Statistics & Probability Letters 2 (1984) 353–355
North-Holland December 1984

ON THE LIMITING DISTRIBUTION OF THE NUMBER OF 'NEAR-MATCHES'

S. Rao JAMMALAMADAKA
University of California, Santa Barbara, CA 93106, USA

Svante JANSON
Uppsala University, Uppsala, Sweden

Received March 1984
Revised July 1984

Abstract: When $(R_1,\dots,R_n)$ is a random permutation of the numbers $(1,\dots,n)$, a 'near-match' at the ith place is defined to have occured if $|R_i - i| < k$, for some fixed integer k. This note studies the asymptotic distribution of the number of 'near-matches' when k is fixed and when k is allowed to go to infinity with n.

Introduction

Let $(R_1,\dots,R_n)$ denote a random permutation of the natural numbers $(1, 2,\dots,n)$ so that all the $n!$ possible permutations of $(1,\dots,n)$ are equally likely. For any fixed nonnegative integer k we say that a 'near-match' has occured at the ith place if $|R_i - i| \leqslant$ k. Let $M_n = M_n(k)$ denote the number of near-matches. The case $k = 0$ corresponds to the classical matching problem (cf. Feller [2]). For $n \to \infty$, this note shows that if k is fixed and finite, M_n has a Poisson limit with mean $(2k + 1)$ whereas if $k \to \infty$, M_n has a normal distribution.

In many practical problems, the number of near-matches may be of more interest than perfect-matches i.e. with $k = 0$. For instance, if (i, R_i) denotes the ranks, say given by two judges to the ith contestant, $i = 1,\dots,n$, the number M_n serves as a measure of consistency (or association) of these two judges. One may also consider the measure

$$\eta_n = \left[M_n - (n - M_n)\right]/n = 2\left(\frac{M_n}{n}\right) - 1$$

which lies between -1 and $+1$, with values near $+1$ indicating a larger measure of agreement between the two judges. The measure η_n is similar to the definition of Kendall's τ, which is based on the difference in the number of 'concordances' and the number of 'discordances'. In this sense it is a competitor to the Spearman's rank correlation and the Kendall's τ [3] and the relative performance of these measures will be studied elsewhere.

Poisson limit for finite k

For k finite, the exact distribution of the number of near matches is a combinatorial problem which may be treated by the inclusion–exclusion argument (cf. Feller [2]), much as the classical matching problem. Our proof of the Poisson limit uses the probability generating function, say $P_n(t)$ of M_n, i.e.

$$P_n(t) = \sum_{m=0}^{\infty} t^m P(M_n = m).$$

Define A_i as the event 'near-match at the ith place', $i = 1,\dots,n$. For the sake of convenience, we define matches circularly, i.e., integers modulo n. This clearly makes no difference in the asymptotics since k is finite. Consider (here $I(\cdot)$ denotes the indicator function)

$$\begin{aligned}
&P_n(1+t)\\
&= \sum_{m=0}^{\infty} (1+t)^m P(\text{exactly } m \text{ near matches})\\
&= E\left[\sum_{l\leqslant m}\sum \binom{m}{l} t^l \sum_{i_1,\ldots,i_m} I(A_{i_1},\ldots,A_{i_m} \text{ and no others})\right]\\
&= E\left[\sum_{l=0} t^l \sum_{i_1,\ldots,i_l} I(A_{i_1}\cap\cdots\cap A_{i_l})\right]\\
&= \sum_{l=0} t^l \sum_{i_1,\ldots,i_l} P(A_{i_1}\cap\cdots\cap A_{i_l})\\
&= \sum_{l=0}^{\infty} t^l\cdot\frac{a_l}{l!}, \quad \text{say.}
\end{aligned}$$

It is easy to see, by inclusion–exclusion, that

$$a_l \leqslant (2k+1)^l$$

and also

$$\begin{aligned}
a_l &\geqslant (2k+1)^l - l(l-1)(2k+1)^{l-1}\cdot\frac{2k+1}{n}\\
&= (2k+1)^l\left[1-\frac{l(l-1)}{n}\right].
\end{aligned}$$

Thus

$$\begin{aligned}
&\left|P_n(1+t)-e^{(2k+1)t}\right|\\
&\leqslant \sum_{l=0}^{\infty}\frac{|t|^l}{l!}\left|(2k+1)^l - a_l\right|\\
&\leqslant \sum_{l=0}^{n}\frac{|t|^l}{l!}(2k+1)^l\frac{l(l-1)}{n} + \sum_{n+1}^{\infty}\frac{|t|^l(2k+1)^l}{l!}\\
&\leqslant \frac{1}{n}\sum_{l=0}^{\infty}\frac{|t|^l(2k+1)^l}{(l-2)!}\\
&= e^{(2k+1)t}\cdot\frac{(2k+1)^2|t|^2}{n}
\end{aligned}$$

which converges to zero as $n\to\infty$. Thus the probability generating function of M_n is

$$P_n(t)\to e^{(2k+1)(t-1)}$$

which proves that

$$M_n \xrightarrow{\text{d}} \text{Po}(2k+1).$$

Remark. When $k=0$, $M_n(0)$ is the number of (perfect) matches and has a Po(1) limit. This implies the classical result that P(at least one match) tends to $(1-e^{-1})$ for large n.

