ASTROİSTATİSTİK

2. KONU

Hazırlayan: Doç. Dr. Tolgahan KILIÇOĞLU

2. VERİLERİN SINIFLANMASI VE DAĞILIMLARININ SUNULMASI

İstatistiktikte öncelikle çalışılacak verinin tanımlanmasına ihtiyaç duyulur. Veri tamamen rastgele midir yoksa belirli bir dağılım sergilemekte midir? Bu dağılımın şekli nasıldır? Bu noktada verinin gerekiyorsa gruplanması ve frekans dağılımının yapılması gerekir. Frekans dağılımı yapıldığında, örneğin, verinin iyi bilinen çan şekilli bir dağılım sergileyip sergilemediği ortaya konabilir. Tanımlamaların daha kolay yapılabilmeleri için tablolar ve grafikler kullanılır.

2.1 Sınıflanmamış Veri ve Sınıflanmış Veri

Çizelge 2.1'de Astronomi ve Uzay Bilimleri Bölümü erkek öğrencilerinin kiloları verilmektedir. Bu çizelgeye bakıldığında verilerin oldukça dağınık olduğu görülür. Böyle bir veriyi yorumlamak gerçekten zordur. İlk bakışta yetmişli rakamların biraz ağır bastığı görülse de net birşey söylemek ilk bakışta mümkün değildir.

Çizelge 2.1 Astronomi ve Uzay Bilimleri Bölümü'nde okuyan erkek öğrencilerin kiloları

73	76	69	73	73	68	87	72	70
89	77	71	100	78	102	71	81	75
103	77	61	87	77	66	62	87	71
71	73	112	78	94	70	85	75	72
82	75	78	73	69	79	73	70	62
93	79	70	82	69	75	79	71	

Çizelge 2.2 Her kiloda kaç kişi var?

Kilo	Frek.	Kilo	Frek.	Kilo	Frek.	Kilo	Frek.	Kilo	Frek.	Kilo	Frek.	Kilo	Frek.
61	1	69	3	77	3	85	1	93	1	101	0	109	0
62	2	70	4	78	3	86	0	94	1	102	1	110	0
63	0	71	5	79	3	87	3	95	0	103	1	111	0
64	0	72	2	80	0	88	0	96	0	104	0	112	1
65	0	73	6	81	1	89	1	97	0	105	0		
66	1	74	0	82	2	90	0	98	0	106	0		
67	0	75	4	83	0	91	0	99	0	107	0		
68	1	76	1	84	0	92	0	100	1	108	0	**Top.**	53

Bu veriyi daha iyi anlamlandırmak adına her kiloda kaç kişinin olduğunu (frekans dağılımı) yazabiliriz (Çizelge 2.2). **Sınıflanmamış veri**den oluşan bu yeni dağılımda verinin yapısı biraz daha net gözükmektedir. Çizelgede en çok rastlanılan kilo 73 kg dir. Bununla beraber 77 ile 79 kg arasında da oldukça öğrencinin olduğu görülmektedir. Ayrıca kiloların büyük bir bölümünün 69 ile 79 kg arasında olduğu görülmektedir. Kiloları tek tek ele almak işlerimizi kolaylaştırsa da aralığımız 1 kg olduğu için veriler hala çok dağınık gözükmektedir. Bunun nedeni 61 kg'dan 112 kg'a kadar 56 farklı değerin bulunmasıdır. Eğer verideki değişkenin alabileceği değerlerin sayısı 20'nin üzerindeyse Çizelge 2.2'ye benzer bir dağılım verinin anlaşılması adına çok faydalı değildir. Bu verilerin öncelikle **sınıflanması** uygun olacaktır. Eğer daha geniş kilo aralıkları için (örn., 65 ile 70 kg arasında olanlar gibi) bu dağılımı yaparsak veriler daha da anlamlı hale gelir. Eğer değişkenin alabileceği değerler 20'den azsa (örneğin, bir anketteki "Astroistatistik dersini faydalı buluyororum." ifadesine Çok katılıyorum / katılıyorum / emin değilim / katılmıyorum / hiç katılmıyorum şeklinde 5 cevap verilebiliyorsa değişken sadece 5 değer alabilir) veri sınıflamasına gerek yoktur.

