ASTROİSTATİSTİK

3. KONU

Hazırlayan: Doç. Dr. Tolgahan KILIÇOĞLU

3. VERİLERDE ORTA DEĞER BULMA

Bir veriyi tek bir değerle temsil etmeniz istenirse aklınıza hangi değer gelir? Böyle bir temsil yapılabilmesi için elbette veride yer alan değerlerin ortasını bulmanız gerekir. Bir bisküvi fabrikası düşünelim. Fabrika yoğun çalıştığı zamanlara ürettiği paket sayısı artmaktadır. Tatil olduğunda (örneğin Pazar günleri) ise hiç bisküvi üretmemektedir. Böyle bir fabrikanın ne miktarda bisküvi ürettiğini anlamak için günlük ortalamasına bakmak faydalı olacaktır. Fabrikanın günde ortalama 500 paket bisküvi üretildiği söylendiğinde fabrikanın büyüklüğü ve üretim kapasitesine ilişkin kullanışlı bir bilgi ederiz. Benzer şekilde Türkiye'de her 2 saniyede ortalama 5 bebeğin doğduğu ifadesi bebek doğum oranı hakkında oldukça net bir bilgi vermektedir.

Bir verinin ortasını bulmanın veri türüne ve amaca uygun farklı türleri bulunmaktadır. Bu türlere geçmeden önce orta değerini bulmak üzere örnek bir veriyi ele alalım. Çizelge 3.1'de galaksimizde bulunan açık yıldız kümelerinden rastgele seçilmiş 57 tanesinin çapları parsek cinsinden verilmektedir.

Çizelge 3.1 Açık yıldız kümelerin çapları (pc)*
Kaynak: van Den Berg 2006. AJ, 131, 1559.

Küme adı	D (pc)	Küme adı	D (pc)	Küme adı	D (pc)	Küme adı	D (pc)
NGC 2343	2	Haffner 6	5	Haffner 8	2	Pismis 17	6
Collinder 268	3	NGC 7788	3	IC 1805	11	NGC 744	2
Platais 1	4	Collinder 228	9	Stock17	1	NGC 4337	1
NGC 7044	6	Waterloo 2	1	Lynga 2	3	IC 2581	4
King 5	3	Ruprecht 79	3	NGC 2527	2	IC 4996	3
NGC 6405	3	NGC 6087	4	Collinder 261	6	Harvard 10	10
Blanco 1	5	Pismis 11	2	NGC 2281	4	Pismis 5	1
Hogg 10	2	NGC 2682	7	NGC 6242	3	Berkeley 12	4
NGC 3255	1	Bochum 5	4	NGC 957	5	NGC 689	3
Tombaugh 1	4	Basel 12	2	Trumpler 35	2	Basel 20	6
IC 348	1	Berkeley 19	6	Ruprecht 32	8	NGC 6494	5
Trumpler 15	8	NGC 2671	3	Pismis 1	5	Melotte 22	2
NGC 6994	2	Pismis 2	4	NGC 189	1	NGC 6253	2
NGC 6871	13	Sher 1	2	King 10	4	Stock 8	7
NGC 3572	3						

*Tüm listeye şu linkten ulaşılabilir: https://www.univie.ac.at/webda/vdb.html

3.1 Ortalama

Ortalamanın 3 alt türü vardır: Aritmetik, harmonik ve geometrik. İstatistikte kullanılan en yaygın ortalama **aritmetik ortalama**dır. Bu nedenle bir verinin ortalaması dendiğinde aksi belirtilmedikçe aritmetik ortalama anlaşılır. Aritmetik ortalama dışındaki diğer ortalamaları ilerleyen konularda kullanmayacağız; ancak bu kavramların ifadelerinin bilinmesinde fayda olduğundan bu başlık altında diğer ortalama türlerinden de söz edeceğiz.

3.1.1 Aritmetik ortalama

Aritmetik ortalama en yaygın kullanılan ortalama türüdür. Bu satırları okumadan önce dahi hayatınızda muhtemelen birçok defa belirli sebeplerle aritmetik ortalama almış olmalısınız. **Aritmetik ortalama bir verideki tüm değerlerin toplamının eleman sayısına bölümü**yle elde edilir:

$$\overline{x} = \frac{\sum_{i=1}^{n} x}{n}$$

Burada $\overline{x}$ gösterimi aritmetik ortalama için kullanılır. Toplam değer sayısı yine n ile gösterilmiştir. Çizelge 3.1'de verilen veri için aritmetik ortalama hesaplanırsa;

$$\overline{x} = \frac{228}{57} = 4\,\text{pc}$$

olduğu görülür.

