응용통계연구 제 6권 1호
1993년 3월 pp. 125 - 143

스캔 통계량의 발전 과정과 응용에 대한 고찰

김병수[1], 김기한[2]

요 약

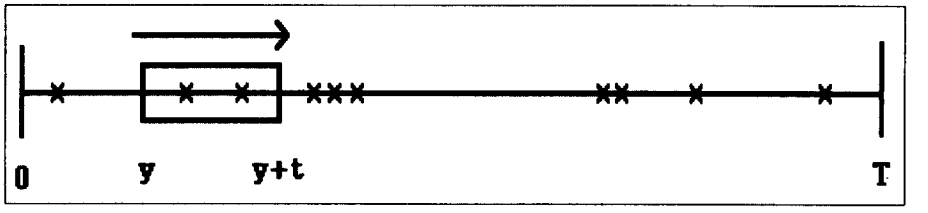
관측치가 (0,T]의 구간에서 균일하게 분포한다는 가설에 대하여, 관측치의 집락화를 검정하는 과정에서 스캔 통계량을 사용할 수 있다. 본 논문에서는 스캔 통계량의 확률분포의 근사분포가 어떠한 이론적 배경으로 개선되어 왔는 지를 고찰하고, 실제로 응용된 예를 살펴보기로 한다.

광물 매장을 조사하기 위한 항공탐사, 두 개의 아미노산 염기서열(amino-acid sequence)을 비교하는 과정에서 스캔 통계량은 사용되어 왔다. 지놈(genome)의 連鎖(sequence)에서 돌연변이가 발생한 위치에 대하여 집락의 가능성을 검색하는 방법으로 스캔 통계량을 이용할 수 있음을 보이고, 이에 대한 구체적인 문제 구성은 추후 연구과제로 제시한다.

1. 서 론

전체구간에서 발생한 관측치의 총수를 N이라 하고, 임의의 부분구간을 정하여, 이 부분구간을 연속적으로 혹은 離散的으로 이동시키면서 전체구간을 탐사(scan)하는 과정을 생각해 보자. 아래 <그림 1>에서처럼 길이 t의 부분구간을 정하고, 이 부분구간을 연속적으로 이동하여 (0,T] 구간을 탐사하면서 그 부분구간내에서 발생하는 관측치의 수를 조사한다고 하자.

<그림 1> 전체구간 (0,T] 위에서 발생한 N개의 관찰치



[1] (120-749) 서울특별시 서대문구 신촌동 134번지, 연세대학교 응용통계학과
[2] (100-014) 서울특별시 중구 충무로4가 126-1 일흥빌딩 6층 누리기획 마케팅 전략연구소

이 경우 길이 t의 부분구간에서 발생하는 관측치의 수는 부분구간의 始点인 y의 위치에 따라 0, 1,…, N 의 가능한 값들을 취할 수 있는데, 이 관측치 수의 최대값을 스캔 통계량이라고 한다. 관측치는 균일하게 분포한다는 가설에 대하여, 관측치의 집락화를 검정하는 통계량으로서 스캔 통계량을 사용할 수 있다.

창연(bismuth-214)이라는 광물에서 나오는 방사능은 우라늄의 매장량과 밀접한 관계가 있다. 항공탐사에 의하여 창연에서 나오는 방사능을 기록함으로써, 다른 지역에 비하여 비정상적으로 많은 방사능이 기록되는 지역을 찾는다. 스캔 통계량을 이용하여 이러한 지역들의 집락화 여부를 검정하고, 집락이 이루어진 지역에 대하여는 우라늄의 채광 여부를 검토한다.

포아송 사건의 발생시점은 (0,T]의 구간에서 균일하게 분포한다는 가정하에서, Naus(1966)에 의해 스캔 통계량의 확률분포가 유도되었다. 그러나 Naus(1966)가 유도한 스캔 통계량의 확률분포는 많은 행렬식을 포함하고 있어 계산하기가 힘들기 때문에, 계산이 편리한 근사분포를 유도하기 위한 연구가 계속되어 왔다. 이러한 문제를 해결하기 위하여 Naus(1982)는 계산이 편리한 근사분포를 유도하였고, 또한 Wallenstein and Neff(1987)와 Glaz(1989)에 의하여 더욱 계산이 쉬운 근사분포가 유도되었다. Glaz(1989)는 Hunter(1976)의 상한값(upper bound)보다 더욱 개선된 상한값을 유도하였고, 이를 만족하는 근사분포를 유도하였다.

Ederer, Myers and Mantel(1964)은 백혈병 발생의 집락 여부를 검정하는 문제에서 Feller(1957)의 점유문제를 이용하여 백혈병이 랜덤하게 발생한다는 가설아래에서 스캔 통계량의 평균과 분산을 구하고, 피어슨의 카이제곱을 이용한 적합도 검정을 하고 있다. 또한, Tango(1984)는 환경적 요인에 의한 질병의 순환적 집락이나 전염병의 경우와 같은 질병의 시간적 집락 현상을 검색하는 절차로서 새로운 지표(index)를 제안하면서, Wallenstein(1980)의 삼염색체성(trisomy) 자료를 이용하여 Tango의 새로운 지표와 스캔 통계량, 카이제곱통계량의 검정력을 비교하고 있다. 이외에도 시간적 집락 현상을 검색하는 여러 절차들에 대한 검정력 비교는 Tango(1983)에서 언급하는 참고문헌에서 찾아볼 수 있다.

본 논문에서는 스캔 통계량의 확률분포의 근사분포가 어떠한 이론적 배경으로 개선되어 왔는 지를 살펴보고, 실제로 응용된 예를 다룬다. 그리고 분자생물학의 분야에서 스캔 통계량을 이용하여 돌연변이 발생 위치의 집락 가능성을 규명할 수 있는 과제를 추후 연구로 제시한다.

2. 스캔 통계량의 정의와 분포

2-1. 스캔 통계량의 정의

$X_1,\cdots,X_N$ 은 전체구간 $(0,T)$ 에서 독립적으로 균일하게 분포하는 관측치라고 가정하고, $X_1\le X_2\le\cdots\le X_N$이라 가정한다. $0\le y\le T-t$ 에 대하여, 길이 t의 부분구간 $(y,\ y+t)$ 를 길이 t의 스캔 윈도우(scan window)라 하고, 길이 t의 스캔 윈도우에 포함되는 관측치의 수를 $N_{y,y+t}$로 정의한다. 길이 t의 스캔 윈도우에 포함되는 관측치의 최대수를 $n(t)$로 표기하고, 다음과 같이 정의한다.

$$n(t)=\mathrm{Max}\{N_{y,y+t}\}\ ,\ 0\le y\le T-t \tag{2.1.1}$$

식 (2.1.1)에서 정의한 $n(t)$를 스캔(scan) 통계량이라 한다. 스캔 통계량은 관측치 X_i $(i=1,2,\cdots,N)$ 가 균일하게 분포한다는 가설에 대하여, 관측치의 집락화를 검정하기 위해서 사용한다.

