Statistics 583, Final Exam

Wellner; 6/5/2006

Instructions: This is an "in class" and "closed-book" exam. Please do it completely on your own with no books or notes.

1. (40 points) **Define** the following terms.
   (a) A Fréchet - differentiable functional $T : \mathcal{F} \to \mathbb{R}$ with respect to a metric d_* on $\mathcal{F}$.
   (b) A Hadamard - differentiable functional $T : \mathcal{F} \to \mathbb{R}$ with respect to a metric d_* on $\mathcal{F}$.
   (c) A metric d between distribution functions which is compatiable with respect to the empirical distribution function.
   (d) The kernel estimator of a density function f on $\mathbb{R}$.

2. (30 points). Give a complete *statement* of **two** of the following results or theorems:
   (a) An example of a functional $T(F)$ which is *not* weakly continuous.
   (b) A limit theorem for the the bootstrap empirical process $\sqrt{m}(\mathbb{F}_m^* - \mathbb{F}_n)$ when $m \wedge n \to \infty$.
   (c) Some version of Hoeffding's theorem about the distribution of the rank vector $\underline{R}$ under alternatives.
   (d) Any theorem about asymptotic normality of an estimator via differentiability of the corresponding statistical functional.

3. (40 points). Suppose that $X_1, \ldots, X_m$ are i.i.d. with continuous d.f. F on $\mathbb{R}^+$ and $Y_1, \ldots, Y_n$ are i.i.d. with continuous d.f. G where G and F are related by $G = F^\Delta$ where $\Delta > 1$. Let $\underline{R} = (R_1, \ldots, R_n)$ denote the ranks of $\underline{Z} = (Z_1, \ldots, Z_N) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ where $N = m + n$, and let $\underline{Q} = (Q_1, \ldots, Q_n)$ denote the ordered Y ranks.
   (a) If $m = 1$, $n = 2$, and $\Delta = 1$, calculate $P_\Delta(\underline{Q} = \underline{q})$ for all possible values of $\underline{q}$.
   (b) If $m = 1$, $n = 2$, and $\Delta = 2$, calculate $P_\Delta(\underline{Q} = \underline{q})$ for all possible values of $\underline{q}$.
   (c) What is the locally most powerful rank test S_N for testing $H : F = G$ versus $K : G = F^\Delta$, $\Delta > 1$ (for general m and n)? [Hint: you may use the fact that the order statistics $Y_{(i)}$ of N i.i.d. exponential(1) random variables satisfy $Y_{(i)} \stackrel{d}{=} \sum_{j=1}^{i} V_j/(N - j + 1)$ where $V_1, \ldots, V_N$ are i.i.d. exponential (1).]
   (d) If $m = 1$ and $n = 2$ as in (a) and (b), and you observe $\underline{Q} = (2, 3)$, what is the p-value of the locally most powerful rank test in (c)?
   (e) What is the asymptotic distribution of S_N under the null hypothesis?

4. (40 points). Suppose that $X_1, \ldots, X_n$ are i.i.d. F with density function f having $p > 2$ continuous derivatives at a point x. Suppose that we use a "kernel estimator" $\widehat{f_n}(x)$ based on the bandwidth $h = h_n$ and kernel k *of order* p: i.e. k satisfies

$$\int k(z)dz = 1, \qquad \int zk(z)dz = 0, \ldots, \int z^{p-1}k(z)dz = 0,$$
$$\int |z|^p k(z)dz < \infty, \qquad \int k^2(z)dz < \infty.$$

Such a kernel cannot be a probability density function since the condition $\int z^2 k(z)dz = 0$ forces k to take negative values; thus such kernels are sometimes called "higher-order kernels". For example, $k(x) = 8^{-1}(9 - 15x^2)1_{[-1,1]}(x)$ is a kernel of order $p = 4$.

(a) Use the same method as in class and homework to show that the resulting estimator $\widehat{f_n}(x)$ has bias given by

$$E\{\widehat{f_n}(x)\} - f(x) = \frac{h_n^p}{p!}(-1)^p \int z^p k(z) f^{(p)}(x - h_n z)dz$$

(b) Use the same methods as in class and homework to show that $\widehat{f_n}(x)$ has variance

$$Var(\widehat{f_n}(x)) = \frac{f(x)}{nh_n} \int k^2(z)dz + o((nh_n)^{-1}).$$

(c) Combine (a) and (b) to find an asymptotic expression for $E(f(x) - \widehat{f_n}(x))^2$, and hence show that $\widehat{f_n}(x)$ achieves the optimal rate of convergence $n^{p/(2p+1)}$ with optimal bandwidth choice $h_{n,opt} = n^{-1/(2p+1)}$.

