



VERİ BİLİMİ İÇİN TEMEL İSTATİSTİK hafta-10

CEMİLE YILDIZÇAKAR

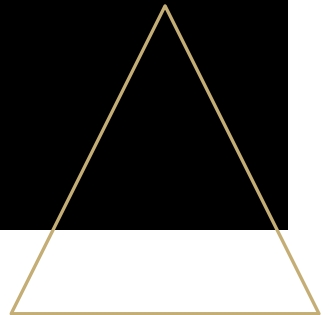
26.02.2021





Sürekli Olasılık Dağılımları(Continuous Probability Distributions)

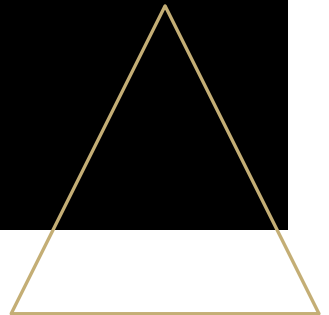
- GAMMA DAĞILIMI
- ÜSTEL DAĞILIM (Exponential Distribution)
- NORMAL DAĞILIM





GAMMA DAĞILIMI

- Rasgele değişken bir poisson işlem içinde k kadar değer oluncaya kadar ki mesafe aralığı gamma dağılımına sahiptir.
- Gamma dağılımı iki parametrelili bir olasılık dağılımıdır.
- Parametrelerden biri ölçek (k),
diğeri ise şekil parametresidir.



- Gamma fonksiyonu:

