MATHEMATICAL STATISTICS
BANACH CENTER PUBLICATIONS, VOLUME 6
PWN—POLISH SCIENTIFIC PUBLISHERS
WARSAW 1980

RANK TESTS FOR THE HYPOTHESES CONCERNING UNIDENTIFIABLE OBJECTS

E. KHMALADZE and E. PARSADANISHVILY

Institute of Economics and Rights of Georgian Acad. Sc., Tibilisi, Georgian SSR

1. Consider n pairs of balls with diameters X_{ij}, $i = 1, \ldots, n$, $j = 1, 2$, being mutually independent random variables (r.v.'s). In each pair, let one ball be of, say, green ($j = 1$) colour and the other ball of blue ($j = 2$) colour. Consider the hypothesis that the diameters of the green and the blue balls are equally distributed ($F_1(x) \equiv F_2(x)$) against the alternative that the diameters of the blue balls are stochastically larger than the diameters of the green ones.

The question is how to test these hypotheses when it is not possible to distinguish between the colours.

Examples of such situations one can find, e.g., in radiolocation or in cytogenetics, in the investigation of homological chromosomes.

2. Let $Y_{i1} = \min(X_{i1}, X_{i2})$ and $Y_{i2} = \max(X_{i1}, X_{i2})$, and let S_{ij} denote the rank of r.v. X_{ij} among all X's. Then the rank of Y_{i1} is $\min(S_{i1}, S_{i2}) = R_{i1}$ and the rank of Y_{i2} is $\max(S_{i1}, S_{i2}) = R_{i2}$.

The fact that it is not possible to distinguish between the colours means formally that a (rank) test should be some function of Y_{i1} and Y_{i2} (R_{i1} and R_{i2}), $i = 1, \ldots, n$, i.e. for all $i = 1, \ldots, n$ it should be a symmetric function of X_{i1} and X_{i2} (S_{i1} and S_{i2}).

As a particular example of test statistics one can consider the statistics

$$\min_i R_{i2} = \min_i \max_j S_{ij}. \tag{1}$$

The possibility of testing our hypotheses is based, roughly speaking, on the following fact: although, for all i and k, r.v. R_{i2} is stochastically larger than r.v. R_{k1} under both the hypothesis and the alternative, still this "increase" is "larger" under the alternative. Just this may be tested statistically.

3. One might mention here some exact distributions of rank statistics. For example, marginal distributions of R_{ij} are

$$P\{R_{i1} = r\} = \frac{2n-r}{n(2n-1)} \quad \text{and} \quad P\{R_{i2} = r\} = \frac{r-1}{n(2n-1)} \tag{2}$$

and the distribution of a vector $R_1 = (R_{11}, \dots, R_{n1})$ is

$$(3)\quad P\{R_{11} = r_1, \dots, R_{n1} = r_n\} = \frac{(2n-r_n)(2n-r_{n-1}-2)\cdot\ldots\cdot(2-r_1)}{(2n)!}(2!)^n, \quad r_1 < r_2 < \ldots < r_n,$$

if all values in brackets are positive, and 0 otherwise. The distribution of statistics (1) is

$$(4)\quad P\{\min R_{i2} = r\} = \frac{n!\,(2n-r)!\,2^{r-1}(r-1)}{(2n)!\,(n-r+1)!}.$$

Now it is easy to see that

$$(5)\quad P\{R_{i2} \leqslant 2tn\} \to P\{\max(U_1, U_2) \leqslant t\}$$

and

$$(6)\quad P\Big\{\min_i R_{i2} \leqslant t/\sqrt{2n}\Big\} - P\Big\{\min_i \max_j U_{ij} \leqslant t/\sqrt{2n}\Big\} \to 0,$$

where U with different indices denotes independent r.v.'s uniformly distributed over $[0, 1]$.

According to what has been said above one may prove that under the alternative r.v. R_{i2} tends weakly to some r.v., which is stochastically larger than $\max(U_1, U_2)$.

