T.C.
KARADENİZ TEKNİK ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
ENDÜSTRİ MÜHENDİSLİĞİ BÖLÜMÜ

Mühendislik İstatistiği-II #Dağılım (Değişkenlik) Ölçüleri

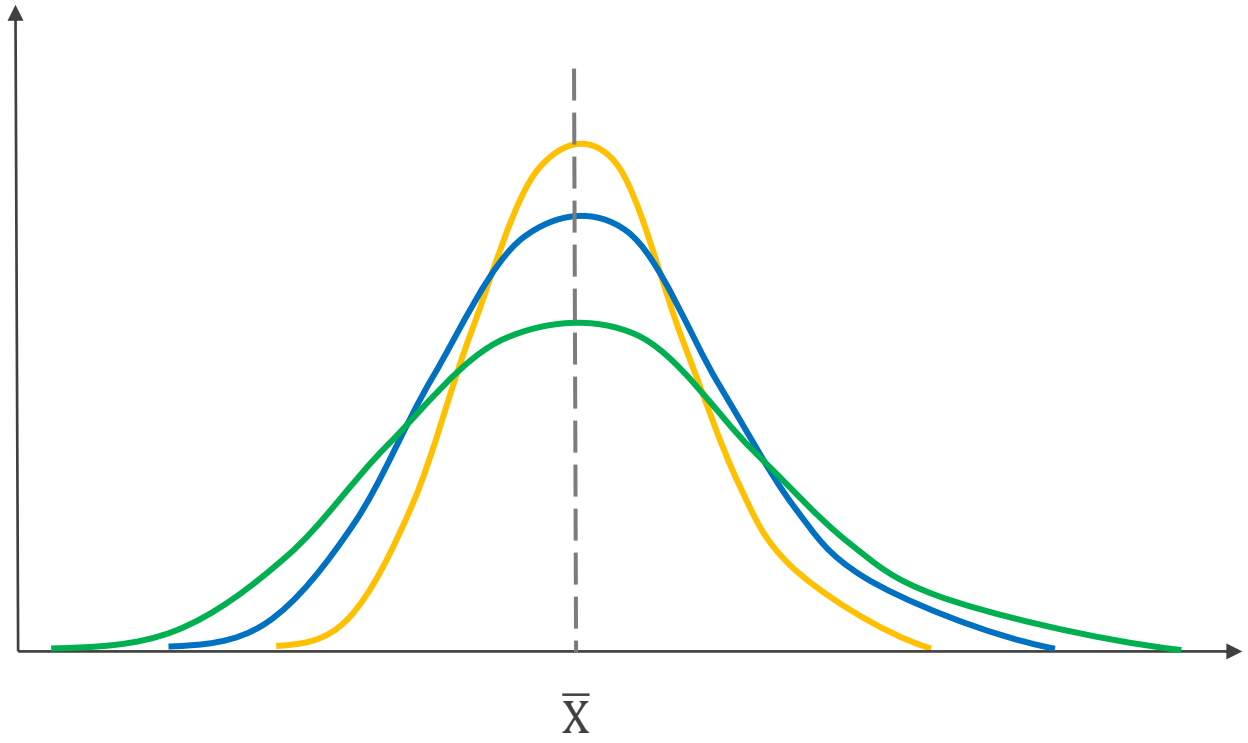
DR. ÖĞR. ÜYESİ FATMA BETÜL YENİ

Dağılım (Değişkenlik) Ölçüleri Nedir?

Dağılım ölçüleri, bir veri kümesinin ne kadar yayıldığını ölçen istatistiksel ölçümlerdir. Bu ölçümler, veri setinin merkezi eğilim ölçümleri (örneğin ortalama, medyan) gibi diğer özelliklerini tamamlayıcıdır. Veri kümesinin dağılımı, veri noktalarının ne kadar yayıldığına bağlıdır. Sıklıkla kullanılan dağılım ölçüleri;

- Değişim Aralığı
- Çeyrek Ayrılış
- Ortalama Ayrılış
- Varyans ve Standart Sapma
- Değişim Katsayısı
- Kutu Grafiği

Dağılım Ölçüleri Neden Kullanılır?



Dağılım Ölçüleri Neden Kullanılır?

Dağılım ölçüleri, birçok farklı alanda kullanılır. Örneğin;

- İşletme ve ekonomide
- Tıp ve sağlıkta
- Eğitimde alanında
- Bilimsel araştırmalarda

Bu nedenlerle, dağılım ölçümleri, birçok alanda verilerin analiz edilmesinde ve yorumlanmasında önemli bir rol oynamaktadır.

