CHAPTER 7.

STATISTICAL FUNCTIONALS AND THE DELTA METHOD

1. **Estimates as functionals of $I\!F_n$ or $I\!P_n$**
2. **Continuity of Functionals of F or P**
3. **Metrics for weak convergence**
4. **Differentiability of Functionals of F or P: Gateaux, Hadamard, and Frechet Derivatives**
5. **Higher order derivatives**

1. Estimators as functionals of $I\!F_n$ or $I\!P_n$

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 $T(I\!F_n)$ or $T(I\!P_n)$ where $I\!F_n$ and $I\!P_n$ denote the empirical distribution function and empirical measure of the data.

Notation: Suppose that $X_1, \cdots, X_n$ are iid P on $(\mathbf{X}, \mathbf{A})$. We let

$$I\!P_n \equiv \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \equiv \text{the empirical measure of the sample},$$

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

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

Here is a list of examples:

Example 1.1. The mean: $T(F) = \int x\, dF(x)$. $T(I\!F_n) = \int x\, dI\!F_n(x)$.

Example 1.2. The rth moment: for r an integer, $T(F) = \int x^r\, dF(x)$. $T(I\!F_n) = \int x^r\, dI\!F_n(x)$.

Example 1.3. The variance:

$$T(F) = Var_F(X) = \int \left(x - \int x\, dF\right)^2 dF(x) = \frac{1}{2}\int\int (x - y)^2\, dF(x)\, dF(y).$$

$$T(I\!F_n) = S_n^2 = \int \left(x - \int x\, dI\!F_n\right)^2 dI\!F_n(x).$$

Example 1.4. The median: $T(F) = F^{-1}(1/2)$. $T(I\!F_n) = I\!F_n^{-1}(1/2)$, the sample median.

Example 1.5. The $\alpha-$trimmed mean. $T(F) = (1-2\alpha)^{-1}\int_\alpha^{1-\alpha} F^{-1}(u)\, du$ for $0 < \alpha < 1/2$. $T(I\!F_n) = (1-2\alpha)^{-1}\int_\alpha^{1-\alpha} I\!F_n^{-1}(u)\, du$.

Example 1.6. The Hodges-Lehmann location functional:

$$T(F) = \frac{1}{2}\{F * F\}^{-1}(1/2).$$

$$T(I\!F_n) = \frac{1}{2}\{I\!F_n * I\!F_n\}^{-1}(1/2) = \underset{i,j}{med}\, \{(X_i + X_j)/2\}.$$

Example 1.7. The Mann - Whitney functional: $T(F,G) = \int F\,dG = P_{F,G}(X \le Y)$ if X, Y are independent with df's F, G respectively. $T(\mathbb{F}_m, \mathbb{G}_n) = \int \mathbb{F}_m\, d\mathbb{G}_n$ (based on two independent samples $X_1, \cdots, X_m$ iid F with empirical df $\mathbb{F}_m$ and $Y_1, \cdots, Y_n$ iid G with empirical df $\mathbb{G}_n$.

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

Example 1.9. Multivariate cross second moments: for P on $(R^k, \mathbf{B}^k)$: $T(P) = \int x\,x^T\, dP(x)$; here $T : \mathbf{P} \to R^{k\times k}$. $T(\mathbb{P}_n) = \int x\,x^T\, d\mathbb{P}_n(x) = n^{-1}\sum_i X_i X_i^T$.

