TANIMLAYICI İSTATİSTİKLER DEĞİŞİM VEYA DAĞILIŞ ÖLÇÜLERİ

DEĞİŞKENLİK NEDİR?

- Bir ana kütleyi (popülasyonu) tanıtmak için, veya başka ana kütlelerle karşılaştırabilmek için merkezi ölçülerin yanında dağılışın genişliğini, değişkenliğin büyüklüğünü gösteren bir başka tipik değerin verilmesi gerekmektedir.

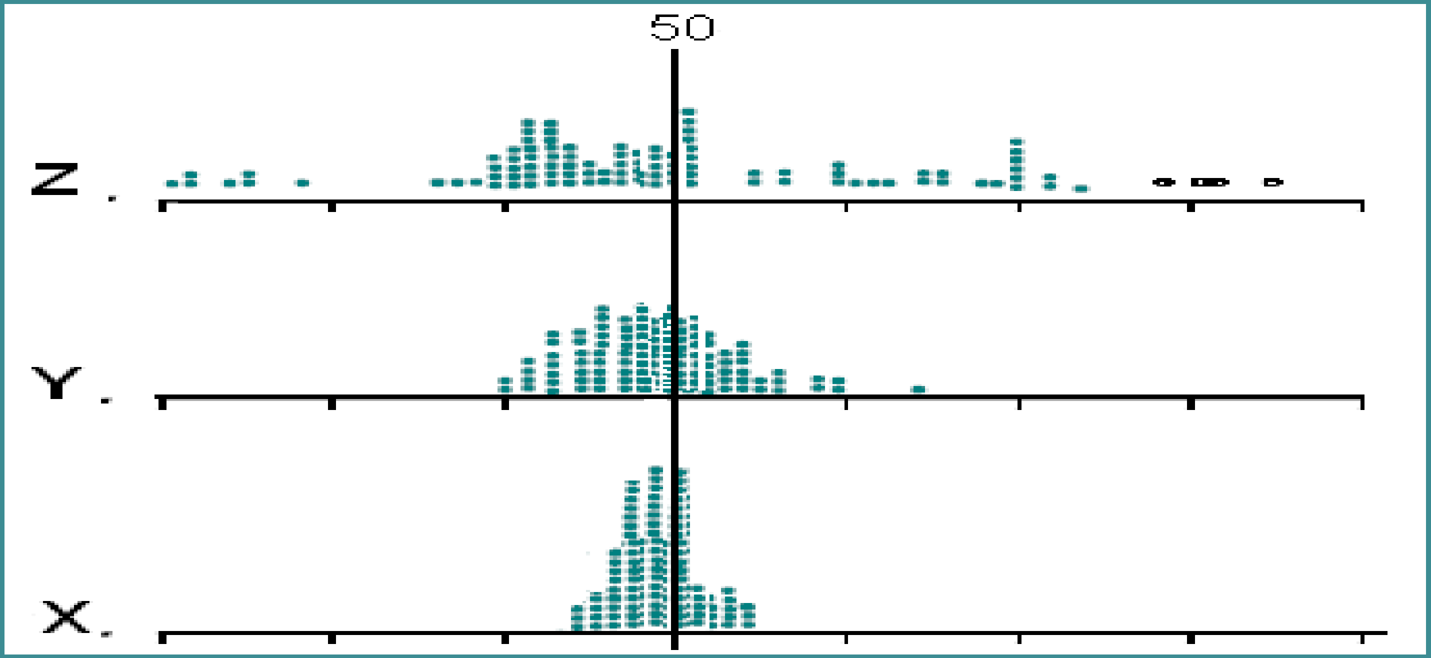
Bu tipik değer değişkenlik ölçüsüdür.

Örnek:

- ► X_i: { 49,49,49,50,51,52 } ; $\bar{X}$=AO(X)= 50
- ► Y_i : { 35,41,50,55,58,61 } ; $\bar{Y}$=AO(Y)= 50
- ► Z_i: { 15,21,33,49,90,92 } ; $\bar{Z}$=AO(Z)= 50

- Bu üç değişkenin ortalaması aynı olduğu halde X, Y ve Z değişkenlerinin aldığı değerlerin   en küçük ve en büyük değerlerine bakıldığında birbirlerinden çok farklıdır.
- Bir popülasyonun gözlem değerlerinde görülen bu değişimin bir istatistikle ifade edilmesi gerekir.

  Dağılış veya değişimleri ifade için kullanılan bu ölçümlere dağılış veya değişim ölçüleri denir.



X değerleri 50 nin etrafında çok yakın kümelendiği halde, Y değerleri biraz daha dağınık, Z değerleri 50 den çok uzakta yer almakta, yani daha büyük değişim göstermektedir.

Dolayısıyla değişkenin ortalama değeri 50 dir demek değişkeni tanımlamak için yeterli olmuyor. Bunun yanında başka ölçütlerde vermek gerekmektedir. Bu ölçüt değişkenlik ölçütü olabilir.

