

OLASILIK VE İSTATİSTİĞE GİRİŞ

Mühendisler ve Fenciler İçin

Introduction to
Probability and Statistics for Engineers and Scientists

4. Basımdan Çeviri

Çeviri Editörleri: Prof. Dr. Salih Çelebioğlu - Prof. Dr. Reşat Kasap

Sheldon M. ROSS



BETİMLEYİCİ İSTATİSTİK



VERİ KÜMELERİNİ BETİMLEME

Bir amaç için derlenen verilerin tamamının olduğu, veri kümelerindeki birimlerin sayısal değerlerinden faydalananarak açık ve net bir şekilde ilgilenilen özellik hakkında bilgi edinilip, bu çalışmaların tablo, grafik ve merkezi eğilim ölçüleri kullanılarak sunulmasına veri kümelerinin betimlenmesi denir.

VERİ KÜMELERİNİ BETİMLEME

Frekans tabloları ve grafikler

Nispeten az sayıda farklı değere sahip bir veri kümelerinin betimlenmesinde *frekans tablosu* kullanılabilir. Frekans tabloları verilerin aldığı değerlerin karşısına her bir verinin frekansının (tekrar sayısının) yazılması ile oluşturulur.

TABLE 2.1 *Starting Yearly Salaries*

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

VERİ KÜMELERİNİ BETİMLEME

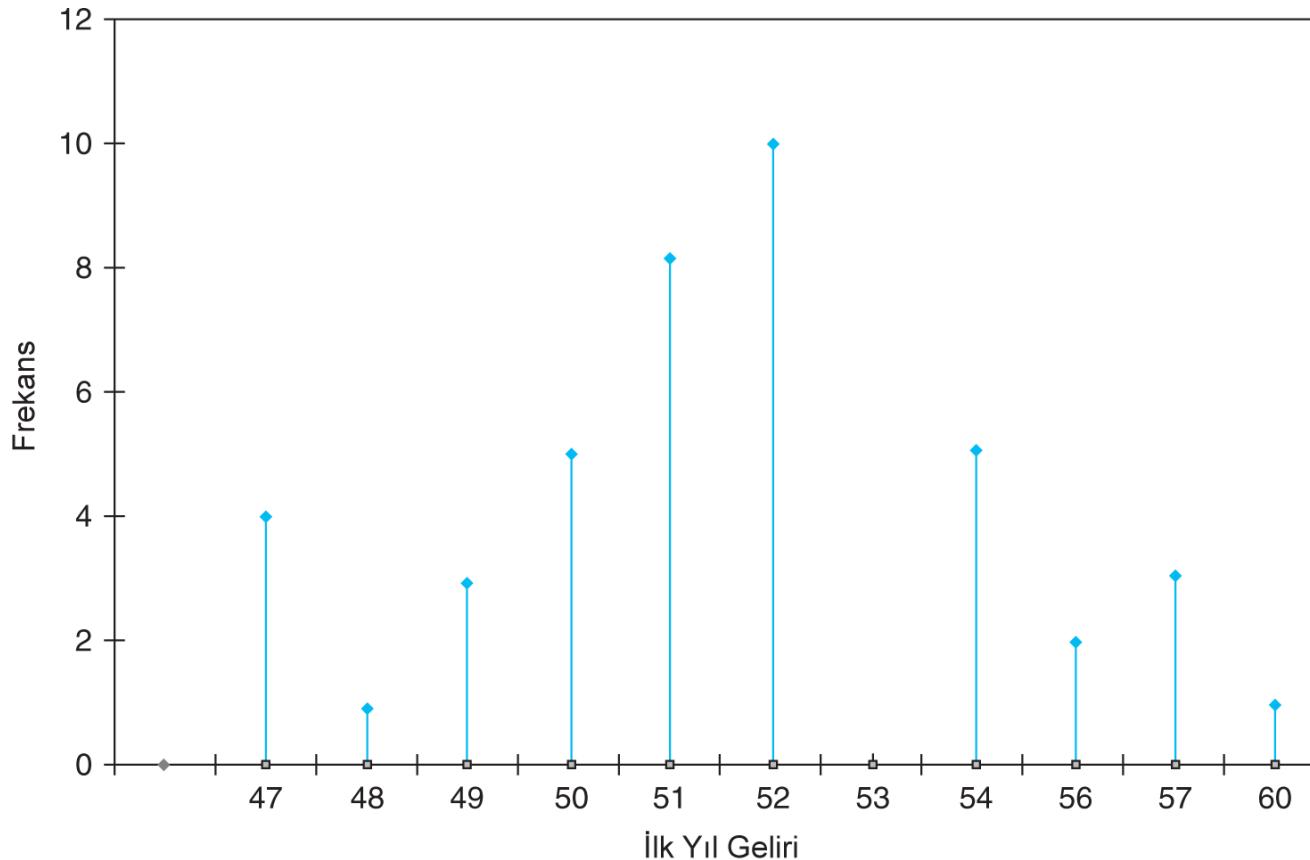
Frekans Tabloları ve Grafikler

Frekans tablosundaki yer alan farklı veri değerlerini yatay eksende yazarak her bir değişkeninin frekansını düşey eksende yüksekliklerle gösterdiğimiz grafiklere çizgi grafiği denir.

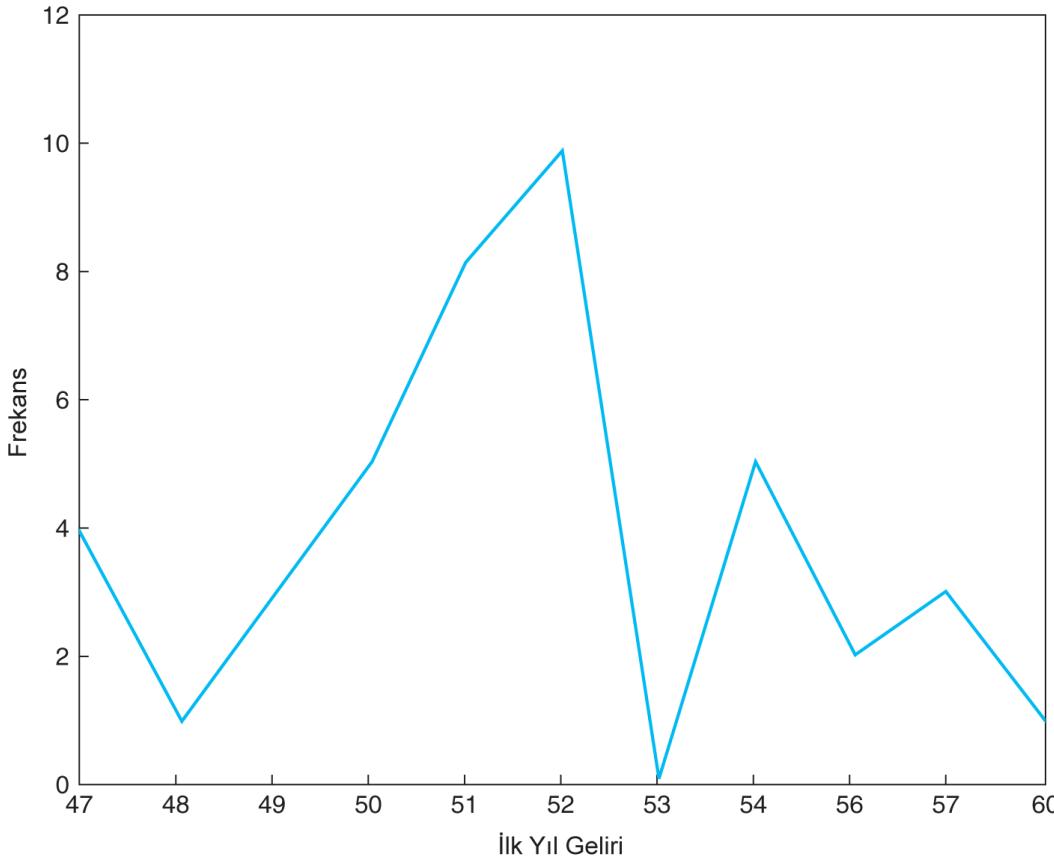
Frekans tablosundaki verileri temsil eden bir başka grafik türü ise frekans çokgenleridir. Frekans çokgenleri her bir değişkenin değerinin karşılık geldiği noktaların düz çizgiler ile birleştirilmesi ile bulunur.

