ANKARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

TARAMA İSTATİSTİKLERİ VE BAZI UYGULAMALARI

Fürüzan KÖKTÜRK

İSTATİSTİK ANABİLİM DALI

ANKARA
2007

Yrd. Doç. Dr. İhsan KARABULUT danışmanlığında Fürüzan KÖKTÜRK tarafından hazırlanan **"Tarama İstatistikleri ve Bazı Uygulamaları"** adlı tez çalışması 26/06/2007 tarihinde aşağıdaki jüri tarafından oybirliği ile Ankara Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Başkan : Prof. Dr. Hamza GAMGAM
Gazi Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü

Üye : Yrd. Doç. Dr. Halil AYDOĞDU
Ankara Üniversitesi Fen Fakültesi İstatistik Bölümü

Üye : Yrd. Doç. Dr. İhsan KARABULUT
Ankara Üniversitesi Fen Fakültesi İstatistik Bölümü

Yukarıdaki sonucu onaylarım.

Prof. Dr. Ülkü MEHMETOĞLU
Enstitü Müdürü

ÖZET

Yüksek Lisans Tezi

TARAMA İSTATİSTİKLERİ VE BAZI UYGULAMALARI

Fürüzan KÖKTÜRK

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Yrd. Doç. Dr. İhsan KARABULUT

Sıra istatistikleri arasındaki ardışık farklardan hareketle tanımlanan tarama istatistikleri, zaman ve konum boyutlarında rasgele oluşların kümelenmesinin analizinde son yıllarda kullanılmaktadır. Kullanım alanları arasında endüstri, tıp, epidemiyoloji, biyoloji ve özellikle genetik araştırmalar, madencilik, meteoroloji sayılabilir. Bu çalışmanın amacı, tarama istatistiklerinin tanıtımını yapmak ve kullanımlarına örnekler vererek ihtiyaç duyabilecek araştırmacılara bir kaynak sunmaktır.

2007, 58 sayfa

Anahtar Kelimeler: Sıra istatistikleri, ardışık sıra istatistikleri arasındaki farklar (spacings), tarama istatistikleri

ABSTRACT

Master Thesis

SCAN STATISTICS AND SOME APPLICATIONS

Fürüzan KÖKTÜRK

Ankara University
Graduate School of Natural and Applied Sciences
Department of Statistics

Supervisor: Asst.Prof.Dr. İhsan KARABULUT

Scan statistics originated from the differences of contigious order statistics is used at time and space dimensions to analyze of the clustering of random occurences. It can be used in industry, medicine, epidemiology, biology, especially genetic researches, mining and meteorology. The aim of this study is to give information about scan statistics and to present a source to researchers also by giving some examples.

2007, 58 pages

Key Words: Order statistics, spacings, scan statistics

TEŞEKKÜR

Tez çalışmam sırasında bilgisi ve tecrübesi ile bana her konuda ve her zaman destek veren, yardımlarını esirgemeyen danışman hocam sayın Yrd. Doç. Dr. İhsan KARABULUT'a teşekkürü bir borç bilirim.

Ayrıca çalışmalarım esnasında manevi desteklerini ve ilgilerini esirgemeyen aileme, tüm çalışmam sırasında hep yanımda ve yardımcı olan çok sevgili arkadaşlarıma teşekkür ederim.

Fürüzan KÖKTÜRK
Ankara, Haziran 2007

İÇİNDEKİLER

SİMGELER DİZİNİ

λ	Lambda
$X_{(i)}$	i. sıra istatistiği
S_w	Tarama istatistiği
W_k	Tarama istatistiği
W_{r+1}	r inci tarama istatistiği
P	Geçmişe dönük tarama istatistiklerine ilişkin olasılık
P^*	İleriye dönük tarama istatistiklerine ilişkin olasılık
$\approx$	Benzer
$\cong$	Eş
DNA	Deoksiribonükleik asit
RNA	Ribonükleik asit
HCMV	İnsanda bulunan Stomegalovirüs
A	Adenin
C	Sitozin
G	Guanin
T	Timin
PLP	Palindrom örüntü

ÇİZELGE DİZİNİ

1. GİRİŞ

Son zamanlarda istatistik kuramında sıra istatistiklerinin çok önemli olduğu görülmüştür. Ardışık sıra istatistikleri ve bu sıra istatistikleri arasındaki farkların oluşturduğu aralıkların (spacings) dağılımı üzerine pek çok araştırma yapılmış ve değişik sonuçlar bulunmuştur. Bu çalışmalardan Venter (1967), Seth (1950), Lieblein (1952) ve Pyke (1965)'nin çalışmaları dikkat çekici olanlarıdır. İkinci bölümde sıra istatistikleri ve ardışık sıra istatistikleri arasındaki farklarla oluşan aralıkların dağılımları düzgün, üstel ve genelleştirilmiş olmak üzere üç kısımda incelenerek rasgele örneklemin alındığı yığının dağılımının düzgün ve üstel olmaması durumlarında hesaplamadaki zorluklar vurgulanmıştır.

Üçüncü bölümde ardışık sıra istatistikleri ile bağlantılı olarak tanımlanan tarama istatistikleri, zaman ve mekan boyutlarında oluşların rasgeleliğinin araştırılması ve bu rasgeleliğin ait olduğu yığının saptanması konularını ele almaktadır. Örneğin, belirli bir zaman ve yerde bir hastalığın yaygınlığının gerçekten düzgün bir rasgele oluş mu olduğu yoksa belirli bir rasgelelik kanununa mı uyduğu sorgulanabilir. Ya da belirli bir genin DNA yapısı içindeki konumu bir araştırmaya konu olabilir. Pek çok alanda kullanılmaya başlanılan tarama istatistiklerinin bunlar ve benzeri türden araştırmalarda başvurulabilecek bir yöntem olduğu düşünülmektedir. Bu bölüm hazırlanırken ağırlıklı olarak Glaz *et al.* (2001)'ın çalışmasından faydalanılmış ve tarama istatistiklerine ait olasılıkların hesaplanmasındaki zorluklar okuyucuya gösterilmeye çalışılmıştır.

2. ARDIŞIK SIRA İSTATİSTİKLERİ

2.1 Sıra İstatistiklerinin Dağılımları

$X_1, X_2, \ldots, X_n$ bağımsız ve aynı dağılımlı rasgele değişkenler ve X_i, $i = 1, 2, \ldots, n$, rasgele değişkenlerinin sıra istatistikleri,

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$$

olsun.

Buna göre $X_{(i)}$, rasgele örneklemin *i*. sıra istatistiğidir.

$X_1, X_2, \ldots, X_n$ bağımsız rasgele değişkenleri olasılık yoğunluk fonksiyonu $f(x)$, dağılım fonksiyonu $F(x)$ olan sürekli bir dağılımdan alınmış olsun. $X_{(k)}$ istatistiğinin olasılık yoğunluk fonksiyonu $f_k(x)$ ve dağılım fonksiyonu $F_k(x)$ olarak gösterilsin. Buna göre *k*. sıra istatistiğinin olasılık yoğunluk fonksiyonu,

$$f_k(x) = \frac{n!}{(k-1)!(n-k)!}\left[F(x)\right]^{k-1}\left[1-F(x)\right]^{n-k} f(x) \quad , \qquad -\infty < x < \infty$$

ve *n* tane sıra istatistiğinin ortak olasılık yoğunluk fonksiyonu,

$$f_{1,2,\ldots,n}(x_1, \ldots x_n) = n!\prod_{i=1}^{n} f(x_i) \qquad , \qquad -\infty < x_1 < x_2 < \ldots < x_n < \infty$$

dır.

Örneğin, U_i rasgele değişkenleri birbirinden bağımsız ve (0,1) aralığındaki düzgün dağılımdan alınmış n çaplı bir örneklem için $U_{(1)}, U_{(2)}, \ldots, U_{(n)}$ rasgele değişkenleri bunlara ait sıra istatistiklerini göstersin. Bu durumda $U_{(i)}$ sıra istatistiklerinin ortak olasılık yoğunluk fonksiyonu,

$$f_{1,2,\ldots,n}(u_1, u_2, \ldots, u_n) = \begin{cases} n! & , \quad 0 \le u_1 \le u_2 \le \cdots \le u_n \le 1 \\ 0 & , \quad d.y \end{cases}$$

dir.

Genel olarak en küçük ve en büyük sıra istatistiklerinin olasılık yoğunluk fonksiyonları sırası ile,

$$f_1(x) = n\{1 - F(x)\}^{n-1} f(x) \quad , \quad -\infty < x < \infty$$

ve

$$f_n(x) = n\{F(x)\}^{n-1} f(x) \quad , \quad -\infty < x < \infty$$

dir.

Dağılım fonksiyonları da sırası ile ,

$$F_1(x) = 1 - \{F(x)\}^n \quad , \quad -\infty < x < \infty$$

ve

$$F_n(x) = \{F(x)\}^n \quad , \quad -\infty < x < \infty$$

olarak bulunur (Günay ve İnal 1993).

2.2 Ardışık Sıra İstatistikleri Arasındaki Farkların (Spacings) Dağılımı

Bu bölümde kısaca ardışık sıra istatistikleri arasındaki farkların dağılımı üzerine Pyke (1965) tarafından yapılan incelemeler üzerinde durulmuştur. Ardışık sıra istatistikleri arasındaki farkların oluşturduğu aralıkların (spacings) dağılımı, düzgün, üstel ve genelleştirilmiş olmak üzere üç durumda ele alınacaktır.

Bu üç durumda n gözlemden oluşan bir örneklem için $n+1$, n veya $n-1$ tane fark oluşturulabilmektedir. Düzgün dağılımlı rasgele değişkenler, sınırlı bir reel sayı aralığında değer alan rasgele değişkenlere; üstel dağılım, aralığının bir tarafı sınırlı diğer tarafı sınırsız bir aralıkta değer alan rasgele değişkenlere örnek oluşturur. Genelleştirilmiş durum ise normal dağılımda olduğu gibi değer aralığının her iki tarafı da sınırsız olan rasgele değişkenlere ilişkin fark istatistiklerini konu almaktadır.

$X_1, X_2, \ldots, X_n$ birbirinden bağımsız ve aynı dağılıma sahip rasgele değişkenler olsun. Bu rasgele değişkenlere ait sıra istatistikleri,

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)}$$

olmak üzere,

$$X_{(i+m)} - X_{(i)} \qquad , \qquad 1 \le i < i+m \le n$$

bir *m – fark (m-spacings)* olarak ifade edilir (Miller 2003).

2.2.1 Düzgün dağılım durumunda ardışık sıra istatistikleri arasındaki farkların dağılımı

$X_1, X_2, \ldots, X_n$, bağımsız ve (0,1) aralığında düzgün dağılıma sahip rasgele değişkenler olmak üzere, $\underline{X} = (X_1, X_2, \ldots, X_n)$ rasgele vektörünün ortak olasılık yoğunluk fonksiyonu,

$$f_{\underline{X}}(x_1, x_2, \ldots, x_n) = \begin{cases} 1 & , \quad 0 \le x_i \le 1, 1 \le i \le n \quad için \\ 0 & , \quad d.y \end{cases}$$

dir.

$\underline{U} = (U_1, U_2, \ldots, U_n)$ bunlara ait sıra istatistiklerini göstermek üzere; ortak olasılık yoğunluk fonksiyonu,

$$f_{\underline{U}}(u_1, u_2, \ldots, u_n) = \begin{cases} n! & , \quad 0 \le u_1 \le u_2 \le \ldots \le u_n \le 1 \\ 0 & , \quad d.y. \end{cases}$$

olur.

$U_0 = 0$ ve $U_{n+1} = 1$ olmak üzere ardışık sıra istatistikleri ile oluşturulan farklar, $D_i = U_i - U_{i-1}$, $1 \le i \le n+1$ olsun. $D_1 + D_2 + \ldots + D_{n+1} = 1$ dir. Farklardan oluşan $\underline{D} = (D_1, D_2, \ldots, D_n)$ vektörünün ortak olasılık yoğunluk fonksiyonu,

$$f_{\underline{D}}(d_1, d_2, \ldots, d_n) = \begin{cases} n! & , \quad d_i \ge 0 \; ve \; d_1 + d_2 + \ldots + d_{n+1} = 1 \\ 0 & , \quad d.y \end{cases}$$

olur.

Düzgün dağılımlı yığından rasgele örneklemle oluşturulan ardışık sıra istatistikleri arasındaki farklar yer değiştirebilen (exchangeable) rasgele değişkenlerdir. Dolayısıyla herhangi bir D_i aralığının olasılık yoğunluk fonksiyonu D_1'in olasılık yoğunluk fonksiyonuna ve $(D_i, D_j)\ (i \neq j)$ çiftinin ortak olasılık yoğunluk fonksiyonu da (D_1, D_2)'nin ortak olasılık yoğunluk fonksiyonuna eşittir. Buna göre, $x, y \geq 0$ ve $x + y \leq 1$ için dağılım fonksiyonları,

$$F_{D_i}(x) = F_D(x) = F_{U_1}(x) = 1 - (1-x)^n$$

ve

$$\begin{aligned} F_{(D_i,D_j)}(x,y) &= P(U_1 \leq x, U_2 - U_1 \leq y) \\ &= n\int_0^x \left\{1 - (1 - \frac{y}{1-u})^{n-1}\right\}(1-u)^{n-1}\,du \\ &= 1 - \left\{(1-x)^n + (1-y)^n - (1-x-y)^n\right\} \end{aligned}$$

dir.

Buradan D_i ve (D_i, D_j) için olasılık yoğunluk fonksiyonları sırasıyla,

$$f_{D_i}(x) = n\,(1-x)^{n-1}$$

ve

$$f_{(D_i,D_j)}(x,y) = n\,(n-1)\,(1-x-y)^{n-2}$$

olarak elde edilir.

$U_{(1)} = u$ verildiğinde $U_{(2)} - U_{(1)}$ in koşullu dağılımı, $[0, 1-u]$ aralığında düzgün dağılıma sahip $n-1$ tane rasgele değişkenin oluşturduğu örneklemden elde edilen ardışık sıra istatistikleri arasındaki farklarla meydana gelen aralıkların dağılımına eşittir.

