Introduction to Statistics

1. İstatistik Nedir?

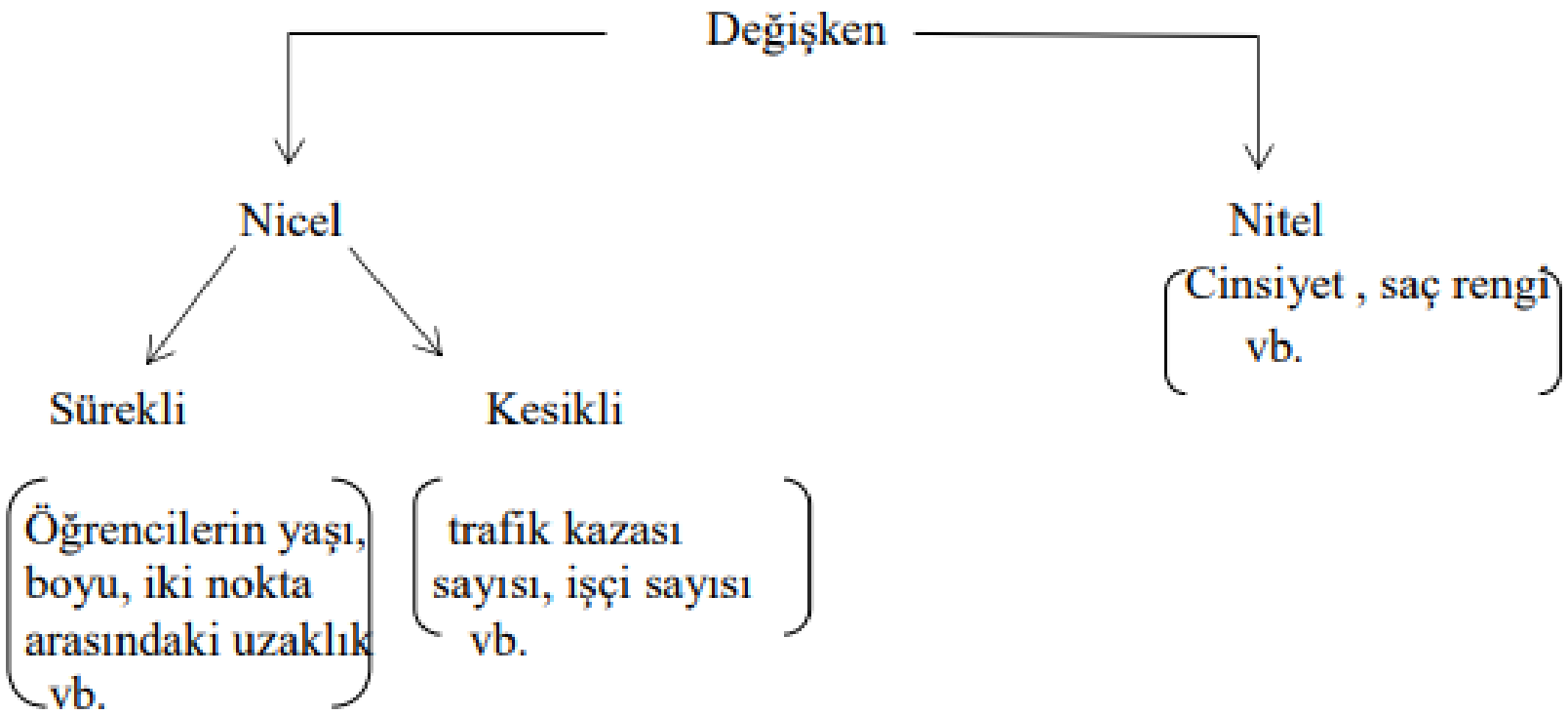
- “İstatistik, rastgelelik içeren olaylar, süreçler, sistemler hakkında modeller kurmada, gözlemlere dayanarak bu modellerin geçerliliğini sınamada ve bu modellerden sonuç çıkarmada gerekli bazı bilgi ve yöntemleri sağlayan bir bilim dalıdır”

- İstatistik; tüm bilim dallarında, çalışma alanlarında bir amaca yönelik, sayısal değerleri derleme, özetleme, tablolar ya da grafikler biçiminde düzenleme, çözümleme, sonuçları yorumlama, parametre tahmini yapma, değişkenler arasındaki ilginin derecesini belirleme, örneklemler arasındaki ayrıcalığı belirleme, deney düzenleme gibi konuları kapsayan bir bilim dalıdır.

İstatistiğin Temel Kavramları

1. **Kitle:** Araştırma kapsamında alınan, aynı özelliği taşıyan birim ya da bireylerin oluşturduğu topluluğa kitle denir.
2. **Örneklem:** Örnekleme yöntemlerinden yararlanılarak bir kitleden seçilen, aynı özellikleri taşıyan ve kitleyi temsil edebilecek nitelikteki ve nicelikteki bireylerin oluşturduğu topluluğa örneklem denir. Örneklem seçmek için kullanılan yöntemler topluluğuna "örnekleme" denir.
3. **Parametre:** Kitlenin özelliklerini belirleyen sayısal değerlerdir
4. **Değişken:** Birim ya da bireylerin her bir özelliğine değişken denir.
5. **Veri:** Üzerinde çalışılan birim ya da bireylerin her birinin aldığı değerlere veri denir. Veriler iki grupta toplanır; nitel veriler, nicel veriler. Nicel veriler de kendi içerisinde kesikli ve sürekli olmak üzere ikiye ayrılır. Nitel veriler ise nitelik belirten değişkenlere ilişkin verilerdir.

İstatistiksel Değişkenler



İstatistiğin Çeşitleri

- İstatistik, teorik istatistik ve uygulamalı istatistik olarak iki başlık altında toplanır.
- Uygulamalı istatistik betimsel istatistik ve çıkarsamalı istatistik olarak iki alanda ele alınır.

- **Betimsel İstatistik:** Verinin tanımlanması ve özetlenmesi.
- **Çıkarsamalı istatistik:** Örneklemden elde edilen bilgilere dayanarak kitle hakkında tahminlerde bulunmak, sonuçlar çıkarmak ve kararlar vermek.

Ölçüm Düzeyleri

- **Sınıflama (Nominal) Ölçme Düzeyi:** Birimlere özelliklerine göre isimler verilir.
- **Sıralama (Ordinal) Ölçme Düzeyi:** Sıralama ölçme düzeyinde değişkenlerin aldığı değerler önem derecesi ya da üstünlüklerine göre sıralanır.
- **Eşit Aralıklı Ölçme Düzeyi (Interval):** Sıcaklık, başarı, performans gibi nicel değişkenleri ölçmek için kullanılır. Bu ölçekte bir başlangıç noktası bulunmaz.
- **Oranlama Ölçme Düzeyi (Ratio):** ekonomik durum, ağırlık, uzunluk, hız, not gibi değişkenleri ölçmek için kullanılır.

