n) = \int x^r d\mathbb{F}_n(x)$.

Example 1.3 The variance:

$$T(F) = Var_F(X) = \int (x - \int x dF(x))^2 dF(x) = \frac{1}{2}\int\int (x-y)^2 dF(x) dF(y),$$
$$T(\mathbb{F}_n) = Var_{\mathbb{F}_n}(X) = \int (x - \int x d\mathbb{F}_n(x))^2 d\mathbb{F}_n(x) = \frac{1}{2}\int\int (x-y)^2 d\mathbb{F}_n(x) d\mathbb{F}_n(y).$$

Example 1.4 The median: $T(F) = F^{-1}(1/2)$. $T(\mathbb{F}_n) = \mathbb{F}_n^{-1}(1/2)$.

Example 1.5 The $\alpha-$trimmed mean: $T(F) = (1-2\alpha)^{-1}\int_\alpha^{1-\alpha} F^{-1}(u)du$ for $0 < \alpha < 1/2$. $T(\mathbb{F}_n) = (1-2\alpha)^{-1}\int_\alpha^{1-\alpha} \mathbb{F}_n^{-1}(u)du$.

Example 1.6 The Hodges-Lehmann functional: $T(F) = (1/2)\{F \star F\}^{-1}(1/2)$ where $\star$ denotes convolution. Then $T(\mathbb{F}_n) = (1/2)\{\mathbb{F}_n \star \mathbb{F}_n\}^{-1}(1/2) = \text{median}\{(X_i + X_j)/2\}$.

Example 1.7 The Mann-Whitney functional. For X, Y independent with distribution functions F and G respectively, $T(F, G) = \int F dG = P_{F,G}(X \leq Y)$. Then $T(\mathbb{F}_m, \mathbb{G}_n) = \int \mathbb{F}_m d\mathbb{G}_n$ (based on two independent samples $X_1, \ldots, X_m$ i.i.d. F with empirical df $\mathbb{F}_m$ and $Y_1, \ldots, Y_n$ i.i.d. G with empirical df $\mathbb{G}_n$.

Example 1.8 Multivariate mean: for P on $(\mathbb{R}^k, \mathcal{B}^k)$: $T(P) = \int x dP(x)$ (with values in $\mathbb{R}^k$), $T(\mathbb{P}_n) = \int x d\mathbb{P}_n(x) = n^{-1} \sum_{i=1}^n X_i$.

Example 1.9 Multivariate cross second moments: for P on $(\mathbb{R}^k, \mathcal{B}^k)$:

$$T(P) = \int x x^T dP(x) = \int x^{\otimes 2} dP(x);$$
$$T(\mathbb{P}_n) = \int x x^T d\mathbb{P}_n(x) = \int x^{\otimes 2} d\mathbb{P}_n(x) = n^{-1} \sum_{i=1}^n X_i X_i^T.$$

Note that $T(P)$ and $T(\mathbb{P}_n)$ take values in $\mathbb{R}^{k \times k}$.

Example 1.10 Multivariate covariance matrix: for P on $(\mathbb{R}^k, \mathcal{B}^k)$:

$$T(P) = \int (x - \int y dP(y))(x - \int y dP(y))^T dP(x) = \frac{1}{2} \int \int (x - y)(x - y)^T dP(x) dP(y),$$
$$T(\mathbb{P}_n) = \int (x - \int y d\mathbb{P}_n(y))(x - \int y d\mathbb{P}_n(y))^T d\mathbb{P}_n(x)$$
$$= \frac{1}{2} \int \int (x - y)(x - y)^T d\mathbb{P}_n(x) d\mathbb{P}_n(y) = n^{-1} \sum_{i=1}^n (X_i - \overline{X}_n)(X_i - \overline{X}_n)^T.$$

Example 1.11 $k-$means clustering functional: $T(P) = (T_1(P), \ldots, T_k(P)$ where the $T_i(P)$'s minimize

$$\int |x - t_1|^2 \wedge \cdots \wedge |x - t_k|^2 dP(x) = \sum_{i=1}^k \int_{C_i} |x - t_i|^2 dP(x)$$

where

$$C_i = \{x \in \mathbb{R}^m : \ t_i \ \text{ minimizes } \ |x - t|^2 \ \text{ over } \ \{t_1, \ldots, t_k\}\}.$$

Then $T(\mathbb{P}_n) = (T_1(\mathbb{P}_n), \ldots, T_k(\mathbb{P}_n))$ where the $T_i(\mathbb{P}_n)$'s minimize

$$\int |x - t_1|^2 \wedge \cdots \wedge |x - t_k|^2 d\mathbb{P}_n(x).$$

Example 1.12 The simplicial depth function: for P on $\mathbb{R}^k$ and $x \in \mathbb{R}^k$, set $T(P) \equiv T(P)(x) = Pr_P(x \in S(X_1, \ldots, X_{k+1}))$ where $X_1, \ldots, X_{k+1}$ are i.i.d. P and $S(x_1, \ldots, x_{k+1})$ is the simplex in $\mathbb{R}^k$ determined by $x_1, \ldots, x_{k+1}$; e.g. for $k = 2$, the simplex determined by x_1, x_2, x_3 is just a triangle. Then $T(\mathbb{P}_n) = Pr_{\mathbb{P}_n}(x \in S(X_1, \ldots, X_{k+1}))$. Note that in this example $T(P)$ is a function from $\mathbb{R}^k$ to $\mathbb{R}$.

Example 1.13 (Z-functional derived from likelihood). A maximum likelihood estimator: for P on $(\mathcal{X}, \mathcal{A})$, suppose that $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ is a regular parametric model with vector scores function $\dot{\mathbf{l}}_\theta(\cdot;\theta)$. Then for general P, not necessarily in the model $\mathcal{P}$, consider T defined by

$$\int \dot{\mathbf{l}}_\theta(x; T(P))dP(x) = 0. \tag{1}$$

Then

$$\int \dot{\mathbf{l}}_\theta(x; T(\mathbb{P}_n))d\mathbb{P}_n(x) = 0$$

defines $T(\mathbb{P}_n)$. For estimation of location in one dimension with $\dot{\mathbf{l}}(x;\theta) = \psi(x-\theta)$ and $\psi \equiv -f'/f$, these become

$$\int \psi(x - T(F))dF(x) = 0 \qquad \text{and} \qquad \int \psi(x - T(\mathbb{F}_n))d\mathbb{F}_n(x) = 0.$$

