IST108
OLASILIK VE İSTATİSTİK

BETİMLEYİCİ İSTATİSTİK

İçerik

Veri Kümelerinin Tanımlanması
Sıklık Tabloları ve Grafikleri
Göreli Sıklık Tabloları ve Grafikleri
Veri Gruplama
Histogram
Birikimli Sıklık Grafiği
Birikimli Göreli Sıklık Grafiği
Kök Yaprak Gösterimi
Veri Kümelerinin Özetlenmesi

İçerik

Örnek Ortalaması
Örnek Ortancası
Örnek Ortalaması ve Örnek Ortancası
Örnek Tepedeğeri
Örnek Varyansı
Örnek Standart Sapması
Chebyshev Eşitsizliği
Eşleştirilmiş Veri Kümeleri
Örnek Korelasyon Katsayısı

Veri Kümelerinin Tanımlanması

Bir çalışmadan sonuç elde edebilmek için veri toplamamız gerekir.

İstatistik
- Verilerin toplanması
- Verilerin betimlenmesi
- Sonuç çıkarım için veri analizi

Yığın denilen genel bir grup hakkında bilgi edinmek isteriz.
- Çok büyük olması durumunda **örnek** denilen bir alt grup seçilir ve genel hakkında bilgi öğrenilmeye çalışılır.

Veri Kümelerinin Tanımlanması

Bir çalışmanın sonuçları, inceleyen kişinin hızlı bir şekilde algılayabilmesini sağlayacak şekilde sunulmalıdır.

Tablolar ve grafikler

- Verinin açıklığı
- Yoğunlaşma derecesi
- Simetrikliği

Sıklık Tabloları ve Grafikleri

Verileri sunmanın en etkin yollarından bir tanesi tablolar kullanmaktır.

En yaygın kullanılan tablo türlerinden bir tanesi sıklık (frekans) tablosudur.

Az sayıda farklı değerlere sahip bir veri kümesi bir **sıklık (frekans) tablosu** ile sunulabilir.

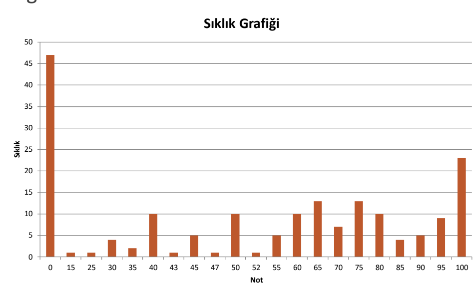
Sıklık Tabloları ve Grafikleri

Yan tarafta bir ödev notuna ait sıklık tablosu görülmektedir.

Not	Sıklık
0	47
15	1
25	1
30	4
35	2
40	10
43	1
45	5
47	1
50	10
52	1
55	5
60	10
65	13
70	7
75	13
80	10
85	4
90	5
95	9
100	23

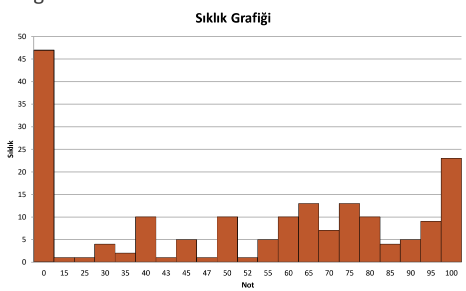
Sıklık Tabloları ve Grafikleri

Çizgi Grafiği



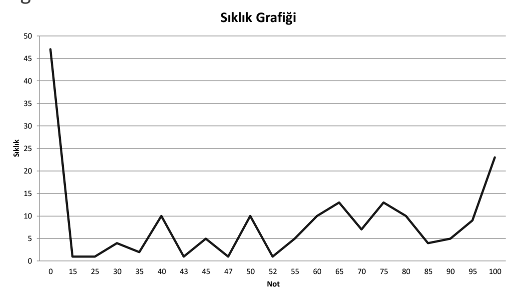
Sıklık Tabloları ve Grafikleri

Sütun Grafiği



Sıklık Tabloları ve Grafikleri

Sıklık Poligonu



Göreli Sıklık Tabloları ve Grafikleri

n adet veri içeren bir veri kümesi ele alalım. f belirli bir değerin sıklık (frekans) bilgisi ise $\frac{f}{n}$ göreli sıklık (frekans) olarak adlandırılır.

Bir veri değerine ait göreli sıklık (frekans), ilgili değere sahip verilerin sayısının toplam veri sayısına oranıdır.