Do **one** of problem 5 **or** problem 6 (but **not both**).

5. (40 points). Suppose that H is a bivariate distribution function of a pair of positive random variables (X, Y) with marginal distribution functions F and G, and with $EX^4 < \infty$, $EY^4 < \infty$, $\mu(F) > 0$, and $\sigma^2(G) = Var_G(Y) > 0$. Consider the functional

$$T(H) = \frac{\sigma(F)/\mu(F)}{\sigma(G)/\mu(G)}$$

the ratio of the marginal *coefficients of variation* $cv(F) \equiv \sigma(F)/\mu(F)$ and $cv(G) \equiv \sigma(G)/\mu(G)$; here $\mu(F) = E_F(X)$, $\sigma^2(F) = Var_F(X)$ and similarly for G. Suppose that we observe i.i.d. pairs (X_i, Y_i) from the distribution H and estimate $T(H)$ by $T_n \equiv T(\mathbb{H}_n)$ where $\mathbb{H}_n$ is the empirical distribution function (or empirical

measure) of the pairs.

(a) Explain how you would use the jackknife to estimate $nVar_H(T_n)$.

(b) Explain how you would use the bootstrap to estimate $nVar_H(T_n)$. Discuss both the ideal bootstrap estimator and the Monte-Carlo implementation thereof.

(c) Do you believe that $\sqrt{n}(T_n - T(H)) \to_d N(0, V^2)$ for some V^2 under the above hypotheses? Why or why not? What transformation g of T_n might lead to a better approximation using this asymptotic distribution?

(d) Will the jackknife estimator of variance "work" in this situation? Will the bootstrap estimator of variance "work" in this situation?

6. (40 points). In the context of testing for a disease, let $X \sim F$ denote the outcome of the test for a diseased individual and let $Y \sim G$ denote the outcome of the test for a non-diseased individual. Assuming that $X > x$ (or $Y > x$) leads to classifying the individual as "diseased", the *Receiver Operating Characteristic* or ROC curve R is a plot of *sensitivity* $\equiv P(X > x) = 1 - F(x) \equiv \overline{F}(x)$ versus $1-$*specificity*$\equiv 1 - P(Y \leq x) = P(Y > x) = 1 - G(x) \equiv \overline{G}(x)$. Thus the ROC curve $R = R_{F,G}$ can be written as

$$R(t) = \overline{F}(\overline{G}^{-1}(t)) = 1 - F(G^{-1}(1-t)), \qquad 0 < t < 1.$$

(a) A good test for a disease has ROC curve with values close to 1 for small t and is everywhere above the line $I(t) = t$. Show that $R(t) \geq t$ for $0 \leq t \leq 1$ with strict inequality for some t if and only if $G <_s F$ (i.e. G is stochastically smaller than F).

(b) Consider $A \equiv \int_0^1 R(t)dt$ as a measure of the quality of the disease test (values close to 1 indicating an excellent test). Show that A can be expressed in terms of the Mann-Whitney-Wilcoxon functional $\int F dG$.

(c) Suppose that $X_1, \ldots, X_m$ are i.i.d. F and $Y_1, \ldots, Y_n$ are i.i.d. G, and consider estimation of the ROC curve $R \equiv R_{F,G}$ on the basis of the data.

(i) Propose a nonparametric estimator $\mathbb{R}_{m,n}(t)$ of $R(t) = R_{F,G}(t)$.

(ii) Give conditions on F and G which imply that your estimator in (i) is consistent for a fixed $t \in (0, 1)$.

(iii) Compute the Gateaux derivatives of the functionals $T_1(F) \equiv R_{F,G}(t)$ for fixed G and t and $T_2(G) \equiv R_{F,G}(t)$ for fixed F and t.

(iv) Give conditions on F and G which imply that your estimator in (i) is asymptotically normal (for a fixed $t \in (0, 1)$). Find the influence function of your estimator (with help from (iii)).

(v) What can you say about your estimator $\mathbb{R}_{m,n}$ as an estimator of the function R uniformly in $0 \leq t \leq 1$?

(vi) Explain how and why (or why not) you could use the jackknife or bootstrap to estimate the variance of the estimator in (i).