BAŞLICA DEĞİŞİM ÖLÇÜTLERİ

- Değişim Aralığı veya Genişliği (Range)
- Çeyrek Ayrılış (IQR)
- Yarı Çeyrek Ayrılış (SIQR)
- Ortalama Mutlak Sapma
- Variyans
- Standart Sapma
- Standart hata
- Değişim (Varyasyon) katsayısı

DEĞİŞİM GENİŞLİĞİ(RANGE)

En basit dağılış ölçüsü Değişim Genişliğidir.

Değişim Genişliği=DG=R=

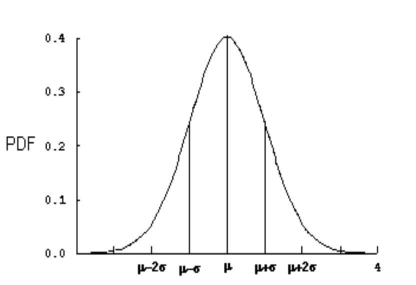
= (En Büyük Gözlem – En Küçük Gözlem)

- $DG(X) = [X_{(enbuy)} - X_{(en\ kuc)}] = 52-49 = 3$
- $DG(Y) = [Y_{(enbuy)} - Y_{(en\ kuc)}] = 61-35 = 26$
- $DG(Z) = [Z_{(enbuy)} - Z_{(en\ kuc)}] = 92-15 = 77$

Olur.

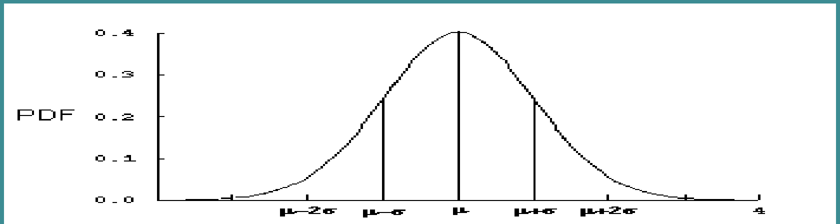
Bu tip değişkenliğe sahip dağılışların grafiklerine bakıldığında aşağıdaki şekilde bir görünüş söz konusudur.

X



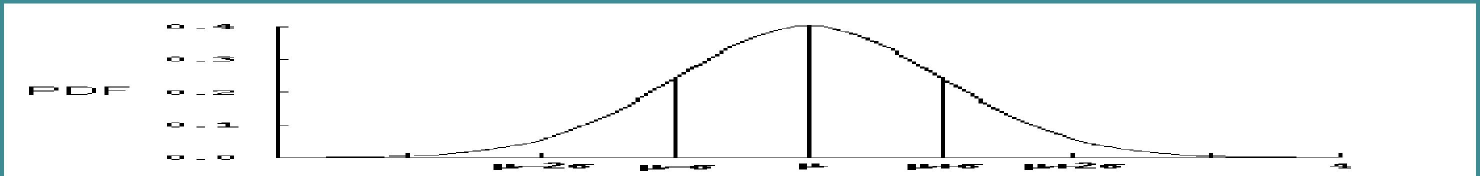
PDF
0.4
0.3
0.2
0.1
0.0
μ-2σ
μ-σ
μ
μ+σ
μ+2σ
4

Y



PDF
0.4
0.3
0.2
0.1
0.0
μ-2σ
μ-σ
μ
μ+σ
μ+2σ
4

Z



PDF
0.4
0.3
0.2
0.1
0.0
μ-2σ
μ-σ
μ
μ+σ
μ+2σ
4

Değişim genişliği veya buna dayanan ölçütler (IQR, SIQR) aslında çok kullanılan ölçülerdir. örneğin şeker hastasının en küçük ve en yüksek değerleri nedir, ateşli bir hastanın en düşük ve en yüksek ateş derecesi nedir, kişiler bu tip değerlerle sürekli ilgilenilir. Özellikle sıralı ölçekteki verilerin tanıtıcı istatistikleri verilirken bunlardan yararlanılır.

- Ancak değişim genişliğinde sadece iki aşırı uç değer kullanıldığı için istikrarlı bir ölçü değildir, bilgi kaybı çoktur, çünkü verinin büyük bir kısmı kullanılmamaktadır.
- Diğer bir sakıncası değişim genişliği örnek büyüklüğüne çok bağlıdır. Örnek büyüdükçe aşırı değerleri kapsama olasılığı artar.

ÇEYREK AYRILIŞ(IQR),
YARI ÇEYREK AYRILIŞ (SIQR)

- Üçüncü ve birinci çeyrekliklerin farkı çeyrek ayrılıştır. IQR $=(Q_3-Q_1)$
- Yarı Çeyrek Ayrılış (SIQR=(Q3-Q1)/2)
- Bu istatistikler genelde sıralama ölçeği ile ölçülen özelliklerin tanımlayıcı istatistiklerinde değişkenliği tanımlarken kullanılır.

Sıralama ölçeğine sahip değişkenlerde merkezi bölgeyi tanımlarken medyan kullanılmasının doğru olacağı daha önce belirtilmişti. Değişkenlik ölçüsü olarak da R, IQR veya SIQR kullanılabilir.

