Statistics 583, Midterm Exam, Solutions

Wellner; 5/13/2015

1. (30 points) **Define** *any three* of the following terms.
   (a) The bounded Lipschitz metric d_{BL^*} on the set of probability measures on a metric space (S, d).
   (b) The Kolmogorov metric d_K on the class of all distribution functions on $\mathbb{R}$.
   (c) A metric d_* that is *compatible* with the empirical distribution function or empirical measure.
   (d) A Hadamard-differentiable functional $T : \mathcal{F} \to \mathbb{R}$ with respect to the Kolmogorov metric d_K.

   Solution: See course notes.

2. (30 points) Give a complete **statement** of *any two* of the following results:
   (a) Hoeffding's formula for the distribution of ranks under the alternative.
   (b) The Wald- Wolfowitz-Noether-Hájek finite sampling central limit theorem.
   (c) Varadarajan's theorem concerning weak convergence of the empirical measure $\mathbb{P}_n$ when $X_1, \ldots, X_n$ are i.i.d. P on a metric space (M, d).
   (d) Any theorem about consistency of an estimator via continuity of statistical functionals.
   (e) Any theorem about asymptotic normality of an estimator via differentiability of the corresponding statistical functional.

   Solution: See course notes.

   Do either 3 or 4.

3. (40 points) Let $h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be symmetric; thus $h(x, y) = h(y, x)$. Suppose that $X_1, \ldots, X_n$ are i.i.d. with $Eh^2(X_1, X_2) < \infty$ and $Eh^2(X_1, X_1) < \infty$, and let $\theta \equiv T(F) = Eh(X_1, X_2)$.
   (a) State a result from van der Vaart chapter 12 that allows one to conclude that the $U-$statistics $U_n = \sum_{i \neq j} h(X_i, X_j)/n(n-1)$ satisfy $\sqrt{n}(U_n - \theta) \to_d N(0, V^2(F))$ where $V^2(F)$ is related to the variance of a certain function $h_1(x)$.
   (b) Use the asymptotic normality you obtained in (a) to show that the corresponding $V-$statistic V_n satisfies $\sqrt{n}(V_n - \theta) \to_d N(0, V^2(F))$.

   Solution: (a) From van der Vaart, Theorem 12.3, with $r = 2$, if $Eh^2(X_1, X_2) < \infty$, then $\sqrt{n}(U_n - \theta - \hat{U}_n) \to_p 0$ where

$$\hat{U}_n \equiv \frac{2}{n} \sum_{i=1}^{n} h_1(X_i)$$

and $h_1(x) = Eh(x, X_2) - \theta$. Thus

$$\begin{aligned}\sqrt{n}(U_n - \theta) &= \sqrt{n}(\hat{U}_n + U_n - \theta - \hat{U}_n) \\ &= \sqrt{n}\hat{U}_n + o_p(1) = 2n^{-1/2}\sum_{i=1}^{n} h_1(X_i) + o_p(1) \\ &\to_d N(0, 4Eh_1^2(X_1))\end{aligned}$$

since $Eh_1^2(X_1) = E\{[Eh(X_1, X_2)]^2\} \leq Eh^2(X_1, X_2) < \infty$ by Jensen's inequality. Another way to identify the structure of "linear term" is by a Gateaux derivative computation as in problem 6 below.
(b) As in problem 3, problem set #2, we let $V_n \equiv n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} h(X_i, X_j)$ and

$$U_n \equiv \frac{1}{\binom{n}{2}}\sum_{i<j} h(X_i, X_j) = \frac{1}{n(n-1)}\sum_{i\neq j} h(X_i, X_j).$$

From van der Vaart's Theorem 12.3 we know that if $Eh^2(X_1, X_2) < \infty$, then $\sqrt{n}(U_n - \theta - \hat{U}_n) \to_p 0$, and hence

$$\sqrt{n}(U_n - \theta) \to_d N(0, 4Var(h_1(X_1)))$$

where $h_1(x) = Eh(x, X_2) - \theta$. But

$$\begin{aligned}V_n - U_n &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} h(X_i, X_j) - \frac{1}{n(n-1)}\sum_{i\neq j} h(X_i, X_j) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} h(X_i, X_i) + \left\{\frac{1}{n^2} - \frac{1}{n(n-1)}\right\}\sum_{i\neq j} h(X_i, X_j) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} h(X_i, X_i) - \frac{1}{n^2(n-1)}\sum_{i\neq j} h(X_i, X_j) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} h(X_i, X_i) - \frac{1}{n}U_n. \qquad (1)\end{aligned}$$

