Statistics 583, Problem Set 1

Wellner; 4/1/2015

Reading: Chapter 6, sections 6.3, 6.4, and 6.5;
Lehmann and Romano, TSH, Chapters 6 and 7.
See also Ferguson, MS, Chapter 5, sections 5.6 and 5.7.
Due: Wednesday, April 8, 2015

1. Let $S_N = \sum_{i=1}^n Q_i$ be the Wilcoxon rank sum statistic derived in class on 30 March.
   (a) Assuming that the null hypothesis $H : F = G$ holds, compute $E(S_N)$ and $Var(S_N)$.
   (b) Show that $(S_N - E(S_N))/\sqrt{Var(S_N)} \to_d N(0,1)$ by applying the Wald-Wolfowitz-Noether-Hájek finite sampling central limit theorem. Carefully specify any hypotheses you need to apply the theorem.
   (c) Now let $U_{m,n} \equiv \int \mathbb{F}_m d\mathbb{G}_n$ denote the Mann-Whitney form of the Wilcoxon statistic. Use the calculations in (a) and (b) and our derivations in class to compute $E(U_{m,n})$ and $Var(U_{m,n})$ under $H : F = G$.

2. (a) What is the locally best rank test of $F = G$ against $G = (e^{\theta F} - 1)/(e^\theta - 1)$, $\theta > 0$?
   (b) What is the locally best rank test of $F = G$ against $G = F/(e^\theta(1-F)+F)$?
   (c) What can you say about the power of these tests (other than the fact that they are locally most powerful)?

3. Suppose that an urn contains N balls with the numbers $z_i = -1 - \log(1 - i/(N+1))$, $i = 1, \ldots, N$ and we sample $n < N$ balls from this urn. Let $\overline{Y}_n = n^{-1}\sum_1^n Y_i$ denote the sample mean of the sampled balls.
   (a) Calculate the mean $\mu_N = E(\overline{Y}_n)$ and variance $\sigma_N^2 = Var(\overline{Y}_n)$ of $\overline{Y}_n$ in terms of $\overline{z}_N \equiv N^{-1}\sum_1^N z_i$ and $\sigma_z^2 \equiv N^{-1}\sum_1^N (z_i - \overline{z}_N)^2$. Find the limits of $\overline{z}_N$ and σ_z^2 as $N \to \infty$.
   (b) Use the Wald-Wolfowitz-Noether-Hájek finite-sampling CLT to prove that $(\overline{Y}_n - \mu_N)/\sigma_N \to_d N(0,1)$.
   (c) What classical two-sample rank statistic is $\overline{Y}_n$ equivalent to under the null hypothesis (of all $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ equal in distribution with a common continuous distribution function F, noting the two different uses of the notation "$Y_1, \ldots, Y_n$")?

4. Suppose that $X_1, \ldots, X_n$ are independent Exponential(1) random variables. Let $Y_i \equiv X_{(i)}$, for $i = 1, \ldots, n$, denote the *order statistics* corresponding to $X_1, \ldots, X_n$.

(a) Show that the vector $(Y_1, \ldots, Y_n)$ has the same joint distribution as $(W_1, \ldots, W_n)$ where $W_i \equiv \sum_{j=1}^{i} Z_j/(n-j+1)$ and $Z_1, \ldots, Z_n$ are i.i.d. Exponential(1).
(b) Use the result of (a) to compute $E(Y_i)$, $Var(Y_i)$, and $Cov(Y_i, Y_j)$ for any fixed i, j.

5. Suppose that, in Example 6.3.17, page 30, $1-F_i = (1-F)^{\Delta_i}$ where $\Delta_i = \exp(\theta z_i)$ and $z_1, \ldots, z_N$ are given real numbers and $\theta \in \mathbb{R}$. Then the distribution of the ranks of $X_1, \ldots, X_N$ (independent with respective d.f.'s $F_1, \ldots, F_N$) is

$$P_\theta(\underline{R} = \underline{r}) = \prod_{i=1}^{N} \frac{e^{\theta z_{d_i}}}{\sum_{j=i}^{N} e^{\theta z_{d_j}}}.$$