Göreli Sıklık Tabloları ve Grafikleri

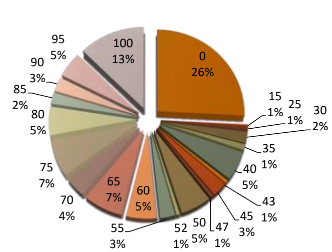
Örnek bir **göreli sıklık tablosu**

yan tarafta görülmektedir.

Not	Sıklık	Göreli Sıklık
0	47	26%
15	1	1%
25	1	1%
30	4	2%
35	2	1%
40	10	5%
43	1	1%
45	5	3%
47	1	1%
50	10	5%
52	1	1%
55	5	3%
60	10	5%
65	13	7%
70	7	4%
75	13	7%
80	10	5%
85	4	2%
90	5	3%
95	9	5%
100	23	13%

Göreli Sıklık Tabloları ve Grafikleri

Örnek bir **göreli sıklık grafiği** aşağıda görülmektedir.



Veri Gruplama

Farklı değer sayısı çok fazla ise yukarıdaki yaklaşımlar çok uygun olmayabilir.

Bu gibi durumlarda değerleri gruplara veya sınıf aralıklarına bölmek daha faydalı olacaktır. Seçilen aralık sayısı önemlidir:

- Az sayıda aralık: Bilgi kaybı.
- Fazla sayıda aralık: Anlamlı olmayan desen.

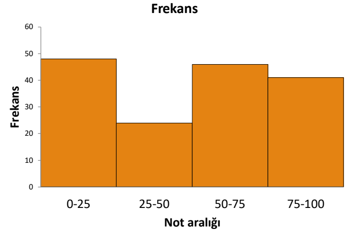
Aralıkların başlangıç ve bitiş noktaları sınıf sınırları olarak ifade edilir. Sol uç dahil gösterimi kullanılacak.

- Örneğin 20-30 aralığı 20 ve 20'den büyük ve 30'dan küçük değerler aralığını gösterir.

Veri Gruplama

Az sayıda aralık seçildiği takdirde bilgi kaybı yaşanır.

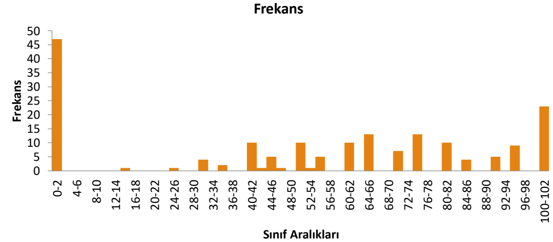
Bir örneği aşağıdaki grafikte gösterilmektedir.



Veri Gruplama

Fazla sayıda aralık seçildiği takdirde anlamlı olmayan desen oluşur.

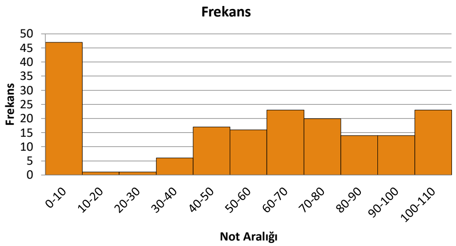
Bir örneği aşağıdaki grafikte gösterilmektedir.



Veri Gruplama

Uygun sayıda aralık seçilmesi önemlidir.

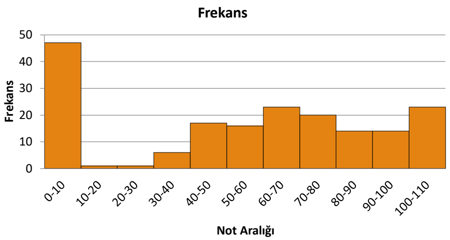
Bir örneği aşağıdaki grafikte gösterilmektedir.



Histogram

Sınıf verilerinin sütun grafiği gösterimine histogram denir.

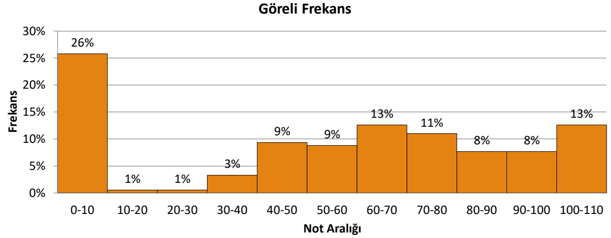
Aşağıdaki grafik bir sıklık (frekans) histogramıdır.