ORTALAMA MUTLAK SAPMA

- Ortalama Mutlak Sapma veya Ayrılış

$$OA = \frac{1}{n}\sum_{i=1}^{n}\left|Xi - \bar{X}\right|$$

- Ortalama mutlak sapma tüm veriyi kullandığı için değişim genişliğinden daha güçlü bir istatistiktir.
- Ancak bunun sakıncası da, üzerinde cebirsel işlemlerin yapılmasının zorluğudur. Yani iki grubun (n) leri ve ortalama mutlak sapmaları verilse, bunlar birleştirildiğinde birleşik örneğin ortalama mutlak sapmasını hesaplayın dendiğinde bu bulunamaz.

Bunların yerine verinin tamamını kullanan, üzerinde cebirsel işlemlerin daha kolay yapılabildiği bir değişkenlik ölçüsünün kullanılması daha faydalı olacaktır.

Bu ölçü variyans ve buna dayanan ölçülerdir.

VARİYANS

Variyans değişkenlerin aldığı değerlerin kendi ortalamasından ayrılışlarının karelerinin toplamının (n-1) bölünmesi ile elde edilir. Eğer popülasyon ortalaması (μ) biliniyorsa n bölünerek te elde edilebilir. Ancak genelde μ bilinmez bu nedenle de yansız tahmin için (n-1) e bölünür.

Xi: { $X_1, X_2, \ldots, X_n$} değerler kümesinin variyansı iki farklı şekilde hesaplanabilir:

$$\text{Variyans} = S^2 = \underbrace{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}}_{\text{ilk hesaplama şekli}} = \underbrace{\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right)}_{\text{ikinci hesaplama şekli}}$$

Eşitliğin sağ tarafındaki sonraki formül ile bir önceki tamamen aynıdır, biri diğerinin açılmış halidir.

Örnek:

X_i:{49,49,49,50,51,52}

değişkeninin variyansı nedir?

İLK FORMÜLÜN KULLANILIŞI

X_i	$\left(X_i - \overline{X}\right)$	$\left(X_i - \overline{X}\right)^2$
49	-1,0	1,0
49	-1,0	1,0
49	-1,0	1,0
50	0,0	0,0
51	1,0	1,0
52	2,0	4,0

n= 6

Toplam= $\sum X_i =$ 300 $\quad \sum_{i=1}^{6}\left(Xi - \overline{X}\right)^2 = 8{,}0$

Ortalama = $\sum X_i / n =$ 50 $\quad S^2 = 8/5 =$ 1,6

İKİNCİ FORMÜLÜN KULLANILIŞI

X_i	(X_i^2)
49	2401,0
49	2401,0
49	2401,0
50	2500,0
51	2601,0
52	2704,0
n= 6	
Toplam= 300	15008,0

Variyans =

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

Variyans =

$$S^2 = \frac{1}{6-1}\left(15008 - \frac{300^2}{6}\right) = \frac{8}{5} = 1{,}6$$

Her iki formülden de aynı sonuç elde edilmektedir.

$S^2 = [(49-50)^2 + (49-50)^2 + (49-50)^2 + (50-50)^2 + (51-50)^2 + (52-50)^2]/(6-1) = (1+1+1+0+1+4)/5$

$= 8/5 = 1.6$

Veya ikinci formül kullanılırsa;

$S^2 = [(49^2 + 49^2 + 49^2 + 50^2 + 51^2 + 52^2) - 300^2/6]/5 = (15008-15000)/5 = 8/5 = 1.6$

X'in variyansı olduğunu ifade için $S_x^2 = 1.6$ yazılabilir. Sonuç aynıdır.

Aynı şekilde Y nin ve Z nin variyansları da hesaplanabilir.

$S_y^2 = 103.2$, $S_z^2 = 1144$ bulunur.

FREKANS(SIKLIK) TABLOSUNDAN VARİYANS HESAPLANMASI

Daha önce merkezi ölçülerin hesaplanmasında verilen örnek kullanılırsa, “yüz yeni doğan çocuğun doğum ağırlıkları (kg) ile ilgili sıklık tablosundaki gruplandırılmış olan veri setine” ait variyans nasıl hesaplanacaktır ?

ÇOCUK DOĞUM AĞIRLIĞI SIKLIK TABLOSU

Sınıf sınırları	Sınıf Değeri (Xi)	Sıklık (frekans= fi)
1.45 - 1.94	1.7	2
1.95 - 2,44	2,2	18
2,45 – 2,94	2,7	24
2,95 – 3,44	3,2	19
3,45 – 3,94	3,7	18
3,95 – 4,44	4,2	9
4,45 – 4,94	4,7	6
4,95 – 5,44	5,2	4
		100

SINIFLANDIRILMIŞ VERİLERDE VARİYANS HESABI

Sınıf sınırları	(Xi)	(fi)	fi*Xi	fi* $(Xi)^2$
1.45 - 1.95	1.7	2	3,4	5,78
1.95 - 2,45	2,2	18	39,6	87,12
2,45 – 2,95	2,7	24	64,8	174,96
2,95 – 3,45	3,2	19	60,8	194,56
3,45 – 3,95	3,7	18	66,6	246,42
3,95 – 4,45	4,2	9	37,8	158,76
4,45 – 4,95	4,7	6	28,2	132,54
4,95 – 5,45	5,2	4	20,8	108,16
	Σ	100	322	1108,3

$$\text{Variyans} = \frac{1}{\sum_{i=1}^{k} fi - 1}\left(\sum fi.Xi^2 - \frac{(\sum fiXi)^2}{\sum_{i=1}^{k} fi} \right)$$

$$S^2 = \frac{1}{100-1}\left(1108.3 - \frac{322^2}{100}\right) = 0.722$$

Variyanstan hesaplanan diğer değişkenlik ölçütleri Standart Sapma, Standart Hata ve Varyasyon(Değişim) Katsayısı çok sık kullanılan değişkenlik istatistikleri olarak karşımıza çıkar.

