İki Boyutlu Veriler İçin Görsel Etkili Bazı Betimsel İstatistikler

Cemal Atakan[1], İhsan Karabulut, Fikri Öztürk

Ankara Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Ankara

Özet: Bir boyutlu veriler için çubuk grafiği, histogram, frekans poligonu, kutu çiti gibi görsel etkili betimsel istatistikler ve sıra istatistikleri veri çözümlemesinde temel araçlardır ve yaygın olarak kullanılırlar. Verinin iki veya daha fazla boyutlu olması durumunda bunların kullanımı çok sınırlıdır. Bu çalışmada, histogram, frekans poligonu, kutu çiziti ve sıra istatistiklerinin iki boyutluya genişletilmesi üzerinde durulmuştur.

Some Descriptive Statistics With Visual Effects For Two-Dimensional Data

Abstract: For the one dimensional data, descriptive statistics with visual effects like bar graph, histogram, frequency poligon and boxplot, order statistics are tools in data analysis and used generally. The use of these tools are very limited in case of two or higher dimensional data. In the current study the extensions of histogram, frequency poligon, boxplot and order statistics to two dimensions are considered.

1. Giriş

Bir boyutlu dağılımlar için aklımıza ilk gelen betimsel istatistikler örneklem ortalaması, örneklem varyansı, örneklem ortancası, örneklem çeyreklikleri, sıra istatistikleri, kutu çiziti, çubuk grafiği, histogram, frekans poligonu, örneklem dağılım fonksiyonu gibi istatistiklerdir. İki boyutlu verilere gelince, akla ilk olarak örneklem ortalaması, örneklem varyans-kovaryans matrisi, örneklem korelasyon matrisi, çapraz tablo ve son zamanlarda çubuk grafiği ile histogram gelmektedir. Görsel bilgi sağlayan kutu çiziti, çubuk grafiği, histogram, frekans poligonu gibi betimsel istatistiker hemen hemen her bilgisayar istatistik paket pragramında yer almaktadır. Örneğin iki boyutlu veriler için çubuk grafiği ile histogram SPSS 13 de yer almaktadır. Ancak kutu çizitinin iki boyutlu verilerde karşılığı olan çanta çiziti ile iki boyutlu veriler için frekans poligonu paket programlarda pek görünmemektedir.

Bir boyutlu dağılımlarda birçok betimsel istatistik esasında sıra istatistiklerinin bir fonksiyonudur. Bir boyutlu dağılımlar reel sayıların Borel cebiri üzerinde olup, reel sayılardaki sıralama doğal olarak sıra istatistiklerini ortaya çıkarmaktadır. Çok boyutlu dağılımlarda, bir boyutlu dağılımlarda olduğu gibi sıra istatistiği tanımlamak mümkün olmamakla birlikte, çok boyutlu verilerin bulunduğu Öklit (Euclide) uzayındaki normlara dayalı bazı sıra istatistikleri

[1] E-mail: atakan@science.ankara.edu..tr

tanımlanabilmektedir. Son yıllarda çok boyutlu veriler için oluşturulan derinlik ve merkez kavramlarına dayalı olarak tanımlanan sıra istatistikleri, bir boyutlu sıra istatistiklerine benzer biçimde bazı işler görebilmektedir [4].

Çalışmanın ikinci kısmında kutu çizitinin iki değişkenli verilere genişletilmesi üzerinde durulmaktadır. Üçüncü kısımda bazı derinlik tanımlamaları hatırlatılacak ve veri analizinde uygulamaları gözden geçirilmektedir. Son kısımda iki boyutlu veriler için histogram örnekleri verilecektir.

2. Kutu çizitinin iki değişkenli verilere genişletilmesi

Tukey (1977) tarafından belirtildiği gibi kutu çizitleri, beş tane istatistik değeri ile verilerin görsel bir betimlemedir. Bu beş istatistik; ortanca (median) alt ve üst menteşeler (hinges) ile uç değerlerdir. Kutu çizitlerinin çok değişik biçimleri sözkonusudur. En yaygın olarak Box-and-whiskers kutu çiziti kullanılır. Bu çizitin kutusunun yan kenarları, birinci ve üçüncü çeyreklik ($Q_1 \text{ve } Q_3$) ve içindeki çizgi ortanca (Q_2) değerindedir. Kutunun sol taraftaki $Q_2 + 4(Q_1 - Q_2)$ değerine alt çit (lower fence) ve sağ taraftaki $Q_2 + 4(Q_3 - Q_2)$ değerine üst çit (upper fence) değeri denir. Kutu dışındaki yatay çizgiler (whiskers) kutudan başlayıp, çit değerleri arasında kalan en küçük (solda) ve en büyük (sağda) gözlem değerlerine kadar uzanır. Çit değerleri dışında kalan gözlemler sıradışı (outliers) olarak adlandırılır [5].