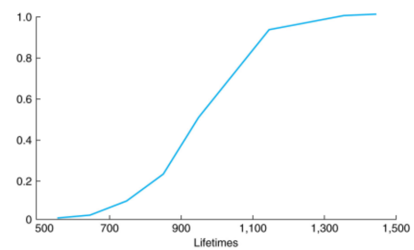
Histogram

Sınıf verilerinin sütun grafiği gösterimine histogram denir.

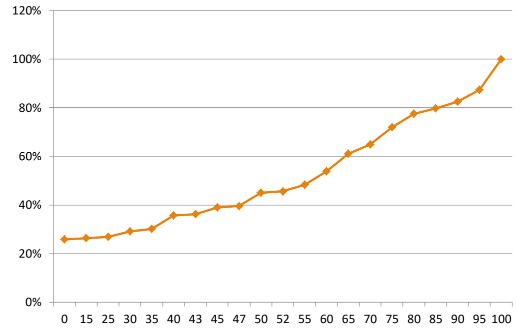
Aşağıdaki grafik bir göreli sıklık (frekans) histogramıdır.



Birikimli Sıklık Grafiği



Birikimli Göreli Sıklık Grafiği



Kök Yaprak Gösterimi

Küçük ve orta ölçekli verileri göstermede kök ve yaprak gösterimi iyi bir yoldur.

Bu gösterim, her bir değeri kök ve yaprak olmak üzere iki parçaya ayırmak ile mümkün olabilir.

Örneğin tüm veri değerleri iki basamaklı ise, onlar basamağı kök ve birler basamağı yaprak olabilir.

Örneğin 62 sayısı için kök 6 ve yaprak 2 olacaktır. Eğer 62 ve 67 sayıları varsa gösterim;

Kök	**Yaprak**
6	2;7

Kök Yaprak Gösterimi

Ödev notlarına ilişkin örnek bir kök yaprak gösterimi aşağıda mevcuttur.

Kök	**Yaprak**
0	0
1	5
2	5
3	0; 5
4	0; 3; 5; 7
5	0; 2; 5
6	0; 5
7	0; 5
8	0; 5
9	0; 5
10	0

Veri Kümelerinin Özetlenmesi

Günümüzde büyük veri kümeleriyle çalışmaktayız.

Büyük bir veri kümesinden anlamlı veriler çıkartılabilmesi için verilerin çeşitli ölçekler kullanılarak özetlenmesi gerekir.

Devam eden yansılarda veri özetlemede kullanılan istatistiklerden bahsedilecektir.

Örnek Ortalaması

n adet sayısal değerden oluşan bir veri kümesi için, örnek ortalaması bu değerlerin aritmetik ortalamasıdır.

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

Örnek ortalamasının hesaplanması genellikle aşağıdaki şekilde basitleştirilebilir.

$y_i = ax_i + b, \quad i = 1, 2, \dots, n$

$\bar{y} = \frac{\sum_{i=1}^{n}(ax_i + b)}{n} = a\bar{x} + b$

Örnek 1

Aşağıdaki veri kümesinin ortalaması nedir?

$280, 278, 272, 276, 281, 279, 276, 281, 289, 280$

Örnek 1

Aşağıdaki veri kümesinin ortalaması nedir?

$280, 278, 272, 276, 281, 279, 276, 281, 289, 280$

Bu veri kümesinin doğrudan ortalamasını hesaplamak yerine aşağıdaki şekilde hesaplamak daha kolay olur.

$y_i = x_i - 280$

$0, -2, -8, -4, 1, -1, -4, 1, 9, 0$

$\bar{y} = -0{,}8$

$\bar{x} = \bar{y} + 280 = -0{,}8 + 280 = 279{,}2$

Örnek Ortalaması

Bazen sıklık tablosunda listelenen $f_1, f_2, \dots, f_k$ sıklıklarına sahip k adet farklı $v_1, v_2, \dots, v_k$ değerinin örnek ortalaması istenebilir. $n = \sum_{i=1}^{k} f_i$ adet veriden oluşan böyle bir veri kümesi için örnek ortalaması

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i v_i}{n}$$

Diğer bir deyişle örnek ortalaması veri kümesindeki farklı değerlerin ağırlıklı ortalamasıdır.

$$\bar{x} = \frac{f_1}{n} v_1 + \frac{f_2}{n} v_2 + \dots + \frac{f_k}{n} v_k$$

Örnek 2

Aşağıdaki tabloda belirli bir gruptaki kişilere ait yaş sıklık bilgileri verilmiştir. Bu grubun yaş ortalaması nedir?