Olasılık ve İstatistik Giriş

İlk yıl geliri verileri

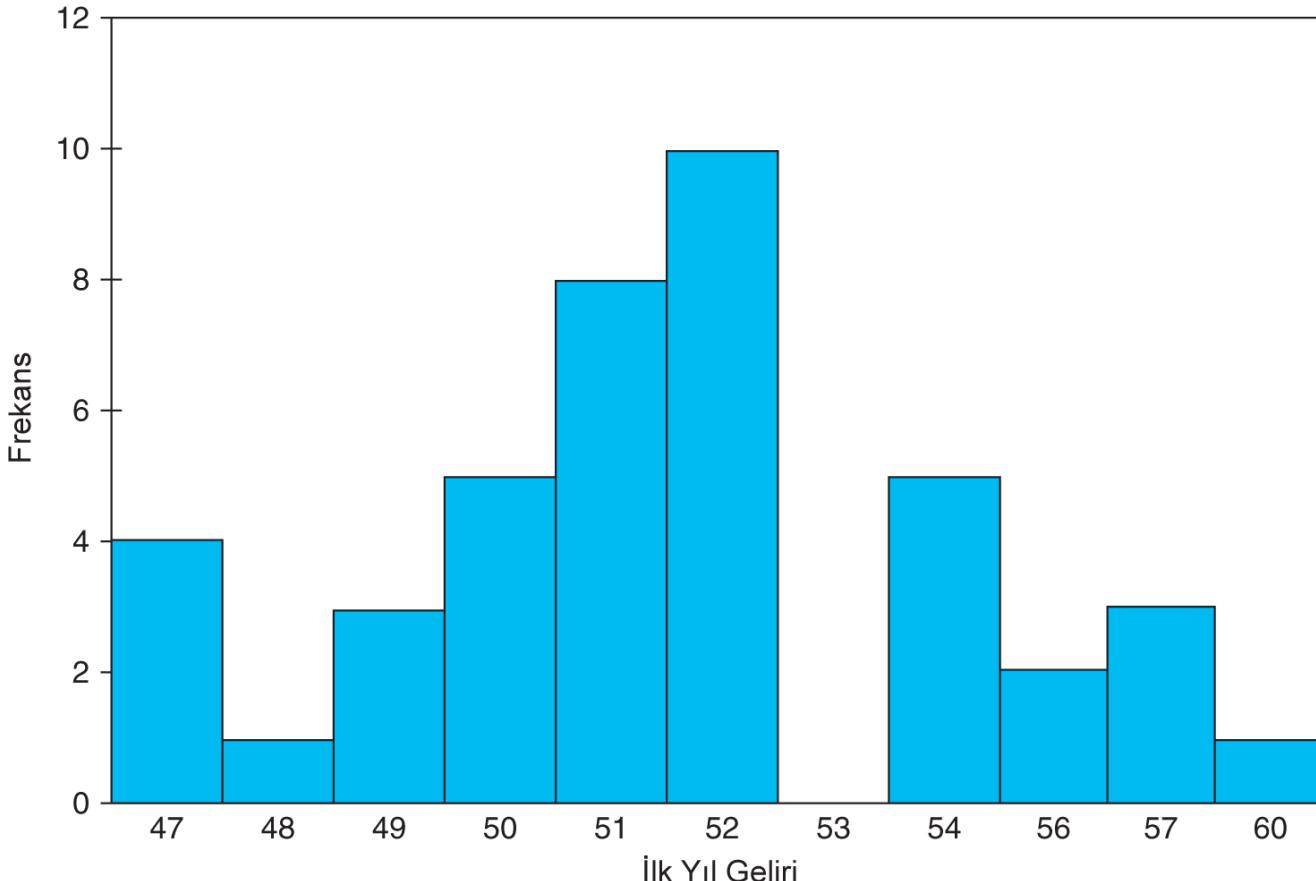


İlk yıl geliri için frekans çokgeni



Olasılık ve İstatistik Giriş

İlk yıl geliri için çubuk grafiği



İstatistik kütlesi n olan bir veri kümesini alalım. f , özel bir değişkenin frekansı olmak üzere, f/n oranına bu değişkenin *göreli frekansı* denir. Bir değişkenin göreli frekansı bu değişkene sahip verilerin oranını verir.

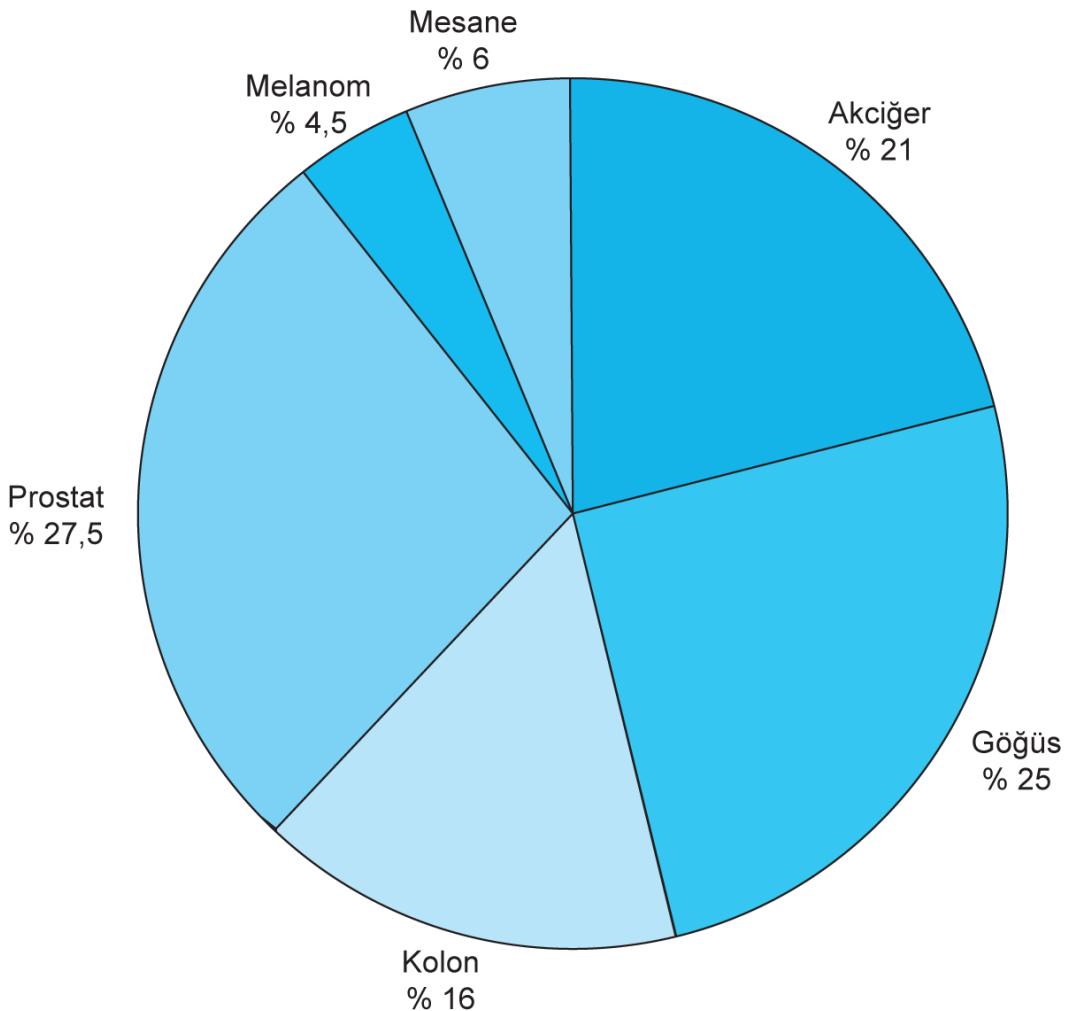
Veriler sayısal yapıda olmadığında göreli frekanslar kullanılır. Göreli frekanslar genellikle *pasta grafikleri* ile gösterilirler.

