Olasılık ve İstatistik

Fırat İsmailoğlu, PhD

Veri Analizi

Sıklık Dağılımı / Histogram

Veri toplandığından yapılacak ilk seylerden biri, verideki her bir değerin (skorun) veride kaç defa görüldüğünü saymaktır. Ortaya çıkan sayılar eldeki verinin ne olduğu hakkında önemli ipucu verir.

Çok basit bir örnek olarak, diyelimki 10 kişilik bir arkadas grubundaki kisilerin eğitim seviyeleri şöyle olsun:

lise,üniversite, y.lisans, y.lisans, lise, universite, universite, universite, ortaokul, y.lisans.

Sıklık Dağılımı / Histogram

Veri toplandığından yapılacak ilk seylerden biri, verideki her bir değerin (skorun) veride kaç defa görüldüğünü saymaktır. Ortaya çıkan sayılar eldeki verinin ne olduğu hakkında önemli ipucu verir.

Çok basit bir örnek olarak, diyelimki 10 kişilik bir arkadas grubundaki kisilerin eğitim seviyeleri şöyle olsun:

lise,üniversite, y.lisans, y.lisans, lise, universite, universite, universite, ortaokul, y.lisans olsun.

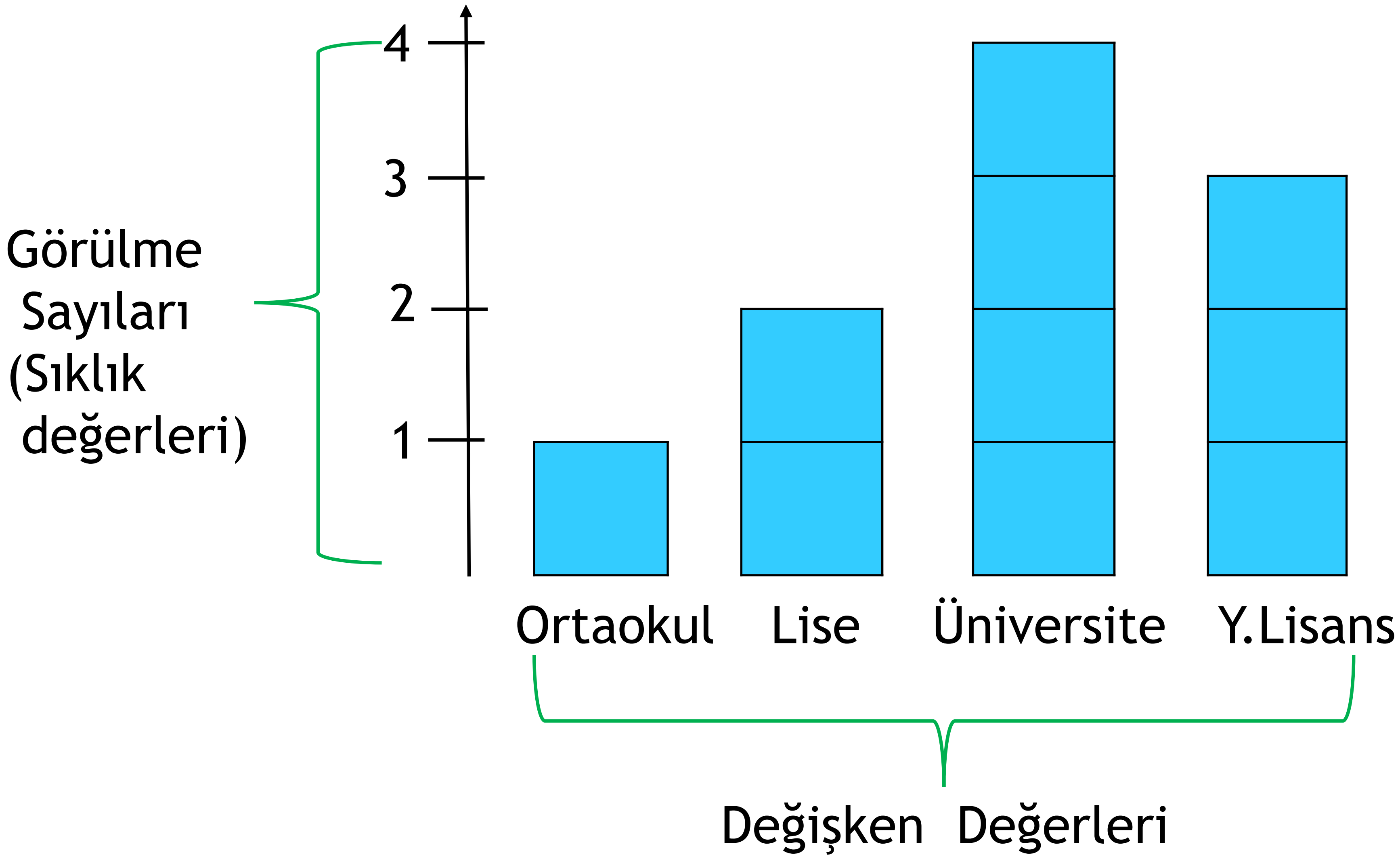
Bu durumda 1 ortaokul, 2 lise, 4 üniversite ve 3 y.lisans mezunu vardır.

Tablo hali:

Eğitim Seviyesi	Sayı
Ortaokul	1
Lise	2
Üniversite	4
Y.Lisans	3

Sıklık Tablosu
(Frequency table)

Sıklık Dağılımı / Histogram



Değişkenin aldığı farklı değerlerin veride kaç defa görüldüğünü gösteren sütun grafiğine **histogram** diyeceğiz.

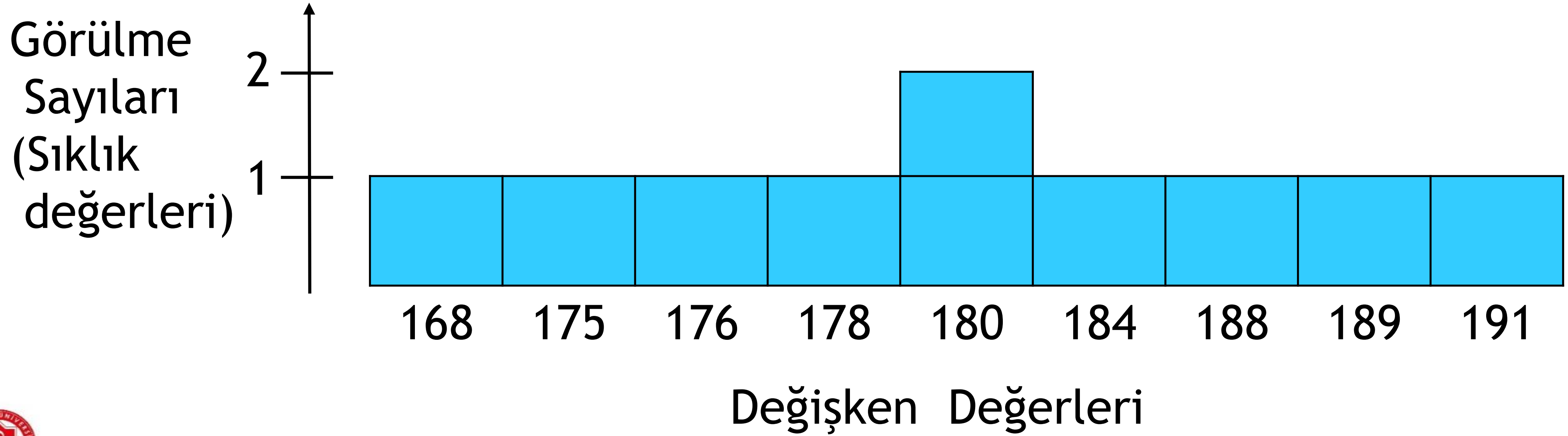
Sıklık Dağılımı / Histogram

Peki ya verideki değişkenimiz sayısal ise? Bu durumda histogram nasıl olur?

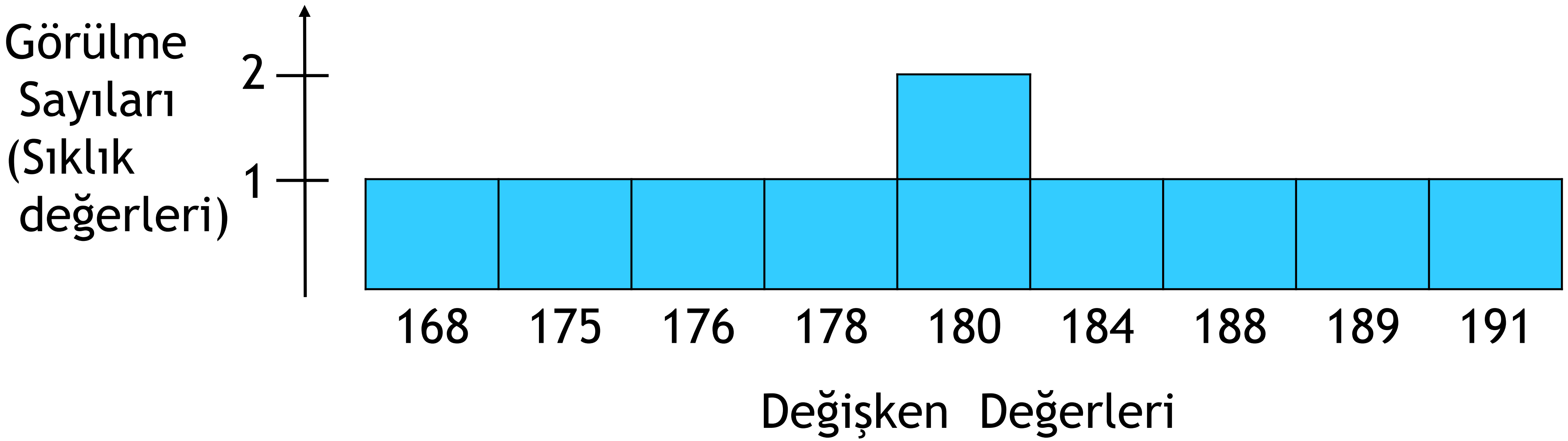
Yine diyelimki 10 kişilik bu arkadaş grubundaki kişilerin boyları cm cinsiden şöyle olsun:

191, 168, 176, 175, 188, 180, 184, 180, 189, 178.

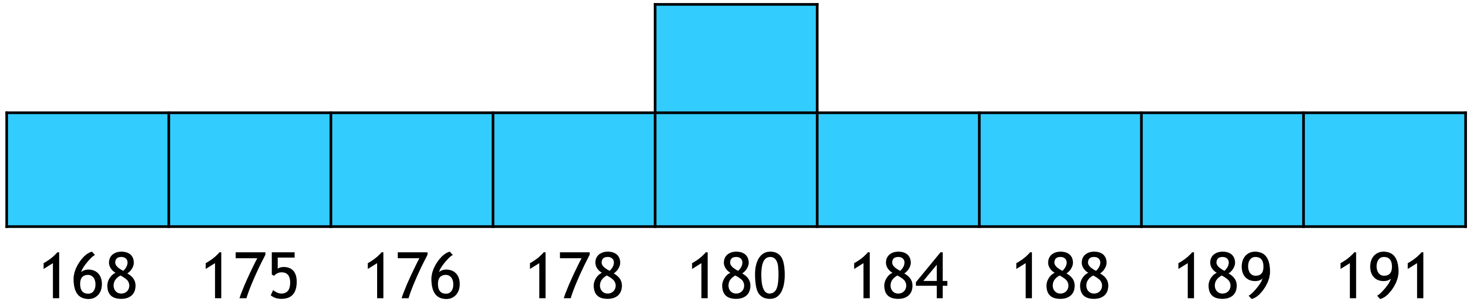
Bu durumda 180cm iki kez, diğer uzunluklar bir kez görülmüş olur. O halde histogram:



Sıklık Dağılımı / Histogram



Sıklık Dağılımı / Histogram



Bu histograma bakarak gruptaki kişilerin boyları hakkında hemen bir kanıya varmak, bir sonuç çıkarmak güçtür.