2.2.2 Üstel dağılım durumunda ardışık sıra istatistikleri arasındaki farkların dağılımı

Fark istatistiklerinin en kullanışlı örnekleri üstel dağılım tarafından sağlanır. $\underline{X} = (X_1, X_2, \ldots, X_n)$ bağımsız rasgele vektörünün olasılık yoğunluk fonksiyonu, $\lambda > 0$ için, $f(x) = \lambda \exp(-\lambda x)$, $x > 0$ olarak verilsin. $\underline{T} = (T_1, T_2, \ldots, T_n)$, $X_1, X_2, \ldots, X_n$ rasgele örnekleminden elde edilen sıra istatistiklerini göstersin ve $T_0 = 0$, $D_i = T_i - T_{i-1}$ $(1 \le i \le n)$ olmak üzere tanımlanan $\underline{D} = (D_1, D_2, \ldots, D_n)$ vektörünün ortak olasılık yoğunluk fonksiyonu $d_i > 0$ $(1 \le i \le n)$ için,

$$
\begin{aligned}
f_{\underline{D}}(d_1,\ldots,d_n) &= n!\,\lambda^n \prod_{i=1}^{n} \exp\{-\lambda(d_1 + \ldots + d_i)\} \\
&= n!\,\lambda^n \exp\left\{-\lambda \sum_{i=1}^{n} (n-i+1) d_i\right\} \\
&= \prod_{i=1}^{n} \lambda\,(n-i+1) \exp\{-\lambda\,(n-i+1) d_i\}
\end{aligned}
$$

dir.

Böylelikle $D_1, D_2, \ldots, D_n$ bağımsız ve parametreleri sırasıyla $\lambda n, \lambda(n-1), \ldots, \lambda$ olan üstel dağılıma sahip rasgele değişkenler olmaktadır. Bu rasgele değişkenlere normalleştirilmiş aralıklar denmektedir. Buradaki önemli nokta, üstel dağılım durumunda, ardışık sıra istatistikleri arasındaki farkların birbirinden bağımsız oluşudur (Pyke 1965).

2.2.3 Ardışık sıra istatistikleri arasındaki farkların genelleştirilmiş dağılımı

$\underline{X} = (X_1, X_2, \ldots, X_n)$, n tane bağımsız ve aynı dağılıma sahip rasgele değişkenden oluşan örneklem ve $\underline{T} = (T_1, T_2, \ldots, T_n)$ bunlara ait sıra istatistiklerinden oluşan vektör olsun. T vektörünün herhangi bir alt vektörü $\underline{T}_{\underline{k}} = (T_{k_1}, T_{k_2}, \ldots, T_{k_m})$ vektörünün ortak olasılık yoğunluk fonksiyonu $1 \le k_1 < k_2 < \ldots < k_m \le n$ için,

$$f_{\underline{T}_{\underline{k}}}(t_1,\ldots,t_m) = \begin{cases} n! \prod_{j=1}^{m+1} \left\{F(t_j) - F(t_{j-1})\right\}^{k_j - k_{j-1} - 1} f(t_j)/(k_j - k_{j-1} - 1)!, & t_1 < t_2 < \ldots < t_m \\ 0 & , d.y \end{cases}$$

dir. Burada $t_0 = -\infty,\ t_{m+1} = +\infty,\ k_{m+1} = n+1$ ve $f(t_{m+1}) = 1$ dir. Özel olarak $\underline{T}$ vektörünün ortak olasılık yoğunluk fonksiyonu,

$$f_{\underline{T}}(t_1,\ldots,t_n) = \begin{cases} n! f(t_1) f(t_2) \ldots f(t_n) & , \quad t_1 < t_2 < \ldots < t_n \\ 0 & , \quad d.y \end{cases}$$

dir.

$D_i = T_i - T_{i-1}\ (2 \le i \le n)$ ve $\underline{D} = (D_1, D_2, \ldots, D_n)$ olmak üzere yukarıdaki lineer dönüşüm,

$$f_{\underline{D}}(d_2, d_3 \ldots, d_n) = \begin{cases} n! \int_{-\infty}^{\infty} \prod_{i=2}^{n} f(x + d_2 + d_3 + \ldots + d_i) dx & , \quad d_i > 0, \quad (2 \le i \le n) \\ 0 & , \quad d.y \end{cases}$$

ya da *i*. fark istatistiğinin marjinal olasılık yoğunluk fonksiyonu,

$$f_{D_i}(y) = \int_{-\infty}^{\infty} f_{(T_{i-1},T_i)}(x, x+y)dx$$

$$= \frac{n!}{(i-2)!(n-i)!} \int_{-\infty}^{\infty} \{F(x)\}^{i-2} \{1-F(x+y)\}^{n-i} f(x) f(x+y)dx$$

olarak elde edilir.

D_i ve D_j 'nin ortak olasılık yoğunluk fonksiyonu, *u, v > 0* için,

$$f_{(D_i,D_j)}(u,v) = n! \int_{-\infty}^{\infty} \int_{x+u}^{\infty} \frac{\{F(x)\}^{i-2}}{(i-2)!} \cdot \frac{\{F(y)-F(x+u)\}^{j-i-2}}{(j-i-2)!}$$

$$\times \frac{\{1-F(y+v)\}^{n-j}}{(n-j)!} f(x) f(x+u) f(y) f(y+v)dydx$$

olmaktadır.

Burada görülebileceği gibi ardışık sıra istatistikleri arasındaki farkların olasılık yoğunluk fonksiyonlarını elde etmek kolay değildir. Bu aralıklar üzerinde yapılan çalışmalarda sıra istatistiklerinin özelliklerinden yararlanılır (Pyke 1965). Ardışık sıra istatistikleri arasındaki farkların uygulamada kullanımlarına ilişkin olarak, Venter (1967), Seth (1950), Lieblein (1952) ve Pyke (1965)'nin çalışmaları örnek olarak gösterilebilir. Bundan sonra ele alınacak olan tarama istatistikleri ile ardışık sıra istatistikleri birbirleriyle bağlantılıdır.

3. TARAMA İSTATİSTİKLERİ

Pek çok alanda araştırmacılar olayların kümelenmesine önem verirler. Örneğin, moleküler biyologlar virüslerin çoğalmasını sağlayan genlerin saptanması için DNA içindeki palindrom gruplarını ya da kalite kontrol uzmanları üretim bandındaki bozukların gruplarını araştırırlar.

Tarama istatistikleri, zamanın ve konumun (mekanın) taranarak, tanımlanan olay gruplarının araştırılması problemlerini konu alır. N tane nokta (0, T) gibi bir zaman aralığında düzgün dağılıma sahip olarak konumlansın. S_w, w uzunluğundaki sabit bir "zaman" aralığında gerçekleşen olayların maksimum sayısıdır ve bu rasgele değişkene *tarama istatistiği (scan statistics)* adı verilir. Bir diğer rasgele değişken W_k, verilen sabit k sayıda olay içeren en kısa zaman aralığıdır. W_{r+1}, en küçük *r. sıralı aralık* (*r-th order gap*) veya *r. tarama istatistiği (r-scan statistics)* olarak adlandırılır. S_w ve W_k istatistiklerinin dağılımları için,

$$P(W_k > w) = P(S_w < k)$$

dır. S_w ve W_k istatistikleri ile tanımlanan olaylar arasındaki bu ilgi, k tane olay içeren en kısa aralık w uzunluğunda ise k veya daha fazla sayıda olay içeren w uzunluğunda bir başka aralığın olmaması biçiminde de ifade edilebilir.

Örneğin, bir şehirde 1991-1995 yılları arasındaki 5 yıllık bir dönemde 19 adet belirli tip kanser vakası tespit edilmiş ve veriler incelendiğinde 14 Nisan 1993 ile 13 Nisan 1994 tarihlerini kapsayan 1 yıllık dönemde 8 olayın gerçekleştiği gözlenmiş olsun.

Yukarıdaki şekilde 1 Ocak 1991 ile 31 Aralık 1995 tarihleri (0,1) birim zaman aralığı olarak gösterilmiş ve bu dönem $w = 0.2$ uzunluklu bir aralıkla taranmıştır. Burada, $N = 19,\ S_{0.2} = 8$ olmaktadır. Bu tarama aralığında gözlenen maksimum kanser vaka sayısı 8 dir.

19 olayın gerçekleştiği 5 yıllık bir dönemde, herhangi bir 1 yıllık zaman aralığının 8 veya daha fazla olay içermesi olağan mıdır (düzgün dağılıma uygun mudur?) sorusu sorulabilir. 19 olayın her biri diğer olaylardan bağımsız olarak bu 1 yıllık zaman aralığının içinde gözlenebilirdi. O halde 8 veya daha fazla olayın bu 1 yıllık zaman aralığında gözlenmesi olasılığı, $N = 19,\ p = 1/5$ alınarak binom olasılığı yardımıyla hesaplanabilir. Fakat bu hesaplama sorunun cevabı değildir. Araştırmacılar belirtilen 14 Nisan 1993 ile 13 Nisan 1994 tarihleri ile tanımlanan özel bir yılı değil, herhangi 1 yıllık zaman aralığının bu kadar olay içermesinin düzgün dağılıma uygunluğunu sorgulamaktadırlar.

Bu problem, 5 yıllık dönem ayrık 1 yıllık zaman aralıklarına bölünerek ve herhangi bir yılda gözlenen maksimum olay sayısının dağılımı kullanılarak da çözülmeye çalışılabilir. Yinede bu sorunun cevabı değildir. Çünkü maksimum olay sayısının gözlendiği yıl ayrılmış iki yılın arasında kalmış olarak gözlenebilir. Araştırmacılar soruya takvim yılı olarak da sınır koymamışlardır.

Bir başka örnek ise şöyledir: Yolcu ve yük taşıyan bir uçak firmasının 25 gün içerisinde görülen 3. kazadan sonra bütün uçuşları askıya alınmıştır. Bir aydan daha kısa bir süre içerisinde görülen bu 3 kaza, 5 yıllık dönem içerisindeki beklenen orandan neredeyse 7 kat fazladır. Bu örnekte $w = 25$, $S_{25} = 3$, $k = 3$, $W_3 = 25$ olmaktadır.

Bir diğer örnek olarak, bir santrale 1 dakika içinde gelen aramaların zaman içindeki dağılımının düzgün olduğu varsayılsın. Her aramanın da 10 saniye süre aldığı bilindiğinde santrale yönlenecek 15 aramanın en az 8 tanesinin aynı zamanda çevrilmesi olasılığı nedir sorusuna cevap arandığı düşünülürse, $(N = 15)$ olmak üzere, $P(S_{1/6} \geq 8)$ olasılığına cevap arandığı anlaşılır (Naus 1965).

Tarama istatistikleri [0, T) gibi sürekli bir zaman aralığında tanımlanmıştır. Yine tarama istatistikleri benzer olarak T denemeli bir zinciri üzerinde de tanımlanabilir. Bunlar kesikli olaylardır. Bernoulli denemelerinin bir zincirindeki en uzun başarı tekrarının sayısı kesikli tarama istatistiklerinin özel bir durumudur. N deneme içerisindeki herhangi bir ardışık m denemedeki başarı sayılarının maksimumu S'_m ile gösterilir ve kesikli tarama istatistiği adını alır.

Kesikli tarama istatistiklerinin önemli uygulamaları protein veya DNA dizilerinin eşleşmelerinde ortaya çıkar. Örneğin, virüs DNA ve konukçu DNA arasındaki sıra dışı büyük eşleşmeler bazı ipuçlarını ve hastalıkların seyrinin anlaşılmasını sağlar.

Tarama istatistikleri iki ve daha yüksek boyutlar için de tanımlanmıştır. Pek çok alanda sıra dışı kümelenmelerin belirlenmesi için iki yada daha çok boyutlu taramalar yapılmaktadır.

3.1 Zaman Boyutunda Olayların Geçmişe Dönük (Retrospective) Taranması

3.1.1 Koşullu durum: Olayların düzgün dağılımı

Ortaya çıkması imkansız gibi görünen olayların sezgisel olarak sıra dışı kabul edilip edilmeyeceği ya da olayların bağımsız ve zaman üzerinde tamamen rasgele olarak dağılıp dağılmadığı araştırmacılar tarafından merak konusudur. Tamamen rasgele

oluşan gözlem kümelerinde göreli sıklıklar bu türden bazı soruların cevaplanması için bir araçtır.

Burada belirli bir rasgele modele göre kümelenmeleri tanımlayan iki istatistik üzerinde durulmaktadır. Rasgele model, olayların oluş zamanlarının birbirlerinden bağımsız olarak dağıldıklarını ve zamanın herhangi bir aralığında gözlenme olasılıklarının eşit olduğunu varsayar. Birbirleriyle ilgili iki istatistik, sabit bir zaman aralığındaki olayların maksimum sayısı (S_w) ve sabit bir sayıda olay içeren en kısa zaman aralığıdır (W_k).

N tane noktanın verilmesi koşulu altında hesaplama yapılması, geriye dönük gözlem sonuçlarıyla değerlendirme yapılıp olasılık hesaplanması durumunda bu tarama işlemi geriye dönük tarama olarak adlandırılır. Bu tarama işlemine ait olasılıklar da geçmişe dönük tarama olasılıkları olmaktadır. Verilen N nokta [0,1) zaman aralığında bağımsız olarak düzgün dağılıma uygun biçimde konumlanmış olsun. S_w, [0,1) aralığının w uzunluğundaki herhangi bir alt aralığında bulunan en fazla sayıdaki olay sayısını gösteren rasgele değişken olsun. W_k, [0,1) aralığının k sayıda olay içeren en küçük alt aralığının uzunluğunu tanımlayan rasgele değişkendir. W_{r+1}, minimum r. inci sıralı aralık olarak adlandırılır. $P(S_w \geq k) = P(W_k \leq w)$ olasılığı, $P(k; N, w)$ ile gösterilecektir ve $Q(k; N, w) = 1 - P(k; N, w)$ olmaktadır (Glaz *et al.* 2001)[1].

Bazı araştırmalarda bu olasılık tam olarak belirlenmek istenirken diğerlerinde yaklaşık hesaplamalar yeterli olabilmektedir.

3.1.2 P(k; N, w) için yaklaşık sonuçlar

S_w ve W_k rasgele değişkenlerine ait olasılık fonksiyonlarının elde edilmesi ve olasılıkların hesabının doğrudan yapılması gözlem sayısının artmasıyla daha da zorlaşır. Bu nedenle S_w rasgele değişkeninin olasılığına ilişkin yaklaşımlar geliştirilmiştir Naus

[1] Bu kaynak bundan sonraki kullanımlarında (GNW 2001) şeklinde gösterilecektir.