2.2 Verilerin Sınıflanması ve Frekans Dağılımı

Bir veri sınıflara ayrıldığında, her sınıfta kaç adet birim olduğunun tespit ve ifade edilmesine **frekans dağılımı** denir. Peki frekans dağılımı için veriler kaç sınıfa bölünmelidir? Sınıfların genişlikleri nasıl olmalıdır? Frekans dağılımı elde edildikten sonra daha iyi nasıl betimlenebilir? İşte şimdi bu sorulara adım adım cevap vereceğiz.

2.2.1 Sınıf sayısının belirlenmesi

Bir verinin kaç adet sınıfa bölüneceğine karar verilmesi gerekir. Gereğinden fazla sınıfa bölünmüş olan bir veri çok fazla bilgi içerir ve zor yorumlanır. Veri gerektiğinden az sınıfa bölünürse de detaylar kaybolacağından önemli bir bilginin gözden kaçırılmasına neden olabilir. Verinin kaç sınıfa bölüneceğine ilişkin kesin bir kural yoktur. Bu araştırmacının ne kadar detay istediğine bağlıdır. Eğer veriyi kaç sınıfa böleceğiniz hakkında hiçbir fikre sahip değilseniz 10 civarında sınıfa bölerek işe başlayabilirsiniz. Sınıf sayısını belirlemede bir diğer yaklaşım ise n veri sayısı olmak üzere $\sqrt{n}+1$ ifadesinin verdiği sayının tam değerini almaktır. Örneğin Çizelge 2.1'de 53 kişinin kilo verisi bulunmaktadır. Bu durumda $\sqrt{53}+1\approx 8$ sınıf kullanılabilir. Ancak burada verilen değerler sadece öneridir. Biz yukarıdaki kilo çizelgesi için 11 adet sınıf kullanacağız.

2.2.2 Sınıfların genişliğinin belirlenmesi

Sınıf sayısı belirlendikten sonra sınıf genişliği aşağıdaki ifade ile belirlenir:

$$[\text{Sınıf Genişliği}]=\frac{[\text{En Büyük Değer}]-[\text{En Küçük Değer}]}{[\text{Sınıf Sayısı}]}$$

Örnek olarak Çizelge 2.1'deki veriler kullanılırsa;

$$\text{SG}=\frac{112-61}{11}=4.63$$

elde edilir. Kolaylık olması açısından sınıf genişliğini bu veri için yaklaşık 5 olarak alalım.

2.2.3 Sınıfların sınırlarının belirlenmesi

Sınıfların sayısı ve genişliği belirlendiğine göre en küçük değerden başlanılarak ve sınıf genişliği kadar atlanılarak sınıfların sınırları belirlenebilir. Örneğe geri dönersek en küçük değeri 61 olan ve sınıf genişliği 5 olan 11 sınıf aşağıdaki şekilde oluşturulabilir:

61 – 65
66 – 70
71 – 75
76 – 80
81 – 85
86 – 90
91 – 95
96 – 100
101 – 105
106 – 110
111 – 115

Çizelge 2.1'de en büyük kilo değeri 112 olmasına karşın yaptığımız sınıflama 115'e kadar gitmektedir. Bunun nedeni 4.63 olan sınıf genişliğini 5'e yuvarlamamızdır. Bu durumun oluşması sakıncalı bir durum değildir.

Sınıfların sınırları belirlenirken iki basit ama önemli kurala her zaman dikkat edilmelidir:

i) Sınıflar bütün değerleri kapsayacak şekilde seçilmeli ve aralarında boşluklar olmamalı (bir değer mutlaka bir sınıfa yazılabilmeli),

ii) Bir sınıfın aralığı bir diğerinin aralığının bir bölümü ile çakışmamalıdır (bir değer sadece tek bir sınıfa yazılabilmelidir).

Artık sınıfları da belirlediğimize göre frekans dağılımını yapabiliriz.

2.2.4 Frekans dağılımının yapılması

Sınıflar belirlendikten artık yapılması gereken tek şey her sınıfa kaç üyenin düştüğünü belirlemektir. Böylece frekans dağılımı çizelgesi oluşturulmuş olur. Çizelge 2.1'deki verilerin sınıflandırılmış frekans dağılımı Çizelge 2.3'de verilmektedir.