길이 t의 스캔 윈도우가 N개의 전체관측치 중에서 n개의 관측치를 포함한 집락이 적어도 한 개 이상일 확률을 $P(n,N,t/T)$로 표기하고, 다음과 같이 정의한다.

$$P(n,N,r)=\Pr\{n(t)\ge n\}\ ,\ r=t/T \tag{2.1.2}$$

대부분의 경우 스캔 통계량의 분포는 포아송 과정에 의하여 생성되는 사건의 발생시점이 균일분포한다는 가정하에서 유도되어 왔다. 따라서, 본 논문에서는 포아송 과정의 조건하에서 스캔 통계량의 분포의 유도과정에 대하여 살펴보기로 한다.

본 논문에서 사용되는 표기의 정의는 다음과 같다.

$n(t)$ = 길이 t의 스캔 윈도우에 포함되는 관측치의 최대수
N = $(0,T)$의 구간에서 발생하는 포아송 사건의 총수
L = T/t
r = $1/L$
λ = 포아송 과정의 단위시간당 평균발생율

2-2. 스캔 통계량의 분포의 종류

$(0,T)$의 구간에서 포아송 과정에 의해 N개의 포아송 사건이 발생하면, 이 N개

의 포아송 사건의 발생시점은 독립적이고 동일한 $(0,T)$의 구간상의 균일분포를 이룬다. 이 N개 포아송 사건의 발생시점을 N개의 관측치로 간주한다. 스캔 통계량의 분포의 유도는 관측치의 총수 N이 주어진 경우와 N이 확률변수인 경우로 구분된다.

첫째, 관측치의 총수 N이 주어진 경우, N개의 관측치는 $(0,T)$의 구간에서 독립적으로 균일분포를 이룬다. 이 경우 유도되는 스캔 통계량의 분포를 조건확률에 대한 스캔 통계량의 분포라 하고, $P(n,N,r)$로 표기한다. 둘째, $(0,T)$의 구간에서 발생하는 관측치의 총수 N을 모르는 경우에는, 단위시간당 평균발생률을 λ라고 할 때, (0,T)의 구간에서 발생하는 관측치의 갯수 N은 평균이 λT인 포아송 확률변수가 된다. 이 경우 유도되는 스캔 통계량의 분포를 무조건확률에 대한 스캔 통계량의 분포라 하고, $P^*(n,\lambda T,r)$로 표기한다.

2-2-1. 조건확률에 대한 스캔 통계량의 분포

$X_1,\cdots,X_N$은 전체구간 (0,1)에서 독립적으로 균일하게 분포하는 관측치라고 가정하고, $X_1 \le X_2 \le \cdots \le X_N$이라 가정한다. 표기의 편의상 구간의 길이를 T=1로 한다. 전체구간을 L개의 겹치지 않는 等區間으로 나누고, $L \ge 2$, $2 \le n \le N$ (n, N, L은 정수)에 대하여 Naus(1966)는 다음과 같은 스캔 통계량의 분포를 유도하였다.

$$\Pr\{n(t) < n\} = N! L^{-N} \sum_{s} det|1/c_{ij}!| \qquad (2.2.1)$$

단, S는 $\{(n_1,\cdots,n_L): \sum_{i=1}^{L} n_i = N,\ 0 \le n_i \le N\}$ 에 대하여 합이 이루어지며, n_i는 구간 $(\frac{i-1}{L}, \frac{i}{L})$ 에 포함되는 관측치의 수를 나타낸다.

그리고 $det|\cdot|$ 는 $L \times L$ 행렬식이며,

$$c_{ij} = (j-i)n - \sum_{k=i}^{j-1} n_k + n_i \ , \ i < j,$$
$$= (j-i)n + \sum_{k=j}^{i} n_k \ , \ i \ge j \text{ 이다.}$$

단, $c_{ij} < 0$, $c_{ij} > N$이면, $1/c_{ij}! = 0$ 이다.

식 (2.2.1)은 이론적으로는 모든 n, N, r에 대해 값을 구할 수 있다. Naus(1966)는 $2 \le n \le N$, $2 \le N \le 10$, $0.1 \le r \le 0.9$ 에 대하여 $\Pr\{n(t) \ge n\}$의 값을 계산하였다. 그러나 많은 N과 작은 r에 대하여는 여전히 많은 행렬식을 포함하고 있어, 계산하기가 쉽지 않다.

Neff and Naus(1980)는 행렬식의 수를 줄임으로써, $0<r<½$, $3\leq n<N\leq 20$에 대하여 $P(n,N,r)$의 정확한 값을 계산하였다. $N>20$에 대하여는, 제 3 절에서 살펴보는 근사분포를 이용하여 값을 구하는 것이 유용하다.

2-2-2. 무조건확률에 대한 스캔 통계량의 분포

단위시간당 평균발생률은 λ이고, 관측치의 숫자인 N이 평균 $E(N)=\lambda T$를 갖는 포아송 분포를 이룰 때, 무조건확률에 대한 스캔 통계량의 분포는 Wallenstein and Neff(1987)에 의하여 다음과 같이 유도되었다.

$$P^*(n,\lambda T,r)=\sum_{N=n}^{\infty} P(n,N,r)\times \frac{e^{-\lambda T}(\lambda T)^N}{N!} \qquad (2.2.2)$$

매우 작은 λT, n, 혹은 $r>½$의 경우를 제외하고는, $P^*(n,\lambda T,r)$의 정확한 값을 구하기가 힘들다. Naus and Neff(1980)는 $3\leq n\leq 9$와 $0.1\leq\lambda\leq 10$에 대하여, $P^*(n,2\lambda,½)$와 $P^*(n,3\lambda,⅓)$의 값을 계산하였고, 또한 Naus(1982)는 모든 n과 λ에 대하여, $P^*(n,2\lambda,½)$와 $P^*(n,3\lambda,⅓)$의 값을 계산하였다. 많은 N과 작은 r에 대하여는 복잡한 컴퓨터 계산이 필요하기 때문에, 제 3 절에서 살펴보는 근사분포를 이용하여 값을 구하는 것이 유용하다. 통상적으로 λ는 알려져 있지 않으므로 총발생건수를 관찰시간으로 나누어 줌으로써 추정이 가능하다.

