Statistics 581, Midterm Exam Solution

Wellner; 11/07/2014

This exam is to be taken without any books or notes.

1. (24 points) **Define** any **three** of the following five terms.
   (a) A *Brownian bridge process* $\mathbb{U}$.
   (b) *Convergence in distribution* of a sequence of random vectors X_n in $\mathbb{R}^k$.
   (c) *A normal random vector* $Y = (Y_1, \ldots, Y_n)$.
   (d) A *non-central chi-square distribution* with m degrees of freedom and non-centrality parameter δ.
   (e) The *Hellinger distance* between two probability measures P and Q on a measurable space $(\mathcal{X}, \mathcal{A})$.

   Solution: See course notes, chapters 1 and 2.

   Do **either** problem 2 **or** problem 3.

2. (40 points).
   (a) State the ordinary (univariate) central limit theorem.
   (b) State the Cramér-Wold device.
   (c) State the multivariate central limit theorem.
   (d) Use the ordinary (univariate) central limit theorem (a) and the Cramér-Wold device (b) to prove the multivariate central limit theorem (c).

   Solution: (a)-(c) See course notes, chapters 1 and 2.
   (d) Let $X_1, \ldots, X_n$ be i.i.d. random vectors in $\mathbb{R}^d$ with $E(X_1^T X_1) < \infty$ so that $\Sigma \equiv E(X-\mu)(X-\mu)^T$ with $\mu \equiv E(X)$ is well-defined. Without loss of generality suppose that $\mu = 0$. We want to show that $\sqrt{n}\overline{X}_n \to_d Z \sim N_d(0, \Sigma)$. To do this let $a \in \mathbb{R}^d$, and consider

$$a^T \sqrt{n}\overline{X}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n a^T X_i \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$$

   where $Y_i = a^T X_i$ are i.i.d. with $EY_1 = 0$ and $E(Y_1)^2 = E(a^T X_1 X_1^T a) = a^T \Sigma a$. Thus it follows from the univariate (Lindeberg) central limit theorem that

$$a^T \sqrt{n}\overline{X}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i \to_d N_1(0, a^T \Sigma a) \stackrel{d}{=} a^T Z.$$

   Thus it follows from the Cramér-Wold device that $\sqrt{n}\overline{X}_n \to_d Z \sim N_d(0, \Sigma)$.

3. (40 points). State and prove the Glivenko-Cantelli theorem.

   Solution: See course notes, chapters 1 and 2.

4. (40 points). Suppose that $X_1, \ldots, X_n$ are i.i.d. with distribution function F having a continuous density function f. Let $\mathbb{F}_n$ be the empirical distribution function of the X_i's, suppose that b_n is a sequence of positive numbers, and let

$$\hat{f}_n(x) = \frac{\mathbb{F}_n(x + b_n) - \mathbb{F}_n(x - b_n)}{2b_n}.$$

   (a) Compute $E\{\hat{f}_n(x)\}$ and $Var(\hat{f}_n(x))$.
   (b) Show that $E\hat{f}_n(x) \to f(x)$ if $b_n \to 0$.
   (c) Show that $Var(\hat{f}_n(x)) \to 0$ if $b_n \to 0$ and $nb_n \to \infty$.
   (d) Use some appropriate central limit theorem to show that (perhaps under some suitable further conditions that you might need to specify)

$$\sqrt{2nb_n}(\hat{f}_n(x) - E\hat{f}_n(x)) \to_d N(0, f(x)).$$

   Hint: Write $\hat{f}_n(x)$ in terms of some Bernoulli random variables and identify $p = p_n$.

   Solution: [This $\hat{f}_n$ is a *kernel density estimator* based on the uniform kernel $k(x) = 1_{[-1,1]}(x)/2$, and can be rewritten as

$$\hat{f}_n(x) = \int_{-\infty}^{\infty} \frac{1}{b_n} k((x - y)/b_n) d\mathbb{F}_n(y);$$

   other kernel density estimators result when the uniform kernel is replaced by some other density function.]
   (a) and (b): First note that

$$2nb_n\widehat{f}_n(x) = n(\mathbb{F}_n(x + b_n) - \mathbb{F}_n(x - b_n)) = \sum_{i=1}^{n} 1_{(x-b_n, x+b_n]}(X_i) \equiv \sum_{i=1}^{n} Y_{ni}$$

   is a Binomial(n, p_n) random variable with $p_n = F(x + b_n) - F(x - b_n)$ since each $Y_{ni} \equiv 1_{(x-b_n, x+b_n]}(X_i) \sim$ Bernoulli(p_n). Hence

$$\begin{aligned} E\widehat{f}_n(x) &= \frac{F(x + b_n) - F(x - b_n)}{2b_n} = \frac{p_n}{2b_n} \\ &= \frac{1}{2}\left\{\frac{F(x + b_n) - F(x)}{b_n} + \frac{F(x) - F(x - b_n)}{b_n}\right\} \\ &\to \frac{1}{2}\{f(x) + f(x)\} = f(x). \end{aligned}$$