Yaş	Sıklık
15	2
16	5
17	11
18	9
19	14
20	13

Örnek 2

Yaş	Sıklık
15	2
16	5
17	11
18	9
19	14
20	13

$$\bar{x} = \frac{2}{54} \times 15 + \frac{5}{54} \times 16 + \frac{11}{54} \times 17 + \frac{9}{54} \times 18 + \frac{14}{54} \times 19 + \frac{13}{54} \times 20$$

$$\bar{x} = \frac{2\times15+5\times16+11\times17+9\times18+14\times19+13\times20}{54} = 18{,}24$$

Örnek Ortancası

Bir veri kümesinin merkezini göstermek için kullanılan bir diğer istatistik ise örnek ortancasıdır. Örnek ortancası, veri kümesi artan sıra ile dizildiğinde tam ortadaki değerdir.

Örnek ortancasını bulmak için veri kümesindeki değerler küçükten büyüğe sıralanır. Veri kümesinin eleman sayısı n olsun.

- n tek ise, $\frac{(n+1)}{2}$ konumundaki değer örnek ortancasıdır.
- n çift ise, $\frac{n}{2}$ ve $\frac{n}{2}+1$ konumlarındaki değerlerin ortalaması örnek ortancasıdır.

Örnek Ortalaması ve Örnek Ortancası

Hem örnek ortalaması hem de örnek ortancası faydalı istatistiklerdir.

Örnek ortalaması, diğer verilere göre çok büyük veya çok küçük olan uç değerlerden etkilenirken örnek ortancası bu değerlerden etkilenmez.

Sabit bir vergi oranı kullanılan bir yerde vergiden gelecek gelir hesaplanırken, vergi verenlerin gelirlerinin ortalamasını almak daha kullanışlıdır.

Fakat, orta sınıf için apartmanlar yapılacaksa ve bu apartmanların fiyatını ödeyebilecek nüfusun oranı tespit edilmeye çalışılıyor ise bu durumda ortanca daha kullanışlıdır.

Örnek 3

Aşağıdaki tabloda belirli bir gruptaki kişilere ait yaş sıklık grafiği verilmiştir. Bu veriye ait ortanca nedir?

Yaş	Sıklık
15	2
16	5
17	11
18	9
19	14
20	13

Örnek 3

Yaş	Sıklık
15	2
16	5
17	11
18	9
19	14
20	13

Veri kümesinde 54 adet veri mevcuttur.

27. ve 28. verinin ortalaması alınarak ortanca hesaplanır.

$$Ortanca = \frac{18+19}{2} = 18{,}5$$

Örnek 4

İki fare grubunda yapılan yaşam sürelerine ait deney sonuçları aşağıdadır. Yaşam sürelerine ait ortalama ve ortancaları hesaplayın.

1	58, 92, 93, 94, 95
2	02, 12, 15, 29, 30, 37, 40, 44, 47, 59
3	01, 01, 21, 37
4	15, 34, 44, 85, 96
5	29, 37
6	24
7	07
8	00

1	59, 89, 91, 98
2	35, 45, 50, 56, 61, 65, 66, 80
3	43, 56, 83
4	03, 14, 28, 32

Örnek 4

1	58, 92, 93, 94, 95
2	02, 12, 15, 29, 30, 37, 40, 44, 47, 59
3	01, 01, 21, 37
4	15, 34, 44, 85, 96
5	29, 37
6	24
7	07
8	00

1	59, 89, 91, 98
2	35, 45, 50, 56, 61, 65, 66, 80
3	43, 56, 83
4	03, 14, 28, 32

1. grup: $\bar{x} = 344{,}07$

2. grup: $\bar{x} = 292{,}32$

Örnek 4

1	58, 92, 93, 94, 95
2	02, 12, 15, 29, 30, 37, 40, 44, 47, 59
3	01, 01, 21, 37
4	15, 34, 44, 85, 96
5	29, 37
6	24
7	07
8	00

1	59, 89, 91, 98
2	35, 45, 50, 56, 61, 65, 66, 80
3	43, 56, 83
4	03, 14, 28, 32

1. grup: $\bar{x} = 344{,}07 \qquad ortanca = 259$

2. grup: $\bar{x} = 292{,}32 \qquad ortanca = 265$

Örnek Tepedeğeri

Merkezi eğilimi gösteren bir başka istatistik ise örnek tepedeğeridir.

