Statistics 581

Problem Set 5 Solutions

Wellner; 11/1/2006

1. Suppose that $X_1, \ldots, X_n$ are i.i.d. random vectors with values in R^k with $E(X_1) = \mu$ and $E(X_1^T X_1) < \infty$ so that $\Sigma = E(X_1 - \mu)(X_1 - \mu)^T$ is well-defined. Thus

$$Z_n \equiv \sqrt{n}(\overline{X}_n - \mu) \to_d Z \sim N_k(0, \Sigma)\,.$$

Suppose that $g : R^k \to R$ is a function, and suppose that $\nabla g = \dot{g}$ exists at μ. Then the delta-method (or g' theorem) tells us that

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) \to_d \nabla g(\mu)^T Z \sim N(0, \nabla g(\mu)^T \Sigma \nabla g(\mu))\,. \tag{1}$$

(a) Show that we can strengthen (1) as follows: Suppose that $\nabla g = \dot{g}$ is continuous at μ. Then $\sqrt{n}(g(\overline{X}_n) - g(\mu))$ is *asymptotically linear* at μ:

$$\begin{aligned}\sqrt{n}(g(\overline{X}_n) - g(\mu)) &= \nabla g(\mu)^T \sqrt{n}(\overline{X}_n - \mu) + o_p(1)\\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(X_i) + o_p(1)\end{aligned}$$

where

$$\psi(x) = \nabla g(\mu)^T (x - \mu) \tag{2}$$

which is called the *influence function* of $g(\overline{X}_n)$ as an estimator of $g(\mu)$, has mean $E\psi(X_i) = 0$ and $Var(\psi(X_i)) = \nabla g(\mu)^T \Sigma \nabla g(\mu)$.
(b) Does the result of (a) apply to the situation considered in problem 2(a) of problem set #4? If not, formulate another result of the same type as in (a) which does apply, and use it to find the influence function of $S_n^2/\overline{X}_n$.

Solution: (a) By Taylor's theorem, for some Y_n^* satisfying $|Y_n^* - \mu| \le |\overline{X}_n - \mu| \to_p 0$ it follows that

$$\begin{aligned}\sqrt{n}(g(\overline{X}_n) - g(\mu)) &= \nabla g(Y_n^*) \sqrt{n}(\overline{X}_n - \mu)\\ &= \nabla g(\mu) \sqrt{n}(\overline{X}_n - \mu)\\ &\qquad + \{\nabla g(Y_n^*) - \nabla g(\mu)\} \sqrt{n}(\overline{X}_n - \mu)\\ &= \nabla g(\mu) \sqrt{n}(\overline{X}_n - \mu) + o_p(1)\end{aligned}$$

since $\nabla g(Y_n^*) \to_p \nabla g(\mu)$ by continuity of ∇g at μ and since $\sqrt{n}(\overline{X}_n - \mu) = O_p(1)$. Now note that

$$\nabla g(\mu) \sqrt{n}(\overline{X} - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla g(\mu)(X_i - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(X_i)$$

with ψ as in (2).

As pointed out by Arseni, the hypothesis of continuity of ∇g can be dropped: consider a new function $h(x) = g(x) - \nabla g(\mu)x$. Then $\nabla h(\mu) = \nabla g(\mu) - \nabla g(\mu) = 0$, and we can write

$$\begin{aligned} \sqrt{n}(g(\overline{X}_n) - g(\mu) - \nabla g(\mu)(\overline{X}_n - \mu)) &= \sqrt{n}(h(\overline{X}_n) - h(\mu)) \\ &\to_d \nabla h(\mu)Z = 0 \cdot Z = 0 \end{aligned} \tag{3}$$

by the delta-method applied to the function h. Since convergence in distribution to a constant implies convergence in probability to the same constant, we conclude from (3) that the left side of (3) converges in probability to 0. But this is just the claimed asymptotic linearity with $\psi(x) = \nabla g(\mu)(x - \mu)$.