Verilerin genellikle normal dağılım ile modellendiği ve iki değişkenli normal dağılımlarda güven bölgelerinin elipsler olduğu göz önüne alınırsa, elipslerin bir boyuttaki kutuların yerine kullanılması doğal görünmektedir. Verilerin %50 'sini içeren ve içte olan bir elips menteşe (hinge) ve sıradışı değerleri ayırd eden dıştaki elips çit (fence) vazifesini görebilir. Bu elipslerin oluşturulması bir tarafa, eliptik veya simetrik olmayan dağılımlar için uygun olmayacakları ortadadır. Böyle dağılımlar için tek parçalı elips yerine, dört farklı elipsin parçalarından oluşan menteşe ve çit önerilmektedir. Bu kısımda, Goldberg ve Iglewicz (1992) tarafından sunulan, elips çiziti (robust elliptic plot, relplot) ile dört elips parçasından oluşan çizit (quarter elliptic plot, quelplot) özetlenmektedir. İki değişkenli dağılımlarda verilerin %50 sinin bulunduğu ve bir değişkenlide kutuya karşılık gelen bölgeye çanta denir. Böylece, kutu çizitinin karşılığı da çanta çiziti olmaktadır[6].

Normal dağılıma sahip olan iki değişkenli (X,Y) rasgele vektörünün dağılımı, değişkenlerin ortalamaları (μ_X , μ_Y), standart sapmaları (σ_X , σ_Y) ve aralarındaki korelasyon katsayısı (ρ) ile belirlenebilir. Bunların örneklem karşılıkları olan $\overline{X}, \overline{Y}, S_X, S_Y \text{ ve } R$ bu parametreler için alışılmış tahmin edicilerdir. Bunların yerine, örneğin uç değerlere karşı dirençli (robust) olan, başka tahmin ediciler de düşünülebilir. Bir boyutlu normal dağılımda, ortalama aynı zamanda konum (location) ve satandart sapma da ölçek (scale) parametresidir.

Goldberg ve Iglewicz (1992) elips çizitini (relplot) aşağıdaki gibi oluşturmaktadır.

Bir (X,Y) rasgele vektörünün marjinal dağılımlarının merkezi eğilim ve saçılım ölçüleri ile değişkenler arasındaki ilişki katsayısı için birer tahmin edici sırasıyla $T_X^*, T_Y^*, S_X^*, S_Y^*, R^*$ olsun. $i = 1,2,...n$ için (X_i, Y_i) gözlemleri,

$$X_{si} = \frac{X_i - T_X^*}{S_X^*} \quad , \quad Y_{si} = \frac{Y_i - T_Y^*}{S_Y^*}$$

olarak standartlaştırılmakta ve gözlemlerin (T_X^*, T_Y^*) noktasından,

$$E_i = \sqrt{\frac{X_{si}^2 + Y_{si}^2 - 2R^* X_{si} Y_{si}}{1 - R^{*2}}} \quad , \quad i = 1,2,...n$$

uzaklıkları hesaplanmaktadır. $T_X^*, T_Y^*, S_X^*, S_Y^*, R^*$ ler sırasıyla örneklem ortalamaları, standart sapmalar ve Pearson korelasyon katsayısı olduğunda E_i uzaklıkları Mahalonobis uzaklıkları olmaktadır. Bu uzaklıkların ortancası E_m ve

$$R_1 = E_m\sqrt{\frac{1+R^*}{2}} \quad , \quad R_2 = E_m\sqrt{\frac{1-R^*}{2}}$$

olmak üzere, $\theta \in [0, 2\pi]$ için,

$$X = T_X^* + (R_1 \cos\theta + R_2 \sin\theta)S_X^*$$

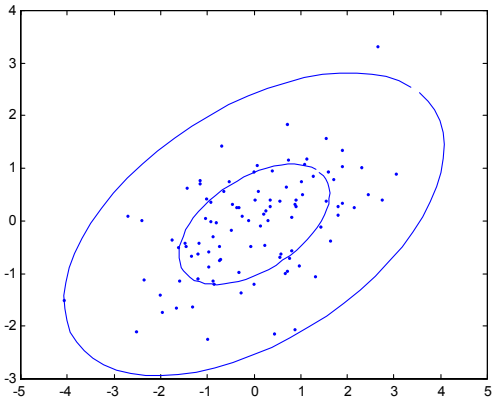
$$Y = T_Y^* + (R_1 \cos\theta - R_2 \sin\theta)S_Y^*$$

olarak elde edilen (X,Y) noktaları çantayı belirleyen iç elipsi oluşturmaktadır. Dış elips ile iç elipsin alanları oranlarının maksimumu D sabiti olmak üzere, yukarıdaki ifadelerde E_m yerine,

$$E_{\max} = \max\left\{E_i : E_i^2 < DE_m^2,\ i = 1,2,...,n\right\}$$

alınmasıyla elde edilen (X,Y) noktaları çit'i belirleyen dış elipsi oluşturmaktadır. D sabiti ile ilgili olarak, Goldberg ve Iglewicz (1992) bir gözlem için %99 luk bir güven sınırı oluşturmak amacıyla, $D = 7$ değerini önermektedirler (normal dağılım durumunda, $E_i \sim [2(n-1)/(n-2)]F_{(2,n-2)}$ olmak üzere, $n = 77$ için $F_{(2,n-2;0.99)} / F_{(2,n-2;0.50)}$ =7 dır).

