Statistics 583, Problem Set 7 Solutions

Wellner; 5/18/2016

1. Let $T(P) \equiv \int\int h(x,y)dP(x)dP(y)$ for a fixed function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with $\int\int |h(x,y)|dP(x)dP(y) < \infty$. The corresponding estimator $T(\mathbb{P}_n)$ is a $V-$statistic, and the natural unbiased estimator is

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

(a) Show that $T(\mathbb{P}_n)$ is a biased estimator of $T(P)$ and compute the bias.
(b) Find the influence function of $T(P)$.
(c) What do you expect for the asymptotic variance of $\sqrt{n}(T(\mathbb{P}_n) - T(P))$?
(d) What is the Hájek projection of U_n? How does it relate to the influence function you calculated in (b)?
(e) How does the result in (c) compare with the limiting distribution of $\sqrt{n}(U_n - T(P))$?
Hint: see van der Vaart, *Asymptotic Statistics*, section 12.1, pages 161 - 163.

Solution: (a) Note that

$$T(\mathbb{P}_n) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n h(X_i, X_j) = n^{-2} \left\{ \sum_{i=1}^n h(X_i, X_i) + \sum_{i \ne j} h(X_i, X_j) \right\}.$$

Thus we compute

$$\begin{aligned} ET(\mathbb{P}_n) &= n^{-2} \{ nEh(X_1, X_1) + n(n-1)Eh(X_1, X_2) \} \\ &= n^{-1} \int h(x,x)dP(x) + \frac{n-1}{n} \int\int h(x,y)dP(x)dP(y), \end{aligned}$$

and it follows that the bias is given by

$$ET(\mathbb{P}_n) - T(P) = \frac{1}{n} \left(\int h(x,x)dP(x) - \int\int h(x,y)dP(x)dP(y) \right).$$

(b) To calculate the influence function, let $P_t = (1-t)P + tQ$. Then

$$\begin{aligned} \frac{d}{dt} T(P_t) \Big|_{t=0} &= \frac{d}{dt} \int\int h(x,y)dP_t(x)dP_t(y) \Big|_{t=0} \\ &= \int\int h(x,y)d(Q-P)(x)dP(y) + \int\int h(x,y)dP(x)d(Q-P)(y) \\ &= \int \left\{ \int h(x,y)dP(y) - T(P) \right\} dQ(x) + \int \left\{ \int h(x,y)dP(x) - T(P) \right\} dQ(y) \\ &= \int \left\{ \int h(x,y')dP(y') + \int h(x',x)dP(x') - 2T(P) \right\} dQ(x). \end{aligned}$$

Thus the influence function of $T(P)$ is

$$\begin{aligned}\psi_P(x) &= \int h(x,y')dP(y') + \int h(x',x)dP(x') - 2T(P) \\ &= h_1(x) + h_2(x) - 2T(P)\end{aligned}$$

where $h_1(x) = \int h(x,y')dP(y') = Eh(x,X_1)$ and $h_2(x) = \int h(x',x)dP(x') = Eh(X_1,x)$.
(c) Thus we expect that the asymptotic variance of $\sqrt{n}(T_n - T(P))$ will be

$$Var_P(\psi_P(X_1)) = E(h_1(X_1) + h_2(X_1))^2 - (2T(P))^2.$$

(d) Assuming symmetry of the kernel h of U_n ($h(x,y) = h(y,x)$ for all x,y), the Hájek projection of $U_n - \theta$ where $\theta = T(P)$ is given by

$$\hat{U}_n = \sum_{k=1}^{n} E\{U_n - \theta | X_k\};$$

see e.g. van der Vaart (1998, pages 161-162. In the present case we compute

$$\begin{aligned}E(U_n - \theta|X_k) &= E\left(\binom{n}{2}^{-1} \sum_{1\le i<j\le n} (h(X_i,X_j) - \theta|X_k)\right) \\ &= \binom{n}{2}^{-1} \sum_{1\le i<j\le n} E\left(h(X_i,X_j) - \theta|X_k\right) \\ &= \frac{2}{n}\left(h_1(X_k) - \theta\right)\end{aligned}$$

where $h_1(x) \equiv Eh(x,X_2) = h_1(x))$ from from (b).
(e) Thus the Hájek projection yields

$$\sqrt{n}(U_n - \theta) = \frac{2}{\sqrt{n}} \sum_{k=1}^{n} \left(2h_1(X_k) - 2T(P)\right) \to_d N(0, 4Var(h_1(X_1)).$$

exactly the same as the distribution we expect for $\sqrt{n}(T(\mathbb{P}_n) - T(P))$.

