Statistics & Probability Letters 1 (1983) 129–135
North-Holland Publishing Company March 1983

# One- and Two-Sample Match Statistics

J.S. Rao

*Department of Mathematics, University of California at Santa Barbara, Santa Barbara, CA 93106, U.S.A.*

Ram C. Tiwari

*University of Allahabad, Allahabad, India*

Received August 1982; revised version received October 1982

*Abstract.* A match statistic considered by Khidr (1981) is interpreted in terms of crossings of the empirical and true distribution functions and a simpler alternate derivation of its distribution provided. This approach can also be used to obtain the distribution of a two-sample match statistic, considered earlier by Takàcs (1971).

*Keywords.* Order statistics, matches, spacing-frequencies, one- and two-sample problems.

## 1. Introduction

Khidr (1981) considered the problem of matching order statistics with intervals and developed a test for goodness of fit based on the number of 'matches'. Let $U_{in}$, $i = 1,\ldots,n$ be the order statistics from a random sample of size $n$ from U(0, 1), the uniform distribution on the unit interval. Given an arbitrary partition of [0, 1] into $n$ intervals, a match is said to occur if the $i$th order statistic falls in the $i$th interval. Among all possible partitions of [0, 1], the expected number of matches is shown there to be a maximum corresponding to that partition whose $i$th interval is $((i-1)/n, i/n]$, $i = 1,\ldots,n$. We consider therefore this statistic $M_n$ to be the total number of matches, where a match is said to occur at the $i$th place if $U_{in}$ falls in $((i-1)/n, i/n]$. In Section 2 of this paper, a simple derivation of the exact distribution of $M_n$ is given. This derivation makes use of the multinomial frequency counts in these fixed intervals after relating it to the number of 'crossings' in the Kolmogorov–Smirnov statistic.

Section 3 considers a similar match statistic for the two-sample problem. Assuming equal sample sizes, a match is said to occur at the $i$th place if the $i$th order statistic of the first sample falls inbetween the $(i-1)$st and $i$th order statistics of the second sample. This two-sample match statistic, $M_n^*$, is related to the number of crossings of the two empirical distribution functions and has been considered in Takàcs (1971). We briefly indicate how the method of Section 2 can be applied to derive the distribution of $M_n^*$. A discussion of some possible alternate definitions of a 'match' as well as some Monte Carlo power comparisons will be considered in a separate paper.

## 2. Matches in the one-sample problem

Let $U_i$, $i = 1,\ldots,n$ denote the *order statistics* in a random sample of size $n$ from U(0, 1) where we omit the second subscript $n$ for notational convenience. Let $I_i = ((i-1)/n, i/n]$, $i = 1,\ldots,n$ be a partition of [0, 1].

0167-7152/83/0000-0000/$03.00 © 1983 North-Holland

We say a match occurs in the $i$th interval if $U_i$ falls in the interval $I_i$. Let $M_n$ denote the total number of matches.

Let $R_i$ be the number of $U_j$'s falling in the interval $I_i$ and let $T_i = \Sigma_{j=1}^{i} R_j$ be the cumulative frequencies, i.e., the total number of $U_j$'s falling below $i/n$ which includes all intervals upto and including the $i$th interval. Clearly the joint distribution of $(R_1, \ldots, R_n)$ is a multinomial with $n$ trials and cell probabilities $(1/n, \ldots, 1/n)$ – this will be denoted by $M(n;\ (1/n, \ldots, 1/n))$. Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables having a Poisson distribution with parameter $\lambda > 0$, i.e., $X_i \sqcap \text{Poi}(\lambda)$. Then it is well known that the joint distribution of $(X_1, \ldots, X_n)$ given $\Sigma_1^n X_i = n$ is $M(n;\ (1/n, \ldots, 1/n))$. Thus the frequencies $(R_1, \ldots, R_n)$ have the same joint distribution as the random variable $(X_1, \ldots, X_n)$ conditional on the event $\Sigma_1^n X_i = n$. In symbols,

$$(R_1, \ldots, R_n) \sqcap \left( X_1, \ldots, X_n \,\middle|\, \sum_1^n X_i = n \right) \tag{1}$$

where '$\sqcap$' stands for 'same in distribution as'.

