ANKARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

DERİNLİKLERE DAYALI ÇOK DEĞİŞKENLİ TARAMA İSTATİSTİKLERİ ÜZERİNE BİR ÇALIŞMA

Elmas Burcu MAMAK EKİNCİ

İSTATİSTİK ANABİLİM DALI

ANKARA
2011

Her hakkı saklıdır

TEZ ONAYI

Elmas Burcu MAMAK EKİNCİ tarafından hazırlanan "**Derinliklere Dayalı Çok Değişkenli Tarama İstatistikleri Üzerine Bir Çalışma**" adlı tez çalışması 14.03.2011 tarihinde aşağıdaki jüri tarafından oy birliği ile Ankara Üniversitesi İstatistik Anabilim Dalı'nda **DOKTORA TEZİ** olarak kabul edilmiştir.

Danışman : Yrd.Doç.Dr. İhsan KARABULUT
Ankara Üniversitesi İstatistik Anabilim Dalı

Jüri Üyeleri :

Başkan: Prof.Dr. M.Akif BAKIR
Gazi Üniversitesi İstatistik Anabilim Dalı

Üye : Prof.Dr. Hamza GAMGAM
Gazi Üniversitesi İstatistik Anabilim Dalı

Üye : Prof.Dr. Fikri ÖZTÜRK
Ankara Üniversitesi İstatistik Anabilim Dalı

Üye : Doç.Dr. Halil AYDOĞDU
Ankara Üniversitesi İstatistik Anabilim Dalı

Üye : Yrd.Doç.Dr. İhsan KARABULUT
Ankara Üniversitesi İstatistik Anabilim Dalı

Yukarıdaki sonucu onaylarım

Prof. Dr. Özer KOLSARICI

Enstitü Müdürü

ÖZET

Doktora Tezi

DERİNLİKLERE DAYALI ÇOK DEĞİŞKENLİ TARAMA İSTATİSTİKLERİ ÜZERİNE BİR ÇALIŞMA

Elmas Burcu MAMAK EKİNCİ

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Yrd. Doç. Dr İhsan KARABULUT

Derinlik kavramı ile tanımlanan çok değişkenli fark istatistikleri temelinde çok değişkenli tarama istatistikleri önerilmiş ve uygulamalara değinilmiştir. Tek boyutlu tarama istatistiklerinin fark istatistiklerine bağlı olarak tanımlanmasına paralel olarak çok değişkenli tarama istatistikleri de çok değişkenli fark istatistikleri aracılığı ile tanımlanmıştır. Çok değişkenli tarama istatistiklerine ait dağılım, düzgün dağılım varsayımı altında elde edilmeye çalışılmıştır. Eldeki örneklem için çok değişkenli düzgün dağılım varsayımı altında derinlik konturlarına karşılık gelen olasılıklar (0,1) aralığında konumlandırılmıştır. Bu yaklaşımda tarama penceresinin şekli doğal olarak yığının geometrik yapısını yansıtan derinlik halkaları tarafından belirlenir.

Çalışmada istatistiksel bir derinlik fonksiyonu olan yarı uzay derinliği kullanılmıştır. Derinliklere dayalı geriye dönük tarama iki boyutlu dağılımlar ile sınırlandırılmıştır.

Mart 2011, 63 sayfa

Anahtar Kelimler: Derinlik Fonksiyonu, Çok Değişkenli Tarama İstatistikleri, Çok Değişkenli Fark İstatistikleri, Yarı Uzay Derinliği, Düzgün Dağılım

ABSTRACT

Ph. D. Thesis

A STUDY ON MULTIVARIATE SCAN STATISTICS BASED ON DEPTH

Elmas Burcu MAMAK EKİNCİ

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

Supervisor: Asst. Prof. Dr. İhsan KARABULUT

Multivariate scan statistics have been proposed by multivariate spacings which are based on depth and applications have been mentioned. In addition to the spacings based on the explanation of one-dimensional scan statistics, multivariate scan statistics have been also defined through multivariate spacings. The study aims to obtain the distribution of multivariate scan statistics based on data depth underlying uniform distribution. For the sample, the probabilities that correspond to depth contours are located in $(0,1)$ interval under the assumption of multivariate uniform distribution. In this approach, the shape of the scanning window is determined by the depth rings which reflect the geometrical structure of the population.

In this study, half space depth is used as a statistical depth function. Retrospective scanning based on depth is restricted by two dimensional distributions.

March 2011, 63 pages

Key Words: Depth Function, Multivariate Scan Statistics, Multivariate Spacings, Half Space Depth, Uniform Distribution

TEŞEKKÜR

Çalışmalarımı yönlendiren, tez çalışmamın her aşamasında bilgi, öneri ve yardımlarını esirgemeyerek bu tezin ortaya çıkmasını sağlayan, akademik gelişmemde üzerimde çok büyük emeği olan ve manevi desteğini hiç esirgemeyen danışmanım sayın Yrd.Doç.Dr. İhsan KARABULUT (Ankara Üniversitesi İstatistik Anabilim Dalı)'a, tez çalışması boyunca tez izleme komitelerinde değerleri görüş ve düşüncelerini esirgemeyen, bilgi ve birikimlerini paylaşan sayın Prof. Dr. Fikri ÖZTÜRK (Ankara Üniversitesi İstatistik Anabilim Dalı) ve sayın Prof. Dr. Hamza GAMGAM (Gazi Üniversitesi İstatistik Anabilim Dalı)'a, çalışmamızda katkısı olan sayın Doç. Dr. Halil AYDOĞDU (Ankara Üniversitesi İstatistik Anabilim Dalı)'ya, sayın Prof. Dr. M. Akif BAKIR (Gazi Üniversitesi İstatistik Anabilim Dalı)'a, Ankara Üniversitesi İstatistik Bölümü'nün değerli hocalarına, çalışmalarım süresince engin hoşgörüsü ile her konuda desteğini ve yardımlarını esirgemeyen değerli hocam sayın Prof. Dr. Özkan ÜNVER (Ufuk Üniversitesi)'e, manevi destekleri ile hep yanımda olan arkadaşlarım Arş.Gör. Z.Birce ERGÖR, Arş.Gör. Özlem TÜRKŞEN, Aslı AŞIK, Selma EŞGÜNOĞLU, Arş.Gör. Hande GÜRDAĞ, Yrd.Doç.Dr. Meltem ANAFARTA ŞENDAĞ, Yrd. Doç. Dr. Funda KUTLU, Yrd. Doç. Dr. Aslıhan ALHAN, Arş.Gör. Aslı YALÇIN, Arş.Gör. Burçin SARICA, Arş.Gör. Gülçim BİLİM, Arş.Gör. Tuba YUMUŞAK'a ve doktora sürecinin her aşamasını birlikte yaşadığımız, aynı heyecanları paylaştığımız arkadaşlarım Arş.Gör. M. Bahar BAŞKIR ve Öğr. Gör. Pelin TOKTAŞ'a, çalışmalarım süresince birçok fedakârlık göstererek beni destekleyen eşim Gökhan EKİNCİ, ablam Dilek MAMAK AYDIN ve Taylan AYDIN'a ve tüm aileme en içten duygularımla teşekkür ederim.

Elmas Burcu MAMAK EKİNCİ

Ankara, Mart 2011

İÇİNDEKİLER

ÖZET.....i

ABSTRACT.....ii

TEŞEKKÜR.....iii

SİMGELER DİZİNİ.....vi

ŞEKİLLER DİZİNİ.....vii

ÇİZELGELER DİZİNİ.....ix

1. GİRİŞ.....1

2.TEMEL TANIM ve KAVRAMLAR.....6

2.1 Derinlik Kavramı 6

2.2 Bazı Derinlik Fonksiyonları.....7

2.3 İstatistiksel Derinlik Fonksiyonlarına İlişkin Özellikler.....14

2.4 Yığın Dağılımı için Derinlik Fonksiyonu.....15

2.5 Çok Değişkenli Yüzdelikler.....19

2.6 Tek Değişkenli Fark İstatistikleri.....22

2.7 Tarama İstatistikleri.....23

3. DERİNLİKLERE DAYALI ÇOK DEĞİŞKENLİ TARAMA İSTATİSTİKLERİ.....29

3.1 Giriş.....29

3.2 Derinliklere Dayalı Çok Değişkenli Tarama İstatistikleri.....32

4. YÖNTEMİN DEĞERLENDİRİLMESİ VE SONUÇ.....44

KAYNAKLAR.....48

EKLER.....52

EK 1 Tek Köşeli Konveks Bölge İçin Derinlik Fonksiyonunun Hesaplanması.....53

EK 2 Huffer ve Lin'in Yaklaşımı 55

EK 3 Teorem 3.1 İspatı(Li and Liu 2008, Theorem 2.1)….……...59

EK 4 Düzgün Dağılıma ait Yığın Derinlik Fonksiyonunun

Olasılık Dağılımının Bulunması……..……....60

ÖZGEÇMİŞ.....62

SİMGELER DİZİNİ

α	Alfa
Δ	Delta
$D(x;F)$	Derinlik fonksiyonu
$D_F(X)$	Yığın derinlik fonksiyonu
$D_{F_n}(X)$	Örneklem derinlik fonksiyonu
D^α	Yığına ait α derinlikli iç bölge
D_n^α	Örnekleme ait α derinlikli iç bölge
inf	İnfimum
$X_{[i]}$	$i.$ sıra istatistiği
$\langle\ \rangle$	İç çarpım fonksiyonu
λ	Lamda
Max	Maksimum
Min	Minimum
$F_X(x)$	X rasgele değişkenine ait dağılım fonksiyonu
$f_X(x)$	X rasgele değişkenine ait olasılık yoğunluk fonksiyonu
S^2	Örnekleme ait varyans kovaryans matrisi
Σ	Yığına ait varyans kovaryans matrisi
ξ_p	Yığına ait $p.$ yüzdelik

ŞEKİLLER DİZİNİ

Şekil 1.1 Birim zaman aralığında sabit w=1/6 aralık uzunluğundaki bir tarama penceresi ile tarama…….2

Şekil 2.1a Herhangi bir x noktası üzerinden geçen yarı uzayın altında ya da üstünde kalan nokta sayısı 5 olduğu durum……11

Şekil 2.1b Herhangi bir x noktası üzerinden geçen yarı uzayın altında ya da üstünde kalan nokta sayısı 3 olduğu durum……11

Şekil 2.2 p olasılığı için derinlik konturları tarafından kapanmış en küçük bölge…….…..……..13

Şekil 2.3 Taralı üçgenin minimum alanı ile ifade edilen (x_0, y_0) noktasının derinliği…….……..17

Şekil 2.4 Bir kare üzerinde düzgün dağılımın 0.05, 0.15, 0.25, 0.35, 0.45 derinlik değerlerinde karşılık gelen derinlik konturları……19

Şekil 3.1 Düzgün dağılımın yığınına ait derinlik konturları içinde 2 boyutlu normal dağılımlı 100 gözlemin saçılımı……..35

Şekil 3.2 İki boyutlu normal dağılımlı 100 gözlemin kendi verisine göre elde edilen 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40 derinlikli derinlik konturları……35

Şekil 3.3 Düzgün dağılımın yığınına ait derinlik konturları içinde 2 boyutlu düzgün dağılımlı 100 gözlemin saçılımı……..36

Şekil 3.4 İki boyutlu düzgün dağılımlı 100 gözlemin kendi verisine göre elde edilen 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40 derinlikli derinlik konturları……36

Şekil 3.5 $\alpha_1 = 0.05$ ve $\alpha_2 = 0.15$ derinlikli derinlik konturları arasında kalan bölgenin alanı(olasılığı)…….40

Şekil 3.6 İki boyutlu düzgün dağılımlı 25 gözlemin düzgün dağılımın yığın derinlik fonksiyonundan elde edilen derinlik değerlerine karşılık gelen olasılıklar……42
Şekil 3.7 İki boyutlu normal dağılımlı 25 gözlemin düzgün dağılımın yığın derinlik fonksiyonundan elde edilen derinlik değerlerine karşılık gelen olasılıklar……42

ÇİZELGELER DİZİNİ

Çizelge 4.1 100 örneklem içinden k gözlem içeren örneklemlerin sayısı……..45

Çizelge 4.2 $k \geq 7$ durumunda simülasyon sonuçlarına göre H_0 hipotezinin reddedilme olasılıkları…….....45

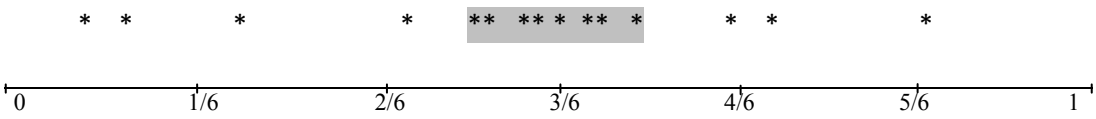
1.GİRİŞ

Çok değişkenli tarama istatistiklerine ait dağılımın derinliklere dayalı olarak elde edilmesi bu çalışmanın konusudur. Tek boyutlu tarama istatistiklerinin fark istatistiklerine (spacings) bağlı olarak tanımlanmasına paralel olarak çok boyutlu tarama istatistikleri de çok değişkenli fark istatistikleri (multivariate spacings) aracılığı ile tanımlanacaktır. Çok değişkenli verilerde tarama için bu yaklaşımın, verinin geldiği yığına ait dağılımın geometrik yapısını dikkate alıyor olması ve tarama penceresinin veriye göre kendiliğinden belirlenmesinin önemli olduğu düşünülmektedir.

Gözlemlerin tek boyutlu rasgele değişkenler olduğu durumlarda tarama istatistiklerinin tanıtımı için Naus (1965a)'da belirtilen uygulama problemini dikkate almak uygun olacaktır. Bir dakikalık sürede en fazla 15 telefon bağlantısının yapılabildiği bir telefon santrali ele alınsın. 10 saniyelik ($1/6$ dakika) herhangi bir zaman aralığında santrale 8 veya daha fazla telefon bağlantısı gerçekleşmesi olasılığı hesaplanmak istenmektedir. Bu olasılık problemini çözmek için söz konusu rasgele olgu aşağıda anlatıldığı gibi modellenebilir.

$X_1, X_2, ..., X_n$, sürekli rasgele değişkenleri $-\infty < a < b < \infty$ aralığında değer alan birbirinden bağımsız ve aynı F dağılımlı bir örneklem olsun. $X_{[1]} \le X_{[2]} \le ... \le X_{[n]}$ bu örnekleme ilişkin sıra istatistiklerini göstersin. Bu örneklem için $X_{[0]} = a$ ve $X_{[n+1]} = b$ olmak üzere $\left(X_{[i-1]}, X_{[i]}\right]$, $i = 1, 2, ..., n+1$ aralıkları veya bu aralıkların uzunlukları $D_i = X_{[i]} - X_{[i-1]}$ ardışık sıra istatistikleri arasındaki fark (spacings) olarak adlandırılır (Pyke 1965, Li ve Liu 2008). Bundan sonraki kısımlarda ardışık sıra istatistikleri arasındaki farklara (spacings) kısaca fark istatistikleri denilecektir. Çoğu durumda $(0, t)$ zaman aralığında ortaya çıkan n tane gözlemden, bu zaman aralığında herhangi bir sabit $0 < w < t$ için w genişliğine sahip tarama penceresi ile tarama yapıldığında, pencere içinde yer alan gözlemlerin maksimum sayısı tarama istatistiği olarak adlandırılır ve S_w ile gösterilir (Glaz vd 2001). Yukarıdaki santral probleminde aranan

olasılık Naus (1965a)'un gösterimi ile $P\left(S_{1/6} \geq 8\right) = P\left(8;15,1/6\right)$ şeklindedir. Bu olasılık sorusunun ardışık sıra istatistikleri ile bağlantısını kurmak için bir dakikalık zaman aralığında santrale gelen telefonların geliş zamanları rasgele değişken olarak kabul edilsin. Dolayısıyla geliş zamanları, geliş sırasına göre (0,1) dakikalık zaman aralığında sıralanmış olacaktır; geliş zamanları rasgele değişkeninin sıra istatistikleri olarak değer almış olacaklardır. k nokta içeren en küçük aralık uzunluğu $W_k = \underset{1\leq i\leq n-k+1}{Min}\left\{X_{[k+i-1]} - X_{[i]}\right\}$, $k=2,...,n$ (Wallenstein ve Naus 1973) ile tanımlandığında bu problem $k=8$ nokta içeren en küçük aralığın $1/6$'dan küçük olması olasılığı $P\left(W_8 \leq 1/6\right)$ olarak da yazılabilir. $w=1/6$ uzunluğu, (0,1) aralığı boyunca 8 gözlemi içeren herhangi bir aralık olabilir.



Şekil 1.1 Birim zaman aralığında sabit $w=1/6$ aralık uzunluğundaki bir tarama penceresi ile tarama. $n=15$ gözlemin meydana geliş zamanları "*" ile gösterilmiştir.

Sıra istatistiğinin tanımlanabileceği her rasgele olgu için tarama istatistiği de tanımlanabilecektir.

Geriye dönük (retrospective) ya da ileriye yönelik (prospective) olarak istatistiksel çıkarımın hedeflendiği durumlarda bütün rasgele değişkenler için tarama istatistikleri tanımlanabilir. Bu konulardaki geniş ve detaylı bilgi ve kaynakça için Glaz ve Balakrishnan (1999), Glaz vd. (2001), Glaz vd. (2009) çalışmalarına bakılabilir. Fark istatistikleri ve tarama istatistikleri istatistiksel çıkarım içinde aykırı gözlemlerin (outliers) belirlenmesi ve mod tahmini amacıyla da kullanılabilir (Seth 1950). Bu amaçlardan da görülebileceği üzere tarama istatistikleri genel olarak örneklemin ortaya koyduğu yığılma ve kümelenmenin anlamlılığını test etme aracı olarak düşünülebilir.

Olgular ve olaylar dünyasında, uygulamada rasgeleliği değerlendirmek rasgeleliğin işevuruk (operasyonel) tanımı yapılmadıkça olanaksızdır. Bununla birlikte, rasgelelikten ayrılış olarak değerlendirebileceğimiz gözlemler, kümelenme veya bir trend yapısına sahipse bu durumu rasgelelikten sapma olarak değerlendirmemiz olanaklıdır. Genel olarak, gözlemlerin bir kümelenmesinin veya trendinin olağan olup olmaması, rasgeleliğin incelenmesi ve gözlemlerin geldiği yığının düzgün olup olmadığının sorgulanması olarak değerlendirilir (Lehmann 1998). Tarama istatistikleri çerçevesinde veride kümelenme ve sıra dışılığın göze çarpması da gözlemlerin düzgün dağılımdan gelmediğini anlamlandıracak istatistikî kanıt olarak değerlendirilmektedir (Glaz ve Balakrishnan 1999, s.14). Bu nedenle çoğu kez S_w istatistiği, düzgün $(0,1)$ dağılımlı yığınından n çaplı $X_1, X_2, ..., X_n$ rasgele örneklemine dayandırılarak tanımlanır.

