Statistics 583, Problem Set 1 Solutions

Wellner; 4/8/2015

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 $H : F = G$.

Solution: (a) Recall that we may write $S_N = \sum_{j=1}^n R_{m+j}$ and that for an urn with balls labelled as $a_N(1), \ldots, a_N(N)$ we have

$$(a_N(R_1), \ldots, a_N(R_n)) \stackrel{d}{=} (Y_1, \ldots, Y_n)$$

where $Y_1, \ldots, Y_n$ is a sample without replacement from the urn. In the present case $(a_N(1), \ldots, a_N(N)) = (1, \ldots, N)$, and it follows that

$$E(S_N) = E(\sum_1^n Y_i) = \sum_1^n E(Y_i) = \sum_1^n E(a_N(R_i)) = n\overline{a}_N = n\frac{N+1}{2}.$$

Similarly

$$\sigma_N^2 = Var(S_N) = Var(\sum_1^n Y_i) = n\sigma_a^2 \left(1 - \frac{n-1}{N-1}\right) = n\frac{(N^2-1)}{12}\left(1 - \frac{n-1}{N-1}\right).$$

To see that the 3rd equality in the last display holds, note that by symmetry

$$\begin{aligned} Var(\sum_1^n Y_i) &= \sum_{i=1}^n Var(Y_i) + \sum_{i \neq j} Cov(Y_i, Y_j) \\ &= n\sigma_a^2 + n(n-1)Cov(Y_1, Y_2). \end{aligned} \tag{1}$$

When $n = N$ this yields

$$0 = Var(\sum_1^N Y_i) = N\sigma_a^2 + N(N-1)Cov(Y_1, Y_2),$$

and hence

$$Cov(Y_1, Y_2) = -\frac{N\sigma_a^2}{N(N-1)} = -\frac{\sigma_a^2}{N-1}.$$

Substituting this in (1) yields

$$Var(\sum_1^n Y_i) = n\sigma_a^2 - \frac{n(n-1)}{N(N-1)}\sigma_a^2 = n\sigma_a^2\left(1 - \frac{n-1}{N-1}\right).$$

(b) We need to verify the Noether condition. But here

$$\frac{\max_{i\le N}(a_N(i) - \overline{a}_N)^2}{\sum_1^N (a_N(i) - \overline{a}_N)^2} = \frac{N^{-1}((N-1)/2)^2}{(N^2-1)/12} \to 0$$

as $N \to \infty$. Thus if $0 < \liminf(n/N) \le \limsup(n/N) < 1$ the Wald-Wolfowitz-Noether-Hájek CLT implies that

$$\frac{S_N - n(N+1)/2}{\sigma_N} \to_d N(0,1).$$

(c) Now $U_{m,n} = (mn)^{-1}\sum_{i=1}^m \sum_{j=1}^n 1\{X_i \le Y_j\}$, so it follows immediately by symmetry that $E(U_{m,n}) = 1/2$. Another derivation follows from the identity $S_N = mnU_{m,n} + n(n+1)/2$:

$$E(U_{m,n}) = \frac{1}{mn}E\left(S_N - n(n+1)/2\right) = \frac{1}{mn}\left(\frac{n(N+1)}{2} - \frac{n(n+1)}{2}\right) = \frac{1}{m}\frac{m}{2} = \frac{1}{2}$$

as before. Similarly,

$$\begin{aligned} Var(U_{m,n}) &= \frac{1}{m^2n^2}Var(S_N) = \frac{1}{m^2n^2}\frac{n(N^2-1)}{12}\frac{m}{N-1} \\ &= \frac{1}{mn}\cdot\frac{N+1}{12}. \end{aligned}$$

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)?

