VERİ BİLİMİ İÇİN TEMEL İSTATİSTİK hafta-3

CEMİLE YILDIZÇAKAR

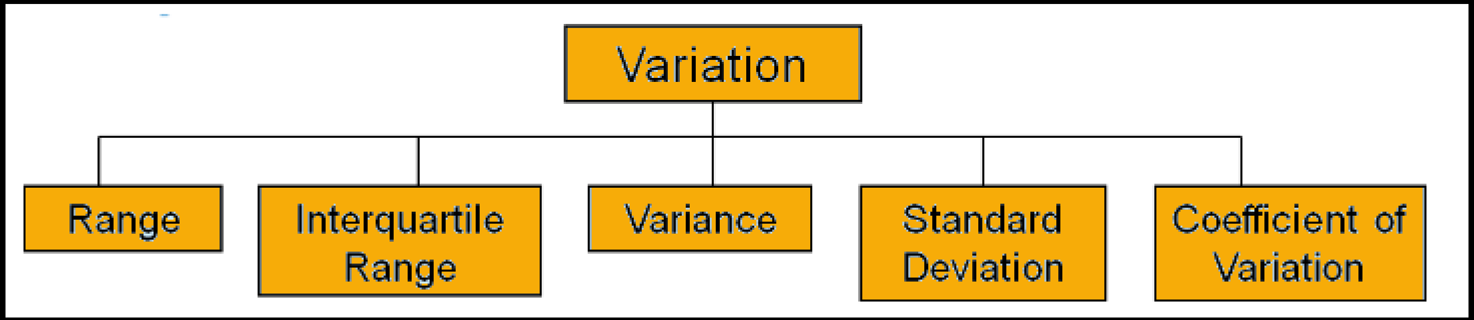
22.12.2020

İçindekiler

Değişim Ölçüleri (Measures of Variabillity)

- Açıklık
- Çeyreklikler
- Varyans
- Standart Sapma
- Değişim Katsayısı

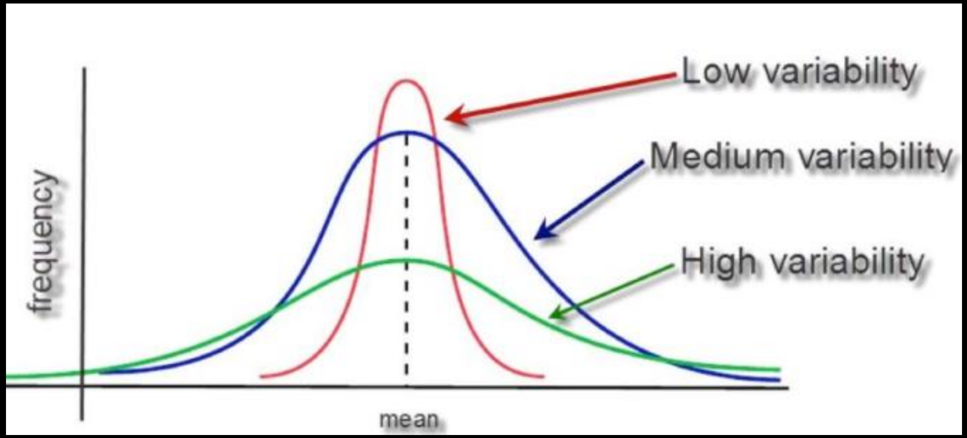
Cemile YILDIZÇAKAR

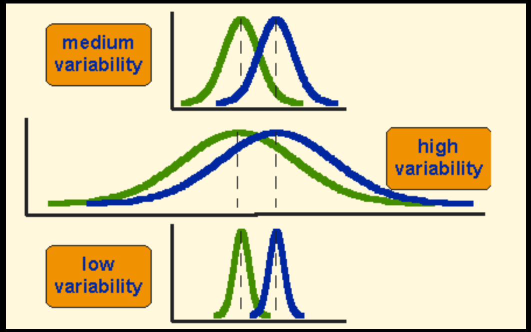

Variation
Range
Interquartile Range
Variance
Standard Deviation
Coefficient of Variation

Değişim Ölçüleri (Measures of Variabillity)

- **<u>Değişim:</u>**

Bir dağılımda ölçümler arasında gözlenen farklılık ve değişikliğe değişim, bir serideki gözlemlerin (birimlerin) birbirinden ya da herhangi bir ortalama değerden uzaklıklarının çeşitli ölçümlerine merkezi değişim ölçüleri denir.





