ASTROİSTATİSTİK
5. KONU

Hazırlayan: Doç. Dr. Tolgahan KILIÇOĞLU

5. MOMENTLER, ÇARPIKLIK VE BASIKLIK

Bir verinin orta değeri ve yayılımına ilişkin ölçütlerin nasıl hesaplandığını gördük. Bu iki ölçütün verilmesi çoğu durumda verinin temsil edilmesi için yeterli olmaktadır. Ancak bu iki ölçüt veriye ilişkin iki bilgiden yoksundur: çarpıklık ve basıklık. Bu bölümde istatistikdeki moment kavramını anlatarak çarpıklık ve basıklığa ilişkin ölçütlerden söz edeceğiz.

Bölüm 4.2'de bir verideki değerlerin ortalama değerden olan sapmalarının toplamının sıfır olduğunu görmüştük. Bu farkların karelerinin ortalaması alındığında ise varyans (s^2) olarak adlandırdığımız ve karekökü alındığında standart sapmayı veren oldukça güçlü bir yayılım ölçütünü elde etmiştik (bkz. Bölüm 4.4). Peki verilerin farklarının küpü veya dördüncü kuvvetinin ortalamasının özel bir anlamı var mıdır? İstatistikte farkların kuvvetlerinin alınması sonucu elde edilen ölçütlere **momentler (μ)** adı verilir. Momentler ile verinin yayılımı, çarpıklığı ve basıklığı arasında önemli ilişkiler bulunur.

Bir veride dağılımının herhangi bir **a** değerine karşılık gelen **r.** momenti aşağıdaki gibi hesaplanır:

$$\mu_r(a) = \frac{\sum_{i=1}^{N} (x_i - a)^r}{N}$$

Burada N popülasyonun üye sayısıdır. Eğer moment hesaplanırken a yerine verinin ortalama değeri μ alınırsa moment **merkezi moment** adını alır. Merkezi moment kısaca μ_r şeklinde gösterilir ve genel ifadesi aşağıdaki gibidir:

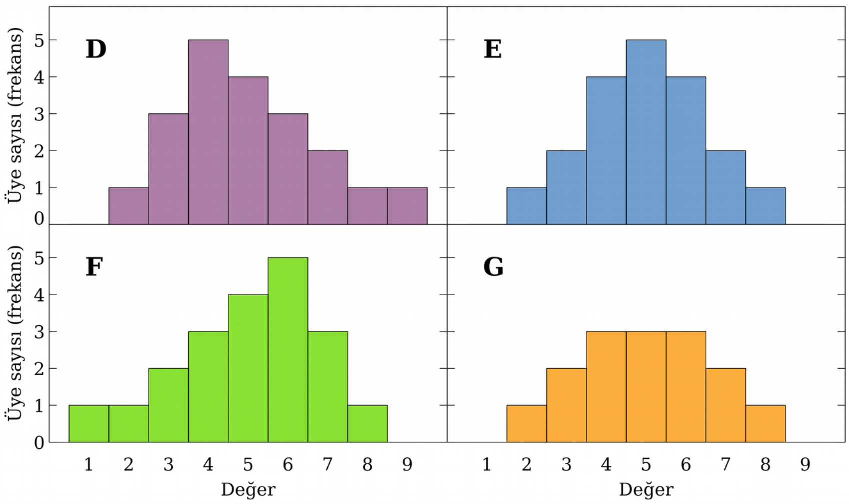
$$\mu_r = \frac{\sum_{i=1}^{N} (x_i - \mu)^r}{N}$$

Burada denklemin sağ tarafındaki μ popülasyonun ortalaması olup soldaki merkezi moment μ_r ile karıştırılmamalıdır. İkisi tamamen farklı kavramlardır.