4. Derive now the statistics of the *locally most powerful* (LMP) test and the LMP rank test for given parametric subhypotheses. Namely, suppose that the distribution functions F_1 and F_2 of X's belong to some parametric family $H, \{F_\theta(x), \theta \in \Theta\}$, where Θ is assumed simply to be a neighbourhood of 0 on a real line. Suppose that under the hypothesis $F_1 = F_2 = F_0$ and under the alternative $F_1 = F_{\theta_1}$ and $F_2 = F_{\theta_2}$. One may now consider the Neyman–Pearson statistics

$$(7)\quad L = \sum_{i=1,\dots,n} \ln \frac{dG_{\theta_1,\theta_2}}{dG_{0,0}}(Y_{i1}, Y_{i2})$$

where $G_{\theta_1,\theta_2}(y_1, y_2) = F_{\theta_1}(y_1)F_{\theta_2}(y_2) + F_{\theta_1}(y_2)F_{\theta_2}(y_1)$, and expect that, as $\theta_1 = c_1/\sqrt{n}$ and $\theta_2 = c_2/\sqrt{n}$ tend to 0, the statistics L should lead to the LMP test while its conditional expectation $E(L|S_1, S_2)$ given ranks $S_j = (S_{1j}, \dots, S_{nj})$, $j = 1, 2$, should correspond to the LMP rank test — just as in the case of usual situations with statistics

$$(8)\quad M = \sum_{\substack{i=1,\dots,n\\ j=1,2}} \ln \frac{dF_{\theta_j}}{dF_0}(X_{ij}).$$

More precisely, under usual regularity conditions, statistics M is asymptotically normal under both the hypothesis and a sequence of alternatives $\theta_1 = c_1/\sqrt{n}$ and $\theta_2 = c_2/\sqrt{n}$ with c_1 and c_2 fixed, and the asymptotic expression of the power of the test $M > c$ of level α is

$$(9)\quad 1 - \varphi\big(t_{1-\alpha} - b\sqrt{c_1^2 + c_2^2}\big),$$

where

$$b^2 = E_0 h^2(X) \quad \text{with} \quad h(x) = \frac{\partial}{\partial\theta} \ln \frac{dF_\theta}{dF_0}(x)\Big|_{\theta=0}$$

and $\varphi(x)$ denotes the standard normal distribution function. The asymptotic expression of the power of the test $E(M|S_1, S_2) > c$ is

$$(10)\quad 1 - \varphi\big(t_{1-\alpha} - b|c_1 - c_2|/\sqrt{2}\big),$$

so that the two tests are asymptotically equivalent iff $c_1 = -c_2$. For other choices of c_1 and c_2 the power given by (9) is greater than that given by (10).

For the statistics (7) the situation is different: if $c_1 \neq -c_2$, then, under usual regularity conditions,

$$(11)\quad L - \frac{1}{2\sqrt{n}} \sum_{\substack{i=1,\dots,n\\ j=1,2}} h(X_{ij}) \cdot (c_1 + c_2) - \text{const} = o_P(1)$$

under both the hypothesis and a sequence of alternatives, so that the asymptotic expression of the power of the test $L > c$ is

$$(12)\quad 1 - \varphi\big(t_{1-\alpha} - b|c_1 + c_2|/\sqrt{2}\big).$$

But

$$(13)\quad E(L|R_1, R_2) = \sum_{r=1,\dots,2n} E(L(X) \mid rgX = r) + o_P(1) = \text{const} + o_P(1),$$

and this implies that for $c_1 \neq -c_2$ rank tests cannot distinguish our hypotheses.