Değişim Aralığı – Ranj (R)

En basit değişim ölçüsü olarak bilinir. Bir veri kümesindeki en büyük gözlem değeri ile en küçük gözlem değeri arasındaki sayısal farktır.

Bu ölçü, gözlem değerlerinin ölçüm birimiyle (gr, cm, m^3 ... gibi) ifade edilir ve örnekteki tek bir gözlem değerinin aşırı büyük ya da küçük olmasından oldukça etkilenir.

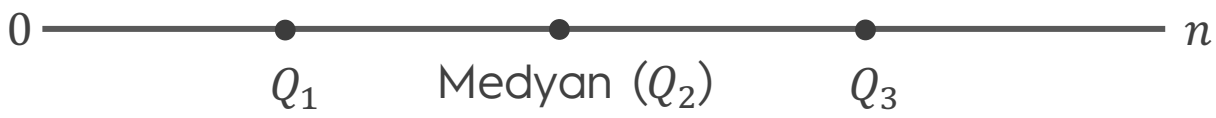
- Az sayıdaki verilerde (ham verilerde); $R = X_{max} - X_{min}$
- Gruplandırılmış verilerde; X_s : Son sınıfın orta değeri

  X_i : İlk sınıfın orta değeri olmak üzere

  $R = X_s - X_i$

Çeyrek Ayrılış – Quartile (Q)

Çeyrekler, küçükten büyüğe doğru sıralı halde dizili gözlem değerlerini çeyrek (dörtte bir) parçalara bölen değerlerdir.



- $\frac{n+1}{4} = 1.\text{çeyrek}\ (Q_1)$
- $\frac{n+1}{2} = 2.\text{çeyrek}\ (Q_2)$
- $\frac{3(n+1)}{4} = 3.\text{çeyrek}\ (Q_3)$

Çeyrek ayrılış (çeyrekler arası açıklık) ise 3. çeyrek ile 1. çeyrek değerleri arasındaki farktır. Veri setinin ortasında yer alan %50'nin aralığını temsil etmektedir.

Çeyrek Ayrılış – Quartile (Q)

Örnek: 50, 75, 90, 110, 125, 140, 142 verilerini ele alalım.

- Veri setinin medyanı = 110

  Q_1 = 75 ve Q_3 = 140 olur.

  $Çeyrek\ ayrılış\ Q$ = Q_3 - Q_1 = 140 – 75 = 65 olarak bulunur

Çeyrek Ayrılış – Quartile (Q)

Gruplandırılmış veriler için;

n_1 : İlk dörtte birliğin bulunduğu sınıftan önceki sınıfların toplam frekansı

J_1 : (n/4) – n_1

f_{Q1} : İlk dörtte birliğin bulunduğu sınıf frekansı

h : Sınıf genişliği

L_1 : İlk dörtte birliğin bulunduğu sınıfın alt değeri olmak üzere;

$$Q_1 = L_1 + \frac{J_1 * h}{f_{Q1}}$$ ile bulunur.

Benzer şekilde; $Q_3 = L_3 + \frac{J_3 * h}{f_{Q3}}$ ile bulunur

Çeyrek Ayrılış – Quartile (Q)

Örnek: n=100 olan bir veri seti için

Sınıf No	Sınıf Limitleri	Gözlem Sayısı (f_i)	Eklemeli Frekanslar
1	0-2	1	1
2	3-5	3	4
3	6-8	8	12
4	9-11	13	25
5	12-14	20	45
6	15-17	25	70
7	18-20	17	87
8	21-23	5	92
9	24-26	3	95
10	27-29	5	100

- n/4 =25 olduğuna göre 4. sınıf Q_1'in olduğu sınıftır.