(1965), Lieblein (1952), Seth (1950). Bazı durumlarda sonuçlar tam olarak hesaplansa da hesaplamaların uzunluğu nedeniyle yaklaşık hesaplamalar önerilmiştir. Tam sonuçlar için Huntington and Naus (1975) ile Huffer and Lin (1997), asimptotik dağılımlar içinde Cressie (1980)'e başvurulabilir.

$P(k; N, w)$ olasılığını hesaplamak için binom dağılımı kullanılarak aşağıdaki basit yaklaşım önerilmiştir:

$$b(k; N, w) = \binom{N}{k} w^k (1-w)^{N-k}, \qquad (3.1)$$

$$G_b(k; N, w) = \sum_{i=k}^{N} b(i; N, w)$$

Bu durumda $P(k;N,w)$ olasılığına önerilen yaklaşım, $\approx$, hesaplamalardaki yaklaşıklığı göstermek üzere,

$$\begin{aligned} P(k; N, w) &\approx (N-k+1)\,b(k-1; N, w) - (N-k-1)\,b(k; N, w) + 2G_b(k+1; N, w) \\ &= (kw^{-1} - N - 1)\,b(k; N, w) + 2G_b(k; N, w) \qquad (3.2) \end{aligned}$$

dır. Bu yaklaşım $P(k;N,w) < 0.1$ için oldukça isabetli olup, $w \le 0.5$ ve $k > N/2$ değerleri için yaklaşık değil tam sonuçlar vermektedir.

Örnek: (HIV virüsü taşıyan diyaliz hastalarının grubu)

Bir diyaliz merkezinde yapılan araştırmada Ocak 1988'den Aralık 1993'e kadar olan süre içerisinde HIV virüsü taşıyan hastaları içeren geçmişe yönelik bir inceleme gerçekleştirilmiştir. Bu 72 aylık dönem boyunca HIV virüsü pozitif olan 13 hastadan 8

tanesinin 1992'nin son 6 ayında bu virüse sahip olduğu tespit edilmiş ve bu çalışma için tarama istatistikleri kullanılmıştır.

Burada, küme büyüklüğü k = 8, noktaların toplam sayısı N = 13 ve aralık uzunluğu $w = 6/72 = 0.0833$ olarak alındığında (3.2) formülü kullanılarak olasılık, $k > N/2$ olduğundan tam olarak

$$P(8;\ 13,\ 0.0833) = 0.00016$$

olarak bulunur. Söz konusu düzgün dağılım varsayımı altında gözlemler ve testin sonucuna göre bu kümelenmenin sıra dışı olduğu söylenmiştir $(0.00016 < 0.0002)$. Bulunan sıra dışı küme ve testin sonucu diyaliz merkezindeki incelemenin derinleşmesine ve işlemler yapılırken hastaların bu ve benzeri virüs almamalarını sağlamaya yönelik önlemlerin alınmasına yardımcı olmuştur (GNW 2001).

3.1.3 Çember üzerindeki tarama istatistikleri

Gezegen yörüngelerinin yaptıkları eğimlerin açıları veya kuş ya da böcek sürülerinin uçuşlarının yönleri araştırmacıları çember üzerindeki tarama istatistiklerinin tanımlanmasına yöneltmiştir. Aşağıdaki notasyonda indislerde yer alan "*c*", rasgele değişkenlerin çember üzerinde tanımlandığına işaret etmek üzere kullanılmıştır.

Verilen N nokta birim çemberin çevresinde bağımsız ve rasgele olarak dağılsın. S_w^c, w uzunluğundaki herhangi bir alt yay parçasında bulunan noktaların maksimum sayısını, W_k, k sayıda nokta içeren en küçük alt yayın uzunluğunu göstersin. Buna göre olasılık,

$$P(S_w^c \geq k) = P(W_k \leq w) = P_c(k;\ N,\ w)$$

ve

$$Q_c(k; N, w) = 1 - P_c(k; N, w)$$

dir.

Aşağıda küçük olasılıklar için basit bir yaklaşım verilmiştir. $b(k; N, w)$, binom olasılığıdır ve (3.1) formülünde tanımlanmıştır. Bu yaklaşıma göre,

$$P_c(k; N, w) \approx b(k; N, w)(k - Nw)/(w(1-w)) \qquad (3.3)$$

dir (GNW 2001).

Örnek: (Ergen intiharlarının mevsimsel kümelenmesi)

Bir araştırmada intihar sayılarındaki mevsimsel değişim incelendiğinde bu sayının bahar sonu yada yaz başında en yüksek noktasına çıktığı görülmüş, bir diğer araştırmada da en yüksek intihar oranının Mart, Nisan, Mayıs ve Haziran aylarında olduğu saptanmıştır.

Ergen (15-19 yaş) intiharlarının mümkün mevsimsel kümeleri 1978 ve 1979 yılları dikkate alınarak incelendiğinde, bu 2 yıllık dönem içerisinde 3474 ergen intiharı gözlenmiştir. 2 yılın her bir günündeki gözlemler birleştirilip veriler 1 Ocaktan 31 Aralığa kadar geçen bir yıllık zaman süresinde birim çember üzerinde gözlenmiş veriler olarak ele alınmış ve 91 günlük (3 ay) bir zaman, tarama aralığı olarak seçilmiştir. Yıl, birim çember olarak düşünüldüğünde tarama aralığının uzunluğu, w = 91/365 = 0.249 olarak dönüştürülerek w uzunluklu aralıklardaki intiharların maksimum sayısı 966 olarak bulunmuştur. Bu aralıklardan ilki 1 Ocaktan 1 Nisana kadar olan dönemi içermektedir.

w = 0.249, k = 966 ve N = 3474 değerleri (3.3) yaklaşımında yerine konulduğunda böyle bir durumla karşılaşma olasılığı,

$$P_c(966; 3474, 0.249) \approx (8.5 \times 10^{-6})(966 - 865)/(0.249(0.751)) = 0.0045$$

olarak bulunur. Bu olasılık anlamlılık düzeyi ile karşılaştırıldığında ergen intiharları kümesinin istatistiksel olarak önemli olduğu söylenebilir.

3.1.4 Tarama istatistiklerinin momentleri

Söz konusu tarama istatistiklerinin beklenen değerlerini bulmak tek başına araştırma konusu olabileceği gibi tarama istatistiklerinin dağılımlarına asimptotik yaklaşımlar için de gerekli olabilecektir.

Tarama istatistiği S_w'nin w ve N değerlerine bağlı olduğunu göstermek üzere bu istatistik $S_w(N)$ ile gösterilsin. $S_w(N)$'nin momentleri için bazı analitik formüllerin bulunduğu bilinmektedir. Formüllerdeki katsayılar, bu hesaplamalar için geliştirilen Neff and Naus (1980)'un katsayılar tablosundan alınmıştır (GNW 2001).

$100 < N < 1000$ arasındaki N değerleri için yaklaşımda kullanılan uygun model,

$$E(S_w(N)) = wN + b_w N^{0.5}$$

ve

$$V(S_w(N)) = a_w N + c_w N^{0.5}$$

dir.

En küçük aralık uzunluğu W_k'nın beklenen değeri ve varyansı $(N+1)/2 < k \leq N$ için,

$$E(W_k) = (k - 2(N - k + 1)b)/(N + 1)$$

ve

$$V(W_k) = (N - k + 1)((N + k + 1) + 2(2k - N - 1)b - 4(N + 2)(N - k + 1)b^2) \\ /(N + 1)^2 (N + 2)$$

dir.

Burada *b*, parametreleri $N - k + 1$ ve 0.5 olan binom olasılığıdır.

3.2 Zaman Boyutunda Olayların Geleceğe Dönük (Prospective) Taranması

Önceki kesimlerde belirli bir zaman aralığında yapılan gözlem sayısı veri olduğunda olasılık hesaplamalarının yapıldığı anlaşılmaktaydı. Gelecekteki belirli bir zaman aralığında tarama istatistiği tanımlamak durumunda kalındığında, bu aralıkta yapılabilecek gözlem sayısının da rasgele olacağı açıktır. Bu durum, koşulsuz tarama istatistikleri olarak sınıflandırılmaktadır.

3.2.1 Olayların Poisson dağılımı

Bu kesimde olayların oluş sayılarının Poisson sürecine uyduğu tarama istatistiklerinin dağılımı konu alınmaktadır. Burada, incelenen zaman aralığındaki olayların sayısına ait rasgele değişken, *N*, her bir birim zamandaki beklenen değeri λ olan Poisson dağılımlıdır. Araştırmacıların tarama istatistiklerini bu şekilde kullanmaları gelecekteki verileri değerlendirmek ya da örneğin bir telefon santralini uygun kapasitede tasarlamak amacıyla olmaktadır.

Poisson süreci, mekan ya da zaman içerisindeki olayların oluş sayılarıyla ile ilgilenen pek çok rasgele olgu için model olarak kullanılmıştır. Uygulamalarından bazıları telefon trafiği, kuyruk modeli problemlerini içermektedir.

Örnek: (Karbonmonoksit Zehirlenmesi)

Bir ilde 8 yıllık bir süre içerisinde birbirlerinden farklı zaman ve konumlarda gözlenen 19 karbon monoksit zehirlenmesi olayı rapor edilmiştir. Bir halk sağlığı çalışanı bu kadar zehirlenmenin sıra dışı olduğunu düşündüğü kümelenmeye dikkat etmiş ve gelecekteki bir 8 yıl içerisinde 8 yada daha fazla olayı kapsayan bir 1 yıllık dönemin olmasının ne kadar mümkün olabileceğini araştırmak istemiştir.

Bu durum, 19 olayın olduğu (koşul olarak verildiği) 8 yıllık dönem içerisindeki herhangi 1 yıllık dönemin 8 veya daha fazla sayıda olay içermesi olasılığı gibi düşünülebilir. Bu durumda yapılan çalışma, N olay sayısının belirli olduğu koşullu durumu içeren geçmişe dönük (retrospective) çalışma olmaktadır. İkinci durum ise 8 yılı içeren dönemde 19 olayın gözlenmesini koşul olarak alıp ileriye dönük (prospective), olası bir çalışma gerçekleştirilmesidir. Birinci durumdaki olasılık P ile, ikinci durumdaki olasılık ise P^* ile gösterilecektir. Bu örnekteki koşullu ve koşulsuz olasılık hesaplamaları neticesinde alınacak farklı kararlar pratikte farklı uygulamalara yön verebilecektir. Koşullu durum için olasılık, $P(8;19,1/8)=0.04811$ bulunurken, koşulsuz durumda olasılık, $P^*(8;19,1/8)=0.09293$ olarak hesaplanmaktadır. Anlamlılık düzeyini 0.05 seçen araştırmacılar P olasılığı ile olayı sıra dışı olarak niteleyecek, P^* olasılığına göre ise olayı sıradan bir olay olarak niteleyeceklerdir.

Aşağıda, $[0,T)$ zaman aralığında ileriye dönük tarama istatistiklerine ilişkin olasılık hesaplamalarının yaklaşık olarak yapılmasını sağlayan formüller sunulacaktır.

3.2.2 [0, T) aralığında koşulsuz tarama istatistiği

Bazı uygulamalarda [0,T) zaman aralığı içerisindeki olayların toplam sayısının sabit bir *N* sayısı olarak verildiği (bilindiği) tarama istatistiklerinin dağılımı yerine bu zaman aralığı içerisindeki olayların sayısının bir rasgele değişken olduğu dağılımlara ihtiyaç duyulur. Poisson süreci, bir zaman aralığı içerisinde gözlenen olay sayısının rasgele olduğu olgular için bir modeldir. Bu süreçte, λ parametresi herhangi bir birim aralıktaki olayların beklenen değerini gösterir. Herhangi bir [t, t+w] aralığındaki olayların sayısı $Y_t(w)$, Poisson dağılımlı olduğunda bu dağılıma ait beklenen değer λw olmaktadır. Bu durumda,

$$P(Y_t(w) = k) = e^{-\lambda w}(\lambda w)^k / k!, \qquad k = 0, 1, 2, \ldots$$

dır.

Bu durumda ayrık aralıklardaki olayların sayısına ait rasgele değişkenler de birbirlerinden bağımsız dağılımlıdırlar. Poisson süreci çeşitli biçimlerde karakterize edilebilmektedir. Poisson süreci için noktalar arasındaki varışlar arası zamanlara ait rasgele değişkenler birbirlerinden bağımsız ve üstel dağılımlıdırlar.

Zaman aralığında rasgele meydana gelen olaylar için, $T_{k,w}$, w uzunluğundaki bir aralıkta en az k olayı gözlediğimiz zamana kadar geçen bekleme süresi rasgele değişkenini göstersin.

$X_{(i+k-1)} - X_{(i)} \le w$ olduğunda en küçük i değeri için, $T_{k,w} = X_{(i+k-1)}$ olmaktadır. S_w, W_k, ve $T_{k,w}$ tarama istatistiklerinin dağılımları aşağıdaki gibi birbirleri ile ilgilidirler:

$$P(S_w \ge k) = P(W_k \le w) = P(T_{k,w} \le T)$$

Her bir birim zamandaki ortalaması λ olan Poisson süreci için olasılık, yani $P(S_w \geq k)$ olasılığı,

$$P^*(k; \lambda T, w/T) = 1 - Q^*(k; \lambda T, w/T)$$

olarak gösterilir (GNW 2001).

3.2.3 P*(k; λT, w/T) için yaklaşık formüller

Olasılığın hesaplanması için elde edilen asimptotik formül:

$$P^*(k; \lambda T, w/T) \approx 1 - \exp\left\{-\lambda^k w^{k-1} T/(k-1)!\right\} \qquad (3.4)$$

dır. Bu yaklaşım, P^* yeterince küçük olduğunda kullanışlı olan, kabaca bir yaklaşım verir fakat asimptotik yakınsaması çok yavaştır.

Oldukça doğru sonuçlar veren bir başka yaklaşım $L = T/w$ ve $\Psi = \lambda\ w$ olmak üzere,

$$P^*(k; \Psi L, 1/L) \approx 1 - Q_2^*(Q_3^*/Q_2^*)^{L-2} \qquad (3.5)$$

formülüdür (Naus 1982).