We expect that often the value $T(P) \in \Theta$ satisfying (1) also satisfies

$$T(P) = \text{argmin}_{\theta\in\Theta} K(P, P_\theta).$$

Here is a heuristic argument showing why this should be true: Note that for many cases we have

$$\begin{aligned} \hat{\theta}_n &= \text{argmax}_\theta n^{-1} l_n(\theta) = \text{argmax}_\theta \mathbb{P}_n(\log p_\theta) \\ &\to_p \text{argmax}_\theta P(\log\theta) = \text{argmax}_\theta \int \log p_\theta(x) dP(x). \end{aligned}$$

Now

$$\begin{aligned} P(\log p_\theta) &= P(\log p) + P\log\left(\frac{p_\theta}{p}\right) \\ &= P(\log p) - P\log\left(\frac{p}{p_\theta}\right) \\ &= P(\log p) - K(P, P_\theta). \end{aligned}$$

Thus

$$\text{argmax}_\theta \int \log p_\theta(x) dP(x) = \text{argmin}_\theta K(P, P_\theta) \equiv \theta(P).$$

If we can interchange differentiation and integration it follows that

$$\nabla_\theta K(P, P_\theta) = \int p(x)\dot{\mathbf{l}}_\theta(x;\theta)d\mu(x) = \int \dot{\mathbf{l}}_\theta(x;\theta)dP(x),$$

so the relation (1) is obtained by setting this gradient vector equal to 0.

Example 1.14 A bootstrap functional: let $T(F)$ be a functional with estimator $T(\mathbb{F}_n)$, and consider estimating the distribution function of $\sqrt{n}(T(\mathbb{F}_n) - T(F))$,

$$H_n(F;\cdot) = P_F(\sqrt{n}(T(\mathbb{F}_n) - T(F)) \le \cdot).$$

A natural estimator is $H_n(\mathbb{F}_n, \cdot)$.

2 Continuity of Functionals of F or P

One of the basic properties of a functional T is continuity (or lack thereof). One important sense in which we might want our functionals T to be continuous is in the sense of weak convergence.

Definition 2.1 A. $T : \mathcal{F} \to \mathbb{R}$ is *weakly continuous* at F_0 if $F_n \Rightarrow F_0$ implies $T(F_n) \to T(F_0)$. $T : \mathcal{F} \to \mathbb{R}$ is *weakly lower-semicontinuous* at F_0 if $F_n \Rightarrow F_0$ implies $\liminf_{n\to\infty} T(F_n) \geq T(F_0)$.
B. $T : \mathcal{P} \to \mathbb{R}$ is *weakly continuous* at $P_0 \in \mathcal{P}$ if $P_n \Rightarrow P_0$ implies $T(P_n) \to T(P_0)$.

Example 2.1 $T(F) = \int x dF(x)$ is weakly discontinuous at every F_0: if $F_n = (1-n^{-1})F_0 + n^{-1}\delta_{a_n}$, then $F_n \Rightarrow F_0$ since, for bounded ψ

$$\int \psi dF_n = (1-n^{-1}) \int \psi dF_0 + n^{-1}\psi(a_n) \to \int \psi dF_0.$$

But

$$T(F_n) = (1-n^{-1})T(F_0) + n^{-1}a_n \to \infty$$

if we choose a_n so that $n^{-1}a_n \to \infty$.

Example 2.2 $T(F) = (1-2\alpha)^{-1}\int_\alpha^{1-\alpha} F^{-1}(u)du$ with $0 < \alpha < 1/2$ is continuous at every F_0: $F_n \Rightarrow F_0$ implies that $F_n^{-1}(t) \to F_0^{-1}(t)$ a.e. Lebesgue. Hence

$$\begin{aligned} T(F_n) &= (1-2\alpha)^{-1} \int_\alpha^{1-\alpha} F_n^{-1}(u)du \\ &\to (1-2\alpha)^{-1} \int_\alpha^{1-\alpha} F_0^{-1}(u)du = T(F_0) \end{aligned}$$

by the dominated convergence theorem.

Example 2.3 $T(F) = F^{-1}(1/2)$ is continuous at every F_0 such that F_0^{-1} is continuous at $1/2$.

Example 2.4 (A lower-semicontinuous functional T). Let

$$T(F) = Var_F(X) = \int (x - E_F X)^2 dF(x) = \frac{1}{2} E_F(X - X')^2$$

where $X, X' \sim F$ are independent; recall example 1.3. If $F_n \to_d F$, then $\liminf_{n\to\infty} T(F_n) \geq T(F)$; this follows from Skorokhod and Fatou.

Here is the basic fact about empirical measures that makes weak continuity of a functional T useful:

Theorem 2.1 (Varadarajan). If $X_1, \ldots, X_n$ are i.i.d. P on a separable metric space (S, d), then $Pr(\mathbb{P}_n \Rightarrow P) = 1$.

Proof. For each fixed bounded and continuous function ψ we have

$$\mathbb{P}_n\psi \equiv \int \psi d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \psi(X_i) \to_{a.s.} P\psi \equiv \int \psi dP$$

by the ordinary strong law of large numbers. The proof is completed by noting that the collection of bounded continuous functions on a separable metric space (S, d) is itself separable. See Dudley (1989, 2002), sections 11.2 and 11.4. □

Combining Varadarajan's theorem with weak continuity of T yields the following simple result.

Proposition 2.1 Suppose that:
(i) $(\mathcal{X}, \mathcal{A}) = (S, \mathcal{B}_{Borel})$ where (S, d) is a separable metric space and $\mathcal{B}_{Borel}$ denotes its usual Borel sigma - field.
(ii) $T : \mathcal{P} \to \mathbb{R}$ is weakly continuous at $P_0 \in \mathcal{P}$.
(iii) $X_1, \ldots, X_n$ are i.i.d. P_0.
Then $T_n \equiv T(\mathbb{P}_n) \to_{a.s.} T(P_0)$.

Proof. By Varadarajan's theorem 2.1, $\mathbb{P}_n \Rightarrow P_0$ a.s. Fix $\omega \in A$ with $Pr(A) = 1$ so that $\mathbb{P}_n^\omega \Rightarrow P_0$. Then by weak continuity of T, $T_n(\mathbb{P}_n^\omega) \to T(P_0)$. □

A difficulty in using this theorem is typically in trying to verify weak-continuity of T. Weak continuity is a rather strong hypothesis, and many interesting functions fail to have this type of continuity. The following approach is often useful.