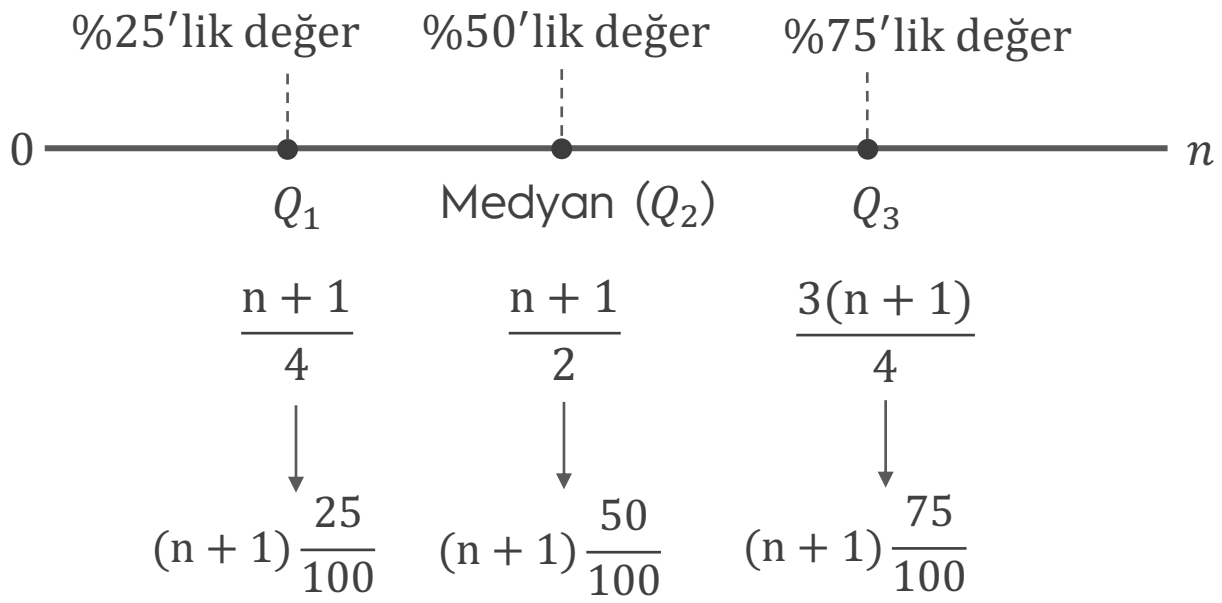
  $L_1 = 9$
  $n_1 = 12$
  $f_{Q1} = 13$
  $h = 3$
  $J_1 = n/4 - n_1 = 25 - 12 = 13$ olmak üzere

$$Q_1 = L_1 + \frac{J_1 * h}{f_{Q1}} = 9 + \frac{13 * 3}{13} = 12$$

Benzer şekilde;

$$Q_3 = 18 + \frac{5 * 3}{17} = 18{,}88 \text{ olarak bulunur}$$

Yüzdelikler



Buna göre örneğin veri setinin %15'lik değerini öğrenmek istersek; $(n+1)\frac{15}{100}$ ile hesaplayabiliriz.

Yüzdelikler

Örnek: 45, 58, 23, 12, 87, 99, 34, 32, 56, 78, 28, 76, 75, 34, 96, 94, 48, 19, 61, 91 (n=20) olan aşağıdaki veri setinin %20'lik ve %90'lık değerleri nelerdir?

12	19	23	28	32	34	34	48	48	56	58	61	75	76	78	87	91	94	96	99

- %20'lik değer; $(20+1)\frac{20}{100} = 4{,}2$

  4. değer (28) ile 5. değerin (32) ortalaması (28 + 32)/2 = 30

- %90'lık değer; $(20+1)\frac{90}{100} = 18{,}9$

  18. değer (94) ile 19. değerin (96) ortalaması (94 + 96)/2 = 95

  Yani veri seti içerisindeki verilerin %20'si 30'dan; %90'ı ise 95'ten küçüktür.

Mutlak (Ortalama) Sapma

Mutlak sapma (MS), bir veri seti içindeki gözlem değerlerinin aritmetik ortalamadan mutlak değer olarak sapmalarının (farklarının) ortalamasıdır.

- Az sayıdaki verilerde; $x_1, x_2, \cdots x_n$ dizisinin aritmetik ortalaması $\overline{X} = \frac{1}{n}\sum x_i$ ol.üz.;

  $MS = \frac{|x_1-\overline{X}|+|x_2-\overline{X}|+\cdots+|x_n-\overline{X}|}{n} = \frac{1}{n}\sum_{i=1}^{n}|x_i-\overline{X}|$ olarak hesaplanır.

- Sınıflandırılmış verilerde; $\overline{X} = \frac{\sum_{j=1}^{k} F_j \cdot x_j}{\sum_{j=1}^{k} F_j}$ ol. üz; $MS = \frac{\sum_{j=1}^{k} F_j \cdot |x_j-\overline{X}|}{\sum_{j=1}^{k} F_j}$ 'dır.