Çizelge 2.3 incelendiğinde artık ilgili öğrencilerin büyük bir bölümünün kilolarının 71 ile 75 kg arasında olduğu rahatlıkla söylenebilir. Ardından 76 – 80 kg ve 66 – 70 kg sınıfları gelir. Diğer kilolara sahip öğrenci sayısı oldukça azdır. Çizelgenin üçüncü ve dördüncü kolonunda birikimli frekans ve oransal frekans olarak adlandırılan iki ölçüt daha mevcuttur.

Sınıf göstergesi sınıfın üst ve alt sınırının ortalamasıdır. Birikimli frekans bulunduğu sınıf ve bu sınıfın öncesinde bulunan sınıfların frekanslarının toplamıdır. Oransal frekans ise bir sınıfın frekansının toplam frekansa bölümü ile elde edilir.

Birikimli frekans: $bf_i = \sum_{j=0}^{i} f_j$ Oransal Frekans: $of_i = \frac{f_i}{n}$ ve $n = \sum_{i=0}^{k} f_i$

Burada k sınıf miktarı n ise toplam gözlem sayısıdır. Oransal frekans verinin yorumlanmasında çok kullanışlıdır. Çünkü oransal frekans 100 ile çarpıldığında ilgili sınıfa giren birim sayısının toplam birim sayısına oranını verir. Örneğin, kiloları 71 ile 75 aralığında olan erkek öğrenciler bölümdeki tüm erkek öğrencilerin %32'sini (0.321×100) oluşturmaktadır. Oransal frekansın tüm gruplar için toplamının 1 olması gerekir. Ancak Çizelge 2.3'te bu toplam 1.001 olarak çıkmıştır. Bunun temel nedeni oransal frekanslar yuvarlanırken hataların birikmesidir. Noktadan sonra yeteri kadar hane bulunduğundan bu durumun pratikte bir zararı bulunmamaktadır.

Çizelge 2.3 Kilo verilerinin sınıflandırılmış frekans dağılımı

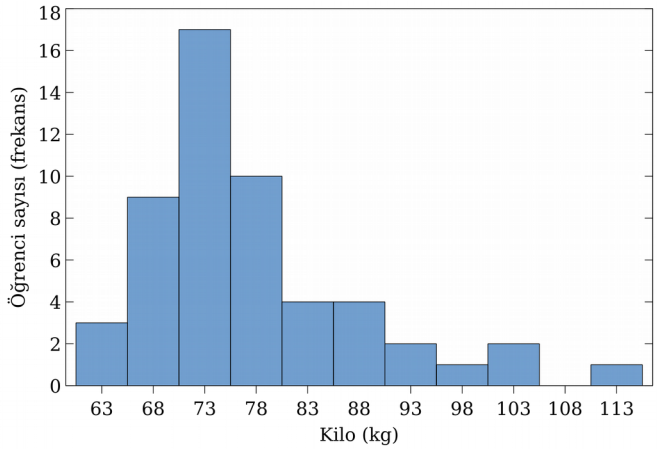
Sınıf	Sınıf Göstergesi	Frekans (f_i)	Birikimli frekans (bf_i)	Oransal frekans (of_i)
61 – 65	63	3	3	0.057
66 – 70	68	9	12	0.170
71 – 75	73	17	29	0.321
76 – 80	78	10	39	0.189
81 – 85	83	4	43	0.075
86 – 90	88	4	47	0.075
91 – 95	93	2	49	0.038
96 – 100	98	1	50	0.019
101 – 105	103	2	52	0.038
106 – 110	108	0	52	0.000
111 – 115	113	1	53	0.019
	Toplam:	**53**		**1.001**