Aritmetik ortalama (ve diğer ortalama türleri) **sadece nicel verilere** uygulanabilir. Nicel veriler aralık ölçümü veya oranlanabilir ölçüm olabilir ve değerleri kesikli veya sürekli olabilir.

3.1.2 Harmonik ortalama

Bazı durumlarda oran, zaman veya hız içeren değişkenlerin ortalamasının alınması için harmonik ortalama kullanılır. **Harmonik ortalama toplam eleman sayısının elemanların değerlerinin çarpmaya göre terslerinin toplamına bölümü**yle elde edilir:

$$\text{HO} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Örnek 1. Katiplik sınavına giren bir aday 200 kelimeyi ilk sınavda 5 dakikada, ikinci sınavda 4 dakikada, üçüncü sınavda ise 6 dakikada yazdığına göre adayın 200 kelimeyi ortalama kaç dakikada yazdığını bulunuz.

$$\text{HO} = \frac{3}{\frac{1}{5}+\frac{1}{4}+\frac{1}{6}} \simeq 4.9 \text{ dk}$$

Örnek 2. Bir araba Ankara'dan Eskişehir'e 80 km s^{-1} ile gidip 60 km s^{-1} ile geri dönüyor. Arabanın ortalama hızını hesaplayınız.

$$\mathrm{HO}=\frac{2}{\frac{1}{80}+\frac{1}{60}}\simeq 68.6\ \mathrm{km\ s^{-1}}$$

Çizelge 3.1'de verdiğimiz verinin harmonik ortalaması hesaplanırsa açık kümelerin çaplarının harmonik ortalamasının;

$$\mathrm{HO}=\frac{57}{21.66}=2.6\,\mathrm{pc}$$

olduğu görülür. Harmonik ortalamanın aritmetik ortalamadan (4 pc) daha küçük olduğu görülmektedir. Bu kural her türlü veri için her zaman geçerlidir.

3.1.3 Geometrik ortalama

İş ve finans sektöründe (örn. yüzde oranlar kullanılarak büyüme oranlarının hesaplanmasında) geometrik ortalamanın kullanılmasına ihtiyaç duyulabilir. **Geometrik ortalama veriyi oluşturan değerlerin çarpımının eleman sayısı mertebesinde kökü**dür:

$$\mathrm{GO}=\sqrt[n]{\prod_{i=1}^{n} x_i}$$

Çizelge 3.1'de verdiğimiz verinin geometrik ortalaması hesaplanırsa açık kümelerin çaplarının geometrik ortalaması aşağıdaki gibi elde edilir:

$$\mathrm{GO}=\sqrt[57]{1.8655\cdot 10^{19}}\simeq 3.3\ \mathrm{pc}$$

Geometrik ortalamanın aritmetik ortalamadan (4 pc) daha küçük olduğu ancak harmonik ortalamadan büyük olduğu görülmektedir. Bu kural her türlü veri için her zaman geçerlidir.

3.2 Mod

Bir veride en sık tekrarlanan değere o verinin **mod**u denir. Çizelge 3.1'de açık kümelere ilişkin verinin modunu bulmaya çalışalım. Çizelge 3.2'de açık yıldız kümelerinin çaplarına ilişkin verinin bir sınıflanmamış frekans dağılımı (her çapın karşısında o çapa sahip olan kümelerin sayısı) görülmektedir. Frekans analizi incelendiğinde verinin mod değerinin (yani en fazla elemana sahip olan çap değerinin) **2 pc** olduğu görülmektedir.

Nitel, sıralama ve nicel verilerin (yani her tür verinin) modu alınabilmektedir. Dahası, nominal nitel verilerin sadece modu alınabilmektedir. Sınıflanmış nicel verilerde en fazla elemana sahip olan sınıf **mod sınıfı** olarak adlandırılabilir. Bu durumda modun tam değeri bilinmez ama mod değerinin mod sınıfı içinde olduğu bilinir. Böylece mod sınıfının temsili değeri (sınıfının alt ve üst sınırının ortalaması) yaklaşık mod değeri olarak kabul edilebilir.