2-2-3. 윈도우의 길이 t의 결정

윈도우의 길이 t 는 스캔 통계량을 적용하는 문제의 성격에 의하여 결정될 수 있다. 예를 들어 어느 지역 병원에서 과거에는 연평균 36건의 암환자가 보고되었는데, 1년간 기록을 조사하던 중 5월 1일 부터 5월 10일 사이에 놀랍게도 5건의 암환자가 발생하였다. 이 경우 $t=10$, $L=365/10\fallingdotseq 36$으로 결정이 될 수 있다. 이외에도 Ederer, Myers and Mantel(1964), Glaz and Naus(1983), Wallenstein(1980)와 Wallenstein and Neff(1987) 등에서 구체적인 예를 다루고 있다.

2-3. 스캔 통계량의 근사분포와 효율성

2-3-1. Naus의 근사분포

제 1 절에서 정의된 $N_{y,y+t}$, $n(t)$, $P(n,N,t/T)$에 대하여, 사상 E_i를 다음과 같이 정의한다.

$$E_i=\{Max\{N_{y,y+t}\}<n\},\ (i-1)t\leq y\leq it,\ i=1,2,\cdots,L-1 \qquad (2.3.1)$$

$\{n(t)<n\}$인 사상을 $E(n,t)$로 정의하면, 다음 식 (2.3.2)가 성립한다.

$$E(n,t)=\bigcap_{i=1}^{L-1} Ei \ , \ L\text{은 2이상인 수이다.} \tag{2.3.2}$$

사상 $E(n,t)$의 확률을 $Q(n,N,1/L)$로 정의하면, 다음 식 (2.3.3)이 성립한다.

$$Q(n,N,1/L)=\Pr\{n(t)<n\}=1-P(n,N,1/L) \tag{2.3.3}$$

식 (2.3.2)와 (2.3.3)에 의하여, Naus(1982)는 다음 식 (2.3.4)를 유도하였다.

$$Q(n,N,1/L)=\Pr\left(\bigcap_{i=1}^{L-1} E_i\right)=\Pr(E_1)\prod_{k=2}^{L-1}\Pr\left(E_k\middle|\bigcap_{j=1}^{k-1} E_j\right) \tag{2.3.4}$$

많은 실제적인 경우에 식 (2.3.4)에서의 $\Pr(E_k|\bigcap_{j=1}^{k-1} E_j)$는 $\Pr(E_k|E_{k-1})$에 의해 근사되고, 또한 경우에 따라서는 대칭성이 성립하여 모든 k에 대해 $\Pr(E_k|E_{k-1})=\Pr(E_2|E_1)=\Pr(E_1\cap E_2)/\Pr(E_1)$이 성립한다. 그러므로 $Q(n,N,1/L)$는 다음과 같이 유도된다.

$$Q(n,N,1/L)=Q_{2t}(Q_{3t}/Q_{2t})^{L-2} \tag{2.3.5}$$

$$\text{단, } Q_{2t}=\Pr(E_1),\ Q_{3t}=\Pr(E_1\cap E_2) \tag{2.3.6}$$

무조건확률에 대하여 식 (2.3.5)는 다음 식 (2.3.7)과 같이 근사된다.

$$Q^*(n,\lambda L,1/L)\fallingdotseq Q^*(n,2\lambda,\tfrac{1}{2})\{Q^*(n,3\lambda,\tfrac{1}{3})/Q^*(n,2\lambda,\tfrac{1}{2})\}^{L-2} \tag{2.3.7}$$

$$\text{단, } Q^*(n,\lambda L,1/L)=1-P^*(n,\lambda L,1/L) \tag{2.3.8}$$

식 (2.3.7)은, Conover, Bement and Iman(1979)에 의해서 유도된 다음의 식 (2.3.9)에 비하여 매우 정확한 근사분포이다.

$$\begin{aligned} Q^*(n,\lambda L,1/L) \fallingdotseq\ & (3/2)P_{k-1}\exp\{-\lambda(L-1)(1-P_{k-1}/P_{k-2})\} \\ & -\tfrac{1}{2}\exp\{-\lambda L(1-P_{k-2})\} \end{aligned} \tag{2.3.9}$$

$$\text{단, } P_k=e^{-\lambda}\sum_{j=0}^{k}\lambda^j/j!$$

n>7인 경우에 식 (2.3.9)는 정확한 근사값을 얻을 수 없지만, 식 (2.3.7)은 식 (2.3.9)보다 더 정확한 근사값을 준다. 또한, 식 (2.3.7)은 L이 정수가 아닐 경우에도 쓰일 수 있다.

2-3-2. Wallenstein and Neff의 근사분포

N이 클 경우에, Naus(1982)의 근사분포는 조건확률일 경우에는 무조건확률일

때보다 정확한 값을 구할 수가 없다. Wallenstein and Neff(1987)는 가능한 큰 N에 대하여, 정확한 근사분포를 유도하였다. 특히, 조건확률일 경우에 정확한 값을 주는 근사분포를 식 (2.3.10)과 같이, 무조건확률일 경우에는 식 (2.3.11)과 같이 유도하였다. 식 (2.3.10), (2.3.11)에서의 Y의 분포는 길이 t의 스캔 윈도우에 포함되는 관측치의 수의 분포와 같다.

(1) 조건확률에 대한 스캔 통계량의 분포

$$P(n,N,r) \fallingdotseq (n/r-N+1)\Pr(Y=n)+2\Pr(Y\geq n+1) \qquad (2.3.10)$$

단, Y는 N과 r을 모수로 하는 이항분포를 이룬다. 즉,

$$\Pr(Y=n)=\binom{N}{n} r^n(1-r)^{N-n}$$

(2) 무조건확률에 대한 스캔 통계량의 분포

$$P^*(n,\lambda T,r) \fallingdotseq \{(n-\lambda t)(1-r)/r+1\}\Pr(Y=n)+2\Pr(Y\geq n+1) \qquad (2.3.11)$$

단, Y는 평균 λT를 갖는 포아송 분포를 이룬다.

식 (2.3.10)은 $P(n,N,r)\leq 0.5$일 경우에 매우 정확한 값을 주지만, $P(n,N,r)$의 큰 값에 대하여는 근사가 좋지 않다.