Cemile YILDIZÇAKAR

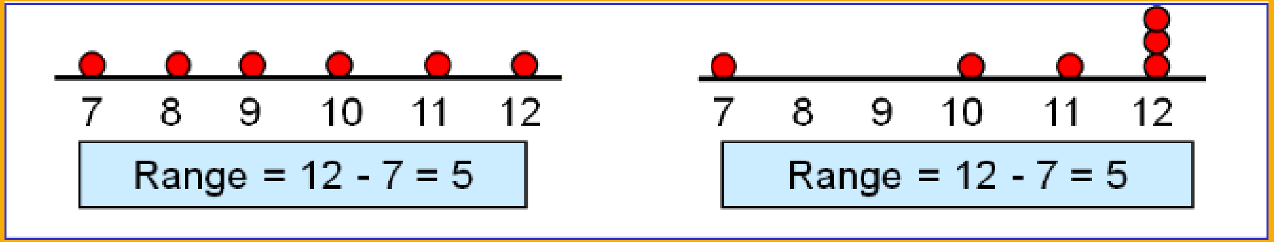
ARALIK/AÇIKLIK (RANGE)

- En basit değilim ölçüsüdür.
- Ölçümlerin en büyüğü ile en küçüğü arasındaki farktır.

Cemile YILDIZÇAKAR

Dezavantajları

- Dağılımın şeklini dikkate almaz.



- Sapan değerlerden etkilenir.

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

Range = 5 - 1 = 4

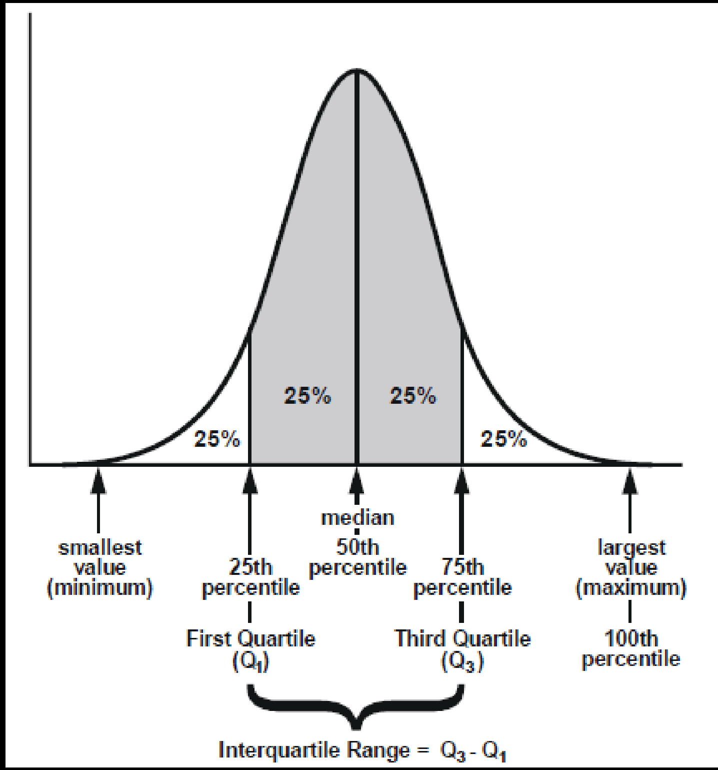
1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