(b) The result in (a) does not quite apply since

$$Z_n \equiv \sqrt{n}(\overline{X}_n - \mu, S_n^2 - \sigma^2)'$$

is not exactly an average of i.i.d. random vectors. But the key features of the proof in (a) do carry through since $n^{-1/2}Z_n \to_p 0$ and $Z_n = n^{-1/2}\sum_{i=1}^n \underline{Y}_i + o_p(1)$ where $\underline{Y}_i = (X_i - \mu, (X_i - \mu)^2 - \sigma^2)'$ are i.i.d. with mean 0 and finite second moment under the assumptions of problem 2(a) of problem set #4. Thus the conclusion continues to hold. Thus with $g(u, v) = v/u$

$$\begin{aligned} \sqrt{n}\left(\frac{S_n^2}{\overline{X}_n} - \frac{\sigma^2}{\mu}\right) &= \nabla g(\mu, \sigma^2)\sqrt{n}(\overline{\underline{Y}}_n - \underline{\mu}_Y) + o_p(1) \\ &= \frac{1}{\mu}(-\sigma^2/\mu, 1)\sqrt{n}(\overline{\underline{Y}}_n - \underline{\mu}_Y) + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{1}{\mu}\left\{(X_i - \mu)^2 - \sigma^2 - \frac{\sigma^2}{\mu}(X_i - \mu)\right\} + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i) + o_p(1) \end{aligned}$$

where

$$\begin{aligned} \psi(x) &= \frac{1}{\mu}\{(x - \mu)^2 - \sigma^2 - (\sigma^2/\mu)(x - \mu)\} \\ &= \frac{\sigma^2}{\mu}\left\{\left(\frac{x - \mu}{\sigma}\right)^2 - 1 - \frac{\sigma}{\mu}\left(\frac{x - \mu}{\sigma}\right)\right\} \end{aligned}$$

has $E\psi(X_i) = 0$ and

$$Var(\psi(X_i)) = \frac{\sigma^4}{\mu^2}\left\{2 + \gamma_2 - 2\frac{\sigma\gamma_1}{\mu} + \frac{\sigma^2}{\mu^2}\right\} = V^2$$

as in the solution of problem 4.2(a).

2. (a) Write out a proof of (10) on page 16 of the Chapter 2 notes.
(b) Write out a proof of the corresponding fact concerning the general empirical process $\mathbb{G}_n$: $\mathbb{G}_n \to_{f.d.} \mathbb{G}$ where $\mathbb{G}_n$ and $\mathbb{G}$ are as defined on page 21 of the chapter 2 notes; i.e. for any $f_1, \ldots, f_k \in L_2(P)$, $(\mathbb{G}_n(f_1), \ldots, \mathbb{G}_n(f_k)) \to_d (\mathbb{G}(f_1), \ldots, \mathbb{G}(f_k))$.

Solution: (a) $\mathbb{U}_n \to_{f.d.} \mathbb{U}$. To see this, let $0 < t_1 < t_2 < \cdots < t_k < 1$. Then define random vectors $\underline{Y}_i$ by

$$\underline{Y}_i = (1_{[0,t_1]}(\xi_i) - t_1, \ldots, 1_{[0,t_k]}(\xi_i) - t_k),$$

for $i = 1, \ldots, n$. Note that $E\underline{Y}_1 = 0$ and

$$\begin{aligned} E\underline{Y}_1\underline{Y}_1' &= \begin{pmatrix} t_1(1-t_1) & t_1 - t_1t_2 & \cdots & t_1 - t_1t_k \\ t_1 - t_1t_2 & t_2(1-t_2) & \cdots & t_2 - t_2t_k \\ . & . & \cdots & . \\ . & . & \cdots & . \\ . & . & \cdots & . \\ t_1 - t_1t_k & t_2 - t_2t_k & \cdots & t_k(1-t_k) \end{pmatrix} \\ &= (t_i \wedge t_j - t_it_j)_{i,j=1}^k \equiv \Sigma. \end{aligned}$$

Thus it follows from the multivariate central limit theorem that

$$(\mathbb{U}_n(t_1), \ldots, \mathbb{U}_n(t_k))' = \sqrt{n}\underline{Y}_n \to_d N_k(0, \Sigma).$$