Definition 2.2 Let $\mathcal{F} \subset L_1(P)$ be a collection of integrable functions. Say that $P_n \to P$ with respect to $\|\cdot\|_\mathcal{F}$ if $\|P_n - P\|_\mathcal{F} \equiv \sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \to 0$. Furthermore, we say that $T : \mathcal{P} \to \mathbb{R}$ is continuous with respect to $\|\cdot\|_\mathcal{F}$ if $\|P_n - P\|_\mathcal{F} \to 0$ implies that $T(P_n) \to T(P)$.

Definition 2.3 If $\mathcal{F} \subset L_1(P)$ is a collection of integrable functions with $\|\mathbb{P}_n - P\|_\mathcal{F}^* \to 0$, we then say that $\mathcal{F}$ is a Glivenko-Cantelli class for P and write $\mathcal{F} \in GC(P)$.

Theorem 2.2 Suppose that:
(i) $\mathcal{F} \in GC(P)$; i.e. $\|\mathbb{P}_n - P\|_\mathcal{F}^* \to_{a.s.} 0$.
(ii) T is continuous with respect to $\|\cdot\|_\mathcal{F}$.
Then $T(\mathbb{P}_n) \to_{a.s.} T(P)$.

3 Metrics Probability Distributions F and P

We have already encountered the total variation and Hellinger metrics in the course of studying Scheffé's lemma, Bayes estimators, and tests of hypotheses. As we will see, as useful as these metrics are, they are too strong: the empirical measure $\mathbb{P}_n$ fails to converge to the true P in either the total variation or Hellinger distance in general. In fact this fails to hold in general for the Prohorov and dual bounded Lipschitz metrics which we introduce below, as has been shown by Dudley (1969), Kersting (1978), and Bretagnolle and Huber-Carol (1977); also see the remarks in Huber (1981), page 39. Nonetheless, it will be helpful to have in mind some some useful metrics for probability measures P and df's F, and their properties.

Definition 3.1 The *Kolmogorov* or *supremum* metric between two distribution functions F and G is

$$d_K(F,G) \equiv \|F-G\|_\infty \equiv \sup_{x\in\mathbb{R}^k} |F(x)-G(x)|.$$

Definition 3.2 The *Lévy metric* between two distribution functions F and G is

$$d_L(F,G) \equiv \inf\{\epsilon>0:\ G(x-\epsilon)-\epsilon \le F(x) \le G(x+\epsilon)+\epsilon \ \text{ for all } \ x\in\mathbb{R}\}.$$

Definition 3.3 The *Prohorov metric* between two probability measures P, Q on a metric space (S,d) is

$$d_{pr}(P,Q) = \inf\{\epsilon>0:\ P(B)\le Q(B^\epsilon)+\epsilon \ \text{ for all Borel sets } \ B\}$$

where $B^\epsilon \equiv \{x:\ \inf_{y\in B} d(x,y)\le\epsilon\}$.

To define the next metric for P, Q on a metric space (S,d) for any real-valued function f on S, set $\|f\|_L \equiv \sup_{x\neq y}|f(x)-f(y)|/d(x,y)$, and denote the usual supremum norm by $\|f\|_\infty \equiv \sup_x |f(x)|$. Finally, set $\|f\|_{BL} \equiv \|f\|_L + \|f\|_\infty$.

Definition 3.4 The *dual - bounded Lipschitz metric* d_{BL^*} is defined by

$$d_{BL^*}(P,Q) \equiv \sup\{|\int f dP - \int f dQ|:\ \ \|f\|_{BL}\le 1\}.$$

Definition 3.5 The *total variation metric* d_{TV} is defined by

$$d_{TV}(P,Q) \equiv \sup\{|P(A)-Q(A)|:\ \ A\in\mathcal{A}\} = \frac{1}{2}\int |p-q|d\mu$$

where $p\equiv dP/d\mu$, $q=dQ/d\mu$ for some measure μ dominating both P and Q (e.g. $\mu=P+Q$).

Definition 3.6 The *Hellinger metric* H is defined by

$$H^2(P,Q) = \frac{1}{2}\int\{\sqrt{p}-\sqrt{q}\}^2 d\mu = 1-\int\sqrt{pq}d\mu \equiv 1-\rho(P,Q)$$

where μ is any measure dominating both P and Q. The quantity $\rho(P,Q)\equiv\int\sqrt{pq}d\mu$ is called the *affinity* between P and Q.

The following basic theorem establishes relationships between these metrics:

Theorem 3.1 A. $d_{Pr}(P,Q)^2 \leq d_{BL^*}(P,Q) \leq 2d_{Pr}(P,Q)$.
B. $H^2(P,Q) \leq d_{TV}(P,Q) \leq H(P,Q)\{2 - H^2(P,Q)\}^{1/2}$.
C. $d_{Pr}(P,Q) \leq d_{TV}(P,Q)$.
D. For distributions P, Q on the real line, $d_L \leq d_K \leq d_{TV}$.

Proof. We proved B in chapter 2. For A, see Dudley (1989) section 11.3, problem 5, and section 11.6, corollary 11.6.5. Also see Huber (1981), corollary 2.4.3, page 33. Another useful reference is Whitt (1974). □

Theorem 3.2 (Strassen). The following are equivalent:
(a) $d_{Pr}(P,Q) \leq \epsilon$.
(b) There exist $X \sim P$, $Y \sim Q$ defined on a common probability space $(\Omega, \mathcal{F}, Pr)$ such that $Pr(d(X,Y) \leq \epsilon) \geq 1 - \epsilon$.

Proof. (b) implies (a) is easy: for any Borel set B,

$$\begin{aligned}[X \in B] &= [X \in B, d(X,Y) \leq \epsilon] \cup [X \in B, d(X,Y) > \epsilon] \\ &\subset [X \in B^\epsilon] \cup [d(X,Y) > \epsilon],\end{aligned}$$

so that $P(B) \leq Q(B^\epsilon) + \epsilon$.

For the proof of (a) implies (b) see Strassen (1965), Dudley (1968), or Schay (1974). A nice treatment of Strassen's theorem is given by Dudley (1989, 2002). □

4 Differentiability of Functionals T of F or P

To be able to prove more than consistency, we will need stronger properties of the functional T, namely differentiability.