Çok boyutlu vektörler ile çalışıldığında da tarama istatistiklerine gereksinim duyulabilir. Örneğin, bir hastalığın belirli bir yerleşim alanında görülme sıklığının incelendiği bir çalışmada hastalığın görüldüğü coğrafi bölgede bir kümelenmenin göze çarpması araştırmacılara önemli bir bilgi sağlayabilir. Böyle bir kümelenmenin, uygun görülen bir tarama penceresi kullanılarak mekânsal (spatial) olarak taranması ile sözkonusu kümelenme bölgesine düşen gözlem sayısının olağan olup olmadığı sorgulanabilir. Eğer olağan ise böyle bir gözlem kümesine rastlama olasılığı hayli yüksek çıkacaktır. Bu durumda araştırmanın yapıldığı mekânda hastalıklı bireyin gözlemlendiği konumun bileşenleri iki boyutlu rasgele bir vektör ile gösterilecektir. Bununla birlikte yukarıdaki tek boyutlu rasgele değişkenlerle tanımlanan tarama istatistikleri, yerini düzlemde tanımlı tarama pencelerine bırakacaktır.

Rasgele değişkenlerde sıra istatistikleri ile yapılan istatistiksel çözümlemeler, R^d'de rasgele vektörlerle ile yapılmak istendiğinde sıralama işlemi rasgele değişkenin sıralamasında olduğu gibi doğal ve anlamlı olarak yapılamayacaktır. Örneğin, tarama istatistiklerinin kullanımı söz konusu ise tarama penceresi, tek boyutlu istatistiksel çözümlemelerde doğal olarak bir aralık iken çok boyutlu çözümlemelerde farklı şekillerde tarama pencereleri şeklinde düşünülebilecektir. Aynı sorun tolerans bölgelerinin oluşturulmasında da ortaya çıkmaktadır (Li ve Liu 2008). Nitekim çok

boyutlu tarama istatistiklerinde çalışmalar tarama pencerelerinin şeklinin seçimi ve sonrasında dağılım hesaplamada karşılaşılan sorunların giderilmesi, kolaylaştırılması üzerinde süregelmektedir. Daha geniş bilgi için Naus (1965b), Cressie (1991)'e bakılabilir. Örneğin, Alm (1999), tek boyutlu bir X Poisson sürecinde iki boyutlu tarama istatistikleri için bazı olasılık yaklaşımları geliştirmiştir. Bir dikdörtgensel bölgede sabit bir tarama penceresine düşen maksimum sayıda gözlemlere ilişkin olasılıkların, seçilen farklı şekillerdeki (karesel, üçgensel ve dairesel) tarama penceresi için değiştiğini göstermiştir. Glaz vd. (2001)'de tarama penceresi şeklinin ve büyüklüğünün, gözlemlerin tarama penceresine düşme olasılıklarını değiştirdiğini belirtmişlerdir. Månsson (1999), koşullu ve sürekli durumda, çoklu tarama istatistikleri (birden fazla değişik tarama pencereleri kullanılarak) çerçevesinde Poisson yaklaşımını incelemiştir; bir Poisson sürecinde iki boyutlu durumda n gözlemden k ve daha fazla gözlem içeren farklı geometrik şekilli tarama pencereleri bulunduğunda tarama, bu tarama pencerelerinden minimum alana sahip olanı ile yapılmaktadır.

Derinlik (depth) kavramı, çok boyutlu verilerin istatistiksel çözümlenmesinde sıralama problemine rasgele değişkenlerdeki doğal bir sıralama kadar olmasa da önemli çözümler sunmaktadır. Bunun yanında rasgele değişkenlerin sıra istatistikleri ile de bağdaşmaktadır.

Bu çalışmada derinliklerin sağlamış olduğu rasgele vektörlerin derinlik değerlerine göre sıralama özelliği kullanılarak çok boyutlu verilerin taramasında karşılaşılan tarama penceresinin şeklinin seçimi ve tarama istatistiğinin dağılımının elde edilmesine bir yaklaşım önerisinde bulunulmuştur. Derinlikler aracılığıyla elde edilen derinlik konturları çok boyutlu verilerde dağılımın "geometrik yapısını" yansıtır (Zuo ve Serfling 2000b). Formal olarak tanımlaması tezin 3. Bölümünde Tanım 3.1 ile verilen çok değişkenli fark istatistikleri (multivariate spacings), ardışık derinlik değerleri ile oluşturulan derinlik konturları arasında kalan gözlemlerin oluşturduğu bölgeler olarak ifade edilir (Li ve Liu 2008). Bu bölgeler "derinlik halkaları" olarak da ifade edilmektedir. Li ve Liu (2008), R^d 'de herhangi bir F dağılımı için çok değişkenli fark istatistiklerinin olasılıklarının ortak dağılımının, düzgün dağılımın tek değişkenli fark istatistiklerinin dağılımı ile aynı dağılımlı olduğunu göstermişlerdir. Bununla birlikte

çok değişkenli fark istatistiklerinin yapısı (geometrik biçimi) dağılımdan dağılıma değişiklik göstermektedir.

Bu çalışmada çok boyutlu verinin gözlemlendiği yığının dağılımının çok boyutlu düzgün dağılım olduğu hipotezi altında gözlemler, Rousseeuw ve Ruts (1999)'da tanımlandığı üzere düzgün dağılımın yığın derinlik fonksiyonuna göre sıralanıp gözlem derinlik değerlerinin yukarıdaki hipotez altında olasılıkları, düzgün dağılıma ait yığın derinliği dağılım fonksiyonu aracılığı ile elde edilecektir. Böylece çok boyutlu veri için sabit olasılıklı derinlik halkaları ile dıştan merkeze doğru tarama yapılabilecektir. Çok boyutlu tarama, tek boyutlu durumdaki aralık ile taramaya indirgenip verinin çok boyutlu düzgün dağılımdan gelip gelmediğinin test edilmesi mümkün olacaktır. Sonuç olarak sabit genişlikli tarama penceresi, verinin gözlemlendiği yığının geometrik yapısına göre belirlenmiş olacaktır. Örneklem için derinlik olasılıklarının dağılımı, dağılım ne olursa olsun düzgün dağılıma sahip olacaktır (Li ve Liu 2008, Theorem 2.1). Tarama halkalarına ait olasılıkların ortak dağılımınında, düzgün dağılımlı örneklem için tek değişkenli fark istatistiklerinin ortak dağılımları ile aynı olması sonucunda yapılan bu taramanın tek boyutlu veriler için yapılan tarama ile bağdaştığı söylenebilir.

Böylece klasik çok boyutlu tarama istatistikleri yaklaşımlarında kullanılan tarama penceresinin şeklinin ne olacağı sorununun ortadan kalkacağı düşünülmektedir. R^d 'deki herhangi bir F dağılımından rasgele seçilen $X_1, X_2, ..., X_n$ örneklemindeki her bir noktanın bu dağılım altındaki derinlik sıralaması konum olarak yığılmayı vermemektedir. Ancak, aynı olasılığa sahip farklı derinlik değerlerine karşılık gelen derinlik halkalarına (multivariate spacings) düşen gözlem sayıları, dağılıma uygunluğu test etme imkânı sağlayacaktır.

Bu çalışmada tek boyutlu ve çok boyutlu tarama istatistiklerinin tanımları yapılmıştır. Ancak uygulamada iki boyutlu (değişkenli) veriler üzerinde çalışılmış ve iki boyutlu verilerin sıralanmasında yarı uzay derinlik fonksiyonu (Tukey depth ya da Halfspace depth) kullanılmıştır. Tarama istatistikleri için sürekli durumda koşullu veya geriye dönük (retrospective) tarama istatistikleri üzerinde çalışılmıştır.

2. TEMEL TANIM ve KAVRAMLAR

2.1 Derinlik Kavramı

Rasgele vektörler ile istatistiksel analiz yapılırken bazen bu verilerin sıralanması ihtiyacı doğar. Örneğin çok değişkenli dağılım yüzdeliklerine ulaşmak için verilerin sıralanması gerekmektedir. Çok değişkenli verilerin sıralanması tek değişkenli verilerdeki sıra istatistikleri gibi doğal yollarla yapılamamaktadır. Bu ihtiyacı "derinlik" kavramı kısmen karşılayabilmektedir. Bu kavram üzerine Liu vd. (1999)'nin önemli ve yol açan çalışması burada temel alınacaktır.

$X_1, X_2, ..., X_n$, d boyutlu rasgele vektörleri birbirinden bağımsız ve sürekli olan aynı F dağılım fonksiyonuna sahip olsunlar.

Derinlik fonksiyonu, R^d'de tanımlı, $\forall x \in R^d$ için ve F dağılım fonksiyonuna göre katman derinliği hariç $[0,1]$ kapalı aralığında ya da bunun bir alt aralığında değer alan bir fonksiyondur ve $D(x;F)$ ile gösterilir. Kolaylık olması açısından bu aşamadan sonra yığın için $D_F(X)$, n çaplı örneklem için $D_{F_n}(X)$ (Li ve Liu 2008 notasyonları) ile gösterilecektir. $D_F(X)$, aynı zamanda yığın derinlik fonksiyonunun da tanımıdır (Rousseeuw ve Ruts 1999).

Derinlik, $x \in R^d$ noktasının, F dağılım fonksiyonunun "merkezine" göre uzaklığının bir ölçüsüdür. Burada merkez, dağılımın ortancası (medyan) alınabilir. Derinlik fonksiyonu, d boyutlu rasgele vektörlerin, n çaplı örnekleminden alınan rasgele vektörlerin sıralamasını yapabileceğimiz yardımcı veya aracı bir fonksiyon olarak görülebilir. Örneklemin en derindeki noktası $X_{[1]}$, en uzaktaki noktası da $X_{[n]}$ olarak gösterilir. Merkeze yaklaştıkça derinlik değerlerinin artması istatistiksel derinlik fonksiyonları için beklenen bir olgudur (Serfling 2002b).

Giriş bölümünde rasgele değişkenlerin sıra istatistikleri için kullanılan gösterim, Liu vd. (1999)'de olduğu gibi rasgele vektörlerin derinliklere göre sıralanan gözlemleri için de kullanılacaktır. Ancak bir farkla, rasgele vektörlerde en derindeki gözlem değeri $x_{[1]}$, merkezden en uzakta olan gözlem değeri için $x_{[n]}$ gösterimi kullanılırken, rasgele değişkenler için en küçük gözlem değeri için $x_{[1]}$ ve en büyük gözlem değeri için $x_{[n]}$ kullanılacaktır.

Veri kümesine göre derinlik, verilen bir $x \in R^d$ noktasının F dağılımlı $X_1, X_2, ..., X_n$ rasgele örneklemine göre hangi derinlikte olduğunun ölçümüdür (Liu vd. 1999). Değişik derinlik fonksiyonları tanımlanmıştır. Hesaba dayalı geometri (computational geometry) çerçevesinde proximity derinliği (proximity depth) gibi farklı derinlik fonksiyonları da söz konusudur (Hugg vd. 2006). Ancak bu çalışmada bunlar irdelenmeyecektir.

2.2 Bazı Derinlik Fonksiyonları

Derinlik fonksiyonlarından yaygın olarak bilinenlerinin tanımları herhangi d boyutlu rasgele vektörler için verilecektir. Derinlik fonksiyonlarına ait simgeler biraz değiştirilerek kullanılacaktır. Her derinlik tanımlaması değişik bir derinlik kavramı ifade etmektedir. Karışıklık olmaması için bu yola başvurulmuştur. Daha sonraki bölümlerde aşağıdaki derinlik kavramlarından sadece yarı uzay derinliği (half-space depth) kullanılacak ve önceden belirlenen derinlik notasyonları kullanılmaya devam edilecektir.

$x \in R^d$ noktasının F dağılımına göre yarı uzay derinliği

$$HD(F; x) = \inf_{H} \left\{ P(H) : H, R^d \text{'de x'i içeren kapalı bir yarı düzlem} \right\} \qquad (2.1)$$

olarak tanımlanır. Yarı uzay derinliğinin örneklem karşılığı

$$\hat{HD}(F;x) = \inf_{H} \left\{ \frac{S\{X_i; X_i \in H, i = 1,2,...,n\}}{n};\ H, R^d \ \text{de } x\text{'i içeren kapalı bir yarı düzlem} \right\} \quad (2.2)$$

olup burada $s\{A\}$, A kümesinin eleman sayısını göstermektedir (Liu vd. 1999). Bu çalışmada örnek olarak alacağımız derinlik fonksiyonu yarı uzay derinliği olacaktır. Burada $a \in R^d$, $a \neq 0$ ve $b \in R$ ise $a_1x_1 + a_2x_2 + ... + a_dx_d \leq b$ eşitsizliğini sağlayan $x \in R^d$ noktaların kümesi $\{x \in R^d : \langle a, x \rangle \leq b\}$, R^d'de bir kapalı yarı uzay olarak adlandırılır (Cameron, 1985).

$X_1, X_2, ..., X_{d+1}$ rasgele örneklemi için $x \in R^d$ noktasının F dağılımına göre simpleks derinliği (simplicial depth)

$$SD(F;x) = P(x \in S[X_1, X_2, ..., X_{d+1}]) \quad (2.3)$$

olarak tanımlanır. $S[X_1, X_2, ..., X_{d+1}]$, köşeleri F dağılımından alınan $d+1$ tane, $X_1, X_2, ..., X_{d+1}$ rasgele noktaları (vektörleri) olan kapalı bir simplekstir.

$C \subset R^d$ boş olmayan konveks küme olmak üzere eğer $x \in C$ ve C içindeki herhangi bir doğrunun iç noktası değilse C'nin bir köşe noktası olarak bilinir. R^d'de C gibi bir konveks çok yüzlü eğer $d+1$ tane köşeye sahipse bu çok yüzlü d boyutlu simplekstir (Cameron 1985). Bilindiği gibi R^2' de simpleksler üçgenlerdir. Simpleks derinliğin örnek karşılığı

$$\hat{SD}(F;x) = \binom{n}{d+1}^{-1} \sum I_{(x \in S[X_1, X_2, ..., X_{d+1}])}(x) \quad (2.4)$$

şeklindedir. Burada I, 0 ve 1 değerlerini alan gösterge fonksiyonu olup, toplama işlemi $X_1, X_2, ..., X_n$ örneklemindeki n tane noktanın (vektörün) simpleks oluşturan $d+1$'li kombinasyonları üzerinden yapılmaktadır (Liu vd. 1999).

F, d boyutlu X rasgele vektörüne ait dağılım fonksiyonu, μ_F ortalama (beklenen değer) vektörü, Σ_F varyans kovaryans matrisi ve $x \in R^d$ bir nokta olmak üzere

$$MD(F;x) = \left[1 + (x - \mu_F)' \Sigma_F^{-1} (x - \mu_F)\right]^{-1} \tag{2.5}$$

değerine x noktasının F dağılım fonksiyonuna göre Mahalanobis derinliği (Mahalonobis depth) denir. Dağılımın merkezi μ_F alınmıştır. Mahalanobis derinliğinin örneklem karşılığı μ_F yerine $\overline{X}$ örnek ortalaması ve Σ_F yerine S^2 örnek varyans kovaryans matrisinin konulmasıyla

$$\hat{MD}(F;x) = \left[1 + (x - \overline{X})'(S^2)^{-1}(x - \overline{X})\right]^{-1} \tag{2.6}$$

şeklinde yazılır. Burada $\overline{X} = n^{-1}\sum_{k=1}^{n} X_k$ ve $\overline{X}_l$, örneklemdeki X_i rasgele vektörlerinin $l.$ bileşenine ait örneklem ortalamalarını göstermek üzere S^2, $d \times d$'lik $i.$ satır ve $j.$ sütun elemanı $S_{ij} = n^{-1}\sum_{k=1}^{n}\left(X_{ik} - \overline{X}_i\right)'\left(X_{jk} - \overline{X}_j\right)$, $i, j = 1,2,...,d$ olan matristir. Örneklem derinliği bir anlamda kitle derinliği için bir tahmin edicidir (Liu vd. 1999).

$x \in R^d$ noktasının F dağılımına göre Oja derinliği (oja depth)

$$OD(F;x) = \left[1 + E\left(Hacim\left(S\left[x, X_1, X_2, ..., X_d\right]\right)\right)\right]^{-1} \tag{2.7}$$

dır. Burada $S[x, X_1, X_2, ..., X_d]$, $x \in R^d$ noktası ve d tane $X_1, X_2, ..., X_d$ rasgele vektörleri ile oluşturulan simplekstir. Oja derinliğinin örneklem karşılığı ise

$X_1, X_2, ..., X_n$ örneğinden elde edilebilecek tüm d'li $(X_{i1}, X_{i2}, ..., X_{id})$ gözlemleri üzerinden alınmak üzere

$$\hat{OD}(F;x) = \left[1 + \binom{n}{d}^{-1} \sum Hacim(S[x, X_{i1}, X_{i2}, ..., X_{id}])\right]^{-1} \tag{2.8}$$

olarak tanımlanır (Liu vd. 1999).

$X_1, X_2, ..., X_n$ veri kümesine göre bu örneklemdeki X_k örneklem noktalarının konveks katman derinliği (convex hull peeling), X_k'ya ait olan konveks tabakanın seviyesidir.

Bir konveks tabaka aşağıdaki gibi tanımlanır. $X_1, X_2, ..., X_n$ örneklem noktalarını kapsayan en küçük konveks katman kurulur. Bunun çevresindeki örneklem noktaları olan birinci konveks tabaka yok edilir ve kalan noktaların konveks katmanı kurulur. Bu da ikinci konveks tabaka olur. Bu süreç tekrarlanır ve en küçük konveks tabaka kurulmuş olur. Gözlemin dahil olduğu katman sıra sayısı büyüdükçe derinliği artıyor demektir. Burada katmanların oluşumu da bir soğanın katlarına benzetilebilir. Her bir katmanca oluşturulan çok yüzlünün çizgileri derinlik çizgisi olarak adlandırılır (Liu vd. 1999).

$x \in R^d$ noktasının F dağılımına göre yarı uzay çoğunluk derinliği (majority depth)

$$MJD(F; x) = P(x \in C(X_1, X_2, ..., X_d)) \tag{2.9}$$

olarak tanımlanır. Burada $C(X_1, X_2, ..., X_d)$, F dağılımlı bağımsız $X_1, X_2, ..., X_d$ rasgele noktalarından (vektörlerinden) geçen; 0.50 ve daha fazla olasılıklı yarı uzayı tanımlamaktadır.

Olasılığı 0.50 ve daha fazla olan yarı uzaylar içinde sıklıkla yer alan bir noktanın daha büyük derinlik değerine sahip olacağı düşünülürse yarı uzay çoğunluk derinliğinin örneklem karşılığı

$$\hat{MJD}(F, x) = \binom{n}{d}^{-1} \sum I_{C(X_{i1}, X_{i2}, ..., X_{id})}(x) \tag{2.10}$$

olarak tanımlanır. Buradaki toplam $X_1, X_2, ..., X_n$ örneğinin d'li kombinasyonları üzerindendir (Liu vd. 1999).