Değişken sayısal iken, bu değişkenin daha çok hangi değerler aldığını göstermek için, yani bu değerin histogramı için, genel yaklaşım değeri gruplara bölmektir: buna ingilizcede (binnning, (kutulamak)) denir.

Bu veriden 3 grup oluşturalım:

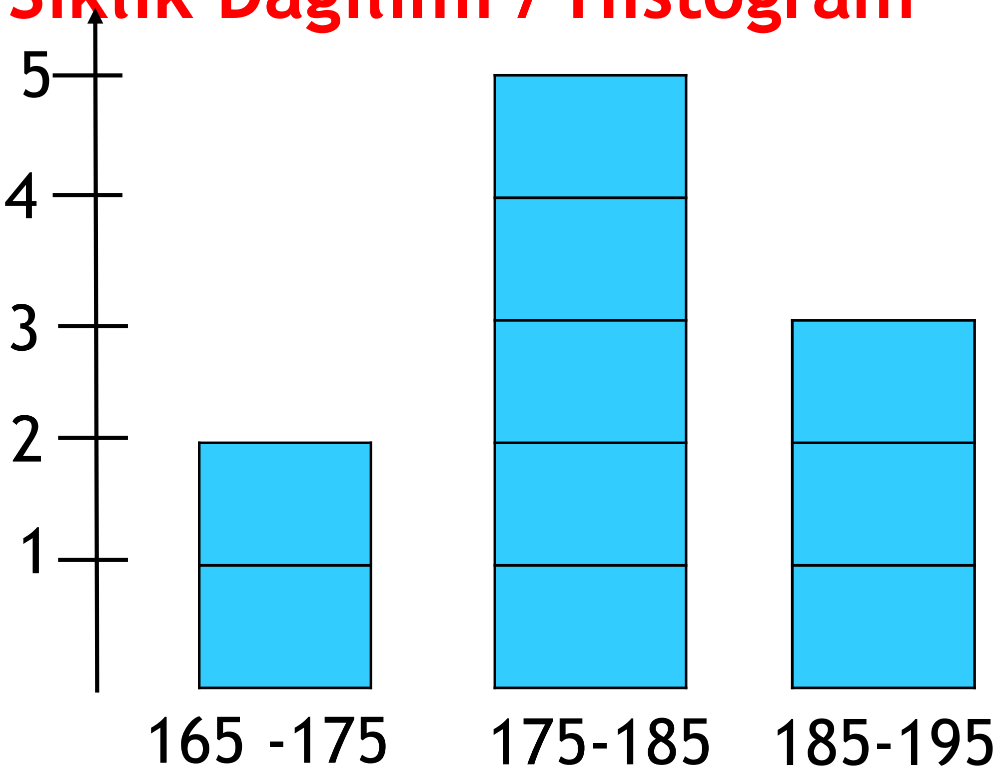
165 – 175 arası değerler: 168, 175 (2 adet)

175 – 185 arası değerler: 176, 178, 180,180,184 (5 adet)

185 – 195 arası değerler: 188, 189, 191 (3 adet)

Not : Yukarıdaki gruplama tamamen kişiseldir. İsteyen değişkenleri başka aralıklarla da gruplayabilir, ama grup aralıklarının aynı olmasına dikkat edilmelidir!

Sıklık Dağılımı / Histogram



Bu histograma bakarak verimizdeki değişken olan boy uzunluğu hakkında <u>hemen</u> bir sonuca varabiliriz: *Gruptaki kişilerin %50'sinin boyu 175 cm ile 185 cm arasındadır.*

Soru 1: Aşağıdaki veri, bir şirkette çalışan 50 kişinin son 6 haftadaki izinli gün sayısını göstermektedir. Bu veriye ait histogramı çiziniz.

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

Göreli Sıklık Dağılımı (Relative Frequency Distribution)

Veride herbir değerin görülme sayısını dikkate aldığımızda, ortaya çıkan sıklık dağılımı o veriye özgü (elimizdeki örneğe (sample)) gibi görülebilir. Verinin alındığı populasyona yönelik daha genel bir sonuca varmak için görülme sayılıarı normalleştirilerek, her bir sayının [0-1] arası bir değer alması sağlanabilir. Bu şekilde oluşturulan sıklık dağılımına göreli sıklık dağılımı (relative frequency table) denilir.

Bir görülme sayısını [0-1] aralğına getirmek için, bu sayıyı toplam görülme sayısına (verideki toplam eleman sayısı) böleriz:

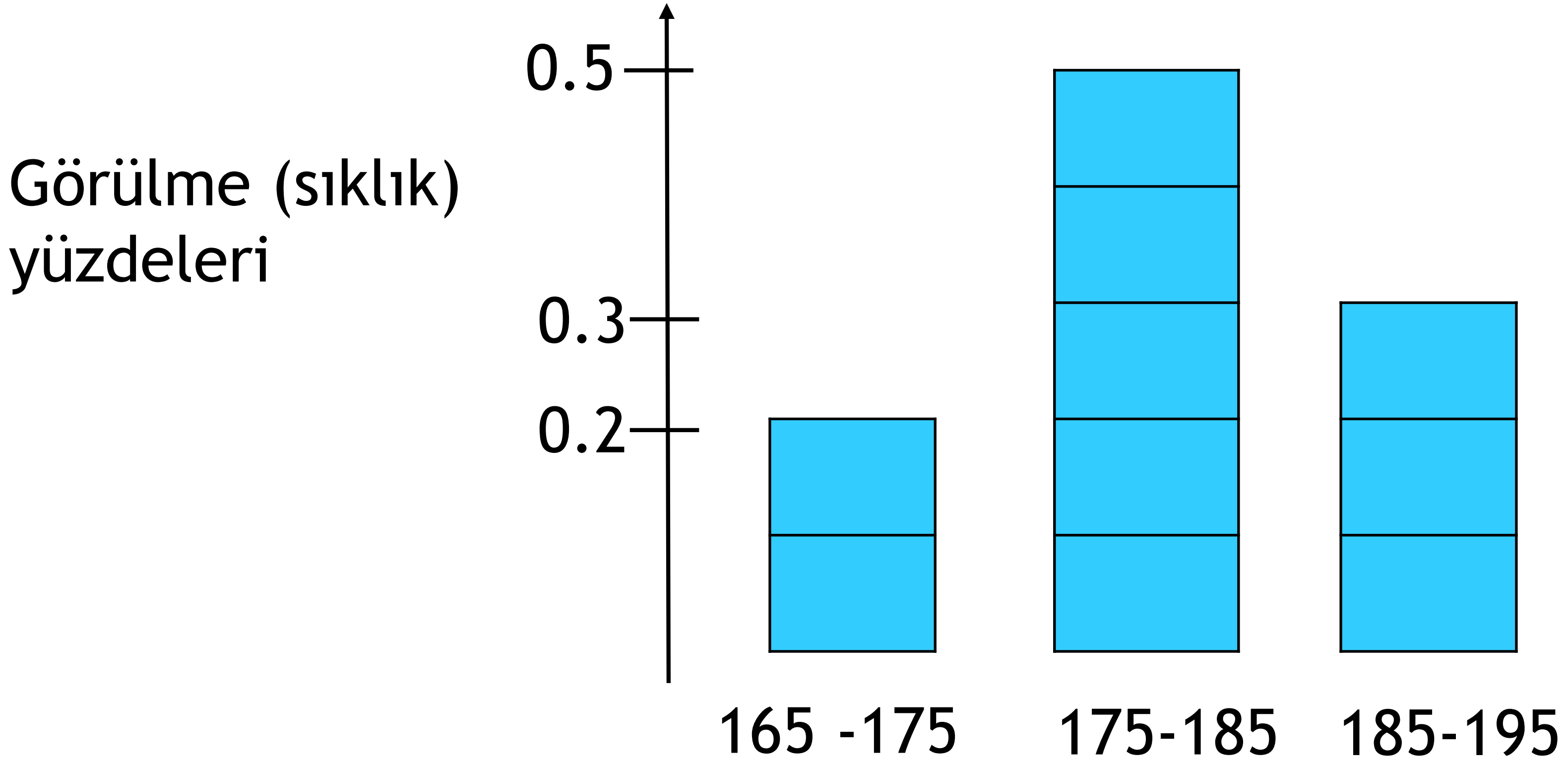
Bir önceki örnekte toplam 10 kişi vardı. Bu 10 kişinin 2'sinin boyu 165-175 arasında idi. O halde 165-175 cm olanların görülme (sıklık) yüzdesi: $\frac{2}{10} = 0.2$

175-185 arası 5 kişinin görülme yüzdesi: $\frac{5}{10} = 0.5$

185-195 arası 3 kişinin görülme yüzdesi: $\frac{3}{10} = 0.5$

Göreli Sıklık Dağılımı (Relative Frequency Distribution)

Histogramı görülme yüzdelerini kullanarak yaparsak:



Soru 2: Aşağıdaki veri, bir şirkette çalışan 50 kişinin son 6 haftadaki izinli gün sayılsını göstermektedir. Bu veriye ait histogramı görülme yüzdelerini gösterecek şekilde çiziniz.

