

Statistics 583, Problem Set 3

Wellner; 4/13/2011

Reading: Chapter 6, section 6.5 (additional optional reading: van der Vaart, AS, chapter 13);
Chapter 7, sections 7.1- 7.4 (to be handed out on Monday, 4/18);
Wasserman, Chapters 1-2, pages 1-24.

Due: Wednesday, April 20, 2011

Reminder: Make-up lecture: Friday, 15 April; room TBD.

- What is the locally best rank test of $F = G$ against $G = (e^{\theta F} - 1)/(e^\theta - 1)$, $\theta > 0$?
 - What is the locally best rank test of $F = G$ against $G = F/(e^\theta(1 - F) + F)$?
 - What can you say about the power of these tests (other than the fact that they are locally most powerful)?
- Suppose that an urn contains N balls with the numbers $z_i = -1 - \log(1 - i/(N + 1))$, $i = 1, \dots, N$ 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.
 - Calculate the mean $\mu_N = E(\bar{Y}_n)$ and variance $\sigma_N^2 = Var(\bar{Y}_n)$ of \bar{Y}_n . Find the limits of \bar{z}_N and σ_z^2 as $N \rightarrow \infty$.
 - Use the Wald-Wolfowitz-Noether-Hajek finite-sampling CLT to prove that $(\bar{Y}_n - \mu_N)/\sigma_N \rightarrow_d N(0, 1)$.
 - 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 , noting the two different uses of the notation “ Y_1, \dots, Y_n ”)?
- 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 .
 - Show that the vector (Y_1, \dots, Y_n) has the same joint distribution as (W_1, \dots, W_n) where $W_i \equiv \sum_{j=1}^i Z_j/(n - j + 1)$ and Z_1, \dots, Z_n are i.i.d. Exponential(1).
 - Use the result of (a) to compute $E(Y_i)$, $Var(Y_i)$, and $Cov(Y_i, Y_j)$ for any fixed i, j .
- Suppose that, in Example 6.3.15, page 29, $1 - F_i = (1 - F)^{\Delta_i}$ where $\Delta_i = \exp(\theta z_i)$ and z_1, \dots, z_N are given real numbers and $\theta \in \mathbb{R}$. Then the distribution of the

ranks of X_1, \dots, X_N (independent with respective d.f.'s F_1, \dots, 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 $a_N(j)$, $j = 1, \dots, N$, 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 $z_1 = \dots = z_m = 0$ and $z_{m+1} = \dots = z_N = 1$, 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

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

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

5. **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.

6. **Optional bonus problem 2:** Let X_1, X_2, \dots, X_N be a sample from a distribution with density $f_\theta(x) = \theta \exp(\theta x) 1\{x < 0\}$, $\theta > 0$, and let $V_{(1)} < V_{(2)} < \dots < V_{(N)}$ denote the order statistics. Show that $Y_1 = V_{(1)} - V_{(2)}$, $Y_2 = V_{(2)} - V_{(3)}$, \dots , $Y_{N-1} = V_{(N-1)} - V_{(N)}$, $Y_N = V_{(N)}$ are independent random variables and that Y_j has density $j\theta e^{j\theta x} 1\{x < 0\}$ for $j = 1, \dots, N$. Use this fact to determine the rejection region of the test you found in problem 5 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 5. Show that the resulting test is a most powerful rank test of $H : F = G$ versus $K : G = F^2$.