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

$$
\begin{aligned}
T(P) &= \int (x - \int y\, dP(y))\,(x - \int y\, dP(y))^T\, dP(x) \\
&= (1/2)\int\int (x-y)(x-y)^T\, dP(x)dP(y) .
\end{aligned}
$$

$$
\begin{aligned}
T(\mathbb{P}_n) &= \int (x - \int y\, d\mathbb{P}_n(y))\,(x - \int y\, d\mathbb{P}_n(y))^T\, d\mathbb{P}_n(x) \\
&= n^{-1}\sum_i (X_i - \overline{X}_n)(X_i - \overline{X}_n)^T .
\end{aligned}
$$

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

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

where

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

Then $T(\mathbb{P}_n) = (T_1(\mathbb{P}_n), \cdots, T_k(\mathbb{P}_n))$ 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 R^k and $x \in R^k$, set $T(P) \equiv T(P)(x) = Pr_P(x \in S(X_1, \cdots, X_{k+1})$ where $X_1, \cdots, X_{k+1}$ are iid P and $S(x_1, \cdots, x_{k+1})$ is the simplex in R^k determined by $x_1, \cdots, x_{k+1}$; e.g. for $k = 2$, the simplex determined by x_1, x_2, x_3 is just a triangle.

Example 1.13. A maximum likelihood estimator: For P on $(\mathbf{X}, \mathbf{A})$; suppose that $\mathbf{P} = \{P_\theta : \theta \in \Theta \subset R^k\}$ is a regular parametric model with (vector) score function $\dot{\mathbf{l}}_\theta(\cdot\,; \theta)$. Then for general P, not necessarily in the model $\mathbf{P}$, consider $T(P)$ defined by $\int \dot{\mathbf{l}}_\theta(x\,; T(P))\, dP(x) = 0$. Then $\int \dot{\mathbf{l}}_\theta(x\,; T(I\!P_n))\, dI\!P_n(x) = 0$ defines $T(I\!P_n)$. For location in one dimension, with $\dot{\mathbf{l}}_\theta(x\,;\theta) = \psi(x - \theta)$ and $\psi \equiv -f'/f$, these become $\int \psi(x - T(F))\, dF(x) = 0$ and $\int \psi(x - T(I\!F_n))\, dI\!F_n(x) = 0$.

Example 1.14. A bootstrap functional: Let $T(F)$ be a functional with estimator $T(I\!F_n)$, and consider estimating the distribution $H_n(F; \cdot) = P_F\{\sqrt{n}(T(I\!F_n) - T(F)) \le \cdot\}$. A natural estimator is $H_n(I\!F_n; \cdot)$.

2. Continuity of Functionals of F or P

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

Definition 2.1. $F_n \rightarrow_d F$ or $F_n \Longrightarrow F$ if $\int \psi \, dF_n \rightarrow \int \psi \, dF$ for all bounded continuous functions ψ. $P_n \Longrightarrow P$ if $\int \psi \, dP_n \rightarrow \int \psi \, dP$ for all bounded and continuous functions ψ.

Definition 2.2. A. $T : \mathbf{F} \rightarrow R$ is *weakly continuous* at F_0 if $F_n \Longrightarrow F_0$ implies $T(F_n) \rightarrow T(F_0)$. $T : \mathbf{F} \rightarrow R$ is *weakly lower-semicontinuous at* F_0 *if* $F_n \Longrightarrow F_0$ *implies* $\liminf_{n \rightarrow \infty} T(F_n) \geq T(F_0)$.
B. $T : \mathbf{P} \rightarrow R$ *is weakly continuous* at $P_0 \in \mathbf{P}$ if $P_n \Longrightarrow P_0$ implies $T(P_n) \rightarrow T(P_0)$.

Example 2.3. $T(F) = \int x \, dF(x)$ is discontinuous at every F_0: if $F_n = (1 - n^{-1}) F_0 + n^{-1} \delta_{a_n}$, then $F_n \Longrightarrow F_0$ since, for bounded ψ,

$$\int \psi \, dF_n = (1 - \frac{1}{n}) \int \psi \, dF_0 + \frac{1}{n} \psi(a_n) \rightarrow \int \psi \, dF_0 .$$

But

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

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

Example 2.4. $T(F) = (1 - 2\alpha)^{-1} \int_\alpha^{1-\alpha} F^{-1}(u) \, du$ with $0 < \alpha < 1/2$ is continuous at every F_0: $F_n \Longrightarrow F_0$ implies that $F_n^{-1}(t) \rightarrow F_0^{-1}(t)$ a.e. Lebesgue. Hence

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

by the dominated convergence theorem.

