Statistics 581, Midterm Exam Solutions

Wellner; 11/15/2006

1. (24 points) **Define any three of the following five terms**.
   (a) A *uniformly integrable* sequence of random variables.
   (b) *Convergence in rth mean* of a sequence of random variables.
   (c) *A normal random vector* $Y = (Y_1, \ldots, Y_n)$.
   (d) The *inverse or quantile function* F^{-1} of a distribution function F.
   (e) The *total variation distance* between two probability measures P and Q on a measurable space $(\mathcal{X}, \mathcal{A})$.

   Solution: See 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) If $X, X_1, \ldots, X_n, \ldots$ are i.i.d. with mean $\mu = E(X)$ and finite second moment $E(X^2) < \infty$, then with $\sigma^2 \equiv Var(X) = E(X-\mu)^2$,

   $$\sqrt{n}(\overline{X}_n - \mu) \to_d N(0, \sigma^2).$$

   (b) A random vector $\underline{Z}_n$ satisfies $\underline{Z}_n \to_d \underline{Z}$ in $\mathbb{R}^d$ if and only if $\underline{a}'\underline{Z}_n \to_d \underline{a}'\underline{Z}$ in $\mathbb{R}$ for all $\underline{a} \in \mathbb{R}^d$.
   (c) If $\underline{X}_1, \ldots, \underline{X}_n, \ldots$ are i.i.d. random vectors in $\mathbb{R}^d$ with mean vector $\underline{\mu} = E(\underline{X})$ and finite second moment $E(\underline{X}'\underline{X}) = E|\underline{X}|^2 < \infty$, then with $\Sigma \equiv E\{(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})'\}$,

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

   (d) By the Cramér - Wold device and (c) it suffices to show that

   $$\underline{a}'\underline{Z}_n \to_d \underline{a}'\underline{Z} \sim N(0, a'\Sigma a) \tag{0.1}$$

   in $\mathbb{R}$. But

   $$\begin{aligned} \underline{a}'\underline{Z}_n &= n^{-1/2} \sum_{i=1}^n \underline{a}'(\underline{X}_i - \underline{\mu}) \\ &= n^{-1/2} \sum_{i=1}^n Y_i \end{aligned}$$

where $Y_i \equiv \underline{a}'(\underline{X}_i - \underline{\mu})$ are i.i.d. real-valued random variables with $E(Y_i) = 0$ and finite second moment $E(Y_i^2) = Var(Y_i) = \underline{a}'\Sigma\underline{a}$. Thus (0.1) holds by the ordinary univariate CLT (a).

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

   Solution: See Chapter 2 notes.

4. (36 points). Let $X_1, \ldots, X_n$ be i.i.d. with exponential density $p_\theta(x) = \theta \exp(-\theta x) 1_{[0,\infty)}(x)$.
   (a) Find a constant c so that $c\mathbb{F}_n^{-1}(p) \to_p \theta^{-1}$.
   (b) For the $c = c_p$ you found in (a), show that $\sqrt{n}(c\mathbb{F}_n^{-1}(p) - \theta^{-1}) \to_d N(0, \sigma^2)$ and find $\sigma^2 = \sigma^2(p)$.
   (c) Show that the asymptotic variance $\sigma^2(p)$ is minimized by p satisfying $2p = -\log(1-p)$.

   Solution: (a) Now $F(x) = 1 - \exp(-\theta x)$, so $F^{-1}(t) = -\theta^{-1}\log(1-t)$ for $0 < t < 1$, and $F^{-1}(p) = \theta^{-1}\{-\log(1-p)\}$. Thus $c\mathbb{F}_n^{-1}(p) \to cF^{-1}(p) = \theta^{-1}$ if we take $c = 1/\{-\log(1-p)\} = 1/\log(1/(1-p))$.
   (b) Recall that

$$\sqrt{n}(\mathbb{F}_n^{-1}(p) - F^{-1}(p)) \to_d Q'(p)\mathbb{V}(p) \sim N(0, p(1-p)Q'(p)^2)$$

where $Q(p) = F^{-1}(p) = -\theta^{-1}\log(1-p)$ and hence $Q'(p) = \theta^{-1}/(1-p)$. Thus the asymptotic variance in the last display is $\theta^{-2}p/(1-p)$. From this it follows easily that

$$\begin{aligned}
\sqrt{n}(c_p\mathbb{F}_n^{-1}(p) - \theta^{-1}) &= c_p\sqrt{n}(\mathbb{F}_n^{-1}(p) - F^{-1}(p)) \\
&\to_d c_p Q'(p)\mathbb{V}(p) \sim N(0, \theta^{-2}c_p^2 p/(1-p)) \equiv N(0, \theta^{-2}v(p))
\end{aligned}$$