Definition 4.1 T is *Gateaux differentiable* at F if there exists a linear functional $\dot{T}(F;\cdot)$ such that for $F_t = (1-t)F + tG$,

$$\begin{aligned}\lim_{t\to 0}\frac{T(F_t)-T(F)}{t} &= \dot{T}(F;G-F) = \int \psi(x)d(G(x)-F(x)) \\ &= \int \psi_F(x)dG(x)\end{aligned}$$

where $\psi_F(x) = \psi - \int \psi dF(x)$ has mean zero under F. Or, $T : \mathcal{P} \to \mathbb{R}$ is Gateaux - differentiable at P if there exists $\dot{T}(P;\cdot)$ bounded and linear such that for $P_t \equiv (1-t)P + tQ$

$$\begin{aligned}\lim_{t\to 0}\frac{T(P_t)-T(P)}{t} &= \dot{T}(P;Q-P) = \int \psi(x)d(Q(x)-P(x)) \\ &= \int \psi_P(x)dQ(x).\end{aligned}$$

Definition 4.2 T has the *influence function* or *influence curve* $IC(x;T,F)$ at F if, with $F_t \equiv (1-t)F + t\delta_x$,

$$\lim_{t\to 0}\frac{T(F_t)-T(F)}{t} = IC(x;T,F) = \psi_F(x).$$

Example 4.1 Probability of a set: suppose that $T(F) = F(A)$ for a fixed measurable set A. Then

$$\frac{T(F_t)-T(F)}{t} = \int\{1_A(x) - \int 1_A(y)dF(y)\}dG(x) = \int \psi_F(x)dG(x)$$

where $\psi_F(x) = 1_A(x) - F(A)$.

Example 4.2 The mean: $T(F) = \int xdF(x)$. Then

$$\frac{T(F_t)-T(F)}{t} = \int\{x - T(F)\}dG(x) = \int \psi_F(x)dG(x)$$

where $\psi_F(x) = x - T(F)$. Note that the influence function $\psi_F(x)$ for the probability functional is *bounded*, but that the influence function $\psi_F(x)$ for the mean functional is *unbounded*.

Example 4.3 The variance: $T(F) = Var_F(X) = \int(x-\mu(F))^2dF(x)$. Now

$$\begin{aligned}\frac{d}{dt}T(F_t)|_{t=0} &= \frac{d}{dt}\int(x-\mu(F_t))^2dF_t(x) \\ &= \int(x-\mu(F))^2d(G-F)(x) + 2\int(x-\mu(F))(-1)\dot{\mu}(F;G-F)dF(x) \\ &= \int(x-\mu(F))^2d(G-F) \\ &= \int\{(x-\mu(F))^2 - \sigma_F^2\}dG(x).\end{aligned}$$

Hence $IC(x;\ T,F) = \psi_F(x) = (x-\mu(F))^2 - \sigma_F^2$.

Example 4.4 $T(F) = F^{-1}(1/2)$, and suppose that F has density f which is positive at $F^{-1}(1/2)$. Then, with F_t $(1-t)F + tG$,

$$\frac{d}{dt}T(F_t)\Big|_{t=0} = \frac{d}{dt}F_t^{-1}(1/2)\Big|_{t=0}.$$

Note that $F_t(F_t^{-1}(1/2)) = 1/2$, and hence

$$\begin{aligned} 0 &= \frac{d}{dt}F_t(F_t^{-1}(1/2))\Big|_{t=0} \\ &= \frac{d}{dt}\{F(F_t^{-1}(1/2)) + t(G-F)(F_t^{-1}(1/2))\}\Big|_{t=0} \\ &= f(F^{-1}(1/2))\dot{T}(F; G-F) + (G-F)(F^{-1}(1/2)) + 0, \end{aligned}$$

so that

$$\begin{aligned} \dot{T}(F; G-F) &= -\frac{(G-F)(F^{-1}(1/2))}{f(F^{-1}(1/2))} \\ &= -\frac{\int(1_{(-\infty,F^{-1}(1/2)]}(x) - 1/2)dG(x)}{f(F^{-1}(1/2))}. \end{aligned}$$

Hence

$$\psi_F(x) = IC(x; T, F) = -\frac{1}{f(F^{-1}(1/2))}\{1_{(-\infty,F^{-1}(1/2)]}(x) - 1/2\}.$$

Example 4.5 The $p-$th quantile, $T(F) = F^{-1}(p)$. By a calculation similar to that for the median,

$$\begin{aligned} \dot{T}(F; G-F) &= -\frac{(G-F)(F^{-1}(p))}{f(F^{-1}(p))} \\ &= -\frac{\int(1_{(-\infty,F^{-1}(p)]}(x) - p)dG(x)}{f(F^{-1}(p))}. \end{aligned}$$

and

$$\psi_F(x) = IC(x; T, F) = -\frac{1}{f(F^{-1}(p)}\{1_{(-\infty,F^{-1}(p)]}(x) - p\}.$$

Now we need to consider other types of derivatives: in particular the stronger notions of derivative which we will discuss below are those of Fréchet and Hadamard derivatives.

Definition 4.3 A functional $T : \mathcal{F} \to \mathbb{R}$ is Fréchet - differentiable at $F \in \mathcal{F}$ with respect to d_* if there exists a continuous linear functional $\dot{T}(F; \cdot)$ from finite signed measures in $\mathbb{R}$ such that

$$(1) \qquad \frac{|T(G) - T(F) - \dot{T}(F; G-F)|}{d_*(G, F)} \to 0 \qquad \text{as} \quad d_*(F, G) \to 0.$$

Here are some properties of Fréchet - differentiation:

Theorem 4.1 Suppose that d_* is a metric for weak convergence (i.e. the Lévy metric for df's on the line; or the Prohorov or dual-bounded Lipschitz metric for measures on a metric space (S, d)). Then:

A. If $\dot{T}$ exists in the Fréchet sense, then it is unique, and T is Gateaux differentiable with Gateaux derivative $\dot{T}$.

B. If T is Fréchet differentiable at F, then T is continuous at F.

C. $\dot{T}(F; G-F) = \int \psi d(G-F) = \int (\psi - \int \psi dF) dG$ where the function ψ is bounded and continuous.

Proof. See Huber (1981), proposition 5.1, page 37. □

Fréchet differentiability leads to an easy proof of asymptotic normality if the metric d_* is "compatible with the empirical df or empirical measure".