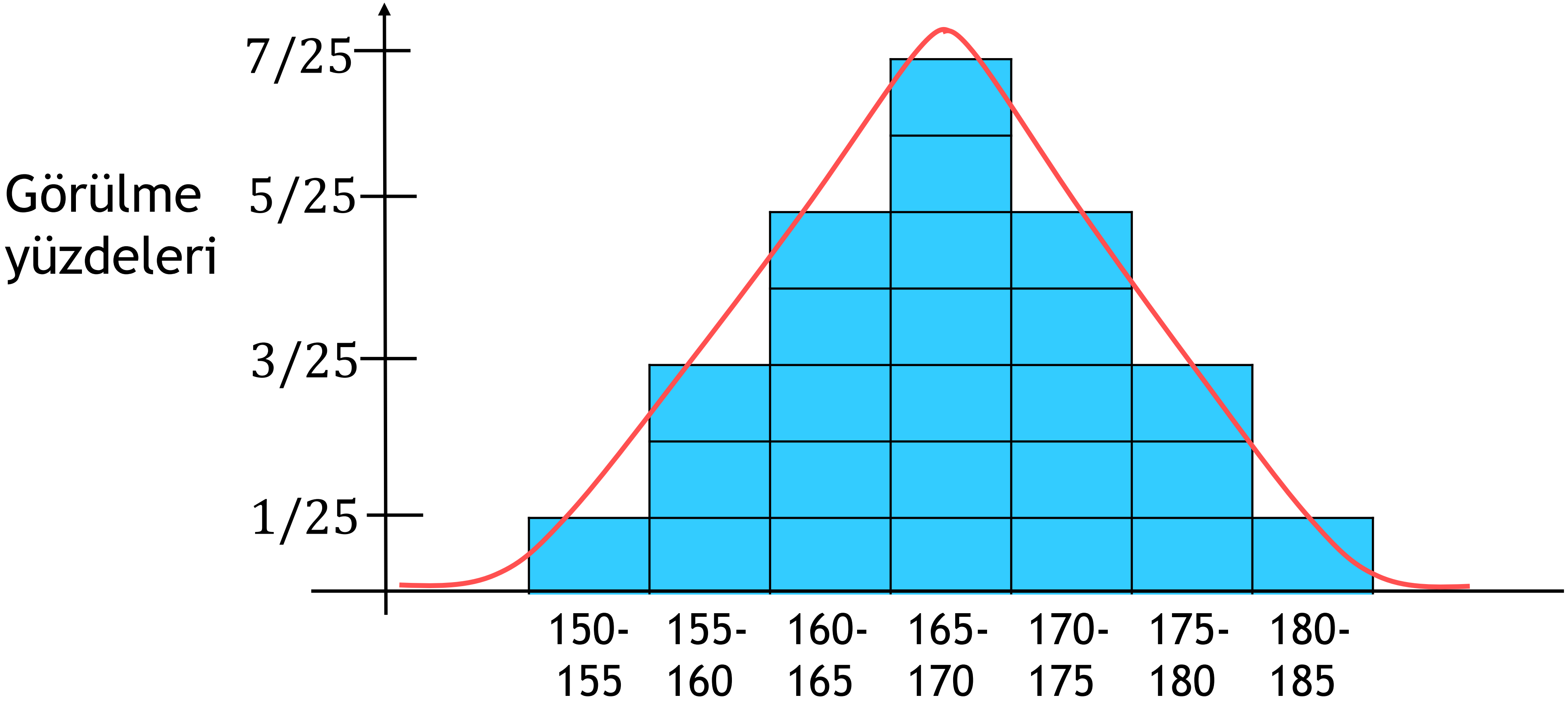
2, 2, 0, 0, 5, 8, 3, 4, 1, 0, 0, 7, 1, 7, 1, 5, 4, 0, 4, 0, 1, 8, 9, 7, 0,
1, 7, 2, 5, 5, 4, 3, 3, 0, 0, 2, 5, 1, 3, 0, 1, 0, 2, 4, 5, 0, 5, 7, 5, 1

Sıklık Dağılımı Ne Zaman Normaldir?

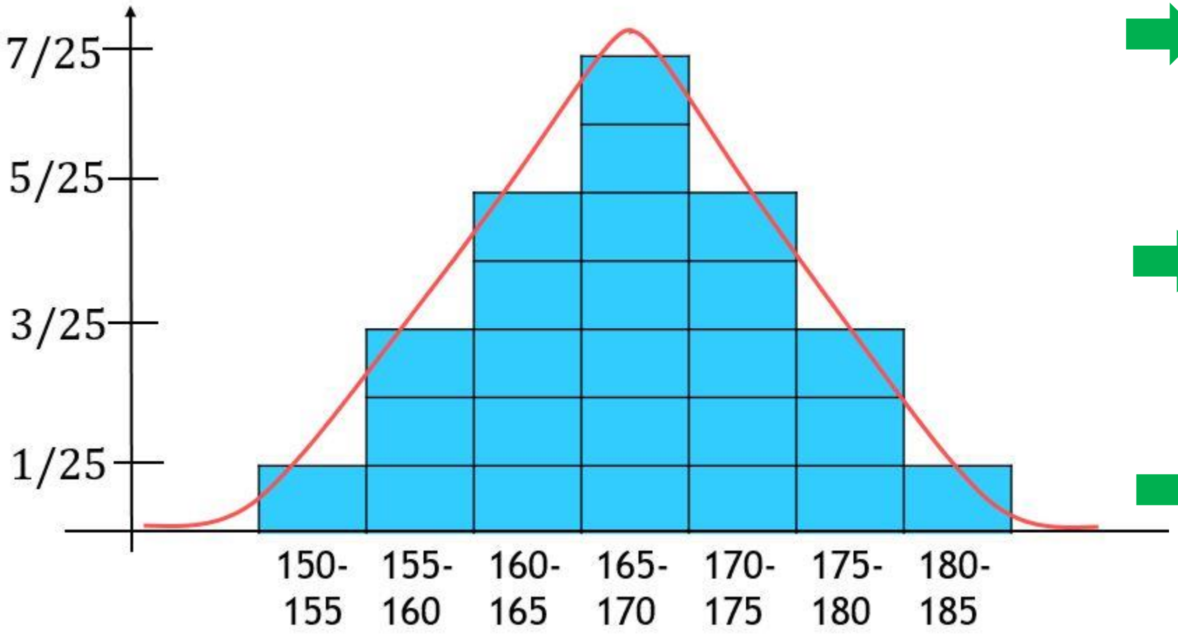
İdeal dünyada veri, merkezi etrafında toplanmıştır. Bu,"normal" olandır.
Verinin bu şekilde merkezi etrafında toplanmasına normal dağılım diyeceğiz.

Normal dağılıma sahip bir verinin histogramı çan eğrisi şeklinde olur.

ör.



Sıklık Dağılımı Ne Zaman Normaldir?

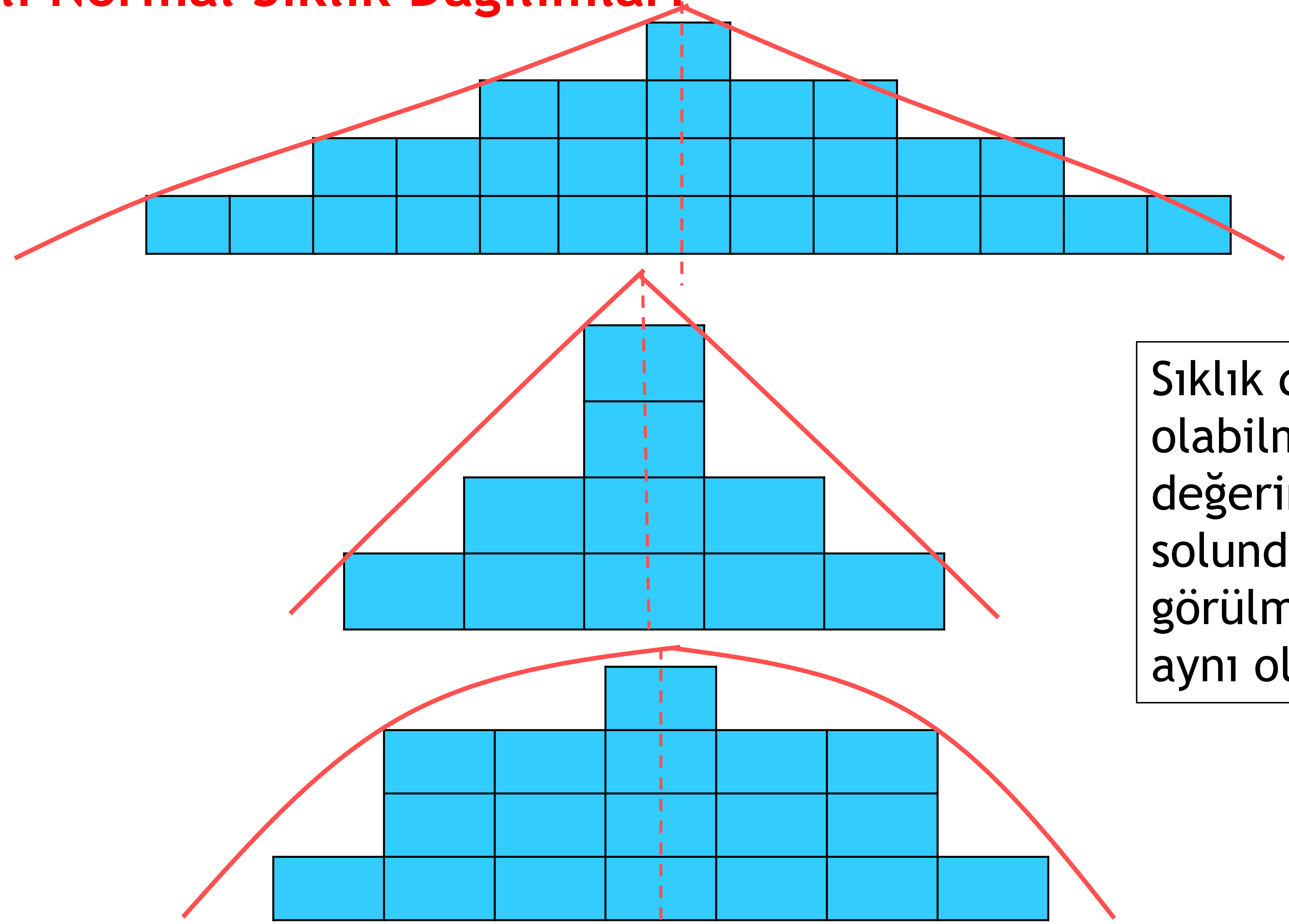


- Ortalama boy olan 165-170 boya sahip kişilerin sayısı en fazladır.
- Uç değerdeki boylara: 150-155 ve 180-185 sahip kişilerin sayıları en azdır.
- Grafikte ortalama boyun sağı ve solu simetriktir.

Sonuç olarak; eğer sıklık dağılımı normalse veri merkezi etrafına toplanmıştır.
Bu şu demektir:

- En fazla ortalama değer görülür.
- Uç değerler (en küçük değerler, en büyük değerler) en az görülür.
- Ortalama değerin altındaki ve ve üstündeki değerlerin görülme sayısı hemen hemen aynıdır.

Farklı Normal Sıklık Dağılımları



Sıklık dağılımının normal olabilmesi için ortalama değerin sağındaki ve solundaki değerlerin görülme sayılarının aynı olması gerekir.

