ASTROİSTATİSTİK

6. KONU

Hazırlayan: Doç. Dr. Tolgahan KILIÇOĞLU

6. FARKLI VERİ TÜRLERİNDE ORTA DEĞER, YAYILIM, ÇARPIKLIK VE BASIKLIK

Bir veri üzerinde istatistiksel bir çalışma yaparken verinin bir popülasyona mı yoksa bir örnekleme mi ait olduğunu bilmek önemlidir. İstatistikteki formüller bu iki veri türü için farklılıklar sergiler.

Örneğin, bir bilgisayar şirketinde çalışan 100 kişinin ortalama maaşları hesaplanmak istiyor. Eğer elimizde bu 100 kişinin ne kadar maaş aldığına ilişkin bir veri varsa bu bir **popülasyon** olur. Eğer sadece **rastgele** seçilmiş 10 kişinin maaşının verisi mevcutsa bu veri bir **örneklem** olur. Popülasyonun analizi sonucunda ortalama maaş bir **parametre** (μ) olarak net bir şekilde ortaya koyulur. Örneklem üzerinde ise bir **istatistik** yapılarak ortalama maaş tahmin edilir ($\bar{x}$).

Çözülmek istenen bir soruya göre bir verinin popülasyon veya örneklem olma durumu da değişebilir. Örneğin, bir restoran sahibi şu ana kadar restoranı ziyaret etmiş olan müşterilerinin yaşları üzerine bir istatistik betimleme yapmak istemektedir. Elinde şu ana kadar restoranı ziyaret eden herkesin yaş bilgisi veri olarak bulunmaktadır. Bu durumda bu yaş verisi bir popülasyon oluşturur. Ancak, restoran sahibi bu verileri kullanarak bundan sonra gelecek olan müşterileri hakkında tahminde bulunmak istiyorsa bu veri artık bir örneklem olarak ele alınmalıdır. Çünkü bu ikinci senaryoda popülasyon restoranın sadece geçmişteki değil gelecekteki müşterilerini de kapsamaktadır.

Astronomide belirli gök cisimlerinden toplanan verilerin çok büyük bölümünün bir örneklem oluşturduğunu rahatlıkla söyleyebiliriz. Örneğin, nötron yıldızları üzerine yapılan bir çalışmada evrendeki nötron yıldızlarının tamamını gözlememiz mümkün değildir. Ancak bir örneklem gözleyerek nötron yıldızlarının tamamına ilişkin bir istatistik yapabiliriz. Çok nadir de olsa bazı veriler popülasyon da oluşturabilir. Örneğin Güneş Sistemi'ndeki gezegenlerin yoğunluklarının ortalaması üzerine yapılan bir çalışmada sisteme üye 8 gezegenin de yoğunluk verisi mevcutsa bu veri bir popülasyondur. Bir başka örnek ise, bir yıldız kümesine üye tüm yıldızların uzay hareketleri biliniyorsa ve bu hareketler kullanılarak kümenin uzaydaki ortalama hareketi belirleniyorsa verinin bir popülasyon olduğu söylenebilir. Ancak, yine çoğu durumda bir kümenin tüm üyelerini gözlemek mümkün olmamakta ve belirli sayısa üyeden bir örneklem oluşturularak kümenin uzaydaki hareketi tahmin edilmektedir. Dolayısıyla elimizde gözlemsel bir veri olduğunda onun çoğu zaman bir örneklem olduğunu aklımızdan çıkarmamamız gerekir.

Veriler daha önce de gördüğümüz gibi **sınıflanmış** ve **sınıflanmamış** olmak üzere de ikiye ayrılırlar. Sınıflandırılmış veriler söz konusu olduklarında istatistiksel ifadelere sadece frekans (f) terimi eklenir. İfadelerin vereceği değerlerin popülasyon-örneklem durumunda olduğu gibi farklılaşması söz konusu değildir.

Bu bölümde istatistikte kullanılan ifadelerin verinin popülasyon/örneklem olması ve sınıflandırılmış/sınıflandırılmamış olması durumlarında nasıl değiştiği gösterilmektedir. Burada verilen ifadelerin büyük bir bölümü önceki konuların tekrarı niteliğini taşır.

