ASTROİSTATİSTİK

4. KONU

Hazırlayan: Doç. Dr. Tolgahan KILIÇOĞLU

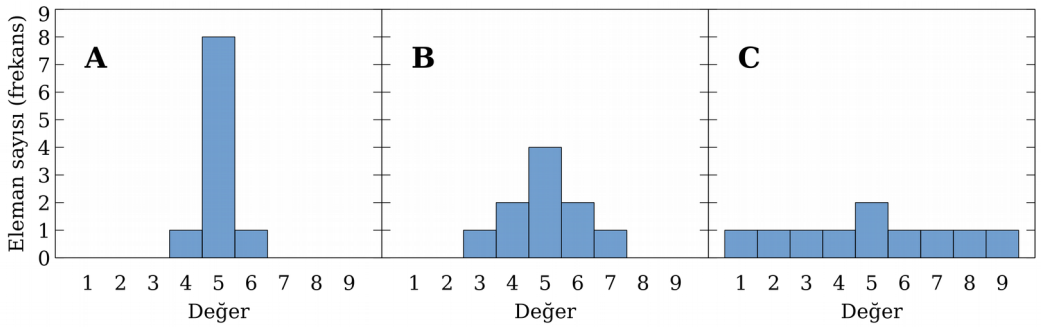
4. VERİLERİN YAYILIMININ BELİRLENMESİ

Bir veri tanımlanırken orta değer (ortalama, medyan veya mod) verilmesinin yanında verilerin yayılımına (saçılmasına) ilişkin de bir bilgi verilmesine ihtiyaç vardır. Öncelikte bu konuda kullanmak üzere birkaç veriyi ele alalım. Çizelge 4.1'de A, B ve C olarak adlandırılan 3 farklı veri yer almaktadır. Her verinin 10 üyesi bulunmaktadır. Şekil 4.1'de ise bu verilerin frekans dağılımları bir histogram üzerinde gösterilmektedir.

Çizelge 4.1 A, B ve C verileri

A						B						C				
5	5	5	5	4		5	5	4	3	6		2	5	4	3	8
5	6	5	5	5		7	5	6	5	4		7	1	5	6	9

Şekil 4.1 A, B ve C verilerinin frekans dağılımlarının histogram grafiği üzerinde gösterimi



Çizelge 4.1'deki A, B ve C verilerinin üçünün de mod, medyan ve ortalama değerleri birbirleri ile aynı ve 5'e eşittir. Ancak, Şekil 4.1'deki frekans dağılımlarına bakıldığında verilerin birbirlerine hiç benzemediği görülür.

Soru 4.1: Bu verilerden hangisi için hesapladığımız ortalama değer daha güvenilir olacaktır?

Cevap 4.1: Şekil 1.1'deki frekans dağılımları incelendiğinde A verisinin daha az saçılmaya sahip (daha duyarlı) olduğu görülür. Bu nedenle sorunun cevabı A verisidir. En duyarsız olanı ise C verisidir.

Bu sorunun cevabından da anlaşılacağı gibi bir veriyi doğru şekilde yorumlamak ve çıkarımlarda bulunmak için verinin ne kadar duyarlı olduğuna (yayılımına) ilişkin başka bir ölçüme de ihtiyaç vardır. Bu konu kapsamında verilerin yayılımının hangi ölçeklerle belirlendiğini göreceğiz.

4.1 Açıklık

Bir verinin değişkenliğini ortaya koymanın kolay yollarından biri verinin açıklığına bakmaktır. Bir başka deyişle, verideki en küçük ile en büyük değer arasındaki farka bakmaktır:

$$[\text{Açıklık}]=[\text{En büyük değer}]-[\text{En küçük değer}]$$

Buna göre veri aralıkları A verisi için 6–4=**2**, B verisi için 7–3=**4** ve C verisi için 9–1=**8** olarak elde edilir. Verilerin aralığı irdelendiğinde en az değişkenlik sergileyen verinin A verisi, en çok değişim sergileyenin ise C verisi olduğu görülmektedir.