Merkezi Eğilim Ölçüleri

- Merkezi eğilim ölçüleri kitleye ilişkin bir değişkenin bütün farklı değerlerinin çevresinde toplandığı merkezi bir değeri gösterirler. Dağılım ölçüleri ise değişkenin aldığı değerlerin birbirinden ne kadar farklı olduğunun ölçüsüdür.

- Aritmetik Ortalama
- Mod (Tepe Değer)
- Medyan
- Çeyreklikler

Aritmetik Ortalama

- μ: kitleye ilişkin aritmetik ortalama
- $\bar{x}$: örnekleme ilişkin aritmetik ortalama
- Simetrik dağılım gösteren verileri en iyi temsil eden bir merkezi eğilim ölçüsüdür

Kitle Ortalaması	Örneklem Ortalaması
$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Aritmetik ortalamanın bize ne anlatır?

- Bir veri seti için sadece bir aritmetik ortalama vardır.
- Nicel verilere uygulanabilir.
- Birim değerlerinde meydana gelen değişim çok küçük olsa bile aritmetik ortalamayı etkiler.
- Aritmetik ortalama ile birim değerleri arasındaki farkların toplamı sıfırdır.
- Aritmetik ortalama ile birim değerleri arasındaki farkların kareleri toplamı minimum bir değerdir.

- Örnek: 3,5,4,3,6,7,1 veri setinin aritmetik ortalaması nedir?

Mod (Tepe Değer)

- Bir veri setinde en çok tekrar eden değere denir.
- Denek sayısı az olduğunda tepe değer güvenilir bir ölçü değildir.
- Bazı örneklemlerde bir tepe değer yerine iki ya da daha çok tepe değer olabilir. Bu durumda ya tepe değerini hesaplamaktan vazgeçilir ya da frekans tablosu tek tepe değerli bir dağılım olacak şekilde yeniden düzenlenir.
- Tepe değer hesaplanırken birimlerin tümü işleme katılmadığı için uç değerlerden etkilenmez.
- Nicel ve nitel verilerin her iki türü için de uygundur

- Örnek: 1,6,3,1,95,2,6,2,1,7,9 veri setinin modu nedir?

Medyan (Ortanca)

- Bir veri gurubundaki değerlerin küçükten büyüğe sıralandığında tam ortaya düşen değer ortanca denir. Örneklem hacmi tek ise ortanca tam ortaya düşen verinin değeridir. Örneklem hacmi çift ise tam ortaya düşen iki değerin aritmetik ortalamasıdır.

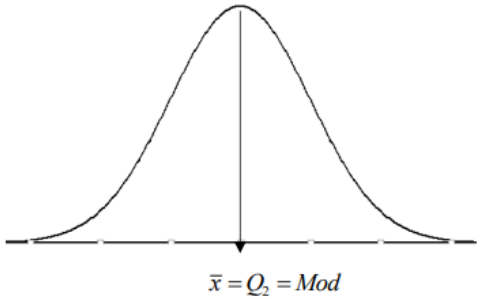
$$\text{Medyan } Q_2 = \begin{cases} \dfrac{x_j + x_{j+1}}{2},\ j = \dfrac{n}{2},\ n \text{ çift ise} \\ x_j\ ,\ j = \dfrac{n+1}{2},\ n \text{ tek ise} \end{cases}$$

- Aşırı uç değerlerden etkilenmez.
- Birim değerleri ile ortanca arasındaki farkın yarısı negatif yarısı pozitiftir.
- $\sum|xi - ortanca| = minimum$ dur.

Aritmetik Ortalama, Tepe Değer ve Ortanca Arasındaki İlişki

- Aritmetik ortalama, ortanca ve tepe değer arasındaki ilişki verilerin dağılımının çarpıklığı hakkında bilgi verir.

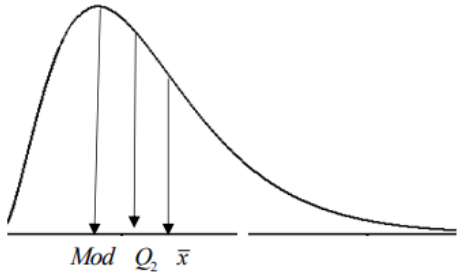
i) Bir sıklık dağılımı göze alındığında, dağılım simetrik olduğunda;



$\bar{X} = Q_2 = TD$

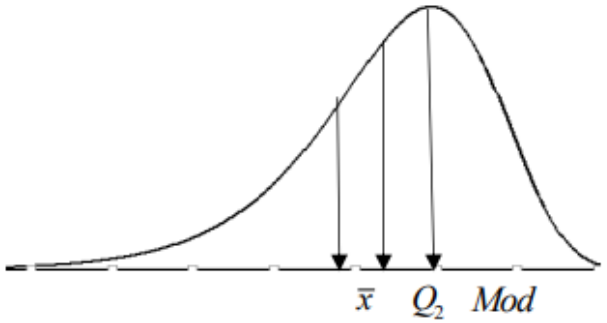
dağılım simetrik

ii) Sağa çarpık ya da pozitif yöne eğimli ise



$TD < Q_2 < \bar{X}$

dağılım sağa çarpık ya da pozitif yöne eğimli



$\bar{X} < Q_2 < TD$

dağılım sola çarpık ya da negatif yöne eğimli

Çeyreklikler

- Küçükten büyüğe doğru sıralanmış verileri dört eşit parçaya bölen değerlere çeyrek değerler denir.
- Birinci çeyreklik (Q1), veriler küçükten büyüğe sıralandığında verilerin %25 ini sağında, %75 ini solunda bırakan değerdir.
- İkinci çeyreklik ortancaya denk gelmektedir.
- Üçüncü çeyrek değer (Q3), veriler küçükten büyüğe sıralandığında verilerin %75 ini sağında, %25 ini solunda bırakan değerdir.
- Yani sıralı verilerde, ortancadan küçük olan değerlerin ortancası birinci çeyrek değer, ortancadan büyük olan verilerin ortancası üçüncü çeyrek değerdir.
- Çeyrekliklerin önemini ilerde IQR yani veri setimizde aykırı değer analizinde göreceğiz.

Değişim Ölçüleri

- Bir sıklık dağılımı hakkında bazı bilgiler edinmek ya da sıklık dağılımlarını karşılaştırmak için sadece konum ölçüleri yeterli değildir. Bu konum ölçüleri etrafındaki yayılma derecesini değişkenin alabileceği değerlerin birbirinden ne kadar farklı olabileceğini gösteren ölçülere de ihtiyaç vardır.
- Çeşitleri;
- Değişim Genişliği
- Çeyrek Sapma
- Varyans & Standart Sapma
- Standart Hata
- Çarpıklık
- Basıklık

- **Değişim Genişliği**
- Bir veri grubunda en büyük değer ile en küçük değer arasındaki farka "değişim genişliği" denir. "R" ile gösterilir.
- R: en büyük değer - en küçük değer
- Aşırı uç değerlerden etkilenir. Örneklem hacmi eşit olmayan iki veri setini karşılaştırmak anlamlı değildir.
- Değişim genişliği, değişim aralığını gösteren bir dağılım ölçüsüdür
- Değişim genişliğinin hesaplanmasında sadece iki uç değer işleme alındığından, diğer değerlerin hiçbir etkisi yoktur.
- Bu nedenle değişim genişliği yaygın olarak kullanılan bir dağılım ölçüsü değildir.