2.3 Verilerin Grafikle Gösterilmesi

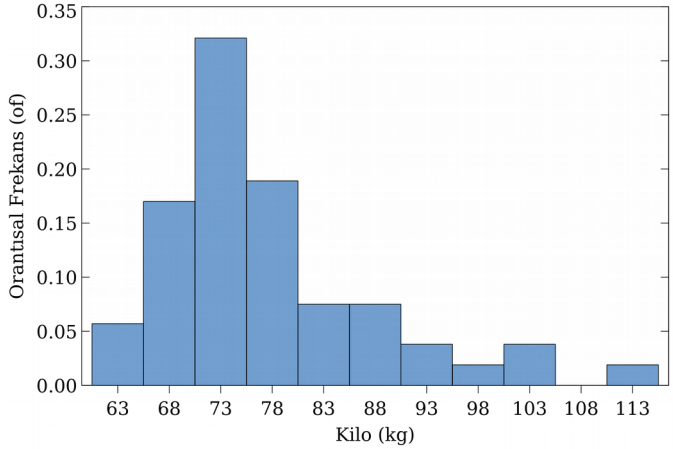
Bir frekans dağılımını anlamanın veya ifade etmenin en güzel yollarından biri onu bir grafiğe aktarmaktır. Bunun için kullanılabilecek birçok farklı grafik türü bulunmaktadır. Ancak burada sadece sık kullanılan birkaç türden söz edeceğiz: Histogram, poligon, pasta dilimi.

2.3.1 Histogram Grafiği

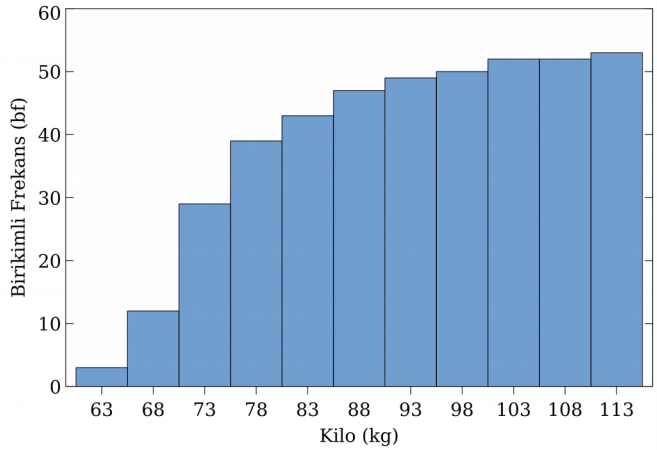
Histogram grafiği oluşturmak için öncelikle her sınıfı temsil eden bir ortalama değer belirlenir. Örneğin, 61 – 65 kg aralığını içeren sınıfı temsil eden ortalama değer (61+65)/2, yani 63'tür. Diğer sınıflar için de benzer şekilde ortalama değerler hesaplanır. Elde edilen bu değerler grafiğin x eksenine yerleştirilir. Y-eksenine ise frekans (Şekil 2.1), oransal frekans (Şekil 2.2) veya birikimli frekans (Şekil 2.3) yerleştirilerek histogram grafiği oluşturulur.



Şekil 2.1 Frekans dağılımının histogram grafiği ile gösterimi



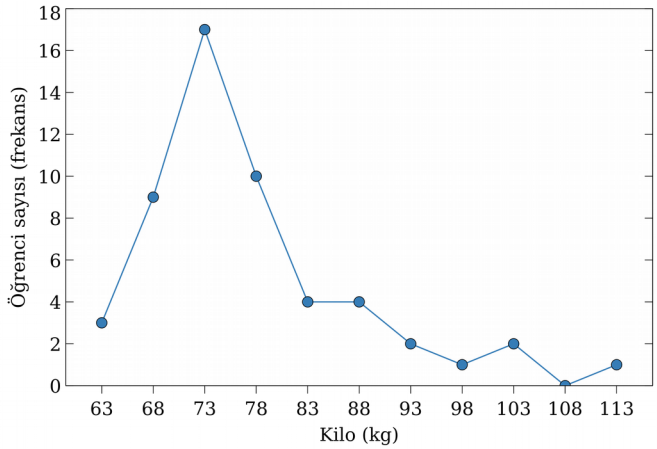
Şekil 2.2 Orantısal frekans kullanılarak oluşturulan histogram grafiği