Burada,

$$Q_2^* = 1 - P^*(k; 2\Psi, 1/2) \text{ ve } Q_3^* = 1 - P^*(k; 3\Psi, 1/3)$$

$$Q^*(k; 2\Psi, 1/2) = (F_p(k-1; \Psi))^2 - (k-1)p(k; \Psi)p(k-2; \Psi) - (k-1-\Psi)p(k; \Psi)F_p(k-3; \Psi)$$

ve

$$Q^*(k; 3\Psi, 1/3) = (F_p(k-1; \Psi))^3 - A_1 + A_2 + A_3 - A_4$$

olup,

$$A_1 = 2\,p(k; \Psi)F_p(k-1; \Psi)\left\{(k-1)F_p(k-2; \Psi) - \Psi F_p\right\}$$
$$A_2 = 0.5(p(k; \Psi))^2\left\{(k-1)(k-2)F_p(k-3; \Psi) - 2(k-2)\Psi F_p(k-4; \Psi) + \Psi^2 F_p(k-5; \Psi)\right\}$$
$$A_3 = \sum_{r=1}^{k-1} p(2k-r; \Psi)(F_p(r-1; \Psi))^2$$
$$A_4 = \sum_{r=2}^{k-1} p(2k-r; \Psi)p(r; \Psi)\left\{(r-1)F_p(r-2; \Psi) - \Psi F_p(r-3; \Psi)\right\}$$

dır. Yaklaşım Alm (1983) tarafından daha da geliştirildiğinde,

$$P^*(k; \lambda T, w/T) \approx 1 - F_p(k-1; \lambda w)\exp\left\{-\left[(k - w\lambda/k)\right]\lambda(T-w)p(k-1; \lambda w)\right\}$$

olmaktadır (GNW 2001).

Yukarıdaki formül λT büyük, ek olarak w/T küçük olduğunda daha basittir.

Örnek: (Bir Sağlık Merkezindeki Vakaların Kümelenmesi)

Epidemiyolojistler ve halk sağlığı çalışanları sıklıkla, kanser olaylarının, intiharların, kazaların ya da diğer hastalıkların veya ölümlerin bir zaman aralığında kümelenmesinin nedenlerine açıklık getirmeye çalışırlar. Bu araştırmacılar olaylara neden olabilecek ortak faktörleri araştırırlar. Öyle bir faktör bulamadıklarında ise olayları tarayarak seçtikleri kümelenmiş vakaları inceleyip sonuç çıkarmaya çalışırlar. Yani sıra dışı

kümelenmeler rasgele oluşan kümelerden ayırt edilmeye çalışılır. Böyle kümeleri seçerken önsezileri kullanmak konu uzmanları için bile yanıltıcı olabilir.

Bir gazetede engelli olan bireylerin bulunduğu 400 yataklı bir sağlık kurumunda, 10 aylık bir zaman içerisinde 11 kişinin öldüğü yazılmıştır. Ölenlerin sayısı yaklaşık olarak beklenen oranın iki katı olduğundan bu ölüm oranı açıklama için yetkilileri harekete geçirecektir. Kimi tanıklar çeşitli nedenlerin varlığından söz etmektedirler.

3-5 yıllık dönem yeniden incelendiğinde, 10 aylık bir dönemde gözlenen 11 ölümün tamamen rasgele olarak gerçekleştiği söylenebilecektir. Kurumda, ortalama, bir 10 aylık dönem içerisinde 5.5 ölümün olduğu bilinirken, tamamen şans eseri olarak, bir 5 yıllık dönemin %38'inde 11 yada daha fazla ölümün olduğu 10 aylık dönemlerin olması beklenebilir. Burada, $k = 11$, $\Psi = 5.5$, $w/T = 10/60 = 1/L$ değerleri (3.5) formülünde yerine konarak olasılık, $P^*(11; 33; 1/6) \approx 0.38$ olarak bulunmuştur. Bu yüksek sayılabilecek olasılık ise söz konusu gözlemin sıra dışı olmadığına dair kuvvetli bir kanıt olarak değerlendirilebilecektir.

3.3 Denemelerin Bir Dizisindeki Başarı Sayılarının Taranması

3.3.1 Olay sayılarının binom dağılımı: Kesikli zaman, koşulsuz durum

Bir çok araştırmada her biri olası iki sonuca sahip rasgele denemelerin dizileri ile ilgilenilir. Bu iki olası sonuç başarı ve başarısızlık olarak adlandırılabilir. Bir kalite kontrol işleminde ardışık gözlemlerin kontrol sınırları içine ya da dışına düşmesi veya bir spor takımının bir sezon içerisinde kazanması ya da kaybetmesi gibi olaylar bu duruma örnek olarak verilebilir.

Kimi zaman araştırmacılar gözlem yapılan böyle bir süreçte değişiklik olup olmadığını belirlemeye çalışırlar. Süreçte bir değişiklik olmadığı bilindiğinde böyle bir süreç için kullanılan basit modelde rasgele deneylerin birbirlerinden bağımsız ve her bir

denemedeki başarı olasılığının sabit olduğu varsayılır. Varsayılan bir süreçteki değişikliklerin özel tiplerini test etmek için bazı istatistiksel ölçütler ve ilkeler geliştirilmiştir. Süreçteki değişiklik tiplerinden biri, sürecin bazı noktalarında başarı olasılıklarının artmış olmasıdır. $N = Lm$ deneme verildiğinde, L bir tamsayı olmak üzere; N denemeyi her birinde ardışık m denemenin olduğu L ayrık kümeye bölmek ve her bir kümedeki başarı sayılarını gözleyerek değişikliği belirlemek bunlardan biridir. Kümelerin herhangi birindeki başarı sayısı çok büyükse bu, varsayılan süreçteki bir değişikliğe işaret edebilir. Basit model varsayımı altında her m denemedeki başarı sayısı bir binom rasgele değişkenidir. Burada, L tane birbirinden bağımsız ve aynı dağılımlı binom rasgele değişkeninin en büyüğünün dağılımı ile ilgilenilir. Başarı sayısındaki bir değişikliğin anlamlılığını değerlendirmek için bütün deneyi dikkate alacak bir anlamlılık düzeyine ihtiyaç duyulacaktır.

Kullanılan bir diğer ilke ise yapılan işlemin N deneme içerisindeki m ardışık denemenin (ayrık yada birbirinin içine geçmiş) bütün kümelerindeki başarı sayılarını gözleyerek süreçteki değişkenliğin anlamlı olup olmadığının test edilmesidir. Bu amaçla, m uzunluklu aralıklarda bulunan denemelerdeki başarı sayısı sayılır. Herhangi bir aralıktaki bu sayı yeterince büyükse (anlamlıysa), bu, varsayılan süreçteki bir değişikliğe işaret eder. N deneme içerisinde bulunan herhangi bir m ardışık denemedeki maksimum başarı sayısı bir rasgele değişkendir ve *tarama istatistiği* adını alıp S'_m ile gösterilir. m, daha önceden sürekli zaman aralığının tarama uzunluğu w işlevindedir. $S'_m = m$ olduğu durum, özel durum olup N denemede birbirine komşu ya da biri diğerinin içine eklemlenmiş (ulanmış) bütün ayrık yada ayrık olmayan m uzunluklu ardışık deneme kümelerindeki maksimum başarı sayısının m olmasıdır. $S'_m = k$ olması ise m ardışık denemede en az k başarılı kota (quota) olarak ifade edilir.

Bazı uygulamalarda, verilen N sayıda denemedeki toplam başarı sayısı bir sabit olarak dikkate alınırken bazı uygulamalarda ise N deneme içindeki başarı sayısı bir rasgele değişken olup bir dağılıma sahip olduğu varsayılır. N denemedeki toplam başarı sayısının bilinen bir sabit olması durumu *koşullu* veya *geçmişe dönük (retrospective)*

durum, N deneme içerisindeki toplam başarı sayısının bir rasgele değişken olarak ele alındığı durum ise *koşulsuz* veya *ileriye dönük (prospective) durum* olarak adlandırılır.

3.3.2 İleriye dönük (koşulsuz) durum için basit bir model: Bernoulli süreci

$X_1, X_2, \ldots, X_N$, $P(X_i = 1) = p = 1 - P(X_i = 0)$ olmak üzere, bağımsız ve aynı dağılımlı kesikli rasgele değişkenler olsun. İlgili sürece de Bernoulli süreci denilsin. m bir tamsayı ve $i = 1, 2, \ldots, N - m + 1$ olmak üzere Y_i rasgele değişkeni,

$$Y_i = \sum_{j=i}^{i+m-1} X_j$$

olarak tanımlansın. Burada Y_i, X_j rasgele değişkenlerinin hareketli bir toplamıdır. Tarama istatistiği, S'_m ise bu hareketli toplamların maksimumudur. Herhangi bir m ardışık denemedeki maksimum başarı sayısı,

$$S'_m = \max_{1 \le i \le N-m+1} \{Y_i\}$$

olarak ifade edilir.

Tarama istatistiği ile ilişkili bir istatistik olan W'_k ise k başarı içeren ardışık en küçük m deneme sayısı olup,

$$W'_k = \min_{k \le m \le N} \{m : S'_m = k\}$$

dir.

Verilen bir Bernoulli sürecinde, $T_{k,m}$, m uzunluğundaki bir aralıkta en az k tane başarının ilk kez gözlemlendiği zamana kadar geçen süreye ilişkin rasgele değişkendir. Üç istatistik, S'_m, W'_k ve $T_{k,m}$ birbirleri ile ilişkilidir.

$$P(S'_m \geq k) = P(W'_k \leq m) = P(T_{k\,m} \leq N)$$

Bernoulli süreci için genel olasılık $P'(k\,;m,N,p)$ olarak gösterilir.

Bu süreç için tanımlanan dördüncü bir istatistik V_r ise ardışık olarak en fazla r başarısızlığın gözlemlendiği deneme sayısı rasgele değişkenidir. Özel olarak $r=0$ olduğunda V_r, ardışık en uzun başarı sayısı olur. N deneme içeren bir dizi için V_r istatistiği, S'_m ile aşağıdaki gibi ilgilidir. Bu ilgi,

$$P(V_r \geq k+r) = P(S'_{k+r} \geq k) = P'(k\,;k+r,N,p)$$

olarak ifade edilir.

Söz konusu Bernoulli sürecine ilişkin yaklaşımlar verildikten sonra bazı uygulamalara örnekler verilecektir.

3.3.3 P^*(k; m, N, p) olasılıklarının hesaplanması

Huntington and Naus (1975) aşağıdaki olasılıkları yaklaşıma gerek duyulmaksızın tam olarak hesaplayabilmektedirler. Hesaplamaların uzunluğu yaklaşık hesaplamaları gerekli kılmıştır. Koşulsuz olan bu olasılıklar için başarı sayısının dağılımı duruma göre poisson veya binom olabilmektedir. Sunulan yaklaşım Naus (1982) tarafından Teorem 2 olarak verilmiştir.

$Q'(k\,;m,N,p)=1-P'(k\,;m,N,p)$ ve $Q'(k\,;m,\ Lm\,,p)$ ifadesinin kısaltması Q'_L olsun. Q'_L için hayli doğru sonuçlar veren bir yaklaşım,

$$Q'_L \approx Q'_2\{Q'_3/Q'_2\}^{((N/m)-2)} \tag{3.6}$$

dir.

$$b(k\,;m,\,p)=\binom{m}{k}p^k(1-p)^{m-k},$$

ve

$$\begin{aligned} F_b(r\,;s\,,p) &= \sum_{i=0}^{r} b(i\,;s\,,p), \qquad r=0,1,\ldots,s,\\ &= 0, \qquad\qquad\qquad r<0. \end{aligned}$$

olmak üzere, $2<k<N$ ve $0<p<1$ için

$$\begin{aligned} Q'_2 &= (F_p(k-1;m,\,p))^2-(k-1)b(k\,;m,\,p)F_b(k-2\,;m,\,p)\\ &\quad + mpb(k\,;m,\,p)F_b(k-3\,;m-1,\,p), \end{aligned}$$

ve

$$Q'_3=(F_p(k-1;m,\,p))^3-A_1+A_2+A_3-A_4$$

olmaktadır.

Burada,

$$A_1 = 2b(k\,;\,m\,,\,p) - F_b(k-1\,;\,m\,,\,p)\{(k-1)\,F_b(k-2\,;\,m\,,\,p) - mp\,F_b(k-3\,;\,m\,,\,p)\}$$

$$A_1 = 0.5\,(b(k\,;\,m\,,\,p))^2\{(k-1)(k-2)\,F_b(k-3;\,m\,,\,p) - 2(k-2)mp\,F_b(k-4;\,m-1,\,p) + m(m-1)\,p^2 F_b(k-5;\,m-2\,,\,p)\}$$

$$A_3 = \sum_{r=1}^{k-1} b(2k-r\,;\,m\,,\,p)\,(F_b(r-1;\,m\,,\,p))^2$$

$$A_4 = \sum_{r=2}^{k-1} b(2k-r\,;\,m\,,\,p)\,b(r\,;\,m\,,\,p)\{(r-1)\,F_b(r-2\,;\,m\,,\,p) - mp\,F_b(r-3\,;\,m-1,\,p)\}$$

dır.

$r=0$ için V_0 rasgele değişkeni, sıfır başarısızlığın gerçekleştiği ardışık en uzun deneme sayısı rasgele değişkenidir ve $P(V_0 \geq m) = P'(m\,;\,m,\,N,\,p)$ dir. Bu olasılığın hesaplanmasına ilişkin ardışık hesaplama bağıntısı,

$$P(V_0 \geq m\,;\,N+1) = P(V_0 \geq m\,;\,N) + (1-p)\,p^m\,(1 - P\{V_0 \geq m\,;\,N-m\,\}) \quad (3.7)$$

olup burada,

$$P(V_0 \geq m\,;\,N) = \sum_{j=1}^{[N/m]} (-1)^{j+1}\{p + ((N-jm+1)q/j)\}\binom{N-jm}{i-1} p^{jm} q^{j-1} \quad (3.8)$$

ile tam hesaplama yapılabilir.