İlk Yıl Gelirleri

İlk Yıl Geliri	Frekans
47	$4/42 = 0,0952$
48	$1/42 = 0,0238$
49	$3 / 42$
50	$5 / 42$
51	$8 / 42$
52	$10 / 42$
53	0
54	$5 / 42$
56	$2 / 42$
57	$3 / 42$
60	$1 / 42$

Kanser Türü	Yeni Vakaların Sayısı	Göreli Frekans
Akciger	42	0,210
Göğüs	50	0,250
Kolon	32	0,160
Prostat	55	0,275
Melanom	9	0,045
Mesane	12	0,060

Olasılık ve İstatistik Giriş



Gruplanmış Veriler, Histogramlar, Diyagramlar ve Kök ve Yaprak Grafikleri

Veri kümesi içindeki birimlerin çok fazla olduğu veya nispeten çok farklı değer sayısı aldığı durumlarda değişkenleri gruptara veya sınıf aralıklarına bölünerek her bir sınıfın frekans değerinin karşısına yarılmasıyla elde edilen tabloya *sınıflandırılmış frekans tablosu* denir. Böyle verilere ise sınıflandırılmış veriler denir.

Bir sınıfın uç noktalarına *sınıfın sınırları* denir. Veri sınıflarında sınıfın sol ucu sınıfa dahil edilirken sağ ucu dahil edilmez ve genellikle sınıf aralıkları zorunlu olmamakla birlikte eşit olarak alınır.

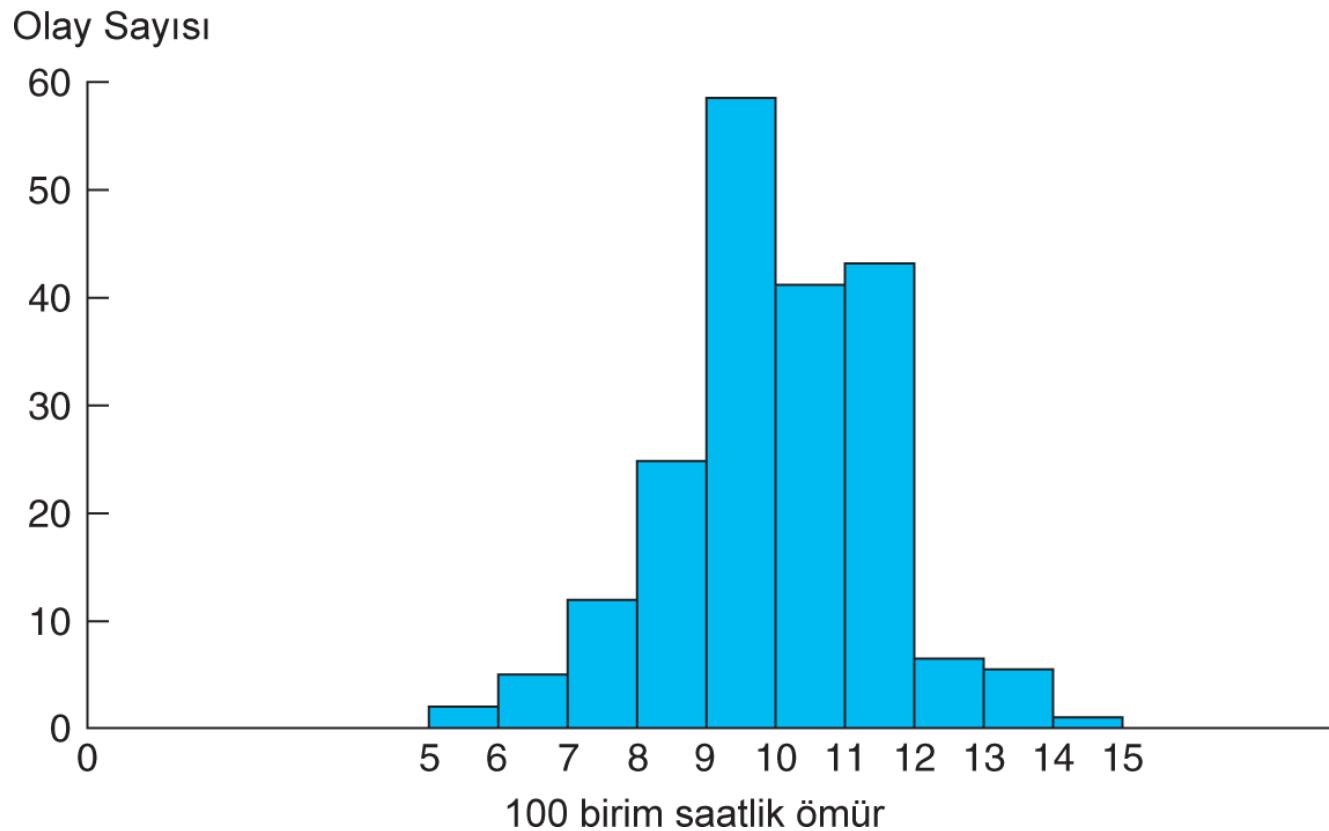
TABLE 2.3 *Life in Hours of 200 Incandescent Lamps*

Item Lifetimes									
1,067	919	1,196	785	1,126	936	918	1,156	920	948
855	1,092	1,162	1,170	929	950	905	972	1,035	1,045
1,157	1,195	1,195	1,340	1,122	938	970	1,237	956	1,102
1,022	978	832	1,009	1,157	1,151	1,009	765	958	902
923	1,333	811	1,217	1,085	896	958	1,311	1,037	702
521	933	928	1,153	946	858	1,071	1,069	830	1,063
930	807	954	1,063	1,002	909	1,077	1,021	1,062	1,157
999	932	1,035	944	1,049	940	1,122	1,115	833	1,320
901	1,324	818	1,250	1,203	1,078	890	1,303	1,011	1,102
996	780	900	1,106	704	621	854	1,178	1,138	951
1,187	1,067	1,118	1,037	958	760	1,101	949	992	966
824	653	980	935	878	934	910	1,058	730	980
844	814	1,103	1,000	788	1,143	935	1,069	1,170	1,067
1,037	1,151	863	990	1,035	1,112	931	970	932	904
1,026	1,147	883	867	990	1,258	1,192	922	1,150	1,091
1,039	1,083	1,040	1,289	699	1,083	880	1,029	658	912
1,023	984	856	924	801	1,122	1,292	1,116	880	1,173
1,134	932	938	1,078	1,180	1,106	1,184	954	824	529
998	996	1,133	765	775	1,105	1,081	1,171	705	1,425
610	916	1,001	895	709	860	1,110	1,149	972	1,002

Bir Sınıf Frekans Tablosu

Sınıf Aralığı	Frekans (Aralıktaki Veri Değerleri Sayısı)
500 – 600	2
600 – 700	5
700 – 800	12
800 – 900	25
900 – 1000	58
1000 – 1100	41
1100 – 1200	43
1200 – 1300	7
1300 – 1400	6
1400 – 1500	1