2-3-3. Glaz의 근사분포

$X_1,\cdots,X_N$ 은 전체구간 (0,1]에서 독립적으로 균일하게 분포하는 관측치라고 가정하고, $X_1\leq X_2\leq\cdots\leq X_N$ 이라 가정한다. 표기의 편의상 구간의 길이를 T=1로 한다. $N>n\geq 3$에 대하여 사상 A_i를 다음과 같이 정의한다.

$$A_i=\{X_{n+i-1}-X_i\leq r\},\ i=0,1,2,\cdots,N-n+1.\ \text{단, } X_0\equiv 0 \text{ 이다.} \qquad (2.3.12)$$

$0<r\leq \frac{1}{2}$ 에 대하여 길이 r의 스캔 윈도우에서 크기 $n\geq 3$ 인 집락을 적어도 한 개 관측할 확률을 $P(n,N,r)$로 표기하면, 다음이 성립한다.

$$P(n,N,r)=\Pr\left(\bigcup_{i=1}^{N-n+1} A_i\right) \qquad (2.3.13)$$

식 (2.3.13)에 대하여 Hunter(1976)는 Bonferroni의 상한값보다 더욱 개선된 다음과 같은 상한값(upper bound)을 유도하였다.

$$P(n,N,r)\leq \sum_{i=1}^{N-n+1} P_i-\sum_{i=1}^{N-n} P_{ii+1}=(N-n+1)P_0-(N-n)P_{01} \qquad (2.3.14)$$

단, $P_i = \Pr(A_i)$, $P_{ij} = \Pr(A_i \cap A_j)$, $i,j = 0,1,\cdots,N$

또한 Glaz(1989)는 식 (2.3.14)보다 더욱 개선된 다음의 상한값을 유도하였다.

$$P(n,N,r) \le (N-n+1)P_0 - (N-n)P_{01} - \sum_{k=3}^{L}(N-n+2-k)(Q_{k-1}^{*} - Q_k^{*}) \quad (2.3.15)$$

단, $P_0 = \Pr(A_0)$

$$Q_k^{*} = \Pr\{A_0 \cap (\bigcap_{j=1}^{k-1} A_j^c)\}$$

식 (2.3.12)에서 정의된 A_i를 사용하여, Q_i를 다음과 같이 나타낼 수 있다.

$$Q_i = \Pr(\bigcap_{j=0}^{i-1} A_j^c),\ 1 \le i \le N-n+1,\ N > n \ge 3 \quad (2.3.16)$$

$1 \le i \le N-n$에 대하여 Q_{N-n+1} 는 다음과 같이 표시된다.

$$Q_{N-n+1} = 1 - P(n,N,r) = Q_i \prod_{j=i}^{N-n}(Q_{j+1}/Q_j) \quad (2.3.17)$$

많은 N과 작은 r에 대하여, 길이 r의 스캔 윈도우에 포함되는 관측치의 평균수, 즉 $\mu_r = N \cdot r$이 일정할 때, Q_{j+1}/Q_j는 모든 j 에 대하여 일정하므로 다음과 같이 근사될 수 있다.

$$Q_{j+1}/Q_j = \Pr(A_j^c | \bigcap_{k=0}^{j-1} A_k^c) \fallingdotseq \Pr(A_j^c | \bigcap_{k=j-L}^{j-1} A_k^c) = Q_{L+1}/Q_L,\ j \ge L \quad (2.3.18)$$

식 (2.3.17)과 (2.3.18)에 의하여, Glaz(1989)는 식 (2.3.15)를 만족하는 다음의 근사분포를 유도하였다.

$$\hat{Q}_{N-n+1}^{(L+1)} \equiv Q_L (Q_{L+1}/Q_L)^{N-n-L+1},\ L = 1,2,\cdots,N-n \quad (2.3.19)$$

식 (2.3.19)은 L이 클 수록 더욱 정확한 근사값을 주지만 계산하기가 힘들기 때문에, L=3일 때의 근사분포를 이용하는 것이 효과적이다.

2-3-4. 효율성 비교

Naus(1982)가 유도한 근사분포는 많은 N에 대하여 계산하기가 어렵고, 조건확률에 대하여는 정확한 값을 주지 못할 때가 있다. Wallenstein and Neff(1987)는 가능한 많은 N에 대하여도 계산이 쉬운 근사분포를 유도하였고, 이 근사분포는 μ_r, 즉 $N \cdot r$이 클 때 매우 정확하다. 그러나 Wallenstein and Neff(1987)의 근사분포는 $P(n,N,r) > 1$인 경우가 있다. Glaz(1989)는 항상 $0 \le P(n,N,r) \le 1$이고, 많은

N과 작은 r에 대하여도 계산이 쉬운 근사분포를 유도하였다. 특히 μ_r이 작을 때 Glaz(1989)의 근사분포는 매우 정확하다.

n, N, r 의 여러 값에 대한 근사값은 Glaz(1989)가 계산한 [표 1]에서 보여준다. [표 1]에서 (2.3.19), (2.3.5)등은 본문의 식번호를 나타낸다. 단, 식 (2.3.19)를 이용하여 계산한 값은 L=3일 때의 근사값이다. $\mu_r \leq 5$ 인 경우, Glaz(1989)의 근사분포가 가장 정확한 근사값을 준다. $\mu_r > 10$인 경우, $0<a<b<1$인 a와 b에 대하여, ① $P(n,N,r)>b$인 경우에는, Glaz(1989)의 근사분포가 가장 좋다. ② $a \leq P(n,N,r) \leq b$인 경우에는, Naus(1982)와 Wallenstein and Neff(1987)의 근사분포가 좋다. ③ $P(n,N,r)<a$인 경우에는, 모두 비슷한 결과를 준다. 일반적으로 μ_r이 감소하면, a는 증가하고 b는 감소한다. 예를 들어, [표 1]에서 μ_r=10일 때 $a \fallingdotseq 0.25$, $b \fallingdotseq 0.4$이고, μ_r=20일 때 $a \fallingdotseq 0.15$, $b \fallingdotseq 0.55$ 이다. 결론적으로 $\mu_r \leq 5$이거나 $P(n,N,r) \geq 0.8$인 경우에는 Glaz(1989)의 근사분포를 이용하고, $\mu_r > 5$ 이고 $P(n,N,r)<0.8$인 경우에는 Naus(1982)와 Wallenstein and Neff(1987)의 근사분포를 이용하는 것이 좋다.