(c) The asymptotic variance in (b) is minimized when the derivative of $v(p) \equiv c_p^2 p/(1-p)$ equals zero, or equivalently when

$$\begin{aligned}
0 = \frac{d}{dp}\log v(p) &= \frac{d}{dp}\{\log p - \log(1-p) - 2\log\{-\log(1-p)\}\} \\
&= \frac{1}{p} + \frac{1}{1-p} - 2\frac{1}{-\log(1-p)}\frac{1}{1-p} \\
&= \frac{1}{1-p}\left\{\frac{1}{p} + \frac{2}{\log(1-p)}\right\},
\end{aligned}$$

and this equality holds when $2p = -\log(1-p)$. Thus the optimal p for estimation of θ^{-1} is $p = p_0 \doteq 0.796812....$

5. (36 points)
Suppose that $X, X_1, \ldots, X_n$ are i.i.d. Exponential(θ) random variables so that $P_\theta(X > x) = \exp(-\theta x) = 1 - F_\theta(x)$ for $x > 0$.
(a) Fix $x_0 > 0$ and let $\mathbb{F}_n(x) = n^{-1}\sum_{i=1}^n 1_{[X_i \le x]} = n^{-1}\sum_{i=1}^n 1_{(-\infty,x]}(X_i)$ denote the empirical distribution function. Show that

$$\sqrt{n}\begin{pmatrix} \overline{X}_n - 1/\theta \\ \mathbb{F}_n(x_0) - F_\theta(x_0) \end{pmatrix} \to_d Y \sim N_2(0, \Sigma)$$

and find Σ.
(b) Let $g(\theta) \equiv F_\theta(x_0) = 1 - \exp(-\theta x_0)$, and consider the two estimators of $F = F_\theta$ given by $T_{n,1} \equiv g(\widehat{\theta}_n)$ and $T_{n,2} \equiv \mathbb{F}_n(x_0)$ where $\widehat{\theta}_n \equiv 1/\overline{X}_n$. Show that

$$\sqrt{n}\begin{pmatrix} T_{n,1} - F_\theta(x_0) \\ T_{n,2} - F_\theta(x_0) \end{pmatrix} \to_d \tilde{Y}$$

and find the distribution of $\tilde{Y}$.
(c) What is the advantage of $T_{n,2} = \mathbb{F}_n(x_0)$ as an estimator even though it is inefficient when the exponential model holds?

Solution: (a) Note that

$$\begin{aligned} \sqrt{n}\begin{pmatrix} \overline{X}_n - 1/\theta \\ \mathbb{F}_n(x_0) - F_\theta(x_0) \end{pmatrix} &= n^{-1/2}\sum_{i=1}^n \begin{pmatrix} X_i - \mu \\ 1_{[0,x_0]}(X_i) - F_\theta(x_0) \end{pmatrix} \\ &\equiv n^{-1/2}\sum_{i=1}^n \underline{Y}_i \end{aligned}$$

where $\underline{Y}_i$ are i.i.d. with $E(\underline{Y}_i) = 0$ and

$$\begin{aligned} E(\underline{YY'}) &= \begin{pmatrix} Var_\theta(X) & E_\theta(X-\mu)1_{[0,x_0]}(X) \\ E_\theta(X-\mu)1_{[0,x_0]}(X) & F_\theta(x_0)(1-F_\theta(x_0)) \end{pmatrix} \\ &= \begin{pmatrix} \theta^{-2} & -x_0 e^{-\theta x_0} \\ -x_0 e^{-\theta x_0} & e^{-\theta x_0}(1 - e^{-\theta x_0}) \end{pmatrix} \equiv \Sigma. \end{aligned}$$