Example 2.5. $T(F) = F^{-1}(1/2)$ is continuous at every F_0 such that F_0^{-1} is continuous at $1/2$.

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

Example 2.6. (A lower-semicontinous functional $T(F)$). 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.

Theorem 2.7. (Varadarajan). If $X_1, \cdots, X_n$ are iid P on a separable metric space (S, d), then $Pr(IP_n \Longrightarrow P) = 1$.

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

$$IP_n \psi \equiv \int \psi \, dIP_n = \frac{1}{n} \sum_{i=1}^{n} \psi(X_i) \to_{a.s.} P\psi \equiv \int \psi \, dP$$

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), section 11.4, pp. 313ff for details. $\square$

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

Proposition 2.8. Suppose that:

(i) $(\mathbf{X}, \mathbf{A}) = (S, \mathbf{B}_{borel})$ where (S, d) is a separable metric space and $\mathbf{B}_{borel}$ denotes its usual Borel sigma - field.

(ii) $T : \mathbf{P} \to R$ is weakly continuous at $P_0 \in \mathbf{P}$.

(iii) $X_1, \cdots, X_n$ are iid P_0.

Then $T_n \equiv T(IP_n) \to_{a.s.} T(P_0)$.

Proof. By Varadarajan's theorem 7.2.6, $IP_n \Longrightarrow P_0$ a.s. Fix $\omega \in A$ with $Pr(A) = 1$ so that $IP_n^{\omega} \Longrightarrow P_0$. Then by weak continuity of T, $T(IP_n^{\omega}) \to T(P_0)$. $\square$

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 this type of continuity. The following approach is often useful:

Definition 2.9. Let $\mathbf{F} \subset L_1(P)$ be a collection of integrable functions. Say that $P_n \to P$ with respect to $||\cdot||_{\mathbf{F}}$ if $||P_n - P||_{\mathbf{F}} \equiv \sup_{f \in \mathbf{F}} |P_n(f) - P(f)| \to 0$. Furthermore, we say that $T : \mathbf{P} \to R$ is continuous with respect to $||\cdot||_{\mathbf{F}}$ if $||P_n - P||_{\mathbf{F}} \to 0$ implies that $T(P_n) \to T(P)$.

Definition 2.10. If $\mathbf{F} \subset L_1(P)$ is a collection of integrable functions with $||I\!P_n - P||^*_{\mathbf{F}} \to 0$, we then say that $\mathbf{F}$ is a Glivenko-Cantelli class for P and write $\mathbf{F} \in GC(P)$.

Theorem 2.11. Suppose that :

(i) $\mathbf{F} \in GC(P)$; i.e. $||I\!P_n - P||^*_{\mathbf{F}} \to_{a.s.} 0$.

(ii) T is continuous with respect to $||\cdot||_{\mathbf{F}}$.

Then $T(I\!P_n) \to_{a.s.} T(P)$.

3. Metrics for Distribution Functions F and probability distributions P

We have already encountered the total variation and Hellinger metrics in the course of studying Bayes estimators and tests of hypotheses. As we will see, as useful as these metrics are, they are too strong: the empirical measure $I\!P_n$ 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 useful metrics for probability measures P and df's F, and their properties.

Definition 3.1. The *Kolmogorov or supremum metric* between two df's F and G is

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

Definition 3.2. The *Levy metric* between two df's F and G is

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

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

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

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 $||f||_{BL} = ||f||_L + ||f||_\infty$.

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

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

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

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

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

Definition 3.6. The *Hellinger metric* d_H (or H) is defined by

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

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

Here is a basic theorem establishing relationships between these metrics:

Theorem 3.7.