- Gruplandırılmış verilerde; $\overline{X} = \frac{\sum_{j=1}^{k} F_j \cdot \overline{X}_j}{\sum_{j=1}^{k} F_j}$ ol. üz; $MS = \frac{\sum_{j=1}^{k} F_j \cdot |\overline{X}_j-\overline{X}|}{\sum_{j=1}^{k} F_j}$ 'dır.

Mutlak (Ortalama) Sapma

Örnek: Bir işletmeye ait günlük üretim aşağıdaki gibidir. Günlük ortalama üretim miktarını ve ortalama ayrılışı hesaplayalım.

Gün	Üretim Miktarı	$\lvert x_i - \overline{X} \rvert$
1	195	23
2	128	44
3	165	7
4	147	25
5	213	41
6	184	12
	Σx_i = 1032	Σ = 152
	$\overline{X}$ = 172	

$$MS = \frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{X}| = \frac{152}{6} = 25{,}3 \quad \text{olur.}$$

Varyans ve Standart Sapma

- Varyans ve standart sapma, bir veri kümesindeki değerlerin ne kadar dağıldığını ölçen ölçütlerdir.
- Varyans, her bir veri noktasının ortalamadan ne kadar uzak olduğunu ölçen karelerin ortalama bir ölçüsüdür. Standart sapma ise varyansın karekökü dür ve aynı birimlerle ifade edilir.
- Daha küçük bir varyans veya standart sapma, veri noktalarının daha az dağıldığını ve daha homojen bir veri seti olduğunu gösterirken, daha büyük bir varyans veya standart sapma, veri noktalarının daha fazla dağıldığını ve daha heterojen bir veri seti olduğunu gösterir.

	Birim Sayısı	Aritmetik Ort.	Varyans	S. Sapma
Yığın (Kitle)	N	μ	σ^2	σ
Örneklem	n	$\bar{X}$	S^2	S

Varyans ve Standart Sapma

Ana kitlenin varyansı ve standart sapması

- x_1 , x_2 , ..., x_N sonlu sayıdaki gözlem değerlerinin aritmetik ortalaması $\mu = \frac{\sum_{x=1}^{N} x_i}{N}$ ol. üz.;

  Kitle varyansı $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$ ve kitle standart sapması $= \sqrt{\sigma^2}$ olarak hesaplanır.

- N gözlemden oluşan bir kitlede x_1 , x_2 , ... , x_k gözlem değerleri ve f_1 , f_2 , ... , f_k frekansları olsun.

  $N = \sum_{j=1}^{k} f_j$ ve $\mu = \frac{\sum_{j=1}^{k} x_j * f_j}{N}$ ol. üz.;

  Kitle varyansı $\sigma^2 = \frac{1}{N}\sum_{j=1}^{k}(x_j - \mu)^2 * f_j$ ve kitle standart sapması $= \sqrt{\sigma^2}$ olarak hesaplanır.

Varyans ve Standart Sapma

Örneklemin varyansı ve standart sapması

- Ham veriler için örneklemin varyansı $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$
- Sınıflandırılmış veriler için örneklemin varyansı $S^2 = \frac{1}{n-1}\sum_{j=1}^{k}(x_j - \bar{X})^2 * f_j$
- Gruplandırılmış veriler için örneklemin varyansı $S^2 = \frac{1}{n-1}\sum_{j=1}^{k}(\bar{X}_j - \bar{X})^2 * f_j$
- Örneklemin standart sapması $S = \sqrt{S^2}$

Varyans ve Standart Sapma

Örneklem hacmi n büyük ise; varyans hesabı (sınıflandırılmamış veriler için) şu şekilde yapılabilir:

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i{}^2 - n\bar{X}^2\right)$$

- İspat: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i{}^2 - 2x_i\bar{X} + \bar{X}^2) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i{}^2 - 2\sum_{i=1}^{n} x_i\bar{X} + \sum_{i=1}^{n}\bar{X}^2\right)$

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ Olduğu bilindiğine göre; $n\bar{X} = \sum_{i=1}^{n} x_i$ ol. üz.;

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i{}^2 - 2n\bar{X}^2 + n\bar{X}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i{}^2 - n\bar{X}^2\right)$$

Varyans ve Standart Sapma

- **Örnek**: Bir sınıftan seçilen rasgele 10 öğrencinin İstatistik dersi final sınavından aldıkları notlar yandaki şekildedir. Bu veri için ortalama, varyans ve standart sapmayı hesaplayalım.