**Lemma 2.1.** *The total number of matches $M_n$ is equal to the number of $T_i = \Sigma_{j=1}^{i} R_j$ which equal $i$, i.e., $M_n = \Sigma_{i=1}^{n} I(T_i = i)$ where $I(\cdot)$ stands for the indicator function.*

**Proof.** Let $F_n(x) = n^{-1}\Sigma_{i=1}^{n} I(U_i \leqslant x)$ be the empirical c.d.f. of the sample. Let $F(x) = x$, $0 \leqslant x \leqslant 1$, denote the c.d.f. corresponding to the U(0, 1). We formulate the proof in terms of the 'crossings' of $F_n$ and $F$, which provides a nice interpretation to the statistic $M_n$. We say an *upcrossing* occurs at the $U_i$ if $F_n(U_i - 0) < F(U_i) \leqslant F_n(U_i)$. Thus an upcrossing occurs at $U_i$ iff $((i-1)/n) < U_i \leqslant (i/n)$, i.e., iff there is a match in the $i$th interval. The total number of matches is then equal to the number of upcrossings. We say a *horizontal crossing* occurs at $U_i$ if the step function, $F_n(x)$ over the interval $[U_i, U_{i+1})$ intersects the line $F(x) = x$. Since $F_n(x) = i/n$ over this interval, clearly a horizontal crossing occurs at $U_i$ iff $F_n(i/n) = i/n$; i.e., $T_j = i$. Note that between any two upcrossings (horizontal crossings) there is a horizontal crossing (upcrossing). Thus the total number of matches $M_n$ is also equal to the total number of horizontal crossings; that is, $M_n = \#\{T_i = i\}$, $i = 1, 2, \ldots, n$. □

**Lemma 2.2.** *For $k = 0, 1, \ldots, n$, define*

$$S_{n,k} = \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant n} \mathbf{P}\left(T_{i_1} = i_1, T_{i_2} = i_2, \ldots, T_{i_k} = i_k\right).$$

*Then*

$$S_{n,k} = \frac{n!}{n^n} \sum_{j=0}^{n-k} \binom{n-j-1}{k-1} \left( \frac{n^j}{j!} \right). \tag{2}$$

**Proof.** Let $V_i = \Sigma_{j=1}^{i} X_j$ where $\{X_i\}$ are i.i.d. Poi($\lambda$) variables. From (1) and the fact that $V_i \cap \text{Poi}(i\lambda)$, we have

$$\begin{aligned} S_{n,k} &= \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant n} \mathbf{P}\left(V_{i_1} = i_1, V_{i_2} = i_2, \ldots, V_{i_k} = i_k \mid V_n = n\right) \\ &= \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant n} \frac{\mathbf{P}\left(V_{i_1} = i_1, V_{i_2} - V_{i_1} = i_2 - i_1, \ldots, V_{i_k} - V_{i_{k-1}} = i_k - i_{k-1}, V_n - V_{i_k} = n - i_k\right)}{\mathbf{P}(V_n = n)} \end{aligned}$$

$$=\frac{n!}{n^n}\sum_{1\leqslant i_1<\cdots<i_k\leqslant n}\frac{i_1^{i_1}(i_2-i_1)^{i_2-i_1}\cdots(i_k-i_{k-1})^{i_k-i_{k-1}}(n-i_k)^{n-i_k}}{i_1!(i_2-i_1)!\cdots(i_k-i_{k-1})!(n-i_k)!}$$