Thus

$$\sqrt{n}(V_n - U_n) = \tfrac{\sqrt{n}}{n^2}\textstyle\sum_{i=1}^{n} h(X_i, X_i) - \tfrac{1}{\sqrt{n}}U_n.$$

Now $U_n \to_p \theta$, so $n^{-1/2}U_n \to_p 0$, and

$$E\{n^{-3/2}\sum_{i=1}^{n} h(X_i, X_i)\} = n^{-1/2}Eh(X_1, X_1) \to 0, \quad \text{and}$$

$$Var\left(n^{-3/2}\sum_{i=1}^{n} h(X_i, X_i)\right) = n^{-3}nVar(h(X_1, X_1)) \to 0$$

since $Eh^2(X_1, X_1) < \infty$ by assumption. Thus the first term in (1) converges to 0 in probability by Chebychev's inequality. It follows that $\sqrt{n}(V_n - \theta) \to_d N(0, 4Var(h_1(X_1)))$.

4. (40 points) Suppose that $\{S_n\}$ and $\{T_n\}$ are arbitrary sequences of statistics (with $E(S_n^2) < \infty$ and $E(T_n^2) < \infty$ for each n) satisfying

$$\frac{Var(S_n - T_n)}{Var(T_n)} \to 0. \tag{2}$$

(a) Show that (2) implies that $Var(S_n)/Var(T_n) \to 1$. [Hint: Show that

$$\frac{Var(S_n)}{Var(T_n)} = 1 + \frac{Var(S_n - T_n) + 2Cov(S_n - T_n, T_n)}{Var(T_n)}$$

and use the Cauchy-Schwarz inequality.]
(b) Show that $2Cov(S_n, T_n) = Var(S_n) + Var(T_n) - Var(S_n - T_n)$.
(c) Let

$$R_n \equiv \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} - \frac{T_n - E(T_n)}{\sqrt{Var(T_n)}}.$$

Use the identity in (b) to show that

$$\begin{aligned} Var(R_n) &= 2 - 2\frac{Cov(S_n, T_n)}{\sqrt{Var(S_n)Var(T_n)}} \\ &= 2 - \sqrt{\frac{Var(S_n)}{Var(T_n)}} - \sqrt{\frac{Var(T_n)}{Var(S_n)}} + \sqrt{\frac{Var(T_n)}{Var(S_n)}} \cdot \frac{Var(S_n - T_n)}{Var(T_n)}. \end{aligned}$$

(d) Use (a) together with the identity you proved in (c) to show that (2) implies $Var(R_n) \to 0$.
(e) Does the conclusion of (d) imply $R_n \to_p 0$?

Solution: (a) Note that

$$Var(S_n) = Var(T_n + S_n - T_n) = Var(T_n) + 2Cov(T_n, S_n - T_n) + Var(S_n - T_n),$$

and hence

$$\frac{Var(S_n)}{Var(T_n)} = 1 + \frac{2Cov(T_n, S_n - T_n)}{Var(T_n)} + \frac{Var(S_n - T_n)}{Var(T_n)}$$

where the third term converges to 0 by the hypothesis and where, by the Cauchy-Schwarz inequality,

$$\frac{|Cov(T_n, S_n - T_n)|}{Var(T_n)} \leq \frac{\sqrt{Var(T_n)Var(S_n - T_n)}}{Var(T_n)} = \sqrt{\frac{Var(S_n - T_n)}{Var(T_n)}} \to 0.$$

Thus

$$\begin{aligned}\left|\frac{Var(S_n)}{Var(T_n)} - 1\right| &\leq 2\frac{|Cov(T_n, S_n - T_n)|}{Var(T_n)} + \frac{Var(S_n - T_n)}{Var(T_n)} \\ &\leq 2\sqrt{\frac{Var(S_n - T_n)}{Var(T_n)}} + \frac{Var(S_n - T_n)}{Var(T_n)} \\ &\to 0\end{aligned}$$

by the assumption $Var(S_n - T_n)/Var(T_n) \to 0$. Therefore

$$\frac{Var(S_n)}{Var(T_n)} \to 1$$

as claimed.
(b) Now

$$Var(S_n - T_n) = Var(S_n) - 2Cov(S_n, T_n) + Var(T_n)$$

and rearranging this identity yields

$$2Cov(S_n, T_n) = Var(S_n) + Var(T_n) - Var(S_n - T_n).$$

(c) With

$$R_n \equiv \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} - \frac{T_n - E(T_n)}{\sqrt{Var(T_n)}},$$

it follows from the first identity in (b) followed by the second identity in (b) that

$$\begin{aligned}Var(R_n) &= Var\left(\frac{S_n - E(S_n)}{\sqrt{Var(S_n)}}\right) + Var\left(\frac{T_n - E(T_n)}{\sqrt{Var(T_n)}}\right) - 2\frac{Cov(S_n, T_n)}{\sqrt{Var(S_n)Var(T_n)}} \\ &= 2\left(1 - \frac{Cov(S_n, T_n)}{\sqrt{Var(S_n)Var(T_n)}}\right) \\ &= 2 - \sqrt{\frac{Var(S_n)}{Var(T_n)}} - \sqrt{\frac{Var(T_n)}{Var(S_n)}} + \sqrt{\frac{Var(T_n)}{Var(S_n)}} \cdot \frac{Var(S_n - T_n)}{Var(T_n)}.\end{aligned}$$