But for a Brownian bridge process $\mathbb{U}$, $(\mathbb{U}(t_1), \ldots, \mathbb{U}(t_k))' \sim N_k(0, \Sigma)$, so we have shown that $(\mathbb{U}_n(t_1), \ldots, \mathbb{U}_n(t_k))' \to_d (\mathbb{U}(t_1), \ldots, \mathbb{U}(t_k))'$. But since this holds for every k and every choice of $t_1, \ldots, t_k$, it follows that $\mathbb{U}_n \to_{f.d.} \mathbb{U}$.
(b) $\mathbb{G}_n \to_{f.d.} \mathbb{G}$. To see this, let $f_1, \ldots, f_k \in L_2(P)$. Then define random vectors $\underline{Y}_i$ by

$$\underline{Y}_i = (f_1(X_i) - Pf_1, \ldots, f_k(X_i) - Pf_k)$$

for $i = 1, \ldots, n$. Note that $E\underline{Y}_i = 0$ and

$$\begin{aligned} E\underline{Y}_1\underline{Y}_1' &= \begin{pmatrix} P(f_1^2) - (P(f_1)^2 & P(f_1f_2) - Pf_1Pf_2 & \cdots & P(f_1f_k) - Pf_1Pf_k \\ P(f_1f_2) - Pf_1Pf_2 & P(f_2^2) - (Pf_2)^2 & \cdots & P(f_2f_k) - Pf_2Pf_k \\ . & . & \cdots & . \\ . & . & \cdots & . \\ . & . & \cdots & . \\ P(f_1f_k) - Pf_1Pf_k & P(f_2f_k) - Pf_2Pf_k & \cdots & P(f_k^2) - (Pf_k)^2 \end{pmatrix} \\ &= (P(f_if_j) - Pf_iPf_j)_{i,j=1}^k \equiv \Sigma. \end{aligned}$$

Thus it follows from the multivariate central limit theorem that

$$(\mathbb{G}_n(f_1), \ldots, \mathbb{G}_n(f_k))' = \sqrt{n}\underline{Y}_n \to_d N_k(0, \Sigma).$$

But for a $P-$Brownian bridge process $\mathbb{G}_P$, $(\mathbb{G}(f_1), \ldots, \mathbb{G}(f_k))' \sim N_k(0, \Sigma)$, so we have shown that $(\mathbb{G}_n(f_1), \ldots, \mathbb{G}_n(f_k))' \to_d (\mathbb{G}(f_1), \ldots, \mathbb{G}(f_k))'$. But since this holds for every k and every choice of $f_1, \ldots, f_k \in L_2(P)$, it follows that $\mathbb{G}_n \to_{f.d.} \mathbb{G}$.

3. Ferguson, ACILST, problem 4, page 93 (modified slightly): suppose that $X_1, \ldots, X_n$ are i.i.d. F with continuous and positive density f in neighborhoods of $F^{-1}(1/4)$, $F^{-1}(1/2)$, and $F^{-1}(3/4)$.
(a) Find the asymptotic distribution of the mid-quartile range $R_n \equiv (X_{(3n/4)} + X_{(n/4)})/2$; i.e. find the asymptotic distribution of $\sqrt{n}(R_n - r)$ where $r = (F^{-1}(3/4) +$

$F^{-1}(1/4))/2$.
(b) Find the asymptotic distributions of the median.
(c) For a general distribution function F, the mid-quartile range and median estimate different parameters, the population mid-quartile range and the population median respectively, but in the case of a distribution function F that is symmetric about some point μ (so $1 - F(x + \mu) = F(x - \mu)$), they both estimate the point of symmetry, μ. Compute the asymptotic relative efficiency of the mid-quartile range relative to the median when: (i) F is Cauchy(μ, σ); (ii) F is Uniform$(0, 2\mu)$.

