

ON THE CONSISTENCY OF A TWO-SAMPLE MATCHING TEST

ÜLKÜ GÜRLER^{1,*} and M. M. SIDDIQUI²

¹*Bilkent University Ankara, Turkey*

²*Colorado State University, Fort Collins, Colorado USA*

(Received: May 25, 1995; Revised: November 28, 1995; Accepted: February 2, 1996)

Let $\{X_k\}$ and $\{Y_k\}$, $1 \leq k \leq n$ be the order statistics of independent random samples from continuous distribution function F and G respectively. To test the null hypothesis $H_0: G = F$, known, against the alternative $H_1: G \neq F$, a test S_n , based on the number of matches between the two samples was suggested by Siddiqui and Gürler (1992). In this note the asymptotic distribution of S_n under the null hypothesis is obtained and its consistency against a fixed alternative is shown.

AMS 1991 Subject Classification: 62G30, 62G10.

KEYWORDS: Matching, two sample test, consistency.

1. INTRODUCTION

Let X_k and Y_k , $1 \leq k \leq n$ be the order statistics of independent random samples from continuous distributions with cdf's F and G respectively. To test the null hypothesis $H_0: G = F$, known, against the alternative $H_1: G \neq F$, where F is specified, Siddique and Gürler (1992) suggested a test which considers the 'matches' between the order statistics of the two samples. This test was an extension of the one investigated by Siddiqui (1982) for the one sample case. The exact and large sample expressions for the first two moments of the test statistic were provided. However, the form of its limiting distribution was left as a conjecture. In this note, the exact and large sample expressions for the r^{th} moment of the test statistic are provided and the conjecture about the limiting distribution is proved. Consistency of the test is then established under a fixed alternative.

For the above hypothesis, without loss of generality, assume that the distributions F and G are concentrated on the unit interval $[0, 1]$, and that $F(x) = x$, $0 \leq x \leq 1$. The test suggested by Siddique and Gürler (1982) is based on the number of matches between the order statistics X_k and Y_k , $k = 1, \dots, n$, from F and G respectively. The event A_k , a "match" occurs if $X_k \in (Y_{k-1}, Y_k]$ for any k , $1 \leq k \leq n$ with $Y_0 = 0$. The test statistic is $S_n = \sum_{k=1}^n I_k$, where I_k is the indicator of A_k . Let P_0 and P_1 be the probability measures and E_0 and E_1 be the expectations under H_0 and H_1 respectively. It is intuitively obvious that, for $r \geq 1$, $P_0(S_n \geq r) \geq P_1(S_n \geq r)$. Hence, for a significance level α , $0 < \alpha < 1$, the critical region for testing H_0 against H_1 is of the

*e-mail:ulku@bilkent.edu.tr

form: $S_n \leq r_{n,\alpha}$. Here, $r_{n,\alpha}$ is the largest integer r such that $P(S_n \leq r) \leq \alpha$. The main problem is, to find $P_0(S_n \leq r)$, $0 \leq r \leq n$.

2. THE NULL DISTRIBUTION AND MOMENTS OF S_n

An exact expression for $P_0(S_n \geq r)$ is provided below, which is of more theoretical interest except for small n . To obtain the large sample distribution of S_n , $E_0 S_n^r$, $r \geq 1$ are computed and it is shown that the moments of $n^{-1/2} S_n$ converge to the moments of a known distribution. Observe that S_n counts the number of events A_k , $1 \leq k \leq n$, and $P(S_n \geq r)$ is the probability that at least r of the n events will occur. Hence we have

$$P(S_n \geq r) = \sum_{m=r}^n \binom{m-1}{r-1} p(n, r) \quad (1)$$

with

$$p(n, r) = \sum' P(A_{k_1}, A_{k_2}, \dots, A_{k_r}) \quad (2)$$

where \sum' denotes the summation over all k_j such that $1 \leq k_1 < k_2 < \dots < k_r \leq n$. The lemma below, which follows from integration by parts of Beta functions will be useful. The details of all the proofs in this manuscript can be found in Gürler and Siddiqui (1995):