Avantajlar: Veri aralığı yönteminin tek avantajlı yanı çok hızlı şekilde hesaplanabilir olmasıdır. Öyle ki, değerlere hiçbir özel işlem uygulamadan sadece göz gezdirerek dahi tespit edilebilir.

Dezavantajlar: Veri aralığı yöntemi sadece verideki en küçük ve en büyük değere bağlıdır. Arada kalan değerlerin dağılımını hiçbir şekilde yansıtmamaktadır. Burada örnek olarak sunduğumuz 3 veri de simetrik bir dağılım sergilemektedir. Ancak, bazı verilerde bu simetriyi bozan aykırı değerler (aşırı büyük veya aşırı küçük) bulunabilir. Bu durumda veri aralığı yöntemi verideki aykırı değerlerden son derece etkilenir. Sonuç olarak veri aralığı bir verinin değişkenliğinin ifade edilmesinde çok güvenilir değildir.

4.2 Ortalamadan sapmalar

Bir verideki her elemanın değerinin ortalamadan ne kadar saptığı bulunur ve bu değerler toplanırsa değişkenliğin temsil edilebileceği bir parametre elde edileceği düşünülebilir. Şimdi C verisi için bu hesabı yapalım. Çizelge 4.2'de ilk sütunda C verisindeki değerler ve ikinci sütunda bu değerlerin ortalamadan olan farkları verilmektedir.

Çizelge 4.2 C verisi ve verideki değerlerin ortalama değerden olan sapmaları

x_i	$(x_i-\bar{x})$
2	2 – 5 = –3
5	5 – 5 = 0
4	4 – 5 = –1
3	3 – 5 = –2
8	8 – 5 = 3
7	7 – 5 = 2
1	1 – 5 = –4
5	5 – 5 = 0
6	6 – 5 = 1
9	9 – 5 = 4
$\sum_{i=1}^{n}(x_i-\bar{x})=$	**0**

Ancak, değerlerin ortalamadan olan sapmaları toplandığında sıfır değeri elde edilir. Bu beklenmedik bir durum değildir; çünkü ortalama değer zaten verilerin tam ortasını temsil eder. Ortalama değere negatif yönden olan uzaklıklar ile pozitif yönden olan uzaklıklar birbirini dengelediğinden toplamları hangi veri için olursa olsun sıfır değerini verecektir. Bu nedenle, ortalamadan olan farkların doğrudan toplamı verinin değişkenliğini temsil etmede kullanılamaz.

4.3 Ortalama Mutlak Sapma

Soru 4.1: Ortalamadan olan sapmaların toplamının sıfır olmasını engellemek için sapmalara nasıl bir işlem yapılabilir?

Cevap 4.1: Ortalamadan daha küçük olan değerlerin ortalamadan olan farkları negatif değerler almaktadır. Eğer bu negatif değerler pozitif olarak alınırsa değerlerin birbirlerini yutması engellenmiş olur. Başka bir deyişle, verideki değerlerin ortalamadan olan sapmalarının mutlak değerlerinin alınması bu problemi çözebilir.

Bir verideki her elemanın değerinin ortalamadan ne kadar saptığı bulunur ve bu değerlerin mutlak değerinin ortalaması alınırsa **Ortalama Mutlak Sapma** değeri elde edilir. Ortalama mutlak sapmanın matematiksel ifadesi şöyledir:

