

## Statistics 583, Problem Set 3 Solutions

Wellner; 4/19/2006

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

**Solution:** A. The mean is

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

upon noticing that  $F^{-1}(t) = -\log(1-t)$  for the standard exponential distribution  $F(x) = 1 - e^{-x}$ ,  $x \geq 0$ , so that  $F^{-1}(U) =_d Y \sim \text{Exponential}(1)$ . The convergence in (1) is justified by noting that the measure  $\mu_N$  corresponds to the uniform measure on the set  $\{1/(N+1), \dots, N/(N+1)\}$ , 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  $\mu_N$ : choosing  $\delta = 1/2$ ,

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

so

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

as  $\lambda \rightarrow \infty$ . Similarly, the variance is

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

where

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

The convergence can again be justified by a uniform integrability argument. B. The Wald-Wolfowitz-Noether-Hájek finite-sampling CLT yields  $(\bar{Y}_n - \mu_N)/\sigma_N \rightarrow_d N(0,1)$  as long as  $0 < \liminf(n/N) \leq \limsup(n/N) < 1$  if we show that the Noether condition holds. But the Noether condition is

$$\eta_N \equiv \frac{\max_{1 \leq i \leq N} |z_i - \bar{z}_N|}{\sum_{i=1}^N (z_i - \bar{z}_N)^2} \rightarrow 0.$$

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

$$N^{-1} \max_{1 \leq i \leq N} |z_i - \bar{z}_N|^2 \rightarrow 0.$$

Now since  $z_i$  increases with  $i$ ,

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

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

$$\begin{aligned} N^{-1} \max_{1 \leq i \leq N} |z_i - \bar{z}_N|^2 &\leq N^{-1} \bar{z}_N^2 \vee N^{-1} (\log(N+1) - \bar{z}_N)^2 \\ &\rightarrow 0 \vee 0 = 0. \end{aligned}$$

C. Under the null hypothesis  $\bar{Y}_n$  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  $Y_i$ ,  $i = 1, \dots, n$  in the combined sample,  $X_1, \dots, X_m, Y_1, \dots, Y_n$ .

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?

**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  $E_{\text{uniform}} \phi(U_{(i)})$  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  $X_1, \dots, X_n$  are independent Exponential(1) random variables. Let  $Y_i \equiv X_{(i)}$ , for  $i = 1, \dots, n$ , denote the *order statistics* corresponding to  $X_1, \dots, X_n$ . (a) Show that the vector  $(Y_1, \dots, Y_n)$  has the same joint distribution as  $(W_1, \dots, W_n)$  where

$$W_i \equiv \sum_{j=1}^i Z_j / (n - j + 1) \quad (2)$$

and  $Z_1, \dots, Z_n$  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  $0 \leq W_1 \leq \dots \leq W_n$  and

$$Z_i = (n - i + 1)(W_i - W_{i-1}), \quad i = 1, \dots, n \quad (3)$$

(with  $W_0 \equiv 0$ ). Let  $g(\underline{Z}) \equiv \underline{W}$  be the map defined by (2) so that  $g^{-1}(\underline{W}) = \underline{Z}$  is given in (3). Then the Jacobian of  $g^{-1}$  has entries  $n, (n-1), \dots, 1$  on the diagonal, entries  $-(n-1), \dots, -2, -1$  below the diagonal, and zero elsewhere. Hence  $\det(J_{g^{-1}}) = \text{tr}(J_{g^{-1}}) = n!$  and the density of  $\underline{W}$  is given by

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

on the set  $0 \leq w_1 \leq \dots \leq w_n < \infty$  where  $f(x) = \exp(-x)1_{[0, \infty)}(x)$  is the standard exponential density. Hence  $\underline{Y} \stackrel{d}{=} \underline{W} \equiv \underline{X}_{(\cdot)}$  where  $X_1, \dots, X_n$  are i.i.d. exponential(1).

- (b) It follows immediately from (a) that

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

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

and

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

for any fixed  $i, j$ .