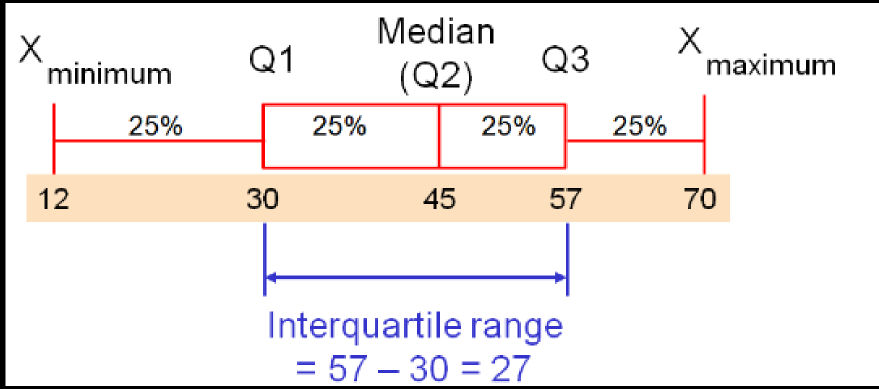
Range = 120 - 1 = 119

Cemile YILDIZÇAKAR

Çeyreklikler

- Sıralanmış veri setini ilk dört parçaya ayıran değerlerdir.





Cemile YILDIZÇAKAR

Çeyrekliklerin konumunun bulunması

Quartile Formula

Lower Quartile (Q1) = (N+1) x (0.25)

Middle Quartile (Q2)= (N+1) x (0.50)

Upper Quartile (Q3) = (N+1) x (0.75)

Sample Ranked Data: 11 12 13 16 16 17 18 21 22

(n = 9)

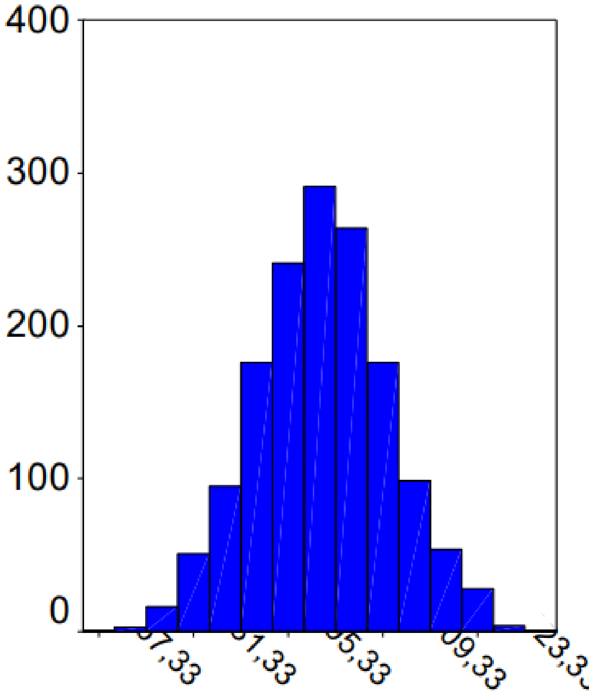
Q_1 = is in the $\boxed{0.25(9+1) = 2.5 \text{ position}}$ of the ranked data

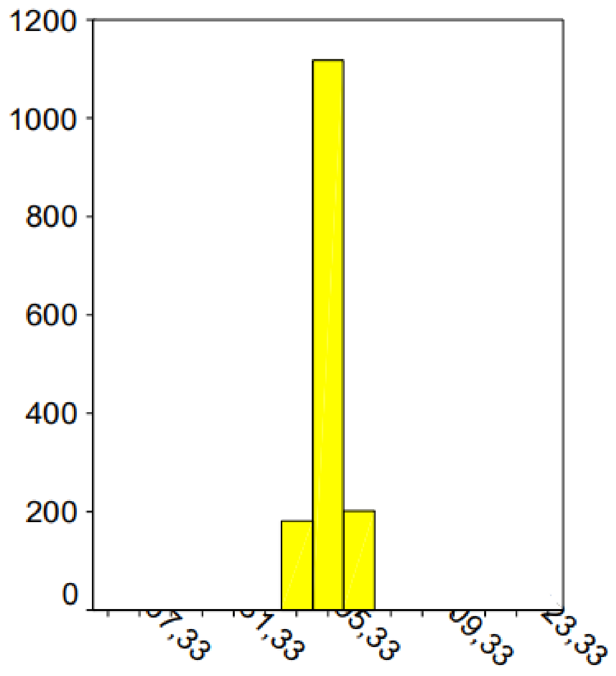
so use the value half way between the 2nd and 3rd values,