Bunların nasıl hesaplandığı üzerinde duralım.

STANDART SAPMA

Standart sapma variyansın kare köküdür. Variyans birimsiz ifade edildiği halde standart sapma ölçülen özelliğin birimi ile ifade edilir.

Ortalama etrafındaki saçılma fazla ise variyans büyük olacak, dolayısıyla standart sapma da büyük çıkacaktır. Variyans hesaplanırken ortalamadan ayrılışların kareleri üzerinden hesap yapıldığı için aşırı değerlerin variyans üzerindeki etkisi ortalamaya yakın değerlerden daha fazla olur.

- Veri seti içinde aşırı değer fazla ise bunların variyansı etkilemesi olağan karşılanabilir, ancak birkaç aşırı değerin variyansı büyütmesi olağan karşılanmaz.
- Aşırı değerlerin az olduğu durumlarda ortalama yerine ortanca, standart sapma yerine de çeyrek ayrılışın kullanılması uygun olur.

Ancak çeyrek sapma hesaplanırken tüm veriler yerine sadece dörtte birlik sınırlar (Q1 ve Q3) kullanıldığından standart sapmadan daha zayıf bir dağılış ölçüsüdür. Ortanca kullanmanın gerekli olduğu hallerde çeyrek sapmada değişim ölçüsü olarak kullanılır.

ÖRNEK:

Xi:{49,49,49,50,51,52} veri seti için variyans daha önce hesaplandı. Bunun standart sapması(S) nasıl bulunur? Variyansın karekökü standart sapmadır.

$$S = \sqrt{Variyans} = \sqrt{S^2}$$

$$\text{Standart sapma} = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}} = \sqrt{= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right)}$$

Standart sapma S veya S_x şeklinde değişken ismi ile birlikte yazılabilir.

$S = \sqrt{Varyans} = \sqrt{S^2}$ $=\sqrt{1.6} =$ **1.26**,

Aynı şekilde Y ve Z değişkenlerinin standart sapmaları hesaplanırsa:

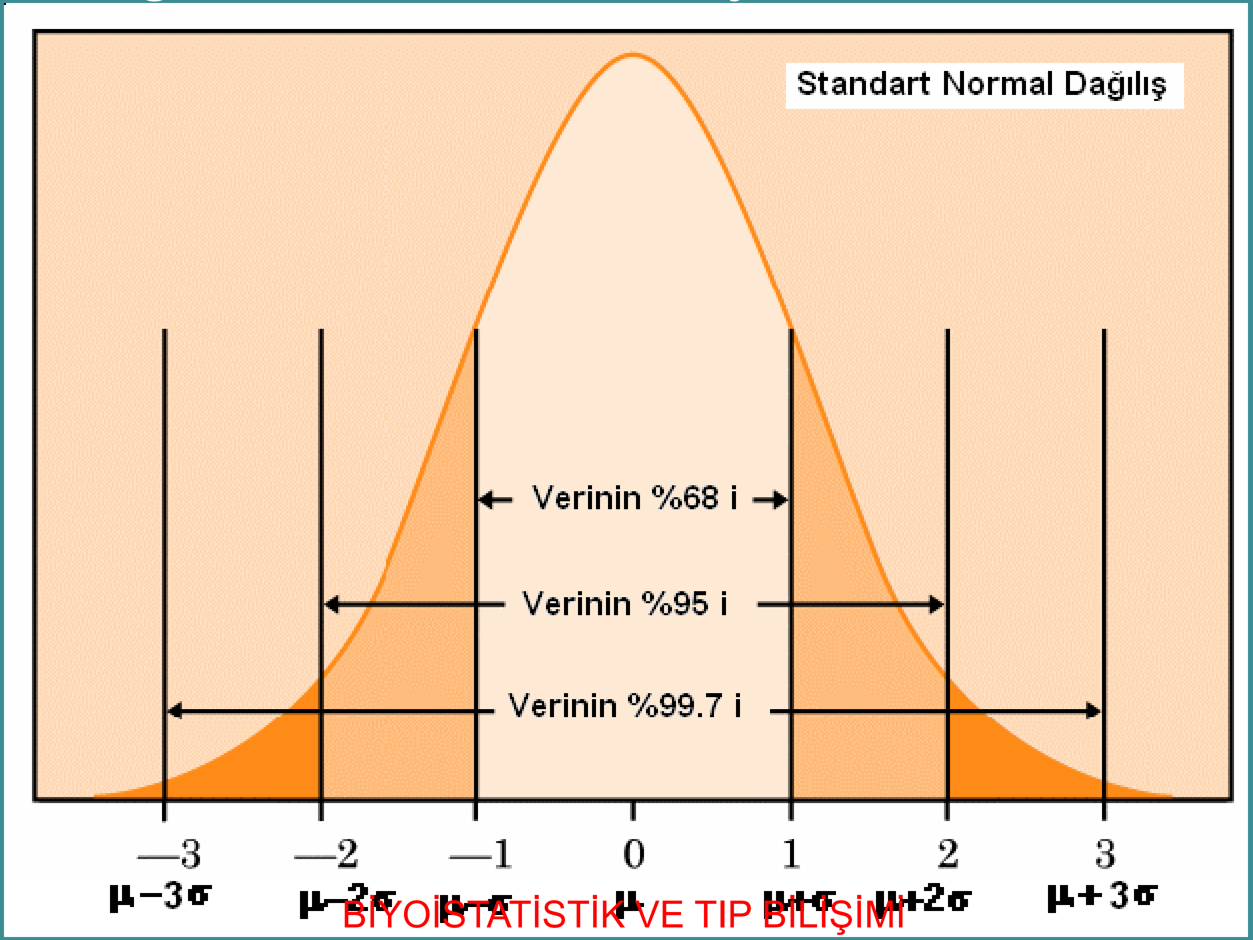
$S_y = \sqrt{103.2} =$ **10.16**,

$S_z = \sqrt{1144} =$ **33.82** bulunur.

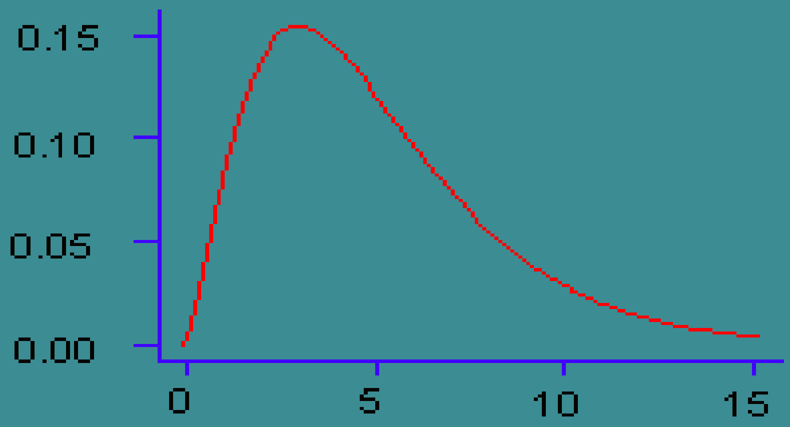
$\bar{X} \mp S_x$ 50±1.26 ifadesi kullanılabilir

Standart sapma <u>gözlemlerin değişkenliği</u> ile ilgili bir göstergedir.

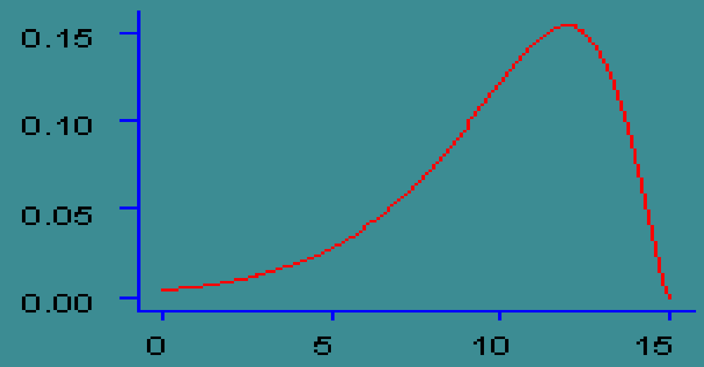
Normal dağılışta gözlemlerin %68.27 si ortalamanın ±1 standart sapma, gözlemlerin %95.45 i ±2 standart sapma, %99.7 si ±3 standart sapma sağında ve solunda yer alır.



ÇARPIK DAĞILIŞLAR



Pozitif(sağa) Çarpık



Negatif(sola) çarpık

Çarpık dağılışlarda bu tip bir ilişkiden bahsetmek söz konusu değildir. Bu nedenle çarpık dağılışlarda güven sınırından bahsetmek anlamsız olur.

STANDART HATA

$$\text{Standart Hata} = \sqrt{\frac{\text{Variyans}}{n}} \Rightarrow S_{\bar{x}} = \sqrt{\frac{S^2}{n}}$$

$$S_{\bar{x}} = \sqrt{\frac{S^2}{n}} = \sqrt{\frac{1.6}{6}}$$

$$= 0.52$$

Standart sapma ve standart hata Normal dağılış gösteren özellikler için oldukça kullanışlıdır.

GÜVEN SINIRI

Ortalamanın dağılışı: $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ ise

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Şeklinde standardize edilir.

Dağılış simetrik ise ana kütle ortalamasına (μ) ait güven sınırı hesaplanabilir.