$$=\frac{n!}{n^n}\sum_{\substack{l_i\geqslant 0,\, j=1,2,\ldots,k+1\\ \sum_{j=1}^{k+1} l_j=n-k}}\left[\prod_{j=1}^{k}\frac{l_j^{l_j}+1}{l_j!}\right]\frac{l_{k+1}^{l_{k+1}}}{l_{k+1}!} \tag{3}$$

by putting $l_1=i_1-1$, $l_2=i_2-i_1-1,\ldots,l_k=i_k-i_{k-1}-1$, and $l_{k+1}=n-i_k$ with the convention $0^0=1$. From the multinomial Abel identities (see Riordan (1968, pp. 24–25), see also Khidr (1981)) it can be shown that

$$\sum_{\substack{l_j\geqslant 0\\ \sum_1^{k+1} l_j=n-k}}\left[\prod_{j=1}^{k}\frac{(l_j+1)^{l_j}}{l_j!}\right]\frac{l_{k+1}^{l_{k+1}}}{l_{k+1}!}=\sum_{j=0}^{n-k}\binom{k+j-1}{k-1}\frac{n^{n-k-j}}{(n-k-j)!}. \tag{4}$$

Thus (3) simplifies to (2) establishing the lemma. □

The final result now follows in exactly the same way as in Khidr (1981, p. 1406) by the inclusion–exclusion arguments. We include the proof for completeness.

**Theorem 2.3.** *For* $r=1,2,\ldots,n$

$$\mathrm{P}(M_n\geqslant r)=\prod_{j=0}^{r-1}\left(1-\frac{j}{n}\right). \tag{5}$$

**Proof.** From the inclusion–exclusion principle (cf. Feller (1968, p. 109)),

$$\mathrm{P}(M_n\geqslant r)=\sum_{j=r}^{n}(-1)^{j-r}\binom{j-1}{r-1}S_{n,j}=\frac{n!}{n^n}\sum_{j=r}^{n}\sum_{i=0}^{n-j}(-1)^{j-r}\binom{j-1}{r-1}\binom{n-i-1}{j-1}\frac{n^i}{i!}.$$

Substituting $l=n-i$ and interchanging the order of summation we get

$$\mathrm{P}(M_n\geqslant r)=\frac{n!}{n^n}\sum_{l=r}^{n}\frac{(l-1)!n^{n-l}}{(r-1)!(l-r)!(n-l)!}\sum_{j=r}^{l}(-1)^{j-r}\binom{l-r}{j-r}.$$

The sum over $j$ is equal to zero if $l\neq r$ and one if $l=r$. Thus

$$\mathrm{P}(M_n\geqslant r)=\frac{n!}{(n-r)!}\frac{1}{n^r}=\prod_{j=0}^{r-1}\left(1-\frac{j}{n}\right).\quad\square$$

**Proposition 2.4.** *The expected value and the variance of the number of matches* $M_n$ *is given by*

$$\mathrm{E}(M_n)=\mu_n=\frac{n!}{n^n}\sum_{r=0}^{n-1}\frac{n^r}{r!},\qquad \mathrm{Var}(M_n)=\mu_n(1-\mu_n)+\frac{2(n!)}{n^n}\sum_{r=0}^{n-2}(n-r-1)\frac{n^r}{r!}. \tag{6}$$

**Proof.** Using the fact that $M_n = \Sigma_{k=1}^n I(T_k = k)$ and the conditional representation (1), we have

$$\mathbf{E}\,M_n = \sum_{k=1}^{n} \mathbf{E}\,I(T_k = k) = \sum_{k=1}^{n} \mathbf{E}\,I(V_k = k \mid V_n = n)$$

$$= \sum_{k=1}^{n} \mathbf{P}(V_k = k,\ V_n - V_k = n-k)/\mathbf{P}(V_n = n) = \frac{n!}{n^n}\sum_{k=1}^{n}\frac{k^k(n-k)^{n-k}}{k!(n-k)!} = \frac{n!}{n^n}\sum_{r=0}^{n-1}\frac{n^r}{r!},$$

