Statistics 583, Midterm Exam Solutions

Wellner; 5/9/2011

1. (30 points) **Define** *any three* of the following terms.
   (a) A maximal invariant with respect to a group G of transformations g on the sample space $\mathcal{X}$).
   (b) A $G-$*invariant test function* ϕ with respect to a group G.
   (c) The Prohorov metric d_{Pr} on the set of probability measures $\mathcal{P}$ on a metric space (S, d).
   (d) A metric d_* compatible with the empirical distribution function or empirical measure.
   (e) A Hadamard-differentiable functional $T : \mathcal{F} \to \mathbb{R}$ with respect to the Kolmogorov metric d_K.

   Solution: See Chapter 6 and 7 notes.

2. (30 points) Give a complete **statement** of *any two* of the following results:
   (a) The defining connection between a group G on a sample space $\mathcal{X}$ and the induced group $\overline{G}$ on a parameter space Θ assuming that $X \sim P_\theta$.
   (b) Any theorem about consistency of an estimator via continuity of statistical functionals.
   (c) Any theorem about asymptotic normality of an estimator via differentiability of the corresponding statistical functional.
   (d) The Wald- Wolfowitz-Noether-Hájek finite sampling central limit theorem.

   Solution: See Chapter 6 and 7 notes.

Do either problem 3 or problem 4.

3. (42 points) Suppose that $X_1, \ldots, X_m$ are i.i.d. exponential(λ) and that $Y_1, \ldots, Y_n$ are i.i.d. exponential(μ). Thus the density of X_1 is $\lambda \exp(-\lambda x) 1_{[0,\infty)}(x)$. Consider testing $H : \lambda \leq \mu$ versus $K : \lambda > \mu$.
(a) Show that this testing problem is invariant with respect to the group of scale changes G given by $g_c(\underline{x}, \underline{y}) = (c\underline{x}, c\underline{y})$ where $c > 0$.
(b) Find the UMP G-invariant test of H versus K. [Hint you may use the fact that the family of distributions $\delta^{-1} F_{r,s} : \delta > 0$ has monotone likelihood ratio.
(c) Specify as exactly as possible how you would carry out the test derived in (b).

Solution: (a) If $X \sim$ exponential(λ), then

$$\begin{aligned} P_\lambda(cX > t) &= P_\lambda(X > t/c) = \exp(-\lambda t/c) \\ &= \exp(-(\lambda/c)t) = P_{\lambda/c}(X > t)\,, \end{aligned}$$

and similarly for $Y \sim$ exponential(μ). Hence the induced group on the parameter space is $\overline{g}(\lambda, \mu) = (\lambda/c, \mu/c)$. Note that for any $\overline{g} \in \overline{G}$ we have $\overline{g}\Theta_0 = \{(\lambda/c, \mu/c) : \lambda \leq \mu\} = \{(\lambda, \mu) : \lambda \leq \mu\} = \Theta_0$, and $\overline{g}\Theta = \{(\lambda/c, \mu/c) : (\lambda, \mu) \in R^+ \times R^+\} = \Theta$. Hence the testing problem is invariant under the group G.
(b) By sufficiency we may reduce to consideration of $(S, T) = (\sum_{i=1}^m X_i, \sum_{j=1}^n Y_j)$. The induced group G^* on the space of the sufficent statistic is given by $G^* = \{g^*(s,t) = (cs, ct) : c > 0\}$, and a maximal invariant for the group G^* is $V = T/S$; a corresponding $\overline{G}-$maximal invariant is $\delta = \lambda/\mu$. Now $2\lambda X_i \sim \chi^2_2$, and similarly $2\mu Y_j \sim \chi^2_2$. Hence $2\lambda S \sim \chi^2_{2m}$ and $2\mu T \sim \chi^2_{2n}$. Hence

$$\frac{m}{n} V = \frac{\lambda}{\mu} \cdot \frac{2\mu T/2n}{2\lambda S/2m} = \delta F_{2n,2m}$$

where $F_{2n,2m}$ has an $F-$distribution with degrees of freedom $2n, 2m$. Note that $H : \lambda \leq \mu$ and $K : \lambda > \mu$ correspond to $H : \delta \leq 1$ and $K : \delta > 1$. Since the family $\delta F_{r,s}$ has monotone (increasing) likelihood ratio, we conclude that the UMP G-invariant test of H versus K is given by "reject H if $mV/n > F_{2n,2m,\alpha}$" where $P(F_{2n,2m} \geq F_{2n,2m,\alpha}) = \alpha$.
(c) See above.