(a) Find the locally most powerful rank test of $H : \theta = 0$ versus $K : \theta > 0$. (Call the statistic S_N and express it explicitly in terms of some scores $a_N(j)$, $j = 1, \ldots, N$, the ranks $\underline{R}$, and the z_j's.)
(b) Compute $E(S_N)$ and $Var(S_N)$ under the null hypothesis $\theta = 0$? How would you carry out the test you found in (a)?
(c) Show that when $z_1 = \cdots = z_m = 0$ and $z_{m+1} = \cdots = z_N = 1$, the test reduces to "reject when $S_N = \sum_{j=1}^{n} a_N(Q_j) > c_{N,\alpha}$" with

$$a_N(i) = 1 - \sum_{j=1}^{i} \frac{1}{N-j+1};$$

this is a close relative for the test we found in Example 3.20, but the current $\theta > 0$ corresponds to $\theta' < 1$ in Example 3.20, so the alternative hypothesis now corresponds to testing $G <_s F$ in the two-sample context.
(d) Let $S_{N,1}(x) \equiv N^{-1} \sum_{i=1}^{N} z_i 1_{[X_i \geq x]}$ and $S_{N,0}(x) \equiv N^{-1} \sum_{i=1}^{N} 1_{[X_i \geq x]}$. Show that the statistic S_N can be rewritten as

$$S_N = N\left(\overline{z} - \int \frac{S_{N,1}(x)}{S_{N,0}(x)} d\mathbb{F}_N(x)\right) = N \int \left(z - \frac{S_{N,1}(x)}{S_{N,0}(x)}\right) d\mathbb{P}_N(x,z)$$

where $\mathbb{F}_N(x) \equiv N^{-1} \sum_{i=1}^{N} 1_{(-\infty,x]}(X_i)$, $\mathbb{P}_N \equiv N^{-1} \sum_{i=1}^{N} \delta_{(X_i,z_i)}$.

6. **Optional bonus problem 1:** In the context of the two sample problem of testing $H : F = G$ versus $K : F <_s G$, consider an exponential family of distributions

$$f(x;\theta) = c(\theta) \exp(\theta x) h(x)$$

and consider the simple null hypothesis $H_0 : \ f(x) = g(x) = f(x;\theta_0)$ versus the simple alternative $H_1 : \ f(x) = f(x;\theta_0)$, $g(x) = f(x;\theta_1)$ with $\theta_0 < \theta_1$. Use the Neyman Pearson lemma to find the best test of H_0 versus H_1 based on the ranks.

7. **Optional bonus problem 2:** Let $X_1, X_2, \ldots, X_N$ be a sample from a distribution with density $f_\theta(x) = \theta \exp(\theta x) 1\{x < 0\}$, $\theta > 0$, and let $V_{(1)} < V_{(2)} < \cdots < V_{(N)}$ denote the order statistics. Show that $Y_1 = V_{(1)} - V_{(2)}, Y_2 = V_{(2)} - V_{(3)}, \cdots, Y_{N-1} = V_{(N-1)} - V_{(N)}, Y_N = V_{(N)}$ are independent random variables and that Y_j has density $j\theta e^{j\theta x} 1\{x < 0\}$ for $j = 1, \ldots, N$. Use this fact to determine the rejection region of the test you found in problem 6 explicitly when the exponential family $f(x;\theta) = \theta \exp(\theta x) 1\{x < 0\}$; i.e. $c(\theta) = \theta$, $h(x) = 1\{x < 0\}$ in problem 6. Show that the resulting test is a most powerful rank test of $H:\ F = G$ versus $K:\ G = F^2$.