6.1 Sembollerdeki Farklılıklar

Ortalama, varyans ve standart sapma ifadeleri için kullanılan semboller verinin popülasyon veya örneklem olmasına göre değişir. Çizelge 6.1'de bu semboller gösterilmektedir.

Çizelge 6.1 Popülasyon ve örneklem için ortalama, varyans ve standart sapma sembolleri

	Popülasyon	**Örneklem**
Ortalama	μ	$\bar{x}$
Varyans	σ^2	s^2
Standart Sapma	σ	s
Eleman sayısı	N	n

Bu ifadelere ek olarak sınıflanmış verilerde sınıf sayısını k sembolüyle, sınıftaki eleman sayısını f sembolüyle ve sınıf göstergesini (sınıfın alt sınırı ile üst sınırının ortalamasını) ise $\hat{x}$ ile göstereceğiz. Ayrıca sınıflanmış bir veri için toplam eleman sayısının (N veya n) sınıfların frekanslarının toplamı olduğunu unutmayınız:

$$N=\sum_{i=1}^{k} f_i \quad \text{veya} \quad n=\sum_{i=1}^{k} f_i$$

6.2 Ortalama

Bir popülasyonun ortalaması ile bir örneklemin ortalaması benzer şekilde hesaplanır. Çizelge 6.2'de bu hesaplamalara ilişkin ifadeler verilmektedir.

Çizelge 6.2 Farklı veri tipleri için ortalama ifadeleri

ORTALAMA	**Popülasyon**	**Örneklem**
Sınıflanmamış	$\mu=\dfrac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x}=\dfrac{\sum_{i=1}^{n} x_i}{n}$
Sınıflanmış	$\mu=\dfrac{\sum_{i=1}^{k} f_i \hat{x}_i}{N}$	$\bar{x}=\dfrac{\sum_{i=1}^{k} f_i \hat{x}_i}{n}$

6.3 Mod

Bir popülasyonun modu ile bir örneklemin modu benzer şekilde hesaplanır. Çizelge 6.3'de bu hesaplamalara ilişkin ifadeler verilmektedir.

Çizelge 6.3 Farklı veri tipleri için mod ifadeleri

MOD	Popülasyon	Örneklem
Sınıflanmamış	Verideki en çok tekrarlayan değerdir.	
Sınıflanmış	$\text{MOD} \simeq L_{mod} + c \cdot \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)$	

Sınıflanmış veride en fazla elemanı olan sınıfa **mod sınıfı** denir. Buna göre:

L_{mod} : Mod sınıfının alt sınırı
Δ_1 : Mod sınıfının frekansı ile bir önceki sınıfın frekansı arasındaki fark (pozitif bir değer)
Δ_2 : Mod sınıfının frekansı ile bir sonraki sınıfın frekansı arasındaki fark (pozitif bir değer)
c : Sınıf genişliği

olarak alınmalıdır.

6.4 Medyan

Bir popülasyonun medyanı ile bir örneklemin medyanı benzer şekilde hesaplanır. Çizelge 6.4'te bu hesaplamalara ilişkin ifadeler verilmektedir.

Çizelge 6.4 Farklı veri tipleri için varyans ifadeleri

MEDYAN	Popülasyon	Örneklem
Sınıflanmamış	Veriler küçükten büyüğe (veya tersine) doğru sıralandığında ortada kalan değerdir.	
Sınıflanmış	$\text{MEDYAN} \simeq L_{medyan} + c \cdot \left(\frac{\frac{N}{2} - \sum_{i=1}^{i_{medyan}-1} f_i}{f_{medyan}}\right)$	$\text{MEDYAN} \simeq L_{medyan} + c \cdot \left(\frac{\frac{n}{2} - \sum_{i=1}^{i_{medyan}-1} f_i}{f_{medyan}}\right)$

Sınıflanmış veride kendisinden önce gelen sınıfların frekanslarının toplamı ile kendisinden sonra gelen sınıfların frekansları toplamının birbirlerine en yakın olduğu sınıfa **medyan sınıfı** denir. Buna göre:

L_{medyan} : Medyan sınıfının alt sınırı

$\sum_{i=1}^{i_{medyan}-1} f_i$: Medyan sınıfına kadar olan (medyan sınıfı hariç) sınıfların frekansları toplamı

f_{medyan} : Medyan sınıfının frekansı

c : Sınıf genişliği

olarak alınmalıdır.