Alınan Not (x_i)	$x_i - \bar{X}$	$(x_i - \bar{X})^2$
45	-28	784
60	-13	169
75	2	4
82	9	81
55	-18	324
90	17	289
85	12	144
80	7	49
93	20	400
65	-8	64
$\bar{X}$ = 73		Toplam= 2308

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

Olmak üzere;

$$S^2 = \frac{2308}{10-1} = 256{,}4$$

$$S = \sqrt{256{,}4} \cong 16$$

Varyans ve Standart Sapma

- **Örnek**: Bir sınıftaki 100 öğrencinin kardeş sayısına göre dağılımı aşağıda verilmiştir. Veri setinin varyans ve standart sapması nedir?

x_j (Kardeş Sayısı)	f_j (Öğrenci Sayısı)	$x_j * f_j$	$x_j - \mu$	$(x_j - \mu)^2$	$(x_j - \mu)^2 * f_j$
0	10	0	-2	4	40
1	20	20	-1	1	20
2	40	80	0	0	0
3	20	60	1	1	20
4	10	40	2	4	40
		Toplam= 200			Toplam= 120

$$\bar{\mu} = \frac{200}{100} = 2$$

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{k}(x_j - \mu)^2 * f_j$$

Olmak üzere;

$$\sigma^2 = \frac{120}{100} = 1{,}2$$

$$\sigma = \sqrt{1{,}2} = 1{,}09$$

Varyans ve Standart Sapma

- **Örnek**: Rasgele seçilen 100 hasta üzerinde bir ilacın etki süresi aşağıdaki gibidir.

x_j (Etki süresi-dk)	f_j (Hasta sayısı)	$\bar{X}_j$ (Sınıf orta noktaları)	$\bar{X}_j * f_j$	$\bar{X}_j - \bar{X}$	$(\bar{X}_j - \bar{X})^2$	$(\bar{X}_j - \bar{X})^2 * f_j$
$20 \leq X < 24$	20	22	440	-5,8	33,64	672,8
$24 \leq X < 28$	40	26	1040	-1,8	3,24	129,6
$28 \leq X < 32$	20	30	600	2,2	4,84	96,8
$32 \leq X < 36$	15	34	510	6,2	38,44	576,6
$36 \leq X \leq 40$	5	38	190	10,2	104,04	520,2
			Toplam= 2780			Toplam= 1996

$$\bar{X} = \frac{2780}{100} = 27{,}8$$

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{k}(\bar{X}_j - \bar{X})^2 * f_j$$

Olmak üzere;

$$S^2 = \frac{1996}{100-1} = 20{,}16$$

$$S = \sqrt{20{,}16} \cong 4{,}5 \text{ dk}$$

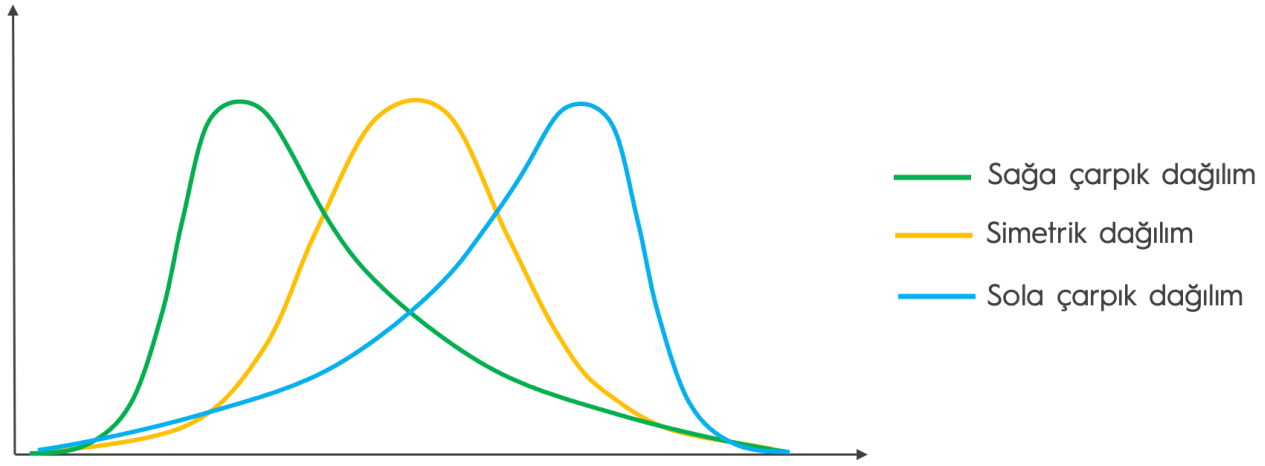
Çarpıklık Ölçütü

- S^2 örneklem varyansı, $(x_i - \bar{X})^2$ sapmaların ortalamasıdır. Bu nedenle varyansa, ortalamaya göre ikinci moment denir ve m_2 ile gösterilir.
- Veri seti için karşılık gelen çarpıklık ölçütü $(x_i - \bar{X})^3$ kübik sapmaların ortalamasının alınması ile bulunur ve m_3 ile gösterilir.
- Çarpıklık ölçütü şu şekilde hesaplanır;