Solution: By Hoeffding's formula

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

where

$$\psi_\theta(u) = G_\theta \circ F^{-1}(u) = \frac{e^{\theta u} - 1}{e^\theta - 1}$$

for the first alternatives, and

$$\psi_\theta(u) = G_\theta \circ F^{-1}(u) = \frac{u}{e^\theta(1-u) + u}$$

in the case of the second type of alternative. In either case the locally most powerful rank test rejects for those values $\underline{q}$ of $\underline{Q}$ which make

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

as large as possible. Hence it remains only to calculate

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

and $E_{uniform}\phi(U_{(i)})$ for the two alternatives in question.
(i) In the first case,

$$\psi'_\theta(u) = \frac{\theta e^{\theta u}}{e^\theta - 1}\,,$$

and straightforward calculation yields

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

By applying L'Hopital's rule twice, we find that

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

Since $E(U_{(i)}) = i/(N+1)$, the locally most powerful rank test of H versus this alternative K is the Wilcoxon test "reject H if $S_N = \sum_{j=1}^{n} Q_j > k_\alpha$".
(ii) In the second case,

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

Hence

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

and again the locally most powerful rank test is the Wilcoxon rank sum test.
As for interpretations of these alternatives, first note that the functions $\psi_\theta(u)$ are distribution functions on $[0,1]$ with densities $\psi'_\theta(u)$. (i) This alternative is the simplest exponential family density related to the uniform$(0,1)$ distribution: the density is of the form $p_\theta(u) = \psi'_\theta(u) = c(\theta)\exp(\theta u)1_{[0,1]}(u)$.
(ii) For this family, note that

$$1 - \psi_\theta(u) = \frac{e^\theta(1-u)}{e^\theta(1-u)+u}\,,$$

and hence the *odds ratio* is

$$\frac{1-\psi_\theta(u)}{\psi_\theta(u)} = e^\theta \frac{1-u}{u} = e^\theta \cdot \text{the odds ratio for Uniform(0,1)}\,.$$

Thus this family is one with proportional odds ratios.

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 "$Y_1,\ldots,Y_n$")?

Solution: (a) The mean is

$$\begin{aligned}\mu_N &= E(\overline{Y}_n) = \overline{z}_N = \frac{1}{N}\sum_{i=1}^N \left\{-\log\left(1 - \frac{i}{N+1}\right)\right\} \\ &\to \int_0^1 \{-\log(1-t)\}\,dt = 1\end{aligned}$$

upon noticing that $F^{-1}(t) = -\log(1-t)$ for the standard exponential distribution $F(x) = 1 - e^{-x}$, $x \geq 0$, so that $F^{-1}(U) =_d Y \sim$ Exponential(1). Similarly, the variance is

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

where

$$\begin{aligned}\sigma_z^2 &= \frac{1}{N}\sum_{i=1}^{N}(z_i-\overline{z}_N)^2 \to \int_0^1 \{-\log(1-t)-1\}^2 dt \\ &= Var(Y) = 1\,.\end{aligned}$$

(b) The Wald-Wolfowitz-Noether-Hájek finite-sampling CLT yields $(\overline{Y}_n - \mu_N)/\sigma_N \to_d N(0,1)$ as long as $0 < \liminf(n/N) \le \limsup(n/N) < 1$ if we show that the Noether condition holds. But the Noether condition is

$$\eta_N \equiv \frac{\max_{1\le i\le N}|z_i-\overline{z}_N|}{\sum_1^N (z_i-\overline{z}_N)^2} \to 0\,.$$

Upon dividing the numerator and denominator by N, we know from part A that the denominator (divided by N) converges to 1. Hence it suffices to show that

$$N^{-1}\max_{1\le i\le N}|z_i-\overline{z}_N|^2 \to 0\,.$$

Now since z_i increases with i,

$$\begin{aligned}\max_{1\le i\le N}|z_i-\overline{z}_N| &\le \max_{1\le i\le N}(\overline{z}_N - z_i) \vee \max_{1\le i\le N}(z_i-\overline{z}_N) \\ &\le \overline{z}\vee(z_N-\overline{z}_N)\end{aligned}$$

where $z_N = -\log(1-N/(N+1)) = -\log(1/(N+1)) = \log(N+1)$. Thus we have

$$\begin{aligned}&N^{-1}\max_{1\le i\le N}|z_i-\overline{z}_N|^2 \\ &\quad\le N^{-1}\overline{z}_N^2 \vee N^{-1}(\log(N+1)-\overline{z}_N)^2 \\ &\quad\to 0\vee 0 = 0\,.\end{aligned}$$

(c) Under the null hypothesis $\overline{Y}_n$ is equivalent to the "log-rank" statistic

$$T_N \equiv \frac{1}{n}\sum_{i=1}^{n}\left\{-\log\left(1-\frac{R_i}{N+1}\right)\right\}$$

where R_i is the rank of Y_i, $i=1,\ldots,n$ in the combined sample, $X_1,\ldots,X_m,Y_1,\ldots,Y_n$.