where the last step follows from identity (4) with $k = 1$. Similarly

$$\operatorname{Var}(M_n) = \mathbf{E}(M_n^2) - [\mathbf{E}(M_n)]^2 = \mu_n(1-\mu_n) + 2\sum_{1 \leqslant k < l \leqslant n}\frac{\mathbf{P}(V_k = k,\ V_l = l,\ V_n = n)}{\mathbf{P}(V_n = n)}$$

$$= \mu_n(1-\mu_n) + 2\sum_{1 \leqslant k < l \leqslant n}\frac{n!k^k(l-k)^{l-k}(n-l)^{n-l}}{n^n k!(l-k)!(n-l)!}$$

$$= \mu_n(1-\mu_n) + 2\frac{n!}{n^n}\sum_{r=0}^{n-2}(n-r-1)\frac{n^r}{r!},$$

again from identity (4). □

## 3. Matches in the two-sample problem

Let $F$ and $G$ denote the two c.d.f.'s, which, for our purposes, can be assumed without loss of generality, to have support on [0, 1] and that one of them, say $F$, corresponds to the uniform distribution. Let $X_1,\dots,X_n$ and $Y_1,\dots,Y_n$ be the *order statistics* (again with the second subscript $n$ dropped for notational convenience) based on independent random samples of size $n$ each from $F$ and $G$ respectively. To test the null hypothesis $\mathsf{H}_0: F = G$, one may use the following match statistic. We say a match occurs at the $i$th place iff $X_{i-1} < Y_i \leqslant X_i$ for $i = 1,\dots,n$ (with the notation $X_0 = 0$).

Let $M_n^*$ be the total number of matches. Observe that $M_n^*$ is also the number of $i$ for which $G_n(Y_i - 0) < F_n(Y_i) \leqslant G_n(Y_i)$ where $F_n$ and $G_n$ are the empirical distribution functions for $X$'s and $Y$'s respectively. A slightly more general version of $M_n^*$ is discussed in Takács (1971). Our aim is to briefly indicate that the exact distribution of this two-sample statistic $M_n^*$ can also be derived on lines exactly similar to those of Section 2.

Let $S_i$ be the number of $Y_j$'s in $(X_{i-1}, X_i]$, $i = 1,\dots,n$, with $S_{n+1} = (n - \Sigma_{i=1}^n S_i)$. These numbers are called the 'spacing-frequencies' and some general asymptotic theory for statistics based on these has been studied in Holst and Rao (1980). By arguments similar to those in Lemma 2.1 it is easy to verify that $M_n^*$ is equal to the number of $j$ for which $T_j^* = j$ where $T_j^* = \Sigma_{i=1}^j S_i$. Now let $\eta$ denote a geometric ($p$) random variable with probability function $\mathbf{P}(\eta = k) = pq^k$, $k = 0, 1,\dots$ and let $\eta_1, \eta_2,\dots$ be a sequence of such i.i.d. geometric variables. Let $V_j^* = \Sigma_{i=1}^j \eta_i$ which then has a negative binomial distribution denoted by $\mathsf{NB}(j, p)$ with probability function

$$\mathbf{P}(V_j^* = k) = \binom{j+k-1}{k}p^j q^k, \quad k = 0, 1,\dots.$$

The following result gives an analogue to the conditional representation (1) for the two-sample problem.