Çizelge 3.2 Açık yıldız kümelerin çap verisinin frekans analizi

D (pc)	f_i	D (pc)	f_i	D (pc)	f_i
1	7	6	5	11	1
2	12	7	2	12	0
3	11	8	2	13	1
4	9	9	1		
5	5	10	1		

3.3 Medyan

Veriler küçükten büyüğe kadar sıralandığında tam ortada kalan değere **medyan** denir. Çizelge 3.1'de açık kümelere ilişkin verinin medyanını bulalım. Bunun için öncelikle verideki değerleri Çizelge 3.3'de görüldüğü gibi küçükten büyüğa doğru sıraya sokalım.

Çizelge 3.3 Açık yıldız kümelerin çap verisinin büyükten küçüğe doğru sıralanması

Sıra no.	D (pc)	Sıra no.	D (pc)	Sıra no.	D (pc)	Sıra no.	D (pc)	Sıra no.	D (pc)
1	1	*13*	2	*25*	3	*37*	4	*49*	6
2	1	*14*	2	*26*	3	*38*	4	*50*	7
3	1	*15*	2	*27*	3	*39*	4	*51*	7
4	1	*16*	2	*28*	3	*40*	5	*52*	8
5	1	*17*	2	*29*	3	*41*	5	*53*	8
6	1	*18*	2	*30*	3	*42*	5	*54*	9
7	1	*19*	2	*31*	4	*43*	5	*55*	10
8	2	*20*	3	*32*	4	*44*	5	*56*	11
9	2	*21*	3	*33*	4	*45*	6	*57*	13
10	2	*22*	3	*34*	4	*46*	6		
11	2	*23*	3	*35*	4	*47*	6		
12	2	*24*	3	*36*	4	*48*	6		

Veride toplamda 57 kümenin çapı bulunmaktadır. Bu veriler küçükten büyüğe doğru sıralandığında $(\frac{n+1}{2})$ nolu sıraya karşılık gelen değer bu sıralamanın tam ortasına denk gelen değerdir. Bu durumda $(\frac{57+1}{2})=29$ olduğundan 29. sıraya karşılık gelen **3 pc** değeri bu verinin medyan değeridir.

Bir verideki eleman sayısı çift sayı olduğu zaman tam ortaya denk gelen bir veri olmayacaktır. Örneğin, 6 daireden oluşan bir apartmanda her dairede yaşayan kedi sayıları 4, 1, 3, 0, 8, ve 1 olsun. Böyle bir verinin medyanını bulmak için öncelikle verileri küçükten büyüğe doğru sıralayalım:

0	1	1	3	4	8

Bu değerler için $(\frac{n+1}{2})$ ifadesi kullanılırsa $(\frac{6+1}{2})=3.5$ elde edilir. Bu durumda ortada kalan iki değerin (kırmızı renkte gösterilen 3. ve 4. sıradaki değer) ortalaması alınır. Yani üstteki verinin medyanı $\frac{1+3}{2}=2$ olarak bulunur. Medyan her zaman tamsayı olmak zorunda değildir. Tıpkı ortalama almada olduğu gibi kesirli değerler de alabilir.

Medyan **nicel verilere** uygulanabildiği gibi bazı durumlarda **ordinal (sıralama) verileri**ne de uygulanabilir. Nicel veriler aralık ölçümü veya oranlanabilir ölçüm olabilir ve değerleri kesikli veya sürekli olabilir. Sıralama verileri de nitel kavramlar içeriyor olabilir. Örneğin, "Klasik müzikten hoşlanır mısınız?" sorusunun cevapları: "Hiç sevmem, sevmem, normal bulurum, severim, çok severim" olan bir anketin sonuçları hem nitel hem de bir sıralama verisi olup medyanı alınabilirdir.

3.4 Ortalama, Mod ve Medyan Değerlerindeki Farklılıklar

Açık kümelere ilişkin örnek veri için elde edilen ortalama, mod ve medyan çap değerleri aşağıda özetlenmektedir:

$$\text{Ortalama}(x)=4 \text{ pc}$$

$$\text{Mod}(x)=2 \text{ pc}$$

$$\text{Medyan}(x)=3 \text{ pc}$$

Orta değerleri bu veri için küçükten büyüğe doğru mod, medyan ve ortalama olarak sıralayabiliriz. Görüldüğü üzere elde edilen bu üç orta değer birbirleriyle uyuşmamaktadır. Bunun neden kaynaklandığını anlamak için Çizelge 3.2'de oluşturduğumuz frekans dağılımını bir histogram grafiği üzerinde Şekil 3.1'de gösterelim. Bu şekildeki frekans dağılımını nasıl tanımlarsınız? Verinin dağılımı normal bir dağılım mıdır yoksa pozitif veya negatif çarpık bir dağılım mıdır? Şekil 3.1 dikkatli incelendiğinde büyük değerlerde bazı fazla sayımların olduğu görülmektedir. Bu nedenle bu verinin dağılımının pozitif çarpık olduğu söylenebilir.