if $b_n \to 0$.
(a) and (c) Furthermore

$$\begin{aligned} Var(\widehat{f}_n(x)) &= \frac{np_n(1-p_n)}{(2nb_n)^2} \\ &= \frac{1}{2nb_n}\frac{p_n}{2b_n}(1-p_n) \\ &\to 0 \cdot f(x) \cdot 1 = 0 \end{aligned}$$

if $nb_n \to \infty$ and $b_n \to 0$.
(d) Since $2nb_n\widehat{f}_n(x) = \sum_{i=1}^n X_{ni}$ where $X_{ni} \sim$ Bernoulli(p_n), it follows that $\sigma_{ni}^2 = p_n(1-p_n)$ so that $\sigma_n^2 = \text{Var}(\sum_{i=1}^n X_{ni}) = np_n(1-p_n)$, and

$$\begin{aligned} \gamma_n \equiv \sum_{i=1}^n \gamma_{ni} &= \sum_{i=1}^n E|X_{ni} - \mu_{ni}|^3 \\ &= np_n(1-p_n)\{(1-p_n)^2 + p_n^2\} \\ &\leq 2np_n(1-p_n) \end{aligned}$$

so that

$$\gamma_n/\sigma^3 \leq \frac{2}{\sqrt{np_n(1-p_n)}} = \frac{2}{\sqrt{nb_n(p_n/b_n)(1-p_n)}} \to 0$$

if $b_n \to 0$ and $nb_n \to \infty$. Thus, by the Liapunov CLT,

$$\frac{2nb_n(\widehat{f}_n(x) - E\widehat{f}(x))}{\sqrt{np_n(1-p_n)}} \to N(0,1)$$

if $b_n \to 0$ and $nb_n \to \infty$. Thus

$$\begin{aligned} \sqrt{2nb_n}(\widehat{f}_n(x) - E\widehat{f}_n(x)) &= \frac{2nb_n(\widehat{f}_n(x) - E\widehat{f}_n(x))}{\sqrt{np_n(1-p_n)}}\sqrt{\frac{np_n(1-p_n)}{2nb_n}} \\ &\to N(0,1)\sqrt{f(x)} = N(0, f(x)). \end{aligned}$$

Do **either** problem 5 **or** problem 6.

5. (30 points) Suppose that $X_1, \ldots, X_n$ are i.i.d. with distribution function F, and let $\mathbb{F}_n(x) = n^{-1}\sum_{i=1}^n 1_{(-\infty,x]}(X_i)$ be the empirical distribution function of the X_i's. A famous inequality due to Dvoretzky, Kiefer, and Wolfowitz (1956) yields

$$P_F(\sqrt{n}\|\mathbb{F}_n - F\|_\infty > t) \leq C\exp(-2t^2) \tag{0.1}$$

for all F, all n, and all $t > 0$ where, by Massart (1990) $C = 2$ works.
(a) Use the inequality (0.1) to give a conservative $1 - \alpha$ confidence band for F with the dependence on n and α made explicit.
(b) Show that (0.1) implies that for any $r > 0$ we have

$$\limsup_{n\to\infty} E\|\sqrt{n}(\mathbb{F}_n - F)\|_\infty^r \le M_r$$

for some constant M_r depending only on r.

Solution: (a) Now $2e^{-2t^2} = \alpha$ if

$$t = \sqrt{\frac{1}{2}\log\left(\frac{2}{\alpha}\right)}.$$

Thus the DKW inequality yields

$$P_F\left(\|\sqrt{n}(\mathbb{F}_n - F)\|_\infty > \sqrt{\frac{1}{2}\log\left(\frac{2}{\alpha}\right)}\right) \le \alpha.$$

This implies that

$$\begin{aligned} 1-\alpha &\le P_F\left(\|\sqrt{n}(\mathbb{F}_n - F)\|_\infty \le \sqrt{\frac{1}{2}\log\left(\frac{2}{\alpha}\right)}\right) \\ &= P_F\left(\mathbb{F}_n(x) - \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)} \le F(x) \le \mathbb{F}_n(x) + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)} \text{ for all } x\right). \end{aligned}$$

Thus $\mathbb{F}_n(x) \pm \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}$ gives a conservative $1 - \alpha$ confidence band for F.
(b) If Y is a non-negative random variable, then $EY^r = \int_0^\infty r y^{r-1} P(Y \ge y)dy$. Using this formula together with the DKW inequality as improved by Massart yields

$$\begin{aligned} E\|\sqrt{n}(\mathbb{F}_n - F)\|_\infty^r &= \int_0^\infty r t^{r-1} P(\|\sqrt{n}(\mathbb{F}_n - F)\|_\infty \ge t)dt \\ &\le \int_0^\infty r t^{r-1} 2\exp(-2t^2)dt = \frac{2r}{4}\int_0^\infty t^{r-2}\exp(-2t^2)4t dt \\ &= \frac{2r}{4}\int_0^\infty (y/2)^{(r-2)/2}\exp(-y)dy \\ &= \frac{r}{2}2^{-(r-2)/2}2\Gamma(r/2) = 2^{1-r/2}(r/2)\Gamma(r+1) \\ &= 2^{1-r/2}\Gamma(1+r/2) \equiv M_r, \end{aligned}$$

and hence the claim in (b) holds.