so $\boxed{Q_1 = 12.5}$

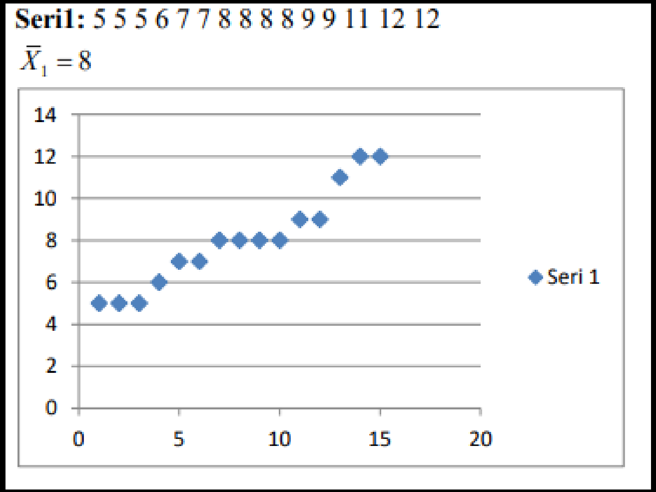
Cemile YILDIZÇAKAR

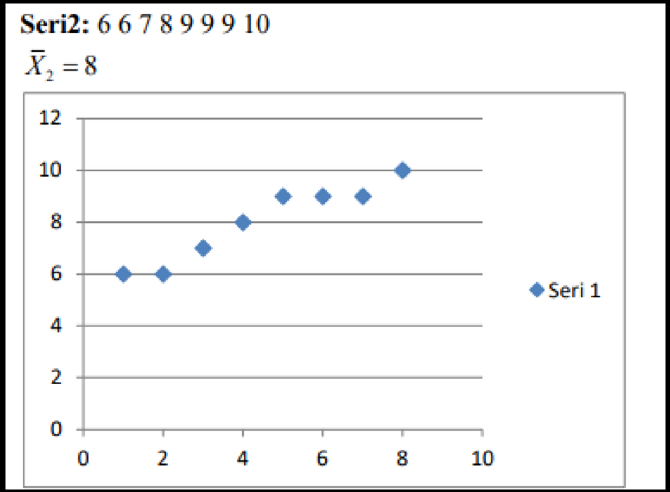
Aşağıdaki iki grafik n = 1500 hacimlik alınan iki farklı örne
doğrultusunda oluşturulan histogramlardır. Her iki örne
ortalaması yaklaşık olarak 100 olduğuna göre iki örneğin ayr
anakütleden alındığı söylenebilir mi?