4. (42 points) Suppose that $X_1, \ldots, X_m$ are i.i.d. $F \in \mathcal{F}_c$, the set of all continuous d.f.'s on R, and that $Y_1, \ldots, Y_n$ are i.i.d. $G \in \mathcal{F}_c$ where, for some $\theta \in R$,

$$\frac{1 - G(x)}{G(x)} = e^\theta \frac{1 - F(x)}{F(x)}$$

for all x. Consider testing $H : \theta = 0$ versus $K : \theta > 0$.
(a) Under what group of transformations G is this testing problem invariant?

(b) What is the maximal invariant $T(\underline{X}, \underline{Y})$ for the group G?
(c) What is the $\overline{G}-$maximal invariant on the parameter space?
(d) What does Hoeffding's formula say about the distribution of the maximal invariant under the alternative K?
(e) Use (d) to find the locally most powerful rank test of H versus K. What is the name of this test statistic?

Solution: (This was part of problem 1, problem set # 3.)
(a) the problem is invariant under the group of strictly monotone increasing functions from R to R (applied to each coordinate of $\underline{Z} \equiv (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$). This follows since for any such function f, $P_F(f(X) \leq x) = P_F(X \leq f^{-1}(x)) = F(f^{-1}(x)) \equiv \tilde{F}(x)$ for a continuous d.f. $\tilde{F}$ and similarly $P_G(f(Y) \leq y) = P_F(Y \leq f^{-1}(y)) = G(f^{-1}(y)) = \tilde{G}(y)$ for a continuous d.f. $\tilde{G}$. Moreover

$$\begin{aligned} \frac{1-\tilde{G}(x)}{\tilde{G}(x)} &= \frac{1-G(f^{-1}(x))}{G(f^{-1}(x))} = e^{\theta}\frac{1-F(f^{-1}(x))}{F(f^{-1}(x))} \\ &= e^{\theta}\frac{1-\tilde{F}(x)}{\tilde{F}(x)}, \end{aligned}$$

so that the proportional odds hypothesis is preserved under the group G.
(b) The maximal invariant under the group G is the vector of ranks $\underline{R} = (R_1, \ldots, R_N)$ where $R_i = N\mathbb{H}_N(Z_i)$ for $i = 1, \ldots, N$.
(c) The $\overline{G}-$maximal invariant on the parameter space (F, G) is $\psi(u) = G \circ F^{-1}(u) = G(F^{-1}(u))$. For the proportional odds model as given,

$$\frac{1-G(x)}{G(x)} = e^{\theta}\frac{1-F(x)}{F(x)},$$

implies that

$$G(x) = \frac{F(x)}{F(x) + e^{\theta}(1-F(x))}$$

after simple algebra, and hence that

$$\psi(u) = G \circ F^{-1}(u) = \frac{u}{u + e^{\theta}(1-u)}.$$

(d) Hoeffding's formula says that

$$P_{\theta}(\underline{Q} = \underline{q}) = \frac{1}{\binom{N}{n}} E_U \left\{ \prod_{j=1}^{n} \psi'_{\theta}(U_{(q_j)}) \right\}.$$