- **Çeyrek Sapma**
- Ortalama yerine medyan kullanıldığında ya da aşırı uç değerler bulunduğunda değişim genişliği yerine çeyrek sapma kullanılır.

$$\theta = \frac{\theta_3 - \theta_1}{2}$$

- Eşitlikte,
- Q: Çeyrek sapma
- $Q1$: Birinci çeyreklik
- $Q3$: Üçüncü çeyrekliktir.
- Dağılımdaki bütün değerler kullanılmadığı için Q yeterli bir dağılım ölçüsü değildir.

2. Olasılığa Giriş

- Bir deney eşit olasılıklı N farklı sonuç verirse ve bu sonuçların M tanesi A olayına uygun ise A olayının P(A) ile gösterilen gerçekleşme olasılığı,

$$P(A)=\frac{\text{Uygun sonuçların sayısı}}{\text{Örnek uzaydaki tüm sonuçların sayısı}}=\frac{M}{N}$$

- Bir zar atıldığında çift sayı gelmesi olasılığı nedir?
- 52'lik bir kart destesinden,
  - Kupa kızının gelmesi,
  - Sinek Valesinin gelmesi,
  - Karo gelmesi olasılıkları nelerdir?

Olasılık yapacağız fakat belirli şartlar altında,

- D bir deney ve S bu deneyin örnek uzayı olsun.
- O halde S deki bir A olayının P(A) olasılığıyla ilgili aşağıdaki aksiyomlarımız vardır.

1. $P(A) \geq 0$
2. $P(S) = 1$
3. $A_1, A_2, ..., A_n, ...$ bir S örnek uzayında sonlu ya da sonsuz sayıda ikişerli ayrık olaylar dizisi olsun. Bu durumda
   $P(A_1 \cup A_2 \cup ..) = P(A_1) + P(A_2) + ...$

- A bir S örnek uzayında herhangi bir olay olsun.
  - P(A') = 1 – P(A) dır.

- Bir parayı iki kez atalım.
  - İki kez Tura
  - İki kez yazı
  - Bir Yazı, bir Tura gelmesi olasılığı nedir?

Sürekli Örnek Uzaylar ve Geometrik Olasılık

S sonlu sayıda elemana sahip bir örnek uzay olsun, $S = \{a_1, a_2, \ldots, a_m\}$ yazalım. Her bir $a_i \in S$ için a_i nin olasılığı denen p_i gerçel sayısı karşılık getirilerek sonlu olasılık uzayı bulunur. Öyle ki,

i) Her $p_i \geq 0, \quad i = 1, 2, \ldots, m$

ve

ii) $p_1 + p_2 + \ldots + p_m = 1$

dir.

S sayılabilir sonsuzlukta eleman içeren bir örnek uzay olsun, yani $S = \{a_1, a_2, \ldots, a_m, \ldots\}$ dir. Sonlu sayıda elemanlı örnek uzaydaki gibi, her bir $a_i \in S$ e onun olasılığı p_i karşılık getirilerek olasılık uzayı bulunur. Öyle ki,

i) $p_i \geq 0$

ii) $p_1 + p_2 + \ldots + p_m + \ldots = \sum_{i=1}^{\infty} p_i = 1$

dir. Bir A olayının P(A) olasılığı onun örnek noktalarının olasılıkları toplamıdır.

Koşullu Olasılık

- Bir deney yapıldığında bir A olayının olasılığı ile ilgilendiğimizi varsayalım öyle ki diğer bir B olayının gerçekleşmiş oluğu bu deney için ek bilgi olarak verilmiş olsun. A nın olasılığının B hakkındaki ek bilgi ile nasıl etkilendiğini bilmek istiyoruz.
- Yani, Bir olayın gerçekleşmesi için başka bir olayın gerçekleşmesi koşuluna bağlı olan olasılıktır.
- B olayı bilindiğinde A olayının gerçekleşme olasılığı;

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0$$

Aşağıdaki örnek kesin matematiksel tanıma geçiş için bize yardımcı olacak.

- **Örnek** 52 kartklık bir desteden rasgele bir kart çekilsin. “Kart bir maçadır“ A olayı olsun, B “kart siyah renklidir” olayını göstersin. B verilmişken A nın gerçekleşmesi olasılığını bulalım.
- P(A) = Kartın maça olması olasılığı = 13/52
- P(B) = Kartın siyah renkli olması olasılığı = 26/52
- P(A n B) = çekilen kartın siyah ve maça olması olasılığı = 13/52
- P(A/B) = P(AnB)/P(B) = (13/52)/(26/52) = 13/26

Örnek:

Bir fabrikada üretilen parçalardan kusursuz 40 tanesi ve kusurlu 10 tanesi bir depoya konuyor. Çekilen yine yerine koyulmaksızın sırayla rasgele iki parça seçildiğinde her iki parçanın da kusurlu olması olasılığı nedir?

Çözüm:

A ve B olayları aşağıdaki gibi tanımlansın

A: İlk seçilen parça kusurlu

B: İkinci parça kusurlu

$P(A) = 10/50 = 1/5$

$P(B/A) = 9/49$

$P(A \cap B) = P(A).P(B/A) = 1/5*9/49 = 9/245$

Örnek:

- Yapılan bir çalışmada hastaların %20 ’si hem asprin hem de noveljin, %40 ’ı sadece asprin ve %30 ’u sadece novaljin kulanmaktadır. Rastgele seçilen bir hastanın asprin kullandığı biliniyorsa novaljin de kullanma olasılığı nedir?

- $P(A) = 0.60$
- $P(A \cap B) = 0.20$
- $P(B) = 0.50$

- $P(B|A) = P(A \cap B) / P(A) = 0.20 / 0.60 = 1/3$

Olasılıkta Bağımsızlık

- B olayı bilindiğinde A olayının gerçekleşme olasılığı;

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0$$

- Eğer A olayının gerçekleşmesi B olayına bağlı değilse A ve B olayları bağımsız olaydır ve

  $P(A \cap B) = P(A)P(B)$ dir.
- Bu durumda,

$$P(A|B) = \frac{P(A).P(B)}{P(B)} = P(A)$$

Şeklinde olacaktır.

Örnek:

- Bir para iki kez atılsın. İki kez de tura gelmesi olasılığı nedir?

- A : İlk atışta tura gelmesi
- B : İkinci atışta tura gelmesi

- İki olay da birbirlerinden bağımsız olduğu için,
- $P(A \cap B) = P(A).P(B)$
- $= 1/2*1/2 = 1/4$

Rassal Değişkenler

- Önceki bölümlerde bir zarın atılışını iyice karıştırılmış bir oyun kartı destesinden bir kart çekimini bir para atılması ve bir kavanozdan top çekilmesi gibi deneylerin düşünmüştük.
- Olasılığın temel kavramları örnek uzaylar ve olayların kullanılmasıyla geliştirildi.
- Bir deney yapıldığında elde edilecek sonuçlar bir zarın atılmasındaki sonuçlarda olduğu gibi sayılsa miktarlar veya kusurlu, kusursuz siyah veya beyaz gibi belirleyici özellikler olabilir.