2. Let $T(F) = \int (F - F_0)^2 dF_0$ Find the first and second order Gateaux derivatives of $T(F)$ at $F = F_0$. What limit distribution do you expect for $nT(\mathbb{F}_n)$?

Solution: With $F_t = (1-t)F_0 + tG$ we have

$$T(F_t) = \int ((1-t)F_0 + tG - F_0)^2 dF_0 = \int t^2 (G - F_0)^2 dF_0$$

and hence

$$\frac{d}{dt}T(F_t) = \int 2t(G-F_0)^2 dF_0, \quad \text{so that} \quad \frac{d}{dt}T(F_t)\Big|_{t=0} = 0,$$

and furthermore

$$\frac{d^2}{dt^2}T(F_t)\Big|_{t=0} \equiv d_2T(F_0; G-F_0) = \int 2(G-F_0)^2 dF_0.$$

This suggests the approximation

$$\begin{aligned} T(\mathbb{F}_n) &\approx \frac{1}{2!}d_2T(F_0; \mathbb{F}_n - F_0) = \frac{1}{2n}d_2T(F_0; \sqrt{n}(\mathbb{F}_n - F_0)) \\ &= \frac{1}{n}\int \{\sqrt{n}(\mathbb{F}_n - F_0)\}^2 dF_0 \end{aligned}$$

which is clearly correct since, by definition of T, it holds with equality. Writing out the sum involved in $\mathbb{F}_n - F_0$ and squaring it yields

$$\begin{aligned} nT(\mathbb{F}_n) &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \int \left((1_{(-\infty,v]}(X_i) - F_0(v))) \cdot (1_{(-\infty,v]}(X_j) - F_0(v)))\right) dF_0(v) \\ &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n h_{F_0}(X_i, X_j) \end{aligned}$$

where

$$h_{F_0}(x,y) = \frac{1}{2}F_0^2(v) + \frac{1}{2}F_0^2(v) - F_0(x \wedge y) + \frac{1}{3}.$$

See van der Vaart (1998), Example 12.13, pp 170-171 and Example 20.6, pp 295-296 for more details. Note that

$$\begin{aligned} nT(\mathbb{F}_n) &= n\int (\mathbb{F}_n - F_0)^2 dF_0 = \int \left\{\sqrt{n}(\mathbb{F}_n - F_0)\right\}^2 dF_0 \\ &\stackrel{d}{=} \int \mathbb{U}_n(F_0)^2 dF_0 = 2\int_0^1 \mathbb{U}_n^2(t)dt \end{aligned}$$

where $\mathbb{U}_n(t) = \sqrt{n}(\mathbb{G}_n(t) - t)$ and $\mathbb{G}_n$ is the empirical d.f. of n i.i.d. Uniform$(0,1)$ rv's, and where we assumed that F_0 is continuous in the last step. This is the form of the *Cramér - von Mises statistic* for testing $H : F = F_0$ versus $K : F \neq F_0$. It is well known that

$$\int_0^1 \mathbb{U}_n^2(t)dt \to_d \int_0^1 \mathbb{U}^2(t)dt \stackrel{d}{=} \sum_{j=1}^\infty \frac{1}{j^2\pi^2}Z_j^2$$

where the $Z_j \sim N(0,1)$ are i.i.d.; see e.g. Shorack and Wellner, (1986, 2009), chapter 5, and van der Vaart (1998) Example 12.13.

3. Let F be a bivariate distribution function and define $T(F) = \int \varphi(F_1, F_2) dF_2$ if F_1 and F_2 are the (one-dimensional) marginal distribution functions of F and $\varphi : [0,1]^2 \to \mathbb{R}$ is a smooth fixed function.
(a) Find the influence function of T.
(b) Write out $T(\mathbb{F}_n)$ where $\mathbb{F}_n$ is the bivariate empirical distribution function of $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. as F.
(c) What asymptotic variance do you expect for $\sqrt{n}(T(\mathbb{F}_n) - T(F))$?