(e) The locally most powerful rank test rejects for those values $\underline{q}$ of $\underline{Q}$ which

make

$$\begin{aligned}\frac{\partial}{\partial\theta}P_\theta(\underline{Q}=\underline{q})\Big|_{\theta=0} &= \frac{1}{\binom{N}{n}}E_U\left\{\sum_{j=1}^{n}\frac{\partial}{\partial\theta}\psi'_\theta(U_{(q_j)})\Big|_{\theta=0}\right\}\\ &= \frac{1}{\binom{N}{n}}\sum_{j=1}^{n}E_U\left\{\frac{\partial}{\partial\theta}\psi'_\theta(U_{(q_j)})\Big|_{\theta=0}\right\}\end{aligned}$$

large. Hence it remains only to calculate

$$\phi(u)\equiv\frac{\partial}{\partial\theta}\psi'_\theta(u)|_{\theta=0}$$

and $E\phi(U_{(i)})$ for the alternative in question. Here we have

$$\psi'_\theta(u)=\frac{e^\theta}{[u+e^\theta(1-u)]^2}.$$

Hence

$$\begin{aligned}\frac{\partial}{\partial\theta}\psi'_\theta(u)|_{\theta=0} &= \psi'_\theta(u)\Big|_{\theta=0}-\left\{2\frac{\psi'_\theta(u)}{u+e^\theta(1-u)}\right\}\cdot\{e^\theta(1-u)\}\Big|_{\theta=0}\\ &= 1-2(1-u)=2u-1.\end{aligned}$$

Since $EU_{(i)}=i/(N+1)$, the locally most powerful rank test of H versus this proportional odds alternative K is the Wilcoxon test "reject H if $S_N=\sum_{j=1}^{n}Q_j>k_\alpha$"; i.e. the locally most powerful rank test is the Wilcoxon rank sum test.

Do either problem 5 or problem 6.

5. (50 points) Consider the functional $T(F)=\iint|x-y|dF(x)dF(y)=E_F|X-Y|$ where X and Y are independent with distribution function F. This is a measure of spread or dispersion of the distribution function F. (The functional $T(F)$ is sometimes called "Gini's mean difference".)
   (a) If $F=N(\mu,\sigma^2)$, compute $T(F)$.
   (b) If $X_1,\ldots,X_n$ are i.i.d. random variables with distribution function F, what is the "principle of substitution" estimator V_n of $T(F)$?
   (c) Give an unbiased estimator U_n of $T(F)$ that is symmetric in all the observations $X_1,\ldots,X_n$.
   (d) Show that the estimator you found in (b) is a biased estimator of $T(F)$, calculate the bias explicitly, and show that $\sqrt{n}(U_n-V_n)=o_p(1)$.
   (e) Compute the influence function of $T(F)$.
   (f) Use the result of (d) and (e) or results from class to show that: $\sqrt{n}(V_n-T(F))\to_d N(0,v^2(F))$, $\sqrt{n}(U_n-T(F))\to_d N(0,v^2(F))$, and compute $v^2(F)$ explicitly.

Solution: (a) If $X, Y \sim N(\mu, \sigma^2)$ are independent, then with $Z \sim N(0,1)$, the difference $X - Y \stackrel{d}{=} \sqrt{2}\sigma Z \sim N(0, 2\sigma^2)$. Hence

$$T(F) = E|X - Y| = \sqrt{2}\sigma E|Z| = 2\sigma/\sqrt{\pi} \approx 1.12838\ldots\sigma$$

since $E|Z| = 2\int_0^\infty z\phi(z)dz = 2\int_0^\infty(-\phi'(z))dz = 2\phi(0) = \sqrt{2/\pi}$.
(b) The principle of substitution estimator is just

$$T(\mathbb{F}_n) = \int\int |x-y| d\mathbb{F}_n(x) d\mathbb{F}_n(y) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n |X_i - X_j|.$$

(c) An unbiased estimator of $T(F)$ is

$$U_n = \frac{2}{n(n-1)}\sum\sum_{i<j}|X_i - X_j|.$$

(d) The estimator $T(\mathbb{F}_n)$ is biased: because the diagonal terms (those with $j = i$) in the sum are zero we have

$$\begin{aligned} E_F T(\mathbb{F}_n) &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n E_F|X_i - X_j| \\ &= \frac{n(n-1)}{n^2}E_F|X_1 - X_2| = \frac{n-1}{n}T(F). \end{aligned}$$