Theorem 4.2 Suppose that T is Fréchet differentiable at F with respect to d_* and that

(2) $\quad \sqrt{n} d_*(\mathbb{F}_n, F) = O_p(1).$

Then

$$\begin{aligned} \sqrt{n}(T(\mathbb{F}_n) - T(F)) &= \int \psi_F d\{\sqrt{n}(\mathbb{F}_n - F)\} + o_p(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_F(X_i) + o_p(1) \\ &\to_d N(0, E\psi_F^2(X)). \end{aligned}$$

Proof. By Fréchet differentiability of T at F,

$$\begin{aligned} \sqrt{n}(T(\mathbb{F}_n) - T(F)) &= \sqrt{n} \int \psi_F d\mathbb{F}_n + \sqrt{n} o(d_*(\mathbb{F}_n, F)) \\ &= \sqrt{n} \int \psi_F d\mathbb{F}_n + \frac{o(d_*(\mathbb{F}_n, F))}{d_*(\mathbb{F}_n, F)} \sqrt{n} d_*(\mathbb{F}_n, F) \\ &= \sqrt{n} \int \psi_F d\mathbb{F}_n + o(1) O_p(1) \end{aligned}$$

by (2) and (1). □

Note that if d_* is the Lévy metric d_L or the Kolmogorov metric d_K on the line, then (2) is satisfied:

$$\sqrt{n} d_L(\mathbb{F}_n, F) \le \sqrt{n} d_K(\mathbb{F}_n, F) = \sqrt{n} \|\mathbb{F}_n - F\|_\infty \stackrel{d}{=} \|\mathbb{U}_n(F)\|_\infty \to_d \|\mathbb{U}(F)\|_\infty.$$

Unfortunately, if $d_* = d_{Pr}$ or $d_* = d_{BL}$, then $\sqrt{n} d_*(\mathbb{F}_n, F)$ is *not* $O_p(1)$ in general; see Dudley (1969), Kersting (1978), and Huber-Carol (1977). Thus we are lead to consideration of other metrics such as the Kolmogorov metric and generalizations thereof for problems concerning functionals $T(P)$ of probability distributions P. While some functionals T are Fréchet differentiable with respect to the supremum or Kolmogorov metric, we can make more functionals differentiable by considering a somewhat weaker notion of differentiability as follows:

Definition 4.4 A functional $T : \mathcal{F} \to \mathbb{R}$ is *Hadamard differentiable* at F with respect to the Kolmogorov distance $d_K = \|\cdot\|_\infty$ (or *compactly differentiable* with respect to d_K) if there exists $\dot{T}(F;\cdot)$ continuous and linear satisfying

$$\frac{|T(F_t) - T(F) - \dot{T}(F; F_t - F)|}{|t|} = o(1)$$

for all $\{F_t\}$ satisfying $\|t^{-1}(F_t - F) - \Delta\|_\infty \to 0$ for some function Δ.

The motivation for this definition is simply that we can write

$$\sqrt{n}(T(\mathbb{F}_n) - T(F)) = \frac{T(F + n^{-1/2}n^{1/2}(\mathbb{F}_n - F)) - T(F)}{n^{-1/2}}$$

where $\sqrt{n}(\mathbb{F}_n - F) \stackrel{d}{=} \mathbb{U}_n(F) \Rightarrow \mathbb{U}(F)$. Hence we can easily deduce the following theorem.

Theorem 4.3 Suppose that $T : \mathcal{F} \to \mathbb{R}$ is Hadamard - differentiable at F with respect to $\|\cdot\|_\infty$. Then

$$\sqrt{n}(T(\mathbb{F}_n) - T(F)) \to_d N(0, E(\dot{T}^2(F; 1_{(-\infty,\cdot]}(X) - F))).$$

Moreover,

$$\sqrt{n}(T(\mathbb{F}_n) - T(F)) - \dot{T}(F; \sqrt{n}(\mathbb{F}_n - F)) = o_p(1).$$

Proof. This is easily proved using a Skorokhod construction of the empirical process, or by the extended continuous mapping theorem. Gill (1989) used the Skorokhod approach; Wellner (1989) pointed out the extended continuous mapping proof. □

One way of treating all the kinds of differentiability we have discussed so far is as follows. Define

$$T(F_t) - T(F) - \dot{T}(F; F_t - F) \equiv \mathrm{Rem}(F + th);$$

Here $h = t^{-1}(F_t - F)$. Let $\mathcal{S}$ be a collection of subsets of the metric space $(\mathcal{F}, d_*)$. Then T is $\mathcal{S}-$differentiable at F with derivative $\dot{T}$ if for all $S \in \mathcal{S}$

$$\frac{Rem(F + th)}{t} \to 0 \qquad \text{as} \;\; t \to 0 \;\; \text{uniformly in} \;\; h \in S.$$

Now different choices of $\mathcal{S}$ yield different degrees of "goodness" of the linear approximation of T by $\dot{T}$ at F. The three most common choices are just those we have discussed:

A. When $\mathcal{S} = \{\text{all singletons of } (\mathcal{F}, d_*)\}$, T is called Gateaux or directionally differentiable.

B. When $\mathcal{S} = \{\text{all compact subsets of } (\mathcal{F}, d_*)\}$, T is called Hadamard or compactly differentiable.

C. When $\mathcal{S} = \{\text{all bounded subsets of } \mathcal{F}, d_*)\}$, T is called Fréchet (or boundedly) differentiable.

Here is a simple example of a function T defined on pairs of probability distributions (or, in this case, distribution functions) which is compactly differentiable with respect to the familiar supremum (or uniform or Kolomogorov) norm, but which is *not* Fréchet differentiable with respect to this norm.

Example 4.6 For distribution functions F, G on $\mathbb{R}$, define T by

$$T(F,G) = \int F dG = P(X \le Y)$$

where $X \sim F$, $Y \sim G$ are independent. Let $\|\tilde{F} - F\|_\infty \equiv \sup_x |\tilde{F}(x) - F(x)| \equiv \|\tilde{F} - F\|_{\mathcal{F}}$ where $\mathcal{F} = \{1_{(-\infty,t]} : t \in \mathbb{R}\}$.