$n = 100$ birimlik iki boyutlu bir veri için serpilme diyagramı ve yukarıdaki yöntemle belirlenen çanta çiziti Şekil 1 deki gibidir. Bu ve aşağıdaki şekillerdeki eksenler marjinal dağılımlardaki değişkenleri göstermektedir.



Şekil 1

Goldberg ve Iglewicz (1992) dört elips parçasından oluşan çiziti (quelplot) aşağıdaki gibi oluşturmaktadır.

$T_X^*, T_Y^*, S_X^*, S_Y^*, R^*$ yanında, elipsin eksenlerinin pozitif yönündeki artıkların toplam standart sapmasının oranını yansıtan asimetri parametreleri P_1 ve P_2 de kullanılarak,

$$Z_{1i} = \frac{Y_{si} + X_{si}}{\sqrt{2(1+R^*)}} \quad , \quad Z_{2i} = \frac{Y_{si} - X_{si}}{\sqrt{2(1-R^*)}}$$

$$F_{1i} = \begin{cases} \dfrac{Z_{1i}}{2P_1} & , Z_{1i} > 0 \\ \dfrac{Z_{1i}}{2(1-P_1)} & , Z_{1i} \le 0 \end{cases} \quad , \quad F_{2i} = \begin{cases} \dfrac{Z_{2i}}{2P_2} & , Z_{2i} > 0 \\ \dfrac{Z_{2i}}{2(1-P_2)} & , Z_{2i} \le 0 \end{cases}$$

$$E_i = \sqrt{F_{1i}^2 + F_{2i}^2} \quad , \quad i = 1,2,...n$$

hesaplanmaktadır.

$$R_1(-1) = 2(1-P_1)E_m\sqrt{\frac{1+R^*}{2}} \quad , \quad R_1(+1) = 2P_1E_m\sqrt{\frac{1+R^*}{2}}$$

$$R_2(-1) = 2(1-P_2)E_m\sqrt{\frac{1-R^*}{2}} \quad , \quad R_2(+1) = 2P_2E_m\sqrt{\frac{1-R^*}{2}}$$

olmak üzere, $\theta \in [0, 2\pi]$ için,

$$X = T_X^* + [R_1(\text{sgn}(\cos\theta))\cos\theta + R_2(\text{sgn}(\sin\theta))\sin\theta]S_X^*$$

$$Y = T_Y^* + [R_1(\text{sgn}(\cos\theta))\cos\theta - R_2(\text{sgn}(\sin\theta))\sin\theta]S_Y^*$$

olarak elde edilen (X, Y) noktaları çantayı oluşturmaktadır. Çit'i oluşturmak için yukarıdaki ifadelerde E_m yerine $E_{\max}$ kullanılmaktadır.

P_1 ve P_2 asimetri parametrelerini belirlemek için önerilen yöntemlerden birisi aşağıdaki gibidir.

$$Z_{1i}^{'} = \begin{cases} P_1(Z_{1i} - \hat{z}_1) & , Z_{1i} < \hat{z}_1 \\ (1-P_1)(Z_{1i} - \hat{z}_1) & , Z_{1i} \geq \hat{z}_1 \end{cases} \qquad Z_{2i}^{'} = \begin{cases} P_2(Z_{2i} - \hat{z}_2) & , Z_{2i} < \hat{z}_2 \\ (1-P_2)(Z_{2i} - \hat{z}_1) & , Z_{2i} \geq \hat{z}_2 \end{cases}$$

olmak üzere, P_1 ile $\hat{z}_1$ değerleri,

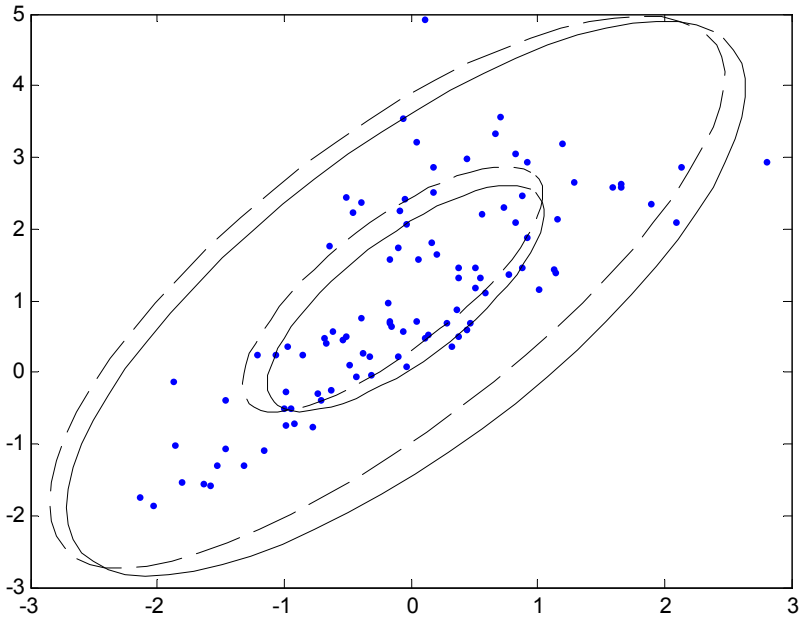
$$\sum_{i=1}^{n} Z_{1i}^{'} = 0 \quad , \quad \sum_{i=1}^{n} Z_{1i}^{'2}\,\text{sgn}(Z_{1i}) = 0$$

ve P_2 ile $\hat{z}_2$ değerleri,

$$\sum_{i=1}^{n} Z_{2i}^{'} = 0 \quad , \quad \sum_{i=1}^{n} Z_{2i}^{'2}\,\text{sgn}(Z_{2i}) = 0$$