[표 1] 근사 분포의 비교*

r	n	P(n,100,r)**	(2.3.19)	(2.3.5)	(2.3.10)	(2.3.15)	(2.3.14)
.001	3	.354	.351	.347	.426	.426	.426
	4	.015	.014	.014	.014	.014	.014
.005	3	.999	.999	.998	>1	>1	>1
	4	.680	.670	.657	>1	>1	>1
	5	.124	.124	.124	.131	.132	.133
.01	4	.998	.997	.992	>1	>1	>1
	5	.727	.706	.694	>1	>1	>1
	6	.213	.209	.208	.232	.233	.238
	7	.037	.037	.037	.038	.038	.038
.05	9	.997	.983	.979	>1	>1	>1
	10	.925	.874	.863	>1	>1	>1
	12	.353	.345	.338	.399	.413	.464
	14	.060	.059	.058	.060	.061	.066
.10	14	.999	.984	.981	>1	>1	>1
	16	.858	.800	.783	>1	>1	>1
	18	.408	.401	.383	.449	.494	.590
	20	.116	.121	.115	.120	.128	.147
	22	.025	.025	.024	.024	.025	.028
.20	26	.900	.839	.830	>1	>1	>1
	28	.585	.569	.542	.619	.775	.994
	30	.272	.285	.263	.277	.325	.403
	32	.098	.107	.098	.099	.112	.135
	34	.027	.031	.029	.029	.032	.037
	35	.014	.016	.015	.015	.016	.018

* Glaz(1989)의 p.564 에서 인용
** P(n,100,r)은 20,000 번의 모의실험에 의해 추정되었다.

3. 스캔 통계량의 응용

Glaz(1989)가 언급하였듯이 스캔 통계량의 중요성은 많은 분야에 응용될 수 있다는 점이다. 원자핵물리학에서 충돌에 의해 발생한 입자의 분열에 따른 집락화 여부를 조사함으로써 새로운 물질의 성질을 규명할 수 있는데 이 경우 스캔 통계량을 사용할 수 있다.(Orear and Cassel, 1971)

Ederer, Myers and Mantel(1964)은 1945년 부터 1959년 사이에 미국의 커넥티컷주에서 발생한 333명의 백혈병 환자를 관찰하여 백혈병 발생의 집락여부를 밝히고 있다. 백혈병은 십만명 중에서 열명 정도가 발병하는 드문 병이지만 급성으로 발병하고 특히 어린이에게서는 치명적이다. 따라서 백혈병 발생의 집락여부는 발병원인을 밝힐 수 있는 가능성을 제시해 줄 수도 있으리라 믿는다.

이외에도 Wallenstein(1980)와 Wallenstein and Neff(1987)는 2년 동안 삼염색체성 자연유산이 79건 발생했을 때, 61일 동안의 윈도우에서 15건 이상 발생할 확률을 계산하고 있다. 또한 Conover, Bement and Iman(1979)은 항공탐사를 통한 우라늄 광맥을 찾는 문제에서 스캔 통계량을 이용하고 있으며 다음 두절에서 이 문제를 자세히 다루고자 한다.

3-1. 집락의 결정 방법

전체적인 방법은 일차원 상황에서 사용된 방법을 이차원 지도 위에 응용한 것에 불과하다. 지도의 전체 크기보다 작은, 미리 결정한 직사각형 윈도우가 지도 위를 횡단하고, 수직 방향으로 k번의 패스(윈도우가 지도상을 한 번 횡단하는 것을 패스라고 한다)를 반복함으로써 각 윈도우에서 관측치를 찾는다. 윈도우내에 포함되는 관측치가 n개 이상일 때 집락이 발생하였다고 하고, 한 번의 패스에서는 집락이 거의 발생하지 않도록 n을 결정한다. 즉, $Q^*(n,\lambda L,1/L)$을 길이 $1/L$의 윈도우내에서 집락이 생기지 않을 확률이라고 정의할 때, $Q^*(n,\lambda L,1/L)$이 1에 가깝도록 n을 선택한다. 한 번의 패스에서 n을 결정하는 방법은 일차원 상황에서 사용된 포아송과정과 같고, 윈도우를 수직방향으로 조금씩 이동함으로써 반복되는 k번의 패스에서도 같은 n값을 사용한다. 단, λ는 윈도우에 포함되는 관측치 수의 평균이고, L은 지도의 전체 폭의 길이를 윈도우의 폭의 길이로 나눈 수이다.

3-2. 항공탐사 자료에 근거한 집락의 결정

항공탐사에 의하여 창연에서 나오는 방사능을 기록함으로써, 다른 지역에 비하여 비정상적으로 많은 방사능이 기록된 지역, 즉 이상치(outliers)가 발생한 77개

의 우라늄 존재가능지역을 찾아 다음 <그림 2>에서처럼 23개의 지도선(map line) 위에 표시하였다. 이에 대한 집락화의 여부를 알아보기 위하여 다음과 같은 방법을 이용한다.

0.2도의 경도를 폭(약 6 마일;전체 폭의 십분지 일, 즉 L=10)으로 하고 3개의 지도선을 높이(약 8.5 마일)로 한 직사각형 윈도우를 이용하여, 동쪽에서 서쪽으로 21번 평행하게 이동한다(각각의 패스는 서로 이웃한 3개의 지도선을 포함한다). <그림 2>에서처럼, 직사각형 윈도우내에 포함되는 이상치의 수가 5개인 경우는 직사각형으로, 이상치가 6개인 경우는 대각선이 있는 직사각형으로, 그리고 이상치가 7개 이상인 경우는 교차선이 있는 직사각형으로 표시한다.

L=10, λ=1일 때, Naus(1966)가 유도한 식 (2.2.1)을 이용하여 계산한 $Q(n,10,1/10)$ 의 값은 다음과 같다.

$$Q(5,10,1/10)=.940,\quad Q(6,10,1/10)=.993,\quad Q(7,10,1/10)=.999$$

다음과 같이 λ의 추정치를 구한 후, Conover, Bement, and Iman(1979)이 유도한 식 (2.3.9)에 의하여 계산한 다음의 $Q^*(n,\lambda L,1/L)$의 값은 식 (2.2.1)보다 더 정확하고 계산하기가 쉽다.

$$\hat{\lambda}=\frac{(\text{관찰한 이상치의 수})\times(\text{윈도우가 포함한 지도선의 수})}{(\text{지도선의 총 수})\times(L)}=(77\cdot3)\div(23\cdot10)\fallingdotseq1.0$$