$$m_3 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^3$$

Çarpıklık Ölçütü

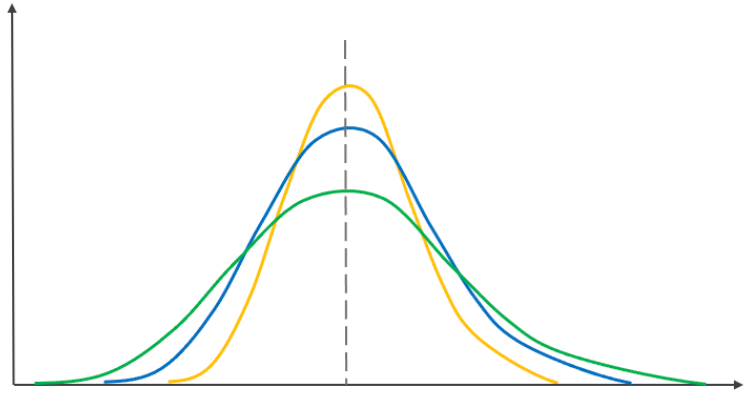
- Çarpıklık ölçütü, veri dağılımının simetrik olması durumunda sıfır değerini alırken, sağa veya sola doğru çarpıklık durumunda pozitif veya negatif değerler alır.
- Pozitif bir çarpıklık değeri sağa doğru eğilimi gösterirken, negatif bir çarpıklık değeri sola doğru eğilimi gösterir.



Sivirilik-Basıklık Ölçütü

$$m_4 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^4$$ Eşitliği sivrilik ölçütü olarak tanımlanmaktadır

- Sivrilik ölçütü negatif değer alırsa; merkeze yakın yerde eğri normal dağılım eğrisine göre daha basıktır denir. Pozitif değer alırsa; eğri merkeze yakın yerde normal dağılım eğrisinden daha dar ve yüksektir.



Değişim Katsayısı (DK)

- Değişim katsayısı, bir veri kümesindeki değişkenliğin (standart sapma) ortalama değere (aritmetik ortalama) oranı olarak ifade edilir.

Örneklem için: $\mathrm{DK} = \frac{S}{\bar{X}}$ veya $\%\mathrm{DK} = \frac{S}{\bar{X}} \cdot 100$

Yığın için: $\mathrm{DK} = \frac{\sigma}{\mu}$ veya $\%\mathrm{DK} = \frac{\sigma}{\mu} \cdot 100$

- Değişim katsayısı, verilerin dağılım şekli hakkında da bilgi sağlar.
- Değişim katsayısı, farklı birimlerle ifade edilen verilerin karşılaştırılması için kullanışlı bir ölçüttür.

Değişim Katsayısı

Örnek: x ve y sırasıyla bir sınıftaki öğrencilerin boy ve ağırlık dağılımını göstersin.

$\overline{X} = 168\ \text{cm}$ ve $S_x^2 = 25\ \text{cm}^2$;

$\overline{Y} = 62\ \text{kg}$ ve $S_y^2 = 16\ \text{kg}^2$ olarak verilmiş olsun.