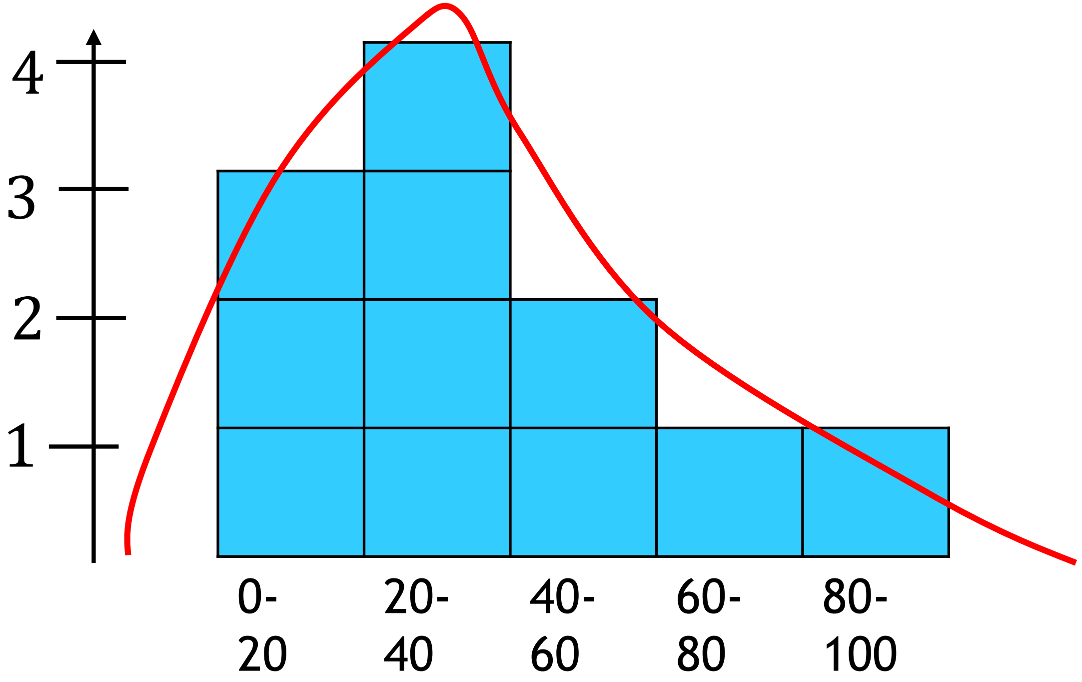
Sıklık Dağılımının Normalden Sapması

Bazen en sık görülen değer ortalama değer olmaz. Örneğin küçük değerler çok daha sık görülebilir.

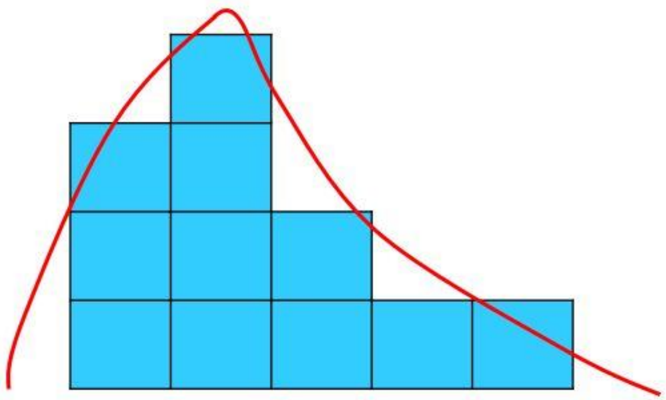
Örneğin 11 kişilik bir arkadaş topluluğunda istatistik dersinden alınan vize notları:

70, 50, 5, 45, 35, 17, 25, 27, 90, 5, 38

olsun. Bu değerleri 20'lik aralıklara bölersek: 0-20 arası 3 değer, 20-40 arası 4 değer, 40-60 arası 2 değer, 60-80 arası 1 değer, 80-100 arası 1 değer görülür.



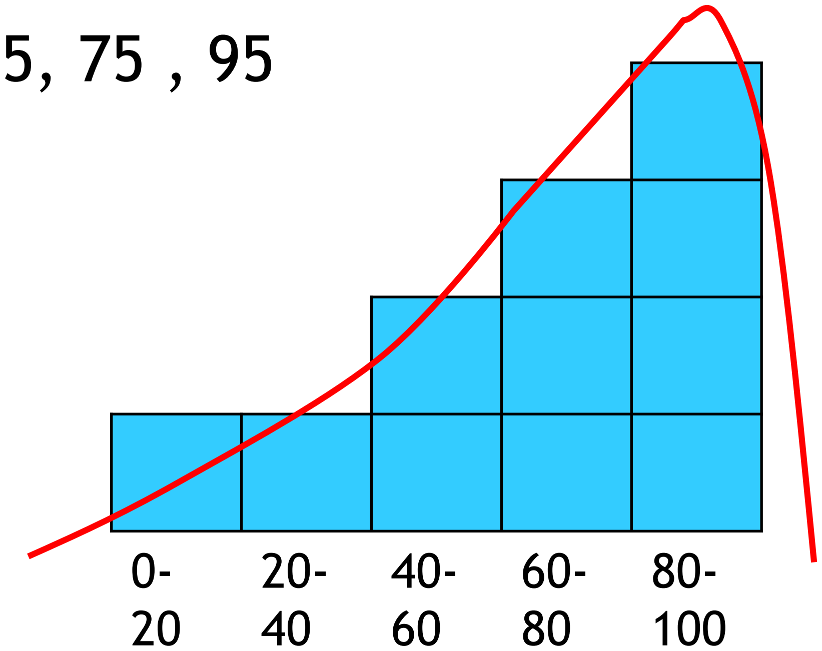
Sıklık Dağılımının Normalden Sapması



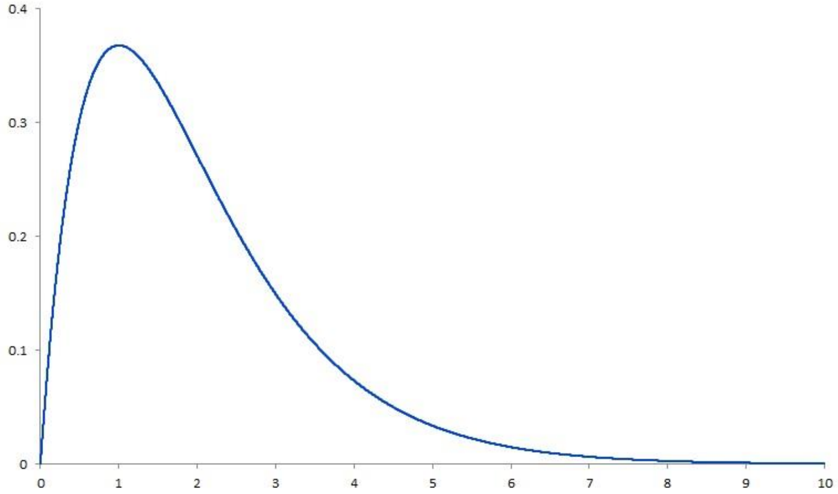
Bu şekilde küçük değerlerin daha çok görüldüğü dağılıma pozitif eğriliğe (positively skewed) sahip dağılım diyeceğiz.

Karşıt şekilde veride eğer büyük değerler daha çoksa, bu sıklık dağılımına negatif eğriliğe (negatively skewed) sahip dağılım denir. Bir önceki örnekte notlar eğer :

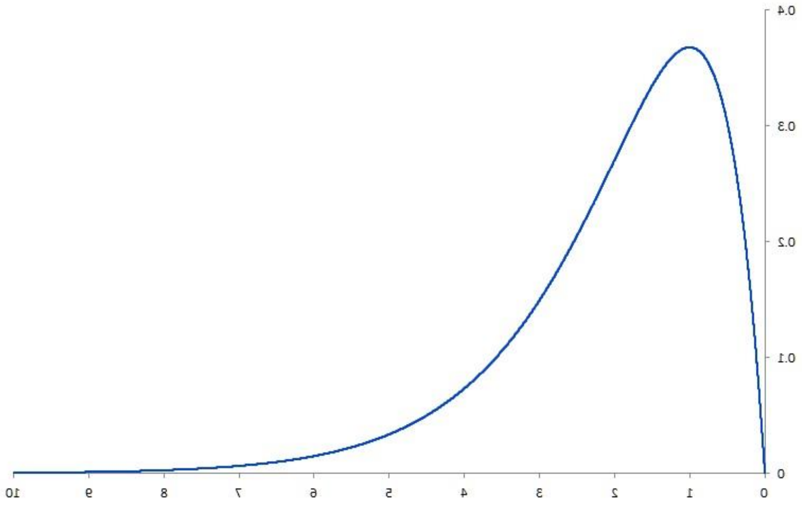
25, 50, 5, 52, 82,100, 95, 78, 65, 75 , 95



Pozitif Eğriliğe Sahip Sıklık Dağılımı



Negatif Eğriliğe Sahip Sıklık Dağılımı



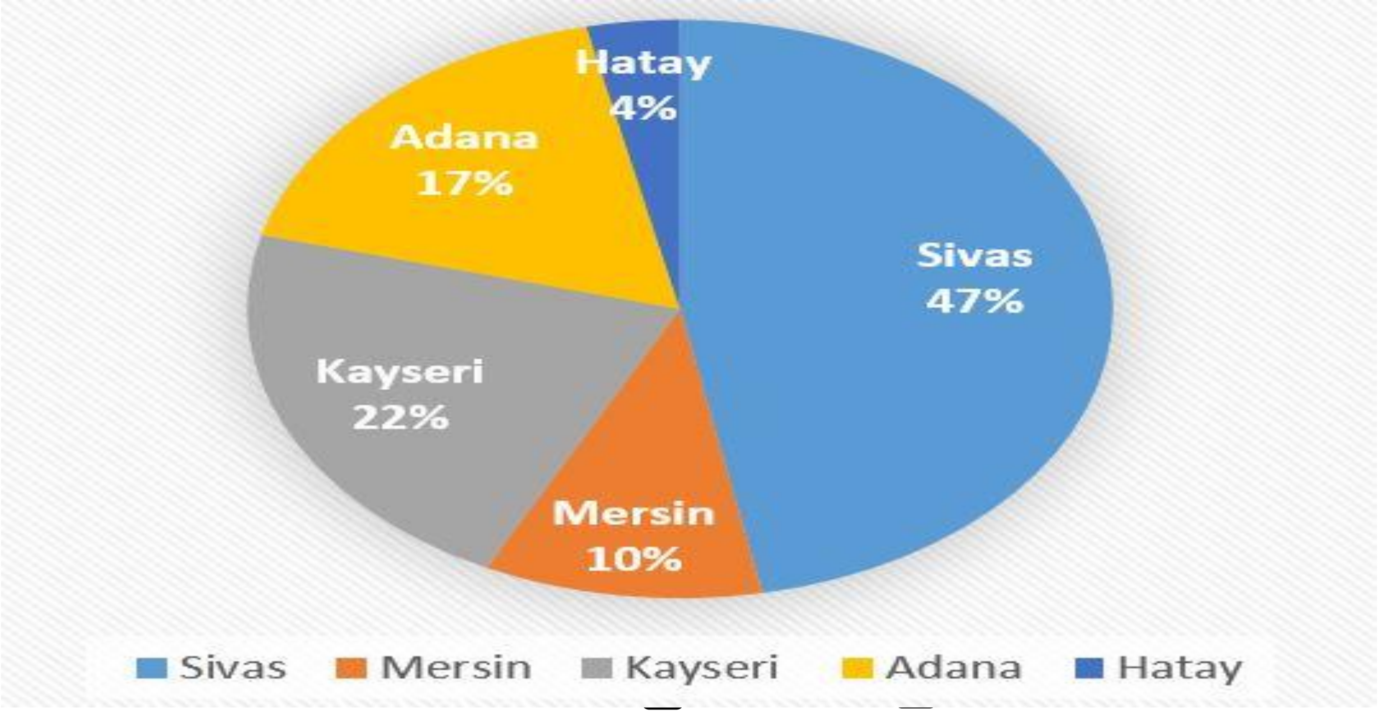
Bu eğriye poztif denmesinin nedeni kuyruğunun sayı doğrusunda pozitif yöne doğru uzaması; benzer şekilde aşağıdaki eğriye negatif denmesinin nedeni kuyruğunun negatif yöne doğru uzamasıdır.

