1. Introduction

We are happy to contribute this paper to the festschrift volume in honor of Dr. H. Rubin. The paper is of practical and theoretical application. I also had the pleasure to develop with Herman an extended analysis concerning a family of distributions in possession of a monotone likelihood ratio (1, 2).

Question about spacings of a marker array and general issues of sequence heterogeneity in our studies of DNA and protein sequences led us to statistical considerations of \( r \)-scan lengths, the distances between marker \( i \) and marker \( i + r \), \( i = 1, 2, 3, \ldots \). It is interesting to characterize the \( r \)-scan lengths harboring clusters or indicating regions of over-dispersion of the markers along the sequence. Applications are reviewed for certain words in the \textit{Haemophilus} genome and the \textit{Cyanobacter} genome.

2. Preliminaries. Minimal \( r \)-scan lengths from a general distribution

In the paper \( 3 \), the asymptotic distribution of the \( k \)th minimum \( r \)-scan length from a general distribution function has been studied by applying the Chen-Stein

---

1Department of Mathematics, Stanford University, Stanford, CA 94305-2125, USA. e-mail: karlin@math.stanford.edu

Keywords and phrases: \( r \)-scan statistics, inhomogeneous Poisson marker array, asymptotic distributions.

AMS 2000 subject classifications: 92B05, 92D20.
method [4]. In that context, an \(r\)-scan process is generated following a piecewise constant function or continuous general density \(f(x)\) with bounded support \((0,T]\).

Thus let \(V_1, V_2, \ldots, V_{n-1}\) be \(n - 1\) i.i.d. samples drawn from the density \(f(x)\), and let \(V_1^* \leq V_2^* \leq \cdots \leq V_{n-1}^*\) be the order statistics corresponding to \(\{V_i\}\). For convenience, let \(V_0^* = 0\) and \(V_n^* = T\). Then the associated \(r\)-scan fragments \(R_i = V_{i+r-1}^* - V_{i-1}^*, i = 1, \ldots, n - r + 1\), and their order statistics \(R_1^* \leq R_2^* \leq \cdots \leq R_{n-r+1}^*\) are defined in the usual way such that \(R_1^* \leq R_2^* \leq \cdots \leq R_{n-r+1}^*\).

For convenience, let \(V_0^* = 0\) and \(V_n^* = T\). Then the associated \(r\)-scan fragments \(R_i = V_{i+r-1}^* - V_{i-1}^*, i = 1, \ldots, n - r + 1\), and their order statistics \(R_1^* \leq R_2^* \leq \cdots \leq R_{n-r+1}^*\) are defined in the usual way such that \(R_1^* \leq R_2^* \leq \cdots \leq R_{n-r+1}^*\).

For an extensive review of \(r\)-scan statistics, see the book [5].

From \(\{R_i\}\), we define the Bernoulli random variables

\[
U_i^-(a) = \begin{cases} 1, & \text{if } R_i \leq a \\ 0, & \text{if } R_i > a \end{cases}
\]

and their sum

\[
N_{n-r+1}^-(a) = \sum_{i=1}^{n-r+1} U_i^-(a).
\]

Denote by \(m_{n,k} = R_k^*\). The asymptotic distribution (as \(n \to \infty\)) for \(m_{n,k}\) is as follows.

**Lemma 1.** For a given positive constant \(\mu\), let \(a_n\) be determined to satisfy

\[
\mu = \frac{(na_n)^r}{r!} n \int [f(x)]^{r+1} \, dx.
\]

(1)

Then we have the Poisson approximation

\[
\lim_{n \to \infty} d(N_{n-r+1}^-(a_n), Po(\mu)) = 0,
\]

for \(Po(\mu)\) the Poisson distribution with parameter \(\mu\). Here \(d(\ ,
\ )\) is the total variational distance between two random variables defined by

\[
d(U,V) = \sup_A \left[ \Pr\{U \in A\} - \Pr\{V \in A\} \right].
\]

Moreover, the \(k\)th minimal \(r\)-scan length, \(m_{n,k}\), possesses the asymptotic distribution

\[
\lim_{n \to \infty} \Pr\{m_{n,k} > a_n\} = \sum_{i=0}^{k-1} e^{-\mu} \frac{\mu^i}{i!}.
\]

(2)

Proof of the above lemma is given in [3], Section 8. Here, by adapting the foregoing result, we will determine the asymptotic distribution of the \(k\)th minimal \(r\)-scan length corresponding to an inhomogeneous Poisson process in \((0,t)\), as \(t \to \infty\).

