Statistics 583, Problem Set 3 Solutions

Wellner; 4/18/2007

1. (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.

2. Suppose that an urn contains N balls with the numbers $z_i = -\log(1-i/(N+1))$, $i=1,\ldots,N$ 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 valls.
(a) Calculate the mean $\mu_N = E(\overline{Y}_n)$ and variance $\sigma_N^2 = Var(\overline{Y}_n)$ of $\overline{Y}_n$.
Find the limits of $\overline{z}_N$ and σ_z^2 as $N\to\infty$.
(b) Use the Wald-Wolfowitz-Noether-Hajek 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)?

Solution: (a) The mean is

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

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). The convergence in (1) is justified by noting that the measure μ_N corresponds to the uniform measure on the set $\{1/(N+1), \ldots, N/(N+1)\}$, and hence these measures converge weakly to the uniform measure on $(0,1)$, namely Lebesgue measure with df t on $[0,1]$. Furthermore, for each $\delta > 0$ there is a constant $M = M_\delta$ such that $-\log(1-t) \leq M(1-t)^{-\delta}$ for $0 \leq t < 1$, and hence the function $-\log(1-t)$ is uniformly integrable with respect to the sequence μ_N: choosing $\delta = 1/2$,

$$\begin{aligned} &\int_{-\log(1-t)>\lambda} (-\log(1-t))d\mu_N(t) \\ &\leq \int_{M(1-t)^{-1/2}>\lambda} M(1-t)^{-1/2}d\mu_N(t) \\ &= \frac{1}{N}\sum_{i=1}^{N} M(1-\frac{i}{N+1})^{-1/2}1\{i/(N+1) > (\lambda/M)^{-2}\} \\ &\leq \frac{N+1}{N}M\int_{t>1-(\lambda/M)^{-2}}(1-t)^{-1/2}dt = \frac{N+1}{N}M\int_0^{(\lambda/M)^{-2}} s^{-1/2}ds \\ &\leq 4\lambda^{-1}, \end{aligned}$$

so

$$\lim_{N\to\infty}\int_{-\log(1-t)>\lambda}(-\log(1-t))d\mu_N(t) \leq 4\lambda^{-1} \to 0$$

as $\lambda \to \infty$. 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

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

The convergence can again be justified by a uniform integrability argument. (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, $X_1, \ldots, X_m, Y_1, \ldots, Y_n$.

3. 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 $0 \leq W_1 \leq \ldots \leq W_n$ 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 by (**??**) 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

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

on the set $0 \leq w_1 \leq \ldots \leq w_n < \infty$ 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.

4. Let $X_1, X_2, \ldots, X_N$ be a sample from a distribution with density

$$f_\theta(x) = \theta \exp(\theta x) 1\{x < 0\}, \qquad \theta > 0,$$

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 completely independent random variables and that Y_j has density $j\theta e^{j\theta x} 1\{x < 0$ for $j = 1, \ldots, N$.

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 $j = 1, \ldots, N$.

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