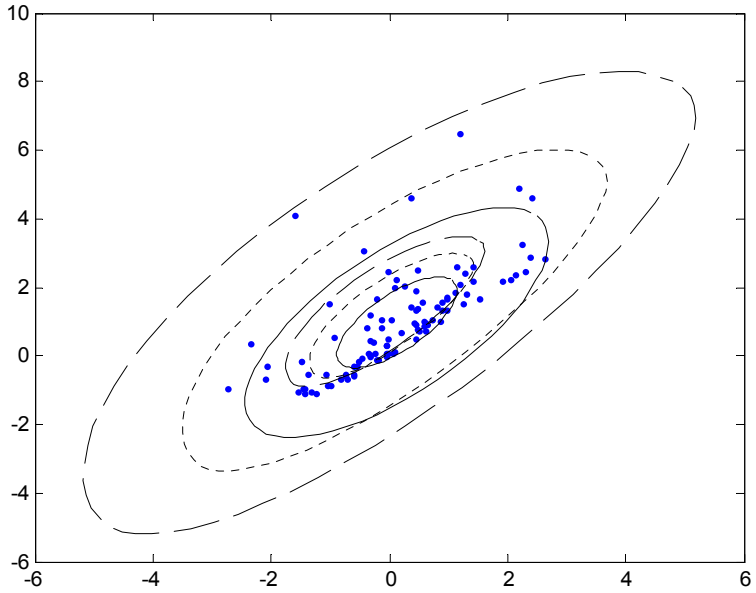
olacak şekilde belirlenmektedir [2].

$n = 100$ birimlik iki boyutlu bir veri için serpilme diyagramı ile tek elips (düz çizgi) ve dört elips parçasından oluşan çanta çiziti (kesikli çizgi) Şekil 2 deki gibidir.



Şekil 2

$T_X^*, T_Y^*, S_X^*, S_Y^*, R^*$ yerine başka istatistiklerin, özellikle konum ve ölçek parametreleri için dirençli (robust) tahmin edicilerin kullanılmasıyla elde edilen çizitler verileri daha iyi bir şekilde betimlemektedir. Bununla ilgili örnekler Goldberg ve Iglewicz (1992) de bulunmaktadır. Örneğin, T_X^*, T_Y^* yerine, $T_X = T_X^* + \frac{\hat{z}_1 - \hat{z}_2}{2} S_X$, $T_Y = T_Y^* + \frac{\hat{z}_1 + \hat{z}_2}{2} S_Y$ alınarak çizilen dört elips parçasından oluşan çanta çiziti (kesikli çizgi) diğerlerine göre (quelplot (noktalı çizgi), tek elips (düz çizgi)) serpilmeyi daha iyi yansıtmaktadır (Şekil 3).



Şekil 3

3. Derinlik kavramına dayalı olarak kutu çizitinin iki değişkenli verilere genişletilmesi

Bir boyutlu olasılık dağılımlarında uçlardaki noktalardan ortancaya doğru gittikçe, dağılımda daha derine doğru gidiyoruz sezgisine dayalı olarak bir derinlik kavramı tanımlanabilir. Örneğin, yüzdeliği (quantile) k olan bir noktanın derinliği $0.5 - |0.5 - k|$ olarak tanımlanırsa, böyle bir derinlik kavramı için birinci çeyreklik ile üçüncü çeyreklik aynı derinliğe sahip ve en derin nokta ortanca olacaktır. En derin noktaya merkez denirse, derinlik kavramına dayalı olarak bir merkez kavramı tanımlanmış olur.

Bir boyutlu dağılımlar için var olan ortanca kavramının çok değişkenli dağılımlara genişletilmesi kolay görünmemektedir. Çok değişkenli dağılımlar için derinlik ve buna dayalı olarak merkez kavramını oluşturmak daha kolay olmaktadır.

$x \in R^d$ verilen bir nokta ve F, d boyutlu X rasgele vektörünün R^d de tanımlı dağılım fonksiyonu olmak üzere, *x* noktasının F nin "merkezine" yakınlığının bir ölçüsü derinlik kavramına dayanılarak yapılabilir. Bunun örneklem karşılığı, $x \in R^d$ noktasının, $X_1, X_2, ..., X_n$ gözlem kümesinin (bulutunun) merkezine yakınlığının ölçüsü olarak ifade edilebilir [4].