- Değeri bir deney sonucuyla belirtilen bir değişkene rassal değişken denir. Rassal değişken örnekleri ise,
    - Bir ailedeki çocuk sayısı, Futbol takımındaki oyundan oyuna kaydedilen gol sayısı, Bir Araba satış mağazasındaki aylık veya haftalık araba satış sayısı, bir bileşiğin içindeki alkol yüzdesi gibi örnekler verilebilir.
- Rassal değişkenler dersin ilk başlarında dediğimiz gibi özelliklerine göre iki farklı şekilde ele alınır. Bunlardan biri kesikli rassal değişkenler bir diğeri ise sürekli rassal değişkenler.

Kesikli Rassal Değişkenler

- Kesikli rassal değişkenleri şu şekilde tanımlayabiliriz.
- X bir rassal değişken olsun, X'in alabileceği değerlerin sayısı sonlu veya sayılabilir sonsuzlukta ise X'e kesikli rassal değişken denir.
- Kesikli oluşturulan olasılık fonksiyonlarına Olasılık fonksiyonudur. (İlerde değineceğiz)
- Bir paranın iki kez atıılması deneyini düşünelim bu deney için örnek uzay,
- S = {YY,TY,YT,TT} dır.
- X Rassal değişkeni "Bulunan" Tura sayısı olsun böylece X'in alabileceği değerler 0,1,2 dir. O halde X sonlu sayıda değer aldığından kesikli rassal değişkendir.

Sürekli Rassal Değişkenler

- Bir diğer rassal değişken türümüz sürekli rassal değişkenlerdir.
- X bir rassal değişken olsun.
- X bir aralıkta ya da birden çok aralıkta her değeri alabiliyorsa X'e sürekli rassal değişken denir.
- Sürekli değişkenlerden oluşan olasılık fonksiyonuna olasılık yoğunluk fonksiyonu diyoruz.

Beklenen an Beklenen Değer

- İsminden de anlaşılacağı üzere bir rassal değişkenin ortalaması ya da kitle dediğimiz popülasyonun ortalamasıdır. Beklenen değeri hesaplamanın rassal değişkenleri ikiye ayırdığımız için iki farklı hesaplanma şekli vardır fakat hesapladığımız formül aynıdır.

- X bir kesikli rassal değişken olsun ve olasılık fonksiyonu da bizler için f(x_i) olsun,

- E(X) yani beklenen değerimizin hesaplanması, formülü şu şekildedir.

$$E(X) = x_1.f(x_1) + \ldots + x_N.f(x_N) + \ldots = \sum_{i=1} x_i f(x_i)$$

- Burada E(X) bizim için beklenen değer, x ise rassal değişkenin alabileceği değerler ve f(x) ise X'in alacağı değerlerin her birinin olasılığıdır.

Örnek

- Düzgün hilesiz bir zarı atıyoruz. Üste gelen yüzdeki noktaların beklenen değerini nasıl hesaplayacağız?

- Zarın üst yüzünde bulunan noktaların sayısını X ile gösterelim. X'in olanaklı değerleri 1,2,3,4,5,6 dır. VE her birinin gelme olasılığı da 1/6 dır.

- O halde beklenen değeri hesaplayalım, tüm bilgiye sahibiz.

- $E(X) = 1.1/6+2.1/6+3.1/6+4.1/6+5.1/6+6.1/6 = 21/6 = 3.5$ olarak bulunur.

Örnek 5.5.2 Düzgün bir para üç kez atılsın bulunan turaların sayısı için beklenen değer nedir?

Çözüm: Bu deney için örnek uzay ve karşılık gelen olasılıklar aşağıdaki tabloda gösterilmiştir.

Tablo 5.5.1

Örnek nokta	Turaların Sayısı	Olasılık
TTT	3	1/8
TTY	2	1/8
TYT	2	1/8
YTT	2	1/8
TYY	1	1/8
YTY	1	1/8
YYT	1	1/8
YYY	0	1/8

Bir paranın üç kez atılmasında bulunan turaların sayısını X ile gösterelim. X in olasılık fonksiyonu Tablo 5.5.2 de verilmiştir.

Tablo 5.5.2

X = x	0	1	2	3
f(x) = P(X = x)	1/8	3/8	3/8	1/8

O halde (5.5.1) denkleminden

$$E(X) = \sum_{i=1}^{3} x_i f(x_i) = 0.(1/8) + 1.(3/8) + 2.(3/8) + 3.(1/8)$$

$$= \frac{12}{8} = \frac{3}{2} \text{ dir.}$$

Örnek 5.5.3

İçinde üç beyaz ve iki siyah top bulunan bir kavanozdan yine yerine koyarak rasgele iki top çekilmiştir. Çekilen her beyaz top için 100 lira kazanılacak ve çekilen her siyah top için 50 lira kaybedilecektir. Bu oyunda beklenen kâr nedir?

Çözüm:

		İKİNCİ TOP	
İLK TOP		B	S
	B	$\frac{9}{25}$:200 Lira	$\frac{6}{25}$:50 Lira
	S	$\frac{6}{25}$:50 Lira	$\frac{4}{25}$:-100Lira

Tablo 5.5.4

X rasgele değişkeni kazanılan liraların sayısı olsun. Tablo 5.5.3 de olanaklı sonuçların örnek uzayı, karşılık gelen olasılıklar ve kâr gösterilmiştir.

X rasgele değişkeninin beklenen değeri, bu oyundaki kâr

$$E(X) = 200.\left(\frac{9}{25}\right) + 50.\left(\frac{6}{25}\right) + 50.\left(\frac{6}{25}\right) - 100.\left(\frac{4}{25}\right)$$
$$= 80 \text{ liradır.}$$

Sürekli Değişkenlerde Beklenen Değer

- X bir boyutlu sürekli rassal bir değişken olsun, f(x) X in olasılık yoğunluk fonksiyonu olmak üzere, X ‘in beklenen değeri,

$$E(x) = \int_{-\infty}^{+\infty} x.f(x).dx \qquad -\infty < x < +\infty$$

Örnek:

- f(x) = x/2, 0<x<2 olsun,

- E(X) nedir?

Beklenen Değerin Özellikleri

- a ve b sabit sayılar ve X rassal değişken ise

$$E(aX+b) = aE(X)+b$$

dir.

- Eğer X ve Y rassal değişkenleri birbirinden bağımsız ise,

$$E(X,Y) = E(X)\,E(Y)$$ dir.

- Yani iki bağımsız rassal değişkenin çarpımının ortalaması, ortalamalarının çarpımına eşittir.