Solution: Let $F_t = (1-t)F + tG$ for bivariate distribution functions F and G. We write $F_{1,t}$, $F_{2,t}$ for the first and second marginal distribution functions of F_t, and F_1, F_2 for the first and second marginal distribution functions of F. Then $T(F_t) = \int \varphi(F_{1,t}, F_{2,t}) dF_{2,t}$. Furthermore, with φ_1 and φ_2 denoting the first partial derivatives of φ with respect to x and y respectively,

$$\begin{aligned}
\frac{d}{dt} T(F_t) &= \int \varphi(F_{1,t}, F_{2,t}) d(G_2 - F_2) \\
&\quad + \int \varphi_1(F_{1,t}, F_{2,t})(G_1 - F_1) dF_{2,t} + \int \varphi_2(F_{1,t}, F_{2,t})(G_2 - F_2) \\
&\overset{t=0}{=} \int \varphi(F_1, F_2) d(G_2 - F_2) \\
&\quad + \int \varphi_1(F_1, F_2)(G_1 - F_1) dF_2 + \int \varphi_2(F_1, F_2)(G_2 - F_2) dF_2 \\
&= \int \varphi(F_1, F_2) d(G_2 - F_2) \\
&\quad + \int \varphi_1(F_1, F_2)(v) \int_{\mathbb{R}} 1_{[x \le v]} d(G_1 - F_1)(x) dF_2(v) \\
&\quad + \int \varphi_2(F_1, F_2)(v) \int_{\mathbb{R}} 1_{[y \le v]} d(G_2 - F_2)(x) dF_2(v) \\
&\equiv I + II + III.
\end{aligned}$$

Now by Fubini's theorem we can write

$$\begin{aligned}
II &= \int_{\mathbb{R}} \left(\int \varphi_1(F_1(v), F_2(v)) 1_{[x \le v]} dF_2(v) \right) d(G_1 - F_1)(x), \quad \text{and} \\
III &= \int_{\mathbb{R}} \left(\int \varphi_2(F_1(v), F_2(v)) 1_{[y \le v]} dF_2(v) \right) d(G_2 - F_2)(y).
\end{aligned}$$

Thus we find that

$$\begin{aligned} I+II+III &= \int \left\{ \varphi(F_1(y),F_2(y)) + \int \varphi_2(F_1(v),F_2(v)) 1_{[y\le v]} dF_2(v) \right\} d(G_2(y)-F_2(y)) \\ &\quad + \int \left\{ \int \varphi_1(F_1(v),F_2(v)) 1_{[x\le v]} dF_2(v) \right\} d(G_1(x)-F_1(x)) \\ &\equiv \int \{\psi_Y(y) - E\psi_Y(Y)\}\, dG_2(y) + \int \{\psi_X(x) - E\psi_X(X)\}\, dG_1(x) \end{aligned}$$

where

$$\psi_X(x) \equiv \int \varphi_1(F_1(v),F_2(v)) 1_{[x\le v]} dF_2(v), \tag{1}$$

$$\psi_Y(y) \equiv \varphi(F_1(y),F_2(y)) + \int \varphi_2(F_1(v),F_2(v)) 1_{[y\le v]} dF_2(v). \tag{2}$$

Thus the influence function is

$$\psi_F(x,y) = \psi_Y(y) + \psi_X(x) - E_F\{\psi_Y(Y) + \psi_X(X)\}$$

where ψ_X and ψ_Y are given by (1) and (2).
(b) Writing out $T(\mathbb{F}_n)$ and $\sqrt{n}(T(\mathbb{F}_n) - T(F))$ we find that

$$T(\mathbb{F}_n) = \int \varphi(\mathbb{F}_{n,1}(v), \mathbb{F}_{n,2}(v)) d\mathbb{F}_{n,2}(v)$$