Çizelge 5.1'de konu anlatımında kullanmak üzere dört tane basit D, E, F ve G popülasyon verisi verilmekte ve bu verilerin frekans dağılımı Şekil 5.1'de gösterilmektedir. Tüm verilerin ortalaması (μ) aynı ve 5'e eşittir. Verilerin standart sapmaları da birbirlerine çok yakın olup 1.5 ile 1.8 arasındadır. Ancak verilerin frekans dağılımına bakıldığında şekillerinin birbirlerinden farklı olduğu görülmektedir. Bu kez farklılık yayılımdan çok çarpıklıkların ve basıklıkların farklı olmasından kaynaklanmaktadır.

Çizelge 5.1 D, E, F ve G verileri

D					E					F					G				
5	3	6	4	7	4	5	6	7	8	3	5	7	4	6	5	3	4	6	8
8	6	2	5	4	6	2	5	4	7	6	8	5	6	1	6	7	5	2	4
4	5	3	4	6	5	6	3	4	5	7	6	3	4	5	4	3	7	5	6
7	5	4	9	3	3	4	6	5		5	2	4	7	6					



Şekil 5.1 D, E, F ve G verilerinin frekans dağılımının histogram grafiği ile gösterimi

Şimdi farklı moment türlerine ve bunların anlamlarına göz atmadan önce aşağıdaki iki soruya cevap veriniz.

Soru 5.1 D, E, F, G verilerinin simetrik, negatif çarpık ve pozitif çarpık durumlarından hangisine uyduklarını söyleyiniz.

Soru 5.2 E ve G verilerinden hangisinin en basık olduğunu söyleyiniz.

5.1 Birinci Merkezi Moment (μ_1)

Genel merkezi moment ifadesinde $r=1$ alınırsa **birinci merkezi moment**in ifadesi elde edilir:

$$\mu_1 = \frac{\sum_{i=1}^{N}(x_i-\mu)}{N}$$

Ancak Bölüm 4.2'de bu ifadeyi hesapladığımızda sıfıra eşit olduğunu görmüştük. Hatırlamak gerekirse, bir verideki değerlerin ortalamadan olan farkları toplanırsa sıfır değeri elde edilir. Çünkü ortalama değer bu verilerin tam ortasındadır. Bu durumda birinci merkezi moment için;

$$\mu_1 = 0$$

yazılır ve bu merkezi moment bize bir bilgi sağlamaz.

Burada dikkat edilmesi gereken nokta birinci **merkezi** momentin sıfır olmasıdır. Eğer ortalama μ değeri yerine başka bir a değeri alınırsa birinci moment sıfırdan farklı bir değere sahip olacak ve a değerinin konumuna ilişkin bir bilgi verecektir. Ancak şimdilik bunun ayrıntılarına girmeyelim ve sadece merkezi momentleri düşünelim.

5.2 İkinci Merkezi Moment (μ_2) ve Varyans

Genel merkezi moment ifadesinde $r=2$ alınırsa **ikinci merkezi moment**in ifadesi elde edilir:

$$\mu_2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Bu ifadeyi hatırladınız mı? Evet, bu ifade popülasyon için verilen varyans ifadesidir. Bu durumda ikinci merkezi moment bize **varyansı** vermektedir. Varyansı daha önce s^2 ile göstermiştik. Ancak s^2 gösterimi bir örneklemin varyansı için kullanılır. Popülasyonun varyansı ise σ^2 ile gösterilir. Bu durumda;

$$\mu_2 = \sigma^2$$

yazılabilir. Bu durumda σ popülasyonun standart sapması olacaktır. Görüldüğü gibi ikinci merkezi moment daha önce gördüğümüz bir kavram olan varyansa karşılık gelmektedir. Yani, μ_2 verinin yayılımı (saçılması) ile ilişkilidir.