Normal limit for infinite k

The near-match statistic M_n in this case can be expressed as

$$M_n = \sum_{i=1}^{n} I(|R_i - i|\leqslant k) = \sum a_n(i, R_i)$$

where

$$a_n(i,j) = \begin{cases} 1 & \text{if } d\leqslant k \text{ where } d=(i-j)\bmod n\\ 0 & \text{otherwise,}\end{cases}$$

is defined circularly for convenience. Therefore combinatorial central limit theorems of the Wald–Wolfowitz–Noether type (see Hajek–Sidak [3]) can be employed to study the asymptotic normality of M_n. The following result due to Motoo [4] is useful. See also von Bahr [1] and Hajek and Sidak [3].

Theorem (Motoo, 1957). *Let $\boldsymbol{R}=(R_1,\ldots,R_n)$ be a random vector which takes every permutation of $(1,\ldots,n)$ with equal probabilities $1/n!$. Let $S_n=\sum_{i=1}^{n} a(i,R_i)$ and define*

$$\bar a(i,\cdot) = n^{-1}\sum_{j=1}^{n} a(i,j),$$

$$\bar a(\cdot,j) = n^{-1}\sum_i a(i,j),$$

$$\bar a(\cdot,\cdot) = n^{-2}\sum_i\sum_j a(i,j)$$

and

$$b(i,j) = a(i,j) - \bar a(i,\cdot) - \bar a(\cdot,j) + \bar a(\cdot,\cdot).$$

Then

$$ES_n = n\cdot\bar a(\cdot,\cdot)$$

and

$$\text{Var}(S_n) = \frac{1}{(n-1)}\sum_i\sum_j b^2(i,j).$$

Let $c(i,j)=b(i,j)/\sqrt{\operatorname{Var}(s_n)}$. *Then*

$$\frac{S_n-ES_n}{\sqrt{\operatorname{Var}(S_n)}}\xrightarrow{\mathrm{d}}N(0,1)$$

if $\lim_{n\to\infty}\frac{1}{n}\sum_{|c(i,j)|>\tau}c^2(i,j)\to 0$

for any $\tau>0$. □

The asymptotic normality of M_n can be established as a consequence of this theorem. From the definition above of $a(i,j)$, using again the circular interpretation, it is easy to check

$$\bar{a}(i,\cdot)=\bar{a}(\cdot,j)=\bar{a}(\cdot,\cdot)=\frac{(2k+1)}{n}=p_{k,n},\quad\text{say.}$$

Thus

$$b(i,j)=[a(i,j)-p_{k,n}]$$

and

$$E(M_n)=\mu_n=n\cdot\bar{a}(\cdot,\cdot)=(2k+1),$$

$$\operatorname{Var}(M_n)=\sigma_n^2=\frac{1}{(n-1)}\sum_i\sum_j b^2(i,j)$$

$$=\frac{n^2}{(n-1)}\cdot p_{k,n}\cdot q_{k,n}$$

where $q_{k,n}=(1-p_{k,n})$. Hence

$$c(i,j)\leqslant\frac{b(i,j)}{(np_{k,n}q_{k,n})^{1/2}}$$

and the sufficient condition for asymptotic normality reduces to

$$\frac{\Sigma\Sigma_{|b(i,j)|>\tau\sqrt{np_{k,n}q_{k,n}}}b^2(i,j)}{\Sigma_i\Sigma_j b^2(i,j)}\to 0\quad\text{for every }\tau>0.$$

Clearly this happens whenever k and $(n-k)$ go to infinity, since the numerator eventually becomes zero. Thus we have

Theorem. *As* k *and* $n-k$ *approach infinity, with* $k<(n/2)$,

$$\frac{M_n-(2k+1)}{\sqrt{np_{k,n}q_{k,n}}}\xrightarrow{\mathrm{d}}N(0,1).\quad\square$$

Remark 1. The denominator on the LHS can be replaced by $\sqrt{(2k+1)}$ if $k/n\to 0$. Compare this with the case of Poisson limit.

Remark 2. The results for the circular and linear cases are identical as long as $k=o(n^{2/3})$. This is because the reduction in $\mu_n=(2k+1)$ for the linear case is

$$2\cdot\left(\frac{1}{n}+\frac{2}{n}+\cdots+\frac{k-1}{n}\right)=\frac{k(k-1)}{n}$$

and hence

$$\frac{E|M_n(\text{circular})-M_n(\text{linear})|}{\sigma_n}=\frac{k(k-1)}{n\sigma_n}\to 0$$

for $k=o(n^{2/3})$. For larger k, one can compute the mean and variance of M_n and establish its asymptotic normality using the same combinatorial limit theorem. But we omit the details.

Remark 3. One can also derive the normal and Poisson limits, exactly on similar lines, for the 'one-sided near matches', i.e. if a match at the ith place is defined whenever $i\leqslant R_i\leqslant i+k$. In the circular case, this obviously gives the same distribution as two-sided matches with k replaced by $k/2$.

References

von Bahr, B. (1976), Remainder term estimate in a combinatorial limit theorem, *Z. Wahrsch. Verw. Gebiete* **35**, 131–139.

Feller, W. (1968), An introduction to probability theory and its applications, Vol. 1 (John Wiley, New York, London, Sydney).

Hajek, J. and Z. Sidak (1967), *Theory of Rank Tests* (Academic Press, New York).

Motoo, M. (1957), On the Hoeffding's combinatorial central limit theorem, *Ann. Inst. Statist. Math. Tokyo* **8**, 145–154.