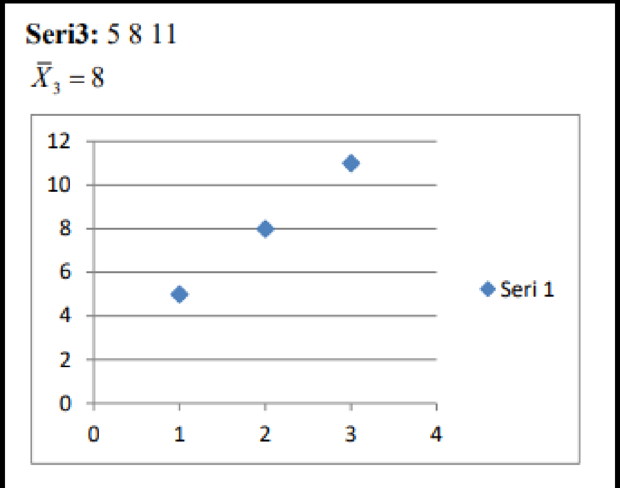




Cemile YILDIZÇAKAR





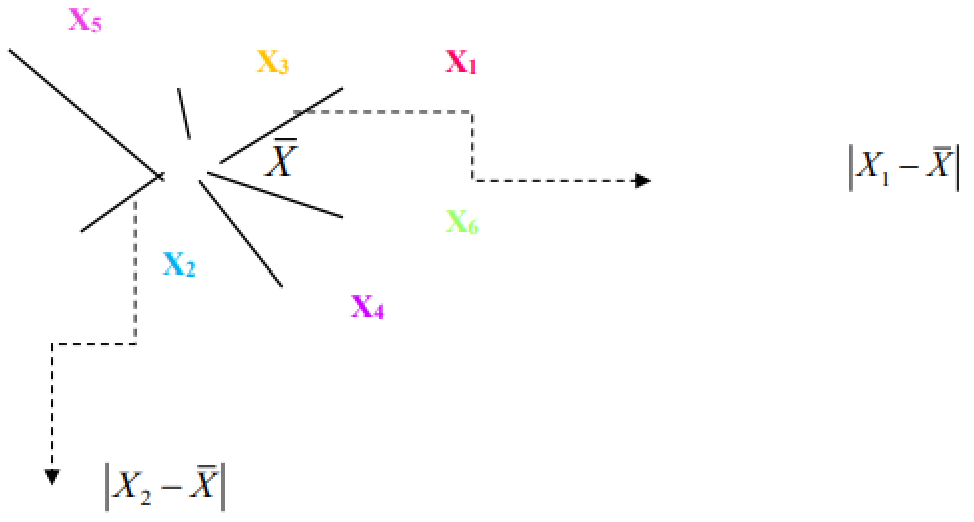


Her üçünün de ortalaması birbiriyle aynı olmasına rağmen, veri sayıları, değişimleri, dağılımları birbirinden farklıdır.

Cemile YILDIZÇAKAR

Verilerin dağılımını ölçmede, gözlem değerlerinin ortalamadan sapmalarını $|X_i - \bar{X}|$ kullanabiliriz. Bu sapmalar yani ne kadar büyükse, Xi gözlemi ortalamadan o denli uzakta demektir

$$\sum (X_i - \bar{X}) = 0$$



Cemile YILDIZÇAKAR

VARIANCE /STANDART DEVIATION

- **<u>VARYANS /STANDART SAPMA :</u>**

  Kitle setindeki tüm değerleri hesaba katan değişkenlik ölçüsüdür.

  Standart sapma varyansın kareköküne eşittir.

  Standart sapma ortalama etrafındaki değişimi gösterir.

  Standart sapmanın birimi veri setindeki değerlerin birimi ile aynıdır.

σ^2 : Kitle varyansı

s^2 : Örneklem varyansı

Cemile YILDIZÇAKAR

Basit seriler İçin:

Populasyon Varyansı: $$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

μ : Populasyon Ortalaması N : Populasyon Hacmi

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

Cemile YILDIZÇAKAR

Gruplanmış Seriler İçin:

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

Sınıflanmış Seriler İçin :

$$s^2 = \frac{\sum_{i=1}^{k} f_i (m_i - \bar{x})^2}{\sum_{i=1}^{k} f_i - 1}$$

Cemile YILDIZÇAKAR

Basit seriler İçin:

Populasyon Standart Sapması: $\sigma = \sqrt{\dfrac{\sum (x_i - \mu)^2}{N}}$

μ : Populasyon Standart Sapması N : Populasyon Hacmi

Örnek Standart Sapması : $s = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

Cemile YILDIZÇAKAR

Gruplanmış Seriler İçin:

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

Sınıflanmış Seriler İçin :

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i (m_i - \bar{x})^2}{\sum_{i=1}^{k} f_i - 1}}$$

Örnek: İstatistik I dersini alan 10 öğrencinin vize notları aşağıdaki gibi sıralanmıştır. Buna göre vize notları için varyans ve standart sapmayı hesaplayınız.

30,41,53,61,68,79,82,88,90,98

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{30+41+....+98}{10} = 69$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(30-69)^2 + (41-69)^2 + \ldots + (98-69)^2}{9}$$

$$= \frac{4538}{9} \approx 504,22$$

$$s^2 \approx 504,22 \quad \rightarrow s = \sqrt{s^2} = \sqrt{504,22} \approx 22,45$$

İstatistik I vizesinden alınan notların ortalama etrafında yaklaşık olarak 22 puan değiştiği görülmektedir.

Cemile YILDIZÇAKAR

CHEBYSHEV TEOREMİ

Herhangi bir veri setinde, verilerin ortalamanın K standart sapma uzağında bulunması oranı $1-1/K^2$ dır. Burada K, birden büyük pozitif sayıdır.

K=2 ve K=3 için;

•Verilerin en az 3/4' ü (%75) ortalamanın 2 standart sapma uzagında bulunur.
•Verilerin en az 8/9' u (%89) ortalamanın 3 standart sapma uzağında bulunur.