Veri kümesi içerisinde en yüksek sıklık değerine sahip veriye **tepedeğer** denir.

Diğer bir deyişle veri kümesinde en sık görünen değer tepedeğerdir.

Veri kümesinde en yüksek sıklık değerine sahip birden fazla veri mevcutsa bu verilere **tepederler** denir.

Örnek Varyansı

Bir veri kümesinin merkezi eğiliminin yanı sıra, veri kümesindeki verilerin yayılımı veya değişkenliği hakkında da bilgi edinmek isteriz.

Değerlerin ortalamadan uzaklıklarının karesinin ortalamasını veren değer, örnek varyansı, bize böyle bir bilgi sunar ve s^2 ile gösterilir.

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

Örnek Varyansı

Aşağıda verilen eşitlikler varyansın hesaplanmasında faydalı olabilir.

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)\\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2\\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2\end{aligned}$$

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

Örnek Varyansı

Aşağıda verilen eşitlikler varyansın hesaplanmasında faydalı olabilir.

$i = 1, 2, \ldots, n$ için $y_i = ax_i + b$ ise $\bar{y} = a\bar{x} + b$'dir.

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

$s_y^2 = a^2 s_x^2$

Örnek Standart Sapması

Standart sapma, varyansın pozitif kareköküdür.

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

Örnek 5

Aşağıdaki veri kümelerinin varyansını hesaplayınız.

$A \to 3, 4, 6, 7, 10$

$B \to -20, 5, 15, 24$

Örnek 5

$A \to 3, 4, 6, 7, 10$

$\bar{A} = 6$

$s^2 = \frac{(3-6)^2+(4-6)^2+(6-6)^2+(7-6)^2+(10-6)^2}{5-1} = \frac{-3^2+-2^2+0^2+1^2+4^2}{4} = 7{,}5$

$B \to -20, 5, 15, 24$

$\bar{B} = 6$

$s^2 = \frac{(-20-6)^2+(5-6)^2+(15-6)^2+(24-6)^2}{4-1} = \frac{-26^2+-1^2+9^2+18^2}{3} \approx 360{,}67$

Örnek 6

Aşağıdaki veri kümesinin varyansını hesaplayınız.

$25, 20, 21, 18, 13, 13, 7, 9, 18$

Örnek 6

Bu veri kümesini x ile temsil edelim.

$x \to 25, 20, 21, 18, 13, 13, 7, 9, 18$

Her bir değerden 18'i çıkartarak yeni bir veri kümesi oluşturalım. Bu yeni veri kümesi ile daha kolay hesaplama yapabiliriz.

$y = x - 18$

$y \to 7, 2, 3, 0, -5, -5, -11, -9, 0$

y veri kümesinin varyansı, x veri kümesinin varyansına eşittir. Dolayısıyla y'nin varyansını bulduğumuzda sonucu bulmuş oluruz.

Örnek 6

$y = x - 18$

$y \to 7, 2, 3, 0, -5, -5, -11, -9, 0$

$\bar{y} = -\frac{18}{9} = -2$

$\sum_{i=1}^{9} y_i^2 = 314$

$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1} = \frac{314 - 9 \times (-2)^2}{9-1} = 34{,}75$

Chebyshev Eşitsizliği

Ortalaması ve standart sapması verilmiş bir veri kümesi için Chebyshev eşitsizliği şunu söyler:

Herhangi bir $k \geq 1$ değeri için,

verinin en az $\%100\left(1 - \frac{1}{k^2}\right)$'si

$\bar{x} - ks$ ile $\bar{x} + ks$ arasındadır.

Chebyshev Eşitsizliği

Örneğin $k = \frac{3}{2}$ olsun. Bu durumda;

$$\%100\left(1-\frac{1}{(3/2)^2}\right) = \%100\left(1-\frac{4}{9}\right) = \%100(0,5556) = \%55,56$$

$$ks = \frac{3}{2}s = 1,5s$$

$k = \frac{3}{2}$ için, Chebyshev eşitsizliğine göre, verinin en az $\%55,56$'sı örnek ortalamasından en fazla $1,5s$ uzaklıktadır.