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.

Solution: (a) Note that $0 \le W_1 \le \ldots \le W_n$ and

$$Z_i = (n - i + 1)(W_i - W_{i-1}), \qquad i = 1, \ldots, n \tag{2}$$

(with $W_0 \equiv 0$). Let $g(\underline{Z}) \equiv \underline{W}$ be the map defined in (a) of the problem statement so that $g^{-1}(\underline{W}) = \underline{Z}$ is given in (2). Then the Jacobian of g^{-1} has entries $n,\ (n-1), \ldots, 1$ on the diagonal, entries $-(n-1),\ \ldots\ , -2, -1$ below the diagonal, and zero elsewhere. Hence $\det(J_{g^{-1}}) = \mathrm{tr}(J_{g^{-1}}) = n!$ and the density of $\underline{W}$ is given by

$$\begin{aligned}
f_{\underline{W}}(\underline{w}) &= f_{\underline{Z}}(g^{-1}(\underline{w}))\det(J_{g^{-1}}) \\
&= n! \prod_{i=1}^{n} \exp(-(n - i + 1)(w_i - w_{i-1})) \\
&= n! \exp(-\sum_{i=1}^{n}(n - i + 1)(w_i - w_{i-1})) \\
&= n! \exp(-\sum_{i=1}^{n}(\sum_{j=i}^{n} 1)(w_i - w_{i-1})) \\
&= n! \exp(-\sum_{j=1}^{n}\sum_{i \le j}(w_i - w_{i-1})) \\
&= n! \exp(-\sum_{j=1}^{n} w_j) = n! f(w_1) \cdots f(w_n)
\end{aligned}$$

on the set $0 \le w_1 \le \ldots \le w_n < \infty$ where $f(x) = \exp(-x)1_{[0,\infty)}(x)$ is the standard exponential density. Hence $\underline{Y} =_d \underline{Y} \equiv \underline{X}_{(\cdot)}$ where $X_1, \ldots, X_n$ are i.i.d. exponential(1).
(b) It follows immediately from (a) that

$$E(Y_i) = E(\sum_{j=1}^{i} \frac{Z_j}{n - j + 1}) = \sum_{j=1}^{i} \frac{1}{n - j + 1}\,,$$

$$Var(Y_i) = Var(\sum_{j=1}^{i} \frac{Z_j}{n - j + 1}) = \sum_{j=1}^{i} \frac{1}{(n - j + 1)^2}\,,$$

and

$$\begin{aligned} Cov(Y_i, Y_j) &= Cov\left(\sum_{k=1}^{i} \frac{Z_k}{n-k+1}, \sum_{k'=1}^{i} \frac{Z_{k'}}{n-k'+1}\right) \\ &= \sum_{k=1}^{i \wedge j} \frac{1}{(n-k+1)^2}\,. \end{aligned}$$

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

Solution: (a) The locally most powerful rank test reject for large values of

$$\frac{\partial}{\partial \theta} P_\theta(\underline{R} = \underline{r})\big|_{\theta=0},$$

or, equivalently, for large values of

$$\begin{aligned}
\frac{\partial}{\partial\theta}\log P_\theta(\underline{R}=\underline{r})\big|_{\theta=0} &= \frac{\partial}{\partial\theta}\sum_{i=1}^{N}\left\{\theta z_{d_i}-\log\left(\sum_{j=i}^{N}\exp(\theta z_{d_j})\right)\right\}\big|_{\theta=0}\\
&= \sum_{i=1}^{N}\left\{z_{d_i}-\frac{\sum_{j=i}^{N}z_{d_j}}{N-i+1)}\right\}\\
&= N\overline{z}-\sum_{i=1}^{N}\frac{\sum_{j=i}^{N}z_{d_j}}{N-i+1}\\
&= N\overline{z}-\sum_{j=1}^{N}\sum_{i=1}^{N}1_{[j\geq i]}\frac{z_{d_j}}{N-i+1}\\
&= \sum_{j=1}^{N}a_N(j)z_{d_j}\quad\text{where}\quad a_N(j)\equiv 1-\sum_{i=1}^{j}\frac{1}{N-i+1}\\
&= \sum_{j=1}^{N}a_N(r_j)z_j
\end{aligned}$$