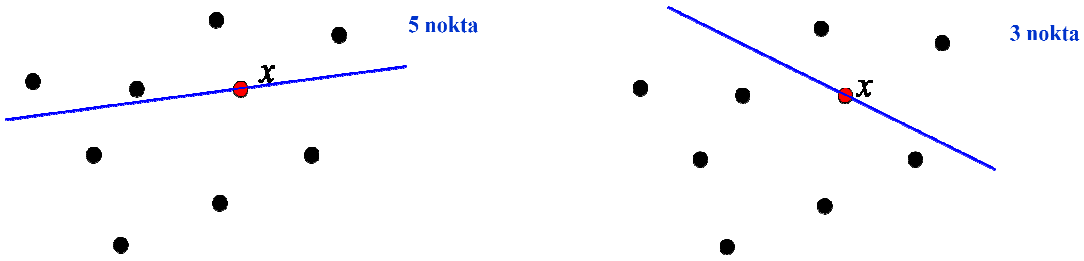
Yarı uzay örneklem derinliğine ilişkin bir örneği aşağıda vermek uygun olacaktır.

Herhangi bir F dağılımı için(yani yığına ait) yarı uzay derinliği (2.1)'de

$$HD(F; x) = \inf_H \{P(H) : H, R^d \text{'de x'i içeren kapalı bir yarı düzlem}\}$$

ile tanımlanmaktadır. Herhangi bir F dağılımından geldiği varsayılan R^2'de değer alan rasgele vektörlerin $n=9$ çaplı bir örnekleminin gözlem değerleri $x_1, x_2, ..., x_9$ içinden herhangi bir x_i noktasının bu veri bulutuna (yani $x_1, x_2, ..., x_9$ gözlemlerine) göre yarı uzay derinliğinin hesaplanması:

Yarı uzay derinliğinin örneklem karşılığı, x noktası için (2.2)'deki tanımdan ,x'in üzerinden geçen her doğru için, bu doğrunun altında ve üstünde kalan minimum sayıdaki noktalar ile hesaplanır. Şekil 2.1.a.b iki boyutlu uzayda 9 gözlemlik bir veri içindeki x noktasının bu veriye göre yarı uzay derinliği için bilgi vermektedir.



Şekil 2.1a Herhangi bir x noktası üzerinden geçen yarı uzayın altında ya da üstünde kalan nokta sayısı 5 olduğu durum. b. Herhangi bir x noktası üzerinden geçen yarı uzayın altında ya da üstünde kalan nokta sayısı 3 olduğu durum

Şekil 2.1a'da x noktası üzerinden geçen yarı uzayın altında veya üzerinde kalan nokta sayısı 5'dir. Bu örnek için minimum olduğu durum şekil 2.1b ile gösterilmiştir. x noktasının bu veriye göre yarı uzay derinliği,

$$\hat{HD}(F;x) = \inf_{H} \left\{ \frac{s\{X_i; X_i \in H\}}{n};\ H, R^d \ \ de\ x'i\ içeren\ kapalı\ bir\ yarı\ düzlem \right\}$$
$$= 3/9 = 0.333$$

olarak bulunur.

Daha önce verilen derinlik ölçülerinin bazılarında derinlik değerleri [0,1] aralığındadır. Derinliği en büyük olan noktaya *merkez* denilebilir. Birden fazla örneklem noktasının en derin değere sahip olduğu durumlarda, en derin nokta bu noktaların aritmetik ortalaması olarak alınır. Derinlik sıralamasında, eş derinlikli gözlemlerin olması halinde sıra istatistiklerinde olduğu gibi işlem yapılmaz; aynı derinliğe sahip olan gözlemlere birbirlerini takip eden derinlik sıra numarası verilir. Bu durumda gözlem sayısı kadar derinlik sıra numarası söz konusudur. Aynı derinlik değerine sahip gözlemlerin oluşturduğu kümeye denk derinlikli noktalar kümesi (*depth-equivalance class*) adı verilir (Liu vd. 1999).

Yukarıda değinilen derinlik kavramı ve başka derinlik fonksiyonlarına ait özellikler, Zuo ve Serfling (2000a)'de incelenmektedir. Zuo ve Serfling (2000a), yarı uzay derinliği, simpleks derinliği ve projeksiyon derinliğinin potansiyel olarak iyi özelliklere sahip olduğunu ve yarı uzay derinliği ile projeksiyon derinliğinin *istatistiksel derinlik fonksiyonları* olarak adlandırılabileceğini söylemektedirler.

Derinlikler kavramını uygulamaya geçirmek için aşağıdaki notasyonlara ve tanımlara ihtiyaç duyulmaktadır.

F , R^d, $d \geq 2$ üzerinde sürekli dağılım fonksiyonu olsun ve $X_1, X_2, ..., X_n$'nin , F'den alınan rasgele bir örneklem olduğunu varsayalım.

Tanım 2.1 $\{x \in R^d : D(x;F) = \alpha\}$ kümesine α derinlikli düzey kümesi veya α derinlik konturu adı verilir.

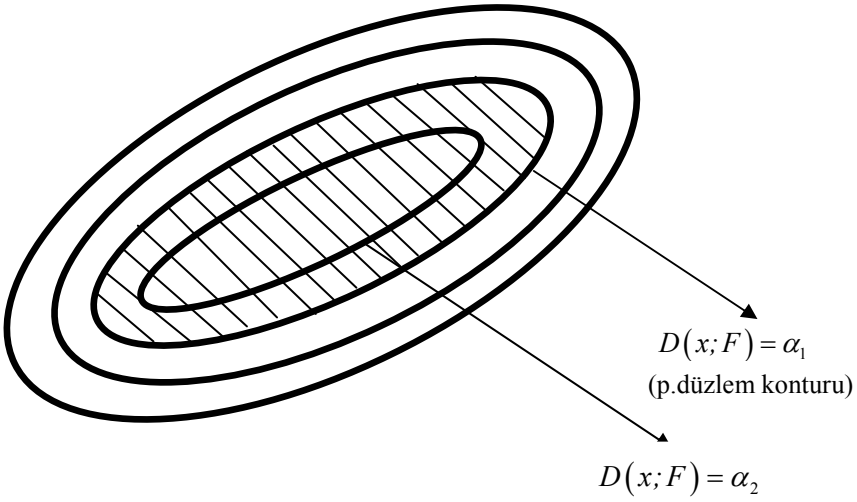
Tanım 2.2 $R(\alpha) = \{x \in R^d : D(x;F) > \alpha\}$ kümesi, α derinlik konturu tarafından sınırlanmış bölge olarak tanımlanır.

Tanım 2.1-2.2 ile belirlenen küme Serfling (2002)'de iki kümenin birleşimi olan

$$I(\alpha, D, F) = \{x \in R^d : D(x;F) \geq \alpha\} \tag{2.11}$$

gösterimi ile α derinlikli iç bölge olarak adlandırılır. $I(\alpha, D, F)$ bölgesinin sınırı $\partial I(\alpha, D, F)$ da α derinlik konturunu gösterir. Kolay kullanım açısından α derinlikli iç bölge gösteriminde $I(\alpha, D, F)$ yerine D^{α} kullanılacaktır.

Tanım 2.3 $C_p = \bigcap_{\alpha} \{R(\alpha) : P_F(R(\alpha)) \geq p\}$ kümesi $p \in (0,1)$ olmak üzere $p.$merkezi bölge olarak adlandırılır. Diğer bir anlatımla C_p, p olasılığı için derinlik konturları tarafından kapanmış en küçük bölgedir. C_p'nin sınırı en küçük $p.$ düzlem konturunu gösterir. $Q(p)$ ve $Q_F(p)$ ile gösterilir (F dağılımı altında).



Şekil 2.2 p olasılığı için derinlik konturları tarafından kapanmış en küçük bölge

Taralı bölge, p ve daha büyük olasılığa sahip α_1 derinlik konturu tarafından kapsanan bölge diğer bir deyişle p olasılığı için derinlik konturları tarafından kapanmış en küçük bölgedir. $P\{x \in R^d : D(x) \geq t_p\} = p$ olduğunda $Q_F(p), \{x \in R^d : D(x) = t_p\}$ derinlik konturudur. Eğer F mutlak sürekli ve f yoğunluk fonksiyonu her yerde sıfırdan faklı ise, $P_F(R(t_p)) = p$ olması koşulu ile $C_p = R(t_p)$ dir (Liu vd. 1999).

2.3 İstatistiksel Derinlik Fonksiyonlarına İlişkin Özellikler

R^d üzerinde ilgilenilen dağılım fonksiyonlarının sınıfı $\mathcal{F}$ ile gösterilsin ve X, dağılım fonksiyonu F_X ile verilen bir rasgele vektör olsun.

Tanım 2.4 R^d'de tanımlı ve $F_X \in \mathcal{F}$ dağılım fonksiyonuna sahip X rasgele vektörü için $D_F(.)$ derinlik fonksiyonu afin değişmez, merkezde en büyük derinliğe sahip, en derin noktaya göre monoton ve derinlik değeri $\|x\| \to \infty$ iken $D_F(x) \to 0$ ise istatistiksel derinlik fonksiyonu olarak adlandırılır.

Yukarıda tanımlanan istatistiksel derinlik fonksiyonlarının özellikleri daha formal olarak aşağıdaki gibi ifade edilebilir (Zuo ve Serfling 2000a).

i. Afin değişmez (Affine invariance): R^d'deki herhangi bir rasgele X vektörü için A, $d \times d$ boyutunda tekil olmayan bir matris ve b, $d \times 1$ boyutunda bir vektör olmak üzere $D(Ax+b; F_{AX+b}) = D(x; F_X)$'dir.

ii. Merkezde en büyük derinliğe sahip olmalıdır (Maximality at center): Herhangi bir $F \in \mathcal{F}$ için θ merkez olmak üzere $D(\theta; F) = \sup_{x \in R^d} D(x; F)$'dir.

iii. En derin noktaya göre monoton (Monotonicity relative to deepest point): Herhangi bir $F \in \mathcal{F}$ için en derin nokta θ olmak üzere (en büyük derinlik noktası), $D(x; F) \le D(\theta + \alpha(x-\theta); F)$, $\alpha \in [0,1]$ olur.

iv. Sonsuzda sıfırlanan derinlik (Vanishing at Infinity): Her $F \in \mathcal{F}$ için $\|x\| \to \infty$ iken $D(x; F) \to 0$ olur.

Eğer X rasgele değişkeninin dağılımı $X \sim E_d(h, \mu, \Sigma)$ ile gösterilen eliptik dağılımları ailesine ait ise örneklem derinlik konturları ve yakınsamalarına ait bazı önemli özellikler de aşağıdaki gibi özetlenebilir. Sırasıyla yığın ve örneklem derinlik

fonksiyonu $D_F(X)$ ve $D_{F_n}(X)$, α derinlikli iç bölge yığın ve örneklem için D^α ve D_n^α ile gösterilsin.

i. Yarı uzay, Mahalonobis ve projeksiyon derinliklerinin derinlik konturları eliptik yüzeylerdir.

ii. Yarı uzay, simpleks ve projeksiyon derinlikleri için $n \to \infty$ iken $\alpha_n \to \alpha$ olacak şekilde herhangi bir α_n dizisi için $D_n^{\alpha_n} \to D^\alpha$ hemen hemen her yerde, $n \to \infty$ ve ∂D^α ile gösterilen eliptik α derinlik konturları ile dağılım konturları aynıdır (Liu vd. 1999), (Zuo ve Serfling 2000a).

2.4 Yığın Dağılımı için Derinlik Fonksiyonu

Yığın dağılımının derinlik fonksiyonu konusu ile ilgili derinlemesine bilgi içeren ve ulaşılabilinen tek kaynak olan Rousseeuw ve Ruts (1999)'un çalışmasından edinilmiş ve uyarlanan bilgilerdir, bu nedenle sürekli kaynak gösterilmeyecektir.

Rousseeuw ve Ruts (1999) çalışmalarında, R^d 'de bir ölçüm uzayında tanımlı herhangi pozitif ölçü veya keyfi seçilmiş bir olasılık ölçüsü temelinde yığına ait yarı uzay derinlik fonksiyonu üzerinde durmuşlardır.

R^d 'deki kümeler üzerinde tanımlı herhangi bir pozitif ölçü μ olsun. $\mu(R^d)$ ölçüsü sonlu veya sonsuz olabilir. R^d 'deki herhangi bir x noktası için bu noktanın derinliği bu ölçüye göre tanımlanabilir. Bu çalışmada (2.1)'de tanımlanan yarı uzay derinliği, söz konusu μ ölçüsü yerine P olasılık ölçüsüne göre tanımlanmıştır. Elde edilen bu fonksiyon, afin değişmezdir (afine invariance) ve yarı-konkav (quasi-concavity) özelliklerini sağlar. Aşağıdaki teorem maksimum derinliğe sahip nokta için yeterli koşulu vermektedir.

Teorem 2.1 Işın Baz (Ray Basis) Teoremi (Rousseeuw ve Ruts 1999, Proposition 8)

P olasılık ölçüsü olmak üzere bir x^* noktası için

$$\begin{cases} \forall_j : P(H_{x^*,u_j}) = D_P(x^*) \\ \bigcup_{j\in J} H_{x^*,u_j} = R^d \end{cases} \quad \text{ise} \tag{2.12}$$

olacak şekilde birim vektörlerin bir kümesi $J = \{u_1,...\}$ varsa $D_P(x^*) = \max_x D_P(x)$

ve

$$\max_x D_P(x) \geq 1/(d+1)$$

dır.

Eğer P dağılımı olasılık yoğunluk fonksiyonuna sahip ise maksimum derinlik için en küçük sınır $1/(d+1)$ dır. Bu durumda $\#J \leq d+1$ olacak şekilde birim vektörlerin kümesi $J = \{u_1,...\}$ var olduğunda x^* maksimum derinliğe sahip nokta ise Teorem 2.1'in tersi de doğrudur(Ters Işın Baz Teoremi(Inverse Ray Basis Theorem)).

Önerme 2.1 Olasılık yoğunluk fonksiyonuna sahip herhangi bir P için

$$\max_x depth(x) \leq \frac{1}{2}$$

olarak elde edilir.

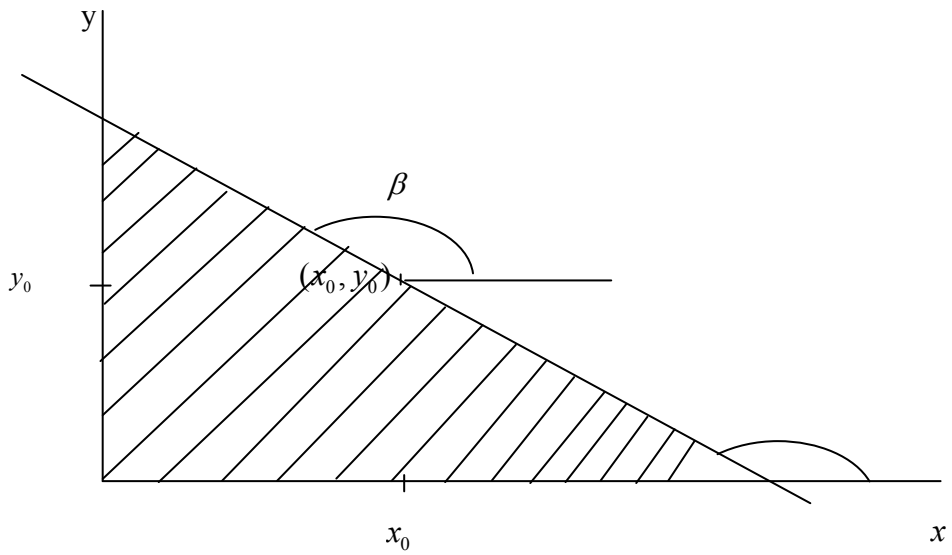
R^2 üzerinde alınan pozitif bir μ ölçüsü, Lebesgue ölçüsüne göre mutlak sürekli olduğu durumlar için f olasılık yoğunluk fonksiyonuna sahip olsun. f, kapalı konveks tanım kümesi Q'nun içinde, pozitifdir. Q nun dışında f=0 dır. Tanım kümesi Q, sınırlı veya sınırsız olabilir.

Tek köşeli ve dört köşeli konveks bölgeler için yığın derinlik fonksiyonları tanım kümesi Q üzerinde $\mu = U_Q$ ile gösterilerek düzgün μ ölçüsüne göre elde edilecektir.

Bunun anlamı c bir sabit olmak üzere $f(z)=cI\left(z\in Q\right)$ yoğunluk fonksiyonuna sahip bir ölçü olmasıdır.

Q , aynı noktadan çıkan iki ışın ile oluşturulan konveks bölge olsun. Q 'nun bir köşesi ve bir iç açısı vardır. Bölge konveks olduğundan iç açı π 'den küçük olmalıdır. Afin değişmezlik özelliği kullanılarak, köşe noktası (0,0) ve ışınlarda pozitif x ekseni ve pozitif y olarak kabul edilebilir. Bu durumda,

$\mu=U_Q$, $Q=\left\{\left(x,y\right);x\geq 0 \text{ ve } y\geq 0\right\}$ olacaktır. $\lambda\left(Q\right)=\infty$, μ , $\mu\left(Q\right)=\infty$ olan pozitif bir ölçüdür.



Şekil 2.3 Taralı üçgenin minimum alanı ile ifade edilen $\left(x_0,y_0\right)$ noktasının derinliği

Şekil 2.3 ile gösterilen tek köşeli konveks bir bölgede, düzgün ölçü üzerinde herhangi bir $\left(x_0,y_0\right)$ noktasının derinlik fonksiyonu taralı alanı minimize edecek β açısına göre bulunur. Bu alanı minimize eden β açısı $\beta=\text{Arctan}(-y_0/x_0)$ (**Ek 1**) ile elde edilir. Buna göre minimum alan $2x_0y_0$ olarak bulunur.

$\forall\ (x,y) \in Q$ noktası için derinlik fonksiyonu $D(x,y) = 2xy$ ile ifade edilir. Böylece herhangi bir $\alpha > 0$ için derinlik bölgesi,

$$D_\alpha = \left\{ (x,y) \in Q; y \geq \frac{\alpha}{2x} \right\} \text{ dir. } \alpha = 0 \text{ için } D_\alpha = Q$$

olur.