6.5 Varyans

Bir popülasyonun varyansı ile bir örneklemin varyansı arasındaki fark örneklem için varyans hesaplanırken paydanın $n-1$ alınmasıdır. Çizelge 6.5'de bu hesaplamalara ilişkin ifadeler verilmektedir.

Çizelge 6.5 Farklı veri tipleri için varyans ifadeleri

VARYANS	Popülasyon	Örneklem
Sınıflanmamış	$\sigma^2=\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$	$s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
Sınıflanmış	$\sigma^2=\frac{\sum_{i=1}^{k} f_i(\hat{x}_i-\mu)^2}{N}$	$s^2=\frac{\sum_{i=1}^{k} f_i(\hat{x}_i-\bar{x})^2}{n-1}$

6.6 Standart Sapma

Bir popülasyonun standart sapması ile bir örneklemin standart sapması arasındaki fark örneklem için standart sapma hesaplanırken paydanın $n-1$ alınmasıdır. Çizelge 6.6'de bu hesaplamalara ilişkin ifadeler verilmektedir.

Çizelge 6.6 Farklı veri tipleri için standart sapma ifadeleri

STANDART SAPMA	Popülasyon	Örneklem
Sınıflanmamış	$\sigma=\sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$	$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$
Sınıflanmış	$\sigma=\sqrt{\frac{\sum_{i=1}^{k} f_i(\hat{x}_i-\mu)^2}{N}}$	$s=\sqrt{\frac{\sum_{i=1}^{k} f_i(\hat{x}_i-\bar{x})^2}{n-1}}$

6.7 Çarpıklık

Çizelge 6.7'de çarpıklığın hesabında kullanılabilecek ifadeler sunulmaktadır. Popülasyon ve örneklem için verilen çarpıklık ifadelerinin ilk etapta birbirlerine çok benzediği düşünülebilir. Ancak paydaya dikkat edildiğinde örneklemde s^3 ifadesi bulunduğu görülür. Burada s örneklemin standart sapması olup $n-1$ terimini içinde barındırmaktadır. Popülasyonun standart sapması (σ) ise N değerinden hesaplanmaktadır. Ayrıca, örneklem için verilen basıklık ifadesinin başında n 'e bağlı bir standartlaştırma ifadesi bulunmaktadır. Bu nedenle popülasyon ve örneklem için hesaplanan çarpıklık değerlerinin birbirlerinden bir miktar farklı olması beklenir.

Çizelge 6.7 Farklı veri tipleri için çarpıklık ifadeleri

ÇARPIKLIK	Popülasyon	Örneklem
Sınıflanmamış	$ç = \dfrac{\sum_{i=1}^{N}(x_i-\mu)^3}{N\sigma^3}$	$ç = \dfrac{n}{(n-1)(n-2)} \cdot \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{s^3}$
Sınıflanmış	$ç = \dfrac{\sum_{i=1}^{k} f_i(\hat{x}_i-\mu)^3}{N\sigma^3}$	$ç = \dfrac{n}{(n-1)(n-2)} \cdot \dfrac{\sum_{i=1}^{k} f_i(\hat{x}_i-\bar{x})^3}{s^3}$

6.8 Basıklık

Çizelge 6.8'de basıklığın hesabında kullanılabilecek ifadeler sunulmaktadır. Popülasyon ve örneklem için verilen basıklık ifadelerinin ilk etapta yine birbirlerine çok benzediği düşünülebilir. Ancak, paydaya dikkat edildiğinde örneklemde s^4 ifadesi bulunduğu görülür. Burada s örneklemin standart sapması olup $n-1$ terimini içinde barındırmaktadır. Popülasyonun standart sapması (σ) ise N değerinden hesaplanmaktadır. Ayrıca, örneklem için verilen basıklık ifadesinin başında tıpkı çarpıklıkta olduğu gibi n 'e bağlı bir standartlaştırma ifadesi bulunmaktadır. Bu nedenle popülasyon ve örneklem için hesaplanan çarpıklık değerlerinin birbirlerinden bir miktar farklı olması beklenir.