5.3 Üçüncü Merkezi Moment (μ_3) ve Çarpıklık

Genel merkezi moment ifadesinde $r=3$ alınırsa **üçüncü merkezi moment**in ifadesi elde edilir:

$$\mu_3 = \frac{\sum_{i=1}^{N} (x_i - \mu)^3}{N}$$

Önceki bölümde, ikinci merkezi momenti (varyansı) hesaplarken değerlerin ortalamadan olan sapmalarının karelerinin alınması negatif olan değerlerin pozitife dönüşmesini sağlamaktadır. Bu sayede yayılıma sahip bir verinin varyans değeri hiçbir zaman sıfır veya negatif değerler almaz. Ancak, üçüncü merkezi momenti hesapladığımız ifadede farkların küpü alınmaktadır. Böylece negatif değere sahip olan sapmalar yine negatif kalmaktadır. İlk bakışta üçüncü merkezi momentin de tıpkı birinci merkezi moment gibi sıfır değerine eşit olması gerektiğini düşünebilirsiniz. Ancak, bu durum sadece ortalama etrafında simetrik bir dağılımın olması durumunda geçerlidir. Eğer verinin dağılımı ortalamanın etrafında simetrik değilse üçüncü merkezi moment sıfırdan sapar. Bu nedenle üçüncü merkezi moment çarpıklığın bir ölçütüdür.

Şimdi bir örnek olarak Çizelge 5.1'de değerleri yer alan F verisinin üçüncü merkezi momentini hesaplayalım:

$$\mu_3 = \frac{\sum_{i=1}^{N} (x_i - \mu)^3}{N} = \frac{(3-5)^3 + (5-5)^3 + (7-5)^3 + ... + (4-5)^3 + (7-5)^3 + (6-5)^3}{20} = -2.7$$

Benzer işlemler diğer veriler için de yapılırsa D, E ve F verilerinin üçüncü merkezi momentleri sırasıyla 2.7, 0 ve –2.7 olarak elde edilir. Görüldüğü gibi pozitif çarpık D verisinin üçüncü merkezi momenti pozitif, negatif çarpık F verisinin üçüncü merkezi momenti ise negatif değerlere sahiptir. Simetrik olan E verisi için ise üçüncü merkezi moment sıfır olarak elde edilmiştir.

5.3.1 Çarpıklık

Üçüncü merkezi moment de tıpkı varyansta olduğu gibi kullanışsız bir birime sahiptir. Örneğin, verideki değerler metre biriminde ise μ_3 metreküp biriminde olur. Bu uyumsuzluk verinin yorumlanmasını zorlaştırmaktadır. Bilindiği gibi standart sapma verideki değerlerle aynı birime sahiptir. Böylece, üçüncü merkezi momenti birimsiz hale getirmek için standart sapmanın küpüne bölebiliriz:

$$ç = \frac{\mu_3}{\sigma^3}$$

Bu sayede elde ettiğimiz birimsiz $ç$ ölçeği standartlaştırılmış bir çarpıklık ölçeği olur. Çarpıklık ölçeğinin $ç = 0$ olması dağılımın simetrik olduğunu, $ç > 0$ olması dağılımın pozitif (sağa) çarpık olduğunu, $ç < 0$ olması ise dağılımın negatif (sola) çarpık olduğunu gösterir.

D, E ve F verilerinin standart sapmaları sırasıyla 1.76, 1.49 ve 1.76 dır (bu değerleri hesaplayarak doğru olup olmadıklarını kontrol ediniz). Buna göre yine örnek olarak F verisinin çarpıklığını hesaplayalım:

$$ç_F = \frac{\mu_3}{\sigma^3} = \frac{-2.7}{1.76^3} = \frac{-2.7}{5.46} = -0.5$$

Diğer veriler için de benzer işlemler yapıldığında D, E ve F verilerinin çarpıklık değerleri sırasıyla 0.5, 0 ve –0.5 olarak bulunur. Pozitif çarpık D verisinin çarpıklık ölçeği pozitif, negatif çarpık F verisinin çarpıklık ölçeği ise negatif değerlere sahiptir. Simetrik olan E verisi için ise çarpıklık ölçeği beklenildiği gibi sıfırdır.

5.3.2 Diğer Pratik Çarpıklık Ölçekleri

Çarpıklığı ölçmenin kolay yollarından biri ortalama, mod ve medyan değerlerini kullanmaktır. Ancak bu ölçütler **sadece veri normal dağılıma yakın bir şekle sahipse** kullanılabilir. Daha önceki konularda simetrik olan bir dağılım için ortalama, mod ve medyan değerlerinin aynı olması gerektiğini ifade etmiştik. Bu noktada ortalama, mod ve medyan birbirlerinden sapmasından da çarpıklık ortaya konulabilmektedir.