R^2'de dört köşeli konveks bölge üzerinde düzgün ölçü olduğunda yani Q=(0,1) x (0,1) olan birim alanı için $f(x,y) = I\left((x,y) \in Q\right)$ yoğunluk fonksiyonu, μ'nün olasılık dağılımını tanımlar. Derinliği hesaplamak için, kare şeklindeki Q bölgesi, dört tane ayrı tek köşeli durumun kesişimi olarak düşünülsün. Buradan,

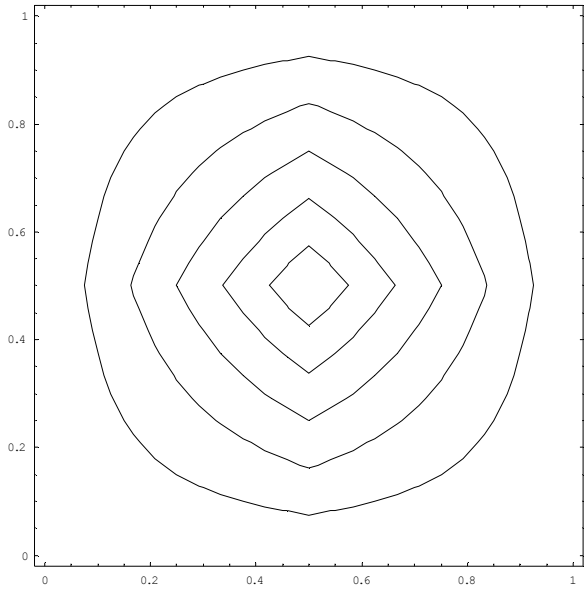
$$D(x,y) = 2\min(x,1-x)\min(y,1-y), \quad (x,y) \in Q$$

ve (2.13)

$$D_\alpha = \left\{ (x,y) \in Q; \min(x,1-x)\min(y,1-y) \geq \frac{\alpha}{2} \right\}$$

dır.

Maksimum derinlik $\alpha^* = D_F\left((x,y)^*\right)$=1/2 ve Tukey medyanı ise (1/2, 1/2) noktasıdır.



Şekil 2.4 Bir kare üzerinde düzgün dağılımın 0.05, 0.15, 0.25, 0.35, 0.45 derinlik değerlerinde karşılık gelen derinlik konturları

Şekil 2.4 ile ileriki kısımlarda sıkça karşılaşılacaktır.

2.5 Çok Değişkenli Yüzdelikler

Einmahl ve Mason (1992), d boyutlu rasgele vektörler için çok değişkenli yüzdelikleri, kümelerin bir sınıfına indeksli, genelleştirilmiş yüzdelik süreçleri olarak tanımlamışlardır.

R^d uzayında değer alan $X_1, X_2, ..., X_n$, $n \geq 1$, rasgele vektörleri için $B(R^d)$ ile gösterilen Borel kümeleri'nin bir alt kümesi olan $\mathbf{A}$ sınıfı seçilsin. Çok değişkenli p. yüzdelik için reel değerli uygun seçilen bir λ fonksiyona bağlı olarak olasılığı p ve daha büyük olan Borel kümeleri üzerinden yardımcı bir fonksiyon $0 < p < 1$ olmak üzere

$$U(p) = \inf_{A \in \mathrm{A}} \{\lambda(A) : P_X(A) \geq p, A \in \mathrm{A}\} \tag{2.14}$$

olarak tanımlanmıştır. Burada $U(p)$ yüzdelik fonksiyonu olarak isimlendirilecektir. P_X, (Ω,U,P) ölçülebilir uzayda tanımlı X vektörünün olasılık dağılımıdır.

Örnekleme dayalı yüzdelik fonksiyonunu tanımlamak için ise $0 < p < 1$ $(inf\ \phi = \infty)$ ve $B \in B(R^d)$ için örneklemden olasılık tahmini, $X_i = (X_{i1}, X_{i2}, ..., X_{id})$ olmak üzere,

$$P_n(B) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_B(X_i) \tag{2.15}$$

ile verildiğinde,

$$U_n(p) = \inf_{A\in \mathrm{A}} \{\lambda(A) : P_n(A) \geq p, A \in \mathrm{A}\} \tag{2.16}$$

gözleme dayalı (empirical) yüzdelik fonksiyonu olarak tanımlanır. $\mathbf{A}$ sınıfı ve λ uygun olarak belirlenir. λ için doğal bir seçim R^d'de λ_d Lebesgue ölçüsü olabilir ki bu durumda $U_n(p)$, $\mathbf{A}$ sınıfında bulunan eldeki verinin en az p. yüzdeliğini içeren kümelerin hacimlerinin en küçüğüdür.

$d = 1$ durumunda $\mathbf{A} = \{ (-\infty, x] : x \in R \}$ ve $\lambda((-\infty, x]) = x$ olarak belirlenirse yukarıdaki tanımlar R'de değer alan rasgele değişkenler için yüzdelik ve örneklem yüzdelik fonksiyonlarına denk olacaktır.

Chen ve Welsh (2002) ise iki boyutlu yüzdelikleri iki boyutlu dağılım fonksiyonları doğrultusunda tanımlamışlardır. Ancak boyut yükseldikçe tanımlamaları işlevsel hale getirmek zorlaşmaktadır.

Einmahl ve Mason (1992)'de bahsedilen genelleştirilmiş yüzdelik süreçleri, Serfling (2002)'de derinlik fonksiyonuna dayalı olarak tanımlanmıştır.

Olasılığı p ve p'den daha fazla en büyük iç derinlikli bölgenin sınır derinliği (α derinlikli kontur)

$$\alpha_p = sup\left\{\alpha : P\left(D^\alpha\right) \geq p\right\} \tag{2.17}$$

dır. Ayrıca $\sup_{x\in R} D\left(x;P\right) = \alpha^*$ ile gösterilecektir. Einmahl ve Mason (1992)'de tanımlanan $\mathbb{A}$ sınıfı, $\forall x \in R^d$ için derinlik fonksiyonu $D\left(x,P\right)$'in belirlediği nokta kümeleri kullanılarak

$$\mathbb{A}= \left\{D^\alpha : \alpha \in \left(0,\alpha^*\right)\right\}$$

olarak tanımlanmıştır.

λ, $\mathbb{A}$ sınıfı üzerinde tanımlı reel değerli bir fonksiyon olmak üzere, λ_d fonksiyonu, $\alpha \in \left(0,\alpha^*\right)$ için $\lambda_d\left(D^\alpha\right)$ sonlu, α'ya göre göre kesin azalan ve sürekli türeve sahip olmak üzere $0 < p < 1$ olmak üzere ilgili genelleştirilmiş yüzdelik fonksiyonu

$$\begin{aligned} U\left(p\right) &= \inf_{\alpha}\left\{\lambda_d\left(D^\alpha\right) : P\left(D^\alpha\right) \geq p\right\} \\ &= \lambda_d\left(D^\alpha\right) \end{aligned} \tag{2.18}$$

olacaktır.

Derinlik fonksiyonları birçok uygulama alanında kullanılmaktadır. Liu ve Singh (1993)'de istatistiksel kalite kontrol süreçleri ve derinliklere dayalı kalite indeksini tanımlamışlardır. Böylece çok değişkenli kalite kontrol süreçlerinde kontrol işlemi (monitoring) derinliklere dayalı olarak tek boyutlu kalite kontrol süreçlerinde olduğu gibi yapılabilmektedir. Regresyon yöntemi altında derinliklerin kullanıldığı uygulamalar da vardır (Rousseeuw ve Hubert 1999). Çok değişkenli dağılımların parametrik olmayan sonuç çıkarımlarında, örneğin; konum, saçılım, simetrik olmama ve basıklık gibi betimsel ölçüler ve Einmahl ve Mason (1992) tarafından tanımlanan çok değişkenli yüzdelikler derinliklere dayalı olarak tanımlanabilmektedir (Liu vd. 1999, Serfling 2002, Serfling 2004, Wang ve Serfling 2005). Derinliklere dayalı olarak çok

değişkenli parametrik olmayan sonuç çıkarımı hakkında geniş bilgi ve kaynakçada Serfling (2006)'de yer almaktadır.

2.6 Tek Değişkenli Fark İstatistikleri (Univariate Spacings)

Derinliğe göre sıralanmış ardışık iki sıra istatistiği (depth order statistics) için fark istatistiği (spacing), bu iki sıra istatistiğinin sahip oldukları derinlik değerleri arasındaki derinlik değerlerine sahip bütün noktaları içeren bölge olarak tanımlanır. Bu şekildeki bir tanım, çok değişkenli tolerans bölgelerinin çoğu kez oluşturulduğu gibi şeklinin ya da geometrik biçiminin nasıl olacağının önceden belirlenmesi sonucunu da ortadan kaldırır. Bu tolerans bölgesi verinin gözlemlendiği yığının olasılık yapısının öngördüğü geometriye de uyar ve tolerans bölgelerinin istenen çoğu özelliklerini (minimum hacimli, minimum alanlı, parçalı olmayan vs.) sağlar.

$X_1, X_2, ..., X_n$, tanım kümesi (a,b) aralığı olan tek değişkenli sürekli bir F dağılımından rasgele bir örneklem olsun. Bu örnekleme ilişkin sıra istatistikleri $X_{[1]} \le X_{[2]} \le ... \le X_{[n]}$ ile gösterilsin. Bu örneklem için $\left(X_{[i-1]}, X_{[i]}\right]$, $i = 1, 2, ..., n+1$ aralıkları ile $X_{[0]} = a$ ve $X_{[n+1]} = b$ olmak üzere bu aralıkların uzunlukları $D_i = X_{[i]} - X_{[i-1]}$ fark istatistikleri(spacings) olarak adlandırılır; ardışık sıra istatistikleri arasındaki fark olarak da özetlenebilir (Pyke 1965, Li ve Liu 2008).

$X_{[1]} \le X_{[2]} \le ... \le X_{[n]}$ sıra istatistikleri için $m < n$ tane sıra istatistiği içeren aralık, m. sıra fark istatistikleri veya m. sıra boşluğu olarak adlandırılır. $X_{[m+i]} - X_{[i]} = D_{i+1} + D_{i+2} + ... + D_{i+m}$, $i = 0,1,...,n-m+1$ ile gösterilir (Lin 1993).

F dağılım fonksiyonuna sahip herhangi bir X sürekli rasgele değişkeni için olasılık integral dönüşümü kullanılırsa $F(X)$ de $U(0,1)$ düzgün dağılımına sahip olur. Bu nedenle $F \sim U(0,1)$ ise

(i) $D_1 + D_2 + ... + D_{n+1} = 1$ ve

(ii) $\left(D_1, D_2, ..., D_{n+1}\right)$'in yoğunluk fonksiyonu

$$f\left(d_1, d_2, ..., d_{n+1}\right) = \begin{cases} n! & eğer\ d_i \geq 0\ ve\ d_1 + d_2 + ... + d_{n+1} = 1 \\ 0 & diğer\ durumlarda \end{cases} \tag{2.19}$$

dir. Böylece f yoğunluk fonksiyonu değişkenlerine göre simetriktir, yani $d_1, d_2, ..., d_{n+1}$'nin olasılık yoğunluk fonksiyonu içinde yer değiştirmeleri halinde olasılık yoğunluk fonksiyonu değişmez. Ortalaması 1 olan üstel dağılımdan $U_1, U_2, ..., U_{n+1}$ örneklemi için $S = U_1 + U_2 + ... + U_{n+1}$ ve $W_i = U_i / S$, $i = 1, 2, ...n+1$ olsun. Bu durumda $\left(W_1, W_2, ..., W_{n+1}\right)$ ve $\left(D_1, D_2, ..., D_{n+1}\right)$ aynı dağılımlıdır (Li ve Liu 2008).

Fark istatistikleri kendi içlerinde yer değiştirebilen rasgele değişkenlerdir. Her bir rasgele değişken $Beta(1, n)$ dağılımına sahiptir. Bu, Huffer ve Lin'in tarama istatistiği hesaplamalarında kullanılan önemli bir özelliktir. Fark istatistikleri için diğer bir önemli özellikte herhangi bir $\Delta \subset \{1, 2, ..., n+1\}$ kümesi için fark istatistiklerinin toplamı $D(\Delta) = \sum_{i \in \Delta} D_i$ ile gösterildiğinde birbiriyle örtüşmeyen(non-overlapping) fark istatistiklerinin toplamlarının ortak dağılımı Dirichlet dağılımına sahip olmasıdır (Pyke 1965, Lin 1993).

2.7 Tarama İstatistikleri (Scan Statistics)

Geriye dönük (retrospective) ya da ileriye yönelik (prospective) olarak istatistiksel çıkarımın hedeflendiği durumlarda amaca göre tarama istatistikleri (scan statistics) tanımlanabilir. Tarama istatistikleri, çoğu kez rasgele olayların zaman ve/veya uzay boyutunda varsa kümelenmenin anlamlı olup olmadığını belirlemek amacıyla kullanılmaktadır. Geçtiğimiz on yıllık süreç içerisinde, uygulamalı olasılık ve istatistik alanında, tarama istatistikleri üzerine yapılan araştırmaların sayısında hayli artış olduğu göze çarpmaktadır. Tarama istatistiklerinin momentleri, kesin ve yaklaşık olasılık hesaplamaları için farklı yaklaşımlar bulunmaktadır. Ayrıca astronomi, genetik,

güvenilirlik ve kalite kontrol, telekomünikasyon, epidemiyoloji, coğrafi bilgi sistemleri gibi birçok teknoloji ve bilim alanında da tarama istatistiklerinin uygulamaları görülmektedir. Tarama istatistikleri kesikli ve sürekli uzayda, tek boyutlu ve çok boyutlu olarak tanımlanabilir.

Kesikli tarama istatistikleri T denemenin bir dizisi üzerinde tanımlanırlar. Kesikli tarama istatistiğinin özel bir durumu, Bernoulli denemelerinin bir serisinde en uzun "başarılı" dizinin uzunluğudur. n deneme içindeki herhangi ardışık m başarılı denemelerin maksimum sayısı S_m ile gösterilen kesikli tarama istatistiğidir (Glaz vd. 2001). $X_1, X_2, ..., X_n$ tam sayı değeri alan rasgele değişkenler ise $2 \le m \le n-1$ için herhangi m ardışık gözlemin toplamı $Y_t = \sum_{i=t}^{t+m-1} X_i$ olmak üzere lineer *koşulsuz* kesikli tarama istatistiği $S_m = \max_{1 \le t \le n-m+1} Y_t$ ile tanımlanır. Koşulsuz durumda dairesel kesikli tarama istatistiği ise $S_m^* = \max_{1 \le t \le n} Y_t$'dir. Kesikli tarama istatistiklerinin kullanımı DNA veya protein dizilerinin eşleştirilmesi gibi uygulama alanlarında sıkça görülmektedir. $X_1, X_2, ..., X_n$ tam sayı değeri alan rasgele değişkenler için $\sum_{i=1}^{n} X_i = a$ koşulu altında, $S_m \left(S_m^* \right)$, lineer-doğru üzerinde anlamında (dairesel) *koşullu* kesikli tarama istatistiği ile ifade edilir. Birbirinden bağımsız 0-1 Bernoulli denemelerinin özel bir durumu için Naus (1974), $P\left(S_m \ge k \left| \sum_{i=1}^{n} X_i = a \right. \right)$ için tam olasılık hesabı elde etmiştir. Kesikli durumda koşullu tarama istatistiklerinin en çok kullanıldığı uygulama alanlarından bazıları meteoroloji, maden arama, moleküler biyoloji, kalite kontrol ve güvenilirlik teorisidir (Glaz ve Balakrishnan 1999).

$\{X_t,\ t \ge 0\}$, λ oranlı bir Poisson süreci olsun. X_t, $(0,t)$ aralığında ortaya çıkan noktaların (olayların) sayısıdır. $(t, t+w)$ aralığında meydana gelen olayların sayısı için tarama süreci $Y_t(w) = X_{t+w} - X_t$, $0 < w < T$ olsun. Sürekli uzayda tek boyutlu durumda *koşulsuz* tarama istatistiği $S_w = S_w(\lambda, T) = \max_{0 < t \le T-w} Y_t(w)$ ile tanımlanır. $(0,T)$ aralığı Poisson sürecinin gözlendiği toplam aralık olarak ifade edilir (Glaz ve Balakrishnan 1999). Gözlenen olay sayısı bir rasgele değişken olarak verildiğinde bu tarama

olasılıklarının hesaplanması ya da taramanın yapılması koşulsuz tarama istatistiği olarak yazılır. Tek boyutlu durumda koşulsuz tarama istatistiği bilim ve mühendislik, epidemiyoloji, moleküler biyoloji, kuyruk teorisi, telekomünikasyon gibi uygulama alanlarında kullanılmaktadır. S_w'nun kesin (exact) dağılımının bulunması hakkındaki genel bilgiler Wallenstein ve Naus (1974) ve Huntington ve Naus (1975)'da yer almaktadır.

$X_1, X_2, ..., X_n$, $(0,1)$ aralığında bağımsız düzgün dağılıma sahip rasgele değişkenler ve $0 < w < 1$ ve $0 < t \leq 1 - w$ için $Y_t(w)$, $(t, t+w)$ aralığındaki gözlemlerin sayısı olsun. Bu durumda lineer *koşullu* tarama istatistiği

$$S_w = \max_{0<t\leq 1-w} Y_t(w) \tag{2.20}$$

ile tanımlanır. W_k, $(0,1)$ aralığı üzerinde k sayıda olay içeren en küçük aralık uzunluğu olmak üzere özel bir durumu W_n, örneklem genişliğidir. W_{r+1} aralığı, en küçük r.sıra boşluğu (gap) veya r. tarama istatistiği olarak adlandırılır. S_w ve W_k istatistiklerinin dağılımları ilişkilidir, $P(W_k \leq w) = P(S_w \geq k)$ (Glaz vd. 2001). S_w istatistiği, gözlemlerin $(0,1)$ aralığında düzgün dağılıma sahip olduğu yokluk hipotezini kümelenmenin var olduğu (düzgün dağılımlı olmadığı) alternatif hipotezine karşı test etmede kullanılır (Glaz ve Balakrishnan 1999, Glaz vd. 2001). Eğer w uzunluğundaki tarama penceresi k ve daha fazla sayıda gözlem içerirse gözlemlerin düzgün dağılıma sahip olduğu yokluk hipotezi reddedilir. Bu test sürecinde doğal olarak $P(k; n, w) = P(S_w \geq k)$ kuyruk olasılıkları değerlendirilmektedir (Glaz ve Balakrishnan 1999). $(0,1)$ aralığında gözlenen sabit n sayıda gözleme ilişkin istatistiksel çıkarımın yapılması geriye dönük tarama istatistiği uygulamaları olarak tanımlanır. Bu olasılık n'nin verilmiş olan değerine bağlıdır. Zaman içindeki geçmişe dönük taramada koşul n üzerinde tanımlanır. Eğer herhangi bir $(0,t)$ zaman aralığında ortaya çıkan olayların sayısı henüz bir sabit sayı olarak belirlenmemiş ise bu zaman sürecindeki noktaların sayısı sabit bir n sayısı olarak görülmez; N rasgele değişkenin aldığı bir değer olarak

görülür. Örneğin, sabit uzunluklu bir zaman aralığında ortaya çıkan toplam gözlem sayısının dağılımı λ ortalamalı Poisson dağılımına sahip olduğu varsayılabilir. Bu tip uygulamalar ise geleceğe dönük (prospective) tarama istatistiği uygulamalarıdır. Geleceğe dönük taramalar koşulsuzdur (Glaz vd. 2001).

