The Annals of Probability
1973, Vol. 1, No. 1, 188–190

PROBABILITIES FOR A kth NEAREST NEIGHBOR PROBLEM ON THE LINE

BY SYLVAN R. WALLENSTEIN[1] AND JOSEPH I. NAUS

Sandoz Pharmaceuticals and Rutgers-The State University of New Jersey

Given N points distributed at random on $[0, 1)$, let p_n be the size of the smallest interval that contains n points. Previous work finds $\Pr(p_n \leqq p)$, for $n > N/2$, and for $n \leqq N/2$, $p = 1/L$, L an integer. This paper finds the distribution of p_n, for all n, N, and p.

1. Introduction. Let $x_1 \leqq x_2 \leqq \cdots \leqq x_N$, be the ordered values of N independent random variables from the uniform distribution on $[0, 1)$. For any $2 \leqq n \leqq N$, let

$$p_n = \min_{1 \leq i \leq N-n+1} \{x_{n+i-1} - x_i\} .$$

The statistic p_n measures the length of the smallest interval that contains n points. Let $P(n; N, p)$ denote the distribution function of p_n.

Let n_p be the largest number of points clustered within an interval of length p. Rothman [7], [8] has shown that rejection of the hypothesis of randomness for large values of n_p is a Uniformly Most Powerful test against alternatives suggestive of clustering. Newell [6], Naus [5], and Ederer, Myers, and Mantel [2] describe various applications of the statistic n_p. Newell [6] relates the distributions of the statistics n_p and p_n;

$$P(n; N, p) = \Pr(p_n \leqq p) = \Pr(n_p \geqq n)$$

and derives asymptotic expressions for a generalization of this probability.

The distribution of p_2, the smallest gap, and of p_N, the sample range are well known. Naus [3] derives explicit expressions for $P(n; N, p)$ for $p \geqq \frac{1}{2}$, and for $p < \frac{1}{2}$, $n > N/2$; Naus [4] derives a formula for $P(n; N, 1/L)$, L an integer. The next section derives a general formula for $P(n; N, p)$ for all n, N, and rational p.

2. A general formula for $P(n; N, r/L)$. We view the unit line divided into L disjoint intervals (cells) each of length $1/L$. Denote the cell occupancy numbers as $n_1, n_2, \cdots, n_L$. Let

$$J(a, b) = \sum_{i=a}^{b} n_i ,$$

and let

$$V_L(N, r) = \{(n_1, \cdots, n_L) \mid n_i \geqq 0, i = 1, \cdots, L; J(1, L) = N, \text{ and } J(i, i+r-1) < n, \text{ for } i \leqq L - r + 1\} . \tag{2.1}$$

Received March 6, 1972; revised May 5, 1972.

[1] This article is based in part on the author's Ph. D. dissertation.

AMS 1970 *subject classifications.* Primary 60; Secondary E05.

Key words and phrases. Coincidences, clusters, nearest neighbor distances, maximum clusters, smallest intervals.

THEOREM. *Given r and L are positive integers with greatest common denominator of one, $0 < r/L < 1$, and given n and N are integers, $2 \leqq n \leqq N$, then*

$$\Pr(n_p \geqq n) = P(n; N, r/L) = 1 - N!\, L^{-N} \textstyle\sum_{V_L(N,r)} \prod_{k=1}^{r} \det D^k , \tag{2.2}$$

where D^k is a square matrix with elements,

$$\begin{aligned} D^k_{a,b} &= 1/[(b-a)n - J(k+1+(a-1)r, k-1+(b-1)r]!\,, && \text{for } a < b\,, \\ &= 1/[(b-a)n + J(k+(b-1)r, k+(a-1)r)]!\,, && \text{for } a \geqq b\,, \end{aligned}$$

subject to the convention that $1/x! = 0$, for $x < 0$. The dimension of the matrix D^k is $e_k + 1$, where

$$\begin{aligned} e_k &= [L/r] - 1\,, && \text{if } kr > L - r\,, \\ &= [L/r]\,, && \text{if } kr \leqq L - r\,, \end{aligned}$$

where $[x]$ denotes the greatest integer in x.