LEMMA 1. Let $\tau = \{0 = y_0 < y_1 < \dots < y_s \leq 1\}$ be a partition of $[0, 1]$. For $v \geq y_s$, $0 \leq m \leq s-1$, define

$$I_\tau(r, m, s; v) = \sum_{r=m+1}^s \frac{(s-m)!}{(r-m-1)!(s-r)!} \int_{y_{r-1}}^{y_r} (u - y_m)^{r-m-1} (v-u)^{s-r} du \quad (3)$$

Then

$$I_\tau(r, m, s; v) = \sum_{r=m+1}^s \binom{s-m}{r-m} (y_r - y_m)^{r-m} (v - y_r)^{s-r}$$

THEOREM 1. Given $\tau = \{0 = y_1 < y_2 < \dots < y_n\}$, let $p(n, r|\tau)$ refer to the probability $p(n, r)$ in 2 for the partition τ . Then

$$p(n, r|\tau) = \frac{n!}{k_1!(k_2 - k_1)! \dots (k_r - k_{r-1})!(n - k_r)!} \cdot y_{k_1}^{k_1} (y_{k_2} - y_{k_1})^{k_2 - k_1} \dots (y_{k_r} - y_{k_{r-1}})^{k_r - k_{r-1}} (1 - y_{k_r})^{n - k_r}$$

Proof. Considering the joint distribution of the order statistics from the uniform distribution,

$$p(n, r|\tau) = \frac{n!}{(k_1 - 1)!(k_2 - k_1 - 1)!(k_r - k_{r-1} - 1)!(n - k_r)!} \cdot \int_{y_{k_{r-1}}}^{y_{k_r}} \int_{y_{k_{r-1}-1}}^{y_{k_{r-1}}} \dots \int_{y_{k_1-1}}^{y_{k_1}} u_1^{k_1-1} (u_2 - u_1)^{k_2 - k_1 - 1} \dots (u_r - u_{r-1})^{k_r - k_{r-1} - 1} (1 - u_r) du_1 \dots du_r$$

where $k_0 = 0$. The proof is completed by repeated application of Lemma 1. \square

The Theorem above provides the distribution of S_n for a fixed partition τ . The result for a uniform random partition is presented below.

THEOREM 2. Let $p_0(n, r) = E_0 p(n, r | \tau)$ where $\tau = \{0 = Y_1 < Y_2 < \dots < Y_n\}$. Then

$$p_0(n, r) = 2^{-r} \binom{2n}{r}^{-1} \sum \binom{2k_1}{k_1} \binom{2k_2 - 2k_1}{k_2 - k_1} \dots \binom{2k_r - 2k_{r-1}}{k_r - k_{r-1}} \binom{2n - 2k_r}{n - k_r}$$

Theorem 2 and Stirling's approximation of factorials imply the following:

COROLLARY

$$(a) \quad P_0(S_n \geq r) = \sum_{m=r}^n (-1)^{m-r} \binom{m-1}{r-1} p_0(n, r)$$

$$(b) \quad \text{For } r \leq n, \lim_{n \rightarrow \infty} n^{-r/2} p_0(n, r) = \frac{\Gamma(r/2 + 1)}{r!}$$

The exact moments of S_n are then obtained via the following lemma which is obtained by multinomial expansion of $S_n^r = (I_1 + I_2 + \dots + I_n)^r$ and noting that $I_j^s = I_j$ for any $j, s > 0$.

LEMMA 2. $E_0 S_n^r = \sum_{k=1}^r C(r, k) p_0(k, n)$, where $C(r, k) = \sum_{j_1, \dots, j_k} \frac{r!}{j_1! \dots j_k!}$ and the summation extends over j_1, j_2, \dots, j_k such that $j_i \geq 1, i = 1, \dots, k$ and $\sum_{i=1}^k j_i = r$.

Theorem 3 below provides the asymptotic distribution of $n^{-1/2} S_n$, from which approximate critical region of the test can be obtained.

THEOREM 3. For $x \geq 0, \lim_{n \rightarrow \infty} P_0(n^{-1/2} S_n \geq x) = e^{-x^2}$.