Proposition 4.1 A. $T(F,G)$ is Hadamard differentiable with respect to $\|\cdot\|_\infty$ at every pair of df's (F,G) with derivative $\dot{T}$ given by

$$\dot{T}((F,G);\alpha,\beta) = \int \alpha dG - \int \beta dF. \tag{3}$$

B. $T(F,G)$ is *not* Fréchet differentiable with respect to $\|\cdot\|_\infty$.

Proof. The following proof of A is basically Gill's (1989), lemma 3. For $F_t \to F$ and $G_t \to G$, define $\alpha_t \equiv (F_t - F)/t$ and $\beta_t \equiv (G_t - G)/t$; for Hadamard differentiability we have $\alpha_t \to \alpha$ and $\beta_t \to \beta$ with resepct to $\|\cdot\|_\infty$ for some (bounded) functions α and β. Now

$$\begin{aligned} \frac{T(F_t,G_t) - T(F,G)}{t} - \dot{T}(\alpha_t,\beta_t) &= t\int \alpha_t d\beta_t \\ &= \int \alpha d(G_t - G) + \int (\alpha_t - \alpha) d(G_t - G). \end{aligned}$$

Since $\dot{T}$ is continuous, it suffices to show that the right side converges to 0. The second term on the right is bounded by

$$\|\alpha_t - \alpha\|_\infty \left\{ \int dG_t + \int dG \right\} \le 2\|\alpha_t - \alpha\|_\infty \to 0.$$

Fix $\epsilon > 0$. Since the limit function α in the first term is right-continuous with left limits, there is a step function with a finite number m of jumps, $\tilde{\alpha}$ say, which satisfies $\|\alpha - \tilde{\alpha}\|_\infty < \epsilon$. Thus the first term may be bounded as follows:

$$\begin{aligned} |\int \alpha d(G_t - G)| &\le |\int (\alpha - \tilde{\alpha}) d(G_t - G)| + |\int \tilde{\alpha} d(G_t - G)| \\ &\le 2\|\alpha - \tilde{\alpha}\|_\infty + \sum_{j=1}^{m} |\tilde{\alpha}(x_{j-1})| |(G_t - G)[x_{j-1}, x_j)| \\ &\le 2\epsilon + 2m\|\tilde{\alpha}\|_\infty \|G_t - G\|_\infty \to 2\epsilon. \end{aligned}$$

Since ϵ is arbitrary, this completes the proof of A.

Here is the proof of B. If T were Fréchet - differentiable, it would have to be true that

$$T(F_n,G_n) - T(F,G) - \dot{T}(F_n - F, G_n - G) = o(\|F_n - F\|_\infty \vee \|G_n - G\|_\infty) \tag{a}$$

for every sequence of pairs of d.f.'s $\{(F_n,G_n)\}$ with $\|F_n - F\|_\infty \to 0$ and $\|G_n - G\|_\infty \to 0$. We now exhibit a sequence $\{(F_n,G_n)\}$ for which (a) fails.

By straightforward algebra using (3),

$$T(F_n,G_n) - T(F,G) - \dot{T}(F_n - F, G_n - G) = \int (F_n - F) d(G_n - G). \tag{b}$$

Consider the d.f.'s F_n and G_n corresponding to the measures which put masses n^{-1} at $0, \ldots, (n-1)/n$ and $1/n, \ldots, 1$ respectively:

$$F_n = n^{-1} \sum_{k=0}^{n-1} \delta_{k/n}, \qquad \text{and} \qquad G_n = n^{-1} \sum_{k=1}^{n} \delta_{k/n}.$$

both of these sequences of df's converge uniformly to the uniform$(0,1)$ df $F(x) \equiv x \equiv G(x)$, and furthermore $\|F_n - F\|_\infty = \|G_n - G\|_\infty = 1/n$. Now

$$(F_n - F)(x) = \sum_{k=1}^{n} \left(\frac{k}{n} - x \right) 1_{[(k-1)/n, k/n)}(x),$$

$(F_n - F)(1) = 0$, and

$$(G_n - G)(x) = (F_n - F)(x) - \frac{1}{n} = \sum_{k=1}^{n} \left(\frac{k-1}{n} - x \right) 1_{[(k-1)/n, k/n)}(x)$$

with $(G_n - G)(1) = 0$. Thus, separating $G_n - G$ into its discrete and continuous parts,

$$\begin{aligned} \int (F_n - F) d(G_n - G) &= \sum_{k=1}^{n} (F_n - F)\left(\frac{k}{n}\right)(1/n) + n \int_0^{1/n} \left(\frac{1}{n} - t \right) \{-dt\} \\ &= n \frac{1}{n} \frac{1}{n} - n \left\{ \frac{1}{n} \frac{1}{n} - \frac{1}{2} \left(\frac{1}{n} \right)^2 \right\} \\ &= \frac{1}{2n} = O(1/n) \\ &\neq o(\|F_n - F\|_\infty \vee \|G_n - G\|_\infty) = o(1/n). \end{aligned}$$

Hence (a) fails and T is not Fréchet - differentiable. □

Remark 4.1 The previous example was suggested by R. M . Dudley. Dudley (1992), (1994) has studied other metrics, based $p-$variation norms, for which this T is *almost* Fréchet - differentiable, and for which some functionals may be Fréchet differentiable even though Fréchet differentiability with respect to $\|\cdot\|_\infty$ may fail.

A particular refinement of Hadamard differentiability which is very useful is as follows: since the limiting $P-$Brownian bridge process $\mathbb{G}_P$ of the empirical process $\mathbb{G}_n \equiv \sqrt{n}(\mathbb{P}_n - P)$ is in $C_u(\mathcal{F}, \rho_P)$ with probability one for any $\mathcal{F} \in CLT(P)$, we say that T is *Hadamard differentiable tangentially* to $C_u(\mathcal{F}, \rho_P)$ at $P \in \mathcal{P}$ if there is a continuous linear function $\dot{T} : C_u(\mathcal{F}, \rho_P) \to \mathbb{B}$ so that

$$\frac{T(P_t) - T(P)}{t} \to \dot{T}(\Delta_0)$$

holds for any path $\{P_t\}$ such that $\Delta_t \equiv (P_t - P_0)/t$ satisfies $\|\Delta_t - \Delta_0\|_{\mathcal{F}} \to 0$ with $\Delta_0 \in C_u(\mathcal{F}, \rho_P)$. Then a nice version of the delta-method for nonlinear functions T of $\mathbb{P}_n$ is given by the following theorem:

Theorem 4.4 Suppose that:
(i) T is Hadamard differentiable tangentially to $C_u(\mathcal{F}, \rho_P)$ at $P \in \mathcal{P}$.
(ii) $\mathcal{F} \in CLT(P)$: $\sqrt{n}(\mathbb{P}_n - P) \Rightarrow \mathbb{G}_P$ (where $\mathbb{G}_P$ takes values in $C_u(\mathcal{F}, \rho_P)$ by definition of $\mathcal{F} \in CLT(P)$).
Then

$$\sqrt{n}(T(\mathbb{P}_n) - T(P)) \Rightarrow \dot{T}(\mathbb{G}_P). \tag{4}$$

Proof. Define $g_n : \mathcal{P} \subset \ell^\infty(\mathcal{F}) \to \mathbb{B}$ by

$$g_n(x) \equiv \sqrt{n}(T(P + n^{-1/2}x) - T(P)).$$

Then, by (i), for $\{\Delta_n\} \subset \ell^\infty(\mathcal{F})$ with $\|\Delta_n - \Delta_0\|_{\mathcal{F}} \to 0$ and $\Delta_0 \in C_u(\mathcal{F}, \rho_P)$,

$$g_n(\Delta_n) \to \dot{T}(\Delta_0) \equiv g(\Delta_0).$$

Thus by the extended continuous mapping theorem in the Hoffmann - Jorgensen weak convergence theory (see van der Vaart and Wellner (1996), Theorem 1.11.1, page 67), $g_n(\mathbb{G}_n) \Rightarrow g(\mathbb{G}_P) = \dot{T}(\mathbb{G}_P)$, and hence (4) holds. □

The immediate corollary for the classical Mann-Whitney form of the Wilcoxon statistic given in example 4.6 is:

Corollary 1 If $X_1, \ldots, X_m$ are i.i.d. F and independent of $Y_1, \ldots, Y_n$ which are i.i.d. G, $0 < P_{F,G}(X \le Y) < 1$, and $\lambda_N \equiv m/N \equiv m/(m+n) \to \lambda \in (0,1)$, then

$$\begin{aligned}
\sqrt{\frac{mn}{N}}\left\{\int \mathbb{F}_m d\mathbb{G}_n - \int F dG\right\} &= \sqrt{\frac{mn}{N}}\left\{T(\mathbb{F}_m, \mathbb{G}_n) - T(F,G)\right\} \\
&\to_d \sqrt{1-\lambda}\int \mathbb{U}(F) dG - \sqrt{\lambda}\int \mathbb{V}(G) dF \\
&\sim N(0, \sigma^2_\lambda(F,G))
\end{aligned}$$

where $\mathbb{U}$ and $\mathbb{V}$ are two independent Brownian bridge processes and

$$\sigma^2_\lambda(F,G) = (1-\lambda)Var(G(X)) + \lambda Var(F(Y)).$$

This is, of course, well-known, and can be proved in a variety of other ways (by treating $T(\mathbb{F}_m, \mathbb{G}_n)$ as a two-sample $U-$statistic, or a rank statistic, or by a direct analysis), but the proof via the differentiable functional approach seems instructive and useful. (See e.g. Lehmann (1975), *Statistical Methods Based on Ranks*, Section 5, pages 362 - 371, and especially example 20, page 365.)

Other interesting applications have been given by Grübel (1988) (who studies the asymptotic theory of the length of the shorth); Pons and Turckheim (1989) (who study bivariate hazard estimators and tests of independence based thereon), and Gill and Johansen (1990) (who prove Hadamard differentiability of the "product integral"). Gill, van der Laan, and Wellner (1992) give applications to several problems connected with estimation of bivariate distributions. Arcones and Giné (1990) study the delta-method in connection with $M-$ estimation and the bootstrap. van der Vaart (1991b) shows that Hadamard differentiable functions preserve asymptotic efficiency properties of estimators.

5 High Order Derivatives

The following example illustrates the phenomena which we want to consider here in the simplest possible setting:

Example 5.1 Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli(p). Then $\sqrt{n}(\overline{X}_n - p) \to_d Z \sim N(0, p(1-p))$, and, if $g(p) = p(1-p)$,

$$\sqrt{n}(g(\overline{X}_n) - g(p)) \to_d g'(p)Z = (1-2p)Z \sim N(0, p(1-p)(1-2p)^2)$$

by the delta-method (or g-prime theorem). But if $p = 1/2$, since $g'(1/2) = 0$ this yields only

$$\sqrt{n}(g(\overline{X}_n) - 1/4) \to_d 0.$$

Thus we need to study the higher derivatives of g at $1/2$. since g is, in fact, a quadratic, we have

$$g(p) = g(1/2) + 0 \cdot (p - 1/2) + \frac{1}{2!}(-2)(p-1/2)^2 = 1/4 - (p-1/2)^2.$$

Thus

$$n(g(\overline{X}_n) - g(1/2)) = -n(\overline{X}_n - 1/2)^2 \to_d -Z^2 \sim -\frac{1}{4}\chi^2_1.$$

This is a very simple example of a more general limit theorem which we will develop below.

Now consider a functional $T : \mathcal{F} \to \mathbb{R}$ as in sections 1 - 4.

Definition 5.1 T is $k-$th order Gateaux differentiable at F if, with $F_t = F + t(G - F)$,

$$d_k T(F; (G-F)) = \frac{d^k}{dt^k} T(F_t)\Big|_{t=0}$$

exists.

Note that $\dot{T}(F; G-F) = d_1(T; G-F)$ if it exists. It is usually the case that

$$\begin{aligned} d_k T(F; G-F) &= \int \cdots \int \psi_k(x_1, \ldots, x_k) d(G-F)(x_1) \cdots d(G-F)(x_k) \\ &= \int \cdots \int \psi_{k,F}(x_1, \ldots, x_k) dG(x_1) \cdots dG(x_k); \end{aligned}$$

here the function $\psi_{k,F}$ is determined from ψ_k by a straightforward centering recipe:

$$\begin{aligned} \psi_{1,F}(x) &\equiv \psi_1(x) - \int \psi_1 dF, \\ \psi_{2,F}(x) &\equiv \psi_2(x_1, x_2) - \int \psi_2(x_1, x_2) dF(x_2) - \int \psi_2(x_1, x_2) dF(x_1) \\ &\quad + \int \int \psi_2(x_1, x_2) dF(x_1) dF(x_2), \end{aligned}$$

and so forth; see Serfling section 6.3.2, lemma A, page 222. A consequence of this is that we can write

$$\begin{aligned} n^{k/2} d_k(T(F; \mathbb{F}_n - F) &= \int \cdots \int \psi_k(x_1, \cdots, x_k) d(\mathbb{F}_n(x_1) - F(x_1)) \cdots d(\mathbb{F}_n(x_k) - F(x_k)) \\ &= n^{k/2} \int \cdots \int \psi_{k,F}(x_1, \ldots, x_k) d\mathbb{F}_n(x_1) \cdots d\mathbb{F}_n(x_k) \\ &= \frac{1}{n^{k/2}} \sum_{i_1=1}^{n} \cdots \sum_{i_k=1}^{n} \psi_{k,F}(X_{i_1}, \ldots, X_{i_k}). \end{aligned}$$

This is exactly $n^{k/2}$ times a "V- statistic" of order k.