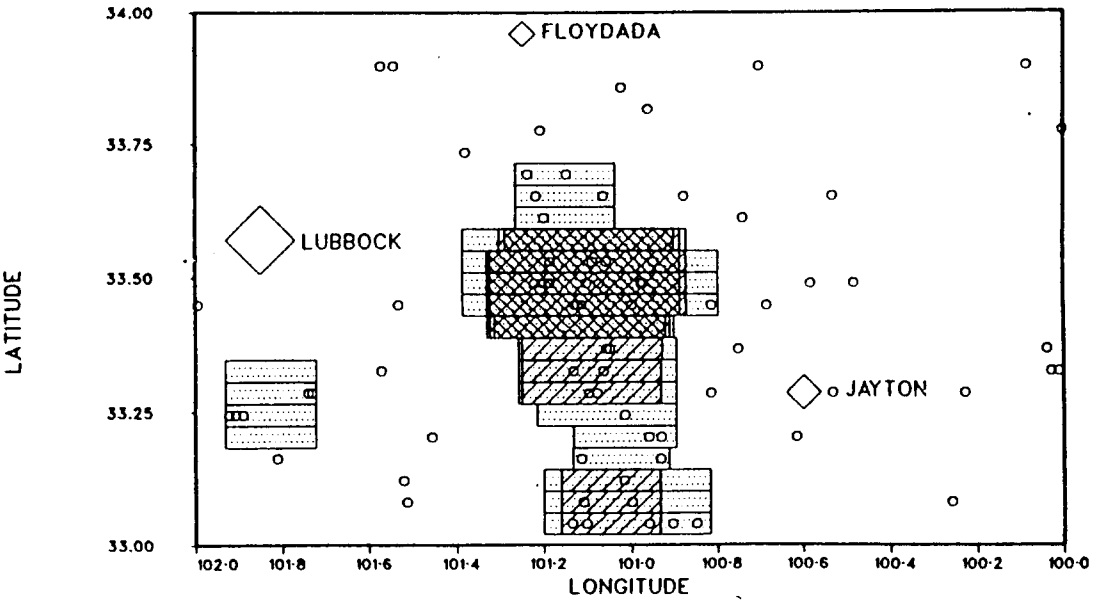
$$Q^*(5,10,1/10)=.888,\quad Q^*(6,10,1/10)=.976,\quad Q^*(7,10,1/10)=.996$$

탐사 방법에서 가장 중요한 것은 자료를 관찰하기 전에 임의로 L의 크기를 정하는 것이다. L=5, 7일 때의 결과는 각각 다음과 같고, L의 크기가 감소하면 집락이 발생할 확률도 비례하여 감소하는 것을 알 수 있다.

$$Q^*(5,7,1/7)=.920,\quad Q^*(6,7,1/7)=.982,\quad Q^*(7,7,1/7)=.997$$
$$Q^*(5,5,1/5)=.946,\quad Q^*(6,5,1/5)=.990,\quad Q^*(7,5,1/5)=.998$$

위의 방법에 의해 집락화의 여부를 알아본 결과, 77개의 우라늄 존재가능지역 중 밀집된 지역, 즉 교차선이 있는 직사각형으로 표시된 지역은 실지답사를 하여 채광여부를 검토할 필요성을 제시한다.

<그림 2> 77개 이상치의 분포*



* Conover, Bement, and Iman(1979)의 p.281 에서 인용

4. 추후 연구과제

분자생물학자는 단백질을 구성하고 있는 아미노산의 염기서열을 관찰함으로써, 특히 두 개의 염기서열을 비교함으로써 종(species)의 진화과정을 밝혀 왔다. $(Y_1, Y_2, \cdots, Y_T)$와 $(Z_1, Z_2, \cdots, Z_T)$를 일렬로 정열된 두 개의 아미노산 염기서열이라 가정한다. Y와 Z의 염기가 위치 i에서 일치될 때, 즉 $Y_i = Z_i$일 때 $X_i = 1$ 이라 하고, 그렇지 않으면 $X_i = 0$이라 한다. 이때 $X_1, X_2, \cdots$은 0과 1을 實現値로 가지는 離散확률변수가 된다. 분자생물학자들은 인간과 마우스(mouse), 그리고 토끼(rabbit)의 아미노산 염기서열을 비교함에 있어서, 소위 "알파벳"으로 불리워지는 "자연적" 분류방법에 따라 분류를 하고 대응(matches)이 연속해서 이루어지는 경우를 찾는다(Karlin and Ghandour, 1985). 이 경우 電荷알파벳(charge alphabet)을 사용하게 되면 $X_i = -1, 0, 1$을 취하게 된다.

4-1. 이산확률변수에 대한 스캔 통계량의 근사분포

$X_1, X_2, \cdots$ 는 다음 식 (4.1.1)을 만족하는 독립적으로 동일하게(identically) 분포하는 이산확률변수라고 가정하고, P_j 와 $Y_{r,t}$를 각각 다음과 같이 정의한다.

$$P(X_i=j) = P_j, \quad j=0, 1, 2, \cdots, c \\ = 0, \quad \text{그렇지 않으면} \tag{4.1.1}$$

단, $\sum_{j=0}^{c} P_j = 1$

$$Y_{r,t} = \sum_{i=r}^{t} X_i, \quad t \geq r \geq 1 \\ = 0, \quad \text{그렇지 않으면} \tag{4.1.2}$$

연속한 m개의 X_i ($i=1,2,\cdots$)의 합을 $Y_{t-m+1,t}(t=m,m+1,\cdots)$로 정의한다. $Y_{t-m+1,t}$의 최대수를 $N_{m,T}$로 표기하고, 다음과 같이 정의한다.

$$N_{m,T} = Max\{Y_{t-m+1,t}\}, \ m \leq t \leq T \tag{4.1.3}$$

길이 m의 스캔 윈도우에 포함되는 X_i (i=1,2,$\cdots$) 의 합이 k만큼 클 때까지 걸리는 시간을 $\tau_{k,m}$ 로 표기하고, 다음과 같이 정의한다.

$$\tau_{k,m} = Min\{Y_{\max(1,t-m+1),t} \geq k\}, \ t \geq 1 \tag{4.1.4}$$

$N_{m,T}$와 $\tau_{k,m}$는 스캔 통계량을 나타내는 두가지 방법인데 $(N_{m,T}<k) = (\tau_{k,m}>T)$ 임을 쉽게 알 수 있다. 또한 $G_{k,m}(T)$와 $f_{k,m}(t)$를 다음과 같이 정의한다.

$$G_{k,m}(T) = P(\tau_{k,m}>T) = P(N_{m,T}<k) = G(T) \tag{4.1.5}$$

$$f_{k,m}(t) = P(\tau_{k,m}=T) = f(t) \tag{4.1.6}$$

Glaz and Naus(1991)는 이산확률변수에 대한 스캔 통계량의 분포에 대하여, 다음과 같은 상한값(upper bound)과 하한값(lower bound)을 유도하였다.

$T \geq im$ 에 대하여,

$$G(T) \leq G(im)(1-A_{j,n})^{T-im}, \ T \geq (n+1)m \tag{4.1.7}$$

$$G(T) \geq G(im)/(1+B_{j,n})^{T-im}, \ T \geq nm \tag{4.1.8}$$

단, $A_{1,n} = f((n+1)m)$

$A_{j,n} = A_{1,n}(1-A_{j-1,n})^{-nm+1}$

$B_{1,n} = f(nm)/G((n+1)m-1)$

$B_{j,n} = f(nm)(1+B_{j-1,n})^{nm}$

$c<k$; i, j, n은 1이상의 정수이다.