6. (36 points).
Suppose that $X, X_1, \ldots, X_n$ are i.i.d. with distribution function F given by $P(X > x) = 1 - F(x) = 1/x^3$, $x \geq 1$, $F(x) = 0$, $x \leq 1$.
(a) For what values of $r > 0$ is $E|X|^r < \infty$? If they are finite compute $\mu = E(X)$ and $\sigma^2 = Var(X)$.
(b) Compute $F^{-1}(t) = Q(t)$, the quantile function corresponding to F and its derivative $Q'(t)$.
(c) Which of the following are true? (Briefly indicate why or why not.)

(i) $\sum_{i=1}^n X_i = O_p(n^{1/2})$.

(ii) $n^{1/3}(\overline{X}_n - \mu) = o_p(1)$.

(iii) $n^{3/4}(\overline{X}_n - \mu) = O_p(1)$.

(iv) $g(n^{1/3}(\overline{X}_n - \mu)) \to_p 1/2$ where $g(x) = 1/(1 + e^{-x})$.

(v) $h(n^{1/2}(\overline{X}_n - \mu)) = O_p(1)$ with $h(x) = 1/|x|$.

(vi) $\sqrt{n}(\mathbb{F}_n^{-1}(1/2) - F^{-1}(1/2)) \to_d N(0, (1/4)/[4(1/2)^{4/3}]^2)$.

Solution: (a) $E|X|^r = EX^r = \int_1^\infty x^r 3x^{-4} dx = 3\int_1^\infty x^{r-4} dx = 3/(3-r) < \infty$ if $r < 3$. If $r \geq 3$, then $EX^r = \infty$. Thus taking $r = 1$ yields $EX = 3/2$ and taking $r = 2$ yields $EX^2 = 3/1 = 3$. Hence $Var(X) = E(X^2) - (EX)^2 = 3 - (3/2)^2 = 12/4 - 9/4 = 3/4$.
(b) Now $F(x) = 1 - x^{-3}$ for $x \geq 1$, so solving $t = F(x) = 1 - x^{-3}$ for x gives $x = F^{-1}(t) \equiv Q(t) = (1-t)^{-1/3}$ and $Q'(t) = (1/3)(1-t)^{-4/3}$.
(c)
(i) False; since $n^{-1}\sum_1^n X_i = \overline{X}_n \to_p E(X) = 3/2$, $\sqrt{n}\overline{X} = n^{-1/2}\sum_1^n X_i \to_p \infty$.
(ii) True; since $\sqrt{n}(\overline{X}_n - \mu) \to_d N(0, \sigma^2)$, it follows that $n^{1/3}(\overline{X}_n - \mu) = n^{-1/6}\sqrt{n}(\overline{X}_n - \mu) \to_d 0 \cdot N(0, \sigma^2) = 0$ by Slutsky's theorem, and hence also this holds in probability.
(iii) False; since $n^{2/3}(\overline{X}_n - \mu) = n^{1/6}\sqrt{n}(\overline{X}_n - \mu) = n^{1/6}Z_n$ where $Z_n \to_d Z \sim N(0, \sigma^2)$, this is not $O_p(1)$.
(iv) True; since $n^{1/3}(\overline{X}_n - \mu) \to_p 0$ by (ii) and g is continuous, the continuous mapping theorem yields $g(n^{1/4}(\overline{X}_n - \mu)) \to_p g(0) = 1/2$.
(v) True; since $Z_n \equiv \sqrt{n}(\overline{X}_n - \mu) \to_d Z \sim N(0, \sigma^2)$ and h is continuous a.s. P_Z $h(Z_n) \to_d h(Z)$, and this implies that $h(Z_n) = O_p(1)$.
(vi) False (by a smidge); $F^{-1}(1/2) = (1/2)^{-1/3}$ and $f(x) = 3x^{-4}$ for $x \geq 1$. Thus $f(F^{-1}(1/2)) = 3(1/2)^{4/3} \neq 4(1/2)^{4/3}$. (Note that the asymptotic variance of the sample median $\mathbb{F}_n^{-1}(1/2)$ is $(1/4)/[3(1/2)^{4/3}]^2 = 2^{2/3}/9 \doteq 0.1764\ldots$ in comparison to the asymptotic variance of the sample mean, $Var(X_1) = 3/4$.)