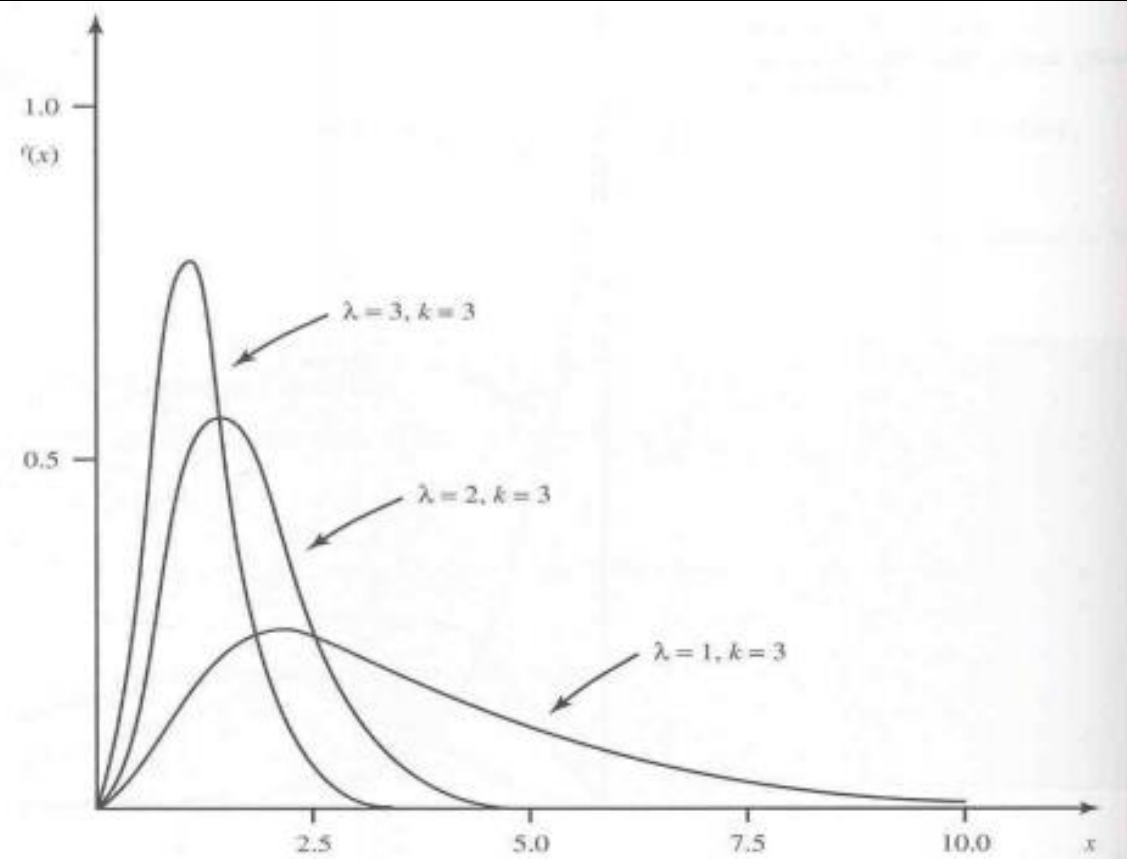
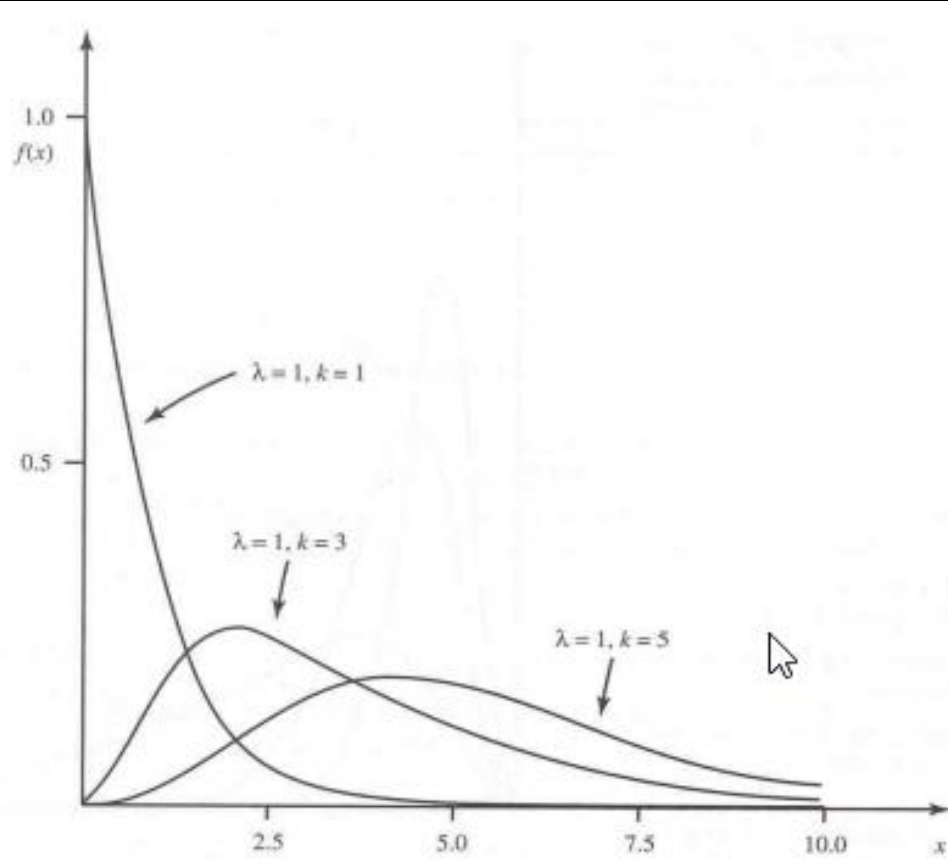
$$\Gamma(k) = \int_0^{\infty} x^{k-1} e^{-x} dx \quad \text{for } k > 0$$

- $k > 0$ ve $\lambda > 0$ için Gamma olasılık yoğunluk fonksiyonu:

$$f(x) = \frac{\lambda (\lambda x)^{k-1} e^{-\lambda x}}{\Gamma(k)} \quad \text{for } x \geq 0$$

- Ortalama ve varyans:

$$E(X) = k / \lambda \quad \text{and} \quad V(X) = k / \lambda^2$$



- Gamma rasgele değişkeninin özellikleri:

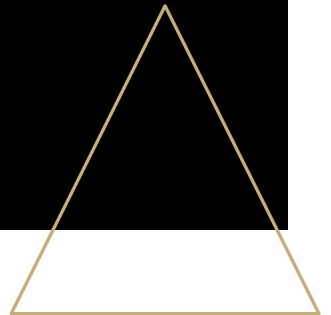
Eğer $X_i, i = 1, \dots, n$, (k_i, λ) , parametreleri ile bağımsız gamma rasgele değişkenleri ise o zaman $\sum_{i=1}^n X_i$, $(\sum_{i=1}^n k_i, \lambda)$. parametresi ile gamma dağılır.