(3.8) eşitliğinde $q = 1-p$ ve $[y]$ tam değeri göstermektedir. (3.7) ve (3.8) formülleri hilesiz bir paranın 200 kez atılması sonucunda ardışık olarak en az k tane tura elde edilmesi olasılığını hesaplamak için kullanılabilir. Pratikte pek çok durum için N sayısı çok büyük, m sayısı çok küçük olmaktadır. Örnek olarak, hilesiz bir paranın 200 000

kez atılışında en az 25 turanın ardı ardına gelmesi olasılığı bulunmak istendiğinde (3.8) formülü, $[N/m]=8000$ terimin toplamını içerir. Büyük N değerleri için çeşitli yaklaşık hesaplama formülleri geliştirilmiştir. Buna göre, yukarıdaki Bernoulli deneyleri için,

$$P'(m;\ m,\ N,\ p) \approx 1-\exp\{-Nqp^m\}$$

yaklaşımı kullanılabilir ve önceki verilen yaklaşım da

$$P(V_0 \geq m;\ N) = 1-Q'(m;\ m,\ N,\ p) \approx 1-Q_2'\{Q_3'/Q_2'\}^{((N/m)-2)} \qquad (3.9)$$

olarak kullanılabilir.

Burada,

$$Q_2' = 1-P'(m;\ m,\ 2m,\ p) = 1-p^m(1+mq)$$
$$Q_3' = 1-P'(m;\ m,\ 3m,\ p) = 1-p^m(1+2mq)+.5\,p^{2m}(2mq+m(m-1)q^2)$$

olmaktadır.

Örnek: (Ardışık en uzun başarı sayısının dağılımı)

Madeni para ile yapılan bir sınıf deneyinde öğrencilerin bir kısmı parayı 200 kez atmış ve sonuçları kaydetmişlerdir. Diğer kısmı ise para atmayıp bir simülasyon yaparcasına bir dizi oluşturmuşlardır. Bu iki dizinin ilk 50 terimi aşağıda gösterilmiştir. İlk dizi hilesiz paranın atılışından elde edilen dizi, ikinci dizi ise diğer yolla elde edilmiş olan dizidir.

Dizi 1: YTYTTYYYTTYTY**TTTTTTTT**YYYTTYYTYYYTYY
YTYYTTYTTYTYTY

Dizi 2: TTYTYYYTYYYYTTYTYYYTYY**TTTTT**YYTYTYTYY
YYTTYYTTYYTTTY

Birinci dizide dikkat çeken 8 turadan oluşan en uzun ardışık başarının varlığıdır. Yani ilk dizide tura ya da yazıdan oluşan en uzun ardışık başarı 8 tura içermektedir. İkinci dizideki en uzun ardışık başarı ise 5 turadan oluşmaktadır. İlk dizideki ardışık en uzun başarı sayısının 8 olması sıra dışı olarak uzun mudur ya da ikinci dizideki ardışık en uzun başarı sayısının 5 olması sıra dışı olarak kısa mıdır sorularını irdeleyelim:

Yukarıdaki (3.9) formülü uygulanarak turaların en uzun ardışık başarı sayısına ilişkin olasılık sorularına yaklaşık cevap bulunabilir. En uzun ardışık tura ya da yazı gözlenme sayısının dağılımını yaklaşık olarak bulabilmek için T ve Y harflerinden oluşmuş bir dizide, eğer harf bir önceki ile aynı ise A, farklı ise F harflerini yazarak A ve F harflerinden oluşan yeni bir dizi yaratılabilir. Böyle bir kurgu ile N denemeden oluşan orijinal dizideki k tane ardışık başarı sayısı, $N-1$ denemeden oluşan yeni dizide $k-1$ ardışık başarı sayısı olur. Böylece, 200 denemeden oluşan orijinal bir dizide k uzunluklu herhangi bir ardışık başarı sayısının elde edilmesi olasılığı, 199 denemeden oluşan yeni dizideki $k-1$ uzunluğuna sahip ardışık tura sayısının elde edilmesi olasılığına eşit olmaktadır.

Çizelge 3.1 Hilesiz bir zarın 200 kez atılışına ilişkin en uzun ardışık tura sayısının olasılık dağılımı

k	*P(en uzun ardışık gözlenen tura sayısı ≥ k)*
5	0.97
6	0.80
7	0.54
8	0.32
9	0.17
10	0.09

(3.9) formülü kullanılarak hesaplanmış çizelge 3.1'de ardışık olarak gözlenen en uzun başarı sayısının beşten az olması olasılığının sadece 0.03 olasılığa sahip olduğunu görebiliriz. Ardışık en uzun tura sayısının 8'den küçük olması olasılığı ise 0.68'dir. Böylece dizi 2'de bulunan ardışık en uzun tura gözlenme sayısı sıra dışı kısalıktadır denilebilir. Dizi 1'deki ardışık en uzun tura sayısının 8 olması ise sıra dışı değildir.

Üç veya daha fazla sonuçlu olaylarda olasılıkların hesaplanmasına örnek olarak amino asitlerin yük problemleri verilebilir. Amino asitlerin bir zinciri olarak tanımlanabilecek protein dizileri üzerinde çalışan bilim adamları, belirli yük dizilişleri ile proteinlerin yapısal özellikleri ve fonksiyonlarının ortaklığı arasındaki bağlantıyı araştırma konusu almaktadırlar. Amino asitlerin bazıları pozitif yüke sahipken bazıları da negatif yüklü, bazıları da yüksüzdür. Bu açıdan bakıldığında protein, amino asitlerin yüklerinin dizilişine göre, -1, +1 ve 0'ların bir dizisi olarak düşünülebilir.

$X_1, X_2, \ldots, X_N$ bağımsız ve aynı dağılımlı kesikli rasgele değişkenlerin bir dizisi olsun. $P(X_i = -1) = p_{-1}, P(X_i = 0) = p_0, P(X_i = 1) = p_1$, m bir tamsayı ve $t = 1, 2, \ldots, N - m + 1$ olmak üzere $Y_t(m)$ rasgele değişkeni, X_j rasgele değişkenlerinin hareketli bir toplamı,

$$Y_t(m) = \sum_{i=t}^{t+m-1} X_j$$

olarak tanımlansın.

Tarama istatistiği S'_m'de bu hareketli toplamların maksimumu

$$S'_m = S'_m(N) = \max_{1 \le t \le N-m+1} \{Y_t(m)\}$$

olacaktır.

$G_{k,m}(N)$ olasılığı,

$$G_{k,m}(N) = P(S'_m(N) < k)$$

olarak tanımlandığında bu olasılığı yaklaşık olarak veren formül (3.5)'deki ile aynıdır.

$$G_{k,m}(T) \approx G_{k,m}(2m)\{G_{k,m}(3m)/G_{k,m}(2m)\}^{((N/m)-2)}$$

dir. Konuyla ilgili bir problem Karwe and Naus (1997)'da yer almaktadır (GNW 2001).

3.3.4 Olay sayılarının binom dağılımı: Kesikli zaman, koşullu durum

Bölümün ilk kısmında N denemenin birbirinden bağımsız ve N denemedeki toplam başarı sayısının bir rasgele değişken olarak ele alındığı model üzerinde durulmuştu. Bu durum koşulsuz yada ileriye dönük durum olarak adlandırılmıştı.

Bazı uygulamalarda toplam başarı sayısı bilinen bir değer olarak karşımıza çıkar. Bu durum ise koşullu veya geçmişe dönük durum olarak adlandırılır. N denemede a başarı olduğu bilindiğinde herhangi bir m ardışık denemede gözlenen başarıların maksimum sayısına *tarama istatistiği* adı verilir ve S'_m ile gösterilir. $S'_m \geq k$ olduğu genel durum, m ardışık deneme içerisinde en az k başarının gerçekleştiği durumdur. Bu bölümde, N deneme içerisinde tam olarak a başarının olduğu verildiğinde S'_m tarama istatistiğinin dağılımına ait olasılıkların hesaplanması ve uygulamalarına yer verilecektir. S'_m tarama istatistiği, a başarının ve $N-a$ başarısızlığın bütün $\binom{N}{a}$ dizilerinin eşit olasılığa sahip olduğu basit olasılık modeli için hesaplanır. Bu durum için $P(S'_m \geq k)$ olasılığı $P(k\,;m,N,a)$ ile gösterilmektedir.

$k > a/2$ ve $N/m = L$ (L bir tamsayı) olduğunda bu olasılık tam olarak

$$P(k\,;m,N,a) = 2\sum_{s=k}^{a} H(s,a,m,N) + (Lk-a-1)\,H(k,a,m,N) \qquad (3.10)$$

formülü ile hesaplanabilmektedir.

Burada,

$$H(s,a,m,N) = \binom{m}{s}\binom{N-m}{a-s} \Big/ \binom{N}{a}$$

olup, hipergeometrik olasılık fonksiyonudur.

Örnek: (Genelleştirilmiş bir doğum günü problemi)

Klasik doğum günü problemi şu soruya yanıt arar: 23 kişiden oluşan bir grup içerisinde aynı doğum gününe sahip (en az) iki kişinin bulunması olasılığı nedir? Çoğu insan bunun pek mümkün olamayacağını düşünür. Fakat gerçekte bu olasılık %50'den daha fazladır.

Yılın $N = 365$ gününde doğumun eşit olasılığa sahip olduğu varsayıldığında verilen a insan için eşleşme olmama olasılığı $(N!/(N-a)!)/N^a$ dır. $a = 23, N = 365$ değerleri için en az bir eşleşmenin var olması olasılığı yaklaşık 0.51'dir. Problem ve çözüm için Mosteller (1987)'e bakılabilir.

Doğum günü problemlerinin bir diğer çözümü ise herhangi ardışık m gün içinde iki doğum gününün olmaması olasılığını bulmaktır. Bu durumda, N günlük dönemin doğru yada çember üzerinde konumlandığı düşünülebilir. Aşağıdaki formüllerde $\Re_c$, çembersel N günlük dönem için, $\Re$ doğru üzerinde konumlanmış günler için ardışık doğum günlerinin çakışmaması olasılıklarını vermektedir. Buna göre,

$$\Re_c = (N - am + a - 1)!/(N - am)!N^{a-1}, \qquad N \geq am \qquad (3.11)$$

ve

$$\Re = (N - (a-1)(m-1)!/(N - (a-1)(m-1) - a)!N^a, \qquad N \geq (a-1)(m-1)$$

dir. Aşağıdaki örnek GNW (2001)'den alınmıştır.

Örnek:

Anne, baba ve iki çocuktan oluşan bir aile verilsin. Bu ailedeki bireylere ait dört doğum günü ve bir de anne ve babanın evlenme tarihlerinin birbirlerinden bağımsız oldukları varsayıldığında beş tarihten ikisinin ardışık yedi günlük dönem içerisine düşmesi olasılığı nedir sorusuna cevap aranabilir. Yukarıdaki (3.11) formülüne göre $N = 365, a = 5, m = 7$ olarak alındığında olasılık 0.31 olarak bulunur. Buradan beş tarihten ikisinin yedi günlük bir dönem içerisine düşmesinin sıra dışı olmadığı sonucuna varılabilir.

Örnek: (Satrançtaki galibiyetlerin kümesi)

Bir satranç ustası 1 yıl içinde 20 turnuvada oynamış ve dokuz tanesini kazanmıştır. Kazanılan turnuvalardan yedi tanesi 10 ardışık turnuvada gerçekleşmiştir. Dokuz galibiyetin birbirinden bağımsız ve 20 turnuva içerisinde tamamen rasgele dağıldığı varsayıldığında 10 ardışık turnuvada en az yedi galibiyetin gerçekleşmesi olasılığı nedir sorusuna yanıt aranabilir. Burada, $a = 9, N = 20, k = 7$ ve $m = 10$ değerleri alındığında formül (3.10)'a göre olasılık,

$$P(7\,;10\,,20\,,9) = 2\sum_{s=7}^{9} H(s\,,9\,,10\,,20) + (14-9-1)\,H(7\,,9\,,10\,,20) = 0.20$$

olarak hesaplanır. 20 turnuva içinde dokuz galibiyet olduğu bilindiğinde 10 ardışık turnuvada yedi galibiyetin gerçekleşmesi sıra dışı değildir sonucuna ulaşılır.

3.3.5 r harflik bir dizide herhangi bir harfin ardışık en uzun tekrar sayısı

N bağımsız denemeden oluşan bir dizideki her biri eşit olasılığa sahip r harften herhangi birinin en uzun tekrar sayısının dağılımı bulunmak istenebilir. Suman (1994) tarafından da çalışılmış olan bu problem $N-1$ Bernoulli denemesindeki en uzun başarı

sayısı ile ilişkilendirilebilir (GNW 2001). r harften oluşan bir alfabenin dizisinde her bir harfin kullanılması olasılığının $p = 1/r$ olduğu varsayılsın,

ζ_r; N bağımsız denemeden oluşan bir dizide, her biri eşit olasılığa sahip r harfli bir alfabeden herhangi bir harfin ardışık en uzun tekrar sayısı olmak üzere beklenen değer ve varyans,

$$E(\zeta_r) \approx +0.5 + \log_e \{(N-1)(r-1)/r)\} + \gamma) / \log_e r \qquad (3.12)$$

ve

$$V(\zeta_r) \approx \left\{(\pi^2)/2(\log_e r)^2\right\} + (1/12)$$

dir. Burada $\gamma = 0.577\ldots$ dir (Euler sabiti).

N Bernoulli denemesindeki en uzun başarı tekrarının beklenen sayısı her bir harfin kullanılması olasılığı $p = 1/r$ olmak üzere,

$$E(V_0) \approx -0.5 + \log_e \{N(r-1)/r)\} + \gamma) / \log_e r,$$

ve

$V(V_0) \approx V(\zeta_r)$ olmaktadır.

3.3.6 Tarama istatistiklerinin beklenen değerleri

Tarama istatistiklerinin çeşitli derecelerden beklenen değerlerini bilmek bu istatistiklere ait olasılıklara üst sınırlar oluşturmak veya bu olasılıklara yaklaşımda bulunmak bakımından önemlidir. GNW (2001)'de tarama istatistiklerinin dağılımına normal

dağılım yaklaşımı olarak $E(S_w)$ ve $V(S_w)$ 'nin kullanımı örnek verilmiştir. Huffer and Lin (1997) değişik dereceden beklenen değerleri kullanarak hem Markov zincirleri hem de poisson yaklaşımları elde etmişlerdir. Eğer dağılım var ve yaklaşık olarak hesaplamalar yapılabiliyorsa beklenen değerlere de yaklaşımlarda bulunulabilir.