Yarı Uzay Derinliği (Half-space Depth**):** $x \in R^d$ noktasının F dağılımına göre yarı düzlem derinliği,

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

olarak tanımlanır[4]. Yarı uzay derinliğinin örneklem karşılığı,

$$H\hat{D}(F;x) = \inf_{H}\left\{\frac{s\{X_i; X_i \in H\}}{n}; H, R^d \text{ de } x \text{ i içeren kapalı bir yarı hiperdüzlem}\right\}$$ biçi

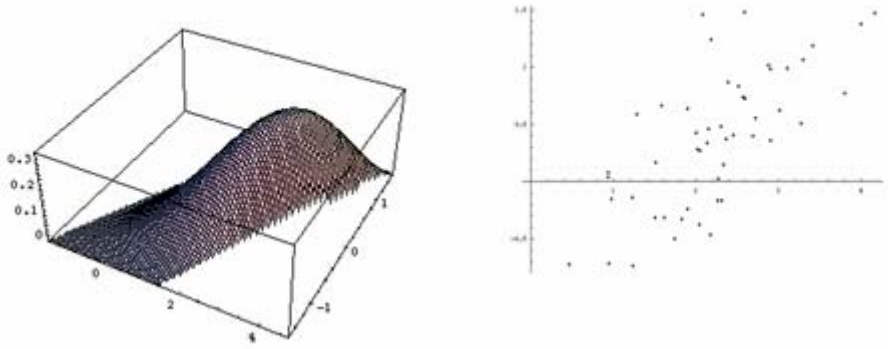
mindedir. (*s(A)*, *A kümesinin* eleman sayısını göstermektedir.)

Konveks Katman Derinliği (Convex Hull Peeling Depth**):** Birçok değişkenli dağılımdan alınan gözlem değerlerini içeren en küçük konveks küme bir çokyüzlü olmak üzere bu konveks kümenin köşe noktaları eldeki gözlemlerin birinci katmanı, ilk katman gözlemleri kaldırılıp geriye kalan gözlemlerin birinci katmanı gözlemler için ikinci katman olarak adlandırılsın ve takip eden katmanlarda aynı şekilde oluşturulsun. Buna göre $X_1, X_2, ..., X_n$ örneğinde X_k noktasının bu veri kümesine göre derinliği, X_k noktasının dâhil olduğu katmanın düzeyi (katman sıra sayısı), olarak adlandırılır [4]. Gözlemin, dâhil olduğu katman sıra sayısı büyüdükçe derinliği artıyor demektir. Burada katmanların oluşumu da bir soğanın katlarına benzetilebilir. Sadece örneklem için düşünülen bu derinliğin sürekli kitle dağılımları için karşılığı tanımsızdır.

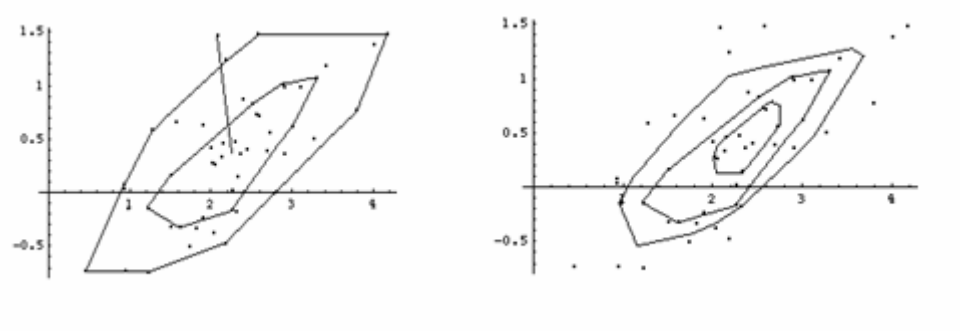
Yukarıda verilen derinlik ölçülerinden başka derinlik ölçüleri de vardır [3,4]. Derinliği en büyük olan noktaya *derinlik merkezi* ya da kısaca *merkez* denir. En büyük derinliğe sahip birden çok nokta bulunduğunda bunların ortalaması merkez olarak alınmaktadır. Derinlik sıralamasında, eşderinlikli gözlemlerin olması halinde sıra istatistiklerinde olduğu gibi işlem yapılmaz; aynı derinliğe sahip olan gözlemlere birbirlerini takip eden derinlik sıra numarası verilir (gözlem sayısı kadar derinlik sıra numarası söz konusudur).

$D(F;x)$ herhangi bir derinlik ölçüsü olmak üzere, $t \in [0,1]$ için $\{x : D(F;x) = t\}$ kümesine t derinlikli kontur veya seviye kümesi denir. $R(t) = \{x : D(F;x) > t\}$ kümesi t derinlikli kontur ile çevrili bölge olmak üzere, $C_p = \bigcap_t \{R(t) : P(R(t)) \geq p\}$ kümesine $p.$ merkezi bölge denir [4]. Bunun $\hat{C}_p$ örneklem karşılığı, np tamsayı olduğunda np tane, olmadığında $\lfloor np \rfloor + 1$ tane en derin gözlemi kapsayan en küçük konveks kümedir.