Daire Grafiği (Pie Chart)

İlgilendiğimiz değişkenim sembolik türde bir kategorik değişken ise (yani bu değişkenler sözel ve sıralanamıyorsa) bu değişkenin farklı değerlerinin veride ne sıklıkta görüldüğünü görselleştirmek için daire grafiği (pie chart) çizeriz.

ör. Diyelimki bir bölümünde 54 kişi Sivas'lı, 12 kişi Mersin'li, 25 kişi Kayseri'li, 20 kişi Adana'lı ve 4 kişi Hatay'lı olsun.

Bu veriden elde edilen daire grafiği bölümdeki kişilerin memleketlerini ve bu memleketlerin görülme yüzdelerini verir:



Merkezi Eğilim (Central Tendency)

Bir veride değişkenin merkezinin nerede olduğunu hesaplayabiliriz. Bu merkez veride en çok görülen değerdir ve bütün veriyi özetlerken kullanacağımız iki büyüklükten biridir (diğeri varyasyon).

Merkezi eğilim 3 şekilde hesaplanabilir: Ortalama, Medyan ve Mod.

1. Ortalama (Mean - Average)

En bilindik merkez hesaplama yöntemidir. Basitçe tüm değerler toplanır; toplam kaç tane değer varsa o sayıya bölünür.

Formül olarak: Verideki N tane değişkenimizi x_i $(i \in \{1, \ldots, N\})$ şeklinde gösterelim.

Şu halde ortalama değer:

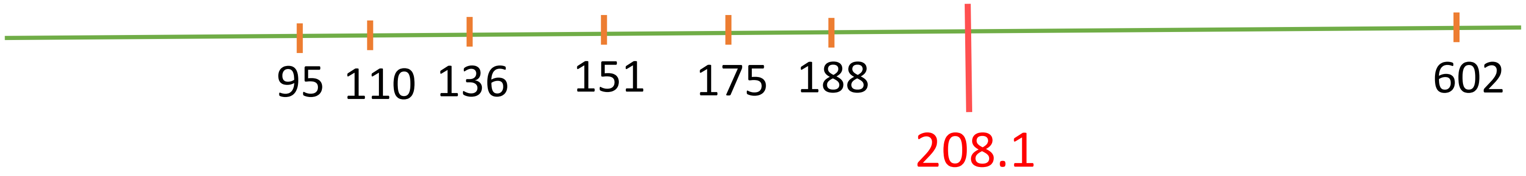
$$\boxed{\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}}$$

olur.

Ortalama (Mean - Average)

ör. 7 kişilik bir gruptaki kişilerin takipçi sayıları: 95, 110, 136, 151, 175, 188, 602 olsun. Bu gruptaki kişilerin ortalama takipçi sayısı:

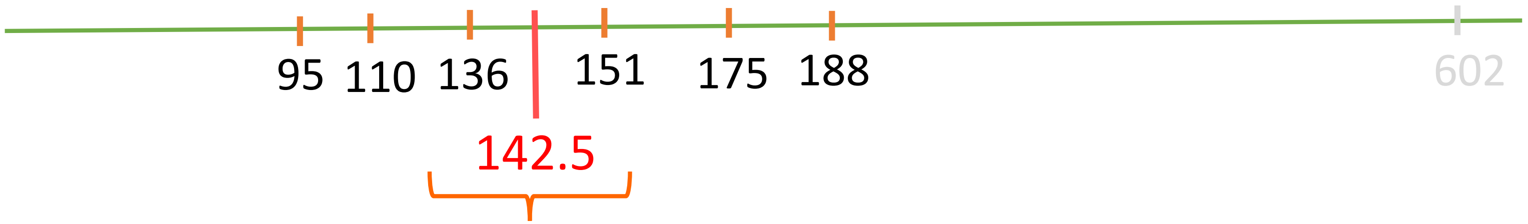
$$\bar{x} = \frac{95 + 110 + 136 + 151 + 175 + 188 + 602}{7} = 208.1$$



Ortalama (Mean - Average)

ör. Takipçi sayıları: 95, 110, 136, 151, 175, ve 188 olsun. Bu gruptaki kişilerin ortalama takipçi sayısı:

$$\bar{x} = \frac{95 + 110 + 136 + 151 + 175 + 188}{6} = 142.5$$



Ortalama gibi ortalama!

Ortalama (Mean - Average)

ör. 7 kişilik bir gruptaki kişilerin takipçi sayıları: 95, 110, 136, 151, 175, 188, 602 olsun. Bu gruptaki kişilerin ortalama takipçi sayısı:

$$\bar{x} = \frac{95 + 110 + 136 + 151 + 175 + 188 + 602}{7} = 208.1$$

olur. Fakat dikkat edilirse bulunan ortalama değer verinin tamamını özetlemekten biraz uzaktır. Çünkü verinin yaklaşık %85'i (7 değerin 6'sı) [95 – 188] aralığındadır ama bulunan ortalama değer bu aralık içersinde yer almaz. Bu, uç (extreme) bir değer olan 602'nin ortalamayı epey yukarı çekmesinden dolayı olşmuşur. Yani aşırı popüler kişi ortalamayı alt üst etmiştir.

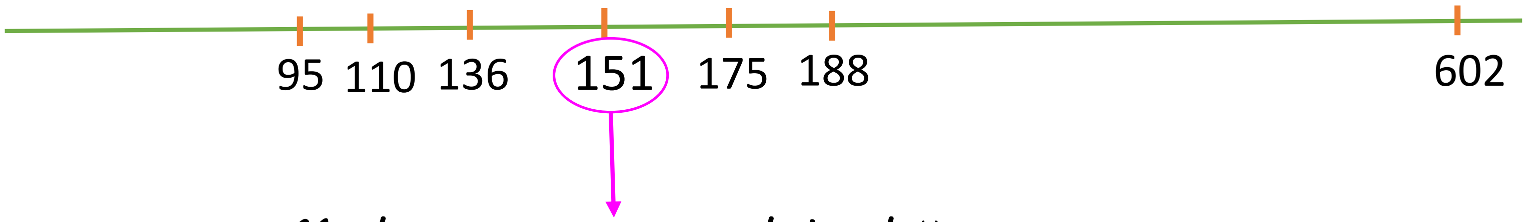
Sonuç: Ortalama hesabı, verideki uç değerlerden çokca etkilenir! Bundan kaçınmak için, istisna değerler olan uç değerler veriden çıkartılıp, ortalama yeniden hesaplanabilir:

$$\bar{x} = \frac{95 + 110 + 136 + 151 + 175 + 188}{6} = 142.5$$

2. Medyan (Ortadaki Değer)

ör. 7 kişilik bir gruptaki kişilerin takipçi sayıları: 95, 110, 136, 151, 175, 188, 602 olsun. Bu gruptaki kişilerin medyanını bulalım.

$$\bar{x} = \frac{95 + 110 + 136 + 151 + 175 + 188 + 602}{7} = 208.1$$



Medyan: tam ortadaki değer

2. Medyan (Median)

Verinin merkezini bulmanın bir başka yolu medyanı hesaplamaktır. Basitçe medyan, verideki değerler küçükten büyüğe sıralandığında ortadaki değerdir.

ör 95, 110, 136, 151, 175, 188, 602 takipçi sayılarının medyanı:

$$\underbrace{95,\ 110,\ 136}_{3\ \text{değer}},\ 151\ \underbrace{175,\ 188,\ 602}_{3\ \text{değer}}$$

Medyan

Medyan, ortalamaya göre daha az uç değerlerden etkilenir. Öte yandan medyan yalnızca sıralamaya dayandığından, medyan hesabında verideki değerleri tam olarak kullanıyoruz diyemeyiz.

17, 45, 148, 151 ,270, 612, 1780

Medyan

Yukarıdakinden çok farklı bu verinin de medyanı 151'dir!

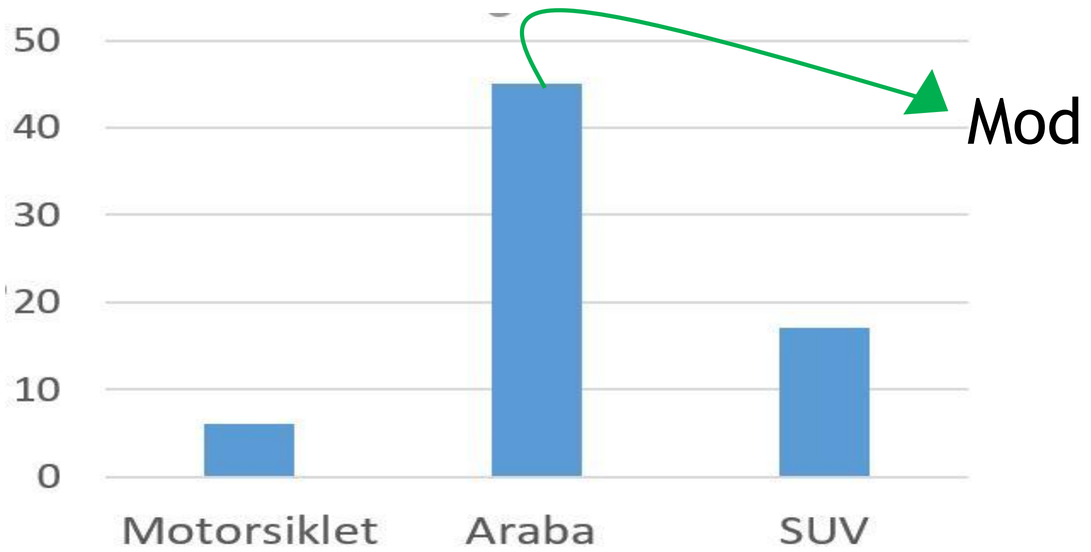
3. Mod (Mode)

Mod, veride en sık görülen değerdir. Verinin merkezini hesaplarken kullanabiliriz.

Değişkenimiz sayısal türde ise, aynı değerin birden fazla görülmesi pek mümkün olmaz. Genelde tüm değerler bir kez görülmüş olur. Örneğin bir önceki örnekte herkesin takipçi sayılası farklıdır. Şu halde en çok görülen değerden pek söz edemeyiz.

Öte yandan diyelimki değişkenimiz kategorik olsun. Bu durumda verinin merkezi için mod hesabını kullanabiliriz.

ör. Diyelimki bir garajdaki araçların 45'i araba, 17'si SUV ve 6'sı motorsiklet olsun. Bu durumda bu verideki araç değişkenin modu araba olur.



ör. 1976-1977 yıllarında Los Angles'ta 770 tane kayitlara gecen motorsiklet kazasi olmuştur. Bu kazalarin 331'inde sürücüler kask takmış, 439'unda ise sürücüler kask takmamıştır. Tablo 1, kazanın önem derecesini ve derecelere karşılık gelen yorumları göstermektedir; Tablo 2 kask takanların ve takmayanların kaza derecelerine göre dağılımlarını göstermektedir.