since d is the inverse permutation of r. Thus the locally most powerful rank test of $H:\theta=0$ versus $K:\theta>0$ is of the form "reject H if $S_N\equiv\sum_{j=1}^{N}a_N(R_j)z_j > k_N(\alpha)$" where $k_N(\alpha)$ satisfies $P_{\theta=0}(S_N>k_N(\alpha))\approx\alpha$.
(b) Now with $a_N(i)\equiv 1-b_N(i)$ for $1\leq i\leq N$ with $b_N(i)=\sum_{j=1}^{i}(N-j+1)^{-1}$,

$$\begin{aligned}
E(S_N) &= N\overline{z}-\sum_{j=1}^{N}Eb_N(R_j)z_j = N\overline{z}-\sum_{j=1}^{N}\left(\sum_{i=1}^{N}N^{-1}b_N(i)\right)z_j\\
&= N\overline{z}\left(1-\overline{b}_N\right)\\
&= 0
\end{aligned}$$

since

$$\begin{aligned}
\overline{b}_N &= N^{-1}\sum_{i=1}^{N}b_N(i) = N^{-1}\sum_{i=1}^{N}\left(\sum_{j=1}^{i}\frac{1}{N-j+1}\right)\\
&= N^{-1}\sum_{j=1}^{N}\frac{1}{N-j+1}\sum_{i=1}^{N}1\{j\leq i\}\\
&= N^{-1}\sum_{j=1}^{N}\frac{1}{N-j+1}(N-j+1) = N^{-1}\sum_{j=1}^{N}1=1.
\end{aligned}$$

To calculate $Var(S_N)$, first note that $\sum_{j=1}^N a_N(R_j) = \sum_{i=1}^N a_N(i)$, so, by symmetry

$$\begin{aligned} 0 &= Var(\sum_{i=1}^N a_N(i)) = Var(\sum_{j=1}^N a_N(R_j)) \\ &= \sum_{j=1}^N Var(a_N(R_j)) + \sum_{j,j'=1, j\neq j'}^N Cov(a_N(R_j), a_N(R_{j'})) \\ &= NVar(a_N(R_j)) + N(N-1)Cov(a_N(R_j), a_N(R_{j'})) \\ &= N\sigma_a^2 + N(N-1)Cov(a_N(R_j), a_N(R_{j'})) \end{aligned}$$

where $\sigma_a^2 \equiv N^{-1}\sum_{j=1}^N (a_N(j) - \overline{a}_N)^2$. Thus we find

$$Cov(a_N(R_j), a_N(R_{j'})) = -\frac{1}{N-1}\sigma_a^2.$$

Now we calculate

$$\begin{aligned} Var(S_N) &= Var(\sum_{j=1}^N a_N(R_j) z_j) \\ &= \sum_{j=1}^N z_j^2 Var(a_N(R_j)) + \sum_{j=1}^N \sum_{j'=1, j'\neq j}^N z_j z_{j'} Cov(a_N(R_j, R_{j'}) \\ &= \sum_{j=1}^N z_j^2 \sigma_a^2 - \frac{1}{N-1}\sum_{j=1}^N \sum_{j'=1, j'\neq j}^N z_j z_{j'} \sigma_a^2 \\ &= \sigma_a^2 \left\{ \sum_{j=1}^N z_j^2 - \frac{1}{N-1} \sum_{j'=1, j'\neq j}^N z_j z_{j'} \right\} \\ &= \sigma_a^2 \left\{ \sum_{j=1}^N z_j^2 - \frac{1}{N-1} \left(\left(\sum_{j=1}^N z_j \right)^2 - \sum_{j=1}^N z_j^2 \right) \right\} \\ &= \frac{N^2}{N-1}\sigma_a^2 \cdot \sigma_z^2 \end{aligned} \tag{3}$$

where $\sigma_z^2 \equiv \frac{1}{N}\sum_{i=1}^N (z_i - \overline{z})^2$. Theorem 4.1 of Hájek (1961), page 513, says that if