Çok boyutlu veriler için doğal bir sıralama sözkonusu olmamakla birlikte, yukarıda tanımlanan derinlik ölçüleri, gözlemleri dağılımın merkezinden dışarıya doğru sıralamaktadır. İki boyutlu sıralanmış gözlemlerin merkeze yakın olan %50'sini içeren konveks çokgene çanta denir. Çanta, tek boyutlu verilerin betimlenmesindeki kutunun karşılığıdır. Çantanın çevre noktalarının merkeze olan uzaklıklarını 3 ile çarpıp merkezden uzaklaştırarak çit (fence) elde edilir. Çitin dışında kalan noktalar sıradışı gözlem olarak nitelendirilir. Sıradışı gözlemler dışındaki gözlemleri içeren en küçük konveks çokgen yastık (bolster) olarak adlandırılır. Çanta koyu, etrafındaki yastık daha açık olarak renklendirilir ve çit görüntülenmeyebilir. Olasılık yoğunluk fonsiyonun grafiği Şekil 4 'de solda olan dağılımdan üretilen 50 birimlik bir örneklem için serpilme diyagramı aynı şeklin sağında olmak üzere, katman derinliğine göre çanta çiziti Şekil 5 'de soldadır. Çanta sınırından bir gözleme olan uzaklık, merkezden çanta sınırına olan uzaklığın 3 katından fazla olduğunda bu gözlem bir sıra dışı gözlem olarak nitelendirilir. Böyle bir gözlem, merkez ile birleştirilmiş çizginin ucundaki gözlemdir. Şekil 5 'in sağında çeyreklik çizgileri yer almaktadır. En içteki çeyreklik çizgisi gözlemlerin %25'ini, ortadaki %50'sini ve dıştaki %75'ini içermektedir [1,5].



Şekil 4



Şekil 5

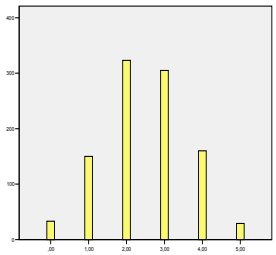
4. İki boyutlu veriler için çubuk grafiği ve histogram

Bir boyutlu veriler için histogram ile frekans poligonu örneklemin alındığı dağılımın olasılık yoğunluk fonksiyonunun biçimi hakında fikir vermektedir. Histogramlar, aralık ile oran ölçme düzeyinde (interval level of measurement, ratio level of measurement) gözlenen ve kitle dağılımı sürekli olan verilere uygulanır. Çubuk grafikleri, isimlendirme (nominal), sıralama (ordinal) , oran, aralık ölçme düzeyinde gözlenen ve kitle dağılımı kesikli olan verilere uygulanır. Bilindiği gibi çubuk grafiklerinde yatay eksende ölçülen özelliğin gözlenen değerleri, düşey eksende de bunların frekansları bulunmaktadır (Şekil 6 b,c). Bir boyutlu verilerde histogram; sınıf aralıkları üzerinde yükseklikleri o sınıfın frekansı olan bitişik dikdörtgenlerden oluşmaktadır (Şekil 7 b,c).

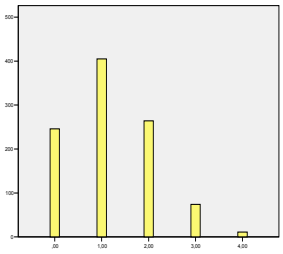
Görsel etki yaratacak şekilde, çubuk grafiği, histogram ve frekans poligonunu iki boyutlu verilere genişletmek mümkün olabilmektedir. İki boyutlu veriler için çubuk grafiği, üç boyutlu bir koordinat sisteminde, yatay düzlemde veriler için hazırlanan çapraz tablo (Şekil 6 a) ve düşey eksende göze frekansları olacak şekilde kolayca görüntülenebilir. Aşağıdaki çapraz tablo için çubuk grafiği Şekil 6 d ’dedir.

	0	1	2	3	4	
0	8	17	8	0	0	33
1	33	56	48	11	2	150
2	85	122	81	28	7	323
3	69	137	76	21	2	305
4	43	62	42	13	0	160
5	8	11	9	1	0	29
	246	405	264	74	11	1000

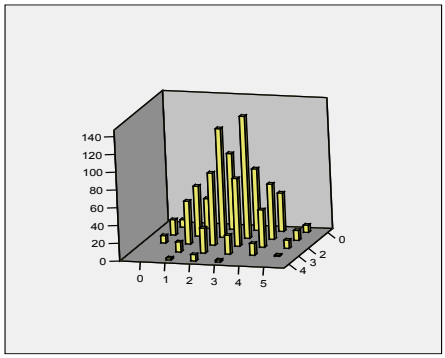
(a)



(b)



(c)