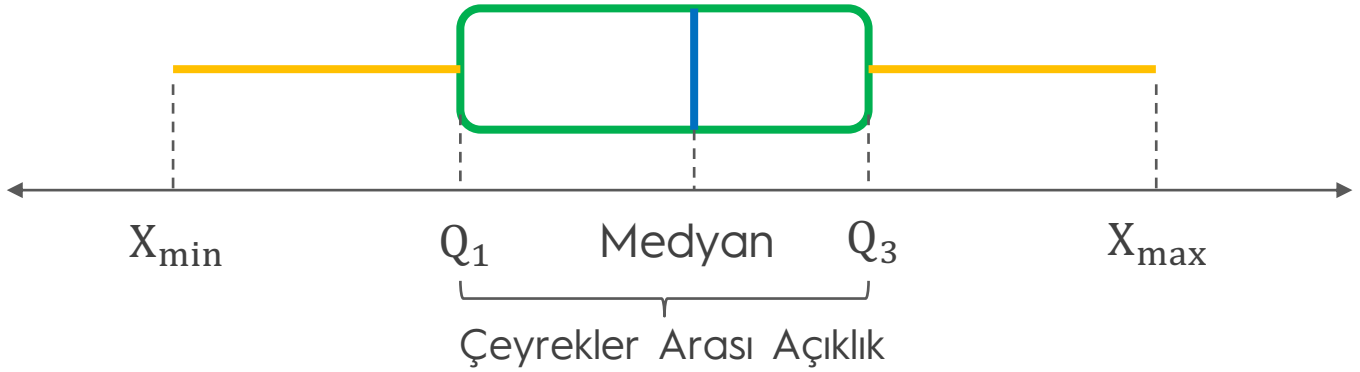
- $(DK)_x = \frac{5}{168} = 0{,}029$
- $(DK)_y = \frac{4}{62} = 0{,}064$

$(DK)_x < (DK)_y$ olduğundan, sınıftaki öğrencilerin uzunluk bakımından birbirine daha benzer olduğu söylenebilir.

Kutu Grafiği

- Kutu grafiği nicel değişkenleri tanımlamak için yaygın olarak kullanılan bir araçtır.
- Bu grafik verilerin, merkezi konumu, dağılımını, çarpıklığı ve basıklığı hakkında bilgi verirken aynı zamanda veride yer alan aykırı değerlerin tespit edilmesinde de kullanılır.
- Grafiğin çizilebilmesi için değişken değerlerine ait,
  - Minimum
  - Maksimum
  - Birinci çeyrek (Q_1)
  - Medyan ve
  - Üçüncü çeyrek (Q_3) değerlerinin hesaplanması gerekir.

Kutu Grafiği



- Kutunun içindeki medyan çizgisi (mavi çizgi), verinin konumu hakkında bilgi verir.
- Çeyrekler arası açıklık(yeşille çizilmiş kutu), verinin dağılımı hakkında bilgi verir. Bu açıklığın büyümesi verinin geniş bir aralıkta yer aldığını ifade eder.
- Medyan çizgisinin Q_1'e yaklaşması dağılımın sağa çarpıklığını, Q_3'e yaklaşması dağılımın sola çarpıklığını ve tam ortada olması ise dağılımın simetrik olduğunu gösterir.
- Kutunun genişliği, çizginin (turuncu çizgi) genişliğine yaklaştığında dağılımın basık, aksi durumda ise dağılımın sivri olduğu söylenebilir.

Kutu Grafiği

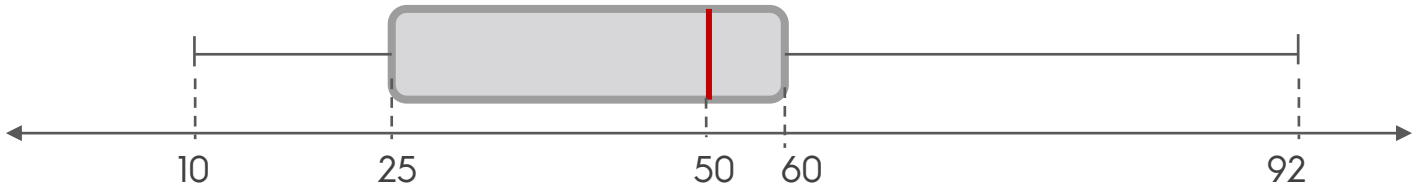
- Kutu grafikleri aykırı değerlerin tespit edilmesinde de kullanılabilir. Grafikte, $Q_1 - 1{,}5(Q_3 - Q_1)$ ve $Q_3 + 1{,}5(Q_3 - Q_1)$ sınırlarının dışına düşen değerler aykırı değerler olarak belirlenir.

Örnek: Bir sınıftaki 20 öğrencinin istatistik dersine ilişkin notları aşağıda verilmiştir. Kutu grafiğini çizerek sonucu yorumlayalım.