$$\max_{1\leq i\leq N} \frac{|a_N(i) - \overline{a}_N|}{N\sigma_a^2} \to 0, \quad \text{and} \quad \max_{1\leq i\leq N} \frac{|z_i - \overline{z}_N|}{N\sigma_z^2} \to 0, \tag{4}$$

then $(S_N - E(S_N))/\sqrt{Var(S_N)} \to_d N(0,1)$ if and only if

$$\sum\sum_{\{(i,j):\ \sqrt{N}|a_i - \overline{a}||z_j - \overline{z}| > \epsilon N^2 \sigma_a^2 \sigma_z^2\}} \frac{|a_i - \overline{a}|^2 |z_j - \overline{z}|^2}{N^2 \sigma_a^2 \sigma_z^2} \to 0 \tag{5}$$

for every $\epsilon > 0$. It is easily verified that $\{a_N(i) : 1 \le i \le N\}$ satisfies the first part of (4), and we will assume that the second part of (4) holds for $\{z_i : 1 \le i \le N\}$. We will also assume that (5) holds. Then it follows from our computation of the mean and variance of S_N that we can carry out our test approximately for large N by rejecting H if $S_N > z_\alpha \sqrt{Var(S_N)} = z_\alpha N \sigma_a \sigma_z / \sqrt{N-1}$ where $P(Z > z_\alpha) = \alpha$ for $Z \sim N(0,1)$.

(c) When $z_1 = \cdots = z_m = 0$ and $z_{m+1} = \cdots = z_N = 1$, the test reduces to the test "reject when $S_N = N\{n/N - \sum_{j=1}^n a_N(R_{m+j}) > k_{N,\alpha}$", or, equivalently , "reject when $\tilde{S}_N \equiv \sum_{j=1}^n a_N(Q_j) < \tilde{k}_{N,\alpha}$"; this is closely related to the Savage test of Example 3.20, but with the direction of the test reversed because $\theta > 0$ in our current setting corresponds to $\Delta > 1$, and this corresponds to the parameter θ of Example 3.20 being less than 1.

(d) To see that S_N can be rewritten as claimed, note that

$$\begin{aligned}
\sum_{i=1}^N \frac{\sum_{j=i}^N z_{d_j}}{N-i+1} &= \sum_{i=1}^N \frac{\sum_{j=1}^N z_{d_j} 1\{X_{(j)} \ge X_{(i)}\}}{\sum_{j=1}^N 1\{X_{(j)} \ge X_{(i)}\}} \\
&= N \int_{-\infty}^{\infty} \frac{N^{-1}\sum_{j=1}^N z_j 1\{X_j \ge x\}}{N^{-1}\sum_{j=1}^N 1\{X_j \ge x\}} \, d\mathbb{F}_N(x) \\
&= N \int_{-\infty}^{\infty} \frac{S_{N,1}(x)}{S_{N,0}(x)} d\mathbb{F}_N(x)
\end{aligned}$$

where $S_{N,1}$ and $S_{N,0}$ are as defined in the problem statement.

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.

Solution: Under H_0

$$P_0(\underline{Q} = \underline{q}) = 1/\binom{N}{n}, \qquad \underline{q} = (q_1, \ldots, q_n)$$

with $1 \le q_1 < q_2 < \cdots < q_n \le N$. Under H_1 it follows from Hoeffding's formula

that

$$\begin{aligned} P_1(\underline{Q}=\underline{q}) &= \frac{1}{\binom{N}{n}} E_0 \prod_{i=1}^{n} \frac{f(V_{(q_j)};\theta_1)}{f(V_{(q_j)};\theta_0)} \\ &= \frac{1}{\binom{N}{n}} E_0 \left\{ \left(\frac{c(\theta_1)}{c(\theta_0)} \right)^n \exp((\theta_1-\theta_0)\sum_{j=1}^{n} V_{(q_j)}) \right\} \\ &= \left(\frac{c(\theta_1)}{c(\theta_0)} \right)^n \frac{1}{\binom{N}{n}} E_0\{\exp((\theta_1-\theta_0)\sum_{j=1}^{n} V_{(q_j)}\} \end{aligned}$$

where $V_{(1)} < \cdots < V_{(N)}$ are order statistics of a sample $V_1, \ldots, V_N$ i.i.d. with density $f(\cdot;\theta_0)$. Thus by the Neyman -Pearson lemma, the most powerful rank test of H_0 versus H_1 is of the form

$$\phi(\underline{q}) = \begin{cases} 1, & \text{if } E\exp((\theta_1-\theta_0)\sum_{j=1}^{n} V_{(q_j)}) > k, \\ \gamma, & \text{if } E\exp((\theta_1-\theta_0)\sum_{j=1}^{n} V_{(q_j)}) = k, \\ 0, & \text{if } E\exp((\theta_1-\theta_0)\sum_{j=1}^{n} V_{(q_j)}) < k, \end{cases}$$

where $E_0\phi(\underline{Q}) = \alpha$ determines k and γ.