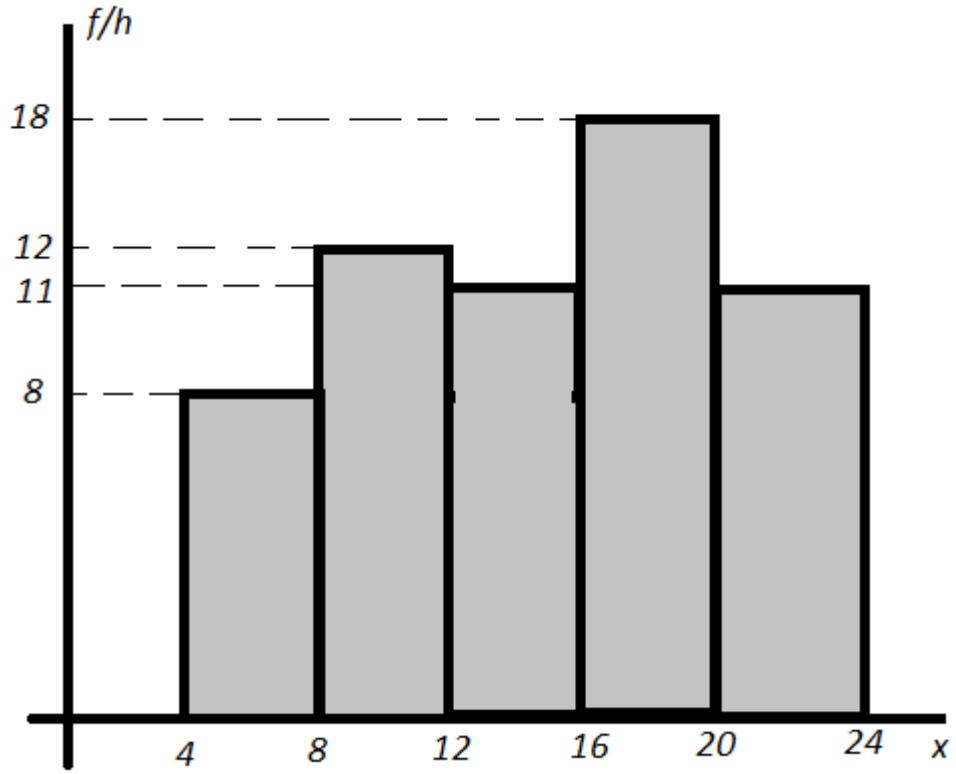
Bir frekans histogramı



Sınıf verilerinin birbirine bitişik çubuklar ile çizimine *histogram* denir. Bir histogram grafiğinin düşey ekseni ya sınıf frekansı yada sınıfın görelî frekansı ile temsil edilebilir.

Sınıf frekansının verildiği grafikler *frekans histogramı*, görelî frekansların verildiği grafikler ise *görelî frekans histogramı* olarak adlandırılır.

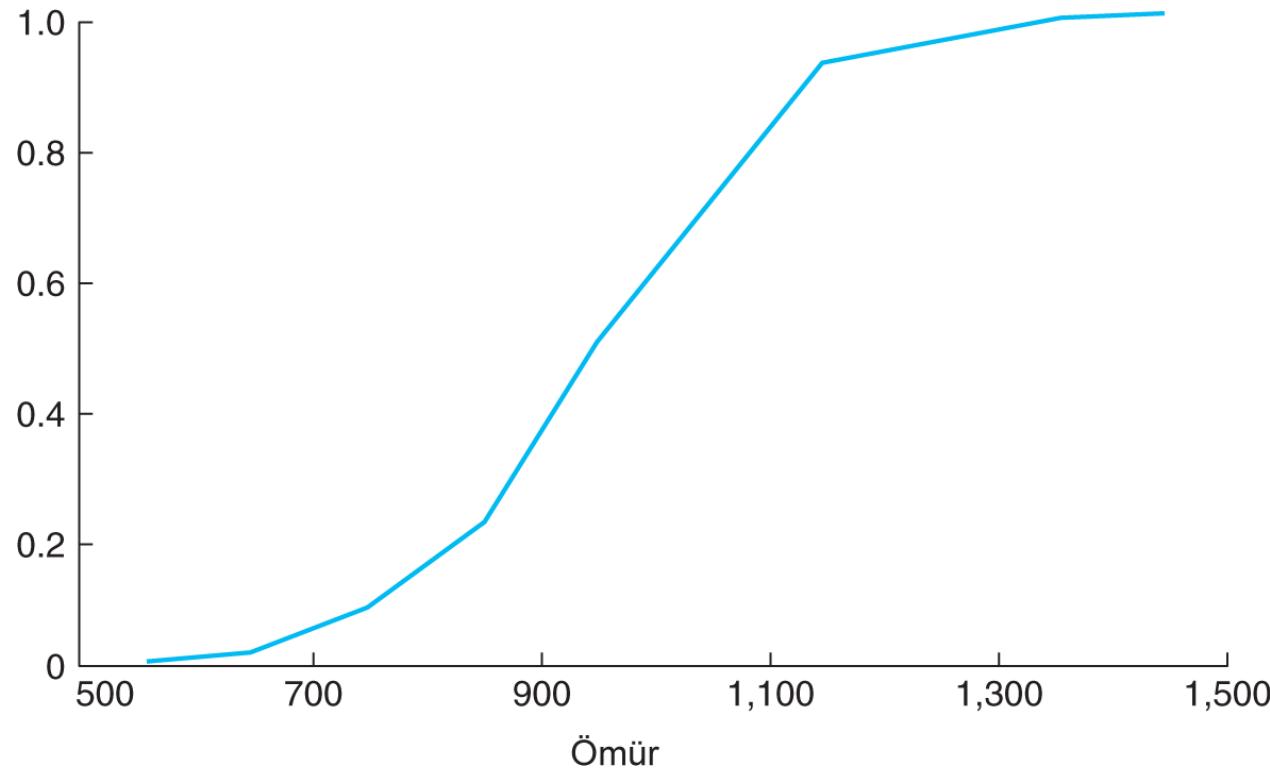
Olasılık ve İstatistik Giriş



Değişkenlerin kendileri ve varsa kendinden önceki değişkenlerin frekans yada göreli frekanslarının toplamına değişkenin *birikimli frekansı* denir.

Birikimli frekansların değişkenlerin karşısına yazılmasıyla birikimli frekans tabloları oluşturularak veya *diyagramlar* yardımıyla veri kümeleri temsil edilebilir.

Birikimli bir frekans grafiği



Olasılık ve İstatistik Giriş

<i>Yaş</i>	<i>Frekans</i>
15	2
16	5
17	11
18	9
19	14
20	13

<i>Yaş</i>	<i>– den fazla</i>	<i>Yaş</i>	<i>– den az</i>
15	54	15	2
16	52	16	7
17	47	17	18
18	36	18	27
19	27	19	41
20	13	20	54

VERİ KÜMELERİNİ ÖZETLEME

Örnek Ortalaması, Örnek Ortancası ve Örnek Tepe değeri

n tane x_1, x_2, \dots, x_n sayısal değerinden oluşan bir veri kümelerinin *örnek ortalaması* bu değerlerin aritmetik ortalaması ile bulunur. \bar{x} in aritmetik ortalaması

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

ile tanımlanır.

Türkiye Süper Ligi 2002-2016 yılları arasında futbol takımlarının şampiyon olma puanları sırasıyla aşağıda verilmiştir.

78, 85, 76, 80, 83, 70, 79, 71, 75, 82, 77, 71, 74, 77, 79

Buna göre ortalama şampiyon olma puanını belirleyiniz.

x	f
x_1	f_1
x_2	f_2
\vdots	\vdots
x_n	f_n

Frekans tablosunun aritmetik ortalaması

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$$

<i>Yaş</i>	<i>Frekans</i>
15	2
16	5
17	11
18	9
19	14
20	13

n hacimli veri kümесinin değerleri küçükten büyüğe doğru sıralanarak, n tek ise $(n + 1)/2$. değer, n çift ise $n/2$. ve $(n/2) + 1$. değerlerin ortalaması veri kümесinin *örnek ortancasıdır* (medyan).