Cemile YILDIZÇAKAR

Örnek: X değişkeni bir sınıftaki İstatistik I dersinin başarı notlarını göstermek üzere, örnek ortalamasının 60 varyansının 100 olduğu bilindiğine göre, verilerin ¾ 'ü hagi aralıkta değişir?

$$1-\frac{1}{k^2}=\frac{3}{4} \quad \rightarrow k=2$$

$$(\bar{x} \mp 2s)$$

$$(60 \mp 2.10)$$

$$(40,80)$$

Kyanak: https://kisi.deu.edu.tr/s.ucdogruk/ist1de%c4%9fi%c5%9fkenlik_%c3%b6l%c3%a7%c3%bcleri(Bolum3).pdf

Cemile YILDIZÇAKAR

Ortalama ve standart sapmanın birlikte kullanıldığı bir durum da, verilerin yüzde kaçının hangi aralıkta bulunduğunu ölçmeye yarayan Chebyshev Teoremi'dir. Ortalaması μ, standart sapması σ olan bir veri kümesindeki gözlemlerin $\mu \pm k\sigma$ aralığına düşenlerin oranı en az $1-\frac{1}{k^2}$ 'dir. **Burada k, 1'den büyük bir sayıdır.**

Örnek: Bir lisedeki öğrencilerin IQ ortalaması 105, standart sapması 9'dur. Öğrencilerin en az % kaçı 85,2 ile 124,8 arasında IQ'ya sahiptir?