Mod çarpıklığı aşağıdaki gibi ifade edilir:

$$\varsigma_{mod} = \frac{[\text{ortalama}] - [\text{mod}]}{\sigma}$$

Medyan çarpıklığı ise aşağıdaki gibi ifade edilir:

$$\varsigma_{medyan} = \frac{3([\text{ortalama}] - [\text{medyan}])}{\sigma}$$

Her iki parametrenin de yorumu standart çarpıklık ölçeği ς ile benzerdir; ancak bu üç ölçütün birbirlerine eşit olmaları gerekmez.

Soru 5.3 D, E ve F verilerinin mod ve medyan değerlerini hesaplayınız.

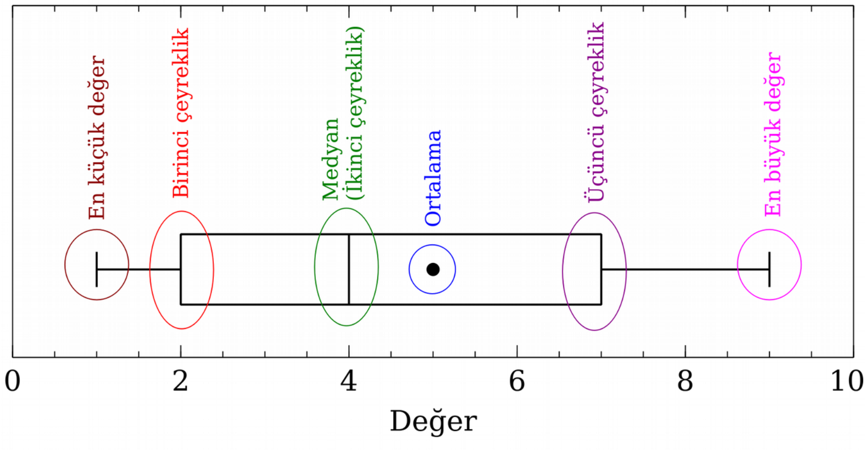
Cevap 5.4 Şekil 5.1 incelendiğinde D, E ve F verilerinin mod değerlerinin sırasıyla 4, 5 ve 6 olduğu hemen görülmektedir. Yine gerekli işlemler yapıldığında verilerin medyan değerlerinin birbirleriyle aynı ve 5'e eşit olduğu görülür. Bu aynı zamanda verilerin ortalama değerlerine eşittir.

D, E, ve F örnek verilerinin yukarıdaki cevapta elde edilen mod değerleri kullanıldığında verilerin mod çarpıklıklarının (ς_{mod}) sırasıyla 0.6, 0, ve –0.6 olduğu görülür. Bu değerler gerçektende standart çarpıklık ölçeğinden hesapladığımız değerlere çok yakındır (bu değerlerin standart ölçekle kıyaslandığında çok daha basit bir ifadeden hesaplandığını unutmayınız).

Her ne kadar mod çarpıklığı D, E, ve F verileri için standart çarpıklığa yakın değerler sunmada başarılı olsa da, medyan çarpıklığı için aynı şey söylenemez. Verilerin medyan ve ortalama değerleri aynı olduğundan hepsinin medyan çarpıklığı (ς_{medyan}) sıfıra eşittir Elbette bu çarpıklık her zaman sıfır olmak zorunda değildir. Burada bu durumun oluşmasının temel nedenleri verilerin kesikli olması (sadece tamsayılardan oluşması) ve verilerdeki çarpıklıklığın medyanı değiştirecek kadar kuvvetli olmamasıdır. Sonuç olarak, mod ve medyan çarpıklıkları her ne kadar pratik çarpıklık göstergeleri olsalar da her zaman standart çarpıklık ifadesi kadar güvenilir değillerdir.

