Statistics 583, Final Exam Solutions
Wellner; 6/8/98

1. (32 points) **Define** *any four* of the following terms. In each case, provide an appropriate context for your definition.
(a) An *unbiased test* ϕ.
(b) A *uniformly most powerful unbiased (UMPU) test* ϕ.
(c) A *Similar On the Boundary* (SOB) test ϕ.
(d) A maximal invariant with respect to a group G.
(e) An *invariant test* ϕ with respect to a group G.
(f) A *continuous functional* $T(F)$ with respect to the Kolmogorov metric d_K on distribution functions.
(g) A *Fréchet differentiable functional* $T(F)$ with respect to a metric d.

   Solution: See notes, chapters 6-9.

2. (32 points) **State** *any four* of the following results:
(a) Varadarajan's theorem concerning weak convergence of the empirical measure $\mathbb{P}_n$.
(b) An example of a functional $T(F)$ which is *not weakly continuous*.
(c) The Wald-Wolfowitz-Noether-Hajek finite sampling central limit theorem.
(d) Hoeffding's formula for the distribution of ranks (under the alternative).
(e) A Central Limit Theorem for a functional $T(F)$ which is Fréchet - differentiable with respect to a metric which is compatible with the empirical distribution function (or empirical measure).
(f) Any large sample theorem for Efron's nonparametric bootstrap.

   Solution: See notes, chapters 6-9.

3. (48 points) Suppose that $X \sim \text{Binomial}(m, p_1)$ and $Y \sim \text{Binomial}(n, p_2)$ are independent. Show how to find the UMP unbiased test of size $\alpha \in (0,1)$ of $H : p_1 \geq p_2$ versus $K : p_1 < p_2$. Describe how you would carry out the test, give a name for a key parameter in the relevant conditional distribution, name the conditional distribution obtained under the null hypothesis, and describe this distribution in terms of an urn model.

   Solution: First write

$$\begin{aligned} P_{p_1,p_2}(X = x,\ Y = y) &= \binom{m}{x}\binom{n}{y} q_1^m q_2^n \exp(x \log(\frac{p_1}{q_1}) + y \log(\frac{p_2}{q_2})) \\ &= \binom{m}{x}\binom{n}{y} q_1^m q_2^n \exp\left\{ y \left(\log(\frac{p_2}{q_2}) - \log(\frac{p_1}{q_1}) \right) + (x+y) \log(\frac{p_1}{q_1}) \right\} \\ &= c(p_1, p_2) h(x,y) \exp\left(\theta U(x,y) + \xi T(x,y)\right) \end{aligned}$$

where

$$\theta \equiv \log\left(\frac{p_2/q_2}{p_1/q_1}\right) \equiv \log(\text{odds ratio}) \equiv \log(\rho) \qquad \xi \equiv \log(\frac{p_1}{q_1})\,,$$

$$U(x,y) \equiv y\,, \qquad T(x,y) \equiv x+y\,.$$

Then the conditional distribution of $U(X,Y) = Y$ given $T(X,Y) = X+Y$ does not depend on ξ, and is given by

$$P_\rho(Y = y \Big| X+Y = t) = C_t(\rho)\binom{m}{t-y}\binom{n}{y}\rho^y, \qquad y = 0,\ldots,t\,,$$

where

$$C_t(\rho)^{-1} = \sum_{y'=0}^{t}\binom{m}{t-y'}\binom{n}{y'}\rho^{y'}$$

which is a one-dimensional exponential family with monotone likelihood ratio. Moreover, the hypotheses H and K are equivalent to testing $H : \theta \le 0$ versus $K : \theta > 0$ (or to $H : \rho \le 1$ versus $K : \rho > 1$). When $\rho = 1$, the family P_ρ becomes the well-known Hypergeometric(n, N, t) family. Hence the UMP unbiased test of H versus K is of the form

