Merkezi Eğilim ve Dağılım Ölçüleri

Merkezi eğilim ölçüleri, veri setini tanımlamak için tüm elemanları dikkate alarak veri setini özetleyen ölçülerdir.

Aritmetik ortalama: Veri setinde olan tüm elemanların toplanıp eleman sayısına bölünmesiyle elde edilir.

Medyan: Veri grubundaki elemanlar küçükten büyüğe sıralandığında elemanları ortadan ikiye ayıran değerdir. Yani ortadaki elemandır.

Elimizdeki veriler simetrik bir dağılım göstermiyorsa *aritmetik ortalama* yerine *medyan* ölçüsünü kullanıyoruz.

A 15, 7, 9, 25, 13, 2, 30, 11, 3, 18, 5, 2, 4, 35, 11

B 1, 2, 1, 3, 3, 4, 5, 7, 2, 80, 1, 3, 4, 50, 24

Aritmetik Ortalama: 12.6

B sınıfının sınav notları kötüdür. Ama A sınıfıyla aynı başarıda olduğu gözüküyor. Sınıftan bir kişinin yüksek not alması başarı oranını arttırmış. İşte bu gibi durumlarda verilerimizin aykırı değerlerden etkilenmemesi için *medyan* ölçüsünü kullanıyoruz.

Medyanlarını bulalım:

A 2, 2, 3, 4, 5, 7, 9, 11, 11, 13, 15, 18, 25, 30, 35

B 1, 1, 1, 2, 2, 3, 3, 3, 4, 4, 5, 7, 24, 50, 80

Aritmetik Ortalama: 12.6

A sınıfının medyan değeri 11, aritmetik ortalamaya çok yakın. Buradan sınıftaki öğrencilerin başarı seviyesinin aynı olduğunu anlayabiliyoruz.

B sınıfının medyan değeri ise 3, aritmetik ortalamaya biraz daha uzak. Yani sınıfın seviyesi biraz daha düşük.

Eğer elimizdeki verilerin sayısı çift sayı ise ortadaki iki elemanın ortalaması bizim medyan değerimizdir.

$$2, 2, 3, 4, 5, 7, 9, \boxed{11, 11,} 13, 15, 18, 25, 30, 35, 42$$

$$11$$

Mod: Veri setinde en çok tekrar eden değere denir. Yani mod ile en fazla tekrar eden verileri öğrenebiliriz. Modu daha çok kategorik verilerde kullanıyoruz.

Merkezi dağılım ölçüleri, veri setindeki elemanların dağılımı normal olmadığı durumlarda veri setini özetleyen ölçüler.

Kimi zaman karşılaştırma yaptığımız verilerin sadece aritmetik ortalaması değil medyan ölçüsü de aynı olabilir. Bu yüzden dağılım ölçülerini hesaplarız.

211 cm, 112 cm, 128cm, 203 cm, 109 cm	156 cm, 147 cm, 150cm, 145 cm, 165 cm
A.O: 152.6 cm	A.O: 152.6 cm
211 - 109 = 102 cm	165 - 145 = 20 cm

Boy uzunlukları verilen iki grubun ortalamalarına baktığımızda birbirine benzer olduğunu söyleyebiliriz. Fakat ilk grubun ölçülerine baktığımızda en büyük değer ile en küçük değer arasında 1 metreye yakın fark olmasına rağmen diğer grupta da 20 santim fark var. Bu yüzden ilk gruptaki veriler daha dağınık olduğu için dağılım ölçülerine göre karşılaştırma yapmalıyız.

Varyans: Her bir değerin ortalamaya uzaklığının karesinin ortalamasını verir. Varyans, sigma kare ile gösterilir. Tüm değerlerin ortalamaya uzaklığının karesini toplayıp eleman sayısına böldüğümüzde varyans ölçüsünü bulmuş oluyoruz.

$$\sigma^2 = \frac{(211 - 152.6)^2 + ... + (109 - 152.6)^2}{5}$$

$$= 2021.04 \text{ cm}^2$$

$$\sigma^2 = \frac{(156 - 152.6)^2 + ... + (165 - 152.6)^2}{5}$$

$$= 52.24 \text{ cm}^2$$

İki grubun varyans ölçüleri değerleri şekildeki gibidir. Sonuca göre ikinci grup yani varyans ölçüsü değeri 52.24 olan daha az dağınık verilerden oluşuyordur.

Uzunluklarına baktığımızda birimleri farklıdır. Varyans ölçülerinde bu sorunla karşılaştığımız için dağılım ölçülerinde daha çok standart sapmayı kullanıyoruz. Varyans ile standart sapma birbirine bağlıdır. Varyans sigma karedir. Standart sapma ise sigmadır. Standart sapmayı bulmak için varyansın karekökünü alıyoruz.

$$\sigma^2 = 2021.04 \text{ cm}^2 \qquad \sigma = \sqrt{\sigma^2} = 44.95$$

$$\sigma^2 = 52.24 \text{ cm}^2 \qquad \sigma = \sqrt{\sigma^2} = 7.22$$

Standart sapma ile değerlerin ortalamadan ne kadar saptığını daha iyi görebiliyoruz. Şekilde ikinci grubun standart sapması daha düşük olduğu için ikinci grubun boy uzunlukları daha benzerdir.

Bu ölçüleri *Verileri Hazırlama* aşamasında kullanıyoruz. Yani eksik ve hatalı olan verileri düzenliyoruz. Çoğunlukla düzeltme işleminde mod veya medyan gibi merkezi eğilim ölçülerini kullanıyoruz. Böylece geliştirilen yapay zekâ uygulamasında daha doğru sonuçlar verecektir.

KAYNAKÇA

Bilgeiş “Herkes için Yapay Zekâ I” eğitimi.