5.3.3 Çeyreklikler ve Kutu-Bıyık Gösterimi

Bölüm 4.6'da bir verinin çeyrekliklerinin nasıl hesaplandığını görmüştük. Bir veri için çeyreklikler arası açıklık hesaplanarak verinin yayılımına ilişkin bir ölçüt elde ediliyordu. Ancak, çeyreklikler **kutu-bıyık grafiği** denilen bir grafik üzerinde gösterildiklerinde verinin çarpıklığına ilişkin de bilgi verirler. Kutu-bıyık grafiği, içinde bir nokta ve çizginin bulunduğu bir dikdörtgenden (kutu) ve bu dikdörtgenin iki tarafına doğru uzanan iki çizgiden oluşur. Şekil 5.2'de örnek bir kutu-bıyık grafiği ve grafik üzerindeki bölgelerin anlamları gösterilmektedir.



Şekil 5.2. Kutu-Bıyık grafiği ve üzerindeki göstergelerin temsil ettiği değerler

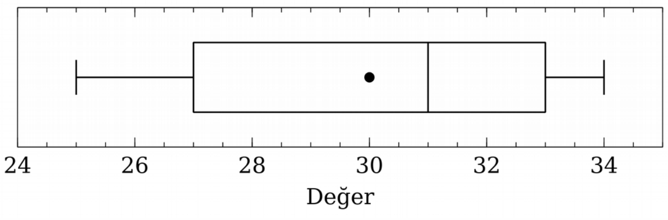
Soru 5.4 Aşağıdaki verinin kutu-bıyık grafiğini oluşturarak verinin çarpıklığını yorumlayınız.

30 25 34 25 29 31 32 34 31

Cevap 5.4 Değerlere gerekli işlemler yapıldığında aşağıdaki sonuçlar elde edilir:

Ortalama	= 30
1. Çeyreklik	= 27
2. Çeyreklik (Medyan)	= 31
3. Çeyreklik	= 33
En küçük değer	= 25
En büyük değer	= 34

Yukarıdaki değerler için kutu-bıyık diyagramı:



Kutu-bıyık diyagramına bakıldığında verinin negatif (sola) çarpık olduğu görülmektedir (bu yargıya nasıl vardığımızın cevabını siz veriniz).

5.4 Dördüncü Merkezi Moment (μ_4) ve Basıklık

Genel merkezi moment ifadesinde $r=4$ alınırsa **dördüncü merkezi moment**in ifadesi elde edilir:

$$\mu_4=\frac{\sum_{i=1}^{N}(x_i-\mu)^4}{N}$$

Dördüncü merkezi moment bir verinin basıklığının ortaya konmasında kullanılır. Basıklık verinin dağılımının tepe noktasının ne kadar sivri veya düz olduğunun bir ölçütüdür.

Örnek olarak Çizelge 5.1'deki G verisinin dördüncü merkezi momentini hesaplayalım:

$$\mu_4=\frac{\sum_{i=1}^{N}(x_i-\mu)^4}{N}=\frac{(5-5)^4+(3-5)^4+(4-5)^4+...+(7-5)^4+(5-5)^4+(6-5)^4}{15}=15.46$$