Varyans

- Varyans, Bir rassal değişkenin beklenen değeri ya da ortalaması olasılık fonksiyonunun merkezi hakkında biz bilgi verir. Fakat ortalama değer bir deneyden bir diğerine rassal değişkenin değerlerinin dağılımı, değişimi ya da yayılması ile ilgili bilgi vermez. İşte bu varyans dağılımını değişimi ya da yayılma ölçüsü olarak kullanılır.

(Varyans) X, olasılık fonksiyonu Tablo 5.6.1 deki gibi verilmiş olan kesikli rasgele değişken olsun. X in ortalaması $E(X) = \mu$ ise X in varyansı, $Var(X)$ veya σ^2_x aşağıdaki gibi tanımlanır.

$$\sigma^2_x = Var(X) = E[(X-\mu)^2] \qquad (5.7.1)$$

ya da eşdeğer olarak, Tanım 5.6.1 in kullanılmasıyla

a) X kesikli rasgele değişken ise

$$\sigma^2_x = Var(X) = \sum_{i=1}^{N} (x_i - \mu)^2 . f(x_i) \qquad (5.7.2)$$

dir.

b) X sürekli rasgele değişken ise,

$$\sigma^2_x = Var(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 . f(x).dx$$

dir. Yani, X in varyansı, X in kendi ortalamasından sapmasının karesinin ortalamasıdır.

X, $E(X) = \mu$ ortalamalı ve $Var(X) = \sigma^2$ varyanslı bir rasgele değişken ise

$$\sigma^2 = E(X^2) - [E(X)]^2 = E(X^2) - \mu^2 \qquad (5.7.4)$$

dir.

Standart Sapma

X, μ ortalamalı kesikli ya da sürekli bir rasgele değişken olsun. X in standart sapması, σ_x, varyansın karekköküdür, ve aşağıdaki eşitlikle verilmiştir.

$$\sigma_x = \sqrt{\text{Var}(X)} = \sqrt{E(X-\mu)^2} \qquad (5.7.3)$$

Varyansın Özelliği

- a ve b bir sabit ve X bir rassal değişken olsun,

$$Var(aX+b) = a^2*Var(X) \text{ dir.}$$

Acil Örnek,

- X rassal değişkeninin varyansı 0.5 dir.
- A)2X
- B)X/2+2579
- C)10X/5
- Rassal değişkenlerinin varyansı nedir?

Kovaryans

X ve Y rasgele değişkenlerinin ortalamaları μ_x ve μ_y olmak üzere

$$E[(X - \mu_x).(Y - \mu_y)]$$

beklenen değerine X ve Y arasındaki **kovaryans** denir ve σ_{xy} ya da Cov(X,Y) ile gösterilir.

X ve Y arasındaki kovaryansı hesaplamak için

$$Cov(X, Y) = E(XY) - E(X).E(Y) = E(XY) - \mu_x.\mu_y)$$

formülü tercih edilir. Kovaryans X ve Y nin birbiriyle nasıl bir ilişkiye sahip olduğunu gösterir. X ve Y istatistiksel olarak birbirinden bağımsız iseler $E(XY) = E(X).E(Y)$ olacağından, bu durumda $Cov(X, Y) = 0$ bulunur. Bunun tersi doğru değildir. Yani, $Cov(X, Y) = 0$ olması X ve Y nin bağımsızlığını gerektirmez. Bu

Eğer değişkenler bağımsızsa kovaryans 0'dır, Eğer kovaryans 0 ise değişkenler bağımsızdır diyemeyiz.

Olasılık Dağılımları fakat Rassal Değişken Kesikli

- 1. Bernouilli Dağılımı (**X ~ Bernouilli(p)**)
- Bir rastgele deney yapıldığında bu deneyin sadece iki mümkün sonucu elde ediliyorsa böyle bir deneye Bernoulli deneyi denir. Bernoulli deneyinde elde edilecek sonuçlardan biri 'başarı' olarak nitelendiriliyorsa diğeri ise 'başarısızlık' olarak nitelendirilir. Başarı sonucu elde edildiğinde x =1, başarısızlık sonucu elde edildiğinde x = 0 değerini alan X rastgele değişkenine 'Bernoulli değişkeni' denir. Bu değişkenin olasılık dağılımına Bernoulli dağılımı denir.
- Deney sonucunda başarı elde etme olasılığı 'p' ise X rassal değişkeninin olasılık fonksiyonu,

$$f_X(x) = P(X = x) = \begin{cases} p^x (1-p)^{1-x}, & x = 0,1 \\ 0, & d.y.(diyer\ yerlerde) \end{cases}$$

şeklinde gösterilebilir. $0 < p < 1$, $q = 1 - p$.

- E(X) = p
- Var(X) = pq

Örnek:

- Bir kişinin Pazar günü sinemaya gitme olasılığı 0.3'tür. Buna ilişkin olarak fonksiyonu tanımlayınız.

p = 0.3, q = 0.7

$$f(x)=\begin{cases} p^x q^{1-x}, & x=0,1 \\ 0, & d.y. \end{cases}$$

$$P(X=x)=(0.3)^x.(0.7)^{1-x},\ x=0,1$$

$$P(X=0)=(0.3)^0.(0.7)^{1-0}=0.7$$

$$P(X=1)=(0.3)^1.(0.7)^{1-1}=0.3$$

$$E(X)=0.3$$

$$Var(X)=(0.3)(0.7)=0.21$$

2. Binom Dağılımı (X ~ Binom(n,p))

- Kesikli rastgele değişkenin oluşturduğu bir dağılımdır. İki olası sonuç ile ilgilenilir. Erkek/kadın, sağlam/bozuk, olumlu/olumsuz, yazı/tura vs. Bu tür iki sonuçlu deneyler n kez tekrar edildiğinde, X rastgele değişkeni gelen başarıların sayısı olsun. X rastgele değişkeninin olasılık fonksiyonu;

$$f_X(x) = P(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0,1...,n \\ 0, & dy \end{cases}$$

$$0 < p < 1,\ q = 1 - p$$

$$E(X) = np$$

$$Var(X) = npq$$

Örnek:

- 5 seçenekli 20 soruluk bir test sınavında sorular rastgele işaretlendiğinde;
- a) En az 10 doğru cevap tutma olasılığı nedir?
- b) Tutturulan doğru cevap sayısının beklenen değeri nedir?