$$\phi(X,Y) = \left\{\begin{array}{ccc} 1 & \text{if} & U(X,Y) > c_\alpha(T) \\ \gamma(T) & \text{if} & U(X,Y) = c_\alpha(T) \\ 0 & \text{if} & U(X,Y) < c_\alpha(T) \end{array}\right\} = \left\{\begin{array}{ccc} 1 & \text{if} & Y > c_\alpha(T) \\ \gamma(T) & \text{if} & Y = c_\alpha(T) \\ 0 & \text{if} & Y < c_\alpha(T) \end{array}\right.$$

where $c_\alpha(t)$ is chosen so that

$$P_1(Y > c_\alpha(t)\Big|X+Y=t)+P_1(Y = c_\alpha(t)\Big|X+Y=t) = \sum_{y=c_\alpha(t)+1}^{t}\frac{\binom{m}{t-y}\binom{n}{y}}{\binom{m+n}{t}}+\frac{\binom{m}{t-c_\alpha(t)}\binom{n}{c_\alpha(t)}}{\binom{m+n}{t}}\,.$$

The hypergeometric distribution which arises here is the distribution of the number of red balls in drawing n balls without replacement from an urn containing $N = m+n$ balls, n of which are red balls and m of which are black balls.

4. (50 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 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 2, problem set #4.)
(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 $P_G(f(Y) \leq y) = P_G(Y \leq f^{-1}(y)) = G(f^{-1}(y)) = \tilde{G}(y)$ 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 alternative under present consideration,

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

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

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

as large as possible. Hence it remains only to calculate

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

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

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

Hence

$$\frac{\partial}{\partial\theta}\psi'_\theta(u)|_{\theta=0}=2u-1,$$

Since $EU_{(i)}=i/(N+1)$, the locally most powerful rank test of H versus this alternative K is the Wilcoxon test "reject H if $S_N=\sum_1^n Q_j>k_\alpha$"; i.e. the locally most powerful rank test is the Wilcoxon rank sum test.

5. (50 points) Suppose that an urn contains N balls with the numbers $z_N(1),\ldots,z_N(N)$ 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 $\sum_{j=}^n z_N(R_j)$?
(d) Under some condition on the numbers $z_N(i)$, a CLT holds for an appropriately standardized version of $\overline{Y}_n$. State this condition and the theorem.
(e) Does the condition of the theorem you stated in (c) hold if the numbers $z_N(1),\ldots,z_N(N)$ are in fact $(X_1,\ldots,X_m,Y_1,\ldots,Y_n)$ where $N=m+n$ and $X_1,\ldots,X_m$ are i.i.d. F with $E_FX_1^2<\infty$ and $Y_1,\ldots,Y_n$ are i.i.d. G with $E_GY_1^2<\infty$?
(f) Briefly describe the relevance of the CLT in (c) for a permutation test of the difference in means of two populations. Briefly describe the relevance of the CLT in (c) for two-sample linear rank statistics.

Solution: (a)

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

(b)

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

where $\sigma_z^2 = N^{-1}\sum_{j=1}^{N}(z_N(j) - \overline{z}_N)^2$.

(c) $n\overline{Y}_n$ and $\sum_{j=1}^{n} z_N(R_j)$ 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 $Y_i = z_N(R_i)$ for $i = 1, \ldots, n$.

(d) If $0 < \underline{\lim}(n/N) \leq \overline{\lim}(n/N) < 1$, then the Noether condition

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

holds if and only if

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

where σ_N^2 is the variance of $\overline{Y}_n$ as defined in (b).