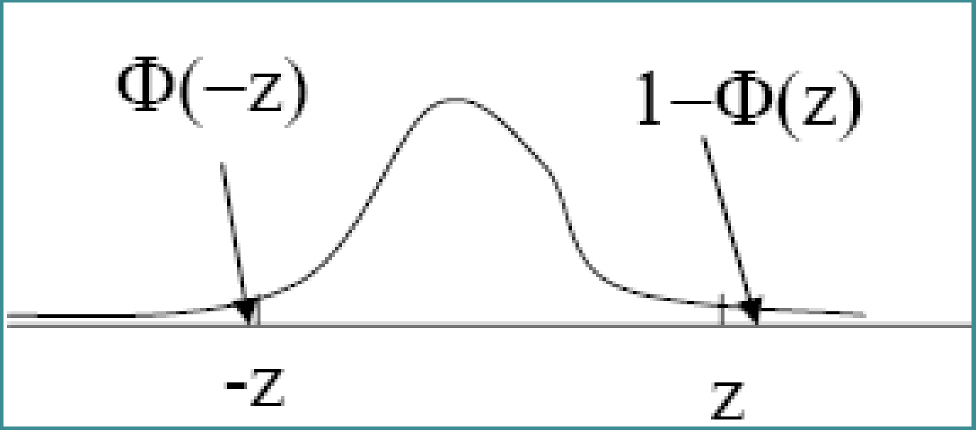
$$P\left[\bar{X} - Z_{(\alpha/2)} \cdot S_{\bar{X}} \leq \mu \leq \bar{X} + Z_{(\alpha/2)} \cdot S_{\bar{X}}\right] = \%(1-\alpha)$$

Burada Z standart normal dağılışın (alfa düzeyindeki) tablo değeridir.

GÜVEN SINIRI

$$P\left[-z_{\alpha/2} \leq Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right] = 1-\alpha$$

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$



X DEĞİŞKENİ İÇİN %95 LİK GÜVEN SINIRLARI

$\overline{X}=50$

$S_{\overline{X}}=0,52$

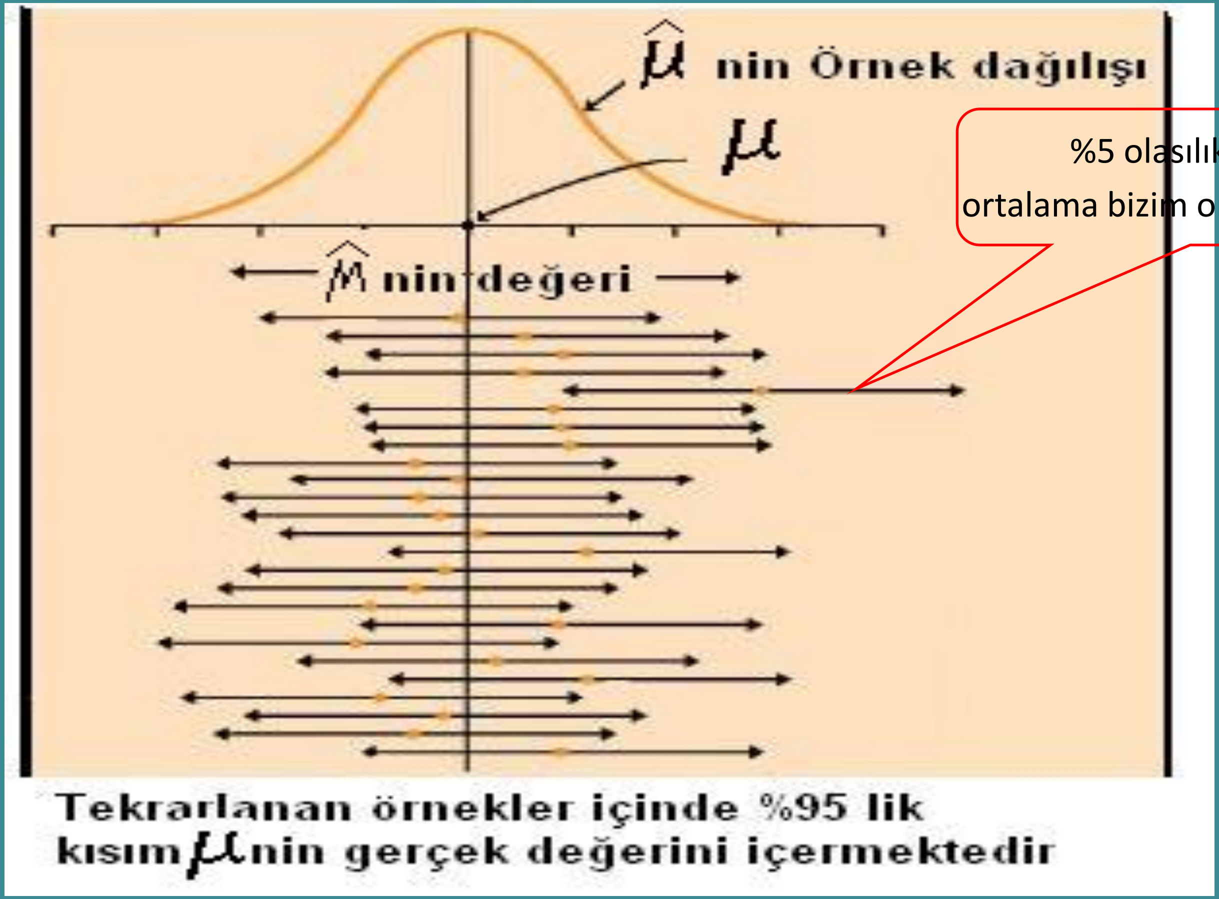
$Z_{(0.025)}=1,96$

$$L=\overline{X}-Z_{(\alpha/2)}\cdot S_{\overline{X}} =50 - 1,96 \times 0,52= 48,98$$

$$U=\overline{X}+Z_{(\alpha/2)}\cdot S_{\overline{X}} =50 + 1,96 \times 0,52= 51,02$$

Alt güven sınırı=48,98, üst güven sınırı=51,02 Olur.