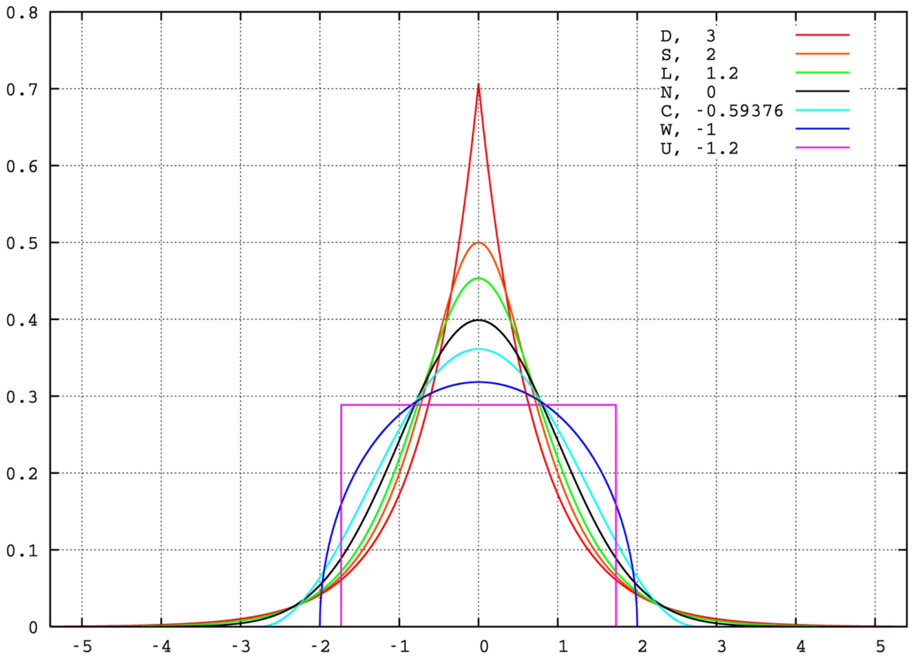
E verisi için benzer bir hesap yapıldığında ise $\mu_4=12.32$ elde edilir. Bu değerlerin şu anki ham halleriyle yorumlanması çok zordur.

5.4.1 Basıklık

Dördüncü merkezi momentin birimi de kullanışsızdır ve standartlaştırılması gerekir. Dördüncü merkezi momenti birimsiz hale getirmek için (üçüncü momentte yaptığımıza benzer şekilde) standart sapmanın dördüncü kuvvetine bölebiliriz:

$$b=\frac{\mu_4}{\sigma^4}-3$$

Burada elde ettiğimiz birimsiz b ölçeği standartlaştırılmış bir basıklık ölçeği olur. İfadeye dikkatli bakıldığında sağ tarafta sonuçtan 3 çıkartıldığı görülür. Şimdi bunun nedenini açıklayalım. Bilindiği gibi değerlerin 4. kuvveti alındığında (tıpkı karelerinin alınması gibi) negatif değerler pozitife dönecektir. Bu nedenle μ_4/σ^4 ifadesi hiçbir zaman negatif değer alamaz. Ayrıca bu değer tam 3'e eşit olduğunda verinin basıklığının normal dağılıma uygun olduğu görülür. Bu nedenle veriden 3 çıkarılarak normal dağılıma yakınsayan durumun 0 değerinde gerçekleşmesi sağlanır. Böylece, basıklık ölçeğinin $b=0$ olması normal dağılıma benzer bir basıklığın olduğunu, $b>0$ olması dağılımın normal dağılıma nazaran daha sivri olduğunu, $b<0$ olması ise dağılımın normal dağılıma nazaran daha basık olduğunu gösterir. Şekil 5.3'te farklı basıklık değerlerine karşılık gelen dağılımlar gösterilmektedir (D: Laplace Çift Üstel dağılım, S: Hiperbolik Sekant dağılımı, L: Lojistik dağılım, N: Normal dağılım, C: Yükseltilmiş Kosinüs dağılımı, W: Wigner yarıçember dağılımı, U: Tekdüze dağılım). Şekli incelerken gösterilen dağılımların hepsinin ortalamalarının ve standart sapmalarının birebir aynı olduğunu, çarpıklıklarının sıfır olduğunu ve sadece basıklığın değiştiğini göz önünde bulundurun.



Şekil 5.3 Farklı basıklık değerlerine sahip dağılımlar

Çizelge 5.1´deki verilerden E ve G nin basıklığı hesaplanırsa (burada standart sapmaların 4. kuvveti alındığından değerler biraz daha hassas yazılmıştır):

$$b_E = \frac{\mu_4}{\sigma^4} - 3 = \frac{12.32}{1.49^4} - 3 = -0.5$$

$$b_G = \frac{\mu_4}{\sigma^4} - 3 = \frac{15.46}{1.63^4} - 3 = -0.8$$

$b_G < b_E$ olduğundan G verisinin E verisinden çok daha basık olduğunu matematiksel olarak ortaya koymuş olduk.