Tarama istatistiklerinin olasılık hesaplamalarında farklı yöntemler bulunmaktadır. Bunlardan bazıları sıra istatistikleri ve direk integrasyon, rasgele yürüyüş (random walk) ve yansıma kuralıdır (reflection principle). Belirli durumlarda ($k = 2$, k=n vb.), Parzen (1960), Frosini (1981), Huffer ve Lin (1997)'in tarama istatistiklerinin dağılımları için sıra istatistikleri ve direk integrasyon yaklaşımı bulunmaktadır. Naus (1965a), $k > (n+1)/2$ olduğunda $w \geq 1/2$ ve $w \leq 1/2$ için $P(k;n,w)$ olasılığı için alternatif bir kombinatorik yaklaşım elde etmiştir. Naus (1965a), bir doğru üzerindeki noktaların maksimum büyüklükteki kümesinin dağılımının bulunması problemini; $(0,1)$ aralığı üzerinde düzgün dağılımdan bağımsız olarak çekilen n nokta için $(0,1)$ üzerindeki herhangi bir w uzunluklu aralığın en az k tane gözlem içermesi olasılığını $P(k;n,w)$ olarak tanımlamıştır. Naus (1965a), $P(k;n,w)$ olasılığına yaklaşımı olan $P'(k;n,w)$'yi aktarabilmek için aşağıdaki gösterimler gerekmektedir:

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

$$F_b(k;n,w) = \sum_{i=0}^{k} b(i;n,w)$$

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

$$H(k;n,w) = \frac{n(1-w)}{w} F_b(n-k;k-1,w/1-w) - (n-k+1) F_b(n-k+1;k,(1-w)/w)$$

gösterimleri kullanıldığında aşağıdaki teoremde yer alan $C(k;n,w)$ ve $R(k;n,w)$ tanımlanacaktır:

$$
\begin{aligned}
C(k;n,w) &= (n-k+1)\,b(k-1;n,w) - (n-k)\,b(k;n,w) \\
&\quad + G_b(k;n,w) + G_b(k+1;n,w) \\
&= (n-k+1)\left[G_b(k-1;n,w) + G_b(k+1;n,w)\right] \\
&\quad - 2(n-k)\,G_b(k;n,w)
\end{aligned}
$$

ve

$$
R(k;n,w) = \sum_{y=k}^{n} b(y;n,w)\,F_b(n-k;y,(1-w)/w) + H(k;n,w)\,b(k;n,w)
$$

Teorem 2.2 (Naus 1965a, Theorem I) $P(k;n,w)$ olasılığı,

$$
\begin{aligned}
P'(k;n,w) &= C(k;n,w) - R(k;n,w), \quad w \geq 1/2, \quad k > (n+1)/2 \\
&= C(k;n,w), \qquad\qquad\qquad\quad w \leq 1/2, \quad k > n/2
\end{aligned}
\tag{2.21}
$$

dır.

Wallenstein ve Naus (1973) tüm k,n ve w için $P(k;n,w)$ olasılığının hesaplanması için genel bir formül vermişlerdir.

Birim uzunluklu doğru, her biri eşit $1/L$ uzunluğunda ayrık L parçaya bölünsün. Hücreler içinde bulunan gözlem sayıları $k_1, k_2, ..., k_L$ ile gösterilmek üzere

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

ve

$$
V_L(n,r) = \left\{ (k_1,k_2,...,k_L) \,\middle|\, k_i \geq 0, i = 1,...,L; J(1,n) = n \text{ ve } i \leq L\text{-}r\text{+}1 \text{ için } J(i,i+r-1) < k \right\}
$$

olsun.

Teorem 2.3 (Wallenstein ve Naus 1973) En büyük ortak bölenleri 1 olan r ve L pozitif tamsayılar olmak üzere $0 < r/L < 1$ ve $2 \le k \le n$ olan $k,\ n$ tamsayıları için elemanları

$$D_{a,b}^{k} = 1/\left[(b-a)k - J\left(s+1+[a-1]k\right), s-1+[b-1]r\right]!, \quad a < b$$
$$= 1/\left[(b-a)k + J\left(s+[b-1]r, s+[a-1]r\right)\right]! \quad , \quad a \ge b$$

olan D^k matrisi ve kolaylık açısından $x < 0$ için $1/x! = 0$ alınmak üzere;

$$P\left(k; n, r/L\right) = 1 - n!\, L^{-n} \sum_{V_L(n,r)} \prod_{s=1}^{r} det\, D^k$$

dır.

$$e_k = [L/r] - 1, \ kr > L - r$$

$$= [L/r] \quad , \ kr \le L - r$$

olmak üzere D^k matrisinin boyutu, $e_k + 1$'dir, burada $[x]$, x'in tam kısmıdır.

$P(k; n, w)$ olasılığı için değişik yaklaşım yöntemleri bulunmaktadır. Bu yaklaşımlarla ilgili detaylı bilgi için diğerleri yanında Berman ve Egleson (1985), Gates ve Wescott (1984), Glaz (1989,1992), Glaz, Naus, Roos, Wallenstein (1994), Naus (1982), Wallenstein and Neff (1987) ve Huffer ve Lin (1997)'in çalışmalarına bakılabilir. Huffer (1988)'ın düzgün dağılıma sahip n nokta için, fark istatistiklerinin doğrusal birleşimlerinin ortak dağılımları, kesin olasılık hesapları ve düzgün dağılıma uyum testleri ile ilgili çalışmaları bulunmaktadır.

3. DERİNLİKLERE DAYALI ÇOK DEĞİŞKENLİ TARAMA İSTATİSTİKLERİ

3.1 Giriş

$X_1, X_2, ..., X_n$ birbirinden bağımsız ve aynı F dağılım fonksiyonuna sahip d boyutlu rasgele vektörlerinin n çaplı bir örneklemi olsun. Örneğin bir coğrafi bölgede kanser gözlemlenen bireylerin bulundukları konumlar $d = 2$ boyutlu rasgele bir vektör olarak değerlendirilebilir. İki veya daha çok boyutta rasgele gözlemlerin olağan dışı kümelenmeleri için birçok alanda tarama istatistikleri kullanılabilir. Kanser vakalarının mekânsal kümelenmesi, belirli bir bölgede maden rezervlerinin konumları gibi araştırma konularında iki boyutlu tarama istatistikleri söz konusu olmaktadır. Tek boyutlu tarama istatistiği uygulamalarında sürekli zaman aralığında ya da aralık yerine kesikli denemelerin söz konusu olduğu durumlarda tarama penceresi olarak, belirlenen bir aralık kullanır. İki boyutlu tarama istatistiği uygulamalarında ise tarama pencereleri kare, dikdörtgen, çember, üçgen biçimli seçilebilirler. Tarama penceresi olarak birim kare, birim aralığın doğal bir genelleştirmesidir. İki boyutlu tarama istatistiği, şekli ve boyutları belirlenen bir tarama penceresine düşen gözlemlerin maksimum sayısı olarak tanımlanır (Glaz vd. 2001).

Tipik bir iki boyutlu tarama istatistiği betimlemek için birim kare üzerinde rasgele seçilen n nokta olduğu düşünülsün. $S_{u,v}$ tarama istatistiği, birim karenin kenarlarına paralel u uzunluklu, v genişlikli herhangi bir dikdörtgendeki maksimum nokta sayısı olsun. İki boyutlu durumda bu tarama penceresinde k veya daha fazla gözlem bulunması olasılığı diğer bir ifade ile tarama istatistiğinin olasılığı $P(k;n,u,v) = P(S_{u,v} \geq k)$ ile gösterilir. Naus (1965b), küçük u ve v değeri için $P(k;n,u,v)$ olasılığına sınırlar önermiştir ve $P(k;n,u,v)$ olasılığı $P(k;n,u,v) \cong k^2 \binom{n}{k} (uv)^{k-1}$ ile hesaplamıştır. Loader (1990, 1991) ise kuyruk olasılıklarına (large deviation theory) dayanarak bu olasılığa aşağıdaki yaklaşımı önermiştir:

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

$$w=uv$$

$$E=(k/nw)-1,$$

$$C=\{nv(1-u)E/(1-w)\}+\{nu(1-v)E^2/(1+E)(1-w)^2\}$$
$$+\{(1+E)(1-w)/E\}$$

olmak üzere,

$$P(k;n,u,v)\cong\left(\{n^2w(1-u)(1-v)E^3/(1-w)^3(1+E)\}+C\right)b(k;n,w) \qquad (3.1)$$

dır.

İki boyutlu tarama istatistiklerinin bir uygulaması Glaz vd. (2001) (özgün hali Hjalmars vd. 1996'da verilmiştir)'de yer almaktadır. Sözkonusu bu uygulamada 20 yıllık bir süreçte İsveç'te 15 yaşından küçük çocuklarda görülen lösemi vakalarının kümelenmesi incelenmiştir ve farklı büyüklük ve şekillerdeki tarama pencereleri için sonuçlar elde edilmiştir. Glaz vd. (2001)'de iki boyutlu tarama istatistiğinin kullanımı aynı veriler için şöyle özetlemişlerdir: 20 yıllık bir sürede 15 yaşın altındaki toplam çocuk nüfusu 1.703.235 olan İsveç'te 1534 tane lösemi vakasının görüldüğü saptaması yapıldıktan sonra İsveç'in güneybatısındaki Okome'de yaşayan 133 çocuktan 3 tanesinde lösemi vakasının görülmesinin olağan olup olmadığı sorgulanmıştır.

Toplam nüfusun İsveç haritası üzerinde eşit olarak dağılmadığı görülmektedir. Ancak Glaz vd. (2001) bu uygulamayı düzgün dağılımlı bir İsveç haritası üzerindeymişcesine yapmışlardır. Birim karede tanımlı iki boyutlu düzgün dağılım kurgusuna uyum sağlaması için bu harita, her birey 1 birim kareyi temsil edecek şekilde kenar uzunluğu 1305 br olan bir kare $\left(\sqrt{1.703.235}=1305\right)$ olarak düşünülmüştür. Aynı düşünce ile Okome bölgesi de, bu karesel bölge içinde kenar uzunluğu 11.5 birim olan bir kare

$\left(\sqrt{133}=11.5\right)$ olarak alınmıştır. Sonrada tüm kare birim kareye dönüştürülmüştür. Böyle yaparak lösemili çocukların görüldüğü konum, birim kare de tanımlı düzgün dağılımlı olduğu varsayılmıştır. Bu bölgenin altkümesi olmak üzere kenar uzunlukları $u=v=11.5/1305=0.0088$ olan ve uygulamada tarama penceresi olarak kullanılacak bir kare elde etmişlerdir. "İsveç'te ortaya çıkan 1534 lösemi vakası içinde Okome bölgesinde 3 lösemi vakasının kümelenmesi olağan mıdır?" sorusu $P\left(3;1534,0.0088,0.0088\right)$ olasılığı yaklaşık olarak hesaplanarak cevaplandırılmaya çalışılmıştır. $P\left(k;1534,0.0088,0.0088\right)$ olasılığı (3.1)'de tanımlanan Loader yaklaşımı ile $k=3,4$ için 1'den büyüktür. $P\left(5;1534,0.0088,0.0088\right)\cong 0.052$ ve $P\left(6;1534,0.0088,0.0088\right)\cong 0.0015$ dir. 133 çocuğun olduğu bir yığında lösemi hastası 3 çocuğun olması istatistiksel olarak anlamlı değildir. Yani o bölge için olağan dışı bir durum söz konusu olmadığı sonucuna varılmıştır. Hjalmars vd. (1996), olağan dışı kümelenmenin olmadığı (gözlemlerin konumlarının düzgün birim düzlemde düzgün dağıldığı) yokluk hipotezinin alternatifine karşı testinde p değerini de 0.697 olarak hesaplamışlardır.

Dörtgensel bir bölgede yer alan noktaların sayısı, λ ortalamalı Poisson dağılımına sahip bir rasgele değişken olduğu durumda tarama, tarama pencereleri sabit alanlı fakat değişken şekillere sahip bölgeler olacak şekilde yapılabilir. Bu durumda aynı tarama problemi için aynı alana sahip farklı şekillerdeki tarama pencerelerinin seçilmesinin tarama olasılıklarını değiştirdiği gözlemlenmiştir. Örneğin, Alm (1999), sürekli ve koşulsuz (geleceğe dönük) durumda gözlemlerin iki boyutlu bir $A=\left[0,T_1\right]\times\left[0,T_2\right]$ gibi bir tarama bölgesinde λ parametreli X Poisson sürecinde iki boyutlu tarama istatistikleri için bazı yaklaşımlar geliştirmiştir ve karesel, üçgensel ve dairesel tarama pencereleri için yaklaşık tarama olasılıklarını bulmuştur. Farklı şekillerdeki tarama pencereleri için bu olasılıkların farklılaştığı gözlemlenmiştir.

3.2 Derinliklere Dayalı Çok Değişkenli Tarama İstatistikleri

Çalışmanın bir sonraki bölümünde ifade edilecek olan çok değişkenli veriler için tarama istatistiklerine ilişkin kesin (exact) olasılıkların hesaplanmasında tarama pencereleri, derinliklere dayalı olarak tek boyutlu verilerde olduğu gibi aralık şeklinde elde edilmiştir. Lin (1993) ve Huffer ve Lin (1997), fark istatistiklerinin toplamlarının minimumlarının (ya da maksimumlarının) dağılımını bulmak için genel bir yöntem geliştirmişlerdir. Bu çalışmada da söz konusu yaklaşım, tarama istatistiklerinin dağılımının bulunmasında kullanılmaktadır. Bu yaklaşımın tanımlanması ve bir örnek **Ek 2**'de yer almaktadır.

Bu çalışmada çok boyutlu tarama istatistiklerinin hesaplanmasında tarama penceresinin seçimi için bir yaklaşım önerilmiştir. Bu önerinin bir tarama istatistiğinin işlevlerini yerine getirdiği gibi bir ölçüde dağılıma uygunluk testi işlevine de sahip olduğu düşünülmektedir. İlk olarak çok boyutlu tarama istatistikleri, Li ve Liu (2008)'nun çok değişkenli tolerans bölgeleri için tanımlamış oldukları çok değişkenli fark istatistiklerine bağlı olarak tanımlanmıştır. Kümelenmenin varlığı ve sıra dışılığın saptanması, gözlemlerin düzgün dağılımdan gelmediğinin istatistiksel olarak gösterilmesi anlamında değerlendirilmiştir.

Çok boyutlu veriler için tarama istatistiklerinin kullanımında en önemli problemlerden biri tarama penceresinin seçimidir. Glaz vd. (2001)'de tarama penceresi şeklinin ve büyüklüğünün, gözlemlerin tarama penceresine düşme olasılıklarını değiştirdiği, tarama penceresinin $u = v$ olacak şekilde bir kare alındığında $P(k;n,u,v)$ olasılığının diğer tarama şekillerine göre en yüksek olasılığa sahip olduğu belirtilmiştir (s.65). Anderson ve Titterington (1997)'da farklı şekildeki tarama pencereleri için tarama istatistikleri olasılıkları Monte Carlo yaklaşımı ile elde edilmiştir.

Yığın için tanımlanan çok değişkenli fark istatistikleri (MS_i) çoğu kez gözlemlenen verinin yönlendirdiği şekilde bir halka olarak biçimlenmiş olduğundan bunlara derinlik halkaları da denilecektir.

Tanım 3.1 $X \in R^d$, $d \geq 2$ olan F dağılım fonksiyonlu n çaplı rasgele örneklemin derinliklere göre sıralanmış istatistikleri $X_{[1]}, X_{[2]}, ..., X_{[n]}$, $i = 1, 2, ..., n+1$ ve $i.$ gözlemin yığın derinlik değeri $Z_i = D_F\left(X_i\right)$ ve $i.$sıradaki istatistiğin derinlik değeri $Z^{[i]} = D_F\left(X_{[i]}\right)$olsun. $Z^{[0]} = sup_x\left\{D_F\left(x\right)\right\}$ ve $Z_{[n+1]} = 0$ olmak üzere,

$$MS_i = \left\{X : Z^{[i-1]} \geq D_F\left(X\right) > Z^{[i]}\right\},\ i = 1, 2, ..., n+1 \tag{3.2}$$

kümesi yığın için çok değişkenli fark istatistikleri olarak adlandırılır. Örneklem için örneklem derinlikleri $D_{F_n}\left(X\right)$ kullanılarak elde edilen fark istatistikleri (rasgele element olarak fark kümeleri) $\hat{Z}^{[0]} = sup_x\left\{D_{F_n}\left(X\right)\right\}$ ve $\hat{Z}^{[1]} \geq \hat{Z}^{[2]} \geq ... \geq \hat{Z}^{[n]}$ için $\hat{Z}_i = D_{F_n}\left(X_i\right)$, $i = 1, 2, ..., n$ olmak üzere

$$\hat{MS}_i = \left\{X : \hat{Z}^{[i-1]} \geq D_{F_n}\left(X\right) > \hat{Z}^{[i]}\right\},\ i = 1, 2, ..., n$$

ve (3.3)

$$\hat{MS}_{n+1} = \left\{X : D_{F_n}\left(X\right) \leq \hat{Z}^{[n]}\right\}$$

dır.

Sonuçta çok değişkenli fark istatistikleri (spacings) içerden dışarıya doğru (tıpkı kuru soğan katmanları gibi) kabuk katmanları olacaktır. Aşağıdaki teorem gerek tolerans bölgelerinin oluşturulmasına gerekse fark istatistiklerinin (spacings) bir formu olan tarama istatistiklerine (scan statistics) ilişkin bir çalışmada önemli bir işleve sahip olacaktır (Li ve Liu 2008).

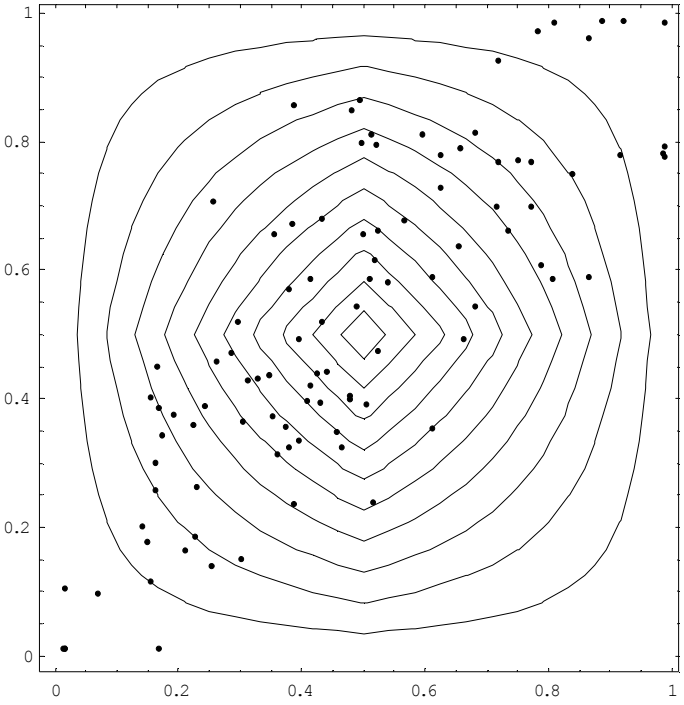
Teorem 3.1 (Li and Liu 2008, Theorem 2.1) $X_1, X_2, ..., X_n$, $F \in R^d$'de tanımlı sürekli bir dağılım fonksiyonundan bir örneklem olsun. Yukarıda tanımlanan çok değişkenli fark istatistiklerinin elde edilmesinde kullanılan derinlik fonksiyonlarının afin değişmez olduğunu varsayalım. Bu durumda çok değişkenli farkların kapsama olasılıkları $P_F\left(MS_1\right), ..., P_F\left(MS_{n+1}\right)$ yani gözlemin bu bölgeye düşme olasılığı tek

değişkenli düzgün dağılımın fark istatistiklerinin $D_1,...,D_{n+1}$ dağılımı ile aynıdır (**Ek 3.** Teoremin İspatı).