Also note that, by Taylor's formula for a function of one real variable t, we have

$$T(F_t) - T(F) = \sum_{j=1}^{k} \frac{1}{j!} d_j T(F; G - F) + \frac{1}{(k+1)!} \frac{d^{k+1}}{dt^{k+1}} T(F_t)\Big|_{t=t^*}$$

for some $t^* \in [0, t]$. To analyze the asymptotic behavior of T in terms of $d_1 T$, $d_2 T$, ..., it is typically the first non-zero term $d_m T$ which dominates.

Serfling's condition A_m: suppose that

$$Var_F(\psi_{k,F}(X_1, \ldots, X_k)) \begin{cases} = 0 & \text{for } k < m \\ > 0 & \text{for } k = m. \end{cases}$$

$$R_{mn} \equiv T(\mathbb{F}_n) - T(F) - \frac{1}{m!} d_m T(F; \mathbb{F}_n - F)$$

satisfies $n^{m/2} R_{mn} = o_p(1)$.

This condition will be invoked with first $m = 1$ and then $m = 2$ in the following two theorems.

Theorem 5.1 (Serfling's theorem A). Suppose that $X_1, \ldots, X_n$ are i.i.d. F, and suppose that T satisfies A_1. Let $\mu(T, F) = E_F \psi_{1,F}(X_1) = 0$ and $\sigma^2(T, F) = Var(\psi_{1,F}(X_1))$ and suppose that $\sigma^2(T, F) < \infty$. Then

$$\sqrt{n}(T(\mathbb{F}_n) - T(F)) \to_d N(0, \sigma^2(T, F)).$$

Proof. Now by (ii) of condition A_1,

$$\begin{aligned} \sqrt{n}(T(\mathbb{F}_n) - T(F)) &= o_p(1) + n^{1/2} d_1 T(F; \mathbb{F}_n - F) \\ &= o_p(1) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{1,F}(X_i) \\ &\to_d N(0, \sigma^2(T, F)). \end{aligned}$$

This is essentially the same as in our earlier proofs of asymptotic normality using differentiability, but here we are *hypothesizing* that the remainder term is asymptotically negligible (and thus the real effort in using the theorem will be to verify the second part of A_1). □

Theorem 5.2 (Serfling's theorem B). Suppose that $X_1, \ldots, X_n$ are i.i.d. F, and suppose that T satisfies A_2 with $\psi_{2,F}(x,y) = \psi_{2,F}(y,x)$ and $E_F\psi_{2,F}^2(X_1,X_2) < \infty$, $E_F|\psi_{2,F}(X_1,X_1)| < \infty$, and $E_F\psi_{2,F}(x,X_2) = 0$. Define $A : L_2(F) \to L_2(F)$ by

$$Ag(x) = \int \psi_{2,F}(x,y)g(y)dF(y), \qquad g \in L_2(F),$$

and let $\{\lambda_k\}$ be the eigenvalues of A. Then

$$n\{T(\mathbb{F}_n) - T(F)\} \to_d \frac{1}{2}\sum_{k=1}^{\infty}\lambda_k Z_k^2$$

where $Z_1, Z_2, \ldots$ are i.i.d. $N(0,1)$.

Sketch of the proof: By condition A_2 we can write

$$\begin{aligned} n(T(\mathbb{F}_n) - T(F)) &= n\left\{T(\mathbb{F}_n) - T(F) - \frac{1}{2!}d_2(F;\mathbb{F}_n - F)\right\} + \frac{n}{2!}d_2(F;\mathbb{F}_n - F) \\ &= o_p(1) + \frac{n}{2}\int\int \psi_{2,F}(x_1,x_2)d\mathbb{F}_n(x_1)d\mathbb{F}_n(x_2). \end{aligned} \tag{1}$$

Now denote the orthonormal eigenfunctions of the (Hilbert-Schmidt) operator A by $\{\phi_k\}$ and the corresponding eigenvalues by $\{\lambda_k\}$: thus $A\phi_k = \lambda_k\phi_k$. Then it is well-known that

$$\psi_{2,F}(x,y) = \sum_{k=1}^{\infty}\lambda_k\phi_k(x)\phi_k(y)$$

in the sense of $L_2(F \times F)$ convergence. Hence

$$\begin{aligned} & n\int\int \psi_{2,F}(x_1,x_2)d\mathbb{F}_n(x_1)d\mathbb{F}_n(x_2) \\ &= n\int\int \sum_{k=1}^{\infty}\lambda_k\phi_k(x)\phi_k(y)d\mathbb{F}_n(x)d\mathbb{F}_n(y) \qquad (2) \\ &= n\sum_{k=1}^{\infty}\lambda_k\left\{\int \phi_k d\mathbb{F}_n\right\}^2 = \sum_{k=1}^{\infty}\lambda_k\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_k(X_i)\right\}^2 \\ &\to_d \sum_{k=1}^{\infty}\lambda_k Z_k^2 \qquad (3) \end{aligned}$$

where the Z_i's are i.i.d. $N(0,1)$ since $E_F\phi_k(X_i) = 0$, $E_F\psi_k^2(X_i) = 1$, and $E_F\phi_j(X_i)\phi_k(X_i) = 0$. Combining (1) and (3) completes the heuristic proof. The reason that this is heuristic is because of the infinite series appearing in (2) and (3). The complete proof entails consideration of finite sums and the corresponding approximation arguments; see Serfling (1981), pages 195 - 199 for the $U-$ statistic case. But note that the $V-$statistic argument on page 227 just involves throwing the diagonal terms back in, and is therefore really easier. □

Remark 5.1 It seems to me that Serfling's $\mu(T,F) = 0$ as formulated above. It also seems to me that he has missed the factor of $1/2$ appearing in the limit distribution.

Remark 5.2 Gregory (1977) gives related but stronger results which *do not* require $E_F\psi_{2,F}(X_1, X_1) = \sum_{k=1}^{\infty} \lambda_k < \infty$ and apply to some interesting statistics with $\lambda_k = 1/k$. Note that the infinite series

$$\sum_{k=1}^{\infty} \frac{1}{k}(Z_k^2 - 1)$$

defines a proper random variable since the summands have mean 0 and variances $2/k^2$; these actually arise as limit distributions of the popular Shapiro - Wilk tests for normality; see e.g. DeWet and Venter (1972), (1973).