Veri kümесinin en çok tekrar eden veya frekans değeri en büyük olan değişkene örnek *tepe değeri* denir (mod).

Olasılık ve İstatistik Giriş

<i>Yaş</i>	<i>Frekans</i>
15	2
16	5
17	11
18	9
19	14
20	13

Örnek Varyansı ve Örnek Standart Sapması

Ele alınan veri kümесinin merkezi eğilimini betimleyen istatistiklerin yanında, aynı zamanda veri değerlerinin dağılımı veya değişkenliğini betimleyen istatistiklerde kullanılır. n değişkenden oluşan x_1, x_2, \dots, x_n veri kümесinin *örnek varyansı*

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

ile tanımlanır. Örnek varyansın pozitif kareköküne *örnek standart sapması* denir.

A: 3, 4, 6, 7, 10

B: -20, 5, 15, 24

Örnek ortalaması, örnek tepedeğeri (mod), örnek ortancası(medyan) veri kümelerinin merkezi eğilim ölçülerini tanımlar iken varyans ve standart sapma veri değerlerinin yayılışını veya değişkenliğini betimleyen istatistiklerdir.

Örnek Yüzdebirlikleri ve Kutu Grafikleri

$0 \leq p \leq 1$ olmak üzere, $100p$. örnek yüzdebirliği kendisinden verilerin yüzde $100p$ sinin küçük veya eşit kaldığı veri değeridir.

n hacimli bir veri kümesinin $100p$. örnek yüzdebirliğini belirlemek için;

Bu değerden küçük veya eşit kalan en az np sayıda değerin bulunması gereklidir.

Bunun için;

- np bir tamsayı değilse; np den büyük en küçük tamsayıncı en küçük veri
- np bir tamsayı ise; $np.$ ve $np + 1.$ en küçük verilerin ortalaması

$100p.$ yüzdebirliği verir.

Örneğin;

$n = 22$ olan veri kümesi için $p = 0.8$ alalım.

$100p = 100 \times 0.8 = 80$. örnek yüzde birliği

$np = 22 \times 0.8 = 17.6$ olur.

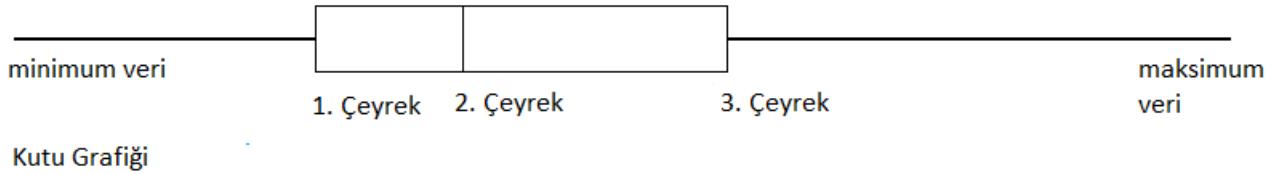
18. en küçük veri 80. örnek yüzde birliği verir.

Bir veri kümesinin;

- 25. örnek yüzdebirligine birinci çeyreklik,
- 50. örnek yüzdebirligine ikinci çeyreklik,
- 75. örnek yüzdebirligine üçüncü çeyreklik

denir.

Olasılık ve İstatistik Giriş



82, 89, 94, 110, 74, 122, 112, 95, 100, 78, 65, 60, 90, 83, 87, 75, 114, 85
69, 94, 124, 115, 107, 88, 97, 74, 72, 68, 83, 91, 90, 102, 77, 125, 108, 65

6	0, 5, 5, 8, 9
7	2, 4, 4, 5, 7, 8
8	2, 3, 3, 5, 7, 8, 9
9	0, 0, 1, 4, 4, 5, 7
10	0, 2, 7, 8
11	0, 2, 4, 5
12	2, 4, 5

36 farklı zamanda yapılmış olan gürültü büyüğünü ölçümlerinin çeyrekliklerini



TABLE 2.1 *Starting Yearly Salaries*

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1



En büyük veri ile en küçük veri arasındaki fark *açıklık*, üçüncü çeyreklik ile birinci çeyreklik arasındaki fark *çeyreklikler arası açıklık* olarak adlandırılır.

CHEBYSHEV EŞİTSİZLİĞİ

Bir veri kümесinin ortalaması \bar{x} ve standart sapması s olsun.

$k \geq 1$ olmak üzere verilerin yüzde $100(1 - \frac{1}{k^2})$ den fazlası

$(\bar{x} - ks, \bar{x} + ks)$ aralığındadır. Buna Chebyshev eşitsizliği denir.

En Çok Satan Araçlar

Nisan 2008

1.	Ford F Series	44,813
2.	Toyota Camry	40,016
3.	Chevrolet Silverado.....	37,231
4.	Honda Accord Hybrid	35,075
5.	Toyota Corolla Matrix	32,535
6.	Honda Civic Hybrid	31,710
7.	Chevrolet Impala.....	26,728
8.	Dodge Ram	24,206
9.	Ford Focus	23,850
10.	Nissan Altima Hybrid	22,630

$$\bar{x} = 31,879.4 \quad s = 7,514.7$$

$$\begin{aligned}
 k &= 3/2 \text{ alınırsa verilerin yüzde} \\
 100(1 - 1/k^2) &= 100(5/9) \\
 &= 55.55 \text{ den fazla}
 \end{aligned}$$

$$\left(\bar{x} - \frac{3}{2}s, \bar{x} + \frac{3}{2}s \right) = (20\,607,43\,151)$$

aralığındadır.

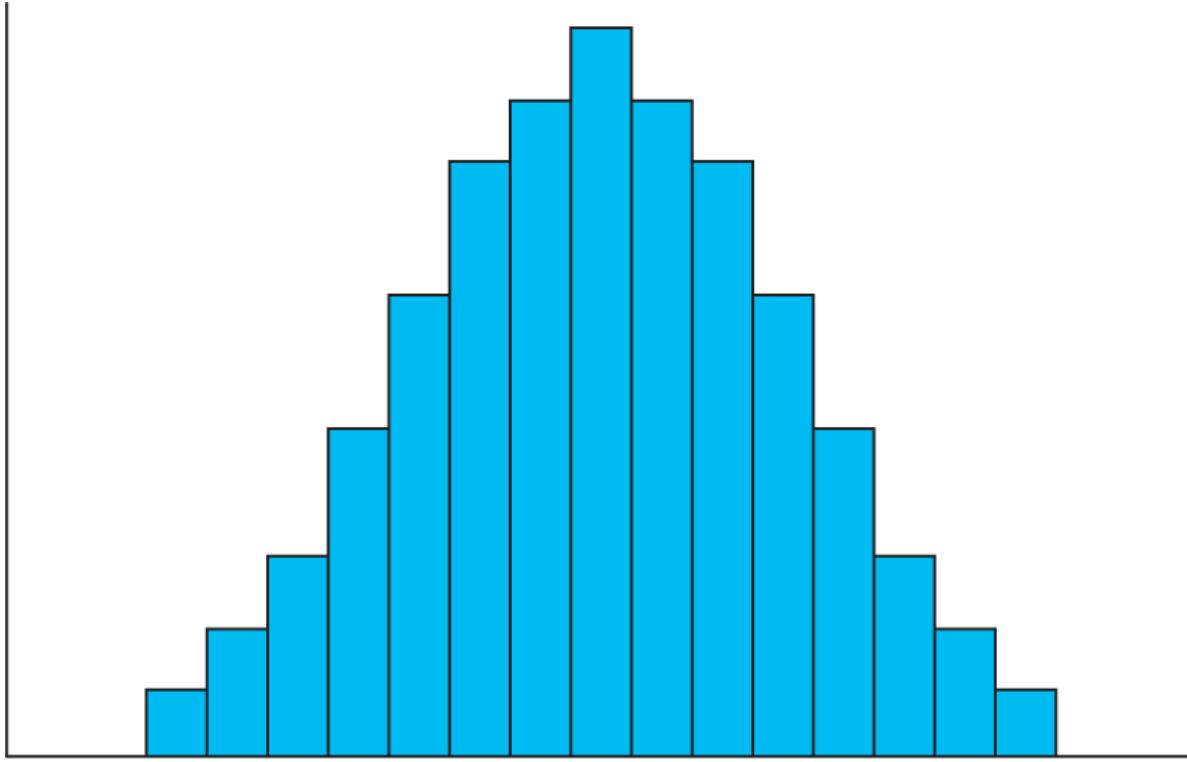
Gerçekte bu oran yüzde 90 dır.