Çizelge 6.8 Farklı veri tipleri için basıklık ifadeleri

BASIKLIK	Popülasyon	Örneklem
Sınıflanmamış	$b = \dfrac{\sum_{i=1}^{N}(x_i-\mu)^4}{N\sigma^4} - 3$	$b = \dfrac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{s^4} - 3\dfrac{(n-1)^2}{(n-2)(n-3)}$
Sınıflanmış	$b = \dfrac{\sum_{i=1}^{k} f_i(\hat{x}_i-\mu)^4}{N\sigma^4} - 3$	$b = \dfrac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \dfrac{\sum_{i=1}^{k} f_i(\hat{x}_i-\bar{x})^4}{s^4} - 3\dfrac{(n-1)^2}{(n-2)(n-3)}$

6.9 Bir Sınıflanmış Veri ile Hesaplamalar

Çizelge 6.9'de Gıda Mühendisliği öğrencilerinin Astronomi dersinden aldığı notların sınıflanmış frekans dağılımı yer almaktadır. Öncelikle her sınıfın sınıf göstergesini yanına yazınız. Daha sonra, bu dağılımı kullanarak verinin ortalama, mod, medyan, standart sapma, çarpıklık ve basıklık değerlerini hesaplayınız.

Çizelge 6.9 Astronomi dersi sınav notlarının sınıflanmış frekans dağılımı

Not Aralığı (Sınıflar)	**Sınıf Göstergesi** $(\hat{x}_i)$	**Frekans (öğrenci sayısı)** (f_i)
1 – 10	5.5	1
11 – 20	15.5	1
21 – 30	25.5	1
31 – 40	35.5	3
41 – 50	45.5	4
51 – 60	55.5	9
61 – 70	65.5	10
71 – 80	75.5	14
81 – 90	85.5	8
91 – 100	95.5	1

Öncelikle verinin popülasyon mu yoksa örneklem mi olduğunu tespit edelim. Astronomi dersini alan Gıda Mühendisliği öğrencilerinin hepsinin notları bu veride yer almaktadır (daha doğrusu bir kısmının olduğuna ilişkin soruda bir ibare bulunmamaktadır). Ayrıca bu notlar kullanılarak başka bir durumun tahmini yapılmak istenmemekte, sadece veriye ilişkin bazı ölçütlerin hesaplanması istenmektedir. Bu bilgiler ışığında verinin bir popülasyondan geldiğini rahatlıkla söyleyebiliriz. Verilerin aynı zamanda sınıflanmış olduğu da görülmektedir. Bu durumda hesaplamalarımızda popülasyon ve sınıflanmış veri için olan bağıntıları kullanacağız.

Ortalama

Öncelikle frekanslar toplamından toplam öğrenci sayısını belirleyelim:

$$N=\sum_{i=1}^{k} f_i = 1+1+1+3+4+9+10+14+8+1=52$$

Şimdi popülasyonun ortalamasını hesaplayalım:

$$\mu=\frac{\sum_{i=1}^{k} f_i \hat{x}_i}{N}=\frac{(1\cdot 5.5)+(1\cdot 15.5)+(1\cdot 25.5)+\ldots+(14\cdot 75.5)+(8\cdot 85.5)+(1\cdot 95.5)}{52}=63.9615$$

Mod

Frekansın en fazla olduğu sınıf (yani mod sınıfı) "71 – 80" sınıfıdır. Öncelikle mod hesabı için gereken aşağıdaki değerleri belirleyelim:

L_{mod}	: 71	*Mod sınıfının alt sınırı*
Δ_1	: 14 – 10 = 4	*Mod sınıfının frekansı ile bir önceki sınıfın frekansı arasındaki fark*
Δ_2	: 14 – 8 = 6	*Mod sınıfının frekansı ile bir sonraki sınıfın frekansı arasındaki fark*
c	: 10	*Sınıf genişliği*

Bu değerler kullanılırsa mod değeri

$$\text{MOD} \simeq L_{mod} + c \cdot \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) = 71 + 10 \cdot \left(\frac{4}{4+6}\right) = 75$$

olarak elde edilir. Buradaki mod değerinin bir yaklaşık değer olduğu unutulmamalıdır.