### 3. Minimal \(r\)-scan lengths for an inhomogeneous Poisson process

The asymptotic theorem for the minimal \(r\)-scan length will be derived from the distributional property of \(\tilde{N}_r^-(a)\), where \(\tilde{N}_r^-(a)\) is the number of \(r\)-scan segments of lengths \(\leq a\) over the interval horizon \((0,t)\). It is clear that if \(\tilde{N}_r^-(a) < k\), the \(k\)th minimal \(r\)-scan length \(m_{t,k}\) exceeds the level \(a\). Thus if the Poisson approximation holds for \(\tilde{N}_r^-(a)\), we can access the asymptotic law for \(m_{t,k}\). Here the \(r\)-scan process of interest is generated from an inhomogeneous Poisson process \(\Pi\) with an intensity function \(\lambda(s), 0 \leq s < \infty\). The main theorem is as follows.
Theorem 1. Assume \( \lambda(s) \) defined for \( s \geq 0 \) satisfies

\[
\int_0^t \lambda(s) \, ds \to \infty, \quad \text{as } t \to \infty.
\] (3)

For a given positive constant \( \mu \), let \( a_t \) be determined to satisfy the equation

\[
\mu = \frac{a_t}{r!} \int_0^t \lambda^{r+1}(s) \, ds.
\] (4)

Then we have the Poisson approximation

\[
\lim_{t \to \infty} d(\tilde{N}_t^- (a_t), \text{Po}(\mu)) = 0.
\]

Moreover, the \( k \)th minimal \( r \)-scan length, \( m_{t,k} \), possesses the asymptotic distribution

\[
\lim_{t \to \infty} \Pr \{ m_{t,k} > a_t \} = \sum_{i=0}^{k-1} e^{-\mu_i} \mu_i!.
\]

Proof of Theorem 1. If \( n_t \) denotes the point count of the Poisson Process \( \Pi \) in \((0, t)\), then

\[
E[n_t] = \int_0^t \lambda(s) \, ds, \quad \text{Var}(n_t) = \int_0^t \lambda(s) \, ds.
\]

For convenience, let \( \bar{n}_t = \lfloor E[n_t] \rfloor \). Thus the Berry–Esseen estimate assures

\[
\Pr \left\{ \left| \frac{n_t}{\bar{n}_t} - 1 \right| > \sqrt{\frac{\ln \bar{n}_t}{\bar{n}_t}} \right\} = O \left( \frac{1}{\sqrt{\bar{n}_t}} \right).
\]

Therefore

\[
d(\tilde{N}_t^- (a_t), \text{Po}(\mu)) \leq d(N_{\bar{n}_t-r-1}^- (a_t), \text{Po}(\mu)) + O \left( \sqrt{\frac{\ln \bar{n}_t}{\bar{n}_t}} \right).
\]

If \( n_t = \bar{n}_t \), the \( \bar{n}_t \) points in \((0, t)\) are distributed independently according to \( g(x) \), with

\[
g(x) = \frac{\lambda(x)}{\int_0^t \lambda(x) \, dx}, \quad 0 \leq x \leq t.
\] (5)

Following the result of Lemma 1, we have

\[
\lim_{n \to \infty} d(N_{\bar{n}_t-r-1}^- (a_n), \text{Po}(\mu)) = 0
\]

for

\[
a_n = \sqrt{\frac{r! \mu}{n^{r+1} \int_0^t g^{r+1}(x) \, dx}}.
\] (6)

Since

\[
\bar{n}_t = \int_0^t \lambda(s) \, ds \to \infty,
\]

we replace \( n \) with \( \bar{n}_t \) in formula (6) and \( g(x) \) with \( \frac{\lambda(x)}{\int_0^t \lambda(x) \, dx} \) to verify equation (4). On this basis, we obtain

\[
\lim_{t \to \infty} d(\tilde{N}_t^- (a_t), \text{Po}(\mu)) = 0,
\]
with

\[ a_t = \sqrt{\frac{r! \mu}{\int_0^t \lambda^{r+1}(s) \, ds}}. \]

This completes the proof of Theorem 1.

4. Examples

*Haemophilus influenza* is a bacterium which engenders an infection in the lungs of humans [6]. The study of the USSs (uptake signal sequences) AAGTGCGGT (USS+) and its inverted complement (USS-) in the *H. influenza* genome (length of 1.83 Mb, Rd strain) provides opportunities for characterizing global genomic inhomogeneities. The result of homogeneous \( r \)-scan tests for \( r = 1, 2, \ldots, 6 \) shows a significant even spacings between the markers such that the USS+ and USS- are remarkably evenly spaced around the genome such that both USS+ positions and USS- positions have respective minimum spacings higher than expected by chance with probability 0.001. This rare possibility may suggest that the homogeneity assumption doesn’t fit the real distributions of the markers and an inhomogeneous \( r \)-san test should be applied for the marker array.

Another example is the distribution of the palindrome GGCGATCGCC labeled HIP1 (highly iterated palindrome) in the genome of the organism *Synechocystis* (3.6 Mb). *Synechocystis* is thought to be the evolutionary precursor of vascular plant plastids [7]. The photosynthetic endosymbiont became dependent on host genetic information for maintenance and evolved into an organelle specialized for \( \text{CO}_2 \) fixation. The \( r \)-scan analysis of the genome shows in this case a significantly even distribution. The observed minimal \( 1 \)-scan spacing is 52 bp (base pair) which is much larger than the threshold of 9 bp with the probability of 0.001. Similar conclusions apply to the \( r \)-scan tests of \( r = 2, \ldots, 6 \). The even spacing of HIP1 in *Synechocystis* is more dramatic than the situation of USSs in *H. influenza*.

References