10	20	20	20	20	30	50	50	50	50
50	60	60	60	60	60	68	90	92	92

Kutu Grafiği

- Min = 10
- Max = 92
- N = 20 ve Medyan = $\frac{n+1}{2}$ =10,5 olduğu için $(X_{10} + X_{11})/2 = 50$
- $Q_1 = \frac{n+1}{4} = 5{,}25$ olduğu için $(X_5 + X_6)/2 = (20 + 30)/2 = 25$
- $Q_3 = \frac{3(n+1)}{4} = 15{,}75$ olduğu için $(X_{15} + X_{16})/2 = (60 + 60)/2 = 60$



Kutu Grafiği

Örnek: 20 kadın ve 20 erkeğe, bir günde ortalama TV seyretme süreleri (dakika olarak) anketle araştırılmış ve aşağıdaki veriler toplanmıştır.

Kadın	100	120	80	95	110	235	36	75	90	98	120	30	48	60	65	80	115	240	50	70
Erkek	30	45	60	20	10	30	65	68	90	18	35	25	83	65	70	130	90	90	46	80

- Çözüm için önce verileri küçükten büyüğe sıralayalım:

Kadın	30	36	48	50	60	65	70	75	80	80	90	95	98	100	110	115	120	120	235	240
Erkek	10	18	20	25	30	30	35	45	46	60	65	65	68	70	80	83	90	90	90	130

Kutu Grafiği

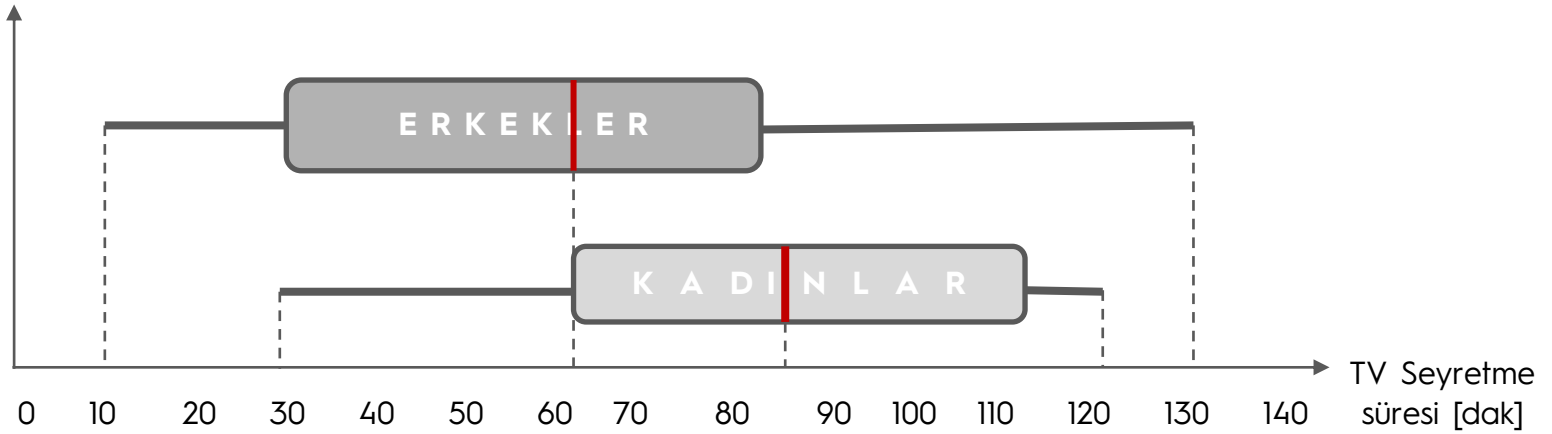
Kadınlar için;

- $X_{min} = 30$
- $X_{max} = 240$
- $\text{MEDYAN} = \frac{80+90}{2} = 85$
- $Q_1 = \frac{60+65}{2} = 62{,}5$
- $Q_3 = \frac{110+115}{2} = 112{,}5$

Erkekler için;

- $X_{min} = 10$
- $X_{max} = 130$
- $\text{MEDYAN} = \frac{60+65}{2} = 62{,}5$
- $Q_1 = \frac{30+30}{2} = 30$
- $Q_3 = \frac{80+83}{2} = 81{,}5$

Kutu Grafiği



- Kadınların TV seyretme sürelerinin dağılımı incelendiğinde, 235 ve 240 değerlerinin diğerlerinden farklı olduğu görülmektedir.
- Bu iki değer de $Q_3 + 1{,}5(Q_3 - Q_1)$ sınırı olan 112,5 + 1,5 * (112,5 - 62,5) = 187,5 değerinden oldukça büyüktür.
- Bu nedenle bu iki değer aykırı veri olarak nitelendirilmiş ve grafikte dikkate alınmamıştır.