Medyan

Kendisinden önce gelen sınıfların frekanslarının toplamı ile kendisinden sonra gelenlerinkinin toplamının birbirlerine en yakın olduğu sınıf (yani medyan sınıfı) "61 – 70" sınıfıdır. Buna göre;

L_{medyan}	: 61	*Medyan sınıfının alt sınırı*
$\sum_{i=1}^{i_{medyan}-1} f_i$	: 1+1+1+3+4+9 =19	*Medyan sınıfına kadar olan sınıfların frekansları toplamı*
f_{medyan}	: 10	*Medyan sınıfının frekansı*
c	: 10	*Sınıf genişliği*

Bu değerler kullanılırsa medyan değeri

$$\text{MEDYAN} \simeq L_{medyan} + c \cdot \left(\frac{\frac{N}{2} - \sum_{i=1}^{i_{medyan}-1} f_i}{f_{medyan}}\right) = 61 + 10 \cdot \left(\frac{\frac{52}{2} - 19}{10}\right) = 68$$

olarak elde edilir. Burada medyan değeri yine yaklaşık bir medyan değeridir.

Standart Sapma

Standart sapma hesaplanırken toplam işaretinin olduğu ifadede parantez içinde kalan fark $(\hat{x}_i - \mu)^2$ değerlerinin önceden hesaplanması işlemlerin daha kolay yapılmasını sağlar.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i (\hat{x}_i - \mu)^2}{N}} = \sqrt{\frac{1 \cdot (5.5 - 63.9615)^2 + \ldots + 9 \cdot (55.5 - 63.9615)^2 + \ldots + 1 \cdot (95.5 - 63.9615)^2}{52}} = 18.74778$$

Çarpıklık

Çarpıklık hesaplanırken toplam işaretinin olduğu ifadede parantez içinde kalan fark $(\hat{x}_i-\mu)^3$ değerlerinin önceden hesaplanması işlemlerin daha kolay yapılmasını sağlar.

$$ç=\frac{\sum_{i=1}^{k} f_i(\hat{x}_i-\mu)^3}{N\,\sigma^3}=\frac{1\cdot(5.5-63.9615)^3+\ldots+9\cdot(55.5-63.9615)^3+\ldots+1\cdot(95.5-63.9615)^3}{52\cdot(18.74778)^3}=-0.98$$

Çarpıklık değeri verinin negatif (sola) çarpık olduğunu göstermektedir.

Basıklık

Basıklık hesaplanırken toplam işaretinin olduğu ifadede parantez içinde kalan fark $(\hat{x}_i-\mu)^4$ değerlerinin yine önceden hesaplanması işlemlerin daha kolay yapılmasını sağlayacaktır.

$$b=\frac{\sum_{i=1}^{k} f_i(\hat{x}_i-\mu)^4}{N\,\sigma^4}-3=\frac{1\cdot(5.5-63.9615)^4+\ldots+9\cdot(55.5-63.9615)^4+\ldots+1\cdot(95.5-63.9615)^4}{52\cdot(18.74778)^4}-3=0.86$$

Basıklık değerinin sıfırdan büyük olması dağılımın normal dağılıma nazaran daha sivri olduğunu gösterir.

Son olarak Çizelge 6.10'da hesapladığımız değerler ile veri sınıflandırılmadan önceki ham verilerle elde edilen gerçek değerlerin bir karşılaştırılması verilmiştir.

Çizelge 6.10 Sınıflanmış verilerden hesaplanan ölçütler ile sınıflanmamış verilerden hesaplananların karşılaştırılması

	Sınıflanmış veriden hesaplanan	**Gerçek değer**	**Mutlak Fark**
Ortalama	63.9615	65.1346	~1.2
Mod	75	80	5
Medyan	68	67	1
Standart Sapma	18.74778	19.93136	~1.2
Çarpıklık	–0.98	-0.95	0.03
Basıklık	0.86	1.15	0.29

Sınıflanmış veriden elde ettiğimiz ölçütler sınıflanmamış veriyi oldukça başarılı şekilde temsil etmektedir. Çizelge 6.9'daki sınıflanmış verileri bir histogram grafiğinde göstererek bulduğumuz bu parametreleri gözle denetleyiniz.