Thus the bias of $T_n = T(\mathbb{F}_n)$ is

$$\text{bias}_n(F) = E_F(T_n) - T(F) = \left(\frac{n-1}{n} - 1\right)T(F) = -\frac{1}{n}T(F).$$

Furthermore,

$$\begin{aligned} U_n - V_n &= \frac{1}{n(n-1)}\sum_{i\neq j}|X_i - X_j| - \frac{1}{n^2}\sum_{i\neq j}|X_i - X_j| \\ &= \frac{1}{n}\left(\frac{1}{n-1} - \frac{1}{n}\right)\sum_{i\neq j}|X_i - X_j| \\ &= \frac{1}{n}U_n \geq 0. \end{aligned}$$

Thus $\sqrt{n}E(U_n - V_n) = n^{-1/2}E(U_n) = n^{-1/2}T(F) \to 0$, and this implies that $\sqrt{n}(U_n - V_n) \to_p 0$ by Markov's inequality.
(e) To calculate the Gateaux derivative, we write $F_\epsilon = (1-\epsilon)F + \epsilon G$ and then

compute

$$\begin{aligned} T(F_\epsilon) &= \iint |x-y| dF_\epsilon(x) dF_\epsilon(y) \\ &= \iint |x-y| dF(x) dF(y) + \epsilon \iint |x-y| d(G-F)(x) dF(y) \\ &\quad + \epsilon \iint |x-y| dF(x) d(G-F)(y) \\ &\quad + \epsilon^2 \iint |x-y| d(G-F)(x) d(G-F)(y). \end{aligned}$$

Hence we find that the Gateaux derivative $\dot{T}(F; G-F)$ is given by

$$\begin{aligned} \dot{T}(F; G-F) &= \frac{d}{d\epsilon} T(F_\epsilon)\Big|_{\epsilon=0} \\ &= \iint |x-y| d(G-F)(x) dF(y) + \iint |x-y| dF(x) d(G-F)(y) \\ &= 2 \iint |x-y| d(G-F)(x) dF(y). \end{aligned}$$

Taking $G = \delta_x$ yields the influence function for T:

$$\begin{aligned} IC(x; T, F) &= 2 \left(\int |x-y| dF(y) - \iint |x-y| dF(x) dF(y) \right) \\ &= 2 \left(\int |x-y| dF(y) - T(F) \right). \end{aligned}$$

(e) The influence function calculations in (c) lead us to expect that

$$\sqrt{n}(T(\mathbb{F}_n) - T(F)) \equiv \sqrt{n}(V_n - T(F)) \to_d N(0, V^2)$$

where the variance V^2 is given by

$$\begin{aligned} V^2 &= E_F \psi_F^2(X) = 4 \int \left(\int |x-y| dF(y) - T(F) \right)^2 dF(x) \\ &= 4 \left\{ \iint |x-y| dF(y) \iint |x-y'| dF(y') - T^2(F) \right\} \\ &= 4 \left\{ \iiint |x-y||x-y'| dF(x) dF(y) dF(y') - T(F)^2 \right\}. \end{aligned}$$

This can be proved rigorously under the assumption that $0 < E_F X^2 < \infty$, and hence $0 < E_F |X-Y|^2 < \infty$: note that the corresponding $U-$statistics $U_n \equiv (2/(n(n-1))) \sum_{i<j} |X_i - X_j|$ satisfy

$$\sqrt{n}(U_n - T(F)) \to_d N(0, V^2)$$

via Hájek projection arguments (as was proved in class) and then using $\sqrt{n}(V_n - U_n) = o_p(1)$ as shown in (d).