X rasgele değişkeni 20 soruda doğruların sayısını göstersin.

$$X \sim Binom(n = 20,\ p = \frac{1}{5})$$

$$\text{a) } P(X \geq 10) = \sum_{x=10}^{20} P(X = x)$$

$$= \sum_{x=10}^{20} \binom{20}{x} \left(\frac{1}{5}\right)^{x} \cdot \left(1 - \frac{1}{5}\right)^{20-x} = 0.0026$$

$$\text{b) } E(X) = np = 20\frac{1}{5} = 4$$

3. Geometrik Dağılım (Geometrik (p))

- Arka arkaya n kez tekrarlanan bir Bernoulli deneyinde, ilk istenen sonucun (başarı ya da başarısızlık) elde edilmesi için yapılan deney sayısı X rastgele değişkeni olmak üzere X 'e 'geometrik rastgele değişkeni' denir. Bu değişkenin dağılımına 'Geometrik Dağılım' denir.

$$P(X=x)=\begin{cases} pq^{x-1}, & x=1,2,3... \\ 0, & dy \end{cases}$$

$$E(X)=\frac{1}{p}$$

$$Var(X)=\frac{q}{p^2}$$

Örnek:

- Bir atıcının her atışta hedefi vurma olasılığı aynı ve 2/3 olasılıkla gerçekleşiyor. Arka arkaya yapılan atışlar sonucunda hedefi ilk kez vurmak için gereken atış sayısı X rastgele değişkeni olduğuna göre;
- X ~ Geometrik(p=2/3)

a) X rastgele değişkeninin olasılık fonksiyonunu bulunuz.

$$P(X=x)=\begin{cases}\frac{2}{3}\left(\frac{1}{3}\right)^{x-1}, & x=1,2,3...\\ 0, & dy\end{cases}$$

b) Hedefi ilk kez 2. atışta vurma olasılığını hesaplayınız.

$$P(X=2)=\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)=\frac{2}{9}$$

c) Hedefi ilk kez en çok 3. atışta vurma olasılığını hesaplayınız.

$$P(X\le 3)=\sum_{x=1}^{3}P(X=x)=P(X=1)+P(X=2)+P(X=3)$$

$$=\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^{1-1}+\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^{2-1}+\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^{3-1}=\frac{2}{3}+\frac{2}{9}+\frac{2}{27}=0.96$$

d) Hedefi ilk kez en az 5. atışta vurma olasılığını hesaplayınız.

$$P(X \geq 5) = \sum_{x=5}^{\infty} P(X = x)$$

$$= 1 - P(X < 5)$$

$$= 1 - \sum_{x=1}^{4} P(X = x)$$

$$= 1 - \left[0.96 + \frac{2}{3}\left(\frac{1}{3}\right)^{4-1} \right] = 0.016$$

e) Hedefi ilk kez vuruncaya kadar ortalama kaç atış gerekir?

$$E(X) = \frac{1}{p} = \frac{3}{2} = 1.5$$

f) $Var(X)$ değerini hesaplayınız.

$$Var(X) = \frac{q}{p^2} = \frac{3}{4}$$

4. Poisson Dağılımı (Poisson~(λ))

- Sürekli ortamlarda (zaman, alan, hacim vs.) kesikli sonuçlar veren ve bazı özelliklere sahip deneyleri modellemede kullanılır. Örneğin; belli bir zaman aralığında bir yoldan geçen arabaların gözlenmesi, belli bir zaman aralığında mağazaya gelen müşterilerin gözlenmesi vb.

Olasılık fonksiyonu

$$P(X=x)=\begin{cases}\dfrac{e^{-\lambda}\lambda^{x}}{x!}, & x=0,1,2...\\ 0, & dy\end{cases}$$

- Burada λ belli bir zaman aralığında ya da belli bir yerde istenen olayın ortalama ortaya çıkma sayısıdır. Poisson dağılımının beklenen değeri ve varyansı,

$$E(X)=\lambda$$
$$Var(X)=\lambda$$

Örnek:

- 5 dakikalık bir zaman aralığında belli bir telefon santraline ortalama 3 müracaat yapılıyor. ($X \sim Poisson(\lambda = 3)$)

a) Bu zaman aralığında hiç müracaat olmama olasılığı nedir?

$$P(X=0)=\frac{e^{-3}.3^{0}}{0!}=0.05$$

b) En çok 3 müracaat olma olasılığı nedir?

$$P(X\le 3)=\sum_{x=0}^{3}P(X=x)$$

$$=\frac{e^{-3}.3^{0}}{0!}+\frac{e^{-3}.3^{1}}{1!}+\frac{e^{-3}.3^{2}}{2!}+\frac{e^{-3}.3^{3}}{3!}=0.65$$

c) En az 5 müracaat olma olasılığı nedir?

$$P(X\ge 5)=1-P(X<5)$$

$$=1-\left[0.65+\frac{e^{-3}.3^{4}}{4!}\right]$$

$$=1-\left[0.65+0.168\right]=0.182$$

d) Beklenen değer ve varyans nedir?

$$E(X)=3$$

$$Var(X)=3$$

Sürekli Dağılımlar

- 1. Uniform (Düzgün) Dağılım (X~Uniform(a,b))

X rasgele değişkeninin olasılık yoğunluk fonksiyonu

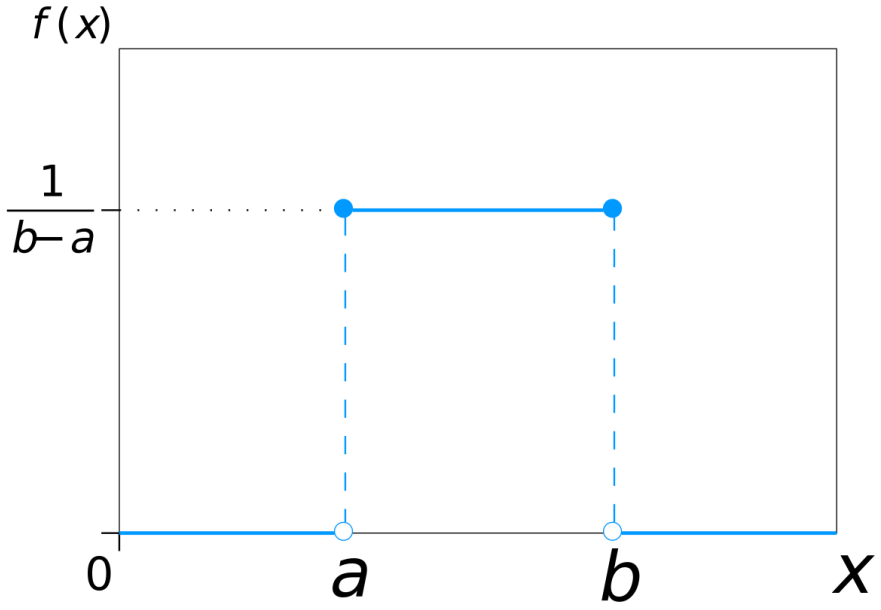
$$f(x) = f(x; a, b) = \frac{1}{b-a} \quad a \le x \le b$$

ise X rasgele değişkeni [a, b] kapalı aralığında düzgün dağılıma sahiptir denir. Yoğunluk fonksiyonu yandaki Şekil 7.5.1 de

X rasgele değişkeni [a, b] aralığında düzgün dağılıma sahipse

$$E(X) = \frac{a+b}{2}, \quad Var(X) = \frac{(b-a)^2}{12}$$

dir.