Şekil 2.3 Birikimli frekans kullanılarak oluşturulan histogram grafiği

2.3.2 Poligon Grafiği

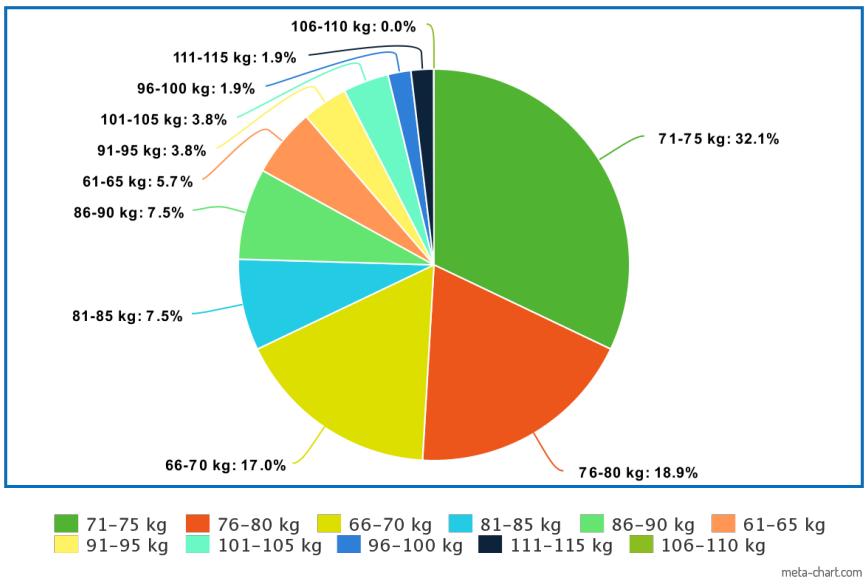
Poligon grafiği de histogram grafiğine çok benzer. Ancak histogram grafiğindeki çubuklar yerine her veri bir nokta ile gösterilir ve noktalar çizgilerle birleştirilir (bkz., Şekil 2.4).



Şekil 2.4 Frekans dağılımının poligon grafiği ile gösterimi

2.3.3 Pasta Dilimi Grafiği

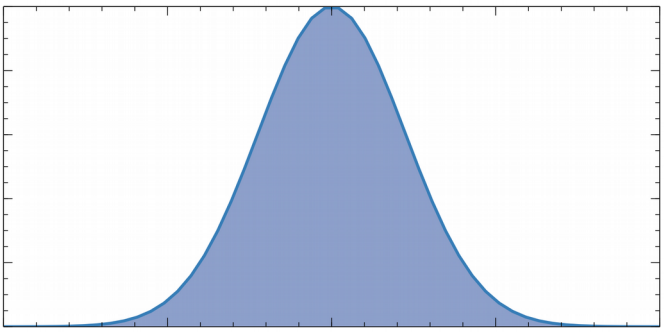
Pasta dilimi grafiği de oldukça yaygın kullanılır. Şekil 2.5'te üstte frekans dağılımını verdiğimiz veri setinin pasta dilimi grafiği gösterilmektedir. Bu gösterimde baskın olan ve zayıf olan sınıflar hemen kendini göstermektedir. Ayrıca bu şekle bakıldığında en fazla öğrenci içeren iki sınıfın toplamının toplam öğrenci sayısının yarısından bir miktar fazlasını kapladığı görülmektedir. Pasta dilimi grafikleri genellikle az sayıda sınıf olduğu durumlarda frekanslar arasındaki farkların daha belirgin şekilde ifade edilmesi için kullanılır. Örneğin bir seçimin sonucunda partilerin dağılımları genellikle bu grafik türü ile verilmektedir.



Şekil 2.5 Frekans dağılımının pasta dilimi grafiği ile gösterimi

2.4 Yaygın Dağılım Türleri

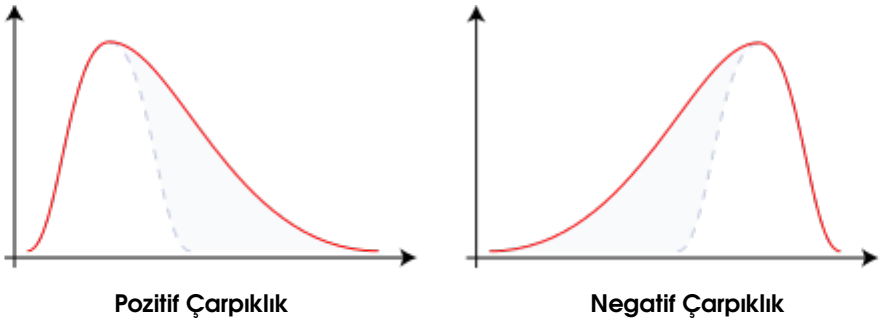
Bir verinin frekans dağılımı yapıldığında dağılım farklı şekillere sahip olabilir. Veri hangi türde olursa olsun en olası durum verideki değişkenin belirli bir değer komşuluğunda daha çok olması ve bu değerden uzaklaştıkça sayının gittikçe azalmasıdır. Bu dağılım **normal dağılım** olarak bilinir ve matematiksel olarak Gauss eğrisi ile gösterilir (Şekil 2.6).