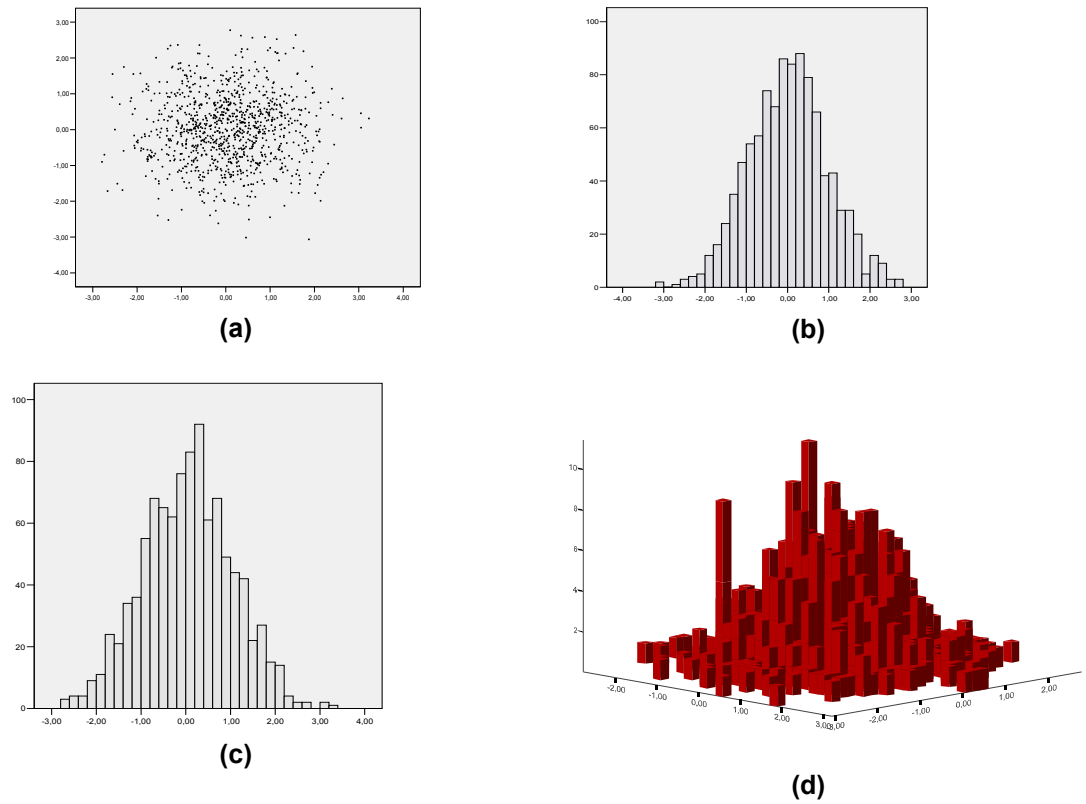


(d)

Şekil 6

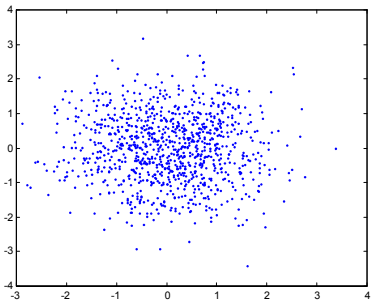
İki boyutlu veriler için histogram; tabanda eşit uzunluklu sınıf aralıklarının kartezyen çarpımı olan dikdörtgenler üzerinde, yükseklikleri o dikdörtgenin frekansı olan prizmalardan oluşturulabilir. Bu prizmaların üst yüzeylerinin konumları görsel etkiyi yaratmaktadır. Ayrıca, bilgisayar grüntülerine renk etkisi de katılabilir (Şekil 7 d ve Şekil 8 b). Bir boyutluda frekans poligonu; histogramdaki yanyana olan dikdörtgenlerin üst kenarlarının orta noktalarını birleştiren kırık çizgi olmak üzere, iki boyutluda; birbirine değen (taban kenarları ortak olan) dört prizmanın üst yüzeylerinin orta noktalarını birleştiren doğru parçalarının oluşturduğu çatı olarak gerçekleştirilebilir. Bunu, sadece çıtaları bulunan kiremitsiz bir çatıya benzetebiliriz. Bu çatı, verilerin alındığı iki boyutlu dağılımın olasılık yoğunluk fonksiyonunun biçimi hakkında görsel bilgi verir (Şekil 8 c). Genelde, histogram ve poligonlar dagılımın olasılık yoğunluk fonksiyonu hakında görsel bilgi vermekle birlikte, olasılık yoğunluk fonksiyonu tahmini oldukça derin bir istatistik teorisi gerektiren çekirdek tahmin yöntemleri ile yapılmaktadır.

İki boyutlu standart normal dağılımdan üretilen $n = 1000$ birimlik bir örneklem için serpilme diyagramı Şekil 7 a 'da, iki boyutlu veri için SPSS de çizilen histogram Şekil 7 d 'deki gibidir.