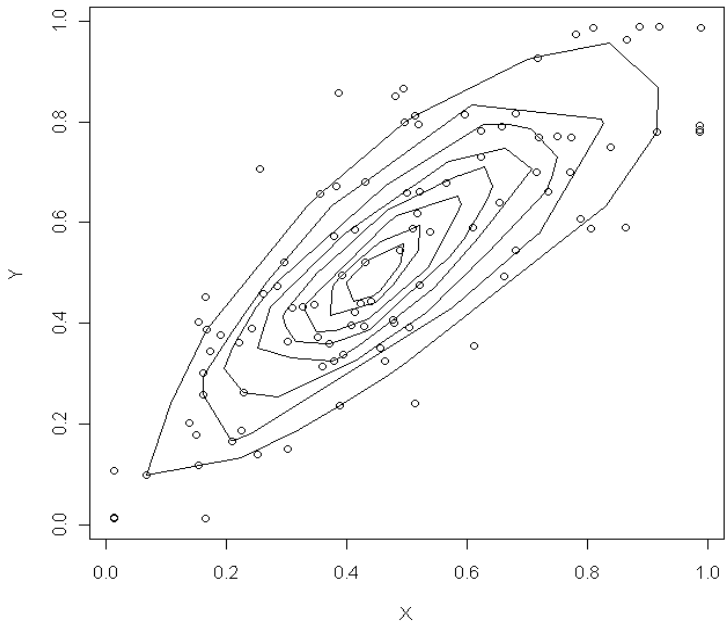
Dağılım ne olursa olsun dağılımın yapısından gelen ardışık derinlik konturları arasında kalan bölgeler için $T_i = P_F\left(X : D_F\left(X\right) > Z_i\right)$ olasılıkları düzgün dağılıma sahiptir. Ancak her çok değişkenli dağılımın derinlik konturlarının geometrik şekli farklıdır. Her veri için kendi derinlik konturları içindeki bu olasılıklar düzgün dağılımlı olacaktır. Bu nedenle gözlemler farklı bir dağılımdan gelmişler ise bu gözlemlerin düzgün dağılımın yığın derinlik konturları içindeki saçılımı bu derinlik konturlarına uymayan, düzgün dağılımlılığı yansıtmayacak şekilde olacaktır.

Düzgün dağılımın yığın derinlik konturunun yapısını, iki boyutlu düzgün ve normal dağılımlardan gelen gözlemlerin düzgün dağılım içindeki konumlanmaları ve dağılımların kendi geometrik yapılarına göre oluşan derinlik konturları şekil 3.1-3.4 ile gösterilmiştir.

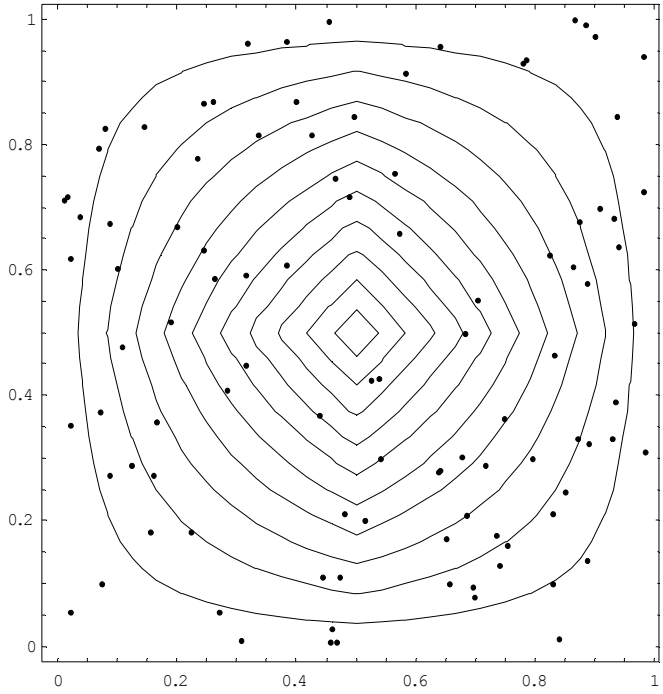
Verilerin ilgili konu çerçevesinde değerlendirilmesi, grafik ve çizelgelerin elde edilmesinde farklı programlardan yararlanılmıştır. Normal ve düzgün dağılımlı örneklemler, dağılımlardan rasgele sayı üretme tekniği ile *Matlab 2006a* programında elde edilmiştir. Düzgün dağılımın yığınına ait derinlik konturlarının çizdirilmesi ve belirli derinliklere karşılık gelen olasılık hesaplamalarında *Mathematica for Students 5.2* programı, düzgün ve normal dağılımlı yığından çekilen örneklemlerin kendi verilerine göre yarı uzay derinlik değerleri, derinlik konturları *R-project* programının *derinlik modülü(depth packages),* uygulama bölümünün simülasyon aşamasında 100 örneklem için sabit bir tarama penceresine düşen gözlem sayılarının bulunması ve genel matematiksel hesaplamalarının yapılmasında *Excel 2007* programı kullanılmıştır.



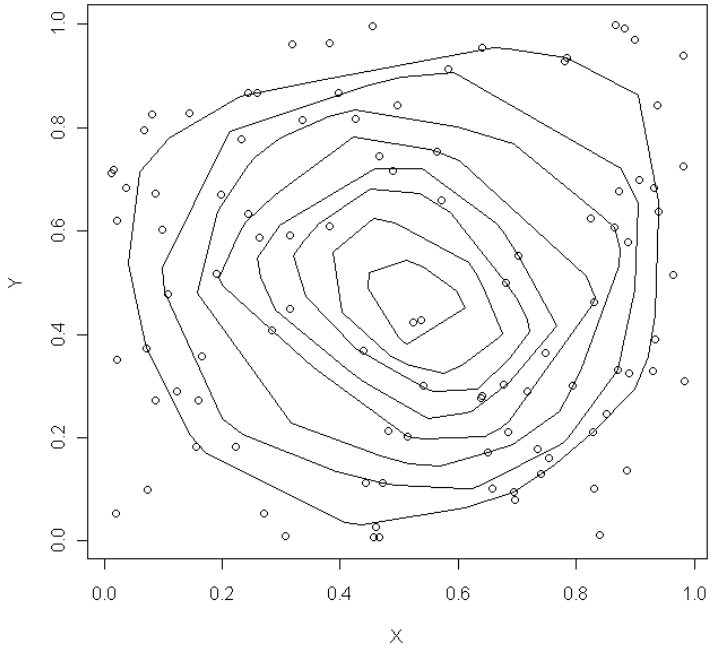
Şekil 3.1 Düzgün dağılımın yığınına ait derinlik konturları içinde 2 boyutlu normal dağılımlı 100 gözlemin saçılımı



Şekil 3.2 İki boyutlu normal dağılımlı 100 gözlemin kendi verisine göre elde edilen 0.05, 0.10, 0.15, 0.20,0.25, 0.30, 0.35, 0.40 derinlikli derinlik konturları



Şekil 3.3 Düzgün dağılımın yığınına ait derinlik konturları içinde 2 boyutlu düzgün dağılımlı 100 gözlemin saçılımı



Şekil 3.4 İki boyutlu düzgün dağılımlı 100 gözlemin kendi verisine göre elde edilen 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40 derinlikli derinlik konturları

Gözlemlenen derinlik değerlerini hipoteze konu olan dağılımın derinlik dağılım fonksiyonuna göre değerlendirmek yetecektir. Herhangi bir dağılımdan gelen bir örneklem düzgün dağılımın yığını içinde değerlendirildiğinde, gözlemlerin düzgün dağılımın yığın derinliklerine göre elde edilen ardışık derinlik konturları arasındaki bölgelerin olasılıkları eşit olmayacaktır.

Eğer örneklemin geldiği yığın düzgün dağılımlı ise örnekleme ilişkin derinlik değerlerinin dağılımı $(0,1)$ aralığında düzgün dağılımlı olacaktır; gözlemler düzgün dağılıma ait derinlik konturları arasında düzgün saçılacaklardır. Aksine örneklemin geldiği yığın düzgün dağılımlı değil ise gözlemler düzgün dağılıma ait derinlik konturları arasında düzgün saçılmayacaklardır. Bu durum şekil 3.1-3.4'deki şekillerde ifade bulmaktadır.

$MS_i = \left\{ X : Z^{[i-1]} \geq D_F\left(X\right) > Z^{[i]} \right\}, \quad i = 1,2,...,n+1$ ile verilen çok değişkenli fark istatistiklerinin tanımı kullanılarak bu ifadenin doğruluğu aşağıdaki gibi elde edilir.

Herhangi bir $F \in R^d$ dağılımından $X_1,...,X_n$ ve herhangi bir $G \in R^d$ dağılımında $Y_1,...,Y_n$ örneklemleri alınsın. F ve G dağılımlarından gelen örneklemlerin i. sıradaki gözlemlerinin derinlik değerleri sırasıyla $Z^{[i]} = D_F\left(X_{[i]}\right), \quad Z^{[i]*} = D_G\left(X_{[i]}\right)$ ve $T_{[i]} = P_F\left(X : D_F\left(X\right) > Z^{[i]}\right)$ ile gösterilsin ve her iki örneklem için derinlik halkalarının (multivariate spacings) kapsama olasılıkları

$$P_F\left(MS_i\right) = P\left\{ X : Z^{[i-1]} \geq D_F\left(X\right) > Z^{[i]} \right\}, \quad i = 1,2,...,n+1$$

ve

$$P_G\left(MS_i^*\right) = P\left\{ X : Z^{[i-1]*} \geq D_G\left(Y\right) > Z^{[i]*} \right\}, \quad i = 1,2,...,n+1$$

olarak tanımlansın. Bu durumda

$$P_F\left(MS_i^*\right) = P_F\left\{ X : D_F\left(X\right) > Z^{[i]*} \right\} - P_F\left\{ X : D_F\left(X\right) \geq Z^{[i-1]*} \right\} \tag{3.4}$$

$$P_G\left(MS_i\right)=P_G\left\{Y:D_G\left(Y\right)>Z^{[i]}\right\}-P_G\left\{Y:D_G\left(Y\right)\geq Z^{[i-1]}\right\} \tag{3.5}$$

şeklinde yazılabilir. Eğer $X\overset{d}{=}Y$ ise $Z^{[i]}=Z^{[i]_*}$ olduğundan (3.4) eşitliği için,

$$T^*_{[i]_F}=P_F\left(X:D_F\left(X\right)>Z^{[i]_*}\right)=P_G\left(Y:D_G\left(Y\right)>Z^{[i]_*}\right)=T^*_{[i]_G}$$
$$T^*_{[i-1]_F}=P_F\left(X:D_F\left(X\right)>Z^{[i-1]_*}\right)=P_G\left(Y:D_G\left(Y\right)>Z^{[i-1]_*}\right)=T^*_{[i-1]_G}$$

dir. Bu durumda,

$$\begin{aligned}P_F\left(MS_i^*\right)&=T^*_{[i]_F}-T^*_{[i-1]_F}\\P_G\left(MS_i^*\right)&=T^*_{[i]_G}-T^*_{[i-1]_G}\end{aligned}\quad\Rightarrow P_F\left(MS_i^*\right)=P_G\left(MS_i^*\right)\text{ dir.} \tag{3.6}$$

Aynı şekilde (3.5) eşitliğinden de

$$P_F\left(MS_i\right)=P_G\left(MS_i\right) \tag{3.7}$$

olduğu gösterilir. Eğer $F\neq G$ ise i.sıradaki gözlemin derinlik değeri $Z^{[i]}\neq Z^{[i]_*}$ olduğundan (3.6) ve (3.7) eşitlikleri sağlanmayacaktır.

Düzgün dağılım varsayımı altında gözlemlerin derinlik değerlerine karşılık olasılıkların konumlanması ve iki boyutlu bir veride kümelenmenin olduğu bölgelerdeki gözlemlerin düzgün olup olmadığının sorgulanması için bazı adımlar izlenmiştir.

İlk olarak, iki boyutlu rasgele bir örneklemdeki gözlemlerin, düzgün dağılım varsayımı altında, düzgün dağılımın yığınına ait yarı uzay derinlikleri elde edilmiştir.

İki boyutlu rasgele gözlem $\left(0,1\right)\times\left(0,1\right)$ birim karede düzgün dağılımlı ise $\forall\left(x,y\right)$ gözlem çifti için yarı uzay derinlikleri (2.13) ile verilen $D(x,y)=2\,min(x,1-x)\,min(y,1-y)\,,\left(x,y\right)\in Q$ derinlik fonksiyonunun aldığı değerler olacaktır.

Daha sonra (2.13)'deki tanımdan hareketle rasgele gözleme ilişkin rasgele $D(X,Y)$ derinliğinin dağılım fonksiyonu elde edilmiştir.

$X \sim U(0,1)$ ve $Y \sim U(0,1)$

ve

$$D(X,Y) = 2\,min(X,1-X)\,min(Y,1-Y)$$

olmak üzere, bu dağılım fonksiyonu,

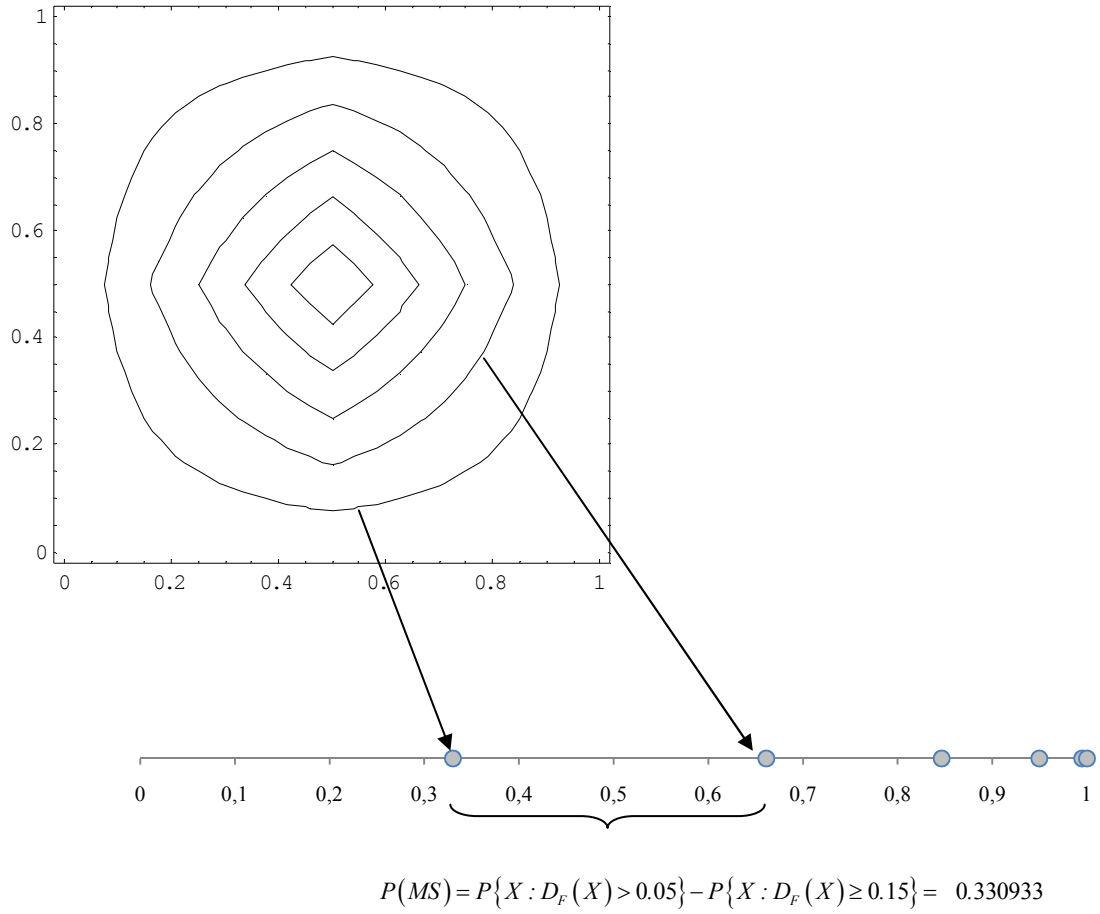
$$P(D \le d) = P(D(X,Y) \le d)$$
$$= P(min(X,1-X)\,min(Y,1-Y) \le d/2)$$

eşitliğinin çözümü ile elde edilir(**Ek 4**).

$D(X,Y)$ fonksiyonun sırasıyla dağılım fonksiyonu ve olasılık yoğunluk fonksiyonu aşağıdadır:

$$F_D(d) = P(D \le d) = \begin{cases} 0 & d < 0 \\ 2d\left(1 - ln(2d)\right) & 0 \le d \le 0.5 \\ 1 & d > 0.5 \end{cases} \quad \text{ve} \qquad (3.8)$$

$f_D(d) = -2\,ln(2d), \;\; 0 \le d \le 0.5$ olarak bulunur.



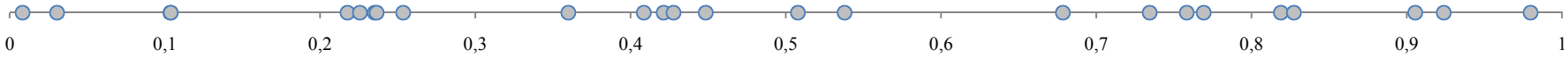
Şekil 3.5 $\alpha_1 = 0.05$ ve $\alpha_2 = 0.15$ derinlikli derinlik konturları arasında kalan bölgenin olasılığı

Düzgün dağılımın yığın derinlik fonksiyonunun dağılımı (3.8) ile verilmiştir.

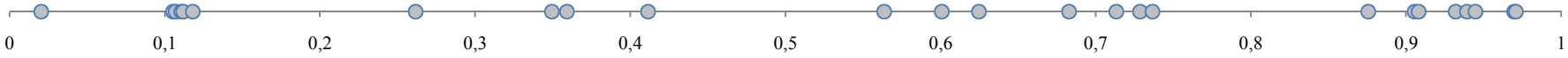
Bir sonraki aşamada ise seçilen sabit uzunluklu bir tarama aralığı (penceresi) için, n çaplı örneklemden bu aralığa düşen maksimum gözlem sayısı belirlenir. Bunun birim karede düzgün dağılım içinde olası bir kümelenmeye işaret edip edemeyeceği $P(k;n,w)$ olasılığı ya tam olarak ya da yaklaşık olarak hesaplanan değerine bakılarak karar verilir.