$$\text{Ortalama Mutlak Sapma} = \frac{\sum_{i=0}^{n} |x_i - \bar{x}|}{n}$$

Çizelge 4.3 A, B ve C verilerinin mutlak sapmaları ve ortalaması

A VERİSİ		B VERİSİ		C VERİSİ	
x_i	$\lvert x_i - \bar{x}\rvert$	x_i	$\lvert x_i - \bar{x}\rvert$	x_i	$\lvert x_i - \bar{x}\rvert$
5	$\lvert 5-5\rvert = 0$	5	$\lvert 5-5\rvert = 0$	2	$\lvert 2-5\rvert = 3$
5	$\lvert 5-5\rvert = 0$	5	$\lvert 5-5\rvert = 0$	5	$\lvert 5-5\rvert = 0$
5	$\lvert 5-5\rvert = 0$	4	$\lvert 4-5\rvert = 1$	4	$\lvert 4-5\rvert = 1$
5	$\lvert 5-5\rvert = 0$	3	$\lvert 3-5\rvert = 2$	3	$\lvert 3-5\rvert = 2$
4	$\lvert 4-5\rvert = 1$	6	$\lvert 6-5\rvert = 1$	8	$\lvert 8-5\rvert = 3$
5	$\lvert 5-5\rvert = 0$	7	$\lvert 7-5\rvert = 2$	7	$\lvert 7-5\rvert = 2$
6	$\lvert 6-5\rvert = 1$	5	$\lvert 5-5\rvert = 0$	1	$\lvert 1-5\rvert = 4$
5	$\lvert 5-5\rvert = 0$	6	$\lvert 6-5\rvert = 1$	5	$\lvert 5-5\rvert = 0$
5	$\lvert 5-5\rvert = 0$	5	$\lvert 5-5\rvert = 0$	6	$\lvert 6-5\rvert = 1$
5	$\lvert 5-5\rvert = 0$	4	$\lvert 4-5\rvert = 1$	9	$\lvert 9-5\rvert = 4$
$\frac{\sum_{i=1}^{10} \lvert x_i - \bar{x}\rvert}{10} = \mathbf{0.2}$		$\frac{\sum_{i=1}^{10} \lvert x_i - \bar{x}\rvert}{10} = \mathbf{0.8}$		$\frac{\sum_{i=1}^{10} \lvert x_i - \bar{x}\rvert}{10} = \mathbf{2.0}$	

Çizelge 4.3'te A, B ve C verileri için mutlak sapmaların değerleri ve sonuçta elde edilen ortalama mutlak sapma değeri sunulmaktadır. A, B ve C verilerinin ortalama mutlak sapmalarının sırasıyla 0.2, 0.8 ve 2.0 olduğu görülmektedir. Bu durumda yine en değişken olan verinin C, en kararlı verinin ise A olduğu sonucuna varılır. Böylece verideki tüm değerleri hesaba katan ve değişkenliği temsil eden kullanışlı bir değer elde ettik.

Avantajlar: Bir verinin değişkenliğini tüm değerleri göz önünde bulundurarak hesaplar. Bu nedenle verilerin duyarlılıklarını karşılaştırmada kullanılabilir.

Dezavantajlar: Ortalama mutlak sapma değerinin işaret ettiği aralıkta verinin yüzde kaçının bulunduğu verinin dağılımına son derece bağlıdır. Örneğin, C verisinin ortalama mutlak sapması 2 dir. Verinin ortalama değeri 5 olduğuna göre 5–2=**3** ve 5+2=**7** değerleri arasında kalan 6 adet değer vardır. Veride toplam 10 eleman olduğuna göre bu aralık verilerin %60'ına karşılık gelmektedir. A verisinde de benzer bir hesap yapıldığında oranın %80 olduğu görülür. Bu değerler %50'nin üzerinde olduğundan miktarlarının yeterli olduğu düşünülebilir. Ancak, normal dağılıma en yakın B verisi için bu oran %40'a düşer! Her ne kadar ortalama mutlak sapma bir verinin değişkenliğini ortaya koymada doğru bir yöntem gibi gözükse de, bazı dağılımlar için aldığı değer verinin saçılmasını ortaya koymada yetersiz kalmaktadır. Bu anlamda bu sapma değeri standart olarak kabul edilmez.