Proof. Using part (b) of the above Corollary and Lemma 2 we have

$$\begin{aligned} \lim_{n \rightarrow \infty} E_0 (n^{-1/2} S_n)^r &= \Gamma\left(\frac{r}{2} + 1\right) \\ &= \int_0^\infty x^r dH(x) \end{aligned} \quad (4)$$

where $H(x) = 1 - e^{-x^2}$. The result now follows from observing that $H(x)$ belongs to the exponential family and is completely determined by its moments. \square

3. CONSISTENCY OF THE TEST

Let \mathcal{F} be a class of distribution functions which are absolutely continuous w.r.t. lebesgue measure and with support on the entire interval $(0, 1)$. Consider the null hypothesis $H_0: G = F = U(0, 1)$ versus $H_1: G \neq U(0, 1), G \in \mathcal{F}$. We require that $G \in \mathcal{F}$, otherwise if G is singular w.r.t. F or if G has a different support than that of F , then the problem will be trivial from a statistical point of view. In fact we will see below that the worst case occurs when F and G coincide at countably infinite (but not dense) number of points in any subinterval on $(0, 1)$.

For $0 \leq y \leq 1$, let $Q(y) = G^{-1}(y) = \inf\{y: G(x) \geq y\}$ be the quantile function. Similarly the quantiles of the empirical d.f. $G_n(x)$ are defined as $Q_n(y) = G_n^{-1}(y) = \inf\{y: G_n(x) \geq y\} = Y_k$ if $y \in (k-1/n, k/n]$, $k=1, \dots, n$. The normed sample quantile process is defined as below, where g is the density of G :

$$\rho_n(y) = \sqrt{ng(Q(y))} [Q_n(y) - Q(y)] \quad 0 < y < 1, n=1, 2, \dots \quad (5)$$

Let $\xi_k = G^{-1}(k/n) = Q(k/n)$, $\rho_{k,n} = \rho_n(k/n)$ and

$$H_k(\rho_k, n) = \left[\xi_k + (\sqrt{ng(\xi_k)})^{-1} \rho_n\left(\frac{k}{n}\right) \right]^k \left[1 - \xi_k - (\sqrt{ng(\xi_k)})^{-1} \rho_n\left(\frac{k}{n}\right) \right]^{n-k} \quad (6)$$

Then we can write:

$$\begin{aligned} P(n, 1|\tau) &= \sum_{k=1}^n \binom{n}{k} \left[\xi_k + (\sqrt{ng(\xi_k)})^{-1} \rho_n\left(\frac{k}{n}\right) \right]^k \left[1 - \xi_k - (\sqrt{ng(\xi_k)})^{-1} \rho_n\left(\frac{k}{n}\right) \right]^{n-k} \\ &= \sum_{k=1}^n \binom{n}{k} H_k(\rho_k, n) \end{aligned} \quad (7)$$

Consistency of the proposed test is established by computing the expectation of $H_k(\rho_k, n)$ as a function of the normed sample quantile process under H_1 and hence derive the unconditional expectation of S_n . The result is stated in Theorem 4, the proof of which utilizes the lemma below; whose proof is calculus based.

LEMMA 3. For a fixed $t \in (0, 1)$, let $S(x) = x^t(1-x)^{1-t} \exp\{1/2(t-x)^2/[t(1-t) + (t-x)^2]\}$. Then $S(t) = t^t(1-t)^{1-t}$ is the unique maximum of $S(x)$ in $0 \leq x \leq 1$.

THEOREM 4. Let $G \in \mathcal{F}$ be an arbitrary but a fixed alternative. If the Lebesgue measure of the set $\{t: G(t) \neq t\}$ is equal one, then the test based on S_n is consistent for any significance level α , $0 < \alpha < 1$.