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 $K:\ G = F^2$.

Solution: Note that $-X_1, \ldots, -X_N$ are i.i.d. $p_\theta(x) = \theta e^{-\theta x}1\{x > 0\}$, and hence $-\theta X_1, \ldots, -\theta X_N$ are i.i.d. Exponential(1). It follows that the vector of order statistics $V_{(1)} < \cdots < V_{(N)}$ of $X_1, \ldots, X_N$ satisfy

$$0 < -V_{(N)} < \cdots < -V_{(1)} < \infty$$

and

$$\begin{aligned} \theta(-V_{(N)}, \ldots, -V_{(1)}) &\stackrel{d}{=} (W_1, \ldots, W_N) \\ &\stackrel{d}{=} \left(\frac{Z_1}{N}, \ldots, \sum_{j=1}^{i} \frac{Z_j}{N-j+1}, \ldots, \sum_{j=1}^{N} \frac{Z_j}{1} \right) \end{aligned}$$

by problem 3. Here $W_1, \ldots, W_N$ are the order statistics of standard exponential(1) random variables, and $Z_1, \ldots, Z_N$ are i.i.d. exponential(1). Thus we find that

$$\begin{aligned}
&\theta V_{(1)} - V_{(2)} \stackrel{d}{=} -W_N + W_{N-1} = -Z_N, \\
&\theta V_{(2)} - V_{(3)} \stackrel{d}{=} -W_{N-1} + W_{N-2} = -\frac{Z_{N-1}}{2}, \\
&\cdots \\
&\theta V_{(N-1} - V_{(N)} \stackrel{d}{=} -W_2 + W_1 = -\frac{Z_2}{N-1}, \\
&\theta V_{(N)} \stackrel{d}{=} -Z_1 = -\frac{Z_1}{N-1},
\end{aligned}$$

are independent. Furthermore,

$$P(\theta(V_{(j)} - V_{(j+1)}) > x) = P(-Z_{N-j+1}/j > x) = P(Z_{N-j+1} < -jx) = 1 - \exp(jx), \qquad x < 0.$$

for $j = 1, \ldots, N$, so

$$P(V_{(j)} - V_{(j+1)} > x) = P(-Z_{N-j+1}/j > \theta x) = P(Z_{N-j+1} < -j\theta x) = 1 - \exp(j\theta x), \qquad x < 0,$$

and hence $V_{(j)} - V_{(j+1)}$ has the claimed density for $j = 1, \ldots, N$.

8. **Optional bonus problem 3:** (Problem 10, page 249, Ferguson, MS) Let $\Theta = \{(\Delta, \pi_1, \ldots, \pi_n) : \Delta \geq 0, \ \pi = (\pi_1, \ldots, \pi_n) \text{ is a permutation of } \{1, \ldots, n\}\}$, 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.

Solution: Let G be the permutation group

$$G = \{g : \ g(x) = (x_{\pi(1)}, \ldots, x_{\pi(n)}), \ \pi \in \Pi\}.$$

Then, if $\underline{X} \sim P_\theta$, for $g = g_{\pi'} \in G$, $g(\underline{X}) \sim P_{\overline{g}(\theta)}$ with $\overline{g}(\theta) = (\Delta, \pi \circ \pi') = (\Delta, (\pi_{\pi'(1)}, \ldots, \pi_{\pi'(n)}))$. Thus the hypotheses are invariant under G. The order statistics are a $G-$MI. But of course the X_i's are *not* identically distributed. The joint density of the X_i's is given by

$$\begin{aligned} f(\underline{x}; \Delta, \pi) &= \prod_{i=1}^n \frac{x_i^{\alpha-1} e^{-\alpha\Delta b_{\pi(i)}}}{\Gamma(\alpha)\beta^\alpha} \exp(-\beta^{-1} e^{-\Delta b_{\pi(i)}} x_i) \\ &= \frac{\exp(-\Delta \sum_{j=1}^n b_j)(\prod_{i=1}^n x_i)^{\alpha-1}}{\Gamma(\alpha)^n \beta^{n\alpha}} \exp\left(-\beta^{-1} \sum_{i=1}^n x_i e^{-\Delta b_{\pi(i)}}\right). \end{aligned}$$