Naus(1982)는 다음의 근사분포를 유도하였다.

$$G(T) \fallingdotseq G(2m)[G(3m)/G(2m)]^{(T/m)-2} \tag{4.1.9}$$

$t \le t_1 \le 3m$에 대하여, Glaz and Johnson(1984)은 다음의 근사분포를 유도하였다.

$$G(T) \fallingdotseq G(t_1-1)[G(t_1)/G(t_1-1)]^{T-t_1+1}, \quad T \ge t_1 \tag{4.1.10}$$

식 (4.1.7)과 (4.1.8)은 i와 n이 클 수록 더 좋은 한계값(bound)을 주지만, $G(T)$와 $f(t)$를 계산하기가 매우 힘들다. 따라서 i=3, n=2일 때의 상한값과 하한값을 이용하는 것이 유용하다. 식 (4.1.9)와 (4.1.10)에 의하여 계산된 값은 Glaz and Naus(1991)가 유도한 상한값과 하한값을 모두 만족한다. 그러나 식 (4.1.9)는 $G(2m)$과 $G(3m)$을 계산할 수 있어야 한다는 조건하에서만 값을 구할 수 있다. 또한, 식 (4.1.10)은 $t \le t_1 \le 3m$에 대하여만 $G(t)$를 계산할 수 있다. 이러한 단점을 극복할 수 있는 근사분포를 유도하는 것은 추후 연구과제로 남겨 놓는다.

Glaz and Naus(1991)는 식 (4.1.7), (4.1.8)의 유도과정에서 $c<k$를 가정하였다. 그러나 다음 4-2절에서 다루는 돌연변이의 집락가능성의 문제에는 $c>k$의 경우이므로 Glaz and Naus(1991)의 결과의 수정이 요구된다.

4-2. 돌연변이의 집락화 여부

核酸은 DNA와 RNA로 구별된다. DNA는 시토신, 티민, 아데닌과 구아닌의 4가지 鹽基가 이중나선구조로 연결되어 있고, RNA는 시토신, 우라실, 아데닌과 구아닌의 4가지 염기가 불완전한 나사구조로 연결되어 있다. 核에서 RNA는 DNA의 염기서열에 따라 그 염기서열이 轉寫되고, 세포질에 와서는 리보솜(ribosome)에 결합되어 단백질이 합성되는 동안에 tRNA(운반 RNA)의 대응 순서를 결정한다. 이때, DNA와 RNA의 4개의 염기는 알파베트의 글자와 매우 흡사하게 작용하고, mRNA(전령 RNA)에 의해 암호문(codon)이 잘못 번역될 때, 돌연변이(mutation)가 일어난다.

예를 들어, RNA에서 4가지의 서로 다른 염기, AUCG 중에서 3개의 염기가 암호화될 때, 그 조합수는 4^3=64이다. [UAU CCA UAU CCA UAU]와 같은 단백질합성에 있어, 3번째 염기 다음에 G염기가 삽입되면 그 점에서 완전히 다른

단백질 [UAU GCC AUA UCC AUA U]이 합성된다.

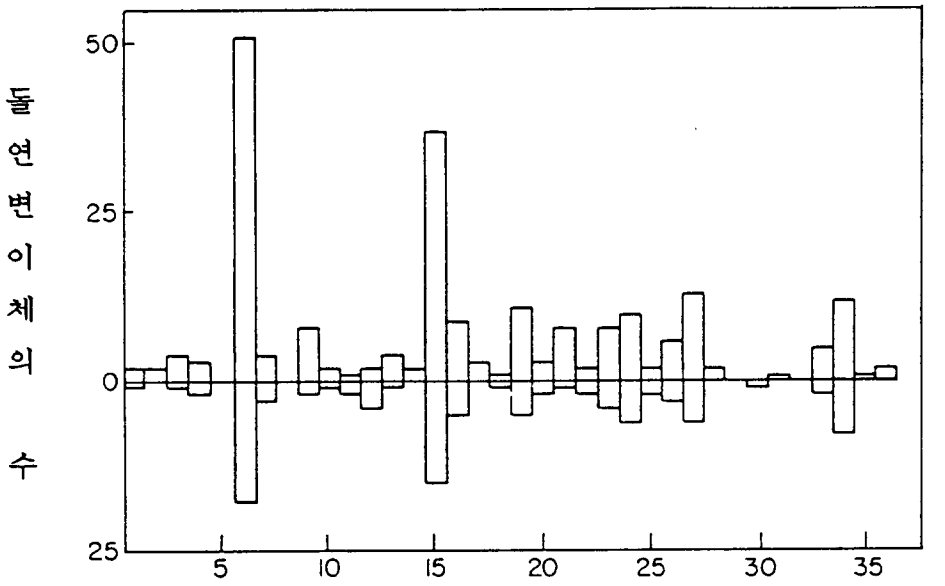
돌연변이는 세포에 異種의 단백질을 생성시키기 때문에, 돌연변이의 기전(mechanism)에 대한 연구가 필요하다. 한 예로서 돌연변이원(mutagen)에 의해서 발생한 DNA의 변화의 형태(type)와 위치(location)의 분석은, 대장균(*Escherichia coli*)에 있어서 돌연변이의 기전을 연구함으로써 가능해진다. Coulondre and Miller(1977)는 대장균에서 발생한 돌연변이체(mutant)의 수를 관찰하였고, Adams and Skopek(1987)은 <그림 3>과 같이 두개의 화학물질에 대한 돌연변이 스펙트럼(mutational spectrum)을 도수분포표로 나타내었다.

전체 돌연변이체의 수에 대한 한 개의 위치 혹은 연속된 n개의 위치에서의 돌연변이체의 수의 비율을 돌연변이율(mutation rate)이라고 하자. <그림 3>의 윗부분을 보면, 한 위치에서의 돌연변이율이 가장 높은 지역은 위치 6이고, 연속된 두 개의 위치에서의 돌연변이율이 가장 높은 지역은 위치 6~7 혹은 15~16으로 생각할 수 있다. 또한 연속된 세 개의 위치에서의 돌연변이율이 가장 높은 지역은 위치 19~21으로 생각할 수 있다.