Şekil 2.6 Normal Dağılım

Normal dağılım günlük hayatta birçok durumda karşımıza çıkabilecek doğal bir dağılımdır. Örneğin, bir atış poligonunda hedefin üzerindeki mermi izlerinin dağılımı, üniversiteye giriş sınavında öğrencilerin notları, elmaların boyutları, insanların boyları veya kiloları (erkek/kadın ayrı ele alınmak üzere), IQ testi sonuçları, bir masanın boyutlarının farklı insanlar tarafından ölçümü, bir gökdelenin uzunluğunun insanlar tarafından tahmini, farklı kişiler tarafından kullanılan aynı marka cep telefonunun ilk kaç gün sonra bozulduğu, üzerinde 300 gr yazan çikolata paketlerinin ağırlığı, hayatınızda geçirdiğiniz günleri 1'den (berbattan) 10'a (müthişe) doğru numaralandırdığınızda oluşan dağılım vs...

Bir veri normal dağılım sergilediğinde oldukça kolay incelenebilmektedir. Bunun nedeni normal dağılımın yapısının iyi bilinmesi yani bir formülünün olmasıdır. Ancak bazen veriler normal dağılımdan saparak pozitif veya negatif çarpıklık sergileyebilirler (Şekil 2.6).

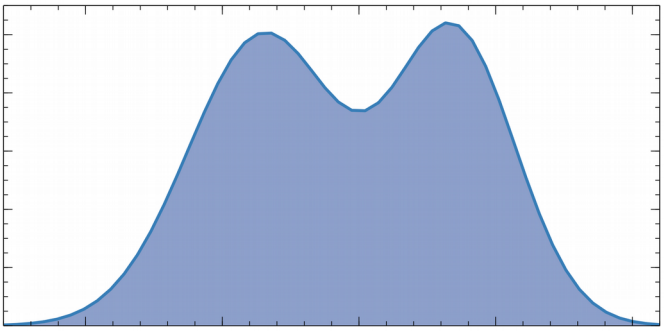


Şekil 2.6 Pozitif ve negatif çarpıklığa sahip dağılımlar

Örneğin bir şehir için *x* ekseninde ev kiralarının miktarlarının olduğu *y* ekseninde evleri kiralayanların sayısının olduğu bir dağılım düşünelim. Eğer zengin insanların çoğunlukta

olduğu bir şehirde bulunmuyorsak bu dağılımın bir miktar negatif çarpık olduğu görülecektir. Negatif çarpıklık, normal dağılımın düşük değerlerinde fazladan sayımların olması anlamına gelir. Pozitif çarpıklık için tersi geçerlidir.

Dağılım türlerinin başka bir yaygın türü ise bimodal (iki modlu) dağılımdır (Şekil 2.7). Hem düz lise mezunlarının hem de bir üniversitenin matematik bölümü mezunlarının birlikte girdiği bir sınav olduğunu düşünün. Böyle bir sınavda üniversite mezunlarının aldığı notların dağıllımı lise mezunlarına göre daha pozitif yöndedir. Ancak her iki grup da birlikte sınava girdiğinden notların dağılımının bimodal olması muhtemeldir. Bir dağılımda birden fazla modun olması genellikle verinin içerisinde birbirlerinden farklı birden fazla grubun olduğunu işaret eder.



Şekil 2.7 Bimodal dağılım

Bunun dışında veri burada sözünü etmediğimiz farklı dağılımlara da sahip olabilir. Genel olarak konuşmak gerekirse bir dağılımın normal dağılımından sapması veride sıradışı bazı durumların olduğunu, veri sayısının yeterli olmadığını veya verinin toplanmasında önyargı yaratan bir etmenin varlığını işaret edebilir. Ancak nadir de olsa bazı verilerin dağılımı doğası gereği normal dağılım sergilemez.