(e) Yes; in this case

$$\begin{aligned} \eta_N &\leq 2 \left\{ \frac{(m/N)m^{-1}\max_{1\leq i\leq m} X_i^2 + (n/N)n^{-1}\max_{1\leq j\leq n} Y_j^2}{(m/N)S_X^2 + (n/N)S_Y^2} + \frac{N^{-1}\{\overline{X}_m^2 + \overline{Y}_n^2\}}{(m/N)S_X^2 + (n/N)S_Y^2} \right\} \\ &\to_{a.s.} 0 \end{aligned}$$

since $E_F X^2 < \infty$ and $E_G Y^2 < \infty$ imply that

$$\frac{1}{m}\max_{1\leq i\leq m}|X_i|^2 \to_{a.s.} 0\,, \qquad \text{and} \qquad \frac{1}{n}\max_{1\leq i\leq n}|Y_i|^2 \to_{a.s.} 0\,,$$

while $\overline{X}_m \to_{a.s.} E_F X_1$, $\overline{Y}_n \to_{a.s.} E_G Y_1$, $S_X^2 \to_{a.s.} Var_F(X)$, and $S_Y^2 \to_{a.s.} Var_G(Y)$.

Because the convergence of the permutation distribution distribution of the two-sample $t-$statistic can be reduced to the WWNH finite-sampling CLT in (d),

it follows that the (conditionally determined, random) upper α critical points of the permutation $t-$ test for the difference of two population means converge a.s. (and in probability) to z_α, the upper α critical point of the standard normal distribution when $E_F X^2 < \infty$ and $E_G Y^2 < \infty$. This implies that the test is asymptotically equivalent to the usual two-sample $t-$test in this case.

6. (48 points) Consider the functional $T(F) = \int\int |x-y| dF(x) dF(y)$ as a measure of spread or dispersion of the distribution function F. (This functional is sometimes called "Gini's mean difference".)
(a) If $X_1, \ldots, X_n$ are i.i.d. random variables with distribution function F, what is the "principle of substitution" estimator of $T(F)$?
(b) Is the estimator you found in (a) an unbiased estimator of $T(F)$? (Calculate the bias explicitly.)
(c) Use the jackknife to suggest an estimator of $T(F)$ with less bias. Can you find an unbiased estimator of $T(F)$?
(d) Calculate the Gateaux derivative of $T(F)$, and use this to find a formula for the asymptotic variance of $\sqrt{n}(T(\mathbb{F}_n) - T(F))$.
(e) Describe how you would use the bootstrap to estimate $nVar_F(T(\mathbb{F}_n))$ and

$$H_n(x, F) \equiv P_F(\sqrt{n}(T(\mathbb{F}_n) - T(F)) \le x)\,,$$

distinguishing clearly in your description between the "ideal bootstrap" and the Monte-carlo implementation thereof.

Solution: (a) The principle of substitution estimator is just

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

(b) The principle of substitution estimator is biased: because the diagonal terms (for which $j = i$) 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 $T_n \equiv T(\mathbb{F}_n)$ is

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

(c) Here we compute

$$T_{n,i} \equiv T(\mathbb{F}_{n-1,i}) = \frac{1}{(n-1)^2} \sum_{j=1, j\neq i}^n \sum_{j'=1, j'\neq i}^n |X_j - X_{j'}|\,.$$

Hence the pseudo-values are

$$\begin{aligned} T^*_{n,i} &= nT_n - (n-1)T_{n,i} = \left\{\frac{1}{n} - \frac{1}{n-1}\right\} \sum_{j=1, j\neq i}^{n} \sum_{j'=1, j'\neq i}^{n} |X_j - X_{j'}| + \frac{2}{n}\sum_{j=1}^{n} |X_i - X_j| \\ &= -\frac{1}{n(n-1)} \sum_{j=1, j\neq i}^{n} \sum_{j'=1, j'\neq i}^{n} |X_j - X_{j'}| + \frac{2}{n}\sum_{j=1}^{n} |X_i - X_j| \,. \end{aligned}$$