6. (50 points) Suppose that an urn contains N balls with the numbers $a_N(1), \ldots, a_N(N)$ written on the balls. Suppose that a sample of n balls is drawn from the urn without replacement: let the numbers on the sampled balls be $Y_1, \ldots, Y_n$, and let $\overline{Y}_n = n^{-1} \sum_{i=1}^n Y_i$.
   (a) What is the mean of $\overline{Y}_n$?
   (b) What is the variance of $\overline{Y}_n$?
   (c) If $\underline{R} = (R_1, \ldots, R_N)$ is a random permutation of $\{1, \ldots, N\}$, what is the relationship between $n\overline{Y}_n$ and $\sum_{j=1}^n a_N(R_j)$?
   (d) Under some condition on the numbers $a_N(i)$, a CLT holds for an appropriately standardized version of $\overline{Y}_n$, and hence also for $\sum_{j=1}^n a_N(R_j)$. State this condition and the theorem.
   (e) If $a_N(j) = j$, $j = 1, \ldots, N$, compute the mean and variance in (a) and (b), and make the conclusion of the CLT in (d) explicit. What is the name of the corresponding rank test and when is it a locally most powerful test based on the ranks?

Solution: (a)

$$\begin{aligned} E(\overline{Y}_n) &= \frac{1}{n}\sum_{i=1}^n E(Y_i) = \frac{1}{n}\sum_{i=1}^n \frac{1}{N}\sum_{j=1}^N a_N(j) \\ &= \frac{1}{N}\sum_{j=1}^N a_N(j) \equiv \overline{a}_N. \end{aligned}$$

(b)

$$\sigma_N^2 \equiv Var(\overline{Y}_n) = \left(1 - \frac{n-1}{N-1}\right)\frac{\sigma_a^2}{n}$$

where $\sigma_a^2 = N^{-1}\sum_{j=1}^N (a_N(j) - \overline{a}_N)^2$.
(c) $n\overline{Y}_n$ and $\sum_{j=1}^n a_N(R_j)$ have exactly the same distribution. (In fact, one way to generate the sample $Y_1, \ldots, Y_n$ is to first generate a random permutation of the first N integers, and then to identify the sample drawn from the urn as $Y_i = a_N(R_i)$ for $i = 1, \ldots, n$.
(d) If $0 < \underline{\lim}(n/N) \leq \overline{\lim}(n/N) < 1$, then the Noether condition

$$\eta_N \equiv \frac{\max_{1\leq i\leq N} |a_N(i) - \overline{a}_N|^2}{\sum_{i=1}^N (a_N(i) - \overline{a}_N)^2} \to 0$$

holds if and only if

$$\frac{\overline{Y}_n - \overline{a}_N}{\sigma_N} \to_d N(0,1)$$

where σ_N^2 is the variance of $\overline{Y}_n$ as defined in (b).
(e) When $a_N(j) = j$ for $j = 1, \ldots, N$, then

$$\begin{aligned}
\overline{a}_N &= N^{-1} \sum_{j=1}^{N} j = N^{-1} N(N+1)/2 = (N+1)/2, \\
\sigma_a^2 &= N^{-1} N(N+1)(2N+1)/6 - (N+1)^2/4 \\
&= \frac{N+1}{2} \left(\frac{2N+1}{3} - \frac{N+1}{2} \right) \\
&= \frac{N+1}{2} \cdot \frac{N-1}{6} = \frac{N^2-1}{12}.
\end{aligned}$$

Thus

$$\sigma_N^2 = \left(1 - \frac{n-1}{N-1}\right) \frac{\sigma_a^2}{n} = \frac{m}{N-1} \frac{N^2-1}{12n} = \frac{m(N+1)}{12n}.$$

Then the CLT in (d) becomes

$$\frac{\overline{Y}_n - (N+1)/2}{\sqrt{m(N+1)/12n}} \to_d N(0,1).$$

This describes the asymptotic behavior of the Wilcoxon rank sum statistic under the null hypothesis. This test statistic is locally most powerful for a two sample location shift problem in which the distribution of the data is (a shift) of a logistic distribution. It is also locally most powerful for the Lehmann alternative given by proportional odds as in the preceding problem.