Proof. From Chebychev's inequality

$$P_1(n^{-1/2}S_n \leq x) \geq 1 - \frac{E_1(S_n)}{\sqrt{nx}}$$

Therefore, a sufficient condition for the test to be consistent is that $\lim_{n \rightarrow \infty} E_1(n^{-1/2}S_n) = 0$. We have

$$\begin{aligned} \log H_k(\rho_k, n) &= k \log \left[\xi_k + (\sqrt{ng(\xi_k)})^{-1} \rho_n\left(\frac{k}{n}\right) \right]^k \\ &\quad + (n-k) \log \left[1 - \xi_k - (\sqrt{ng(\xi_k)})^{-1} \rho_n\left(\frac{k}{n}\right) \right] \\ &\sim k \log \xi_k + (n-k) \log(1 - \xi_k) \\ &\quad + \frac{(k - n\xi_k)\rho_n\left(\frac{k}{n}\right)}{\sqrt{ng(\xi_k)}\xi_k(1 - \xi_k)} - \rho_n^2\left(\frac{k}{n}\right) \frac{(k - 2k\xi_k + n\xi_k^2)}{2ng^2(\xi_k)\xi_k^2(1 - \xi_k)^2} \end{aligned}$$

It follows from Csörgö (1983, p. 13) that $\rho_n(y) \xrightarrow{D} \rho(y) \sim N(0, y(1-y))$ and

$$E\left[\exp\left\{\lambda Y - \frac{b}{2}W^2\right\}\right] = \exp\left\{\frac{1}{2}\frac{\lambda^2\sigma^2}{(1+b\sigma^2)}\right\}/(1+b\sigma^2)^{1/2}$$

where $W \sim N(0, 1)$. Then, with $k = nt$ and $Q_t = Q(t)$,

$$E_1[H_k(\rho_{k,n})] \sim [Q_t^t(1-Q_t)^{1-t}]^n \frac{e^{na(t)}g(Q_t)Q_t(1-Q_t)}{\sqrt{C(t)}}$$

where

$$a(t) = \frac{t(1-t)(t-Q_t)^2}{2C(t)}$$

$$C(t) = g^2(Q_t)Q_t^2(1-Q_t)^2 + t^2(1-t)^2 + t(1-t)(t-Q_t)^2$$

Then

$$\begin{aligned} \lim_{n \rightarrow \infty} n^{-1/2} E_p(n, 1 | \tau) &= \lim_{n \rightarrow \infty} E_1(n^{-1/2} S_n) \\ &= \lim_{n \rightarrow \infty} \frac{1}{\sqrt{2\pi}} \int_0^1 \left[\frac{Q_t^t(1-Q_t)^{1-t}}{t^t(1-t)^{1-t}} e^{a(t)} \right]^n \frac{g(Q_t)Q_t(1-Q_t)t^{-1/2}(1-t)^{-1/2}}{C(t)} dt \end{aligned}$$

Note that if $Q_t = t$, the function in square brackets is one and above limit is

$$\frac{1}{\sqrt{2\pi}} \int_{\{Q_t=t\}} \frac{1}{\sqrt{2\pi}} \frac{1}{t(1-t)} dt$$

which is zero if the measure of $\{t: Q(t) = t\} = \{t: G(t) = t\}$ is zero. For $Q(t) \neq t$, first note that $a(t) \leq \frac{1}{2}(t-Q_t)^2/t(t-t) + (t-Q_t)^2$. Then by Lemma 3, $Q_t(1-Q_t)^{1-t}e^{a(t)} < t^t(1-t)^{1-t}$ and by the dominated convergence theorem, the limit of the integral is again zero. \square

References

1. Csörgö, M. (1983) *Quantile Processes with Statistical Applications*, SIAM, Philadelphia.
2. Gürler Ü. and Siddiqui, M. M. (1995): "Large Sample Behavior of a Two Sample Matching Test", Res. Rep. IEOR 95-23, Dept. of Industrial Engineering, Bilkent University, Turkey.
3. Hájek, J. Šidák, Z. (1967) *Theory of Ranked Tests*, Academic Press, New York.
4. Siddiqui, M. M. (1982) The consistency of a matching test. *J. Statist. Plan. Inf.*, 6, 227-233.
5. Siddiqui, M. M. and Gürler Ü. (1992): A Two Sample Matching Test, in "Order Statistics and Non-parametrics: Theory and Applications". P. K. Sen and I. A. Salama (Eds.), p: 237-243. Elsevier Science Publishers B.V.