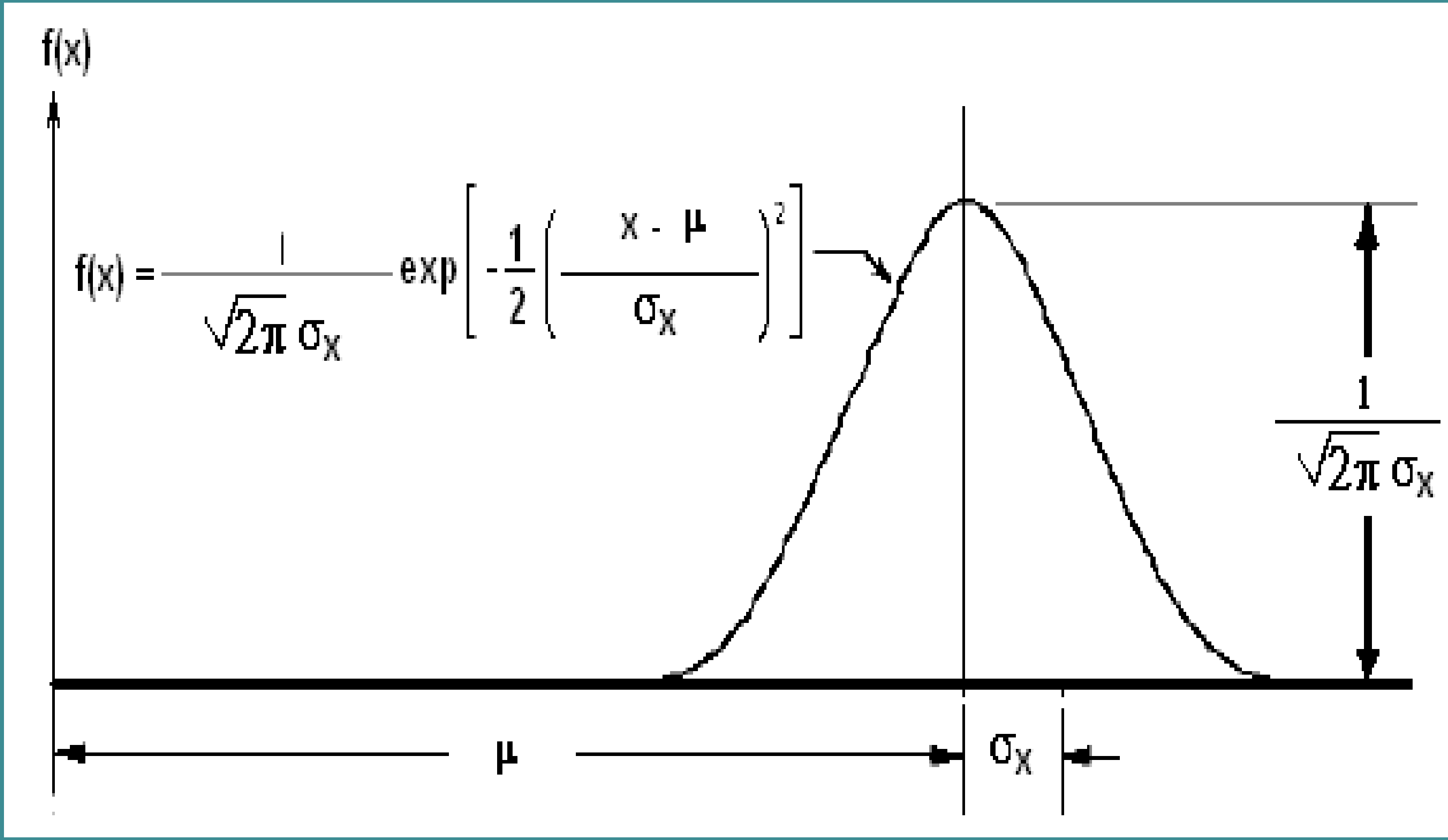
GÜVEN SINIRININ ANLAMI



%5 olasılıkla böyle bir ortalama bizim ortalamamız olabilir!

Bulduğumuz güven sınırının gerçek ortalamayı (μ) kapsadığını %95 güvenle söyleyebiliriz.

Normal Dağılış Olasılık Yoğunluk Fonksiyonu



X, Y ve Z değişkenlerine ait İstatistikler

	N	DG	Ortalama	s^2 veya s_x^2 Variyans	s_x Standart Sapma	$s_{\bar{x}}$ Standart Hata
X	6	3	50	1,6	1,26	0,516
Y	6	26	50	103,2	10,16	4,147
Z	6	77	50	1144,0	33,82	13,808

Bir serinin ortalaması verilirken standart hatası ile birlikte verilir, şöyle ki:

Aritmetik Ortalama ± Standart Hata

$$\bar{X} \mp S_{\bar{X}}$$

X_i serisi için bunun ifadesi: 50 ± 0.52.