- Parametresi $(1, \lambda)$ olan gamma rasgele değişkeni, parametresi λ . olan üstel rasgele değişkene eşdeğerdir.



ÜSTEL DAĞILIM

- Özellikle sanayi ürünlerinin dayanma sürelerinin incelenmesinde yaygın olarak kullanılan sürekli olasılık dağılımıdır.
- Bağımsız olaylar arasındaki zaman aralığını modelleştirirken bir üstel dağılım doğal olarak ortaya çıkar.
- ✓ Müşterilerin gelişleri arasında geçen zaman
- ✓ Bozulmalar arası geçen zaman
- ✓ Belli bir bölgedeki iki deprem arasında geçen zaman



Bir üstel dağılım için **olasılık yoğunluk fonksiyonu** şu şekli alır:

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & , x \geq 0, \\ 0 & , x < 0. \end{cases}$$

Burada $\lambda > 0$ dağılım için tek parametredir ve çok zaman *oran parametresi* olarak anılır. Dağılım için destek $[0, \infty)$ aralığında verilir. Eğer X **rassal değişkeni** bu üstel dağılım gösteriyorsa bu şöyle yazılır:

$$X \sim \text{Üstel}(\lambda).$$

Ancak bir diğer şekilde değişik parametreleme ile ise üstel dağılım için **olasılık yoğunluk fonksiyonu** şöyle ifade edilir:

$$f(x; \beta) = \begin{cases} \frac{1}{\beta} e^{-x/\beta} & , x \geq 0, \\ 0 & , x < 0. \end{cases}$$

Burada $\beta > 0$ bir **ölçek parametresidir** ve yukarıda tanımlanan *oran parametresi* olan λ 'nın bir üstü değeri **çarpım tersi**, yani $\beta = 1/\lambda$ dır. Bu çeşit tanımlamada β *kalım parametresi* çünkü eğer bir **rassal değişken** X bir biyolojik veya mekanik sistem M için ömür geçirme zaman uzunluğu ise ve $X \sim \text{Üstel}(\beta)$ ise

- Örnek: Bir aracın aküsü tükeninceye kadar alınan yol uzunluğu, ortalaması 10.000 mil olan üstel dağılıma sahip olduğunu varsayalım. Bir kişi 5000 mil yolculuk yapmak istiyorsa, aküsünü değiştirmeden yolculuğunu tamamlayabilme olasılığı nedir?
- X bataryanın kalan ömrünü (bin mil olarak) içeren rasgele bir değişken olsun. O zaman,

$$E[X] = 1/\lambda = 10 \Rightarrow \lambda = 1/10$$

$$P\{X > 5\} = 1 - F(5) = e^{-5\lambda} = e^{-1/2} \approx 0.604$$

X , üstel rasgele değişken değilse ne olur?

$$P\{X > t+5 \mid X > t\} = \frac{1 - F(t+5)}{1 - F(t)}$$

Yani, t hakkında ek bir bilgi bilinmelidir

Hafızasızlık özelliği (Lack of memory)

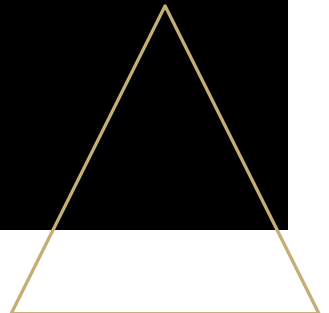
$$\square P(x > 10 \mid x > 3) = P(x > 7 + 3 \mid x > 3) = P(x > 7)$$

$$P(X > s + t \mid X > s) = P(X > t)$$



Üstel dağılım ile Poisson dağılımı arasındaki ilişki

- Poisson olayları arasında geçen süre genellikle üstel dağılım ile açıklanır.
- Kasaya birim zamanda gelen müşteri sayısı: Poisson
- Kasaya gelişler arasında geçen süre: Üstel
- Servis için başvuran müşteri sayısı: Poisson
- Servis için başvuran müşterilerin gelişi arasında geçen süre: Üstel