(d) By (c) and (a) we conclude that

$$Var(R_n) \to 2 - 1 - 1 + 1 \cdot 0 = 0.$$

(e) Since $E(R_n) = 0$ for all n, (d) yields $R_n \to_p 0$ by Chebychev's inequality.

Do either 5 or 6.

5. (40 points): Suppose that $T(F)$ is defined by

$$\int_{\mathbb{R}} \psi(x - T(F))dF(x) = 0$$

where $\psi : \mathbb{R} \to \mathbb{R}$ is monotone non-decreasing and ψ' exists.
(a) Find the Gateaux derivative and the influence function of $T(F)$.
(b) Are there specific choices for ψ (perhaps not satisfying the derivative hypothesis everywhere) that lead to functionals $T(F)$ (and corresponding estimators $T(\mathbb{F}_n)$) with known names?
(c) What asymptotic variance do you expect for $\sqrt{n}(T(\mathbb{F}_n) - T(F))$?
(d) Specialize your result in (c) to the case when $\psi(x) = -(f'/f)(x)$ and $f(x) = e^{-x}/(1 + e^{-x})^2$ is the logistic density.

Solution: (a) Let $F_t \equiv (1 - t)F + tG$ for a distribution function G. Then differentiation across the identity

$$0 = \int \psi(x - T(F_t))dF_t(x)$$

yields, with $\dot{T}(F; G - F)$ denoting the Gateaux derivative,

$$\begin{aligned} 0 &= \int \psi(x - T(F))d(G - F)(x) - \int \psi'(x - T(F))\dot{T}(F; G - F)dF(x) \\ &= \int \psi(y - T(F))d(G - F)(x) - \int \psi'(y - T(F))dF(y) \cdot \dot{T}(F; G - F). \end{aligned}$$

Hence the Gateaux derivative $\dot{T}$ is given by

$$\dot{T}(F; G - F) = \frac{\int \psi(y - T(F))d(G - F)(y)}{\int \psi'(y - T(F))dF(y)}.$$

The influence function is obtained by taking $G = \delta_x$:

$$\begin{aligned} IC(x; T, F) &= \dot{T}(F; \delta_x - F) = \frac{\psi(x - T(F)) - \int \psi(y - T(F))dF(y)}{\int \psi'(y - T(F))dF(y)} \\ &= \frac{\psi_F(x - T(F))}{\int \psi'_F(y - T(F))dF(y)} \end{aligned}$$

with

$$\psi_F(x - T(F)) \equiv \psi(x - T(F)) - \int \psi(y - T(F))dF(y) = \psi(x - T(F))$$

by definition of $T(F)$.
(b) If $\psi(x) = x$, then

$$0 = \int \psi(x - T(F))dF(x) = \int (x - T(F))dF(x) = \int xdF(x) - T(F),$$

and hence $T(F) = \int x dF(x) = E_F(X)$, the mean of F.
If $\psi(x) = 21_{[0,\infty)}(x) - 1$, then

$$\begin{aligned} 0 &= \int \psi(x - T(F))dF(x) = 2\int 1_{[0,\infty)}(x - T(F))dF(x) - 1 \\ &= 2\int 1_{[x-T(F)\geq 0]}dF(x) - 1 \\ &= 2(1 - F(T(F))) - 1, \end{aligned}$$

or, equivalently, $F(T(F)) = 1/2$. Thus $T(F)$ is the (or a version of) the median.
(c) From our theory in chapter 7, we expect that a stronger notion of differentiability will yield

$$\begin{aligned} \sqrt{n}(T(\mathbb{F}_n) - T(F)) &= n^{-1/2}\sum_{i=1}^{n} \frac{\psi_F(X_i - T(F))}{\int \psi'(y) - T(F))dF(y)} + o_p(1) \\ &\to_d N\left(0, \frac{E_F\psi^2(X - T(F))}{\{E_F\psi'(X - T(F))\}^2}\right). \end{aligned} \tag{3}$$

(d) When $\psi(x) = -(f_0'/f_0)(x)$ with $f_0(x) = e^{-x}/(1 + e^{-x})^2$, the logistic density, it follows that

$$\psi(x) = \frac{1 - e^{-x}}{1 + e^{-x}}, \quad \text{and} \quad \psi'(x) = 2f_0(x),$$

so that ψ is strictly increasing from -1 at $-\infty$ to $+1$ at ∞. [Note that f_0 is fixed and not connected to F elsewhere in the problem!] Thus ψ is bounded and continuous and we expect that (3) will hold for any F with ψ as given in the last display.