NORMAL VERİ KÜMELERİ

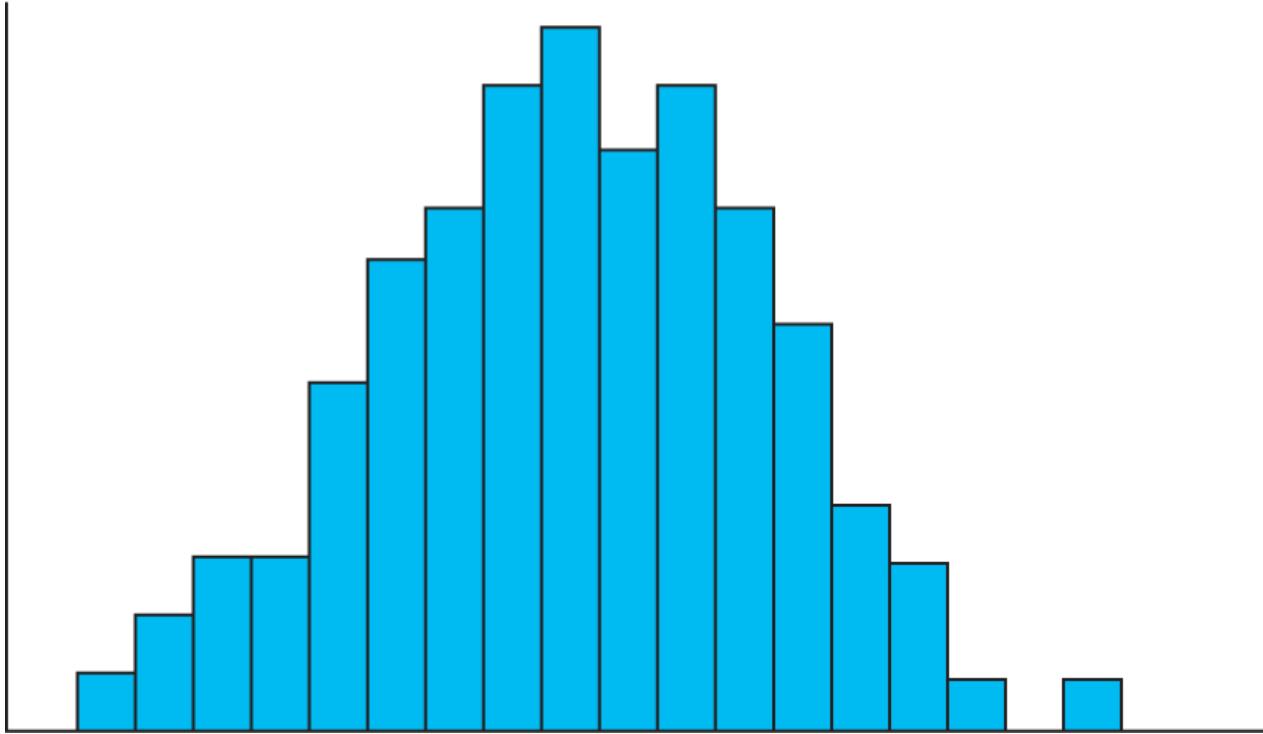
Uygulamada gözlemlenen büyük veri kümelerinin birçoğu benzer biçimde olan histogramlara sahiptir. Bu histogramlar örnek ortancasında zirveye ulaşır ve sonra bu noktanın her iki yanında çan biçimli simetrik bir tarzda azalır.

Böyle veri kümelerinin *normal* olduğu söylenir ve histogramlarına *normal histogramlar* denir.