Solution: (a) Now

$$W_n \equiv \sqrt{n} \begin{pmatrix} \mathbb{F}_n^{-1}(1/4) - F^{-1}(1/4) \\ \mathbb{F}_n^{-1}(3/4) - F^{-1}(3/4) \end{pmatrix} \rightarrow_d \begin{pmatrix} Q'(1/4)\mathbb{V}(1/4) \\ Q'(3/4)\mathbb{V}(3/4) \end{pmatrix} \equiv W,$$

so, with $R_n \equiv (1/2)(\mathbb{F}_n^{-1}(1/4) + \mathbb{F}_n^{-1}(3/4))$, $r = (1/2)(F^{-1}(1/4) + F^{-1}(3/4))$, it follows that

$$\begin{aligned} \sqrt{n}(R_n - r) &= (1/2)\underline{1}^T W_n \rightarrow_d (1/2)\underline{1}^T W \\ &= (1/2)(Q'(1/4)\mathbb{V}(1/4) + Q'(3/4)\mathbb{V}(3/4)) \\ &\sim N(0, \frac{1}{4}(Q'(1/4)^2 \frac{3}{16} + Q'(3/4)^2 \frac{3}{16} + 2Q'(1/4)Q'(3/4)\frac{1}{16})) \\ &= N(0, \frac{1}{64}\left\{3Q'(1/4)^2 + 2Q'(1/4)Q'(3/4) + 3Q'(3/4)^2\right\}). \end{aligned}$$

When F is symmetric, $f(F^{-1}(1/4)) = f(F^{-1}(3/4)$, so $Q'(1/4) = Q'(3/4)$ and the asymptotic variance of $\sqrt{n}(R_n - r$ becomes $Q'(1/4)^2/8$.
(b) For the median we have

$$\sqrt{n}(\mathbb{F}_n^{-1}(1/2) - F^{-1}(1/2)) \rightarrow_d N(0, (1/4)Q'(1/2)^2).$$

(c) It follows from (a) and (b) that the asymptotic efficiency of R_n relative to the median is given by

$$ARE_{R_n,Med}(F) = \frac{(1/4)Q'(1/2)^2}{(1/8)Q'(1/4)^2} = 2\frac{Q'(1/2)^2}{Q'(1/4)^2} = 2\frac{f(F^{-1}(1/4))^2}{f(F^{-1}(1/2))^2}. \tag{4}$$

When F is Cauchy(μ, σ) we have

$$f(x) = \frac{1}{\sigma} f_0\left(\frac{x - \mu}{\sigma}\right), \qquad F(x) = F_0\left(\frac{x - \mu}{\sigma}\right)$$

where

$$f_0(x) = \frac{1}{\pi}\frac{1}{1 + x^2}, \qquad F_0(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x).$$

Thus $F_0^{-1}(t) = \tan(\pi(t - 1/2))$, $F^{-1}(t) = \mu + \sigma F_0^{-1}(t)$, and it follows that $f(F^{-1}(t) = f_0(F_0^{-1}(t))/\sigma$. Therefore we compute $F_0^{-1}(1/4) = \tan(-\pi/4) = -1$, $F_0^{-1}(1/2) = \tan(0) = 0$, and $f_0(F_0^{-1}(1/4) = 1/(2\pi)$, $f_0(F_0^{-1}(1/2)) = 1/\pi$. Thus the ARE computed in (4) above becomes

$$ARE_{R_n,Med}(Cauchy) = 2\frac{\left(\frac{1}{\sigma}\frac{1}{2\pi}\right)^2}{\left(\frac{1}{\sigma}\frac{1}{\pi}\right)^2} = \frac{1}{2}.$$

At the Cauchy distribution, the asymptotic variance of the median is 1/2 of the asymptotic variance of the mid-quartile range.

When F is Uniform$(0, 2\mu)$, $f(x) = (2\mu)^{-1}1_{[0,2\mu]}(x) = (2\mu)^{-1}f_0(x/2\mu)$ where $f_0(x) = 1_{[0,1]}(x)$. Therefore $f(F^{-1}(t)) = f_0(F_0^{-1}(t))/(2\mu) = 1/(2\mu)$ for all t. Thus the ARE computed in (4) above becomes

$$ARE_{R_n,Med}(Uniform) = 2\frac{(2\mu)^{-2}}{(2\mu)^{-2}} = 2.$$

(How does this change if we consider the mid-t-th quantile range defined by $R_n(t) \equiv (\mathbb{F}_n^{-1}(t) + \mathbb{F}_n^{-1}(1-t))/2$ with $0 < t < 1/2$ instead of the mid-quartile range R_n?)