<그림 3>의 경우에서 돌연변이가 빈번하게 발생하는 위치, 즉 위치 6과 같은 경우를 흔히 핫스팟(hot spot)이라고 부른다. 핫스팟의 경우는 쉽게 발견할 수 있다. 그러나 돌연변이가 어느 한 위치에서 많이 발생하면 그 주변의 위치에도 많이 발생하는 경향이 있느냐 하는 것은 생물학적으로 중요한 문제가 될 수 있다. 즉 어느 한 위치에 두드러진 현상을 찾는 것보다는 인접한 여러개 위치들로 구성된 어느 지역(region)에 몰려있는 현상이 있느냐 하는 것을 검정하고자 할 때 스캔 통계량을 이용할 수 있으리라 믿는다. 예를 들어 스캔 윈도우의 길이가 1이라고 할 때는 스캔 통계량이 위치 6에서 50을 취하게 된다. 그러나 윈도우의 길이를 5로 늘릴 경우 위치 13-17, 혹은 위치 23-27에 돌연변이의 빈도가 몰려있는 현상, 즉 집락현상을 발견할 수 있을 것이고 스캔 통계량의 귀무가설 분포를 계산 혹은 근사할 수 있다면 이 지역에서 주어진 집락현상이 발생할 확률을 계산할 수 있다. 따라서 돌연변이가 핫스팟 뿐만 아니라 집락으로도 발생할 수 있다라는 가설을 검정할 수 있을 것이다.

필자들이 알고 있는 한 스캔 통계량을 사용하여 돌연변이의 집락적 발생여부를 검정한 연구보고는 아직 없고, 관련된 통계적 절차로 Adams and Skopek(1987)은 두개의 화학물질이 두개의 돌연변이 도수분포(mutational spectrum)를 나타낼 때 횟셔의 정확한 검정을 다변량초기하분포의 경우로 확장하여 두 도수분포의 동질성을 검정하고 있다.

<그림 3> 돌연변이체에 대한 돗수분포표*



돌연변이의 발생 위치(location)

* Adams and Skopek(1987)의 p.393 에서 인용

5. 감사의 글

본 논문의 초고를 읽고 많은 건설적 비평을 하여 주신 두 익명의 심사위원께 감사를 드리며, 아울러 두 분의 논평과 비평이 본 논문을 개선하는데 많은 도움이 되었음을 밝힌다.

참 고 문 헌

[1] Adams, W. T. and Skopek, T. R. (1987),"Statistical Test for the Comparison of Samples from Mutational Spectra," *Journal of Molecular Biology*, 194, 391-396.

[2] Conover, W.J., Bement, T.R., and Iman, R.L.(1979), "On a Method for

Detecting Clusters of Possible Uranium Deposits," *Technometrics*, 21, 277-282.

[3] Coulondre, C. and Miller, J.H.(1977), "Genetic Studies of the *lac* Repressor : Ⅳ. Mutagenic Specificity in the *lac* I Gene of *Escherichia coli*," *Journal of Molecular Biology*, 117, 577-606.

[4] Ederer, F., Myers, M., and Mantel, N.(1964), "A Statistical Problem in Space and Time :Do Leukemia Cases Come in Clusters?" *Biometrics*, 20, 626-638.

[5] Feller, W.(1957), *An Introduction to Probability Theory and its ApplicaTion*, Vol. 1, 2nd Ed., Wiley, New York, 36-38.

[6] Glaz, J.(1989),"Approximations and Bounds for the Distribution of the Scan Statistic," *Journal of the American Statistical Association*, 84, 560-566.

[7] Glaz, J. and Johnson, B.McK.(1984), "Probability Inequalities for Multivariate Distributions with Dependence Structures," *Journal of the American Statistical Association*, 79, 436-441.

[8] Glaz, J. and Naus, J.I.(1983), "Multiple clusters on the line," *Communications in Statistics Theory and Methodology*, 12, 1961-1986.

[9] Glaz, J. and Naus, J.I.(1991), "Tight Bounds and Approximations for Scan Statistic Probabilities for Discrete Data," *The Annals of Applied Probability*, 1, 306-318.

[10] Hunter, D.(1976), "An Upper Bound for the Probability of a Union," *Journal of Applied Probability*, 13, 597-603.

[11] Karlin, S. and Ghandour, G.(1985), "Multiple-alphabet aminoacid sequence comparisons of the immunoglobulin kappa-chain constant domain," *Proceedings of National Academy of Science U.S.A.*, 82, 8597-8601.

[12] Naus, J.I.(1966), "Some Probabilities, Expectations and Variances for the Size of the Largest Clusters and Smallest Intervals," *Journal of the American Statistical Association*, 61, 1191-1199.

[13] Naus, J.I.(1982), "Approximations for Distributions of Scan Statistics," *Journal of the American Statistical Association*, 77, 177-183.

[14] Neff, N.D. and Naus, J.I.(1980), *Selected Tables in Mathematical Statistics, Vol.6; The Distribution of the Size of Maximum Cluster of Points on a Line*, American Mathematical Society, Providence, Rhode Island.

[15] Orear, J. and Cassel, D.(1971), "Applications of Statistical Inference to Physics," in *Foundation of Statistical Inference*, eds. Godambe, V. and Sprott, D., Holt, Rinehart and Winston, Toronto, 280-289.

[16] Tango, T.(1984), "The detection of disease clustering in time," *Biometrics*, 40, 15-26.

[17] Wallenstein, S.(1980), " A test for detection of clustering over time," *American Journal of Epidemiology*, 3, 367-372.

[18] Wallenstein, S. and Neff, N.D.(1987), "An Approximation for the Distribution of the Scan Statistic," *Statistics in Medicine*, 6, 197-207.

A Review on the Development of a Scan Statistic and Its Applications

Kim, Byung Soo[1] and Kim, Gie Han[1]

Abstract

The primary objective of the paper is to review the development of approximations of the null distribution of a scan statistic and to show how these approximations were improved.

Let $X_1, \cdots, X_N$ be a sequence of independent uniform random variables on an interval $(0,T]$. A scan statistic is defined to be the maximum number of observations in a subinterval of length $t \leq T$, when we continuously (or discretely) move the subinterval from 0 to T. A scan statistic is used to test whether certain events occur in a cluster aganist a null hypothesis of the uniformity. It is difficult to calculate the exact null distribution of a scan statistic. Several authors have suggested approximations of the null distribution of a scan statistic since Naus(1966).

We conceive that a scan statistic can be used for detecting a "hot region" in a mutational spectrum, where a "hot region" is defined to be a region at which the frequencies of mutations are relatively high. A "hot region" may be regarded as a generalized version of a hot spot. We leave it for a further study the concrete formulation of detecting a "hot region" in a mutational spectrum.

[1] Department of Applied Statistics, Yonsei University, Seoul, 120-749.