Kaza Derecesi	Yorum
0	Kafa travması yok
1	Küçük kafa travması
2	Orta derece kafa travması
3	Ciddi, hayat tehlikesi yok
4	Ciddi, hayat tehlikesi var
5	Kritik, her an ölebilir
6	Ölümcül

Tablo 1

Kaza Derecesi	Kask Takanlar	Kask Takmayanlar
0	248	227
1	58	135
2	11	33
3	3	14
4	2	3
5	8	21
6	1	6

Tablo 2

Yukarıdaki tablolara göre, kask takanlarla kask takmayanlar arasında kazanın yol açtığı sonuç bakımından bir fark var mıdır?

Çözüm.

Burada basitçe kaza derecesinin kask takanlar için ortalaması ile kask takmayanlar için ortalamasını bulup, bu iki ortalamayı kıyaslayabiliriz.

Kask takanların 248'inin derecesi 0. O Burdan gelen toplam puan: $248 \times 0 = 0$

Kask takanların 58'inin derecesi 1. O Burdan gelen toplam puan: $58 \times 1 = 58$

Kask takanların 11'inin derecesi 2. O Burdan gelen toplam puan: $11 \times 2 = 22$

Kask takanların 3'ünün derecesi 3. O Burdan gelen toplam puan: $3 \times 3 = 9$

Kask takanların 2'sinin derecesi 4. O Burdan gelen toplam puan: $2 \times 4 = 8$

Kask takanların 8'inin derecesi 5. O Burdan gelen toplam puan: $8 \times 5 = 40$

Kask takanların 1'inin derecesi 6. O Burdan gelen toplam puan: $1 \times 6 = 6$

$+$ ______

143

Tüm puanlar toplanıp, bu toplam kask takan toplam kişi sayısına bölünürse:

kask takanların ortalama kaza derecesi: $\frac{143}{331} = 0.432$

Benzer şekilde kask takmayanlar için de kaza dereceleri toplanıp, bu toplam kask takmayan toplam kişi sayisina bölünürse

kask takmayanların ortalama kaza derecesi: $\frac{396}{331} = 0.902$

olur. Görülüyorki kask takmayanların ortalam kaza derecesi 2 kat daha fazla, yani kask takmayanların sağlığı iki kat fazla tehlikededir.

ödev. Bir zar 30 kez atılıyor. Aşağıdaki tablo bu 30 atışta her bir sayının kaç defa geldiğini göstermektedir. Buna göre 30 atışta gelen sayıların, ortalaması, medyanı ve modu kaçtır?

Değer	Sıklık
1	6
2	4
3	5
4	8
5	3
6	4

Dağılımın Yayılmasını Hesaplamak

Bir verinin merkezini hesaplamak, bu veri hakkında tam bir bilgi edinmeye yetmez. Yani merkez, veriyi tek başına özetleyemez. Merkez hesabının yanında verinin merkezi etrafında ne kadar yayıldığını (dağıldığını) da hesaplamamız gerekir.

Örnek olarak şu iki veri setini ele alalım.

$$A:\ 1\,,2,5,6,6\,; \qquad B:\ -40,0,5,20,35$$

Iki setinde ortalaması aynıdır: 4; fakat B setinde değerler ortalamanın etrafındaki yayılması fazladır, yani bu sette varyasyon fazladır.

Dağılımın merkezden yayılmasını hesaplamanın en doğal yolu, her bir değerin merkezden ne kadar saptığını hesaplayıp, bu sapmaların ortalamasını hesaplamaktır.

Formul olarak, N tane değerimiz olsun ve biz herbirini x_i $(i \in \{1, \dots, N\})$ şeklinde gösterilim. $\bar{x}$ bu N değerin ortalması olmak üzere **varyasyon:**

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

Varyasyon

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

Burada değerlerin merkezden sapma miktarlarının karesinin alınmasının nedeni pozitif sapmalar ile negatif sapmaların birbirini götürmesini engellemektir.

Standart Sapma

Varyasyonun kareköküdür: σ .

ör. $1, 2, 5, 6, 6$ değişkenlerinin varyasyonu:

Bu 5 değerin ortalması: 4'tür. Şu halde varyasyon:

$$\sigma^2 = \frac{1}{4}(1-4)^2 + (2-4)^2 + (5-4)^2 + (6-4)^2 + (6-4)^2 = 5.5$$

ör. $-40, 0, 5, 20, 35$ değişkenlerinin varyasyonu:

Bu 5 değerin ortalması: 4'tür. Şu halde varyasyon:

$$\sigma^2 = \frac{1}{4}(-40-4)^2 + (0-4)^2 + (5-4)^2 + (20-4)^2 + (35-4)^2 = 792.5$$

Teorem 1: Bir veri setindeki değişkenlerin herbirine aynı sabit sayı eklenirse; bu değişkenlerin ortalaması da o sayı kadar artar.

Kanıt: $x_1, \ldots, x_N$ değişkenlerinin ortalaması $\bar{x}$ olsun. Yani $\bar{x} = \frac{x_1 + \cdots + x_N}{N}$

$x_1, \ldots, x_N$ değişkenlerinin herbirine bir $c \in \mathbb{R}$ sabiti ekleyelim. Bu durumda yeni ortalama:

$$\frac{(x_1 + c) + \cdots + (x_N + c)}{N} = \frac{x_1 + \cdots + x_N + Nc}{N} = \frac{x_1 + \cdots + x_N}{N} + c = \bar{x} + c$$

Teorem 2: Bir veri setindeki değişkenlerin herbirine aynı sabit sayı eklenirse; bu değişkenlerin varyansı değişmez.

Kanıt: $x_1, \ldots, x_N$ değişkenlerinin varyansı σ^2 olsun. Yani $\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$.

$x_1, \ldots, x_N$ değişkenlerinin herbirine bir $c \in \mathbb{R}$ sabiti eklenirse Teorem 1 gereği yeni ortalama $\bar{x} + c$ olur.

Yeni varyasyon: $\frac{\left(x_1 + c - (\bar{x} + c)\right)^2 + \cdots + \left(x_N + c - (\bar{x} + c)\right)^2}{N-1} = \frac{(x_1 + c - \bar{x} - c)^2 + \cdots + (x_N + c - \bar{x} - c)^2}{N-1} = \sigma^2$

Bu teoremlerin uygulaması olarak şunları düşünebiliriz.

Diyelimki sınıftaki vize notlarının ortalaması 42, varyasyonu 7.8 olsun.

Eğer sınıftaki herkese 10 puan eklersem ortalama 10 artar 52 olur; fakat varyasyon yine aynı 7.8 olarak kalır. Çünkü varyasyon çeşitlilik demektir. Herekese verilen 10 puan sınıfataki çeşitliliği artırmaz yada çesitliliği azaltmaz; çeşitlilik miktari aynı kalır.

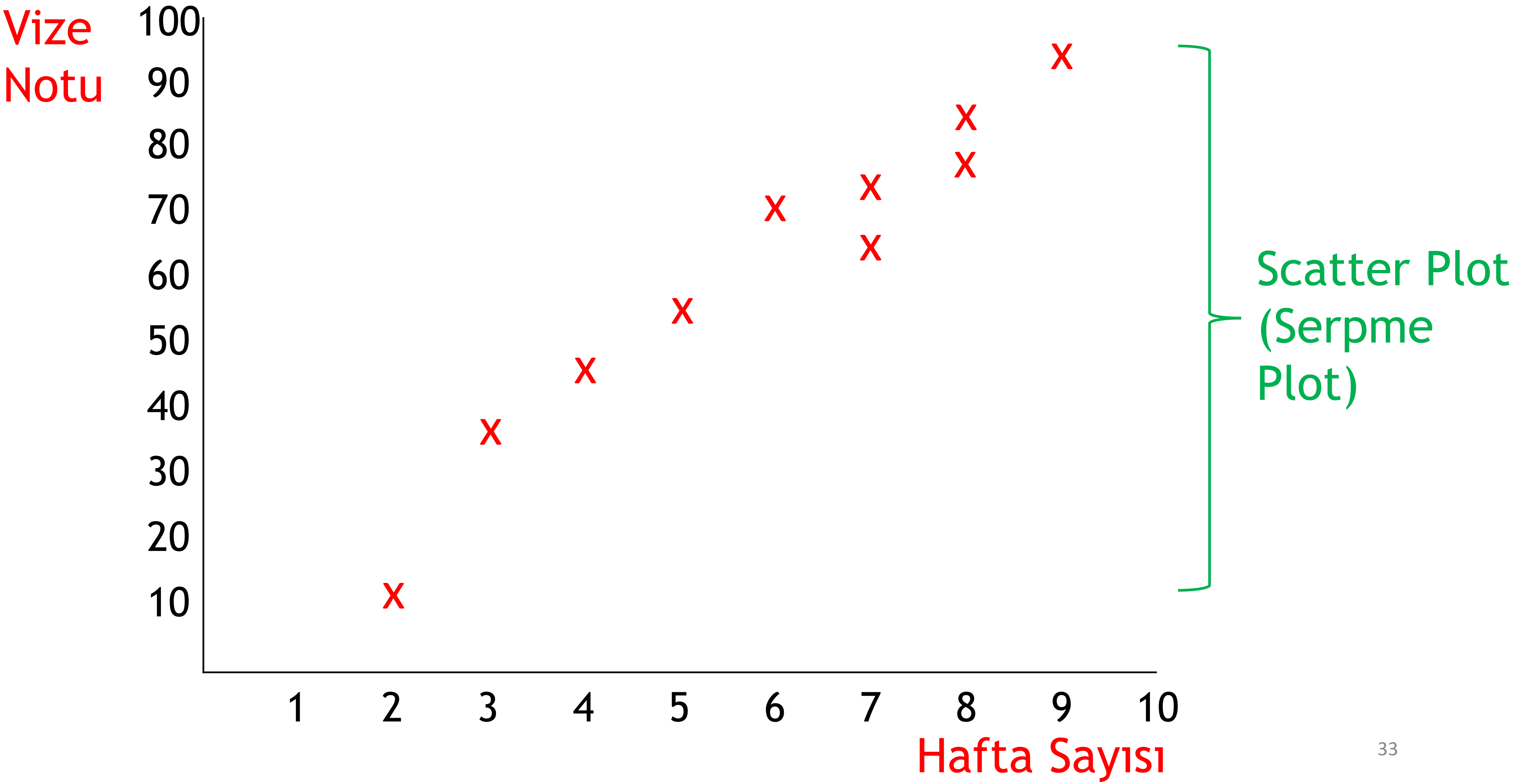
Korelasyon (Correlation)