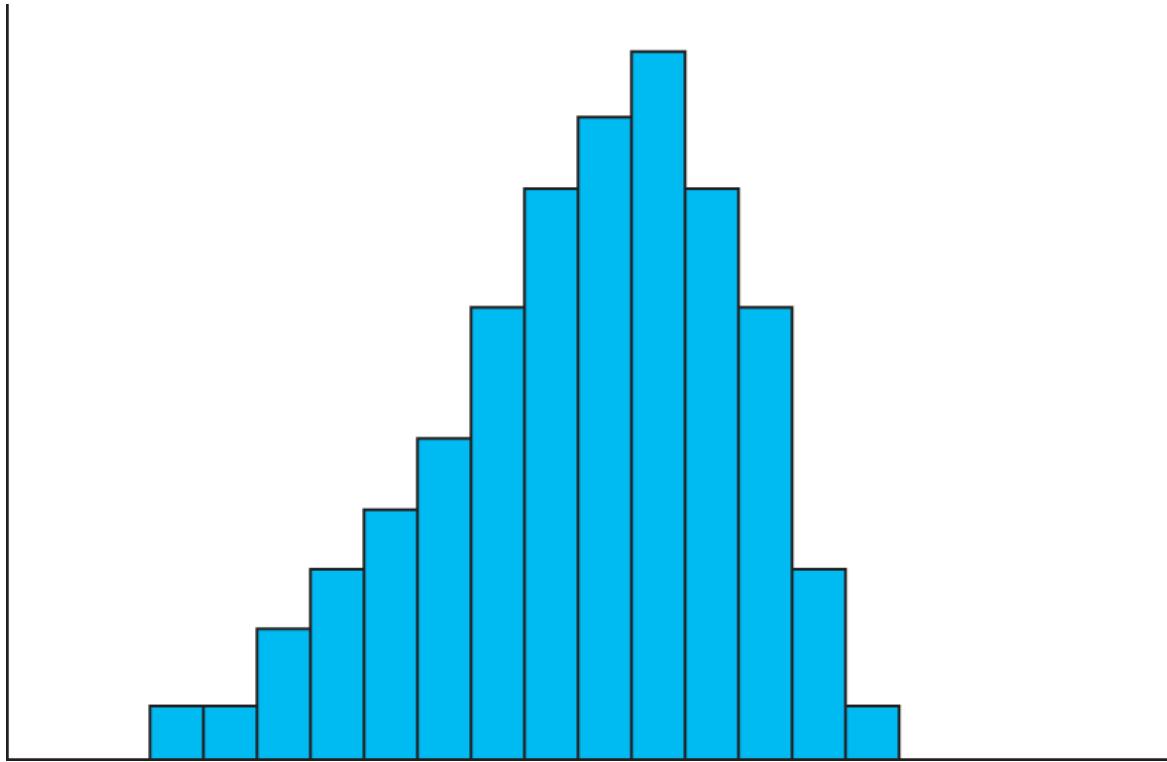
Normal bir veri kumesinin histogramı



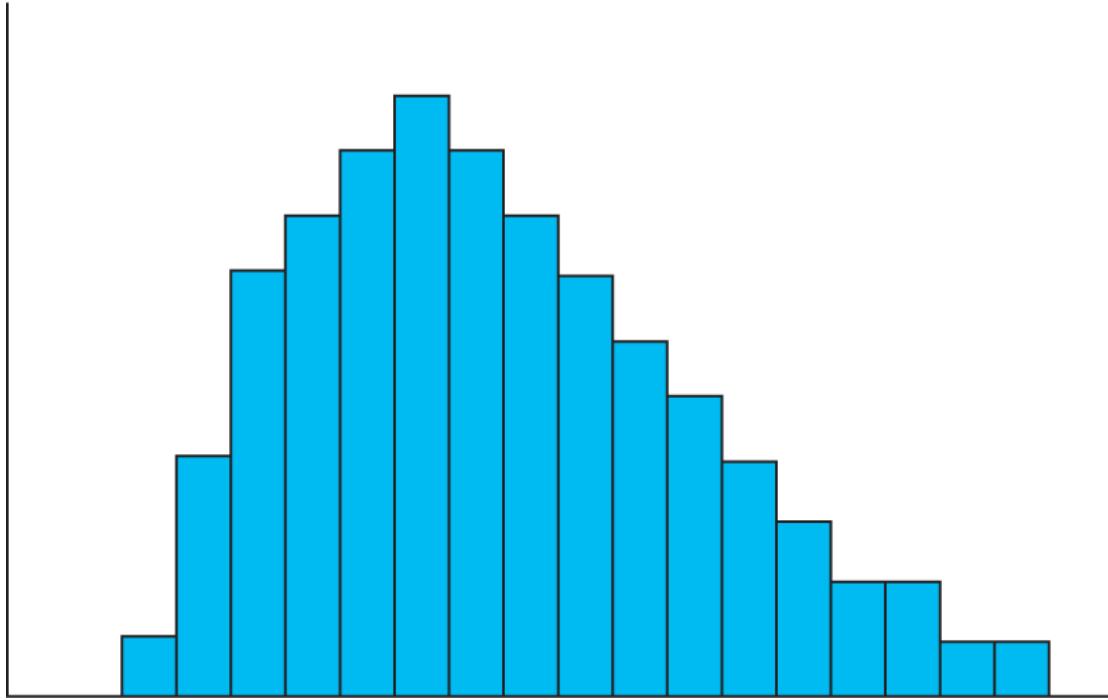
Yaklaşık normal bir veri kümesinin histogramı



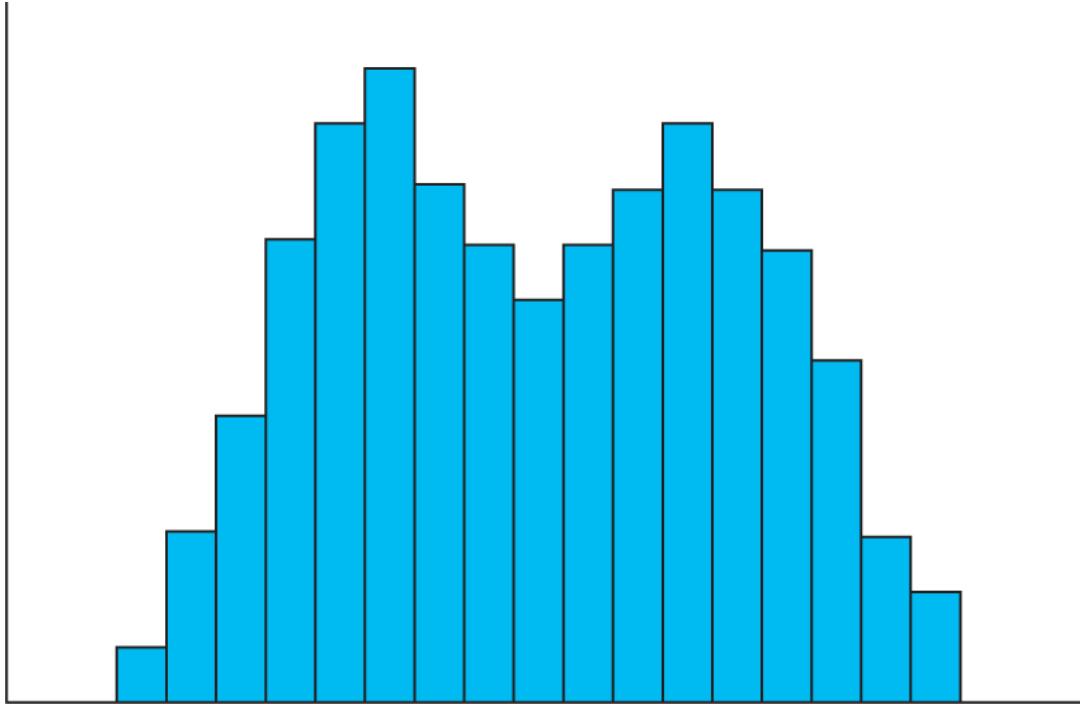
Sola çarpık bir veri kümesinin histogramı



Sağça çarpık bir veri kümesinin historogramı



İki tepedeğerli bir veri kümesinin histogramı



Gözlemsel Kural

Yaklaşık normal dağılım gösteren bir veri kümelerinin ortalaması \bar{x} ve standart sapması s olsun. Bu durumda

- Verilerin yaklaşık %68 i
 $(\bar{x} - s, \bar{x} + s)$
- Verilerin yaklaşık %95 i
 $(\bar{x} - 2s, \bar{x} + 2s)$
- Verilerin yaklaşık %99.7 si
 $(\bar{x} - 3s, \bar{x} + 3s)$

aralığında kalır.

Bir sınıfındaki öğrencilerin istatistik dersinden aldığı notlar;

9		0, 1, 4
8		3, 5, 5, 7, 8
7		2, 4, 4, 5, 7, 7, 8
6		0, 2, 3, 4, 6, 6
5		2, 5, 5, 6, 8
4		3, 6

$$\bar{x} = 70.57 \text{ ve } s = 14.35 \text{ olarak bulunur.}$$

$$(\bar{x} - s, \bar{x} + s) = (56.2, 84.9) \rightarrow \%68 \text{ i} \\ (\text{Gerçek oran} = \%53)$$

$$(\bar{x} - 2s, \bar{x} + 2s) = (41.8, 99.2) \rightarrow \%95 \text{ i} \\ (\text{Gerçek oran} = \%100)$$