A. $d_{Pr}(P,Q)^2 \le d_{BL^*}(P,Q) \le 2\, d_{Pr}(P,Q)$.

B. $d_H^2(P,Q) \le d_{TV}(P,Q) \le d_H(P,Q)\{2 - d_H^2(P,Q)\}^{1/2}$.

C. $d_{Pr} \le d_{TV}$.

D. For distributions P , Q on the real line, $d_L \le d_K \le d_{TV}$.

Proof. We have already proved B in statistics 582. For A, see Dudley (1989) section 11.3, problem 5, and section 11.6, corollary 11.6.5; and see Huber (1981), corollary 2.4.3, page 33. Another useful reference is Whitt (1974). □

Theorem 3.8. (Strassen). The following are equivalent:

(a) $d_{Pr}(P,Q) \le \epsilon$.

(b) There exist $X \sim P$, $Y \sim Q$ defined on a common probability space $(\Omega, \mathbf{F}, Pr)$ such that $Pr(\, d(X,Y) \le \epsilon) \ge 1 - \epsilon$.

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

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

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

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

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

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

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

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

Definition 7.4.2. T has the *influence function* or "influence curve" $IC(x; T, F)$ at F if, with $F_t = (1-t)F + t\,\delta_x$,

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

Examples 7.4.3.

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

B. The mean: $T(F) = \int x\, dF(x)$. 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*.

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

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

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

D. The median: $T(F) \equiv F^{-1}(1/2)$, and suppose that F has density f which is positive at $F^{-1}(1/2)$. Then, with $F_t = (1-t)F + tG$,

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

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

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

so that

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

Hence

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

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

$$\dot{T}(F; G - F) = - \frac{1}{f(F^{-1}(p))} (G - F)(F^{-1}(p))$$

$$= \; - \frac{1}{f(F^{-1}(p))} \{ 1_{(-,F^{-1}(p)]}(x) \; - \; p) \, dG(x)$$

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 derivatives: in particular the stronger notions of derivative which we will discuss below are those of Fréchet and Hadamard derivatives.

Definition 7.4.4. A functional $T : \mathbf{F} \to R$ is $d_* -$ Fréchet differentiable at $F \in \mathbf{F}$ if there exists a continuous linear functional $\dot{T}(F; G - F)$ from finite signed measures into R such that

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

Here are some properties of Fréchet - differentiation:

Theorem 7.4.5. 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 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. $\dot{T}(F; G - G) \; = \; \int \psi \, d(G - F) \; = \; \int (\psi \; - \; \int \psi \, dF) \, dG$ 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 7.4.6. Suppose that T is $d_* -$ differentiable at F and that

(*) $\qquad \sqrt{n}\, d_*(IF_n\,, F) \; = \; O_p(1) \, .$

Then

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

Proof. By Fréchet differentiability of T at F,

$$\sqrt{n}(T(IF_n) - T(F)) = \sqrt{n}\int \psi_F \, dIF_n + \sqrt{n}\, o(d_*(IF_n, F))$$
$$= \sqrt{n}\int \psi_F \, dIF_n + \frac{o(d_*(IF_n, F))}{d_*(IF_n, F)} \sqrt{n}\, d_*(IF_n, F)$$
$$= \sqrt{n}\int \psi_F \, dIF_n + o(1)\, O_P(1) . \quad \square$$

Note that if d_* is the Lévy metric d_L or the Kolmogorov metric d_K on the line, then (*) is satisfied: $\sqrt{n}\, d_L(IF_n, F) \le \sqrt{n}\, d_K(IF_n, F) = \sqrt{n}||IF_n - F||_\infty \stackrel{d}{=} ||IU_n(F)||_\infty \to_d ||IU(F)||_\infty$. Unfortunately, if $d_* = d_{Pr}$ or $d_* = d_{BL}$, then $\sqrt{n}\, d_*(IF_n, F)$ is *not* $O_p(1)$ in general -- see Dudley (1969), Kersting (1978), and Huber - Carol (1977). Thus we are lead to consideration of larger 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 wrt the supremum or Kolmogorov metric, we can make more functionals differentiable by considering a somewhat weaker notion of differentiability as follows:

Definition 7.4.7. A functional $T : \mathbf{F} \to R$ is *Hadamard - differentiable* at F with respect to the Kolmogorov distance $d_K \equiv ||\cdot||_\infty$ (or *compactly differentiable*) if there exists $\dot{T}(F; \cdot)$: continuous and linear satisfying

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

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

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

where $\sqrt{n}(IF_n - F) \stackrel{d}{=} IU_n(F) \Longrightarrow IU(F)$; hence we can easily deduce the following theorem:

Theorem 7.4.8. Suppose that $T : \mathbf{F} \to R$ is Hadamard - differentiable at F wrt $||\cdot||_\infty$. Then

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

Moreover,

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

Proof. This is easily proved using a Skorokod construction of the empirical process, or by the extended continuous mapping theorem. (Gill (1989) used the Skorokod 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

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

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

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

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

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

C. When $\mathcal{S} = \{\text{all bounded subsets of } (\mathbf{F}, d_*)\}$, 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 norm, but which is *not* Fréchet differentiable with respect to this norm.

Example 7.4.9. For distribution functions F, G on R, define T by $T(F, G) = \int F\, dG = P(X \le Y)$ where $X \sim F, Y \sim G$ are independent. Let $||\tilde{F} - F||_\infty := \sup_x |\tilde{F}(x) - F(x)| := ||\tilde{F} - F||_{\mathbf{F}}$ where $\mathbf{F} := \{1_{(-\infty, t]} : t \in R\}$.

Proposition 7.4.10.

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

$$\dot{T}(\alpha, \beta) = \int \alpha dG - \int \beta\, dF. \tag{2}$$

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 respect 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. Now the second term on the right is bounded by

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

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:

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

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

$$
\begin{aligned}
\text{(3)} \qquad T(F_n\,, G_n) \;-\; T(F, G) \;-\; \dot{T}(F_n \;-\; F\,, G_n \;-\; G\,) & \\
&= o(||F_n \;-\; F||_\infty \bigvee ||G_n \;-\; G||_\infty\,)
\end{aligned}
$$

for every sequence of pairs of df'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 (3) fails.

By straightforward algebra using (2),

$$
\begin{aligned}
\text{(4)} \qquad T(F_n\,, G_n) \;-\; T(F, G) \;-\; \dot{T}(F_n \;-\; F\,, G_n \;-\; G\,) & \\
&= \int (F_n \;-\; F)\, d(G_n \;-\; G)\,.
\end{aligned}
$$

Consider the df's F_n and G_n corresponding to the measures which put masses n^{-1} at $0, \cdots, (n-1)/n$ and $1/n\,, \cdots, 1$ 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 $F(x) := x := G(x)$, and furthermore $||F_n - F||_\infty = ||G_n - G||_\infty = 1/n$. Now

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

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

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

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

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

Hence (3) fails and T is not Fréchet - differentiable. [This example was suggested to me by R. M. Dudley in February, 1990.] Dudley (1992), (1994) has studied other metrics, based on the p - variation norms, for which this T is *almost* Fréchet - differentiable, and which enable Fréchet differentiability may hold 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 G_P of the empirical process $\mathbb{G}_n \equiv \sqrt{n}(\mathbb{P}_n - P)$ is in $\mathbf{C}_u(\mathbf{F}, \rho_P)$ with probability one for any $\mathbf{F} \in CLT(P)$, we say that T is *Hadamard differentiable tangentially to* $C_u(\mathbf{F}, \rho_P)$ at $P \in \mathbf{P}$ if there is a continuous linear function $\dot{T} : \mathbf{C}_u(\mathbf{F}, \rho_P) \rightarrow \mathbf{B}$ so that

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

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

Theorem 7.4.11. Suppose that:

(i) T is Hadamard differentiable tangentially to $\mathbf{C}_u(\mathbf{F}, \rho_P)$ at $P \in \mathbf{P}$.