4.4 Varyans

Bir veride değerlerin ortalamadan sapma miktarlarını negatif değerlerden arındırmak için mutlak değerlerini almak yerine **karelerini** de alabiliriz. **Değerlerin ortalamadan sapma miktarlarının karelerinin ortalamasına varyans denir**.

Popülasyonun ve örneklemin varyansı arasında küçük bir fark bulunmaktadır. Bir popülasyonun varyansı;

$$\sigma^2 = \frac{\sum_{i=0}^{n} (x_i - \mu)^2}{N}$$

ifadesi ile hesaplanır. Burada μ popülasyonun ortalama değeridir. Ancak söz konusu örneklem olduğunda hesapladığımız $\bar{x}$ ortalama değeri μ den daha uzakta (ve örneklemdeki değerlere daha yakın) olabilir. Bu nedenle bir örneklemin varyansı popülasyonun varyansından daha küçük çıkacaktır. Bu hatanın düzeltilmesi için Bessel bir örneklem için bulunan varyansın $n/(n-1)$ ile çarpılması gerektiğini bulmuştur. Bu terime **Bessel Düzeltmesi** adı verilir. Bessel düzeltmesi kullanıldığında bir örneklemin varyansı (s^2) için aşağıdaki ifade elde edilir:

$$s^2 = \frac{\sum_{i=0}^{n} (x_i - \bar{x})^2}{n-1}$$

Çizelge 4.4 A, B ve C verilerinin varyanslarının hesaplanması

A VERİSİ		B VERİSİ		C VERİSİ	
x_i	$(x_i-\bar{x})^2$	x_i	$(x_i-\bar{x})^2$	x_i	$(x_i-\bar{x})^2$
5	$(5-5)^2=0$	5	$(5-5)^2=0$	2	$(2-5)^2=9$
5	$(5-5)^2=0$	5	$(5-5)^2=0$	5	$(5-5)^2=0$
5	$(5-5)^2=0$	4	$(4-5)^2=1$	4	$(4-5)^2=1$
5	$(5-5)^2=0$	3	$(3-5)^2=4$	3	$(3-5)^2=4$
4	$(4-5)^2=1$	6	$(6-5)^2=1$	8	$(8-5)^2=9$
5	$(5-5)^2=0$	7	$(7-5)^2=4$	7	$(7-5)^2=4$
6	$(6-5)^2=1$	5	$(5-5)^2=0$	1	$(1-5)^2=16$
5	$(5-5)^2=0$	6	$(6-5)^2=1$	5	$(5-5)^2=0$
5	$(5-5)^2=0$	5	$(5-5)^2=0$	6	$(6-5)^2=1$
5	$(5-5)^2=0$	4	$(4-5)^2=1$	9	$(9-5)^2=16$
$\frac{\sum_{i=1}^{10}(x_i-\bar{x})^2}{9}=\mathbf{0.22}$		$\frac{\sum_{i=1}^{10}(x_i-\bar{x})^2}{9}=\mathbf{1.33}$		$\frac{\sum_{i=1}^{10}(x_i-\bar{x})^2}{9}=\mathbf{6.67}$	

A, B ve C verilerinin bu ifade kullanılarak varyans hesabı Çizelge 4.4'te verilmektedir. A, B ve C verileri için varyans değerlerinin sırasıyla 0.22, 1.33 ve 6.67 olarak elde edilir. Buradan gözükmektedir ki varyans bir verinin değişkenliğine oldukça bağımlı bir parametredir. İfadede farkların kareleri alındığından saçılma arttıkça varyansın değeri hızla artmaktadır. Bu anlamda varyans verilerin ne kadar dağıık olduğunu belirlemede kullanılabilir.

Avantajlar: Bir verinin değişkenliğini tüm değerleri göz önünde bulundurarak hesaplar ve bu değişkenliğe son derece bağlıdır. Bu nedenle verilerin duyarlılıklarını karşılaştırmada kullanılabilir.