- 2. Üstel Dağılım (X~Üstel(α))

(Üstel dağılım) Negatif olmayan değerler alan sürekli X rasgele değişkeninin olasılık yoğunluk fonksiyonu

$$f(x)=\begin{cases}\alpha.e^{-\alpha x} & x>0\\ 0 & \text{başka yerlerde}\end{cases}$$

olsun. Bu takdirde X, $\alpha > 0$ parametresi ile üstel dağılıma sahiptir denir. Görüldüğü gibi

$$\int_0^\infty f(x).dx=\int_0^\infty \alpha.e^{-\alpha x}.dx=-e^{-\alpha x}\Big|_0^\infty=1$$

dir.

b) X in beklenen değeri:

$$E(X)=\int_0^\infty x.\alpha.\alpha^{-\alpha x}.dx$$

dir. Kısmi integrasyon yöntemi uygulanırsa: $\alpha.e^{-\alpha x}.dx = dv$, $x = u$ alarak

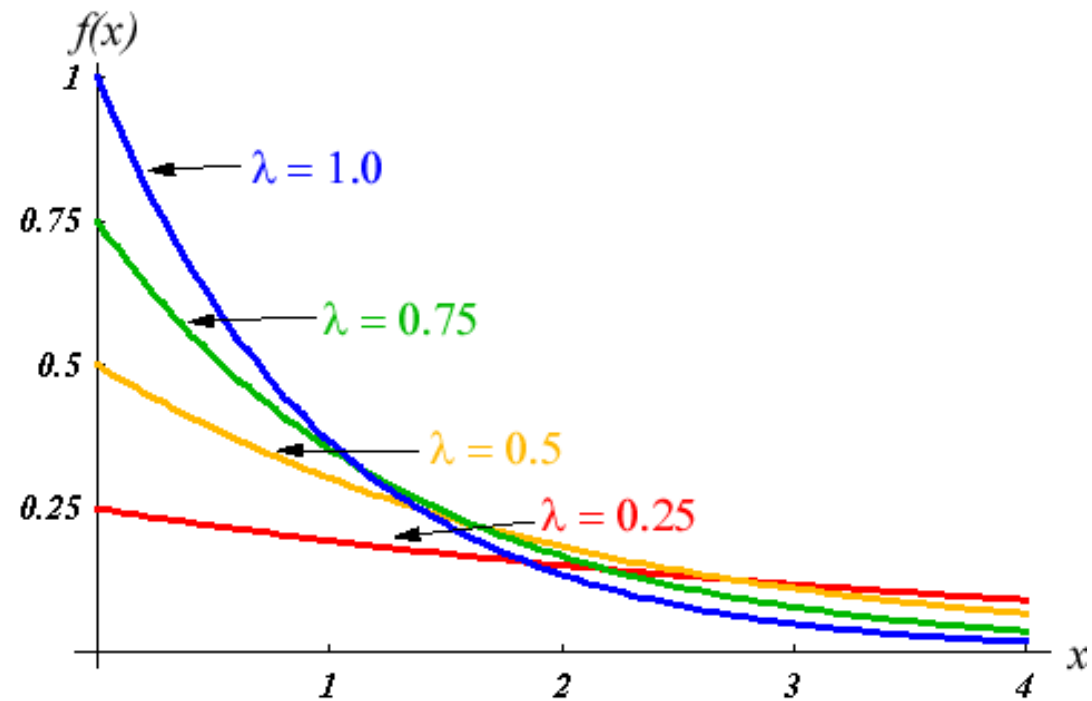
$$E(X)=[-xe^{-\alpha x}]\Big|_0^\infty+\int_0^\infty e^{-\alpha x}.dx=\frac{1}{\alpha}$$

bulunur.

c) X in varyansı benzer integrasyon yöntemiyle bulunur.

$$\sigma^2=E(X^2)-[E(X)]^2=\int_0^\infty x^2.\alpha.^{-\alpha x}.dx-\left(\frac{1}{\alpha}\right)^2$$

$$=\frac{2}{\alpha^2}-\frac{1}{\alpha^2}=\frac{1}{\alpha^2}.$$



- 3. Gamma Dağılımı (X ~ Gamma(r, α))

(Gamma fonksiyonu) Tüm $p > 0$ değerleri için tanımlanan

$$\Gamma(p) = \int_0^{\infty} x^{p-1}.e^{-x}.dx$$

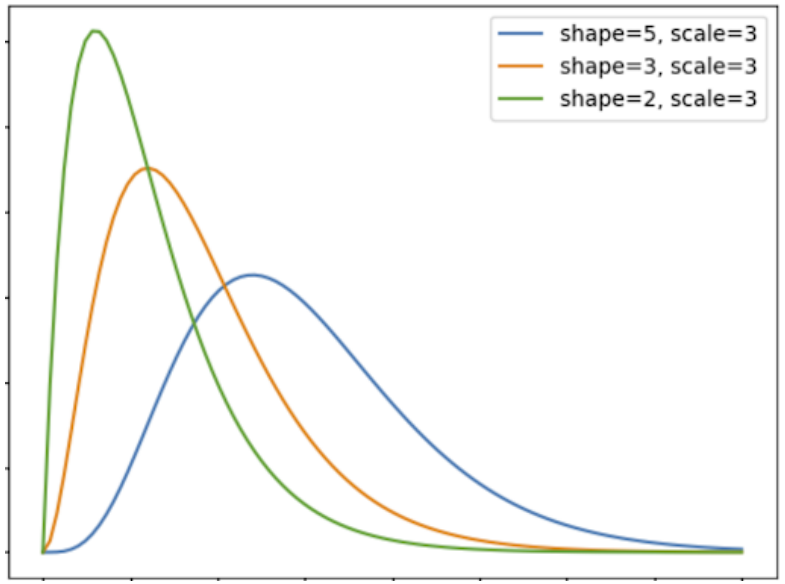
fonksiyonuna Gamma fonksiyonu denir. Parçalara ayırarak

(Gamma olasılık yoğunluk fonksiyonu) X, pozitif değerler alan sürekli rasgele değişken olsun. X in olasılık yoğunluk fonksiyonu

$$f(x;\alpha,r) = \begin{cases} \dfrac{\alpha}{\Gamma(r)}.(\alpha x)^{r-1}.e^{-\alpha x} & x > 0 \\ 0 & x \le 0 \end{cases}$$

ise X, Gamma olasılık dağılımına sahiptir. Bu dağılımda $\alpha > 0, r > 0$ parametreleri vardır. Kolayca gösterilebilir ki

$$E(X) = \frac{r}{\alpha} \qquad Var(X) = \frac{r.(r+1)}{\alpha^2} - \left(\frac{r}{\alpha}\right)^2 = \frac{r}{\alpha^2}$$