Chebyshev Eşitsizliği

Chebyshev eşitsizliğinin resmi tanımı aşağıdaki gibidir.

$$S_k = \{i, 1 \leq i \leq n: |x_i - \bar{x}| < ks\}$$

$$N(S_k) = |S_k|$$

$$\frac{N(S_k)}{n} \geq 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k^2}$$

Eşleştirilmiş Veri Kümeleri

Bazen aralarında ilişki bulunan veri çiftlerine sahip veri kümeleri ile de ilgileniriz.

Böyle bir veri kümesinde, her eleman bir x ve bir de y değerine sahip ise, bu veri kümesindeki i. veri noktasını (x_i, y_i) veri çifti ile ifade ederiz.

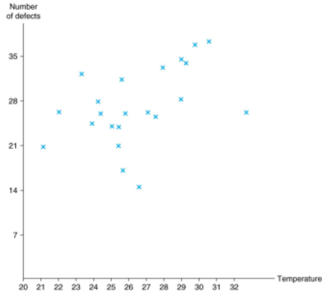
Örneğin günlük sıcaklık ile o günkü arızalı parça sayısını kaydettiğimizi düşünelim. Bu veri kümesi için x_i, i. günkü sıcaklığı ve y_i de i. gündeki arızalı parça sayısını temsil eder.

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 Veri Kümeleri

Eşleştirilmiş bir veri kümesini önceki yansıda gösterildiği gibi tablo ile temsil edebileceğimiz gibi yanda gösterildiği gibi **serpme diyagramı** ile de temsil edebiliriz.



Örnek Korelasyon Katsayısı

Eşleştirilmiş veri kümelerinde x ve y değerleri arasındaki ilişkiyi incelemek isteriz.

- Büyük x değerleri ile büyük y değerleri ve küçük x değerleri ile küçük y değerleri eşleşiyor mu?
- Büyük x değerleri ile küçük y değerleri ve küçük x değerleri ile büyük y değerleri eşleşiyor mu?

Bu soruların cevaplarını kabaca serpme diyagramında görebiliriz.

Ancak eşleşmiş bu veriler arasındaki ilişkiyi sayısal olarak ölçebilmek için bir istatistik bilgi gerekir.

Bu bilgi **örnek korelasyon katsayısı** r ile ifade edilebilir.

Örnek Korelasyon Katsayısı

Örnek korelasyon katsayısı aşağıdaki formüller ile hesaplanabilir.

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

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

$r > 0$ olduğunda örnek veri çiftleri pozitif ilişkilidir denir.

$r < 0$ olduğunda örnek veri çiftleri negatif ilişkilidir denir.

Örnek Korelasyon Katsayısı

Örnek korelasyon katsayısının bazı özellikleri aşağıdaki şekildedir.

1. $-1 \leq r \leq 1$

Örnek korelasyon katsayısı r, her zaman -1 ve $+1$ arasındadır.

Örnek Korelasyon Katsayısı

Örnek korelasyon katsayısının bazı özellikleri aşağıdaki şekildedir.

2. a, b sabit sayılar ve $a > 0$ olmak üzere

$y_i = ax_i + b\ ,\ i = 1, \dots, n$

ise $r = 1$'dir.

Eşleştirilmiş veriler arasında büyük y değerlerini büyük x değerlerine bağlayacak şekilde bir **doğrusal ilişki** var olduğunda r, $+1$ değerine eşit olur.

Örnek Korelasyon Katsayısı

Örnek korelasyon katsayısının bazı özellikleri aşağıdaki şekildedir.

3. a, b sabit sayılar ve $a < 0$ olmak üzere

$y_i = ax_i + b\ ,\ \ i = 1, \ldots, n$

ise $r\ = -1$'dir.

Eşleştirilmiş veriler arasında büyük y değerlerini küçük x değerlerine bağlayacak şekilde bir **doğrusal ilişki** var olduğunda r, -1 değerine eşit olur.

Örnek Korelasyon Katsayısı

Örnek korelasyon katsayısının bazı özellikleri aşağıdaki şekildedir.

4. r, x_i ve y_i $(i = 1, \ldots, n)$ veri çiftlerinin örnek korelasyon katsayısı ise a ve c'nin her ikisi de pozitif veya her ikisi de negatif olmak koşulu ile

$ax_i + b$ ve $cy_i + d$ $(i = 1, \ldots, n)$

veri çiftlerinin örnek korelasyon katsayısı da r'dir.

Örnek Korelasyon Katsayısı