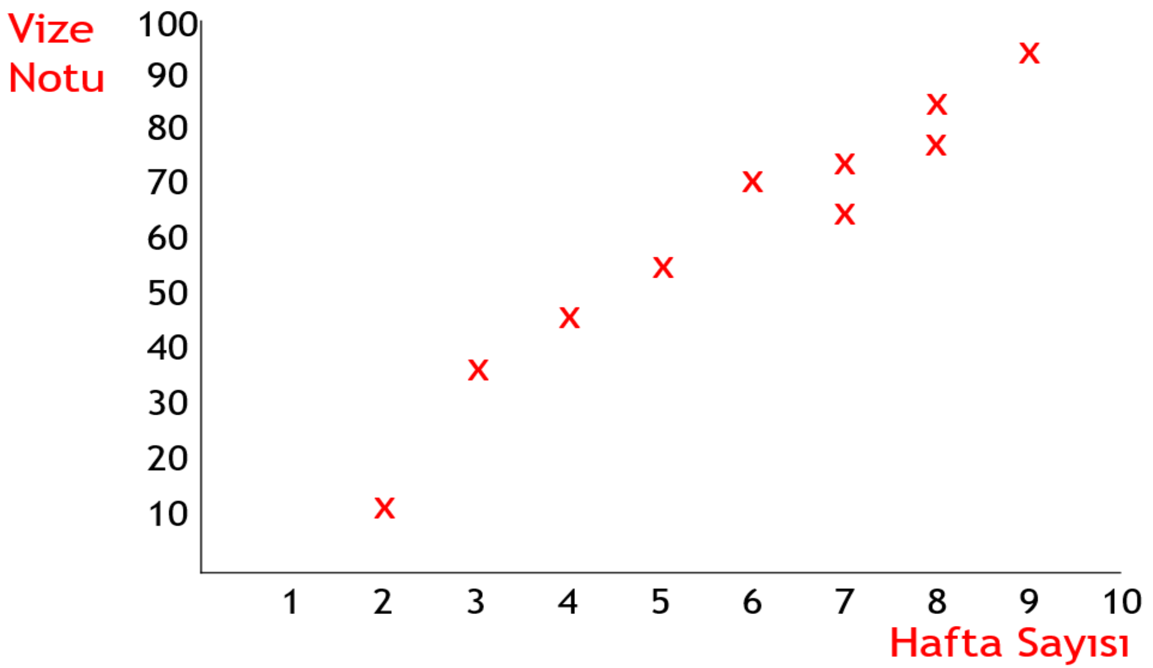
Şimdiye kadar hep tek bir değişken ile ilgilendik (bir değişkenin ortalaması, varyansı…). Şimdi ise diyelimki verimizde birden çok değişken var ve bu değişkenler arasında bir ilişki var mı yok mu diye merak ediyoruz.

Örneğin derse katılınan hafta sayısı ve vize notları arasında bir ilişki var mi diye merak ediyoruz. Bunun için 10 kişiden aşağıdaki veriyi topluyoruz.

	Hafta Sayısı	Vize Notu
1	Hafta Sayısı	Vize Notu
2	5	55
3	3	37
4	2	11
5	8	84
6	9	94
7	7	65
8	7	74
9	4	46
10	6	71
11	8	78
12		

Bu veriyi yataydaki değişken hafta sayısı, dikeydeki değişken vize notu olacak şekilde görselleştirirsek:



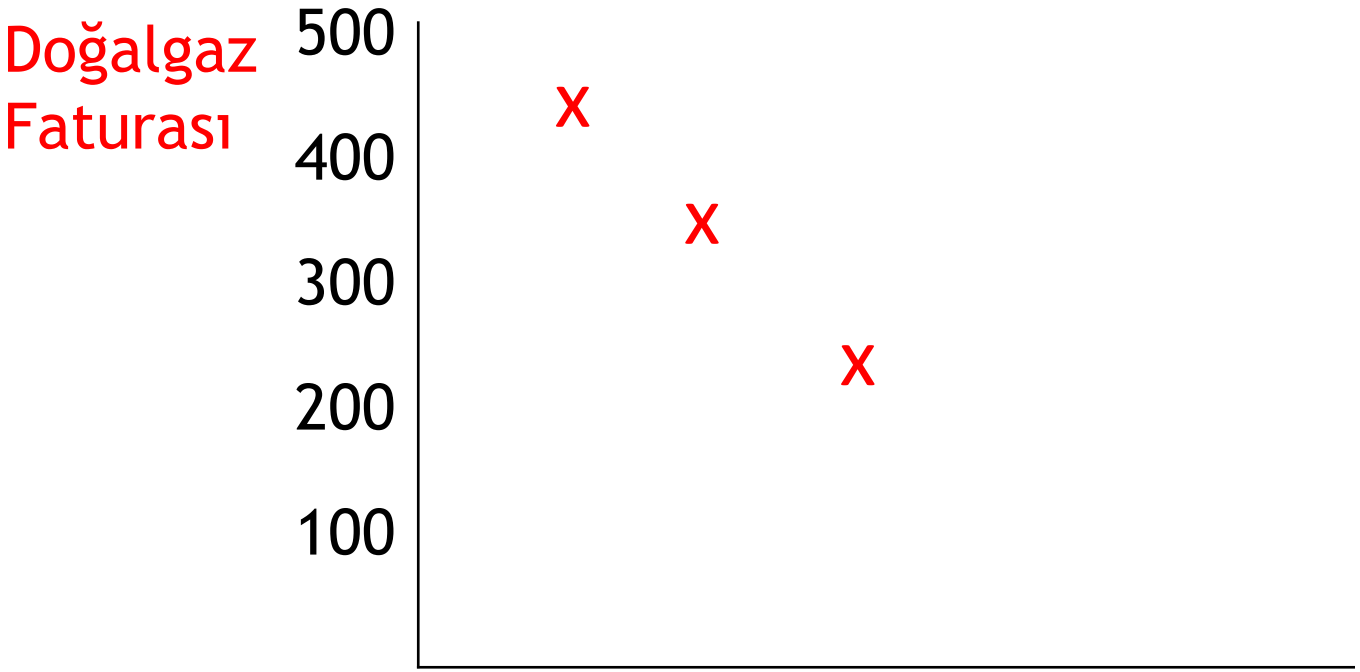


Bu plotta, derse girilien hafta sayısı ile alınan vize notu arasında pozitif bir ilişki görülmektedir. Hafta sayısı arttıkça vize notu artar; hafta sayısı azaldıkça vize notu azalır (iki değişken birlikte hareket ederler). Bu durumda bu iki değişken arasında pozitif korelasyon vardır diyeceğız.

ör.Aşağıdaki tablo aylık ortalama sıcaklık değerlerini ve bu aylardaki doğalgaz faturlarını göstermektedir.

Ortalama Sıcaklık	Doğalgaz Faturası
5	440
9	350
14	220
21	115
26	67

Bu veriyi yataydaki değişken aylık sıcaklık değerleri, dikeydeki değişken doğalfaz faturası olacak şekilde görselleştirirsek:



Bu scatter plotta, aylık ortalama sıcaklık doğalgaz faturasının düştüğü görülür. Burada bir değişken arttıkça diğer değişken azalır (yada bir değişken azalırken diğeri artar). Bu durumda iki değişken arasinda **negatif korelasyon** vardır diyeceğiz.

Korelasyonun Hesaplanması (Korelasyonun Sayısal Olarak Ölçülmesi)

Elimizdeki veri $(x_1, y_1), \ \dots, (x_N, y_N)$ çiftlerinden oluşşsun.

$x_1, \dots, x_N$ değişkenlerinin ortalaması $\bar{x}$; $y_1, \dots, y_N$ değişkenlerinin ortalaması $\bar{y}$ olsun.

(x_i, y_i) çiftini ele alalım.

$(x_i - \bar{x})$, x_i değişkeninin kendi merkezi olan $\bar{x}$ 'den sapmasını verir.

$(y_i - \bar{y})$, y_i değişkeninin kendi merkezi olan $\bar{y}$ 'den sapmasını verir.

$$(x_i - \bar{x})(y_i - \bar{y})$$

çarpımında eğer:

1. $x_i > \bar{x}$ ve $y_i > \bar{y}$ ise bu çarpım pozitif olur (bu durumda her iki değer de merkezinden daha büyük)

Yada

2. $x_i < \bar{x}$ ve $y_i < \bar{y}$ ise bu çarpım yine pozitif olur (bu durumda her iki değer de merkezinden daha küçük).

Korelasyonun Hesaplanması (Korelasyonun Sayısal Olarak Ölçülmesi)

Sonuç olarak iki durumda da; iki değişken benzer hareket ediyorlar (ya ikisi birden merkezlerinden daha büyük, ya ikisi birden merkezlerinden daha küçük). O halde (x_i, y_i) çifti benzerdir; ve x ve y değişkenlerinin korelasyonuna pozitif katkıda bulunur.

Tersi olarak

$$(x_i - \bar{x})(y_i - \bar{y})$$

çarpımında eğer:

1. $x_i > \bar{x}$ ve $y_i < \bar{y}$ ise bu çarpım negatif olur olur (bu durumda birinci değişken merkezinden büyük iken ikinci değişken merkezinden küçüktür; bu iki değişken arasında uyumsuzluk vardır)
2. $x_i < \bar{x}$ ve $y_i > \bar{y}$ ise bu çarpım yine negatif olur olur (bu durumda birinci değişken merkezinden küçük iken ikinci değişken merkezinden büyüktür; bu iki değişken arasında yine uyumsuzluk vardır)

Korelasyonun Hesaplanması (Korelasyonun Sayısal Olarak Ölçülmesi)

O halde

$$(x_i - \bar{x})(y_i - \bar{y})$$

çarpımı negatif iken (x_i, y_i) çifti benzer değildir; ve x ve y değişkenlerinin korelasyonuna negatif katkıda bulunur.

$(x_1, y_1), \ \ldots, (x_N, y_N)$ çiftlerinin x ve y değişkenlerinin korelasyonuna olan katkılarının ortalamasını alırsak:

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

Bu değere **kovaryans** denir σ_{xy} ile gösterilir. Fakat kovaryans değeri $(-\infty, +\infty)$ arasında bir değerdir. Yani herhangi bir sınırlanması yoktur. Veri setinden veri setine çok farklılık gösterebilir.

Elde ettigimiz benzerlik değerinin herkesçe anlaşılır olması için, bir anlam ifade etmesi için standart bir aralık olan $[-1\,, 1]$ aralığında yer alması gerekir. Bunun için kovaryansı x'in ve y'nin standart sapmalarına böleceğiz.

Korelasyonun Hesaplanması (Korelasyonun Sayısal Olarak Ölçülmesi)

x'in standard sapması σ_x :

$$\sigma_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

y'nin standard sapması σ_y :

$$\sigma_y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Kovaryans σ_{xy}, yukarıdaki standart sapmaların çarpımına bölünürse:

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

Korelasyonun Hesaplanması (Korelasyonun Sayısal Olarak Ölçülmesi)

Bu ifade sadeleştirilerek korelasyon hesağlanır:

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

Bu değere Pearson Korelasyon Katsayısı diyeceğiz. Bu değer her zaman $[-1, 1]$ arasındadır.

Eğer x ve y değişkenleri arasında negatif korelasyon varsa korelasyon negatiftir $[-1, 0)$ arası değer alır;

eğer x ve y değişkenleri arasında pozitif korelasyon varsa korelasyon pozitiftir $(0, 1]$ arası bir değer alır;

eğer x ve y değişkenleri arasında herhangi bir korelasyon yoksa, korelasyon 0’dır.

ör. Daha önce gördüğümüz hafta – sayısı vize notu arasındaki korelasyonu hesaplayalım.