4. Suppose that $X_1, \ldots, X_n$ are i.i.d. with continuous distribution function F. Let F_0 be a fixed, specified distribution function. Suppose we want to test $H: \ F = F_0$ versus $K: \ F \neq F_0$. Consider the *Cramér - von Mises statistic* given by

$$C_n^2 \equiv \int_{-\infty}^{\infty} n(\mathbb{F}_n(x) - F_0(x))^2 dF_0(x)\,.$$

(a) Show that

$$C_n^2 =_d \int_0^1 n(\mathbb{G}_n(t) - t)^2 dt\,,$$

where $\mathbb{G}_n$ is the empirical d.f. of n i.i.d. Uniform$(0, 1)$ rv's.
(b) Show that when the null hypothesis is true,

$$C_n^2 \to_d \int_0^1 \mathbb{U}(t)^2 dt$$

where $\mathbb{U}$ is a standard Brownian bridge process.
[Hint: Use the fact that $\mathbb{U}_n \Rightarrow \mathbb{U}$ in $(D[0,1], \|\cdot\|_\infty)$ and the continuous mapping theorem.]
(c) Suppose that the null hypothesis fails. Thus $F \neq F_0$. Show that in this case

$$n^{-1}C_n^2 \to_{a.s.} \int_{-\infty}^{\infty} (F(x) - F_0(x))^2 dF_0(x) > 0\,,$$

and hence the test based on C_n^2 is consistent for all $F \neq F_0$.

Solution: (a) Now $\sqrt{n}(\mathbb{F}_n - F) \stackrel{d}{=} \mathbb{U}_n(F)$ is always true (for any df F), so under the null hypothesis $F = F_0$

$$C_n^2 \equiv \int_{-\infty}^{\infty} n(\mathbb{F}_n(x) - F_0(x))^2 dF_0(x) \stackrel{d}{=} \int_{-\infty}^{\infty} [\mathbb{U}_n(F_0)]^2 dF_0$$

holds. By the change of variable $t = F_0(x)$, the variable t takes on all values in $(0, 1)$ when F_0 is continuous, and

$$\int_{-\infty}^{\infty} [\mathbb{U}_n(F_0)]^2 dF_0 = \int_0^1 [\mathbb{U}_n(t)]^2 dt.$$

Thus the stated conclusion holds.
(b) Now $\mathbb{U}_n \Rightarrow \mathbb{U}$ and $g(x) = \int_0^1 [x(t)]^2 dt$ is a continuous function from $(D[0,1], \|\cdot\|)$ to $\mathbb{R}$ (since

$$|g(x) - g(y)| = |\int_0^1 (x^2(t) - y^2(t))dt| \leq \|x - y\|_\infty \|x + y\|_\infty.$$

Thus by the continuous mapping theorem

$$C_n^2 \stackrel{d}{=} g(\mathbb{U}_n) \to_d g(\mathbb{U}) = \int_0^1 \mathbb{U}^2(t)dt.$$

(c) When $F \neq F_0$, $\|\mathbb{F}_n - F\|_\infty \to_{a.s.} 0$ by Glivenko-Cantelli, so we define $c^2 \equiv c^2(F, F_0) \equiv \int_{-\infty}^{\infty} (F(x) - F_0(x))^2 dF_0(x)$. Then

$$\begin{aligned}
&|n^{-1}C_n^2 - c^2| \\
&= \Big| \int_{-\infty}^{\infty} \{(\mathbb{F}_n(x) - F_0(x))^2 - (F(x) - F_0(x))^2\} dF_0(x) \Big| \\
&\leq \int_{\infty}^{\infty} |(\mathbb{F}_n(x) - F_0(x) - (F(x) - F_0(x))(\mathbb{F}_n(x) - F_0(x) + F(x) - F_0(x))| dF_0(x) \\
&\leq \int_{\infty}^{\infty} |\mathbb{F}_n - F| \{|\mathbb{F}_n - F_0| + |F - F_0|\} dF_0 \\
&\leq 2\|\mathbb{F}_n - F\|_\infty \to_{a.s.} 0.
\end{aligned}$$

Thus we conclude that $n^{-1}C_n^2 \to_{a.s.} c^2$.