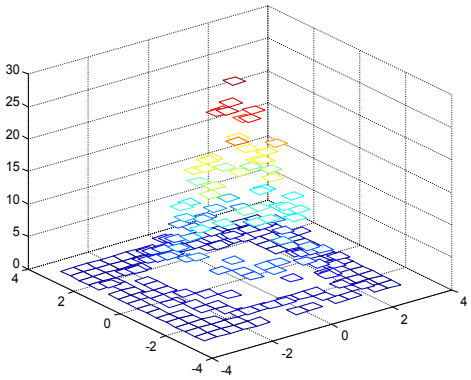


Şekil 7

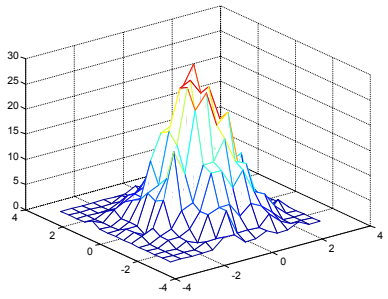
İki boyutlu standart normal dağılımdan üretilen $n = 1000$ birimlik başka bir örneklem için serpilme diyagramı Şekil 6 'da, iki boyutlu veri için histogram Şekil 7 ve poligon Şekil 8 deki gibidir. Bu şekiller ekteki MATLAB programı ile gerçekleştirilmiştir.



(a)



(b)



(c)

Şekil 8

Sonuç

Bir boyutlu veriler için var olan ve kolayca kavranan çubuk grafiği, histogram, frekans poligonu ve kutu çiziti gibi betimsel istatistiklerin iki boyutlu verilere genişletilmesi veri analizinde yararlı görsel bilgi elde edilmesini sağlamaktadır. Maalesef, bu kavramları daha yüksek boyutlara genişletmek görsel olarak bir fayda getirmemektedir. Bir boyutlu veriler için doğal tanımlaması olan sıra istatistiklerinin iki ve daha yüksek boyutlara doğrudan bir genişletilmesi yapılamamakla birlikte, derinlik gibi bazı kavramlar yardımıyla sıra istatistiklerine benzer istatistikler tanımlanabilmektedir. Burada, teorik esaslarına inilmeden yapılan kısa özetlemelerden görüldüğü gibi, iki boyutlu verilerin betimlenmesi oldukça çetin bir matematik altyapıya ve bilgisayar görüntüleme imkanlarına dayalı olduğu söylenebilir.

Bugünkü bilgisayar imkanlarının olmadığı yıllardaki istatistik eğitiminde örneklem ortalaması, örneklem varyansı, tepe değer, örneklem ortancası, örneklem çeyreklikleri, çubuk grafiği, histogram, frekans poligonu, eklemeli frekans poligonu gibi betimsel istatistikler ciddiyetle ve ayrıntılı bir şekilde ele alınmakta idiler. Bunların, adı üstünde, birer istatistik olduklarını gözden kaçırmadan ve bu istatistiklerin dağılım özelliklerini teorik olarak irdeleme gayreti içinde olmamız gerektiğini belirterek, iki boyutlu verilere genişletilmelerini de ele alıp eğitimimizde yeniden eski önemine kavuşturmamız gerekmektedir.

Kaynaklar

1. Aydoğdu,H., İ.Karabulut ve F.Öztürk (2000), Derinlik çizgileri ve çanta çizitleri, 5.Ulusal Biyoistatistik Kongresi Bildiri Kitabı, 191-199, Osmangazi Üniversitesi, Eskişehir.
2. Goldberg, K.M. and B.Iglewicz (1992), Bivariate extensions of the boxplot, Technometrics, Vol. 34, No.3, 307-320.
3. Liu, R.Y. (1990), On a notion of data depth based on random simplices, The Annals of Statistics, Vol. 18, No. 1, 405-414.
4. Liu, R.Y., J.M.Parelius and K.Singh(1999), Multivariate analysis by data depth: Descriptive statistics, graphics and inference(with dicussions), The Annals of Statistics, Vol. 27, No. 3, 783-858.
5. Rousseeuw, P.J. , I.Ruts and J.W.Tukey (1999), The bagplot: A bivariate boxplot, The American Statistician, Vol. 53, N0. 4, 382-387.
6. Tukey, J. W.(1977), Exploratory Data Analysis, Addison Wesley.

EK

```
n=1000;
veri=randn(2,n);
% serpilme diyagramı
figure
    plot(veri(1,:),veri(2,:),'.')

% marjinal dağılımlar için histogram
[frx,sx]=hist(veri(1,:),15);
[fry,sy]=hist(veri(2,:),15);
% sx ile sy marjinal dağılımlar için
% düşünülen 15 sınıfın sınıf ortaları
for i=1:15
  for j=1:15
    x1=sx(i)-(sx(2)-sx(1))/2 ;
    x2=sx(i)+(sx(2)-sx(1))/2 ;
    y1=sy(j)-(sy(2)-sy(1))/2 ;
    y2=sy(j)+(sy(2)-sy(1))/2 ;

    frekans=0;
    for ii=1:n
      if veri(1,ii)<x2
        if veri(1,ii)>=x1
          if veri(2,ii)<y2
              if veri(2,ii)>=y1
                  frekans=frekans+1;
              end,end,end,end
      end

    frpolig(i,j)=frekans;

    %histogram
    x=[x1 x2];
    y=[y1 y2];
    meshgrid(x,y);
    z=frekans*ones(2,2);
    mesh(y,x,z);
    hold on
  end
end

% iki boyutlu veri için poligon
figure
    meshgrid(sx,sy);
    mesh(sy,sx,frpolig);
```