and hence

$$\begin{aligned} \sqrt{n}(T(\mathbb{F}_n) - T(F)) &= \sqrt{n} \left\{ \int \varphi(\mathbb{F}_{n,1}(v), \mathbb{F}_{n,2}(v)) d\mathbb{F}_{n,2}(v) - \int \varphi(F_1(v),F_2(v)) dF_2(v) \right\} \\ &= \sqrt{n} \int \varphi(\mathbb{F}_{n,1}(v), \mathbb{F}_{n,2}(v)) d(\mathbb{F}_{n,2}(v) - F_2(v)) \\ &\quad + \sqrt{n} \int \{\varphi(\mathbb{F}_{n,1}(v), \mathbb{F}_{n,2}(v)) - \varphi(F_1(v),F_2(v))\}\, dF_2(v) \\ &= \sqrt{n} \int \varphi(F_1(v),F_2(v)) d(\mathbb{F}_{n,2}(v) - F_2(v)) \\ &\quad + \sqrt{n} \int \{\varphi(\mathbb{F}_{n,1}(v), \mathbb{F}_{n,2}(v)) - \varphi(F_1(v),F_2(v))\}\, d(\mathbb{F}_{n,2}(v) - F_2(v)) \\ &\quad + \sqrt{n} \int \{\varphi(\mathbb{F}_{n,1}(v), \mathbb{F}_{n,2}(v)) - \varphi(F_1(v),F_2(v))\}\, dF_2(v) \\ &\equiv A_n + B_n + C_n. \end{aligned}$$

By using a Taylor expansion of φ we find that

$$\begin{aligned} C_n &= \int \nabla\varphi(\mathbb{F}^*_{n,1}, \mathbb{F}^*_{n,2})(v)^T(\sqrt{n}(\mathbb{F}_{n,1}(v) - F_1(v)), \sqrt{n}(\mathbb{F}_{n,2}(v) - F_2(v)))dF_2(v) \\ &= \int \nabla\varphi(F_1, F_2)(v)^T(\sqrt{n}(\mathbb{F}_{n,1}(v) - F_1(v)), \sqrt{n}(\mathbb{F}_{n,2}(v) - F_2(v)))dF_2(v) \\ &\quad + \int \left(\nabla\varphi(\mathbb{F}^*_{n,1}, \mathbb{F}^*_{n,2})(v)^T - \nabla\varphi(F_1, F_2)(v)^T\right) \cdot \\ &\qquad \cdot(\sqrt{n}(\mathbb{F}_{n,1}(v) - F_1(v)), \sqrt{n}(\mathbb{F}_{n,2}(v) - F_2(v)))dF_2(v) \\ &\equiv C_{n,1} + C_{n,2} \end{aligned}$$

where, $\|\mathbb{F}^*_{n,j} - F_j\|_\infty \leq \|\mathbb{F}_{n,j} - F_j\|_\infty \to_{a.s.} 0$, and, by Fubini's theorem,

$$\begin{aligned} C_{n,1} &= \int \left\{ \int (\nabla\varphi)_1(F_1, F_2)(v) 1_{[x \leq v]} dF_2(v) \right\} d\left\{\sqrt{n}(\mathbb{F}_{n,1}(x) - F_1(x))\right\} \\ &\quad + \int \left\{ \int (\nabla\varphi)_2(F_1, F_2)(v) 1_{[y \leq v]} dF_2(v) \right\} d\left\{\sqrt{n}(\mathbb{F}_{n,2}(y) - F_2(y))\right\} \end{aligned}$$

and where we expect that $B_n = o_p(1)$ and $C_{n,2} = o_p(1)$. Combining these pieces yields

$$\begin{aligned} &\sqrt{n}(T(\mathbb{F}_n) - T(F)) \\ &= \int \varphi(F_1(v), F_2(v)) d\{\sqrt{n}(\mathbb{F}_{n,2}(v) - F_2(v))\} \\ &\quad + \int \left\{ \int (\nabla\varphi)_1(F_1, F_2)(v) 1_{[x \leq v]} dF_2(v) \right\} d\left\{\sqrt{n}(\mathbb{F}_{n,1}(x) - F_1(x))\right\} \\ &\quad + \int \left\{ \int (\nabla\varphi)_2(F_1, F_2)(v) 1_{[y \leq v]} dF_2(v) \right\} d\left\{\sqrt{n}(\mathbb{F}_{n,2}(y) - F_2(y))\right\} \\ &\quad + o_p(1) \\ &= \int \left\{ \varphi(F_1(y), F_2(y)) + \left\{ \int (\nabla\varphi)_2(F_1, F_2)(v) 1_{[y \leq v]} dF_2(v) \right\} \right\} d\{\sqrt{n}(\mathbb{F}_{n,2}(y) - F_2(y))\} \\ &\quad + \int \left\{ \int (\nabla\varphi)_1(F_1, F_2)(v) 1_{[x \leq v]} dF_2(v) \right\} d\left\{\sqrt{n}(\mathbb{F}_{n,1}(x) - F_1(x))\right\} \\ &\quad + o_p(1). \end{aligned}$$

This yields the same influence function as in (a) (with slightly different notation for the derivatives of φ) and yields a proof of asymptotic normality of $\sqrt{n}(T_n - T(H))$ if we show that $B_n = o_p(1)$ and $C_{n,2} = o_p(1)$.