Frekans dağılımı normal dağılıma (Gauss/çan eğrisine) uyan bir veride ortalama, mod ve medyan değerleri aynıdır (bkz., Şekil 3.2). Bir verinin normal dağılımdan sapması mod, medyan ve ortalama değerlerinin birbirlerinden uzaklaşmasına neden olur. Orta değerlerin birbirlerinden ne kadar uzak olduğu bir anlamda çarpıklığın bir ölçümüdür.

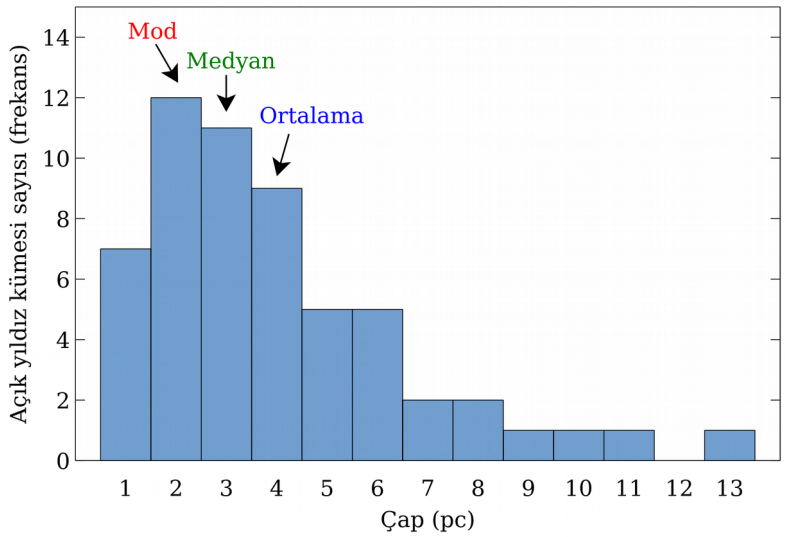
Frekans dağılımı pozitif çarpık olan değerleri için orta değerlerin sıralaması;

Ortalama > Medyan > Mod

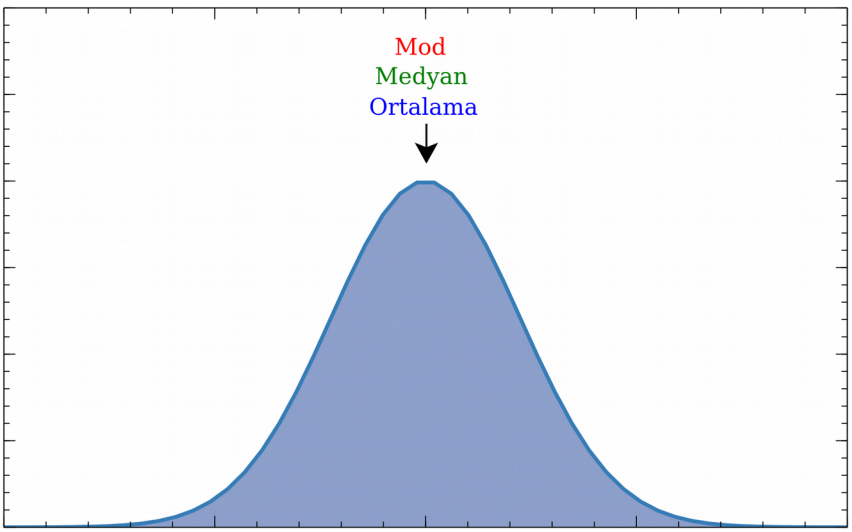
şeklinde iken negatif çarpık bir dağılımda ise