Maksimum sayıda başarı içeren kümedeki deneme sayısının (S'_m) beklenen değeri ve varyansı, $(0, 1)$ aralığındaki düzgün dağılım varsayımı altında aşağıdaki gibi hesaplanabilir (GNW 2001):

$$E(S'_m) = \sum_{k=1}^{\infty} P(S'_m \geq k)$$

(K_1, K_2) göreli olarak dar bir aralık olmak üzere; $P(S'_m \geq k)$ ifadesi, $k < K_1$ için bire, $k > K_2$ için sıfıra yakın olur. Bu aralık, pm 'den oldukça büyük değerler denenerek bulunabilir. Örnek olarak, $N = 10\,000$, $p = 0.1$ ve $m = 59$ için (3.6) formülü $P(S'_m \geq k)$ için aşağıdaki değerleri verir:

k	12	13	14	15	16	17	18	19	20
$P(S'_{59} \geq k)$	1.00	0.980	0.799	0.449	0.183	0.060	0.017	0.0046	0.0011

(3.12) kullanılarak, $E(S'_m) = 12 + (0.980 + 0.799 + \ldots + 0.0011) = 17.7$ olarak bulunur.

$$P(S'_m = k) = P(S'_m \geq k) - P(S'_m \geq k+1)$$

eşitliği kullanılarak $E\left((S'_m)^2\right)$ hesaplandıktan sonra

$$V(S'_m) = E\left\{(S'_m)^2\right\} - \left\{E(S'_m)\right\}^2$$

olarak hesaplanabilir.

Varyans hesaplamasına basit bir yaklaşım olarak aşağıdaki formül de kullanılabilir:

$$V(S'_m) = 2\sum_{k=1}^{\infty} k\, P(S'_m \geq k) - E(S'_m)(1 + E(S'_m))$$

Yine, (0, 1) aralığındaki düzgün dağılım varsayımı altında k başarı içeren en dar aralığın (W'_k) beklenen değeri ve varyansı,

$$E\,(W'_k) = \sum_{m=1}^{\infty} P\,(W'_k \geq m\,) \qquad (3.13)$$

$$V(W'_k) = 2\sum_{k=1}^{\infty} m\, P(W'_k \geq m) - E(W'_k)(1 + E(W'_k)) \qquad (3.14)$$

olmaktadır. Örnek olarak; $p = 0.3, k = 5, N = 100$ alındığında olasılıklar aşağıdaki gibi hesaplanmıştır:

m	5	6	7	8	9	10	11	12
$P(W'_k \geq m)$	1.00	0.8474	0.5799	0.3324	0.1692	0.0806	0.0372	0.0171

m	13	14	15	16	17	18	19
$P(W'_k \geq m)$	0.0080	0.0038	0.0019	0.0009	0.0005	0.0003	0.0001

Buradan (3.13) ve (3.14) formülleri kullanılarak beklenen değer ve varyans,

$$E(W'_k) = 5 + 0.8474 + 0.5799 + \cdots + 0.0001 \cong 7.0784$$

$$V(W_k') = 2(15 + 6(0.8474) + 7(0.5799) + \cdots + 19(0.0001)) - (7.0784)(8.0784) \cong 2.74$$

olarak bulunur (GNW 2001).

Benzer olarak m uzunluklu aralıklarda k tane başarı gözlemek için gerekli bekleme zamanı (deneme sayısı) $(T_{k,m})$'nin beklenen değeri de hesaplanabilir. Bunun için, $P(S_m' \geq k)$ olasılığının N üzerinden sonsuz toplamını alınır ve (3.6)'da yer alan formül kullanılır. Buna göre bekleme zamanının beklenen değerini veren geometrik toplam:

$$E(T_{k,m}) = \sum_{N=0}^{\infty}(1 - P(S_m' < k)) \approx 2m + Q_2'(1 - (Q_3'/Q_2')^{1/m})$$

olur.

3.4 İki ve Daha Yüksek Boyutlu Taramalar

Daha önceki bölümlerde olayların zaman içindeki kümelenmeleri ya da denemelerin bir dizisi üzerinde tek boyutlu tarama istatistiklerinin kullanımı ele alınmıştı. Burada ise iki ya da daha yüksek boyutta tarama istatistikleri ve uygulamaları ele alınacaktır. Bu alandaki ilk çalışmalar Mack (1948) tarafından yapılmıştır.

Pek çok alanda araştırmacılar, sıra dışı kümelenmeler için iki ya da daha çok boyutlu taramalar yapmaktadırlar. Örneğin, bir epidemiyolojist kanser vakalarının coğrafi konumda kümelenmelerini belirlemeye çalışabilir. Bir jeolog belirli türden maden kaynaklarının kümelendiği yerleri bulmak için belirlenen bölgeyi, bir astrofizikçi ise gama ışını yayan patlamaların yoğun olduğu kaynakları belirlemek için gökyüzünü tarayabilir. İki boyutlu taramalar; tıbbi görüntüleme, maden arama ve sistem güvenilirliği gibi konularda sıkça kullanılmaktadır.

Tarama istatistiklerinin geniş bir çoğunluğu iki yada daha yüksek boyut için geliştirilmiştir. Bir boyutlu taramalarda, kesikli deneme dizilerinde ya da sürekli bir zaman aralığının taranması için tarama penceresi olarak bir aralık kullanılır. İki boyutlu taramalarda ise tarama penceresi olarak kare, dikdörtgen, daire, üçgen ve diğer şekiller kullanılabilmektedir. Taranan iki boyutlu bölgeler, dikdörtgensel bölgeleri, bir küre yüzeyini ya da daha genel olarak belli bir şekli olmayan coğrafik alanları içerir. Bu tür tarama istatistiklerinin dağılımları taranan bölgelerin yapısına göre elde edilmiştir. Sürekli yapıya sahip tarama bölgeleri için düzgün, poisson v.s. dağılımları; kafes yapıda (lattice) olanlar için binom, hipergeometrik v.s. dağılımları kullanılır. İki ya da daha yüksek boyutlu tarama problemlerinde probleme iki tür yaklaşımda bulunulabilir. İlki bir boyutlu tarama istatistiklerinin iki yada daha yüksek boyutlu duruma uyarlanmasıdır. Diğeri de doğrudan iki boyutlu yapı korunarak iki boyutlu tarama istatistiklerinin kullanılması durumudur.

İki boyutlu tarama istatistiklerinde N nokta, şekli verilen iki boyutlu bölgede rasgele olarak konumlanır. Bir boyutlu tarama istatistiklerindeki birim aralığın iki boyutlu tarama istatistiklerindeki genelleşmiş hali birim kare veya küre yüzeyidir. Tarama istatistiği S_w, şekli verilen w çaplı (boyutlarında) herhangi bir pencerede bulunan noktaların maksimum sayısı olur. W_k ise k nokta içeren iki boyutlu en küçük tarama penceresinin "çapıdır". S_w ile W_k istatistiklerinin dağılımları birbirleriyle ilintilidir.

İki boyutlu bir bölge taranırken şekli verilen tarama penceresindeki en büyük kümelenme araştırılır. Bu durumda problem, kenarları karenin kenarlarına paralel olan v yüksekliğinde ve u genişliğinde bir alt dikdörtgenle birim karenin taranması şeklinde düşünülür. Koşullu durumda birim kare üzerine dağılmış sabit N sayıda nokta mevcuttur. Koşulsuz durumda ise birim kare üzerine dağılmış noktaların sayısı Poisson rasgele değişkeni olarak düşünülür. Birim karenin taranması, $S \times T$ boyutundaki dikdörtgensel bir bölgenin $a \times b$ boyutundaki bir tarama penceresi ile taranması demektir. Burada, x ekseni üzerinde $S = 1$ ve y ekseni üzerinde $T = 1$ birim olarak alınarak, $u = a / S$ ve $v = b / T$ dönüşümleri yapılır.

3.4.1 Koşullu durum

Verilen N noktanın birim kare üzerinde rasgele dağıldığı bilindiğinde $S_{u,v}$; kenarları birim karenin kenarlarına paralel, v yüksekliğinde ve u genişliğindeki herhangi bir alt dikdörtgende bulunan noktaların maksimum sayısını gösterir. $P(k\,;N\,,u\,,v)$ biçiminde gösterilen $P(S_{u,v} \geq k)$ olasılığı, kenarları karenin kenarlarına paralel, en az k nokta içeren $a \times b$ boyutlu en az bir tarama alt dikdörtgeninin olasılığını ifade etmektedir. Bu olasılık Naus (1965) tarafından,

$$P(k\,;N\,,u\,,v) \cong k^2 \binom{N}{k} (uv)^{k-1}$$

şeklinde ifade edilmiştir.

$k = N$ olduğu durumda olasılığın alt sınırı üst sınırına eşittir. Bu durumda formül,

$$P(N\,;N\,,u\,,v) = P(N\,;N\,,u)\;P(N\,;N\,,v)$$

olur.

$u \leq 0.5$ ve $v \leq 0.5$ olarak verildiğinde $k = N-1$ için bir boyutlu tarama istatistiğinin olasılığı yardımı ile hesaplanan formül,

$$\begin{aligned} P(N-1;N\,,u\,,v) = {} & P(N\,;N\,,u)\;P(N-1;N\,,v) \\ & + \{P(N-1;N\,,u) - B_{12}\}P(N-1;N-1,v) \\ & + \{B_{12} - P(N\,;N\,,u)\}\{2P(N-1;N-1,v) - P(N\,;N\,,v) \\ & - (2(N\,v^{N-1} - (N+1)v^N)/N(N-1))\} \end{aligned} \tag{3.15}$$

olarak yazılır.

Burada, $u \le 0.5$ ve $v \le 0.5$ olmak üzere,

$$B_{12} = 2N u^{N-1}(1-u),$$
$$P(N;\, N,\, u) = N u^{N-1} - (N-1)u^{N},$$
$$P(N-1;\, N-1,\, v) = (N-1)v^{N-2} - (N-2)v^{N-1},$$
$$P(N-1;\, N,\, u) = 2u^{N} + N(N-1)u^{N-2}(1-u)^{2}$$

dır.

Kuyruk olasılıklarına ilişkin bir teori kullanılarak küçük olasılıklı olduğu düşünülen olayların olasılığını hesaplamak için verilen bir diğer yaklaşım ise,

$$P(k;\, N,\, u,\, v) \cong (\{N^2 w(1-u)(1-v)E^3/(1-w)^3(1-E)\} + C)\, b(k;\, N,\, w) \quad (3.16)$$

şeklindedir. Burada,

$$b(k;\, N,\, w) = \binom{N}{k} w^k (1-w)^{N-k}$$
$$w = uv$$
$$E = (k/Nw) - 1$$
$$C = \{Nv(1-u)E/(1-w)\} + \{Nu(1-v)E^2/(1+E)(1-w)^2\} + \{(1+E)(1-w)/E\}$$

olmaktadır.

Örnek: (Kanser olaylarının kümesi)

Bir bölgedeki kanser olaylarının rasgele olan dağılımının düzgün dağılıma uymadığı, bunun yerine hem rasgeleliğin hem de düzenli olmayan yerleşim nedeniyle kümelenmeler oluşturduğu anlaşılmaktadır.

Böyle bir araştırmada bir grup bilim adamı, 1993 yılının sonuna kadar olan 20 yıllık bir dönem içerisinde, İsveç'te tanı konulmuş 15 yaşın altındaki akut lösemi hastası çocukların kümesini araştırmıştır. Verilen dönem içerisinde tüm İsveç'te 1.703.235 çocuk arasından 1543 akut lösemi hastası tespit edilmiş, bunların 133 tanesinin İsveç'in güneybatısında bulunan Okome'de olduğu gözlenmiştir. Bölgedeki oranın İsveç için ortalama oranın yaklaşık 25 katı olduğu bilinmektedir. Bu kümelenmenin sıra dışı olup olmadığı araştırma kapsamındadır. Bölgedeki populasyon düzgün dağılmadığından bir tür yaklaşımla haritaya düzgün dağılımı verecek şekilde dönüştürme işlemleri yapılmıştır. İlk olarak İsveç haritası her bir karede bir kişi olduğu varsayılarak 1.703.235 kareye bölünür. Kolaylık olması açısından, harita 1305×1305 boyutundaki (1305, 1.703.235'in karekökü) bir kare şeklinde düşünülmüş olup Okome'nin populasyonu ise kabaca 11.5×11.5 boyutundaki (11.5, 133'ün karekökü) bir alt karenin içine düşmektedir. Bütün kare içerisindeki 1534 lösemi hastası çocuk göz önüne alındığında, kenarları $u = v = 11.5/1305$ olan bir alt kare içerisinde üç vakanın görülmesi olasılığı nedir sorusu araştırılır. Yani istenen olasılık $P(3; 1534, 0.0088, 0.0088)$ olasılığıdır. (3.16) yaklaşımı kullanılarak hesaplanan olasılık değeri 0.697 olarak bulunur ve 133 akut lösemi hastası çocuktan oluşan bir populasyon içerisinde gözlenen üç akut lösemi hastası çocuğun oluşturduğu kümenin sıra dışı olmadığı sonucuna varılır.

3.4.2 Tarama penceresinin şeklinin etkisi

Vakaların incelenmesi için bir bölge taranırken, tarama penceresinin şeklinin kümelenme olasılığı üzerinde bir fark yaratıp yaratmadığı araştırılmak istenebilir. Örnek olarak; $P(N; N, u, v)$ olasılığının tam hesaplama olasılığı göz önüne alınarak aynı alana sahip biri kare diğeri dikdörtgen iki tarama penceresi kullanıldığında $k = N = 5,\ u = 0.1$ ve $v = 0.9$ değerleri için olasılık $P(5; 5, 0.1, 0.9) = 0.00042$ bulunurken, $u = v = 0.3$ değerleri için $P(5; 5, 0.3, 0.3) = 0.00095$ olarak bulunur. $P(N; N, u, v)$ olasılığı $u = v$ olduğunda maksimum sonuca ulaşır. Tam formül (3.15) kullanıldığında; $P(4; 5, 0.15, 0.4) = 0.00962$, $P(4; 5, 0.2, 0.3) = 0.01018$ ve $P(4; 5, 0.2449489, 0.2449489) = 0.01029$ olarak bulunur. Buna göre, kare tarama

penceresi dikdörtgen tarama penceresine göre kümelenme olasılığını biraz daha yükseltmektedir. Hatta aynı şekillerin farklı yönlendirilmeleriyle yapılan taramalarda da farklı sonuçlar elde edilebileceği gözlemlenmiştir.