Poisson

- number of events in a time interval
- Discrete
 $X = 0, 1, 2, \dots$

number of
↓ ↓ lines




Exponential

- time between two events
- Continuous on an interval



NORMAL DAĞILIM

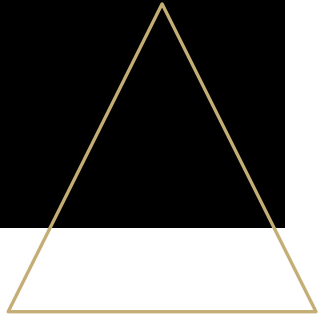




- Normal dağılımın ilk uygulamaları doğada gerçekleşen olaylara karşı başarılı bir biçimde uyum göstermiştir. Dağılımın göstermiş olduğu bu uygunluk adının Normal Dağılım olması sonucunu doğurmuştur.

- İstatistiksel yorumlamanın temelini oluşturan Normal Dağılım, bir çok rassal süreçlerin dağılımı olarak karşımıza çıkmaktadır.

- Normal dağılışı kullanımının en önemli nedenlerinden biride bazı varsayımların gerçekleşmesi halinde kesikli ve sürekli bir çok şans değişkeninin dağılımının normal dağılışa yaklaşım göstermesidir.



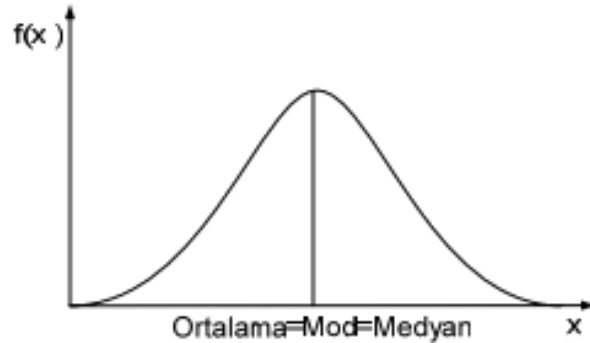
Örnek;

- Yetişkinlerin boy uzunlukları, kütleleri vb.
- • Örneğin 1000 yetişkinin zekâ düzeylerinin dağılımı frekans poligonu üzerinden incelense, gözlemlerin ortalama değeri olan 100 civarında kümелendiği, daha yüksek ve daha düşük IQ'lu birey sayısının daha az olduğu görülecektir.

Normal Dağılımın Özellikleri

- Çan eğrisi şeklindedir.
- Simetrik bir dağılıştır.
- Normal Dağılımın parametreleri,

$$E(x) = \mu \quad Var(x) = \sigma^2$$



Normal Dağılımın Olasılık Yoğunluk fonksiyonu

$$f(x) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} & , -\infty < x < \infty \\ 0 & , \text{diger yerlerde} \end{cases}$$