Bu adımlar sonucu gözlenen durum, iki boyutlu normal ve iki boyutlu düzgün dağılımdan seçilen ve örnek çapları 25 olan rasgele örneklemler için bir örnek üzerinde gösterilmiştir. İki boyutlu normal dağılım için $(x,y) \in R^2$ olduğundan ilk olarak normal

dağılımlı veri konumları değişmeyecek şekilde doğrusal bir dönüşüm ile $(0,1)\mathrm{x}(0,1)$ bölgesine ötelenmiştir. Derinlik fonksiyonu afin değişmez olduğundan dönüştürülmüş veriler ile gerçek verinin derinlik değerleri aynı kalacaktır. İki dağılımdan seçilen örneklemlerin yarı uzay derinlikleri (2.13)'e göre hesaplanmış ve bu derinlik değerlerine karşılık gelen $P\left(D\left(X,Y\right)\le D\left(x,y\right)\right)$ olasılıklarının $(0,1)$ aralığı üzerindeki konumları şekil 3.6-3.7'de gösterilmiştir. Eğer gözlemler iki boyutlu düzgün dağılımdan geliyorsa ilgili olasılıkların dağılımı yine düzgün dağılacaklardır.



Şekil 3.6 İki boyutlu düzgün dağılımlı 25 gözlemin düzgün dağılımın yığın derinlik fonksiyonundan elde edilen derinlik değerlerine karşılık gelen olasılıklar



Şekil 3.7 İki boyutlu normal dağılımlı 25 gözlemin düzgün dağılımın yığın derinlik fonksiyonundan elde edilen derinlik değerlerine karşılık gelen olasılıklar

Önemli not: Tek değişkenli rasgele verilerde n örneklem çapı ile ulaşılan testin gücüne çok boyutlu rasgele vektör durumunda aynı n örneklem çapı ile ulaşmak mümkün olmayabilecektir. Rasgele vektörün boyutunun artması ile seçilen örneklem çapının da bu boyuta bağlı olarak üstel olarak artması gerekmektedir (Wasserman 1959). Dolayısıyla iki boyutlu bir veri için seçilen $n = 25$ birimlik örnek çapının da yığını iyi yansıtmayacağı düşünülmektedir. Ancak tarama istatistiklerinde $P(k;n,w)$ olasılıklarının hesaplamalarında çoğu hesaplama yönteminde k için kısıtlamalar olduğu gibi örneklem çapı büyüdükçe de hesaplama zorlukları olmaktadır. Şekil 3.6'da *kümelerin konumlarının* düzgün olmasının şekil 3.7'ye göre daha belirgin olduğu söylenebilir. Örneklem çapı büyük olsaydı ve/veya simpleks derinlik kavramı kullanılsaydı derinlik değerlerinin çeşitlenmesi beklenebilirdi; eş derinlik değerlerine sahip gözlem sayısı az olması beklenirdi. Böylece yukarıdaki şekillerde gözlemlerin olasılıklarının kümelenmesinin değil de derinlik değerlerine karşılık gelen olasılıkların düzgün dağılımlı olduğu daha da net görülebilecekti.

4. YÖNTEMİN DEĞERLENDİRİLMESİ VE SONUÇ

Yöntemin değerlendirilmesine ışık tutmak amacıyla küçük çaplı bir simülasyon tasarlanmıştır. $\mu = (0 \;\; 0)'$ ve $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ olan standart normal dağılımdan, $\mu = (1 \;\; 1)'$ ve $\Sigma = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$ normal dağılımdan ve birbirinden bağımsız $U_1 \sim U(0,1)$ ve $U_2 \sim U(0,1)$ rasgele değişkenlerinin oluşturduğu iki boyutlu düzgün dağılımdan örneklem çapları 25 olan 100'er tekrar yapılmıştır. Taramalarda $w = 0.10$ uzunluklu tarama pencereleri kullanılmıştır.

Her örneklem için yöntemde sunulan adımlar uygulanmıştır. Buna göre iki boyutlu standart normal, normal ve düzgün dağılımlı örneklemlerin düzgün dağılım içindeki derinliklerine karşılık gelen olasılıklar $(0,1)$ aralığında işaretlenmiştir. Normal dağılım ve düzgün dağılımlı örneklemler için düzgün dağılımın derinlik konturlarına karşılık gelen olasılıklar (gözlemlerin $(0,1)$ aralığı üzerinde konumlandığı yerler) ile oluşturulan sabit tarama penceresine düşen maksimum gözlem sayıları karşılaştırılmıştır. Herhangi bir F dağılımından gelen rasgele bir örneklemin düzgün dağılımın yığınına ait yarı uzay derinlik karşılık gelen olasılıklarının $(0,1)$ aralığında konumlanması ile örnekleme ait kümelenmelerin olağan olup olmadığı (düzgün dağılıp dağılmadığı) araştırılmıştır. Sabit $w = 0.10$ uzunluğunda seçilen tarama penceresi içinde gözlenen en çok gözlem sayısı (k) simülasyona konu olan her dağılımdan için sayım yoluyla saptanmıştır. Çizelge 4.1'de her k değerinin her dağılımda 100 tekrarda kaç defa gözlendiği gösterilmektedir.

Çizelge 4.1 100 örneklem içinden k gözlem içeren örneklemlerin sayısı.

Tarama penceresi $w = 0.10$	Standart normal dağılım	Normal Dağılım $\rho = 0.7$	Düzgün Dağılım
$k = 4$	0	1	1
$k = 5$	12	3	29
$k = 6$	26	12	42
$k = 7$	21	28	21
$k = 8$	25	20	7
$k = 9$	8	14	0
$k = 10$	8	9	0
$k = 11$	0	6	0
$k = 12$	0	5	0
$k = 13$	0	1	0
$k = 16$	0	1	0

Çizelge 4.2 $k \geq 7$ durumunda simülasyon sonuçlarına göre H_0 hipotezinin reddedilme olasılıkları

	Simülasyonun yapıldığı dağılım (Gerçek dağılım)		
Tarama penceresi $w = 0.10$	**Standart normal dağılım**	**Normal Dağılım** $\rho = 0.7$	**Düzgün Dağılım**
H_0:Kümelenmeler düzgün dağılıma göre anlamlı değildir(Yığının dağılımı düzgün dağılımlıdır).	**0.62**	**0.84**	**0.28**

Çizelge 4.1 - 4.2'de verilen simülasyon sonuçları ile şu sonucu çıkartmak olanaklıdır:

Yığın gerçekte yukarıdaki iki değişkenli normal dağılımlı olduğu durumda, bu gözlemlerin birim karede iki değişkenli düzgün dağılımdan gelmediğine, $n = 25$ çaplı

örneklemde $k \geq 7$ olan kümelenmeleri gözleyerek karar verildiğinde, simülasyon çerçevesinde testin gücünün yaklaşık 0.84 olarak gözlemlendiği ifade edilebilir. Gerçekte gözlemler, birim karede düzgün dağılımlı yığından geliyorsa ve yine $k \geq 7$ olan kümelenmelere bakarak karar verip düzgün dağılımlı olmadığı çıkarımı yapılırsa gerçekleşen α değerinin yaklaşık 0.28 olduğu söylenebilecektir. $k \geq 8$ durumlarında ise yukarıda da değerlendirildiği gibi karar sonuçları daha anlamlı olacaktır. Örneklem çapı büyütülüp tekrar sayısı arttırılır ve simpleks derinliği kullanılırsa uygulanan yöntemin daha da anlamlandırılacağı düşünülmektedir.

Tek boyutlu tarama istatistiklerine ilişkin olasılık hesaplamasında Huffer ve Lin (1997) yöntemi ile seçilen tarama penceresine k veya daha fazla gözlem düşmesi olasılığı $P(k; N, w)$ hesaplanmıştır. Bu olasılık http://stat.fsu.edu/~huffer/software/moment adresinde yer alan c++ programında kodlanmış *general_type.c* isimli program kullanılarak elde edilmiştir.

İki boyutlu verilerde tarama penceresinin seçiminde karşılaşılan şekil probleminin, düzgün dağılım varsayımı altında derinlik fonksiyonunun dağılımına dayalı olarak tek boyutluda olduğu gibi aralıklar olarak elde edilerek bir ölçüde çözülebileceği düşünülmektedir. İki boyutlu örnek çapı $n = 25$ olan bir veri için $w = 0.10$ uzunluğundaki tarama penceresine 8 ve daha fazla gözlem düşmesi olasılığı $P(8; 25, 0.10) = 0.101$, $\alpha = 0.10$ seçilmiş ise 8 ve daha fazla gözlem düşmesi olasılığı olağan kabul edilecektir. 11 ve daha fazla gözlem düşmesi olasılığı $P(11; 25, 0.10) = 0.00088$ olarak bulunmuştur. 11 ve daha fazla gözlemin bu aralıkta yer alması olasılığı α=0.0009 için reddedilecektir. Dolayısıyla $(0,1)$ aralığında saçılan 25 gözlemden $w = 0.10$ uzunluğundaki tarama penceresine 11 veya daha fazla gözlem düşmesi durumunun olağandışı olduğu söylenebilecektir.

Çalışmanın amacında da belirtildiği üzere çok boyutlu veriler düzgün dağılımın derinlik dağılım fonksiyonuna göre değerlendirilmiştir. Dolayısıyla veride kümelenme veya sıra dışı bir dizilim varsa, veri düzgün dağılımdan gelmiyorsa, veri sözkonusu düzgün

dağılımın yığınının geometrik yapısına derinlik konturları bakımından uymayacaktır. Herhangi bir dağılımdan gelen bir örneklem düzgün dağılımın yığını içinde değerlendirildiğinde, gözlemlerin düzgün dağılımın yığın derinliklerine göre elde edilen ardışık derinlik konturları arasındaki bölgelerin olasılıkları eşit olmayacaktır. Çizelge 4.1'de görüldüğü üzere normal dağılımlı yığından gelen gözlemler düzgün dağılım içindeki olasılıkları bakımından kümelenme göstermektedirler. Sabit bir aralığa düşen gözlem sayıları, düzgün dağılımlı yığından gelen örnekleme göre bu yöntem farklılığı ortaya koyabilmektedir. Dolayısıyla bu yaklaşım ile herhangi bir dağılımdan gelen çok boyutlu gözlemlerin kümelenmesindeki olağan dışılığı, gözlemlerin çekildiği çok boyutlu dağılımın geometrik yapısını da dikkate alarak tek boyutlu tarama istatistikleri ile değerlendirme imkânı sağlayacaktır.

KAYNAKLAR

Alm, E. A. 1999. Approximations of the distributions of scan statistics of poisson processes. scan statistics: method and applications(statistics for industry and technology), s.113-139, Birkhäuser.

Anderson, N.H. and Titterington, D.M. 1997. Some methods for investigating spatial clustering, with epidemiological applications. Journal Royal Statistical Society, Series A Vol.160, 87-105.

Berman, M. and Eagleson, G.K. 1985. A useful upper bound for the tail probabilities of the scan statistics when the sample size is large. Journal of the American Statistical Associations, Vol.84, 560-566.

Cameron, N. 1985. Introduction to linear and convex programming. Australian Mathematics Society. Lecture Series 1, Cambridge University Press.

Chen, L. A. and Welsh, A. H. 2002. Distribution-function based bivariate quantiles. Journal of Multivariate Analysis, Vol.83, No.1, 208-231.

Cressie, N. 1991. Statistics for spatial data. Wiley, New York.

Einmahl, J.H. J. and Mason, D. M. 1992. Generalized quantile processes. The Annals of Statistics, Vol.20, No.2, 1062-1078.

Frosini, B.V. 1981. Distribution of the smallest interval that contains a given cluster of points. Statistica, Vol.41, 255-280

Glaz, J. and Naus, J. 1983. Multiple clusters on the line. Communications in Statistics: Theory and Methods, Vol.12(17), 1961-1986.

Glaz, J. and Balakrishnan, N. 1999. Scan statistics: Method and applications (Statistics for Industry and Technology). Birkhäuser.

Glaz, J., Naus, J. and Wallenstein, S. 2001. Scan statistics. Springer.

Glaz, J., Pozdnyakov, V. and Wallenstein, S. 2009. Scan statistics: Method and applications (Statistics for Industry and Technology). Birkhäuser.

Hjalmars, U., Kulldorf, M., Gustafsson, G. and Nagarwalla, N. 1996. Childhood leukemia in sweden using GIS and a spatial scan statistics for cluster detection. Statistics in Medicine, Vol.15, 707-715.

Huffer, W.F. 1988. Divided difference and the joint distribution of linear combinations of spacings. Journal of Applied Probability-Applied Probability Trust, Vol.25, 346-354,

Huffer, W.F. and Lin, C.T.1997. Computing the exact distirbution of the extremes of sums of consecutive spacings. Computational Statistics&Data Analysis(Elsevier), Vol.26, 117-132.

Huffer, W.F. and Lin, C.T. 2001. Computing the joint distribution of general linear combinations of spacings or exponential variates. Statistica Sinica, Vol.11, 1141-1157.

Hugg, J., Rafalin, E., Seyboth, K. and Souvaine, D. 2006. An experimental study of old and new depth measures. Eighth Workshop on Algorithm Engineering and Experiments (ALENEX)

Huntington, R. and Naus, J. 1975. A simpler expression for kth nearest neigbor coincidence probabilities. Annals of Probability, Vol.3, 894-896.

Lehmann, E.L. 1998. Nonparametrics: Statistical methods based on ranks. New Jersey: Prentice Hall.

Li, J. and Liu, R.Y. 2008. Multivariate spacings based on data depth: I. Construction of nonparametric multivariate tolerance regions. Annals of Statistics, Vol.36, No.3, 1299-1323.

Lin, C.T. 1993. The computation of probabilities which involve spacings, with applications to the scan statistics. Ph.D. Thesis, The Florida State University, Tallahassee, FL.

Liu, R.Y., Parelius, J. M. and Singh, K. 1999. Multivariate analysis by data depth: Descriptive statistics, graphics and inference(with discussions). The Annals of Statistics, Vol.27, No.3, 783-858.

Liu, R.Y and Singh, K. 1993. A quality index based on data depth and multivariate rank tests. Journal of the American Statistical Association, Vol.88, No. 421, Theory and Methods, 252-260.

Loader, C. 1990. Large deviation approximations to distribution of scan statistics. AT&T Bell Laboratorie Technical Memorandum 11214-900912-12TM.

Loader, C. 1991. Large deviation approximations to distribution of scan statistics. Advanced Applied Probability, Vol.23, 751-771.

Månsson, M. 1999. On poisson approximation for continuous multiple scan statistics in two dimensions. Scan Statistics: Method and Applications(Statistics for Industry and Technology), s.225-247, Birkhäuser.

Naus, J.I. 1965a. The distribution of the size of the maximum cluster of points on a line. Journal of the American Statistical Association, Vol.60, No. 310, 532-538.

Naus, J. I. 1965b. A power comparison of two tests of non-random clustering. Technometrics, Vol.8, 493-517.

Naus, J. I. 1974. Probabilities for a generalized birthday problem. Journal of the American Statistical Association, Vol.69, 810-815.

Parzen, E. 1960. Modern probability theory and its applications. Wiley, New York.

Port, Sidney C. 1994. “Theoretical Probability for Applications”. John Wiley&Sons, Inc.

Pyke, R. 1965. Spacings. Journal of the Royal Society. Series B(Methodological), Vol.27, No.3, 395-449.

Rousseeuw, P. J. and Hubert, M. 1999. Regression Depth. Journal of the Statistical Association, Vol.94, No.446, Theory and Methods, 388-402.

Rousseeuw, P. J. and Ruts, I. 1999. The depth function of a population distribution. Metrika, Vol.49, 213-244.

Serfling, R. J. 2002. Generalized quantile processes based on multivariate depth functions, with applications in nonparametric multivariate analysis. Journal of Multivariate Analysis, Vol.83, No.1, 232-247.

Serfling, R. J. 2004. Nonparametric multivariate descriptive measures based on spatial quantiles. Journal of Statistical Planning and Inference, Vol.123, 259-278

Serfling, R. J. 2006. Depth functions in nonparametric multivariate inference. In Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications (R. Y. Liu, R. Serfling, D. L. Souvaine, eds.). DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, Volume 72, 1-16.

Seth, G.R. 1950. On the distribution of the two closest among a set of three observations. The Annals of Mathematical Statistics. Vol. 21, No.2, 298-301.

Wallenstein, S. R. and Naus, J. I. 1973. Probabilities for a kth nearest neighbor problem on the line. The Annals of Probability, Vol.1, No.1, 188-190.

Wallenstein, S. R. and Naus, J. I. 1974. Probabilities for the size of the largest clusters and smallest intervals. Journal of the American Statistical Association, Vol.69, 690-697.

Wang, J. and Serfling, R. J. 2005. Nonparametric multivariate kurtosis and tailweight measures. Journal of Nonparametric Statistics, Vol.17(4), 441-456.

Wasserman, L. 1959. All of Statistics. Springer.

Zuo, Y. and Serfling, R. 2000a. General notions of statistical depth functions. The Annals of Statistics, Vol.28, No.2, 461-482.

Zuo, Y. and Serfling, R. 2000b. Structural properties and convergence results for contours of sample statistical depth functions. The Annals of Statistics, Vol.28, No.2, 483-499.

Huffer, W.F. and Lin, C.T.1997 Program web sayfası:

http://stat.fsu.edu/~huffer/software/moment

EKLER

EK 1 Tek Köşeli Konveks Bölge İçin Derinlik Fonksiyonunun Hesaplanması

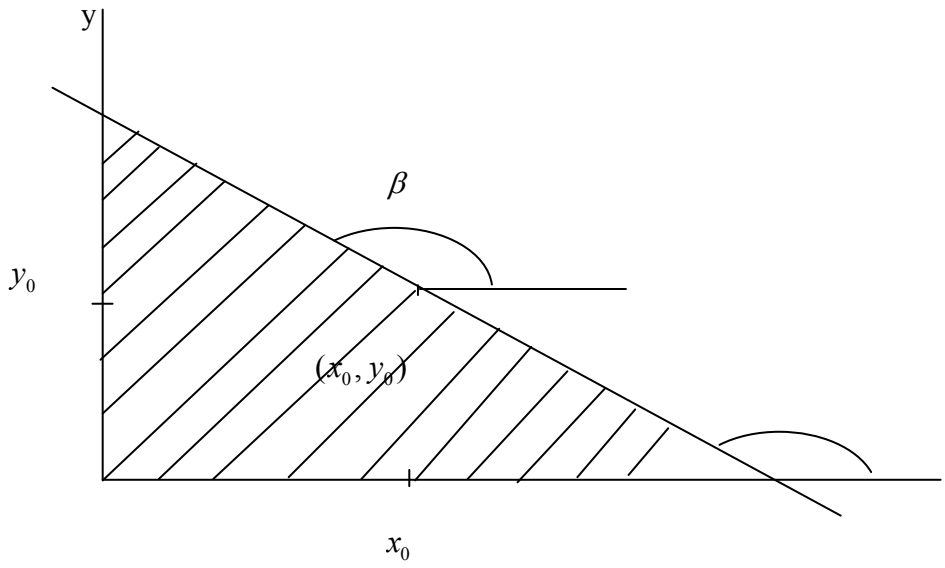
EK 2 Huffer ve Lin'in Yaklaşımı

EK 3 Teorem 3.1 İspatı(Li and Liu 2008, Theorem 2.1)

EK 4 Düzgün Dağılıma ait Yığın Derinlik Fonksiyonunun Olasılık Dağılımının Bulunması

EK 1 Tek Köşeli Konveks Bölge İçin Derinlik Fonksiyonunun Hesaplanması

Q , aynı noktadan çıkan iki ışın ile oluşturulan konveks bölge olsun. Q 'nun bir köşesi ve bir iç açısı vardır. Bölge konveks olduğundan iç açı π 'den küçük olmalıdır. Afin değişmezlik özelliği kullanılarak, Köşe noktası (0,0) ve ışınlarda pozitif x ekseni ve pozitif y olarak kabul edilebilir. Bu durumda $\mu = U_Q$, $Q = \{(x,y); x \geq 0 \text{ ve } y \geq 0\}$ olacaktır. $\lambda(Q) = \infty$ olduğundan μ , $\mu(Q) = \infty$ olan pozitif bir ölçüdür.