6. (40 points): Let $T(P) = \iint h(x, y)dP(x)dP(y)$ where $h : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a fixed function and P is a probability measure on $(\mathcal{X}, \mathcal{A})$. Assume that $Eh^2(X_1, X_2) < \infty$ where $X_1, X_2 \sim P$ are independent.
(a) Find the influence function of $T(P)$.
(b) Evaluate $T(\mathbb{P}_n)$ where $\mathbb{P}_n$ is the empirical distribution of $X_1, \ldots, X_n$ i.i.d. P on $(\mathcal{X}, \mathcal{A})$. Is $T(\mathbb{P}_n)$ an unbiased estimator of $T(P)$? If not, find an unbiased estimator.
(c) What do you expect for the asymptotic variance of $\sqrt{n}(T(\mathbb{P}_n) - T(P))$?
(d) Specialize your result in (c) to the case $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}^d, \mathcal{B}^d)$ and $h(x, y) = \|x - y\|$ where $\|x - y\|$ is the Euclidean distance from x to y.

Solution: Without loss we may assume that h is symmetric: $h(x, y) = h(y, x)$. (If not, replace h by $\overline{h}(x, y) = (h(x, y) + h(y, x))/2$.)

(a) With $P_t = (1-t)P + tQ$, $T(P_t) = \int h(x,y)dP_t(x)dP_t(y)$, and hence

$$\begin{aligned}
\frac{d}{dt}T(P_t)\big|_{t=0} &= \iint h(x,y)d(Q-P)(x)dP(y) + \iint h(x,y)dP(x)d(Q-P)(y) \\
&= \iint h(x,y)d(Q-P)(x)dP(y) + \iint h(y,x)dP(x)d(Q-P)(y) \\
&= 2\iint h(x,y)d(Q-P)(x)dP(y) \quad \text{since} \quad h \quad \text{is symmetric.}
\end{aligned}$$

Taking $Q = \delta_{x_0}$ yields

$$IC(x_0;T,P) = 2\int \left(h(x_0,y) - \int h(x,y)dP(x)\right)dP(y) = 2\left(h_1(x_0) - T(P)\right)$$

where

$$h_1(x) \equiv \int h(x,y)dP(y) = \int h(y,x)dP(y).$$

(b) Note that

$$T(\mathbb{P}_n) = \iint h(x,y)d\mathbb{P}_n(x)d\mathbb{P}_n(y) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n h(X_i,X_j).$$

This is a biased estimator of $T(P)$:

$$\begin{aligned}
E\{T(\mathbb{P}_n)\} &= \frac{1}{n^2}E\left\{\sum_{i=1}^n h(X_i,X_i) + \sum_{i\neq j} h(X_i,X_j)\right\} \\
&= \frac{1}{n^2}\left\{nE_P(h(X_1,X_1) + n(n-1)E_P h(X_1,X_2)\right\} \\
&= \frac{1}{n}E_P(h(X_1,X_1) + \frac{n-1}{n}E_P h(X_1,X_2) \\
&= T(P) + \frac{1}{n}\left\{E_P(h(X_1,X_1) - T(P)\right\} \\
&\neq T(P).
\end{aligned}$$

An unbiased estimator is given by the $U-$statistic

$$V_n = \frac{1}{n(n-1)}\sum_{i\neq j} h(X_i,X_j).$$

(c) Based on the analysis in (a) and a strengthened differentiability argument, or by standard projection theory (as in the solution of problem 3 above), we know that

$$\sqrt{n}(T(\mathbb{P}_n) - T(P)) \to_d N(0, 4Var_P(h_1(X_1)))$$

assuming that $E_P\{h_1^2(X_1)\} < \infty$.
(d) When $h(x, y) = \|x - y\|$ we have $h_1(x) = \int \|x - y\| dP(y) - T(P)$ and the conclusion in (c) holds if $E_P h_1^2(X_1) < \infty$. Note that since the norm $\|x - y\|$ is convex, Jensen's inequality yields

$$h_1(x) \leq \|x - E_P X\|$$

and hence $E_P h_1^2(X) < \infty$ if $E_P \|X\|^2 < \infty$.