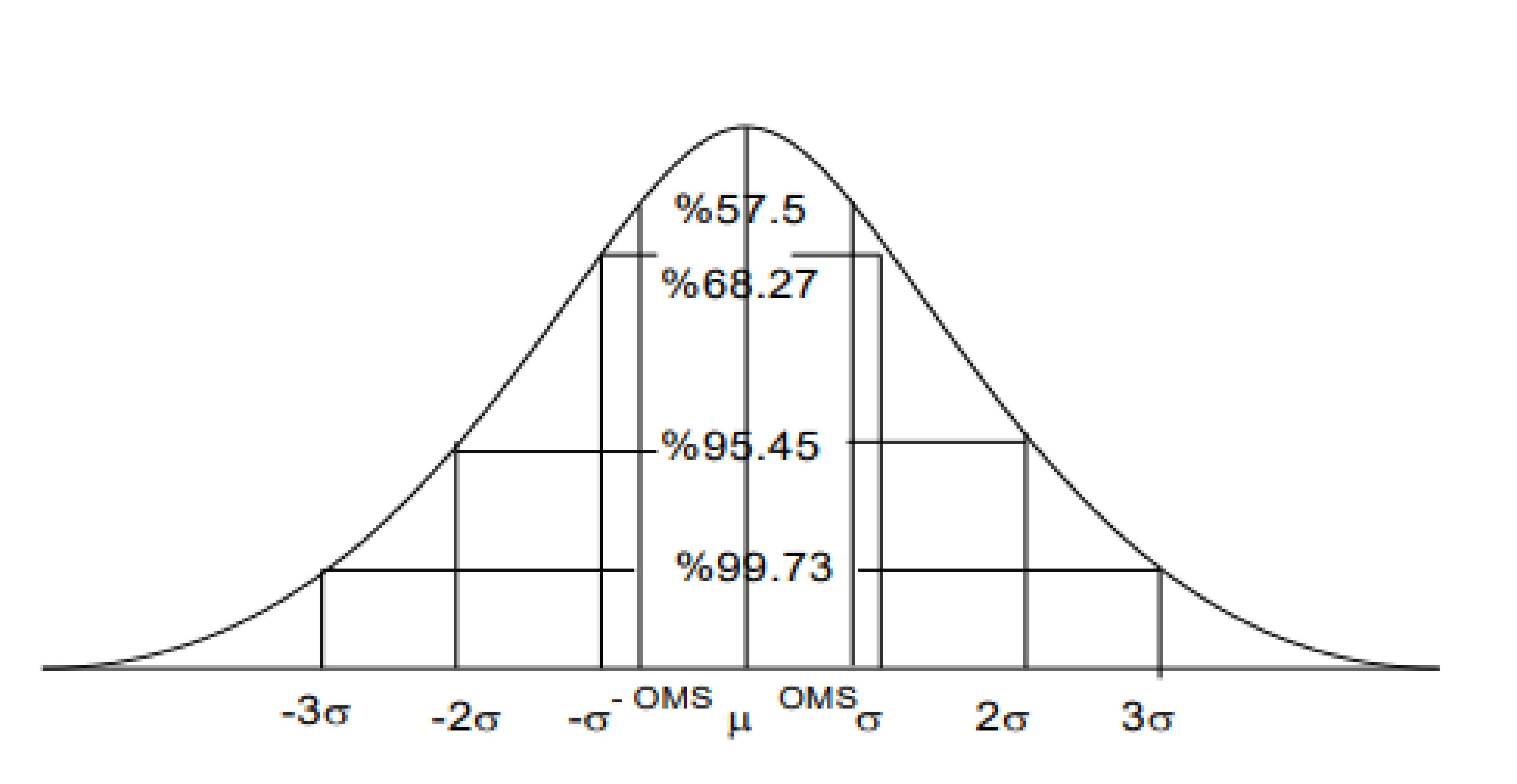
$$\mu - k\sigma = 85,2$$

$$\mu + k\sigma = 124,8$$

Buradan k=2,2 elde edilir.

$$1 - \frac{1}{(2,2)^2} = 0,79$$

Yorum: Öğrencilerin en az %79'u verilen aralıkta IQ skoruna sahiptir.


%57.5
%68.27
%95.45
%99.73
-3σ
-2σ
-σ
- OMS
μ
OMS
σ
2σ
3σ

Değişim Katsayısı (Coefficient of Variation)

- • Standart sapmayı ortalamanın bir yüzdesi olarak ifade eden ve iki veya daha fazla populasyondaki varyasyonu (değişkenliği) karşılaştırmada kullanılan ölçüye varyasyon(değişkenlik) katsayısı denir.
- Yüzdelik değerler alır (%).
- Farklı birimlerde ölçülmüş veri setlerinin değişkenliğini karşılaştırmak için kullanılabilir.

$$CV = \left(\frac{s}{\bar{X}} \right) \cdot 100\%$$

DK'sı küçük olan serilerin, diğerlerine göre daha az değişken olduğu (daha homojen yani ortalamaya yakın dağılıma sahip) birimlerden oluştuğu söylenir. Çünkü standart sapma küçüldükçe DK da küçülür. Standart sapmanın küçük olması ise ancak ortalama etrafındaki saçılımın $\bar{X}$'ya doğru çekilmesiyle mümkündür birimler $\bar{X}$'dan uzaklaştıkça standart sapma da büyümektedir.

Örnek: İki hisse senedi olsun. Bu hisselerin 1 aylık ortalama getirileri ve standart sapmaları aşağıdaki gibi ise, siz bir portföy yöneticisi olarak **riski seven** bir yatırımcıya hangi hisse senedini tavsiye edersiniz?

$\mu_1=70 \quad \sigma_1=6$

$\mu_2=25 \quad \sigma_2=3$

$$DK_1 = \frac{6}{70} * 100 = 8,57$$

$$DK_2 = \frac{3}{25} * 100 = 12$$

DK sı büyük olanı yani 2. hisse senedini tercih etmelidir. Şayet riski sevmeyen yatırımcı olsaydı, ona değişimi görece daha az olan ilk hisse senedini almasını tavsiye ederdik. Borsada riskli senetlerin getirisi de kaybı da büyüktür. Ama riski seven yatırımcı "risk yoksa getiri de yoktur" mantığına sahip olduğu için bu onun tercihidir. Riski düşük olan senetler her ne kadar güvenli olsa da, maalesef getirisi de düşüktür. Elbette bu bir tercih meselesidir.

- Kaynak: http://auzefkitap.istanbul.edu.tr/kitap/kok/istatistikiu155.pdf

Teşekkür Ederim

LinkedIn
https://www.linkedin.com/in/cemile-yıldızçakar-34782248/

Email
yildizcakar.cemile@gmail.com

Cemile YILDIZÇAKAR