PROOF. Let y_p be the number of points in $[y, y+p)$, and let,

$$A = \{\max_a J(a, a+r-1) \geqq n, a \leqq L - r + 1\}$$

and

$$B_i = A^c \cap \{\sup_y y_p \geqq n, (i-1)/L \leqq y < i/L\}\,.$$

For general r and L, there are $(L-r)B_i$'s to consider, and we group these into r sets, $c(1), \cdots, c(r)$, where

$$c(k) = \{i \mid i = k(\text{mod } r), i \leqq L - r\}\,, \tag{2.3}$$

and let e_k be the number of integers in $c(k)$. Let E_k denote the event

$$E_k = \textstyle\bigcap_{i \in c(k)} B_i^c\,.$$

Then, for $\{n_i\}$ in A^c,

$$\Pr(n_p < n \mid \{n_i\}) = \Pr(\textstyle\bigcap_{i=1}^{L-r} B_i^c \mid \{n_i\}) = \Pr(\bigcap_k E_k \mid \{n_i\})\,. \tag{2.4}$$

Conditional on $\{n_i\}$, the events E_k are mutually independent. To see this, let $y_i(t)$ denote the number of points in the interval $[(i-1)/L, (i-1)/L + t)$. Conditional on $\{n_i\}$, the points in cell i are distributed uniformly over that interval, and independent of points in cell j. The quantities $y_i(t)$ and $y_j(t)$, for $i \neq j$, are conditionally independent. The event E_k is equivalent to the event $q_p(k) < n$, where,

$$\begin{aligned} q_p(k) &= \max\{y_p, (i-1)/L \leqq y < i/L, i \in c(k)\} \\ &= \max_{i \in c(k)} \sup_{t \leq 1/L} [J(k + ir - r, k + ir - 1) \\ &\qquad + y_{k+ir}(t) - y_{k+(i-1)r}(t)]\,. \end{aligned} \tag{2.5}$$

Since the statistics $q_p(i)$, $q_p(j)$ depend on disjoint sets of $y_i(t)$'s, the events E_i, E_j are conditionally independent. Thus,

$$\Pr(n_p < n \mid \{n_i\}) = \textstyle\prod_{k=1}^{r} \Pr(E_k \mid \{n_i\})\,. \tag{2.6}$$

Naus [4] notes that for the case $r = 1$, $\Pr(\bigcap_{i=1}^{L-1} B_i^c \mid \{n_i\})$ can be interpreted as an L-candidate ballot probability, and applies a result of Barton and Mallows ([1] page 243) to find this probability. The same result applies here to $\Pr(E_k \mid \{n_i\})$:

$$\Pr(E_k \mid \{n_i\}) = \det D^k \prod_{i=1}^{e_k+1} (n_{k+(i-1)r}!) \,. \tag{2.7}$$

To complete the proof, note that the joint distribution of the n_i is multinomial, and that,

$$\Pr(n_p \geqq n) = 1 - N!\, L^{-N} \sum_{V_L(N,r)} \Pr(n_p < n \mid \{n_i\})/n_1! \cdots n_L! \,. \tag{2.8}$$

Substitute the right-hand side of (2.7) into (2.6), and the resulting expression for $\Pr(n_p < n \mid \{n_i\})$ into (2.8) to find (2.2).

REFERENCES

[1] BARTON, D. E. and MALLOWS, C. L. (1965). Some aspects of the random sequence. *Ann. Math. Statist.* **36** 236–260.

[2] EDERER, F., MYERS, M. H. and MANTEL, N. (1964). A statistical problem in space and time: Do Leukemia cases come in clusters? *Biometrics* **20** 626–636.

[3] NAUS, J. I. (1965). The distribution of the size of the maximum cluster of points on a line. *J. Amer. Statist. Assoc.* **60** 532–538.

[4] NAUS, J. I. (1966a). Some probabilities, expectations, and variances for the size of largest clusters and smallest intervals. *J. Amer. Statist. Assoc.* **61** 1191–1199.

[5] NAUS, J. I. (1966b). A power comparison of two tests of non-random clustering. *Technometrics* **8** 493–517.

[6] NEWELL, G. F. (1963). Distribution for the smallest distance between any pair of kth nearest neighbor random points on a line. *Proceedings of Symposium on Time Series Analysis*, Wiley, New York, 89–103.

[7] ROTHMAN, E. (1967). Tests for clusters in a Poisson process. *Ann. Math. Statist.* **38** 967.

[8] ROTHMAN, E. (1969). Tests for uniformity against regularly spaced alternatives. Technical Report Number 119, John Hopkins University.

[9] WALLENSTEIN, S. R. (1971). Coincidence probabilities used in nearest neighbor problems on the line and circle. Ph. D. Dissertation, Rutgers University.

8 BRIAN ROAD
EAST BRUNSWICK, NEW JERSEY 08816