Dezavantajlar: Varyans değerinin sahip olduğu birim kafa karıştırıcıdır ve yorumlanması zordur. Örneğin veride bulunan değerlerin birimleri metre (m) olsun. Bu durumda hesaplanan varyans değerinin birimi (kare alındığından dolayı) metrekare (m^2) olacaktır. Varyansın sahip olduğu birimle verideki değerlerin birimlerinin birbirleriyle uyuşmaması verinin yorumlanmasını oldukça zorlaştırmaktadır.

4.5 Standart Sapma

Varyansın birimini "kare"den kurtarmanın kolay bir yolu bulunur: varyansın karekökünü almak! Varyansın karekökü istatistikte en sık kullanılan yayılım göstergelerinden biridir ve **standart sapma** (s) olarak isimlendirilir. Aşağıda standart sapma için iki matematiksel ifade bulunmaktadır:

$$[\text{Standart sapma}]=\sqrt{[\text{Varyans}]}$$

$$s=\sqrt{\frac{\sum_{i=0}^{n}(x_i-\bar{x})^2}{n-1}}$$

Çizelge 4.4'de A, B ve C verileri için elde edilen varyans değerlerinin karekökleri alınırsa bu verilerin standart sapmaları sırasıyla 0.5, 1.2 ve 2.6 olarak bulunur. Bu değerler, önceki bölümde 0.2, 0.8 ve 2.0 olarak hesapladığımız ortalama mutlak sapmalardan bir miktar daha fazla olduğu görülmektedir.

Not: *Standart sapmanın popülasyon için hesaplandığı durumlarda paydaya yine* $(n-1)$ *yerine* N *yazılması gerektiğini unutmayınız.*

Avantajlar: Standart sapma bir verinin değişkenliğini tüm değerleri göz önünde bulundurarak hesaplar ve verilerle aynı birimdedir. Standart sapma, ortalama mutlak sapmaya nazaran daha fazla değer aralığını kapsar. Örneğin, B verisinin standart sapması 1.2 ve verinin ortalaması 5 olduğuna göre, 5–1.2=**3.8** ve 5+1.2=**6.2** değerleri arasında 8 eleman bulunur. Veride toplam 10 eleman bulunduğuna göre 5±1.2 standart sapma aralığı verilerin %80'ini kapsamaktadır (ortalama mutlak sapmanın bu veri için %40'da kaldığını hatırlayınız). Burada örnek olarak verdiğimiz veriler kesiklidir ve oldukça az elemandan oluşmaktadır. Gerçekte normal dağılım sergileyen bir verinin %68'inden fazlası standart sapma aralığında kalır. Bu anlamda standart sapmanın değeri ortalama mutlak sapmaya nazaran daha güvenilirdir. Standart sapma verilerin duyarlılıklarını karşılaştırmada kullanılabilecek ideal ölçütlerdendir.

Dezavantajlar: Standart sapma da aykırı değerlere oldukça bağımlıdır.

4.6 Çeyreklikler Arası Açıklık

Bir veride üç tane çeyreklik bulunur. Bu çeyreklikler birinci, ikinci ve üçüncü çeyreklikler olarak adlandırılır. Bir verideki değerler küçükten büyüğe doğru (veya tersine doğru) sıralandığında tam ortaya denk gelen değerin medyan değeri olduğunu daha önce söylemiştik. Bu değere aynı zamanda **ikinci çeyreklik ($Ç_2$)** denir. İkinci çeyreklik verileri ortadan ikiye böler. İkinci çeyrekliğin solunda kalan verilerin medyanına **birinci çeyreklik ($Ç_1$)**, sağında kalan verilerin medyanına ise **üçüncü çeyreklik ($Ç_3$)** denir. Bir başka deyişle, birinci, ikinci ve üçüncü çeyreklik sıralanmış bir veride baştan %25, %50 ve %75 ilerlendiğinde karşılaşılan değerlerdir.