**Lemma 3.1.** *Under the null hypothesis* $\mathsf{H}_0: F = G$, $(S_1,\dots,S_{n+1})$ *has the same distribution as the conditional distribution of* $(\eta_1,\dots,\eta_{n+1})$ *given* $\Sigma_{i=1}^n \eta_i = n$. *Symbolically*

$$(S_1,\dots,S_{n+1}) \sqcap \left(\eta_1,\dots,\eta_{n+1} \,\middle|\, \sum_{1}^{n+1}\eta_i = n\right). \tag{7}$$

**Proof.** A simple direct argument yields the proof. See also Feller (1968, p. 43) for a discussion on Bose–Einstein statistics. Under the hypothesis, the first sample can be considered as obtained by simple random sampling without replacement from the combined sample. Hence

$$\mathbf{P}(S_1 = s_1, \ldots, S_{n+1} = s_{n+1}) = 1\Big/\binom{2n}{n}.$$

On the other hand,

$$\mathbf{P}\left(\eta_1 = s_1, \ldots, \eta_{n+1} = s_{n+1} \,\middle|\, \sum_1^{n+1} \eta_i = n\right) = \frac{(pq^{s_1}) \cdots (pq^{s_{n+1}})}{\binom{n+1+n-1}{n} p^{n+1} q^n} = 1\Big/\binom{2n}{n}. \quad \Box$$

Using representation (7), one can evaluate the distribution of $M_n^*$ in exactly the same way as that of $M_n$ in Section 2. We will only give a brief outline of this derivation. For an alternate derivation using lattice-path counting, see Takács (1971). Analogous to (3), for $k = 0, 1, \ldots, n$,

$$\begin{aligned}
S_{n,k} &= \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant n} \mathbf{P}(T_{i_1}^* = i_1, T_{i_2}^* = i_2, \ldots, T_{i_k}^* = i_k) \\
&= \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant n} \mathbf{P}(V_{i_1}^* = i_1, V_{i_2}^* = i_2, \ldots, V_{i_k}^* = i_k | V_{n+1}^* = n) \\
&= \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant n} \mathbf{P}(V^* = i_1, V_{i_2}^* - V_{i_1}^* = i_2 - i_1, \ldots, V_{i_k}^* - V_{i_{k-1}}^* = i_k - i_{k-1}, \\
&\qquad V_{n+1}^* - V_{i_k}^* = n - i_k)\, \mathbf{P}(V_{n+1}^* = n)^{-1} \\
&= \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant n} \frac{\binom{2i_1 - 1}{i_1}\binom{2(i_2 - i_1) - 1}{i_2 - i_1} \cdots \binom{2(i_k - i_{k-1}) - 1}{i_k - i_{k-1}}\binom{2(n - i_k)}{n - i_k}}{\binom{2n}{n}} \\
&= \frac{1}{2^k\binom{2n}{n}} \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant n} \binom{2i_1}{i_1}\binom{2(i_2 - i_1)}{i_2 - i_1} \cdots \binom{2(i_k - i_{k-1})}{i_k - i_{k-1}}\binom{2(n - i_k)}{n - i_k} \\
&= \frac{1}{2^k\binom{2n}{n}} \sum_{\substack{l_i \geqslant 1, i = 1,2,\ldots,k;\, l_{k+1} \geqslant 0 \\ \sum_1^{k+1} l_i = n}} \binom{2l_1}{l_1}\binom{2l_2}{l_2} \cdots \binom{2l_k}{l_k}\binom{2l_{k+1}}{l_{k+1}}
\end{aligned} \tag{8}$$

where $l_1 = i_1$, $l_2 = i_2 - i_1, \ldots, l_k = i_k - i_{k-1}$ and $l_{k+1} = n - i_k$. Using the generating function (cf. Riordan (1968, p. 130))

$$(1 - 4x)^{-1/2} = \sum_{k=0}^{\infty} \binom{2k}{k} x^k,$$

we obtain the identity

$$\sum_{\substack{l_i \geqslant 1, i=1,\ldots,k; l_{k+1}\geqslant 0 \\ \sum_1^{k+1} l_i = n}} \binom{2l_1}{l_1} \cdots \binom{2l_{k+1}}{l_{k+1}} =$$

$$= \frac{2^{2n}}{n!}\left[\sum_{r=1}^{k} (-1)^{k-r+1}\left(\frac{r}{k-r+1}\right)\binom{k}{r}\left(\tfrac{1}{2}r\right)^{[n]} + \left(\tfrac{1}{2}(k+1)\right)^{[n]}\right] \tag{9}$$

where we use the ascending factorial notation, $a^{[b]} = a(a+1)\cdots(a+b-1)$ if $b>0$ and $=1$ if $b=0$.