$$\pi = 3,14159\dots$$

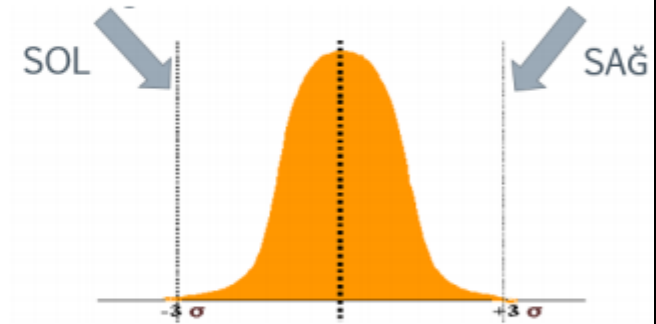
$$e = 2,71828$$

σ = populasyon standart sapması

μ = populasyon ortalaması

Normal Dağılım Eğrisinin Özellikleri

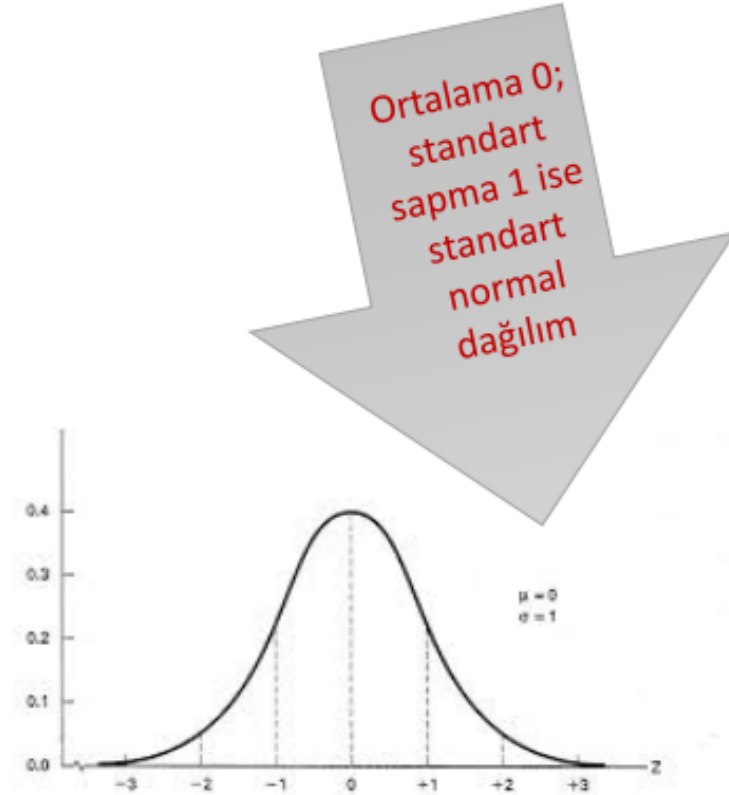
- Eğri, dikey eksene göre **simetriktir**. Puanların yarısı eksenin sağ, diğer yarısı da sol tarafındadır.
- Puanlar merkez etrafında kümelenme eğilimi gösterir.
- Mod, ortanca ve ortalama birbirine eşittir.
- Dağılımın her iki ucu giderek yatay eksene yaklaşır, ancak hiçbir zaman bu eksene değmez (asimptomatik). Normal dağılım eğrisi atındaki alan sınırsızdır.



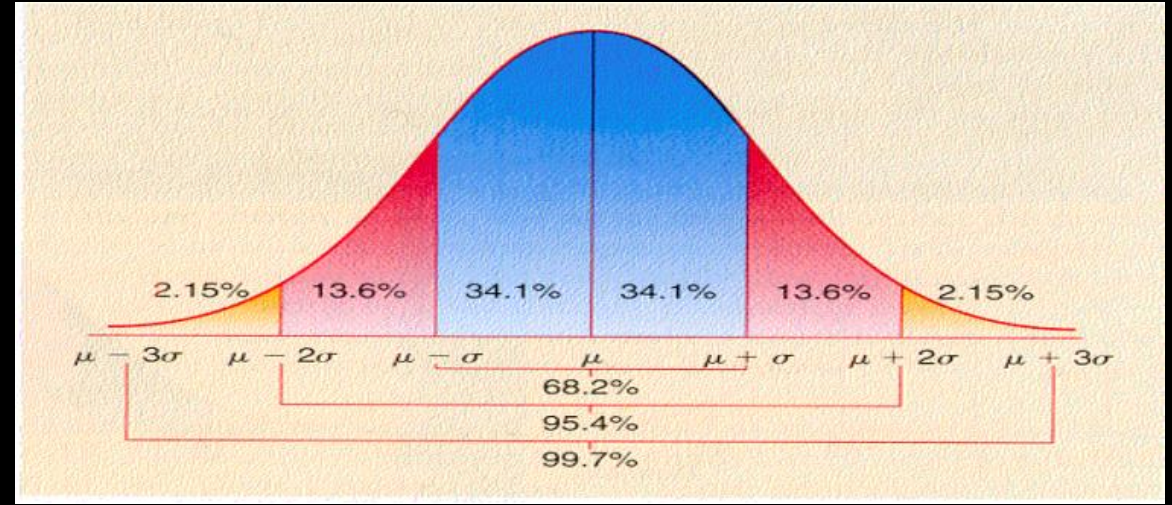
(Ferguson ve Takane, 1989; Ravid, 1994)