Hafta Sayısı	Ortalamdan Sapma	Sapmanın Karesi	Vize Notu	Ortalamdan Sapma	Sapmanın Karesi	Sapmaların Çarpımı
5	(5-5.9)=-0.9	0.81	55	(55-61.5)=-6.5	42.25	-0.9 ·-6.5=5.85
3	(3-5.9)=-2.9	8.41	37	(37-61.5)=-24.5	600.25	-2.9·-24.5=71.05
2	(2-5.9)=-3.9	15.21	11	(11-61.5)=-50.5	2550.25	-3.9·-50.5=196.95
8	(8-5.9)=2.1	4.41	84	(84-61.5)=22.5	506.25	2.1·22.5=47.25
9	(9-5.9)=3.1	15.61	94	(94-61.5)=32.5	1056.25	3.1·32.5=100.75
7	(7-5.9)=1.1	1.21	65	(65-61.5)=3.5	12.25	1.1·3.5=3.85
7	(7-5.9)=1.1	1.21	74	(74-61.5)=12.5	156.25	1.1·12.5=13.75
4	(4-5.9)=-1.9	3.61	46	(46-61.5)=-15.5	240.25	-1.9·-15.5=29.45
6	(6-5.9)=0.1	0.01	71	(71-61.5)=9.5	90.25	0.1·9.5=0.95
8	(8-5.9)=2.1	4.41	78	(78-61.5)=16.5	272.25	2.1·16.5=34.65
Ortalama: 5.9		Toplam: 48.9	Ortalama: 61.5		Toplam: 5526.5	Toplam: 504.5

Şu halde korelasyon $\rho = \frac{504.5}{\sqrt{48.9 \cdot 5526.5}} = 0.97$ (çok yüksek pozitif korelasyon)

ör. Daha önce gördüğümüz ortalama sıcaklık – doğalgaz faturası değişkenlerinin korelasyonuna bakalım.

x_i $(i \in \{1, \ldots, 5\})$ değişkenleri sıcaklıkları göstersin. Bu değişkenlerin ortalaması: $\bar{x} = 15$

Bu değişkenlerin ortalamadan farklarının karelerinin toplamı:

$$(5-15)^2 + (9-15)^2 + (14-15)^2 + (21-15)^2 + (26-15)^2 = 294$$

y_i $(i \in \{1, \ldots, 5\})$ değişkenleri doğalgaz faturalarını göstersin. Bu değişkenlerin ortalaması: $\bar{y} = 238.4$

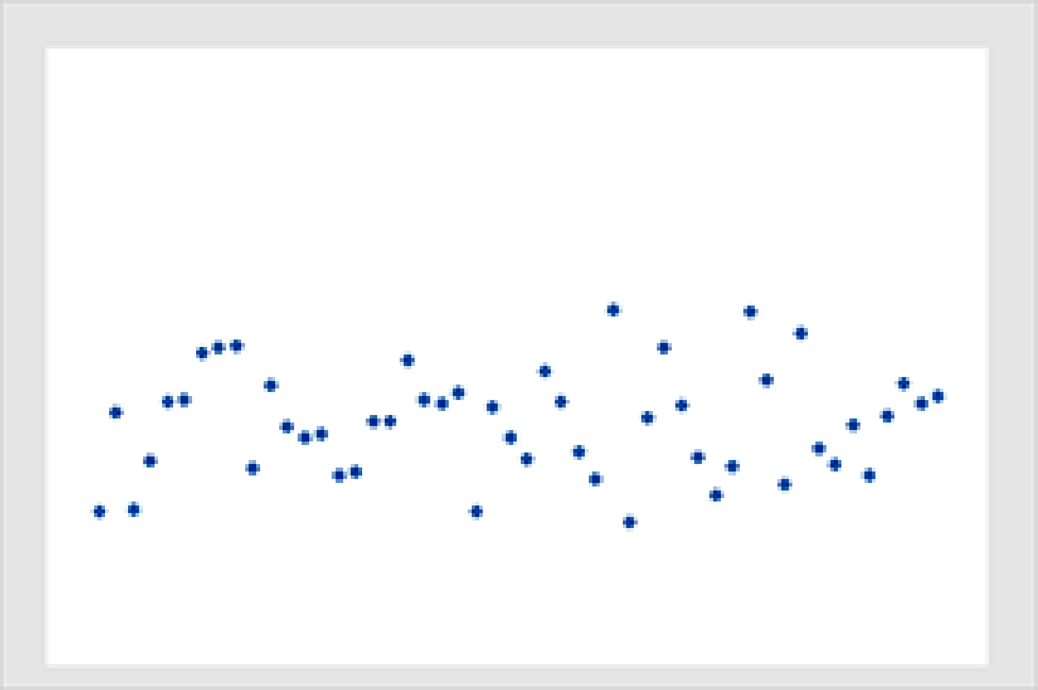
Bu değişkenlerin ortalamadan farklarının karelerinin toplamı:

$$(440-238.4)^2 + (350-238.4)^2 + (220-238.4)^2 + (115-238.4)^2 + (67-238.4)^2 = 98041.2$$

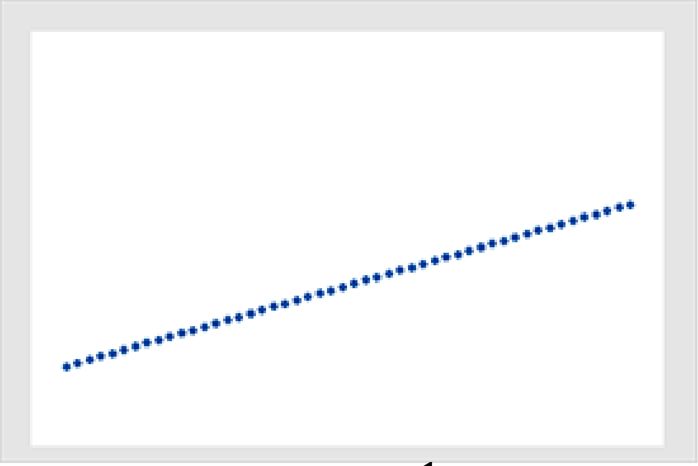
$$(5-15)(440-238.4)+(9-15)(350-238.4)+(14-15)(220-238.4)$$
$$+(21-15)(115-238.4)+(26-15)(67-238.4)=-5293$$

Korelasyon $\rho = \frac{-5293}{\sqrt{294 \cdot 98041.2}} = -0.98$ (çok yüksek negatif korelasyon)

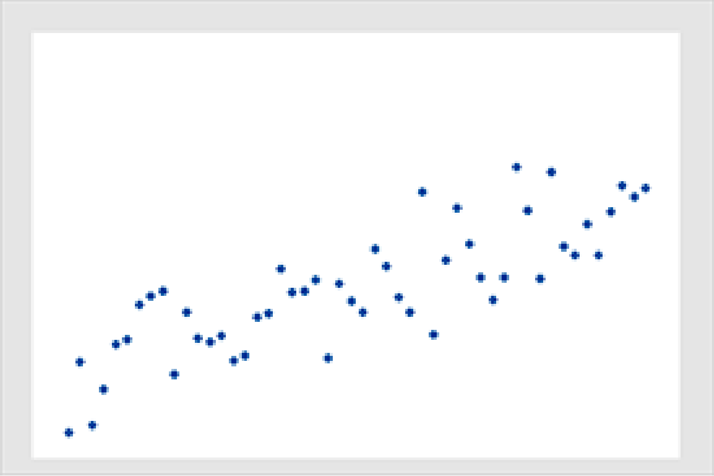
ör. Aşağıda gösterilen scatter plotta iki değişken arasında herhangi bir korelasyon yoktur ($\rho = 0$).



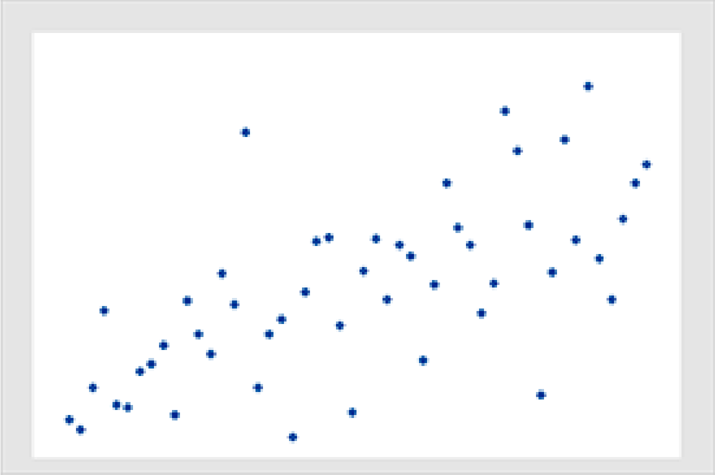
Farklı Büyüklükte Korelasyonlar



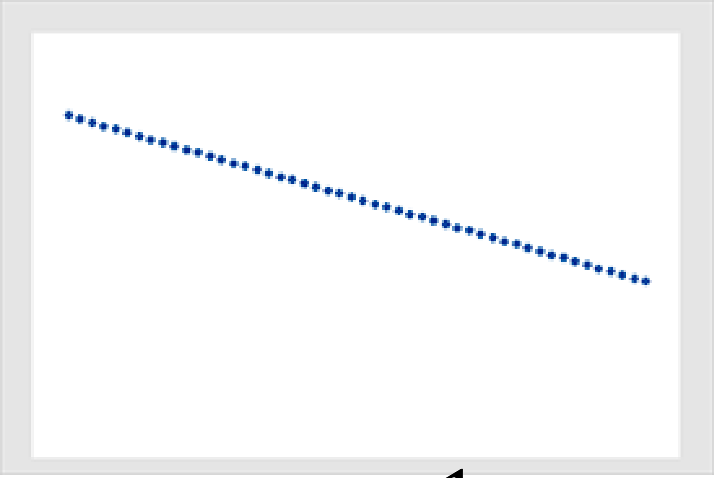
$\rho = 1$



$\rho = 0.8$

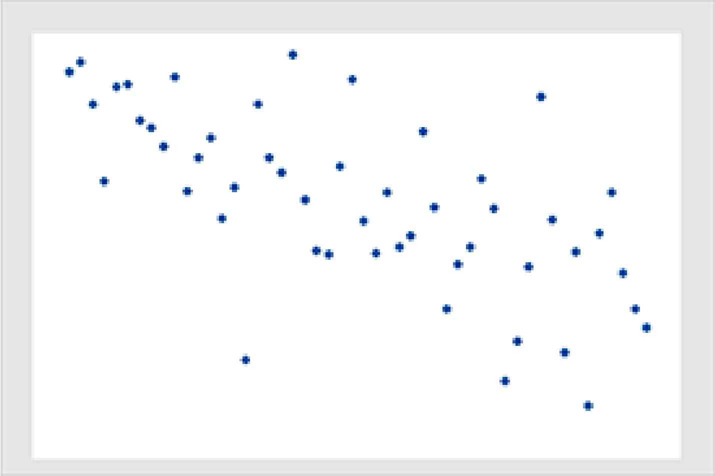


$\rho = 0.6$



$\rho = -1$

$\rho = -0.8$



$\rho = -0.6$