Mod > Medyan > Ortalama

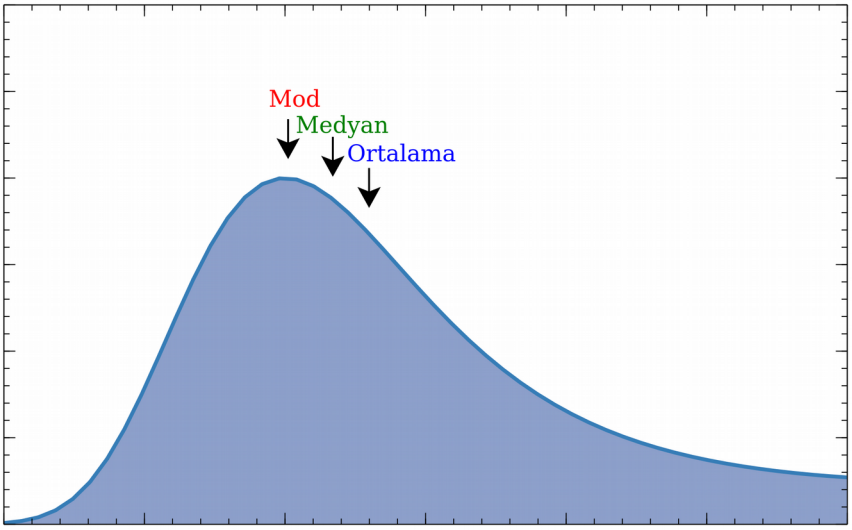
şeklindedir.



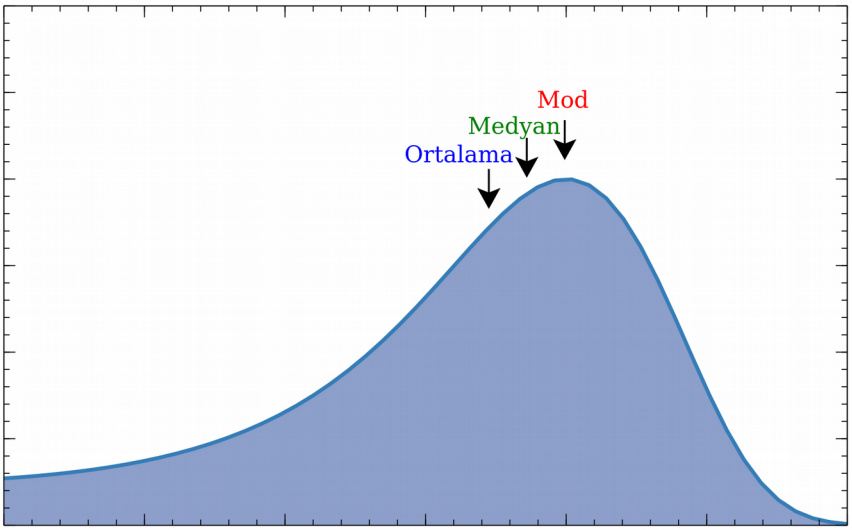
Şekil 3.1 Açık küme çap verilerinin frekans dağılımının histogram grafiği ile gösterimi



Şekil 3.2 Normal dağılım üzerinde mod, medyan ve ortalama



Şekil 3.3 Pozitif çarpık dağılım üzerinde mod, medyan ve ortalama



Şekil 3.4 Negatif çarpık dağılım üzerinde mod, medyan ve ortalama

Ortalama bir verideki bütün değerleri göz önünde bulundurur. Bu nedenle verinin denge noktası olarak düşünülebilir. Ancak veride aykırılık sergileyen değerlerden (varsa) çok etkilenir. Örneğin, açık kümelere ilişkin veride bazı kümelerin 11 ve 13 pc lik büyük çaplara sahip olması ortalama değerinin büyük değerlere kaymasına neden olmuştur. Medyan ise ortalamanın tersine aykırı değerlerden çok daha az etkilenir. Örneğin, açık küme verisinde en büyük çapa sahip olan kümenin çapı 13 pc yerine 40 pc olsaydı ortalama değer oldukça artar; ancak medyan değeri yine aynı kalırdı. Daha önce de belirttiğimiz gibi eğer bir veri normal dağılıma yakın bir dağılım sergiliyorsa ortalama ve medyan değerleri birbirlerine çok yakın olur ve ortalama değer veriyi temsil için kullanılabilir. Ancak dağılımda çarpıklık varsa hem ortalama hem de medyan değerlerinin birlikte sunulması çoğu durumda daha faydalı olacaktır. Mod değeri ise sadece en çok rastlanılan değeri temsil ettiğinden verinin ortasını temsil etmede biraz zayıftır. Ancak, bazı durumlarda (örneğin nominal nitel verilerde) mod değeri kullanılmak zorundadır.