Standart Normal Dağılım

- Standart normal dağılımda, ortalama 0, standart sapma 1'dir.
- Ortalamanın sol tarafındaki (altındaki) birimler negatif, sağındakiler pozitiftir.
- İki standart sapma arasındaki uzaklıklar birbirine eşittir.
- İki standart sapma arasında kalan alanlardan merkeze yakın olanlar, uzak olanlara göre daha fazla puan kapsar (Ravid, 1994).

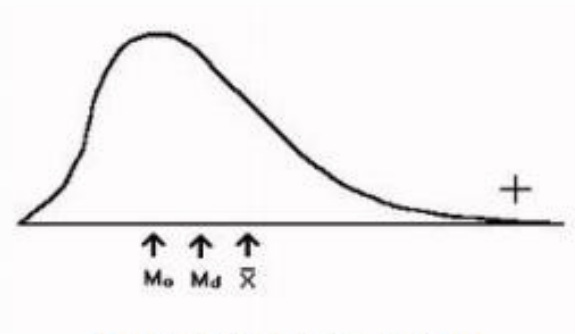


- Ortalama: 25 ve Standart Sapma: 5 olsun;
 - Puanların % 68.3'ü 20 ile 30 puan arasındadır.
 - Puanların % 95.4'ü 15 ile 35 puan arasındadır.
 - Puanların % 99.7'si 10 ile 40 puan arasındadır.
 - Puanların % 47.7'si 25 ile 35 puan arasındadır. (*2'den %95.4'ü 15 ile 35 arası)
 - Puanların % 49.8'i 10 ile 25 arasındadır

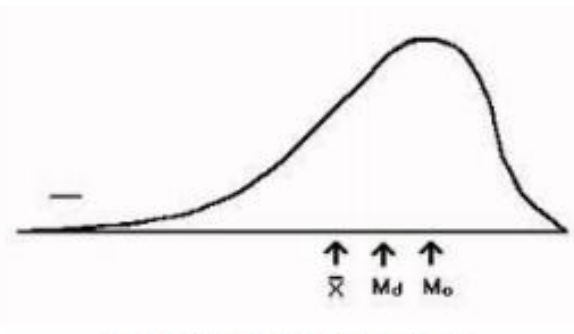


ÇARPIK VE BASIK DAĞILIMLAR

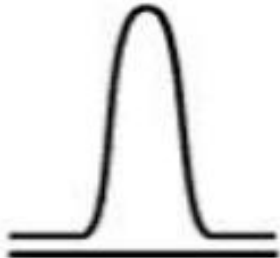
Aşağıda normal dağılımdan farklılaşan dağılımlar, dağılımın şekilleri ile gösterilmiştir.