Tarama istatistikleri üzerinde çalışanlar ayrıca tarama penceresi hangi şekilden olursa olsun iki boyutlu taramanın simülasyonunu hesaplamada etkili bir algoritmanın olması gerektiğini farketmişlerdir. Bunun üzerine iki boyutlu taramalarda tarama olasılığına yaklaşımda bulunma amaçlı bir tür simülasyon algoritması geliştirilmiştir.

3.4.3 Koşulsuz durum

Birim kare içerisindeki noktaların sayısı λ ortalamalı bir Poisson rasgele değişkeni ve birim kare $u \times v$ boyutundaki, kenarları birim karenin kenarlarına paralel bir alt dikdörtgen ile taranıyor olsun. $P^*(k;\lambda,u,v)$, alt dikdörtgenlerden en az birinin en az k gözlem bulundurması olasılığını göstersin. Buna göre bu olasılık,

$$P^*(k;\lambda,u,v) \cong 1-\exp\{-k\lambda(1-u)(1-v)p(k-1;\lambda uv)\}$$

şeklinde verilmektedir.

Burada,

$$p(k-1;\lambda uv) = \exp\{-\lambda uv\}(\lambda uv)^{k-1}/(k-1)!$$

dir.

Verilen diğer bir yaklaşım ise;

$$P^*(k;\lambda,u,v) \cong 1-\exp\left\{-(k-1)^2\,\lambda(1-u)(1-v)\,p(k-1;\lambda uv)/k\right\}$$

olup Alm (1999) tarafından verilen daha iyi bir yaklaşım

$$P^*(k;\lambda,u,v) \cong 1-F_p(k-1;\lambda uv)\exp\left\{-\zeta-(1-(\lambda uv/k))\lambda v(1-u)\,p(k-1;\lambda uv)\right\}$$

şeklindedir (GNW 2001).

Burada,

$$F_p(k-1;\lambda uv)=\sum_{i=0}^{k-1}p(i;\lambda uv)$$

ve

$$\zeta=(1-(\lambda uv/k))\lambda u(1-v)\left\{P^*(k-1;\lambda,v,u)-P^*(k;\lambda,v,u)\right\} \qquad (3.17)$$

olmaktadır. $P^*(k;\lambda,v,u)$ bir boyutlu tarama istatistiğidir. Bu konuda verilen bir başka yaklaşımda,

$$P^*(k;\lambda,u,v) \cong 1-\left\{1-P^*(k-1;\lambda,v,u)\right\}\exp(-\zeta) \qquad (3.18)$$

dir. Yukarıdaki formülde ζ, (3.17) formülünde tanımlandığı gibidir.

Örnek:

$u=v=1/30,\ \lambda uv=5,\ k=19,\ \lambda=4500$ olsun. $P^*(19;4500,1/30,1/30)$ olasılığı bulunmak istenebilir. $P^*(18;150,1/30)=0.00151927$ ve

$P^*(19;150,1/30) = 0.000430589$ olasılıkları (3.5) formülüne göre hesaplandığında formül (3.17) ve (3.18)'e göre,

$$\zeta = (1-(5/19))150(29/30)\{P^*(18;\lambda\ v,u) - P^*(19;\lambda\ v,u)\} = 0.11632$$

ve

$$P^*(19;4500\ ,1/30,1/30) \cong 1-\{1-P^*(18;150,1/30)\}\exp(-\zeta) = 0.11$$

olarak bulunur.

Yukarıdaki yaklaşımlar dikdörtgensel bir bölgenin yine dikdörtgen bir tarama penceresi ile taranması durumlarında kullanılır. $S \times T$ dikdörtgensel bölgenin r yarıçaplı çember şeklinde bir tarama penceresi ile taranması durumunda, $r\pi^{0.5}/S = u$, $r\pi^{0.5}/T = v$ olarak alınır ve X ile Y eksenleri üzerinde yine $S = 1$ ve $T = 1$ alınarak işlem yapılır. $P_c^*(k;\lambda\ ,u,v)$; λuv, r yarıçaplı bir çemberdeki noktaların beklenen sayısını göstermek üzere, tarama penceresindeki noktaların maksimum sayısının en az k olması olasılığıdır. Buna göre bu olasılık,

$$P_c^*(k;\lambda\ ,u,v) \cong 1-\exp\{-k\lambda(1-2u)(1-2v)\,p(k-1;\lambda uv)\}$$

dir.

3.5 DNA ve Protein Dizilerinin Analizinde Tarama İstatistiklerinin Kullanımı

Pek çok alanda çalışan bilim adamları birçok biyolojik kaynaktan DNA ya da protein dizilerini karşılaştırmaktadır. Biyolojik sürecin kontrol edildiği genetik kodları içeren DNA (deoksiribonükleik asit) uzun bir moleküldür. DNA molekül modeli sarmal şekilde kıvrılmış, merdivene benzer bir yapıdadır. Bu çift polinükleotid zincir, birbirine

bazlar arasındaki zayıf hidrojen bağlarıyla bağlanır ve birbirinin tamamlayıcısıdır. DNA molekülünün yapısına katılan organik bazlar adenin, sitozin, guanin ve timin olarak adlandırılır ve sırasıyla A, C, G, T harfleri ile gösterilir. Hidrojen bağlarıyla bağlı iki ipliğin arasındaki baz çiftlerinden bir iplikteki A bazına diğer iplikteki T, bir iplikteki C bazına ise diğer iplikteki G bazı karşılık gelir. Adenin ve guaninden oluşan bazlara "pürin bazları", sitozin, timin ve urasilden (RNA'nın yapısında bulunur) oluşan bazlara ise "pirimidin bazları" adı verilir. Bir iplikteki baz dizilimi bilindiğinde diğer (tamamlayıcı) iplikteki baz dizilimi de bilinmiş olur (Nolan and Speed 2000).

Protein sentezi hücre içinde meydana gelen en önemli olaylardan biridir. Proteinlerin yapı taşları amino asitlerdir. DNA, yapısındaki dört çeşit nükleotid (adenin guanin, sitozin ve timin nükleotidleri) yardımıyla canlı yapısındaki 20 çeşit amino asitin protein molekülündeki sırasını ve sayısını şifreleyebilir. DNA'nın 20 çeşit amino asiti protein yapısına yerleştirebilmesi için en az üçlü nükleotid dizilişini kendi arasında değiştirmesi gerekir. Böylece protein sentezi sırasında kullanılan 64 çeşit üçlü şifre ortaya çıkar. Bu üçlü baz dizilimine "kodon" adı verilir. Bu şifrelerden bazıları başlatma ve durdurma emrini veren kodonlardır. Bazı amino asitlere birden fazla kodon karşılık gelir (Kurşungeçmez, 2005).

DNA ve protein dizileri analizinde baz dizilimi 4 harfli alfabe gibi düşünülebilir. Alfabe 4 harften oluşabileceği gibi (A, G, C, T), 20 amino asitte 20 harfli bir alfabe gibi düşünülebilir. Alfabe, (A, G) pürin bazına karşı (T, C) pirimidin bazının gelmesi gibi iki harfli alfabenin değişik bir türü de olabilir. Belli amino asitlerin pozitif yüklü, diğerlerinin negatif yüklü ya da yüksüz olduğu bilindiğine göre alfabe yükleri göstermek üzere üç harften de oluşabilmektedir. Bu bölüm harflerin tek boyutlu dizilerine tarama istatistiklerinin uygulanmasını ve kelimelerin (ardışık harflerin oluşturdukları örüntüler) araştırılmasını ele almıştır.

3.5.1 DNA ya da protein dizilerindeki örüntü kümelerinin taranması

DNA ya da protein dizileri üzerinde çalışan bilim adamları belirli örüntü tiplerini, net yükleri ya da dizideki olay kümelerini belirlemeye çalışmaktadırlar. Bahsedilen örüntü tipleri biyolojik olarak önemli etkilere sahiptir ya da öyle olduğu varsayılır. Bu örüntülerden bazıları DNA'nın onarım ve çoğalması-kopyalanması ile ilgilidir. Bu tür örüntülerin yer aldığı yerlere DAM bölgeleri denir.

Örnek: (E. Coli DNA'sı içindeki DAM bölgelerinin kümeleri)

E. Coli (bir bakteri cinsi) DNA'sı içindeki 4.7 milyon harf bir zincir olarak görülebilir. Bu durumda DNA; A, C, G, T harflerinden oluşmuş dört harfli bir alfabenin doğrusal bir dizisi şeklinde düşünülür. Araştırmada dizideki GATC örüntüsünün oluşumu araştırılmakta olsun. Bu örüntü DNA'nın kopyalanması ve onarılmasında düzenleyici role sahiptir. Dizide bu örüntünün meydana geldiği noktalar DAM bölgesi olacaktır. Bir E. Coli genom (4.7 milyon harften oluşan dizi) dizisinde 250 harfte ortalama 1.1 dolayında DAM bölge olduğu tahmin edilmektedir. Çalışmalar esnasında dizinin 245 ardışık harften oluşan belirli bir bölümünde sekiz DAM bölgesi gözlemlenmiştir. Araştırmacıların cevaplamaya çalıştıkları soru "4.7 milyon harften oluşmuş bir dizinin herhangi bir yerinde bulunan 245 harfin sekiz DAM bölgesi içermesi sıra dışı mıdır?" sorusudur. 250 harfte bulunan DAM bölgesi sayısının, 1.1 ortalama ile yaklaşık olarak Poisson dağıldığı varsayıldığında, zaman uzunluğunu $T = 4.7$ milyon, tarama uzunluğunu (aralığını) $w = 245,\ k = 8$ ve $\lambda = (1.1/250) = 0.0044$ alınarak (3.4) formülünde yerine konulduğunda istenen olasılık,

$$P*(k = 8;\ \lambda\ ,\ w/T) \approx 1 - \exp\{-0.0044^{k}(245^{k-1})4{,}700{,}000/(k-1)!\} \cong 0.999$$

olarak bulunur. Daha isabetli bir yaklaşım kullanılarak $P^*(k, \Psi L, 1/L) \approx 1 - Q_2^*(Q_3^*/Q_2^*)^{L-2}$ formülü ile $P*(k = 8;\ \lambda\ ,\ w/T) \approx 0.66$ olarak hesaplanır. Bu sonuca göre, E. Coli genomunun 4.7 milyon harften oluşan dizinin

herhangi bir yerinde bulunan 245 ardışık harfin sekiz DAM bölgesi içermesi sıra dışı değildir sonucuna ulaşılır.

Örnek: (İnsanda bulunan Stomegalovirüsün (HCMV) üremesinin incelenmesi için palindrom kümelerinin araştırılması)

Stomegalovirüs (HCMV) ile savaşmak için bilim adamları bu virüsün nasıl çoğaldığını araştırmak üzere virüs DNA'sının kopyalanmasıyla ilişkili örüntüleri incelemektedirler. Bu virüs pek çok insanın vücudunda bulunmakla beraber, virüsün sadece kopyalanma evresine girdiğinde aktif hale geçtiği bilinmektedir. Kopyalanma süreci, belirli başlatıcı proteinler virüs DNA'sının belirli işaret alt dizilerine eklemlendiğinde başlar. Bu alt dizileri içeren bölge virüsün kopyalanma kaynağı olarak adlandırılır.

Herpes virüsü ile ilgili yapılan önceki çalışmalarda kopyalanma kaynaklarının söz konusu DNA'nın simetrik ve tekrarlı örüntüleri ile ilgili olduğu gözlenmiştir. Bu virüs DNA'sının kopyalanma kaynağının uzun bir palindrom olduğu düşünülmektedir.

Genel kullanımı ile bir palindrom, tersten ve düzden okunuşu aynı olan kelime, rakam ya da sözcük öbeğidir (TAKAT, ATA, ÜTÜ, TİRİT, 7557 gibi). DNA dizilerinin analizinde bir palindrom biraz farklı tanımlanır. Bir palindrom örüntünün (PLP) özelliği tersten okunduğu zaman düzden okunan dizinin tamamlayıcısı olmasıdır. Örneğin, CCACGTGG sekiz bazlı bir palindromdur. İki tamamlayıcı iplikteki tamamlayıcı palindromlar:

CCACGTGG
GGTGCACC

şeklindedir. Palindrom örüntüsü her iki doğrultuda da aynı görünümdedir.

Bir “dyad”, palindromun iki simetrik kısmı arasına bazı geçiş bazlarının girmesi ile oluşan palindromik bir örüntüdür. Örnek olarak, CCACtatGTGG verilebilir. Eğer geçiş bazlarının sayısı çok büyük değilse (150’den küçükse) bunlar kısa dyadlar olarak adlandırılır.

İlk olarak en az 10 harf genişliğindeki palindrom örüntüsünü (PLP*) araştırmak üzere tümü 229.354 baz çiftinden oluşan HCMV dizisi taranmıştır. Harf uzunluğunun 10 olarak seçilmesinin nedeni bütün başlama pozisyonlarının kabaca 0.001’inin 10 harf uzunluğundaki bir PLP ile başlamasıdır. 4 harfli bir alfabenin her harfinin eşit olasılıklı olarak kullanıldığı varsayıldığında yukarıdaki gibi 10 harfli bir palindromun olasılığı $(1/4)^5$ olarak bulunur. Araştırma sürecinde palindromların kümelendiği bölgenin belirlenmesi için 1000 bazlı bir aralık ile tüm dizi taranmış ve belirli bir bölgede bu palindromlardan 10 tanesinin yer aldığı bir kümelenme bulunmuştur. Cevaplanması gereken soru, böyle bir kümelenmenin rasgeleliğin bir sonucu oluşup oluşmadığı yani sıradan olup olmadığıdır. Formül (3.5) kullanılarak, $T = 229.354$, $w = 1000$, $N = 296$ olarak belirlenip istenen olasılık, $P(S_w \geq 10) = P(10; 296, 1000) \approx 0.0017$ olarak hesaplanır. Bu sonuca göre palindromların yer aldığı bu kümenin anlamlı olduğu söylenebilir. Bu problem ile ilgili daha detaylı bir inceleme Nolan and Speed (2000)’de yer almaktadır.

3.5.2 DNA dizilerinde eşleştirme

Moleküler biyolojinin gelişmesi ile birlikte çeşitli biyolojik kaynaktan DNA dizilerinin karşılaştırılması büyük önem kazanmıştır. Karşılaştırılan dizilerde genetik materyalin benzerliği ya da DNA’nın söz konusu fonksiyonlarının ortaklığı araştırılmaktadır.