Çeyreklikler belirlendikten sonra çeyreklikler arası açıklık aşağıdaki ifade ile hesaplanır:

$$\text{ÇAA}=Ç_3-Ç_1$$

Soru 4.2 Aşağıdaki verinin çeyrekliklerini hesaplayarak çeyreklikler arası açıklığı bulunuz.

6 8 1 7 5 5 2

Cevap 4.2

i) Öncelikle verileri küçükten büyüğe doğru sıralayalım:

1 2 5 5 6 7 8

ii) Verilerin tam ortasına denk gelen sayı ikinci çeyreklik (yani medyan) olacaktır:

iii) Şimdi ikinci çeyrekliği veride olmadığını düşünelim. Bu durumda ikinci çeyrekliğin solunda kalan verinin medyanı birinci çeyreklik sağında kalan verinin medyanı ise ikinci çeyreklik olacaktır:

iv) Verinin çeyreklikler arası açıklığı hesaplanır:

$$\text{ÇAA} = Ç_3 - Ç_1 = 7 - 2 = 5$$

Soru 4.3 Aşağıdaki verinin çeyrekliklerini hesaplayarak çeyreklikler arası açıklığı bulunuz.

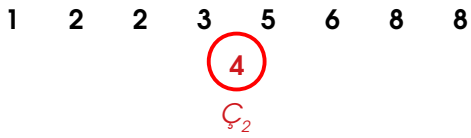
5 6 2 1 3 8 2 8

Cevap 4.3

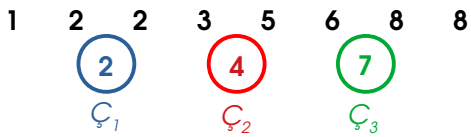
i) Öncelikle verileri küçükten büyüğe doğru sıralayalım:

1 2 2 3 5 6 8 8

i) Verilerin tam ortasına denk gelen sayının ikinci çeyreklik (yani medyan) olması gerekir. Ancak bu veride üye sayısı çift olduğundan orta noktaya iki veri denk gelmektedir. Medyanın bulunması için bu değerin ortalaması alınır:



iii) İkinci çeyreklik değeri veriye ait olmadığından diğer tüm değerler yerinde kalır. İkinci çeyrekliğin solundaki değerlerin medyanı alınarak birinci çeyreklik, sağındaki değerlerin medyanı alınarak ise üçüncü çeyreklik bulunur. Ancak yine medyana karşılık gelen değerler iki adet olduğundan ortalamaları alınır.



iv) Verinin çeyreklikler arası açıklığı hesaplanır:

$$\text{ÇAA} = Ç_3 - Ç_1 = 7 - 2 = 5$$

Soru 4.4 A, B ve C verileri için çeyreklikler arası uzaklığı hesaplayınız.

Cevap 4.4 Gerekli işlemler yapıldığında A, B ve C verilerinin çeyreklikler arası uzaklıklarının sırasıyla 0, 2 ve 4 olduğu elde edilir. Görüldüğü gibi ÇAA da verilerin yayılımını doğru olarak verebilmiştir.

Avantajlar: Çeyreklikler arası açıklık diğer yayılım ölçekleri ile karşılaştırıldığında aykırı değerlerden etkilenmez veya çok az etkilenir. Örneğin, Soru 4.3'de son değerin 8 yerine 8000 olduğunu düşünün. Verinin ÇAA değeri yine aynı olacaktır.

Dezavantajlar: Çeyreklikler arası açıklık ölçeği sadece çeyrekliklerle hesaplanır. Çeyrekliklerin aralarında kalan değerler tam olarak temsil edilmemektedir.

Çeyreklikler arası açıklık genellikle belirgin aykırı değelere sahip olan veriler için kullanılır. Bu ölçeğin tek başına verilmesi yerine standart sapma ile birlikte ifade edilmesi daha anlamlı ve kullanışlıdır.