If $X_{(\cdot)} = (X_{(1)}, \ldots, X_{(n)})$ denotes the order statistics, then by problem 3(c) of problem set #9, Statistics 582,

$$\begin{aligned} & f_{X_{(\cdot)}}(\underline{x}_{(\cdot)}; \Delta, \pi) \\ &= \sum_{\pi' \in \Pi} f(\pi' \underline{x}; \Delta, \pi) \\ &= \frac{1}{\Gamma(\alpha)^n \beta^{n\alpha}} \exp(-\alpha\Delta \sum b_j) \sum_{\pi' \in \Pi} \prod_{i=1}^n x_{\pi'(i)}^{\alpha-1} \exp\left\{-\frac{1}{\beta} \sum_{i=1}^n x_{(\pi'(i))} \exp(-\Delta b_{\pi_i})\right\} \\ &= \frac{1}{\Gamma(\alpha)^n \beta^{n\alpha}} \exp(-\alpha\Delta \sum b_j) \prod_{i=1}^n x_{(i)}^{\alpha-1} \sum_{\pi^* \in \Pi} \exp\left\{-\frac{1}{\beta} \sum_{i=1}^n x_{(i)} \exp(-\Delta b_{\pi_i^*})\right\} \end{aligned}$$

with $\pi^* \equiv (\pi')^{-1} \circ \pi$. Note that this distribution depends only on the $\overline{G}-$MI Δ (and not on π).
(b) Now

$$\begin{aligned} l(\Delta | \underline{X}_{(\cdot)}) &\equiv \log f_{X_{(\cdot)}}(\underline{X}(\cdot); \Delta) \\ &= \log\left\{\sum_{\pi^* \in \Pi} \exp\left\{-\frac{1}{\beta} \sum_{i=1}^n X_{(i)} \exp(-\Delta b_{\pi^*(i)})\right\}\right\} \\ &\quad - \alpha\Delta \sum_{i=1}^n b_i + \text{constant in} \quad \Delta \end{aligned}$$

so that

$$\begin{aligned}
\dot{\imath}_\Delta(\underline{X}_{(\cdot)};\Delta=0) &= \frac{1}{\sum_{\pi^*}\exp(\cdots)}\Big|_{\Delta=0} \\
&\qquad \cdot \sum_{\pi^*\in\Pi}\exp(\cdots)\Big|_{\Delta=0}\left\{\frac{1}{\beta}\sum_{i=1}^n X_{(i)}\exp(-\Delta b_{\pi^*(i)})(-b_{\pi^*(i)})\Big|_{\Delta=0}\right\} \\
&\qquad\qquad -\alpha\sum_{j=1}^n b_j \\
&= \frac{1}{\beta n!}\sum_{\pi^*}\left\{\sum_{i=1}^n X_{(i)}b_{\pi^*(i)}\right\}-\alpha\sum_{i=1}^n b_i \\
&= \frac{1}{\beta n!}\sum_{i=1}^n X_{(i)}\sum_{\pi^*} b_{\pi^*(i)}-\alpha\sum_{i=1}^n b_i \\
&= \frac{1}{\beta}\bar{b}\sum_{i=1}^n X_{(i)}-\alpha n\bar{b} = \frac{n\bar{b}}{\beta}(\overline{X}-\alpha\beta),
\end{aligned}$$

and hence the locally MP invariant test rejects for large values of $\sum_{i=1}^n X_i$.

For an extension to unknown β and to testing $H_0:\Delta=0$ versus $K:\Delta\neq 0$, see Ferguson, problem 11, page 249. In this version of the problem $\sum_1^n b_i = 0$ *is* assumed.

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 $H(x,y)=P(X_1\le x, Y_1\le y)$ 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

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

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.]