8. **Optional bonus problem 3:** (Problem 10, page 249, Ferguson, MS) Let $\Theta = \{(\Delta, \pi_1, \ldots, \pi_n):\ \Delta \geq 0,\ \pi = (\pi_1, \ldots, \pi_n)$ is a permutation of $\{1, \ldots, n\}\}$, and let the distribution of $X_1, \ldots, X_n$ given $\theta = (\Delta, \pi_1, \ldots, \pi_n)$ be as independent random variables with gamma distributions, $X_i \sim \text{Gamma}(\alpha, \beta^{-1} \exp(-\Delta b_{\pi_i}))$ where $\alpha > 0$, $\beta > 0$, and $b_1, \ldots, b_n$ are known real numbers with $\sum_1^n b_i > 0$. Consider testing the hypothesis $H : \Delta = 0$ versus the alternative $K : \Delta > 0$. (This is a Gamma-regression model with covariates or predictors b_i in which the relationship between the responses X_i and the covariates b_i have become scrambled or mixed up: we unfortunately don't know the right pairing of X_i and b_i, but we do know that some permutation of the b_i's is correct. Note that problem 11 in Ferguson, MS, gives a more realistic version of the problem in which β is also unknown. This is a version of a "broken sample" or "record linkage" problem; see e.g. Bai and Hsing, PTRF, 2005.)
(a) Show that this problem is invariant under the group of permutations of $(X_1, \ldots, X_n)$, and that the distribution of the maximal invariant $(Y_1, \ldots, Y_n) \equiv (X_{(1)}, \ldots, X_{(n)})$(the order statistics) has density

$$f_{\underline{Y}}(\underline{y}|\Delta) = \frac{(\prod_1^n y_i)^{\alpha-1} \exp(-\alpha\Delta \sum_1^n b_i)}{\Gamma(\alpha)^n \beta^{n\alpha}} \sum_{\pi \in \Pi} \exp\left\{ -\frac{1}{\beta} \sum_{i=1}^n y_i \exp(-\Delta b_{\pi_i}) \right\}$$

for $y_1 < \cdots < y_n$ and zero elsewhere where $\sum_{\pi\in\Pi}$ denotes the sum over all permutations π of $\{1, \ldots, n\}$.
(b) Show that the locally best invariant test of H versus K (i.e. the test which maximizes the slope of the power function at the null hypothesis) is to reject H when $\sum_{i=1}^n X_i$ is too large.

9. **Optional bonus problem 4:** Suppose that we observe i.i.d. pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ with distribution function $H(x, y) = P(X_1 \leq x, Y_1 \leq y)$ on $\mathbb{R}^2$. Consider testing the null hypothesis $H_0:\ H(x, y) = F(x)G(y)$ for all x, y where $F(x) = H(x, \infty)$ and $G(y) = H(\infty, y)$ are the marginal distributions of H (i.e. X_1 and Y_1 are independent) versus the alternative hypothesis $H_1:\ X_1$ and Y_1 are not independent. Consider the group of transformations $\mathcal{G} = \{g : (\mathbb{R}^2)^n \to (\mathbb{R}^2)^n\} = \mathcal{G}_0^n$

where

$$\mathcal{G}_0 = \left\{ g_0 : \mathbb{R}^2 \to \mathbb{R}^2 \middle| g_0(x,y) = (f_1(x), f_2(y)),\ \ f_1, f_2 \in \mathcal{F} \right\}$$

where $\mathcal{F}$ is the class of all continuous and strictly increasing transformations from $\mathbb{R}$ to $\mathbb{R}$.

(a) Show that the problem of testing H_0 versus H_1 is invariant under $\mathcal{G}$.

(b) Let Θ denote the collection of all distribution functions H on $\mathbb{R}^2$ with continuous marginal distributions F and G. Find the induced group $\overline{\mathcal{G}}$ on the parameter space Θ.

(c) Find the maximal invariant $\nu(\theta) = \nu(H)$ for $\theta = H \in \Theta$.

(d) Consider alternatives H_Δ defined as follows: for (X_0, Y_0) with distribution function $F \cdot G$ for some continuous and strictly increasing univariate distributions F and G, $\Delta \in \mathbb{R}$, and Z independent of (X_0, Y_0), let $(X, Y) \equiv (X_0, Y_0) + \Delta(Z, Z)$. Then H_Δ is the distribution function of (X, Y).

(e) If F and G have densities f and g with respect to Lebesgue measure and Z has distribution function M, find an expression for the density h_Δ of (X, Y) in terms of f, g, M, and Δ. [Hint: see Hájek and Šidák (1967), pages 75-77.]