Remark: It is also interesting to note that $T_n \equiv T(\mathbb{F}_n)$ is essentially a type of

bivariate rank statistic: since $n\mathbb{F}_{n,2}(Y_{(i)}) = i$ for $i = 1, \ldots, n$, and $Y_{(i)} = \mathbb{F}_{n,2}^{-1}(i/n)$

$$\begin{aligned} T(\mathbb{F}_n) &= \int \varphi(\mathbb{F}_{n,1}(v), \mathbb{F}_{n,2}(v)) d\mathbb{F}_{n,2}(v) \\ &= \frac{1}{n}\sum_{i=1}^{n} \varphi(\mathbb{F}_{n,1}(Y_{(i)}), \mathbb{F}_{n,2}(Y_{(i)})) \\ &= \frac{1}{n}\sum_{i=1}^{n} \varphi(\mathbb{F}_{n,1}(\mathbb{F}_{n,2}^{-1}(i/n)), i/n) \\ &= \frac{1}{n}\sum_{i=1}^{n} \varphi(n^{-1}R_{n,i}), i/n) \end{aligned}$$

where $R_{n,i} \equiv n\mathbb{F}_{n,1}(\mathbb{F}_{n,,2}^{-1}(i/n)) = \#\text{of } X_j's \leq \text{the i-th largest } \ Y_i$.

4. (a) Given n distinct data items, show that the probability that a given data item does not appear in a bootstrap sample is $e_n = (1 - 1/n)^n$
(b) Show that $e_n \to e^{-1} \approx .368$ as $n \to \infty$.
(c) Hence show that the probability that each of B bootstrap samples contains an item i is $(1 - e_n)^B$. Evaluate this quantity for $n = 10, 20, 50, 100$ and $B = 10, 20, 50, 100$.
(d) Let $N_n \equiv \sum_{j=1}^{n} 1_{[M_j=0]}$ where $\underline{M} \equiv (M_1, \ldots, M_n) \sim \ \text{Mult}_n(n, \underline{1}/n)$. Show that $E(n^{-1}N_n) = e_n$ as computed in (a).

Solution: (a) The probability that X_i does not appear in a bootstrap sample $X_1^*, \ldots, X_n^*$ from $\mathbb{F}_n$ is just $e_n = P(M_i = 0)$ where $M_i \sim \text{Binomial}(n, 1/n)$. Thus we have $e_n = P(M_i = 0) = \binom{n}{0}(1/n)^0(1 - 1/n)^n = (1 - 1/n)^n$.
(b) Since $(1+x/n)^n \to e^x$ for any x, it follows immediately that $e_n \to e^{-1} \approx .368$.
(c) The probabality that each of B bootstrap samples contains X_i is clearly $(1 - e_n)^B$. The following table gives values of this for $n = 10, 20, 50, 100$ and $B = 10, 20, 50, 100$.

B/n	10	20	50	100
10	.0137	.0118	.0108	.0105
20	.000189	.000139	.000117	.000110
50	4.89×10^{-10}	2.29×10^{-10}	1.47×10^{-10}	1.27×10^{-10}
100	2.39×10^{-19}	5.26×10^{-20}	2.16×10^{-20}	1.61×10^{-20}

(d) N_n/n is the proportion of the original sample not appearing in the bootstrap sample. Since $N_n \equiv \sum_{j=1}^{n} 1_{[M_j=0]}$ where each M_j is marginally Binomial$(n, 1/n)$, it follows immediately that

$$E(N_n/n) = P(M_1 = 0) = (1 - 1/n)^n \to e^{-1}\,.$$

Furthermore, from occupancy theory for urn models,

$$\sqrt{n}(n^{-1}N_n - (1-1/n)^n) \to_d N(0, e^{-1}(1-2e^{-1}))\,;$$

see e.g. Johnson and Kotz (1977), page 317.