Using (9) the expression in (8) for $S_{n,k}$ reduces to

$$S_{n,k} = \frac{2^{2n-k}}{\binom{2n}{n}}\left[\sum_{r=1}^{k}(-1)^{k-r+1}\left(\frac{r}{k-r+1}\right)\binom{k}{r}\frac{\left(\frac{1}{2}r\right)^{[n]}}{n!} + \frac{\left(\frac{1}{2}(k+1)\right)^{[n]}}{n!}\right].$$

Now applying to the inclusion–exclusion formula as in Theorem 2.3 and with considerable simplifications we obtain

$$\mathbf{P}(M_n^* \geqslant k) = \binom{2n}{n+k}\Big/\binom{2n}{n}, \quad k = 1,\ldots,n. \tag{10}$$

This result is contained in Takâcs (1971, equation (6), p. 1158).

**Proposition 3.2.** *Under the hypothesis* $\mathsf{H}_0$: $F = G$,

$$\mathbf{E}(M_n^*) = \mu_n^* = 2^{2n-1}\Big/\binom{2n}{n} - \tfrac{1}{2} \tag{11}$$

*and*

$$\operatorname{Var}(M_n^*) = n + \tfrac{1}{4} - 2^{4n-2}\Big/\binom{2n}{n}^2.$$

**Proof.** Using the conditional representation (7) and the fact that $M_n^* = \sum_{i=1}^n I(T_i^* = i)$, we have

$$\mu_n^* = \sum_{i=1}^{n} \mathbf{P}(V_i^* = i \mid V_{n+1}^* = n) = \sum_{i=1}^{n} \mathbf{P}(V_i^* = i,\ V_{n+1}^* - V_i^* = n-i)/\mathbf{P}(V_{n+1}^* = n)$$

$$= \sum_{i=1}^{n}\binom{2i}{i}\binom{2(n-i)}{n-i}\Big/2\binom{2n}{n}.$$

The expression for $\mu_n^*$ in (11) now follows from identity (9) where we take $k = 1$. Similarly

$$\operatorname{Var}(M_n^*) = \mu_n^*(1-\mu_n^*) + 2\sum_{1\leqslant k<l\leqslant n} \mathbf{E}\, I(T_k^* = k,\ T_l^* = l)$$

$$= \mu_n^*(1-\mu_n^*) + 2\sum_{1\leqslant k<l\leqslant n} \mathbf{P}(V_k^* = k,\ V_l^* = l,\ V_{n+1}^* = n)/\mathbf{P}(V_{n+1}^* = n)$$

$$= \mu_n^*(1-\mu_n^*) + \sum_{1\leqslant k<l\leqslant n}\binom{2k}{k}\binom{2(l-k)}{l-k}\binom{2(n-l)}{n-l}\Big/2\binom{2n}{n}.$$

The second term in the above expression can be simplified again using identity (9) with $k = 2$. It can be further reduced to the form in (11) using this and the expression for $\mu_n^*$. □

## References

Feller, W. (1968), *An Introduction to Probability Theory and its Applications* Vol. 1 (Wiley, New York, 3rd ed.).

Holst, L. and J.S. Rao (1980), Asymptotic theory for some families of two-sample nonparametric statistics, *Sankhyā Ser. A* **42**, 19–52.

Khidr, A.M. (1981), Matching of order statistics with intervals, *Indian J. Pure Appl. Math.* **12**, 1402–1407.

Riordan, J. (1968), *Combinatorial Identities* (Wiley, New York).

Takács, L. (1971), On the comparison of two empirical distribution functions, *Ann. Màth. Statist.* **42**, 1157–1166.