Y_i serisi için 50 ± 4.15,

Z_i serisi için 50 ± 13.8

şeklindedir olur.

Böylece ortalamaya ait değişkenlik te yanında ifade edilmiş olur.

Standart hata ortalamanın güvenirliliği ile ilgili bir ölçüdür. Standart hata ne kadar küçükse ortalama o kadar güvenilirdir.

DEĞİŞİM VEYA VARYASYON KATSAYISI(VK)

Bir serinini standart sapması kendi aritmetik ortalamasının yüzdesi olarak ifade edilirse değişim katsayısı bulunmuş olur.

- V.K= Standart Sapma*100/ Ortalama

$$\text{Varyasyon Katsayisi} = \frac{S}{\overline{X}} * 100$$

- Xi serisi için: V.K= 1.26 *100 /50=%2.52
- Yi için; V.K=10.16*100/50= %20.32
- Zi için; V.K= 33.82*100/50=%67.64 olur.

VARYASYON KATSAYISININ (VK) KULLANIMI

- A ve B gibi iki farklı popülasyondaki değişkenliği karşılaştırmak istersek doğrudan standart sapma veya standart hatalarına bakmak yanıltıcı olabilir.

- Ana kütlelerin ortalamaları büyüklük olarak birbirinden çok farklı ise (fillerin ağırlığımı daha değişkendir, farelerin ağırlığımı ?) standart sapma değerlerini <u>ölçü biriminden bağımsız</u> hale getirmek gerekir. Bunun için Değişim(varyasyon) kullanılır.

Örneğin fillerin ağırlığımı daha değişkendir, farelerin ağırlığımı sorusunun cevabı için V.K. kullanılır.

- İki özellik farklı birimlerle ölçülüyorsa, örneğin üzerlerinde deneme yapılan farelerin kan şekerimi daha değişkendir yoksa vücut ağırlıklarımı sorusu ile karşılaşılsa bunun için varyasyon katsayılarına bakmak gerekir.
- Ortalamanın sıfıra yakın olduğu durumlarda varyasyon katsayısının güvenirliliği azalır. Bu gibi durumlarda dikkatli olmak gerekir.
- %20 nin üzerinde varyasyon katsayısı fazla değişkenlik olduğunun göstergesi olarak görülür.

Örnek:Aynı şahıs üzerinde iki farklı yöntemle (X:autoanalyser, Y:Microenzymatic) ölçülen kolesterol miktarları aşağıda verilmiştir.

- X: { 177,193,195,209, 226} %mg/ml
- Y: {192,197,200,202,209}

Her ikisinin aritmetik ortalaması eşittir.

AO(X)= 1000/5=200, AO(Y)=1000/5=200

V(X)=(201360-200000)/4=340, S=18.44

V(Y)=(200158-200000)/4=39.5, S=6.28

VK(X)=18.44*100/200=%9.22

VK(Y)=6.28*100/200=%3.14 olur.

VARİYANSLA İLGİLİ BAZI ÖNEMLİ HESAPLAMALAR

Veri setini sabit bir sayı ile toplama ve çıkarmanın ortalama ve variyans üzerine etkisi:

C sabit bir sayı ise Yi=Xi+C şeklinde verilerin tamamına C sayısı eklense yeni elde edilen sayının ortalama ve variyansı değişir mi?

- E(Xi)=μ, E(Yi)=E(C+Xi)= E(C)+E(Xi)=C+μ ortalama C kadar artar.
- V(Xi)=σ^2 ise, V(Yi)=V(C+Xi)= V(C)+V(Xi)=0 + σ^2 variyans aynı kalır.

ÖRNEK:

- Xi: { 4, 5, 6, 9} ise

Ortalama(X)= 24/4=6, variyans(X)=4.67

Tüm değerlere c=3 sabiti <u>eklense</u>

- Yi= {7, 8, 9, 12} elde edilir.

Yi nin ortalaması ve variyansı ne olur?

Ortalama(Y)= 36/4=9, variyans(Y)=4.67

Olur.

C sabit bir sayı ise verilerin tamamı C sayısı ile çarpılsa yeni elde edilen sayının ortalama ve variyansı değişir mi?

- $Y_i=C.X_i$ olursa,
- $E(X_i)=\mu$, $E(Y_i)=E(C.X_i)=C.E(X_i)=C\mu$,

ortalama C kat artar.

- $V(X_i)=\sigma^2$ ise, $V(Y_i)= V(C.X_i)=C^2.V(X_i)= C^2. \sigma^2$,

variyans sabit sayının karesi ile çarpımı kadar büyür.

- Xi: { 4, 5, 6, 9} ise
- Ortalama(X)= 24/4=6, variyans(X)=4.67

- Tüm değerlere c=3 sabiti ile çarpılsa
- Yi= {12, 15, 18, 27} elde edilir.
- Yi nin ortalaması ve variyansı ne olur?
- Ortalama(Y)= 72/4=18, variyans(Y)=42,0