Thus we find that

$$\begin{aligned} \overline{T^*_n} &\equiv \frac{1}{n}\sum_{i=1}^{n} T^*_{n,i} \\ &= 2T_n - \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \sum_{j'=1, j'\neq i}^{n} |X_j - X_{j'}| \\ &= 2T_n - \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{j'=1}^{n} |X_j - X_{j'}| 1_{[i\neq j]} 1_{[i\neq j']} \\ &= 2T_n - \frac{1}{n^2(n-1)} \sum_{j=1}^{n} \sum_{j'=1}^{n} \sum_{i=1}^{n} |X_j - X_{j'}| 1_{[i\neq j]} 1_{[i\neq j']} \\ &= 2T_n - \frac{n-2}{n^2(n-1)} \sum_{j=1}^{n} \sum_{j'=1}^{n} |X_j - X_{j'}| \\ &= 2T_n - \frac{n-2}{n-1} T_n = \frac{n}{n-1} T_n \\ &= \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{j'=1}^{n} |X_j - X_{j'}| \,. \end{aligned}$$

This estimator of $T(F)$ is unbiased: $E_F(\overline{T^*_n}) = T(F)$. In fact $\overline{T^*_n}$ is the usual $U-$statistic form corresponding to the $V-$ statistic $T(\mathbb{F}_n)$.

(d) To calculate the Gateaux derivative, we write $F_\epsilon = (1-\epsilon)F + \epsilon G$ and then compute

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

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

$$\dot{T}(F; G-F) = \frac{d}{d\epsilon} T(F_\epsilon)\Big|_{\epsilon=0}$$

$$
\begin{aligned}
&= \int\int |x-y| d(G-F)(x) dF(y) + \int\int |x-y| dF(x) d(G-F)(y) \\
&= 2\int\int |x-y| d(G-F)(x) dF(y)\,.
\end{aligned}
$$

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

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

This leads to the conclusion that the asymptotic variance of $\sqrt{n}(T(\mathbb{F}_n) - T(F))$ will be

$$
\begin{aligned}
E_F \psi_F^2(X_1) &= 4\int \left(\int |x-y| dF(y) - T(F)\right)^2 dF(x) \\
&= 4\left\{\int\int |x-y| dF(y) \int |x-y'| dF(y') dF(x) - T^2(F)\right\} \\
&= 4\left\{\int\int\int |x-y||x-y'| dF(x) dF(y) dF(y') - T^2(F)\right\}\,.
\end{aligned}
$$

(e) The ideal bootstrap estimator of $nVar_F(T(\mathbb{F}_n))$ is $nVar_{\mathbb{F}_n}(T(\mathbb{F}_n^*))$ where $\mathbb{F}_n^*$ is the empirical distribution function of $X_1^*, \ldots, X_n^*$ i.i.d. with distribution function $\mathbb{F}_n$. Similarly, the ideal bootstrap estimator of

$$
H_n(x,F) \equiv P_F(\sqrt{n}(T(\mathbb{F}_n) - T(F)) \le x)\,,
$$

is simply

$$
H_n(x,\mathbb{F}_n) \equiv P_{\mathbb{F}_n}(\sqrt{n}(T(\mathbb{F}_n^*) - T(\mathbb{F}_n)) \le x)\,.
$$

To implement Monte-Carlo approximations of these, we would draw B bootstrap samples of size n from $\mathbb{F}_n$,

$$
X_{j,1}^*, \ldots, X_{j,n}^*\,, \qquad j = 1, \ldots, B,
$$

form the corresponding empirical d.f. $\mathbb{F}_{j,n}^*$ for each $j = 1, \ldots, B$, and calculate the resulting $T_{n,j}^* \equiv T(\mathbb{F}_{j,n}^*)$. Then

$$
\widehat{nVar_{\mathbb{F}_n}(T(\mathbb{F}_n^*))} = n\frac{1}{B}\sum_{j=1}^{B}(T_{n,j}^* - \overline{T}_n^*)^2
$$

while

$$
\widehat{H_n(x,\mathbb{F}_n)} = \frac{1}{B}\sum_{j=1}^{B} 1_{[\sqrt{n}(T_{n,j}^* - T_n) \le x]}\,.
$$