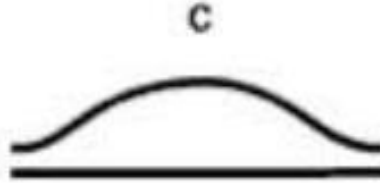
Şekil 1. Sağa Çarpık Dağılım



Şekil 2. Sola Çarpık Dağılım



Şekil 3. Sivri Dağılım



Şekil 4. Basık Dağılım

DAĞILIM NORMALLİĞİNİN İNCELENMESİ

- a) Verilerin normal dağılım gösterip göstermediğini belirlemenin yollarından biri dağılımın grafiğini çizmek ve bu grafiği yorumlamaktır.
- b) Verilerin dağılımının normal dağılım gösterip göstermediğini belirlemenin bir diğer yolu ortalama, mod ve medyan değerlerine bakmaktır. Normal dağılımda bu değerler çakışıktır. Bu istatistikler birbirine yaklaştığı ölçüde dağılım normal dağılıma yaklaşır. Birbirinden uzaklaştığı ölçüde dağılım çarpıklaşır. Fakat bu yakınlığın düzeyi ile ilgili belirli bir standart yoktur. Bu nedenle burada verilen diğer yöntemlerle birlikte değerlendirilmesi önerilir.
- c) Normal dağılımı test etmenin bir diğer yolu da basıklık ve çarpıklık katsayılarına bakmaktır. Çarpıklık (skewness) katsayısı normal dağılımda 0'dır. Negatif çarpıklık katsayısı sağa çarpık dağılıma, pozitif çarpıklık katsayısı sola çarpık dağılıma işaret eder. Basıklık (kurtosis) katsayısı da normal dağılımda 0'dır. Pozitif basıklık katsayısı sivri dağılıma, negatif basıklık katsayısı ise basık bir dağılıma işaret eder. Dağılımın normal dağılımdan manidar düzeyde farklılaşmıyor olması için bu değerlerin (-1, +1) aralığında kalması beklenir.

Normal Dağılımın önemi

- ❑ Birçok istatistiksel test Normal dağılım varsayımına dayanmaktadır. Bunun anlamı, ancak normal dağılım varsayımı sağlandığı takdirde, bu testler en iyi sonucu vermektedir.
- ❑ Sadece güçlü 'robust' olarak tanımlanan testler, normalden sapmaları tolere etmektedir.

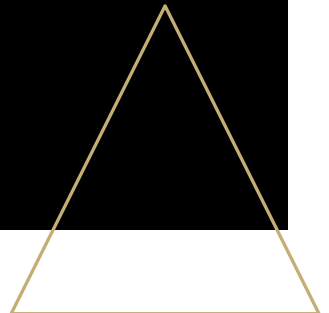
Normallik Sınaması (Assessing Normality)

- Normallik sınaması, verilerin normal dağılıp dağılmadığına karar verme sürecidir.
- Normal dağılan bir kitleden çekilen bir örneklem her zaman normal dağılmayabilir.
- Örneklem her seferinde değiştiğinden , her örneklemin dağılımı da değişir.
- İstatistiksel testler, küçük veri setleri için ($n < 30$) kullanıldığında, kitlenin normal ya da normale yakın dağılım gösterdiği varsayımı yapılır. Küçük veri setlerinin histogramları kitlenin dağılımını her zaman yansıtmayabilir. Bu sebeple, kitlenin normalliğini sınamak için farklı yöntemlere ihtiyaç duyulmaktadır.
- Bununla birlikte, örneklem normal dağılan bir kitleden geliyor ve yeterli büyüklüğe sahip ise, dağılımı normale yakın olabilir.



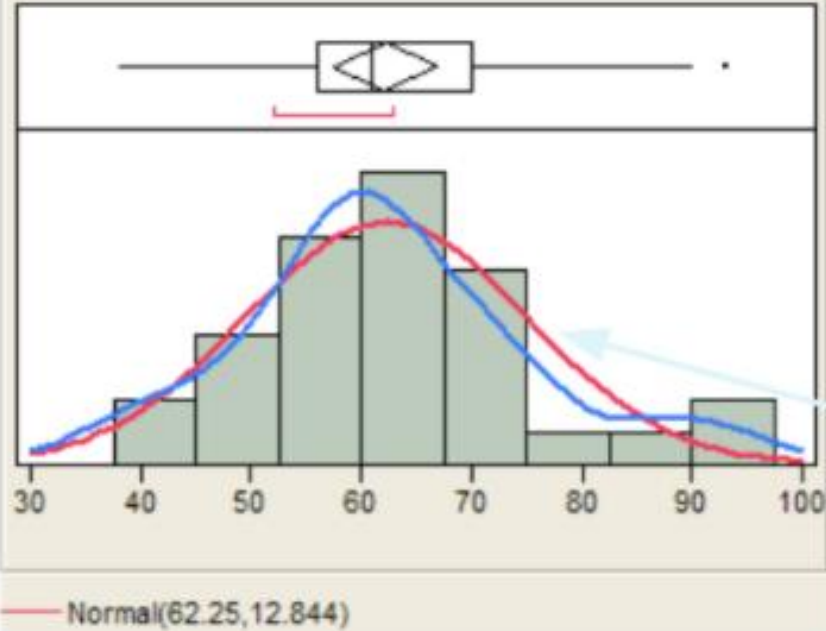
Normallik Sınaması için Kullanılan Yöntemler