EŞLEŞTİRİLMİŞ VERİ KÜMELERİ VE ÖRNEK KORELASYON KATSAYISI

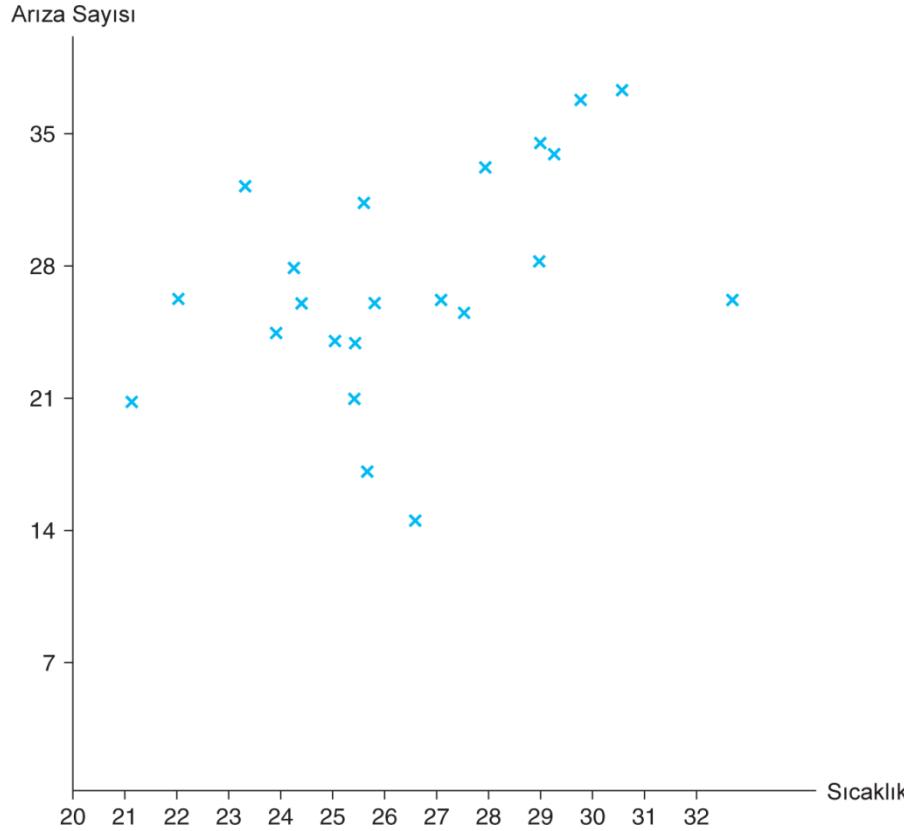
Sıklıkla birbiriyle bazı ilişkilere sahip değer çiftlerinden oluşan veri kümeleriyle ilgilenilir. Böyle bir veri kümesinde her bir eleman bir x ve bir y değerine sahipse, i . veri noktası bir (x_i, y_i) çiftiyle temsil edilir.

TABLE 2.8 *Temperature and Defect Data*

Day	Temperature	Number of Defects
1	24.2	25
2	22.7	31
3	30.5	36
4	28.6	33
5	25.5	19
6	32.0	24
7	28.6	27
8	26.5	25
9	25.3	16
10	26.0	14
11	24.4	22
12	24.8	23
13	20.6	20
14	25.1	25
15	21.4	25
16	23.7	23
17	23.9	27
18	25.2	30
19	27.4	33
20	28.3	32
21	28.8	35
22	26.6	24

Eşleştirilmiş değerlerin veri kümesini resmetmenin yararlı bir yolu, x-ekseni verilerin x değerini ve y-ekseni verilerin y değerini göstermek üzere verileri iki boyutlu bir grafik üzerinde işaretlemektir. Böyle bir işaretlemeye *serpme diyagramı* denir.

Serpme diyagramı



Eşleştirilmiş veri kümelerinde büyük x değerlerinin büyük y değerleri ile eşleşme eğilimini ölçmede *örnek korelasyon katsayısı* istatistiği kullanılır.

s_x , x değişkeninin ve s_y , y değişkeninin örnek standart sapmaları olmak üzere, (x_i, y_i) veri çiftlerinin *örnek korelasyon katsayısı*

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

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

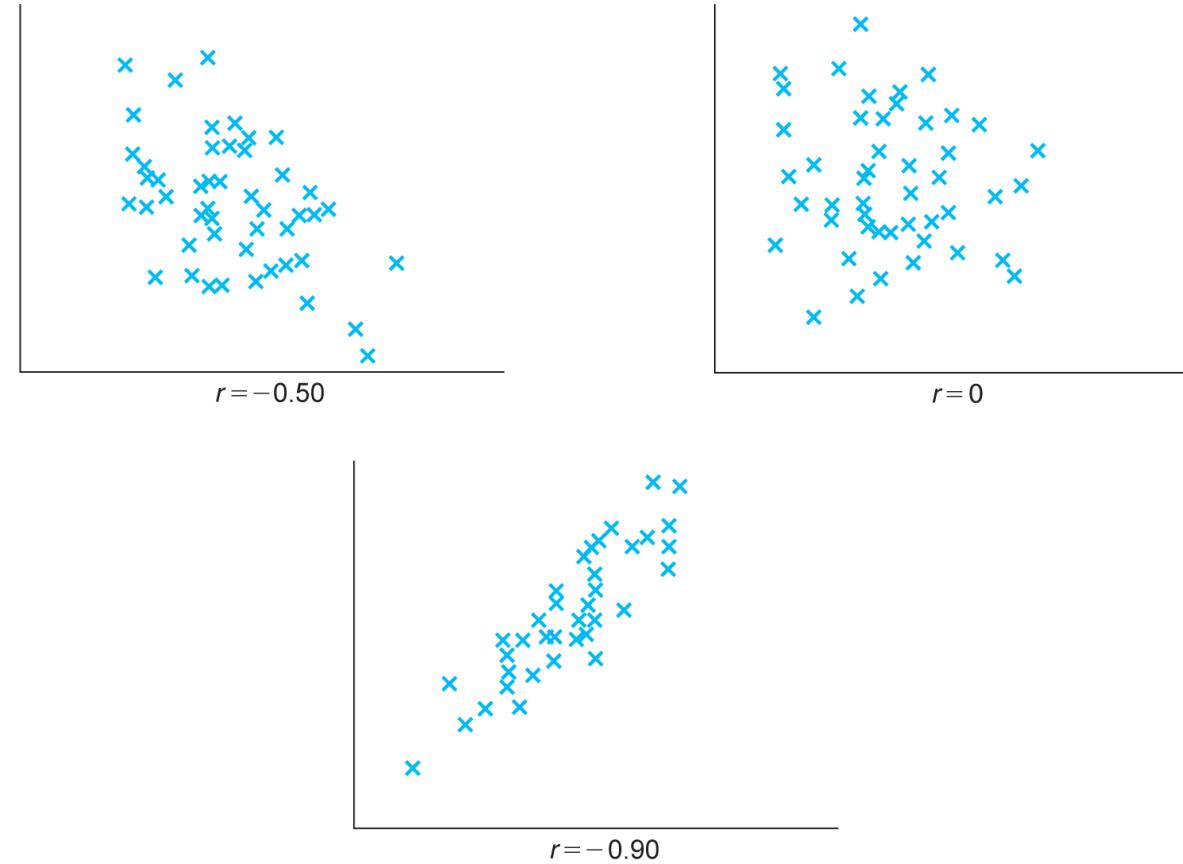
olarak tanımlanır. Veri çiftleri $r > 0$ ise pozitif ilişkili ve $r < 0$ ise negatif ilişkilidir.

Örnek Korelasyonun Özellikleri

- Örnek korelasyonu -1 den büyük veya eşit ve 1 den küçük veya eşittir.
- Büyük x verilerinin büyük y verilerine bağlayacak şekilde bir doğru varsa korelasyon değeri 1 dir. (Doğrusal Korelasyon)
- Büyük x verilerinin küçük y verilerine bağlayacak şekilde bir doğru varsa korelasyon değeri -1 dir.
- Verilerinden herhangi birinin tamamına sabit bir sayı eklenmesi yada sabit bir sayıyla çarpılmasıyla korelasyon sayısı değişmez

Korelasyon katsayısının mutlak değeri değişkenlerin arasındaki doğrusal ilişkinin gücünün göstergesidir.

Örnek korelasyon katsayıları



Örneğin;

Sıcaklık ile arıza sayısı verilerinin korelasyonu $r = 0.4189$ olarak bulunmuştur. Bu gün ortası sıcaklık ile o gün üretilen arızalı parça sayısı arasında görelî olarak zayıf bir korelasyon olduğu anlamına gelir.

Korelasyon Nedenselliği Değil, Birlikteliği Ölçer

Kişi No	1	2	3	4	5	6	7	8	9	10
Okul Yılı	12	16	13	18	19	12	18	19	12	14
Nabız Sayısı	73	67	74	63	73	84	60	62	76	71

$r = -0.7638$ olduğundan bireylerin okul yılı sayısı ve nabız sayısı arasında güçlü bir negatif korelasyon olduğu görülür. yüksek nabız sayısının düşük okul yılıyla ilişkili olduğunu gösterir.

Okul yılı ve nabız sayısı serpme diyagramı