Thus the multivariate CLT yields

$$\sqrt{n}\begin{pmatrix} \overline{X}_n - 1/\theta \\ \mathbb{F}_n(x_0) - F_\theta(x_0) \end{pmatrix} \to_d \underline{Y} \sim N_2(0, \Sigma)$$

where Σ is as given in the last display.
(b) Note that

$$T_{n1} = g(\hat{\theta}_n) = 1 - \exp(-\hat{\theta}_n x_0) = 1 - \exp(-x_0/\overline{X}_n) \equiv h(\overline{X}_n)$$

where $h(u) \equiv 1 - \exp(-x_0/u)$ has $h'(u) = -u^{-2} x_0 \exp(-x_0/u)$. Thus we can use the delta-method to proceed as follows:

$$\begin{aligned}
\sqrt{n}\begin{pmatrix} T_{n,1} - F_\theta(x_0) \\ T_{n,2} - F_\theta(x_0) \end{pmatrix} &= \sqrt{n}\begin{pmatrix} h(\overline{X}_n) - h(1/\theta) \\ \mathbb{F}_n(x_0) - F_\theta(x_0) \end{pmatrix} \\
&\to_d \begin{pmatrix} h'(1/\theta) & 0 \\ 0 & 1 \end{pmatrix} \underline{Y} \sim N_2(0, \tilde{\Sigma})
\end{aligned}$$

where

$$\begin{aligned}
\tilde{\Sigma} &= \begin{pmatrix} h'(1/\theta)^2 \theta^{-2} & -h'(1/\theta) x_0 \exp(-\theta x_0) \\ -h'(1/\theta) x_0 \exp(-\theta x_0) & \exp(-\theta x_0)(1 - \exp(-\theta x_0)) \end{pmatrix} \\
&= \begin{pmatrix} \theta^2 x_0^2 \exp(-2\theta x_0) & \theta^2 x_0 \exp(-\theta x_0) x_0 \exp(-\theta x_0) \\ \theta^2 x_0 \exp(-\theta x_0) x_0 \exp(-\theta x_0) & \exp(-\theta x_0)(1 - \exp(-\theta x_0)) \end{pmatrix}.
\end{aligned}$$

(c) The advantage of $T_{n2} = \mathbb{F}_n(x_0)$ is that it is a consistent estimator of $F(x_0) = P(X \le x_0)$ even when the exponential model fails.

6. (36 points).
Suppose that $X, X_1, \ldots, X_n$ are i.i.d. with distribution function F given by $P(X > x) = 1 - F(x) = 1/x^5$, $x \ge 1$, $F(x) = 0$, $x \le 1$.
(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.
(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 $h(x) = \log|x|$.
(vi) $\sqrt{n}(\mathbb{F}_n^{-1}(1/2) - F^{-1}(1/2)) \to_d N(0, (1/4)/[5(1/2)^{6/5}]^2)$.

Solution:
(a) Since X has density $f(x) = 5x^{-6}$ $E|X|^r = 5\int_1^\infty x^{r-6} dx = 5/(5-r) < \infty$ if $r < 5$. Another way to compute this is to note that since $X \ge 0$ with probability 1 we have

$$\begin{aligned}
EX^r &= \int_0^\infty r x^{r-1}(1 - F(x)) dx = r\int_0^1 x^{r-1} dx + r\int_1^\infty x^{r-6} dx \\
&= 1 + \frac{r}{r-5} x^{r-5}\Big|_1^\infty = 1 + \frac{r}{5-r} \qquad \text{if } \ r < 5 \\
&= \frac{5}{5-r}.
\end{aligned}$$

(b) Since $F(x) = 1 - x^{-5}$ for $x \geq 1$, the quantile function is obtained by solving $t = F(x) = 1 - x^{-5}$ to obtain $x = F^{-1}(t) = (1-t)^{-1/5}$ for $0 < t < 1$.
(c) (i) is false since $E(X) = 5/4 > 0$ from (a). Thus $\sum_{i=1}^{n} X_i = O_p(n)$ but not $O_p(n^{1/2})$.
(ii) is true since $E(X^2) = 5/3 < \infty$ implies that $n^{1/2}(\overline{X}_n - \mu) = O_p(1)$ by the CLT.
(iii) is false for the same reason as in (ii).
(iv) is true by (ii) and continuous mapping since $g(0) = 1/2$.
(v) is true by the continuous mapping theorem since $n^{1/2}(\overline{X}_n - \mu) \rightarrow_d Z \sim N(0, V^2)$ with $V^2 = 5/3 - (5/4)^2$ and $h(x) = \log|x|$ is continuous a.s. with respect to the distribution of Z since it has just one discontinuity (at 0).
(vi) is true by the computation in (b): $Q(t) = F^{-1}(t) = (1-t)^{-1/5}$ so $Q'(t) = (1/5)(1-t)^{-6/5}$ and hence $Q'(1/2) = (1/5)(1/2)^{-6/5}$.