- Histogram
- Boxplot
- Normal Quantile Plot (Normal Olasılık Grafiği)
(Normal probability Plot))
- Uyum İyiliği Testleri (Goodness of Fit Tests)



Histogram ve Boxplot

Cholesterol Levels of Male Heart Patients



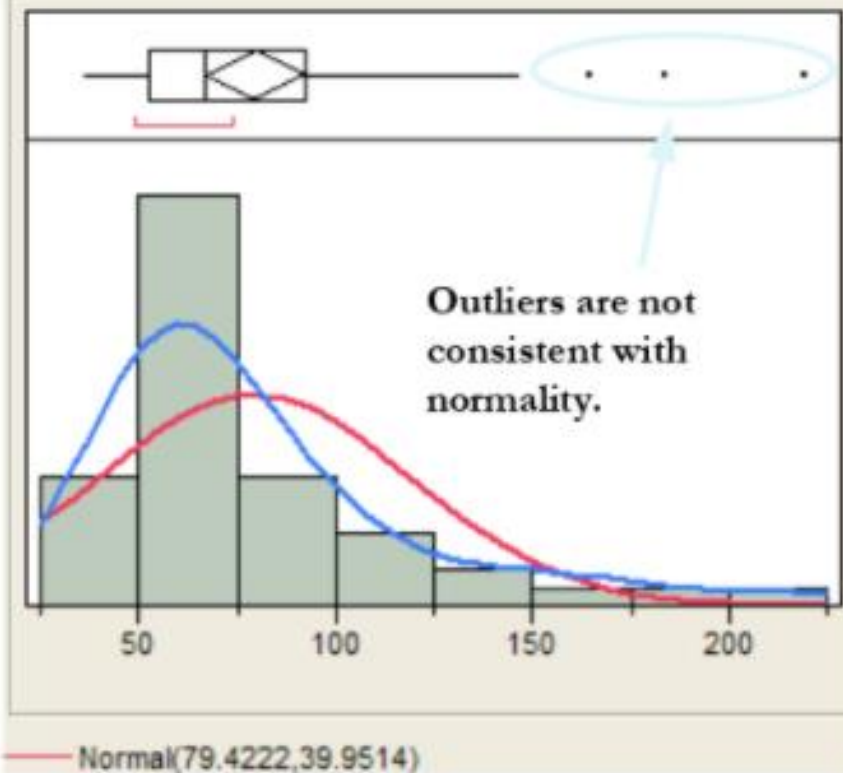
$$\bar{X} = 62.25, \quad s = 12.84$$

Sağdan çarpıklık için ortalamanın ortancadan büyük olması gibi bazı kanıtlar olmasına rağmen hastaların kolesterol seviyesi yaklaşık olarak normal dağılmaktadır.

Kırmızı eğri normal dağılımın bu dağılıma fit edilmiş halidir. Mavi eğri ise bu verilerin olasılık yoğunluk fonksiyonu kestirimidir. Bu veri normal dağılıyor olsaydı, iki eğri üst üste gelecekti.

Histogram and Boxplot

Systolic Volumes of Male Heart Patients



► $\bar{X} = 79.42, s = 39.95$

Histograma göre; Erkek kalp hastalarının sistolik değerleri bu örneklemin sağdan çarpık bir dağılan bir kitleden çekildiğini göstermektedir.

Uyum İyiliği Testleri (Goodness of Fit Tests)

□ Ampirik Dağılım Fonksiyonuna Dayalı Testler

- Kolmogorov-Smirnov Test
- Kuiper's Test
- Lilliefors Test
- Cramér-Von Mises Test
- Anderson-Darling Test
- Watson Test

□ Regresyon ve Korelasyona Dayalı Testler:

- Shapiro-Wilk Test

Teşekkür Ederim



LinkedIn

<https://www.linkedin.com/in/cemile-yildizcakar-34782248/>



Email

yildizcakar.cemile@gmail.com

Cemile YILDIZÇAKAR

A life without love
is like a year
without summer.

A SWEDISH PROVERB