Bu amaçla iki uzun DNA dizisini taramak üzere bilgisayar algoritmaları geliştirilmiştir. Bu algoritmalar sayesinde bölümler ve alt diziler arasındaki mükemmel ya da mükemmele yakın eşleşmeler araştırılmaktadır. İki dizinin eşleştirilmesi sürecinde tamamen rasgeleliğin sonucu bazı eşleşmelerin saptanması beklenebilir.

Bazı uygulamalarda yeni bulunan bir dizinin ardışık bölümleri (contig) eşleşme yapılmak üzere gen bankasındaki ya da diğer gen havuzlarında bulunan daha önceden belirlenmiş dizilerle karşılaştırılır. Eşleşme; mükemmel, belirli sayıda eşleşmeyen kısmın bulunduğu mükemmele yakın ve ekleme/iptal (insertion/deletion=indel) biçimlerinde olabilir. Bu uygulamalarda eşleştirme veri bankasındaki göreli olarak kısa bölümler arasında ya da kısa bölümlerle daha uzun diziler arasında yapılır.

Sıra dışı eşleşmelerin tahmininde kullanılan mevcut asimptotik formüllerin çoğu büyük örneklem yaklaşımına dayandırılmıştır ve bu formüllerin yakınsaması yavaş olabilir. Çoklu eşleştirmelerin yapıldığı durum için bireysel olarak uyum yüksek olsa bile bu formüller duyarlılıklarını yitirebilirler. Eşleşme olasılıklarını hesaplarken doğru yaklaşımı kullanmak faydalı bilgiler elde edilmesini sağlar.

Dizi analizlerindeki önemli noktalardan biride dizilerin karşılaştırılma biçimleridir. İlk durumda eşleştirilmek istenen dizilerin aynı pozisyonlarındaki eşit sayıda harf içeren bölümler alt alta gelecek şekilde çakıştırılır ve bu iki dizide eşleşen ortak alt diziler saptanır. İkinci durumda ise iki dizinin ya da aynı dizinin farklı uzunluktaki bölümleri karşılaştırılır ve eşleşen bölümler araştırılır. Bu durum çakışmasız durumdur. İki durumda da sıra dışı uzun eşleşmeler aranır. Bütün karşılaştırma durumları için mükemmel ya da mükemmele yakın eşleşmelerin olasılıkları çeşitli olasılık modelleri altında doğru bir şekilde tahmin edilmeye çalışılır.

İki dizinin karşılaştırılma sürecinde eşleşen bölümlerin tespiti için 0-1 puanlama sistemi kullanılır (Bu sistemde 0 eşleşme yok, 1 eşleşme var anlamındadır). Bu metotla karşılaştırılan dizilerdeki harflerin benzerliklerine göre 0 ve 1 puanları ile oluşturulmuş bir dizi elde edilir. Daha sonra bu yeni dizi sabit w uzunluklu bir aralık ile taranarak bu aralıklardaki puanların hareketli bir toplamı hesaplanır (GNW 2001, Lange 2002).

Karşılaştırılan N uzunluğundaki iki dizi için, 0-1 puanlama sistemine göre oluşturulmuş dizide m uzunluğundaki herhangi bir aralıkta bulunan sayıların toplamının maksimumu tarama istatistiği S'_m'dir.

Bu puanlama sisteminde çakıştırılan dizilerde mükemmele yakın eşleşmenin olduğu durumda, eşleşen harflerden oluşan sıra dışı uzunluktaki ardışık harf dizisi (kelime) araştırılmak istenebilir. Bu durumda koşulsuz kesikli tarama istatistiği yaklaşımı kullanılır.

DNA dizi analizlerinde 0-1 puanlama sistemi dışında 0-1-2 puanlama sistemi de kullanılmaktadır. Ayrıca benzerlik puanlama sistemi oluşturulurken puanların 0 ve 1 in dışında değer aldığı durumlar da vardır.

Aşağıda ele alınan bir örnek üzerinde rasgeleliğin üç değişik modeli altında eşleşme problemine tarama istatistiklerinin uygulaması yapılmıştır. Örnek, GNW (2001)'den aktarılan bir çalışmadır.

Örnek: (Çakışmış iki amino asit dizisi arasındaki mükemmele yakın eşleşme)

Buğday	A	A	A	C	C	G	G	G	C	A	C	T	A	C	T	T	T	G	A	G	A	C	G	T	G	A
Th. Aquat	A	A	T	C	C	C	C	C	G	T	G	C	C	C	T	T	A	G	C	G	G	C	G	T	G	G
Puan	1	1	0	1	1	0	0	0	0	0	0	0	0	1	1	1	0	1	0	1	0	1	1	1	1	0

Yukarıda iki farklı bitkiye ait r-RNA dizisinin 26 harf içeren bölümleri karşılaştırılmaktadır. Çakıştırılmış harflerin altına eğer eşleşme olmuşsa 1, olmamışsa 0 rakamı konulmuştur. Puanların oluşturduğu dizide en fazla bir eşleşmemenin olduğu en uzun ardışık harf dizisi altı harflidir (altı çizili). Mükemmele yakın eşleşmenin olduğu bu durumda (en fazla 1 eşleşmeme) altı uzunluğunda ardışık bir harf dizisi elde etme olasılığı nedir sorusuna cevap aranacaktır. Bu sorunun cevabı seçilen rasgele modele bağlıdır.

Model 1.

Her biri eşit olasılıkla kullanılan 4 harfin oluşturduğu bir alfabe ve harflerin birbirinden bağımsız olarak kullanıldığı bir harfler dizisinin oluşturuluyor olduğu varsayılsın. O zaman yukarıdaki sıfır ve birlerin dizisi 26 Bernoulli denemesinin sonucu olur. Burada her bir harfin kullanılma olasılığı 0.25 olmaktadır. Her bir aralıktaki başarı sayılarının bir rasgele değişken olduğu varsayıldığında yukarıdaki sorunun cevabı koşulsuz kesikli tarama istatistiği yardımıyla bulunur. Böylece istenen olasılık, $P'(5\,;6,N=26,\,p=0.25)\cong 0.055$ olarak hesaplanır.

Model 2.

Harflerin eşit olasılıkla kullanılmadıkları 4 harfin oluşturduğu bir alfabe ve harflerin birbirinden bağımsız olarak kullanıldığı bir harf dizisinin oluşturuluyor olduğu varsayılsın. Gözlenen sıklıklar kullanılarak her bir dizideki A, C, G, T harflerinin olasılıkları tahmin edilir. Önce birinci bitki için olasılıklar; $P_{A1}=5/26,\,P_{C1}=6/26,\,P_{G1}=7/26,\,P_{T1}=5/26$ ve sonra ikinci bitki için olasılıklar $P_{A2}=3/26,\,P_{C2}=10/26,\,P_{G2}=8/26,\,P_{T2}=5/26$ olarak ayrı ayrı tahmin edilir. Yukarıdaki gibi iki dizinin herhangi bir konumunda herhangi bir harfin eşleşme olasılığı $p=(P_{A1}P_{A2}+P_{C1}P_{C2}+P_{G1}P_{G2}+P_{T1}P_{T2})=0.244$ olur. Model 1'de olduğu gibi her bir aralıktaki başarı sayılarının bir rasgele değişken olduğu varsayıldığında yine koşulsuz kesikli tarama istatistiğine ait yaklaşık hesaplama formülü kullanılarak olasılık, $P'(5\,;6,N=26,\,p=0.244)\cong 0.050$ olarak bulunur.

Model 3.

Yine 4 harfin oluşturduğu ve harflerin birbirlerinden bağımsız olarak kullanıldığı dizilerin oluşturulduğu varsayılsın. Harflerin kullanılma olasılıklarının eşit ve harflerin eşleştirilmeleriyle oluşturulan Bernoulli dizisinde 13 adet başarının olduğu bilinsin. Böylece 26 denemede 13 başarının verildiği koşulu altında bu olasılık, koşullu kesikli

tarama istatistiğine ait hesaplama formülü ile $P'(k\,;m\,,N,a) = P'(5\,;6\,,26,13) = 0.71$ olarak bulunur.

Bu sonuçlara göre böyle bir dizide, en fazla bir eşleşmemenin olduğu, altı uzunluğunda ardışık harf dizisi elde etmek sıra dışı değildir sonucuna varılır. Eğer başarı olasılığı $p = 13/26 = 0.50$ olan koşulsuz tarama istatistiği kullanılsaydı olasılık 0.63 olarak hesaplanacaktı.

İkiden fazla DNA dizisinin eşleşmesi problemlerinde de tarama istatistikleri kullanılır. Ancak başarı olasılıklarının hesaplamaları da buna paralel olarak değişecektir.

4. TARTIŞMA VE SONUÇ

Bu çalışmada, günümüzde pek çok alanda sıklıkla kullanılmaya başlanan tarama istatistiklerinin temel tanım ve kavramları, hesaplama yöntemleri ve uygulama alanlarından örnekler verilerek tarama istatistiklerinin tanıtımı amaçlanmıştır.

Çalışmanın ilk bölümünde ardışık sıra istatistikleri ve ardışık sıra istatistikleri arasındaki farklarla oluşan aralıkların (spacings) tanımları verildikten sonra sıra istatistiklerinin ve sıra istatistikleri arasındaki farklarla oluşan aralıkların dağılımları ele alınmıştır. Bu bölümde n tane rasgele değişkenin alındığı yığının dağılımının düzgün ve üstel olmaması durumlarında karşılaşılan hesaplama zorluklarına dikkat çekilmiştir.

Üçüncü bölümde, önceki bölümde değinilmiş olan sıra istatistikleri ile bağlantılı olan tarama istatistiklerinin temel tanım ve kavramları verildikten sonra geçmişe dönük (retroptective) ve geleceğe dönük (prospective) tarama kavramları üzerinde durularak bu taramalara ait istatistiklerin hesaplama yöntemleri örneklerle gösterilmiştir. Burada P, geçmişe dönük yapılan taramalara ait olasılıkları, P^* ise geleceğe dönük olanları ifade etmektedir. Tarama istatistikleri sadece sürekli bir zaman aralığında tanımlanmamış olduğundan ilerleyen bölümde ardışık deneme dizileri üzerinde tanımlanan tarama istatistikleri konusu ele alınmıştır. Ayrıca çalışmada iki ve daha yüksek boyuttaki tarama istatistiklerinin tanımı yapılarak uygulamalarına örnekler verilmiştir. Son kısım ise son yıllarda gelişen moleküler biyoloji sayesinde önem kazanan DNA ve protein dizileri analizlerinde tarama istatistiklerinin kullanımı konusunu içermekte olup değişik dizilerin ya da aynı dizinin farklı bölümlerinin eşleştirilmesi sonucu ortaya çıkan ortak bölgelerin analizinde tarama istatistiklerinin nasıl ve ne şekilde kullanılacağı örneklerle gösterilmeye çalışılmıştır.

KAYNAKLAR

Cressie, N. 1980. The asymptotic distribution of the scan statistics under uniformity. Annals of Probability, 8(4); 838-840.

Glaz, J., Naus, J. and Wallentein, S. 2001. Scan statistics. Springer-Verlag Inc., 367 p., New York.

Günay, S. ve İnal, C. 1993. Olasılık ve Matematiksel İstatistik. H.Ü. Fen Fakültesi Basımevi, 519 s., Ankara.

Huffer, F.W. and Lin, C.T. 1997. Approximation the distribution of the scan statistic using moments of the number of clumps. Journal of american Statistical Association, 92(440); 1466-1475.

Huntington, R.J. and Naus, J.I. (1975). A simpler expression for k th nearest neighbor coincidence probabilities. Annals of Probability, 3(5); 894-896.

Kurşungeçmez, S. 2005. Biyoloji. Aydoğdu Yayınları, 448 s., Ankara.

Lange, K. 2002. Mathematical and Statistical Methods for Genetic Analysis. Springer-Verlag Inc., 285 p., New York.

Lieblein, J. 1952. Properties of certain statistics involving the closest pair in a sample of three observations. Journal of Research of the National Bureau of Standarts, 48(3); 255-268.

Mack, M.A. 1948. An exact formula for $Q_k(n)$, the probable number of *k-aggregates* in a random distribution of n points. The London, Edinburg and Dublin Philosophical Magazine and Journal of Science, 39(297); 778-790

Miller, E.G. 2003. A new class of entropy estimators for multi-dimensional densities. International Conference on Acoustics, Speech, and Signal processing, 3, 297-300.

Mosteller, F. 1987. Fifty Challenging Problems in Probability with Solutions. Dover Publications Inc., New York.

Naus, J.I. 1965. The distribution of the size of the maximum cluster of points on a line. Journal of American Statistical Association, 60, 532-538.

Naus, J.I. 1982. Approximation for distributions of scan statistics, Journal of the American Statistical Association, 77, 177-183.

Nolan, D. and Speed, T. 2000. Stat Labs (Mathematical Statatistics Through Applications). Springer-Verlag Inc., 282 p., New York.

Pyke, R. 1965. Spacings. Journal of the Royal Statistical Society, Series B, 27(3); 395-449.

Seth, G.R. 1950. On the distribution of the two closest among a set of three observations. Annals of Mathematical Statistics, 21(2); 298-301.

Venter, J.H. 1967. On estimation of the mode. Annals of Mathematical Statistics, 38, 1446-1455.

ÖZGEÇMİŞ

Adı Soyadı : Fürüzan KÖKTÜRK
Doğum Yeri : Çaycuma/ZONGULDAK
Doğum Tarihi : 09.11.1973
Medeni Hali : Bekar
Yabancı Dili : İngilizce

Eğitim Durumu (Kurum ve Yıl)

Lise : Çaycuma Lisesi Matematik Bölümü (1988-1991)
Lisans : Hacettepe Üniversitesi Fen Fakültesi İstatistik Bölümü (1992-1996)
Yüksek Lisans : Ankara Üniversitesi Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı (Şubat 2004-Temmuz 2007)

Çalıştığı Kurum/Kurumlar ve Yıl

Türkiye İstatistik Kurumu (02-05-2007)
İz Yapım Reklam Tasarım ve Organizasyon Tic. Ltd. Şti. (2000-2002)
Kano Basım ve Tanıtım Ltd. Şti. (1997-2000)