(ii) $\mathbf{F} \in CLT(P)$: $\sqrt{n}(I\!P_n - P) \Longrightarrow I\!G_P$ (where $I\!G_P$ takes values in $\mathbf{C}_u(\mathbf{F}, \rho_P)$ by definition of $\mathbf{F} \in CLT(P)$).

Then

$$\sqrt{n}(T(I\!P_n) - T(P)) \Longrightarrow \dot{T}(I\!G_P) . \tag{5}$$

Proof. Define $g_n : \mathbf{P} \subset l^\infty(\mathbf{F}) \rightarrow \mathbf{B}$ by

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

Then, by (i), for $\{\Delta_n\}$ in $l^\infty(\mathbf{F})$ with $||\Delta_n - \Delta_0||_{\mathbf{F}} \rightarrow 0$ and $\Delta_0 \in \mathbf{C}_u(\mathbf{F}, \rho_P)$,

$$g_n(\Delta_n) \rightarrow \dot{T}(\Delta_0) := g(\Delta_0) .$$

Thus by the extended continuous mapping theorem in the Hoffmann-Jorgensen weak convergence theory (see Van der Vaart and Wellner (1990), proposition 1.5.A), $g_n(I\!G_n) \Longrightarrow g(I\!G_P) = \dot{T}(I\!G_P)$, and hence (5) holds. □

The immediate corollary for the classical Mann-Whitney form of the Wilcoxon statistic given in example 7.4.9 is:

Corollary 7.4.12. If $X_1, \cdots, X_m$ are iid F and independent of $Y_1, \cdots, Y_n$ which are iid G , and $\lambda_N := m/N := m/(m+n) \rightarrow \lambda \in (0,1)$, then

$$\begin{aligned}\sqrt{\frac{mn}{N}}\{\int I\!F_m \, dI\!G_n - \int F \, dG\} &= \sqrt{\frac{mn}{N}}\{T(I\!F_m, I\!G_n) - T(F,G)\} \\ &\rightarrow_d \sqrt{1-\lambda}\int I\!U(F)\, dG - \sqrt{\lambda}\int I\!V(G)\, dF \\ &\sim N(0, \sigma_\lambda^2(F,G))\end{aligned}$$

where $I\!U$ and $I\!V$ are two independent Brownian bridge processes and

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

This is, of course, well-known, and can be proved in a variety of other ways (by treating $T(I\!F_m, I\!G_n)$ as a two - sample U-statistic, or as a rank statistic, or by a direct analysis), but the proof via the differentiable functional approach seems instructive and useful.

Other interesting applications of the delta method have recently 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. Higher Order Derivatives

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

Example 7.5.1. Suppose that $X_1, X_2, \cdots, X_n$ are iid Bernoulli(p). Then $\sqrt{n}(\overline{X} - p) \to_d Z \sim N(0, p(1-p))$, and, if $g(p) = p(1-p)$,

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

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

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

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

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

Thus

$$n(g(\overline{X}_n) - \frac{1}{4}) = -n(\overline{X}_n - \frac{1}{2})^2 \to_d -Z^2 \sim -\frac{1}{4}\chi_1^2 .$$

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

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

Definition 7.5.2. T is $k-$ order Gateaux differentiable at F if, with $F_t \equiv F + t(G - F)$,

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

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

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

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

$$\psi_{1,F}(x) \equiv \psi_1(x) - \int \psi_1 \, dF ,$$

$$\psi_{2,F}(x_1, x_2) = \psi_2(x_1, x_2) - \int \psi_2(x_1, x_2)\, dF(x_2) - \int \psi_2(x_1, x_2)\, dF(x_1)$$

$$+ \int\int \psi_2(x_1, x_2)\, dF(x_1)\, dF(x_2) \,,$$

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

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

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

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

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

Condition 7.5.3. (Serfling's condition A_m) Suppose that:

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

(ii) $$R_{mn} \;\equiv\; T(I\!F_n) \;-\; T(F) \;-\; \frac{1}{m!}\, d_m T(F\,; I\!F_n - F)$$

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 7.5.4. (Serfling's theorem A). Suppose that $X_1\,, \cdots, X_n$ are iid 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

$$\sqrt{n}(T(I\!F_n) \;-\; T(F)) \;\to_d\; N(0\,, \sigma^2(T, F)) \,.$$

Proof. Now by (ii) of condition A_1 ,

$$\begin{aligned} \sqrt{n}(T(I\!F_n) \;-\; T(F)) \;&=\; o_p(1) \;+\; n^{1/2} d_1 T(F\,; I\!F_n - F) \\ &=\; o_p(1) \;+\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{1,F}(X_i) \end{aligned}$$

$$\to_d \quad N(0\,,\sigma^2(T\,,F))\;;$$

this is essentially the same as in our earlier proofs of asymptotic normality using differentiability, but here we are hypothesizing that the remainder term goes away. □

Theorem 7.5.5. (Serfling's theorem B). Suppose that $X_1\,,\cdots,X_n$ are iid F , and suppose that T satisfies A_2 with $\psi_{2,F}(x,y) = \psi_{2,F}(y,x)$ and $E_F\psi^2_{2,F}(X_1\,,X_2) < \infty$, $E_F|\psi_{2,F}(X_1\,,X_1)| < \infty$, and $E_F\psi_{2,F}(x\,,X_2) \equiv 0$. Define $A : L_2(F) \to L_2(F)$ by

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

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

$$n\{T(I\!F_n) - T(F)\} \to_d \frac{1}{2}\sum_{k=1}^{\infty}\lambda_k\,Z_k^2$$

where $Z_1\,,Z_2\,,\cdots$ are iid $N(0,1)$.

Sketch of the proof. By condition A_2 we can write

$$\begin{aligned} n(T(I\!F_n) - T(F)) &= n(T(I\!F_n) - T(F) - \frac{1}{2!}\,d_2(F\,;I\!F_n - F)) \\ &\quad + \frac{n}{2!}\,d_2(F\,;I\!F_n - F) \\ \text{(a)} \qquad &= o_p(1) + \frac{n}{2!}\int\!\!\int \psi_{2,F}(x_1\,,x_2)\,dI\!F_n(x_1)dI\!F_n(x_2)\,. \end{aligned}$$

Now denote the orthonormal eigenfunctions of the (Hilbert - Schmidt) operator A by $\{\phi_k\}$ and the corresponding eigenfunctions 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

$$\begin{aligned} & n\int\!\!\int \psi_{2,F}(x_1\,,x_2)\,dI\!F_n(x_1)dI\!F_n(x_2)\,. \\ \text{(b)} \qquad &= n\int\!\!\int \sum_{k=1}^{\infty}\lambda_k\,\phi_k(x)\,\phi_k(y)\,dI\!F_n(x)\,dI\!F_n(y) \\ &= n\sum_{k=1}^{\infty}\lambda_k\,\{\int \phi_k\,dI\!F_n\}^2 = \sum_{k=1}^{\infty}\lambda_k\,\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_k(X_i)\}^2 \\ \text{(c)} \qquad &\to_d \sum_{k=1}^{\infty}\lambda_k\,Z_k^2 \end{aligned}$$

where the Z_i's are iid $N(0,1)$ since $E_F \phi_k(X_i) = 0$, $E_F \phi_k^2(X_i) = 1$, and $E_F \phi_j(X_i)\phi_k(X_i) = 0$. Combining (a) and (c) completes the heuristic proof. The reason that this is heuristic is because of the infinite series appearing in (b) and (c). 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. $\square$

Remark 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 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 (1965) tests for normality; see e.g. DeWet and Venter (1972), (1973).