If $c_1 = -c_2$, or, more precisely, $\theta_1 = -\theta_2$ and we are still interested in local considerations, then (11) is no longer useful: θ's should tend to 0 only as $c/n^{1/4}$ and instead of (11) one may get

$$(14)\quad L - c^2 \frac{1}{2\sqrt{n}} \sum_{\substack{i=1,\dots,n\\ j=1,2}} [h^2(X_{ij}) + k(X_{ij})] - c^2 \frac{1}{2\sqrt{n}} \sum_{i=1,\dots,n} h(X_{i1})h(X_{i2}) - \text{const} = o_P(1),$$

where

$$k(x) = \frac{\partial^2}{\partial\theta^2} \ln \frac{dF_\theta}{dF_0}(x)\Big|_{\theta=0}.$$

Now, without performing routine calculations of the asymptotic power, it may be mentioned that

$$(15)\quad E(L|R_1, R_2) = \sum_{i=1,\dots,n} E(h(X_{i1})h(X_{i2}) \mid R_{i1}, R_{i2}) + \text{const} + o_P(1),$$

since the conditional expectation of the first sum in (14) is constant. The last expression implies that the LMP rank test can now distinguish the hypotheses considered but it remains in general less powerful than the LMP one.

5. In this note only a brief description of the problem has been given. Nothing has been said about the tests based on nonlinear statistics (as that given by (1)), or about tests for alternatives of a different type when, essentially, there is no asymptotic normality of the test statistics. A more detailed discussion of the problems in question will appear in Teor. Verojatnost. i Primenen. vol. 24.

Presented to the semester
MATHEMATICAL STATISTICS
September 15–December 18, 1976

MATHEMATICAL STATISTICS
BANACH CENTER PUBLICATIONS, VOLUME
PWN—POLISH SCIENTIFIC PUBLISHERS
WARSAW 1980

A NOTE ON CONFIDENCE INTERVALS FOR THE KIEFER–WOLFOWITZ PROBLEM

JACEK KORONACKI

Institute of Mathematics, Polish Academy of Sciences, Warszawa, Poland

Introduction

In this simple note we shall be concerned with the so-called Kiefer–Wolfowitz situation; i.e. with the problem of finding a point θ of minimum of a (regression) function f: $R \to R$, R denoting the real line, when the only information available is that we can observe unbiased estimates of function values of f. Namely, we shall apply the method of Farrell [2] to obtain a confidence interval for θ, of length not exceeding any predetermined positive number. Originally, the method was employed for finding a confidence interval for the zero of a regression function. It provides some suitable stopping rule for the experimentation process based upon one of the stochastic approximation procedures (depending on the situation at hand, either on the procedure of Kiefer–Wolfowitz or on that of Robbins–Monro).

Let the sequential procedure for estimating point θ of minimum of a regression function be of the form(1):

$$X_{n+1} = X_n - a_n Y_n, \tag{1}$$

where a_n are positive numbers, X_1 and Y_n are r.v.'s, $n \geqslant 1$. It is well known that under assumptions (A1)–(A5), as well as under (A4)–(A7) (see below), $\lim_{n\to\infty} X_n = \theta$. (Equation (1) with Y_n given by (A4) defines the original Kiefer–Wolfowitz procedure.)

Now, the method can loosely be described as follows. Let $\underline{\theta} \leqslant \theta \leqslant \bar{\theta}$, for some known $\underline{\theta}$ and $\bar{\theta}$. Let $\{X_n^{(i)}, n \geqslant 1\}$, $i = 1, \ldots, 2k$ be $2k$ sequences of r.v.'s obtained by the use of $2k$ simultaneous and independent Kiefer–Wolfowitz procedures starting, respectively, with $X_1^{(i)} = \underline{\theta}$ for $i = 1, \ldots, k$ and $X_1^{(i)} = \bar{\theta}$ for $i = k+1, \ldots, 2k$. These sequences are stopped at such random moments, say M and N, that $X'_M \equiv \min\{X_M^{(1)}, \ldots, X_M^{(k)}\}$ and $X''_N = \max\{X_N^{(k+1)}, \ldots, X_N^{(2k)}\}$ have enabled one to construct a confidence interval for θ, with the confidence level $1-\alpha$ and length $\leqslant L$, α and L given in advance.

(1) All random variables (r.v.'s) are assumed to be defined on a probability space $(\Omega, \mathcal{F}, P)$ and relations between r.v.'s are meant with probability one.