Şekil. Taralı üçgenin minimum alanı (x_0, y_0) noktasının derinliği

R^2 'de $(x_0, y_0) \in Q$ olan bir nokta alınsın ve bu noktanın derinliği hesaplanmak istensin. Bunun için (x_0, y_0) üzerinden bir L_β doğrusu geçsin. Bu doğru ile pozitif x ekseninin yapacağı açı $\beta > \pi/2$ olacaktır. Daha sonra sınırı L_β olan bir H_β yarı uzayı (halfspace) alınsın. L_β doğrusu, Q ile sınırlı bir üçgen oluşturmaktadır. α derinliği, $\mu(H_\beta) = area(H_\beta \cap Q)$ alanının minimum değeri olacaktır.

β açısı tanjant ile aşağıdaki gibi iki şekilde ifade edilebilir.

$\tan(180^\circ - \beta) = -\tan\beta = \dfrac{y - y_0}{x_0}$ (1) ve $-\tan\beta = \dfrac{y}{x}$ (2)

(1) ifadesinden $y = -\left(x_0 \tan\beta - y_0\right)$ olarak elde edilir.

Q tanım kümesi ile H_β yarı uzayı arasında oluşan bölgenin alanı

$Area(H_\beta \cap Q) = \dfrac{yx}{2}$ ile ifade edilir.

$$Area(H_\beta \cap Q) = \frac{yx}{2} = \frac{-\left(x_0 \tan\beta - y_0\right)x}{2}\frac{\mathbf{y}}{\mathbf{y}}$$

(2) ifadesi kullanılarak $\dfrac{x}{y}$ yerine $-\dfrac{1}{\tan\beta}$ yazılırsa,

$$Area(H_\beta \cap Q) = -\left(x_0 \tan\beta - y_0\right)^2 / 2\tan\beta$$

bulunur. Bu alanı minimize edecek β açısı $\beta = \text{Arctan}(-y_0/x_0)$ olarak bulunur. Buna göre minimum alan $2x_0y_0$ elde edilir.

$\forall\ (x,y) \in Q$ noktası için derinlik fonksiyonu $\text{depth}(x,y) = 2xy$ ile ifade edilir. Böylece herhangi bir $\alpha > 0$ için derinlik bölgesi,

$$D_\alpha = \left\{(x,y) \in Q; y \geq \frac{\alpha}{2x}\right\} \text{ dir.}$$

$\alpha = 0$ için $D_\alpha = Q$ olur.

EK 2 Huffer ve Lin'in Yaklaşımı

Huffer ve Lin(1997), $(0,1)$ aralığında birbirinden bağımsız ve düzgün dağılımlı gözlem arasındaki fark istatistiklerine dayanarak tarama istatistiklerini ardışık fark istatistiklerinin toplamı olarak ifade etmişlerdir. Diğer bir ifade ile r tarama istatistiği, r tane ardışık fark istatistiklerinin toplamlarının maksimumu olarak da ifade edilebilir. $X_1, X_2, ..., X_n$, birbirinden bağımsız ve düzgün dağılımlı rasgele değişkenler olsun. Bu örnekleme ilişkin sıra istatistikleri $X_{[1]} \le X_{[2]} \le ... \le X_{[n]}$ ile gösterilsin. $(i-1)$. ve i. noktalar arasındaki uzunluk $D_i = X_{[i]} - X_{[i-1]}$ fark istatistikleri(spacings) olarak adlandırılır. Fark istatistikleri vektörü $\mathbf{D}=(D_1, D_2, ..., D_{n+1})'$ ile gösterilsin. W_k tarama istatistiği, ardışık k tane fark istatistiğinin toplamlarının minimumu olarak da aşağıdaki gibi tanımlanabilir:

$$W_k = \min_{1 \le i \le N-k+1} \left\{ X_{[i+k-1]} - X_{[i]} \right\} = \min_{1 \le i \le N-k+1} \left\{ \sum_{j=1}^{k-1} \left(X_{[j+1]} - X_{[j+i-1]} \right) \right\}$$

Huffer(1988)'ın düzgün dağılıma sahip N nokta için, fark istatistiklerinin doğrusal birleşimlerinin ortak dağılımları, kesin olasılık hesapları ve düzgün dağılıma uyum testleri ile ilgili çalışmaları bulunmaktadır. Lin(1993) ve Huffer ve Lin(1997), fark istatistiklerinin toplamlarının minimumlarının(ya da maksimumlarının) dağılımını bulmak için genel bir yaklaşım geliştirmişlerdir. Bu sonuç, tarama istatistiklerinin dağılımı için parçalı polinom gösterimi ile verilir. Bu yöntem matris gösterimi ile uygulanır. Fark istatistiklerinin lineer birleşimlerinin ortak dağılımı, katsayıların matrisi şeklinde ifade edilir ve bu matris yineleme adım adım daha kolay hesaplanacak alt matrislere parçalanır.

Bu yaklaşım aşağıda ifade edilen yineleme(recursion) işlemine dayanmaktadır.

Γ, $r \times (n+1)$ tipinde bir matris ve $\mathbf{D}=(D_1, D_2, ..., D_{n+1})'$, düzgün dağılıma sahip rasgele değişkenler arasındaki fark istatistiklerinin vektörü olsun. Bu yaklaşımın amacı rasgele

$\Gamma\mathbf{D}$ vektörünün dağılımını elde etmektir. $P(\Gamma)$, $\Gamma\mathbf{D}$'nin olasılık ölçüsü olarak yazılsın. Dolayısıyla $P(\Gamma\mathbf{D} \in B) = (P(\Gamma))(B)$ dir. Bu olasılık daha basit bir gösterim olması açısında $P(\Gamma)$ ile de gösterilecektir.

Teorem 1(Huffer (1988), Theorem)

Herhangi bir $\xi \in R^r$ için Γ matrisinin $i.$sütunu ξ vektörü ile değiştirildiğinde elde edilen $r \times (n+1)$ boyutlu matris $\Gamma_{i,\xi}$, $c = (c_1, c_2, ..., c_{n+1})'$ bir vektör ve $\sum_{i=1}^{n+1} c_i = 1$ olsun. $\xi = \Gamma c$ olmak üzere,

$P(\Gamma) = \sum_{i=1}^{n+1} c_i P(\Gamma_{i,\xi})$ dir.

Bu yöntem ile indirgeme yaparak fark istatistiklerin ortak dağılımı, birbiri ile örtüşmeyen(nonoverlapping) fark istatistiklerinin dağılımları şeklinde ifade edilebilecektir. Fark istatistiklerinin ortak dağılımları Port(1994)(ss.510 Proposition 39.14) tarafından $w_i \geq 0$ sabitleri için ve $(y)_+ = max(y, 0)$ olmak üzere

$$P\left\{\bigcap_{i=1}^{n+1}(D_i > w_i)\right\} = \left(\left(1 - \sum_{i=1}^{n+1} w_i\right)_+\right)^n$$

olarak yazılmıştır.

Bir aralık üzerindeki gözlemlerin kümelenme olasılığının hesaplanması için bir örnek(Glaz vd. 2001, s.133)

$(0,1)$ aralığında düzgün dağılıma sahip X_1, X_2, X_3, X_4 rasgele değişkenleri verilsin. Bu değişkenlerin sıralanmış değerleri $X_{[1]} \leq X_{[2]} \leq X_{[3]} \leq X_{[4]}$ olsun. $P(3; 4, w)$ olasılığı bu yöntem ile aşağıdaki gibi hesaplanır.

$D_1 = X_{[1]}$, $D_2 = X_{[2]} - X_{[1]}$, $D_3 = X_{[3]} - X_{[2]}$,

$D_4 = X_{[4]} - X_{[3]}$, $D_5 = 1 - X_{[4]}$ olsun. Örneğin D_2 ile D_4, D_3 ile D_5 örtüşmeyen (nonoverlapping), D_2 ile D_3 çakışık fark istatistikleridir.

$$1 - P(3;4,w) = P\left\{\left(X_{[3]} - X_{[1]} > w\right) \cap \left(X_{[4]} - X_{[2]} > w\right)\right\}$$

$$= P\left\{\left(D_2 + D_3 > w\right) \cap \left(D_3 + D_4 > w\right)\right\}.$$

Lin(1993)'de son eşitlik için matris gösterimi yazılmıştır. $\mathbf{D}=\left(D_1, D_2, ..., D_{n+1}\right)'$, 5×1 boyutunda bir vektör, w, 2×1 boyutlu $(w,w)'$ vektörü ve $\Gamma_{2\times 5} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$ olsun. $1 - P(3;4,w)$ olasılığı $P(\Gamma \mathbf{D} > w)$ matris gösterimi ile hesaplanacaktır. Bu örnek için Teorem 1 uygulansın. $c = (0,1,-1,1,0)$ olarak seçilirse $\xi = \Gamma c = (0,0)'$ ve

$$P(\Gamma \mathbf{D} > w) = 0P\left(\Gamma_{1,\xi}\mathbf{D} > w\right) + 1P\left(\Gamma_{2,\xi}\mathbf{D} > w\right) - 1P\left(\Gamma_{3,\xi}\mathbf{D} > w\right) + 1P\left(\Gamma_{4,\xi}\mathbf{D} > w\right) + 0P\left(\Gamma_{5,\xi}\mathbf{D} > w\right)$$

elde edilir.

$$\Gamma_{2,\xi} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \qquad \Gamma_{3,\xi} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \qquad \Gamma_{4,\xi} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

Fark istatistikleri negatif olmayan değerler aldıkları için

$P\left(\Gamma_{2,\xi}\mathbf{D} > w\right) = P\left\{\left(D_3 > w\right) \cap \left(D_3 + D_4 > w\right)\right\} = P\left(D_3 > w\right)$,

$P\left(\Gamma_{3,\xi}\mathbf{D} > w\right) = P\left\{\left(D_2 > w\right) \cap \left(D_4 > w\right)\right\}$ ve

$P\left(\Gamma_{4,\xi}\mathbf{D} > w\right) = P\left(D_3 > w\right)$ dir.

Bu şekilde tüm terimler, örtüşmeyen fark istatistiklerinin ortak dağılımları veya ayrık fark istatistiklerinin dağılımları olarak basit halde yazılabilecektir. Bu olasılıklar düzgün fark istatistikleri için aşağıdaki yöntem ile kolayca hesaplanacaktır (Port(1994),Lin(1993), (2.3)).

Böylece,

$P(D_3 > w) = (1-w)^4$ ve $P\{(D_2 > w) \cap (D_4 > w)\} = \left((1-2w)_+\right)^4$ dır. $(0,1)$ aralığında seçilen herhangi bir w uzunluğundaki tarama penceresine 3 ve daha fazla gözlem düşmesi olasılığı,

$1 - P(3;4,w) = 2(1-w)^4 - \left((1-2w)_+\right)^4$ olarak hesaplanır.

Huffer(1988), Γ matrisinin terimlerinin basitleştirilmesi ve hesaplamanın kolaylaşması için tanımı aşağıdaki özellikleri kullanmıştır.

Eğer A matrisinin $i.$ satırın bir elemanı $j.$ satırın aynı elemanına eşit veya büyük ise $i.$ satır silinir.

i. Sütunlar yer değiştirebilir,
ii. Satırlar yer değiştirebilir
iii. Tüm elemanları 0 olan sütun silinir.

Büyük boyutlu matrisler için matrislerin alt matrislere indirgenmesi oldukça karmaşık ve zaman alıcı bir işlem gerektirmektedir. Matris indirme işlemi bu özelliklere dayalı olarak Huffer ve Lin(1997)'de bir algoritma olarak verilmiştir ve buna bağlı olarak $P(\Gamma\mathbf{D} > w)$ olasılığı hesaplanması c++ programı ile elde edilmiştir. Söz konusu program kodları http://stat.fsu.edu/~huffer/software/moment adresinden indirilebilir durumdadır. Bu olasılık *general_type.c* isimli program kodu derlendikten(compile) sonra "Ms Dos" ortamında çalıştırılmıştır.

EK 3 Teorem 3.1 İspatı(Li and Liu 2008, Theorem 2.1)

$i.$ gözlemin derinlik değeri $Z_i = D_F\left(X_i\right)$ ve $T_i = P_F\left(X : D_F\left(X\right) > Z_i\right),\ i = 1,2,...,n$ olsun. T_i'ler $U\left(0,1\right)$ dağılımında alınmış rasgele bir örneklem olarak alınabilir. $T_{[1]}, T_{[2]}, ..., T_{[n]}$, T_i'lerin sıra istatistikleri olmak üzere;

$$T_{[i]} = P_F\left(X : D_F\left(X\right) > Z^{[i]}\right)$$

yazılabilir. Ardışık derinlikli bölgeler

$$MS_i = \left\{X : Z^{[i-1]} \geq D_F\left(X\right) > Z^{[i]}\right\},\ 1,2,...,n+1$$

şeklinde tanımlanmıştı. Bu durumda $i.$ gözlemin kapsama olasılığı

$$P_F\left(MS_i\right) = P_F\left(D_F\left(X\right) \leq Z^{[i-1]}\right) - P_F\left(D_F\left(X\right) \leq Z^{[i]}\right)$$

$$= 1 - P\left(D_F\left(X\right) \geq Z^{[i-1]}\right) - 1 + P\left(D_F\left(X\right) \geq Z^{[i]}\right)$$

$$= T_{[i]} - T_{[i-1]}$$

olarak bulunur.

EK 4 Düzgün Dağılıma ait Yığın Derinlik Fonksiyonunun Olasılık Dağılımının Bulunması:

Rasgele $D(X,Y)$ derinliğinin dağılım fonksiyonu $X \sim U(0,1)$ ve $Y \sim U(0,1)$, $D(X,Y) = 2\,min(X,1-X)\,min(Y,1-Y)$ olmak üzere,

$$P(D \le d) = P(D(X,Y) \le d)$$

$$= P(min(X,1-X)\,min(Y,1-Y) \le d/2)$$

ile gösterilmiştir.

$Z_1 = min(X,1-X) \sim U(0,1/2)$ ve $Z_2 = min(Y,1-Y) \sim U(0,1/2)$

$$P(Z_1 > z_1) = P(min(X,1-X) > z_1) = P(X > z_1, 1-X > z_1)$$

$$= P(X > z_1, X < 1-z_1) = P(z_1 < X < 1-z_1)$$

$= 1-2z_1$ dir. Dolayısıyla,

$$F_{Z_1}(z_1) = \begin{cases} 0, & z_1 < 0 \\ 2z_1, & 0 < z_1 < 1/2 \\ 1, & z_1 > 1/2 \end{cases} \text{ dır.}$$

$$P(D \le d) = \int_0^1 P(D \le d \,|\, X = x) 1 dx$$

$$= \int_0^1 P(min(x,1-x)\,min(Y,1-Y) \le d/2)\,dx$$

$$= \int_0^1 P(Z_1 Z_2 \le d/2)\,dx = 2\int_0^{1/2} P(Z_1 Z_2 \le d/2)$$

integrali hesaplandığında, düzgün dağılımın yığın derinlik fonksiyonunun sırasıyla dağılım fonksiyonu ve olasılık yoğunluk fonksiyonu

$$F_D(d) = P(D \le d) = \begin{cases} 0 & d < 0 \\ 2d\left(1 - ln(2d)\right) & 0 \le d \le 0.5 \\ 1 & d > 0.5 \end{cases} \quad \text{ve}$$

$f_D(d) = -2\,ln(2d), \;\; 0 \le d \le 0.5$ olarak bulunur.

ÖZGEÇMİŞ

Adı Soyadı : Elmas Burcu MAMAK EKİNCİ

Doğum Yeri : Ankara

Doğum Tarihi : 02/01/1980

Medeni Hali : Evli

Yabancı Dili : İngilizce

Eğitim Durumu (Kurum ve Yıl)

Lise : Özel Yüce Fen Lisesi (1996)

Lisans : Ankara Üniversitesi Fen Fakültesi İstatistik Bölümü (2001)

Yüksek Lisans : Ankara Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı (2005)

Çalıştığı Kurum/Kurumlar ve Yıl

Ufuk Üniversitesi-Fen Edebiyat Fakültesi-İstatistik Bölümü (2006-Devam)

Yayınları (SCI ve diğer)

Uluslar arası bilimsel toplantılarda sunulan ve bildiri kitabında (Proceedings) basılan bildiriler:

- Erkan A.F., Ekici B., Demir G., Çehreli S., Töre H.F., **Mamak E.B.**, Candan I., *"The Correlation Between Serum High Density Lipopretein Levels and Duke Treadmill Score"*, 4th Congress of Update in Cardiology and Cardiovascular Surgery, 2008.
- Erkan A.F., Ekici B., Demir G., Çehreli S., Töre H.F., **Mamak E.B.**, Candan I., *"Postprandial Hyperglycemia Predicts Left Ventricular Hypertrophy"*, 4th Congress of Update in Cardiology and Cardiovascular Surgery, 2008.

- Erkan A.F., Ekici B., Demir G., Çehreli S., Töre H.F., **Mamak E.B.**, Candan I., *"The Relationship of C-Reactive Protein Levels to QT Interval Dispersion"*, 4th Congress of Update in Cardiology and Cardiovascular Surgery, 2008.

Yazılan kitaplar veya kitaplarda bölümler:

- Harald, M., Papke, K., Althoff, S. ve Oberwittler, C., "Nörolojide Ölçme Değerlendirme". 1.Baskı, Bölüm 3 "Biyomedikal İstatistiğin Temelleri" (3. Bölümün İngilizce Türkçe Çevirisi), Palme Yayıncılık, 2008

Ulusal bilimsel toplantılarda sunulan ve bildiri kitabında basılan bildiriler:

- Karabulut I., Mamak Ekinci, E.B., "Derinliklere Dayalı Yüzdelik Konturlarının Bootstrap Yöntemi İle Oluşturulması", ODTÜ 7. İstatististik Günleri Sempozyumu, Ankara, Türkiye, Haziran 2010