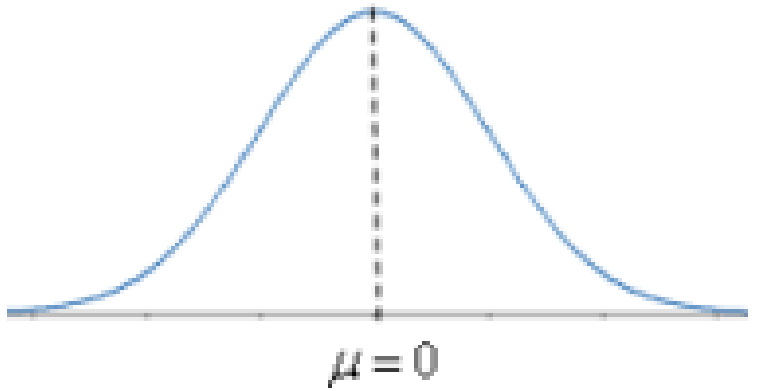


4. Normal Dağılım (X ~ Normal(μ, σ^2))

- Önemli bir dağılım olmasının nedenlerinden biri yapılan bir çok gözlem sonucunun dağılımının bu yapıya benzemesi ve çoğu dağılımın da gözlem sayısı arttıkça normal dağılıma yaklaşmasıdır. X sürekli rastgele değişkeni normal dağılıma sahip ise X rastgele değişkeninin olasılık yoğunluk fonksiyonu;

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma^2 > 0$$

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

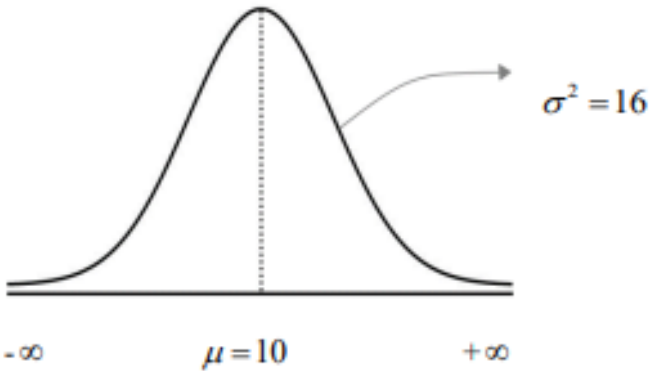


Örnek;

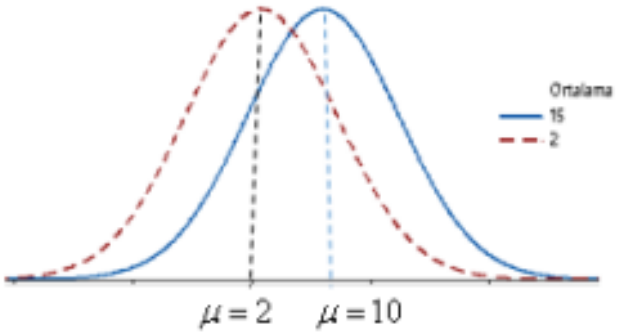
$\mu = 10,\ \sigma^2 = 16$ $\quad -\infty < x < \infty$

$f(x) = \frac{1}{\sqrt{2\pi}4} e^{-\frac{1}{2}\frac{(x-10)^2}{16}}$, $\quad -\infty < \mu < \infty$

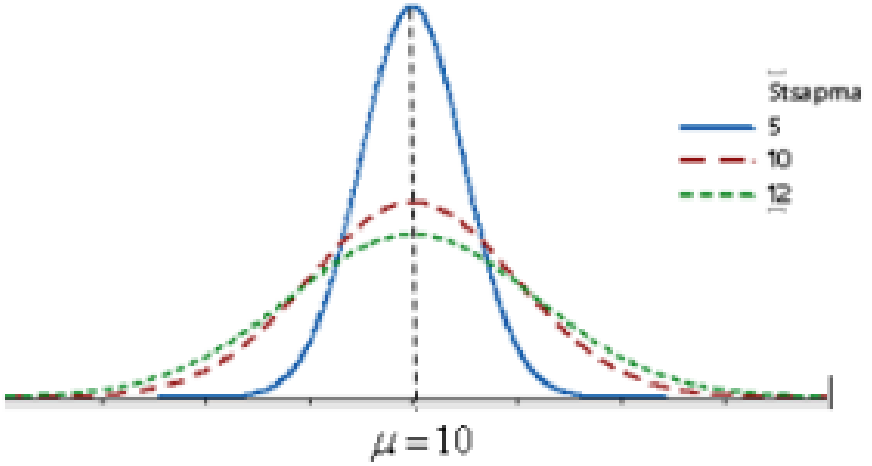
$\sigma^2 > 0$



=>Ortalama değiştiği sürece; $\sigma^2 = 16$



=>Varyans değiştiği sürece;



Standart Normal Dağılım

- X ~ Normal(μ, σ) iken eğer $\mu = 0,\ \sigma^2 = 1$, olan normal dağılıma standart normal dağılım denir. Standart normal dağılıma sahip rasgele değişken genellikle Z harfi ile gösterilir
- Z~N($\mu = 0,\ \sigma^2 = 1$) standart normal dağılıma sahiptir denir.

$$f_z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < \infty$$

- Aklınıza gelecek herhangi bir rassal değişkenin dağılımı her zaman Standart Normal Dağılıma dönebilir nasıl mı?

$$Z = \frac{X - \mu}{\sigma}$$

Z testi

Örnek:

- Belli bir tür bitkinin yaşam süresi N(35,16) olan dağılıma sahip olduğu bilinmektedir.
- X -> Bitkinin yaşam süresi ve X ~Normal(35,16)

a) Rastgele seçilen bir bitkinin yaşam süresinin 45 günden çok olma olasılığı nedir?

$$P(X>45)=P(\frac{X-\mu}{\sigma}>\frac{45-35}{4})=P(Z>2.5)=1-P(Z\le 2.5)=1-0.9938=0.0062$$

b) Aynı tür bitki için alınan 10.000 örnekten kaç tanesinin yaşam süresi 45 günden fazladır?

$$10000\times 0.0062\Rightarrow 62 \text{ tanesi}$$

- https://www.z-table.com

c)

$$P(25 < X < 30) = P\left(\frac{25-35}{4} < Z < \frac{30-35}{4}\right)$$

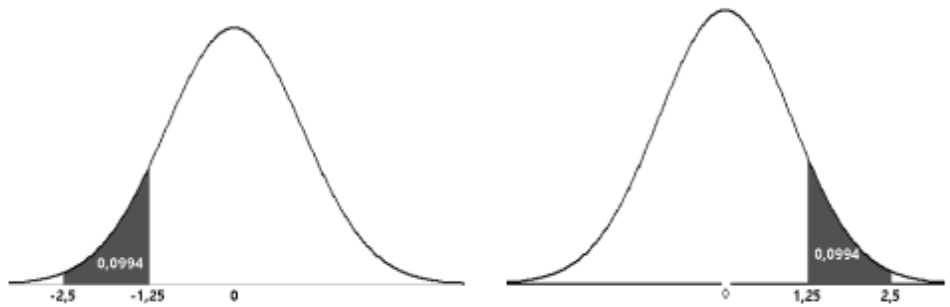
$$= P\left(\frac{-10}{4} < Z < \frac{-5}{4}\right)$$

$$= P(-2.5 < Z < -1.25)$$

$$= P(1.25 < Z < 2.5)$$

$$= P(Z < 2.5) - P(Z < 1.25)$$

$$= 0.9938 - 0.8944 = 0.0994$$



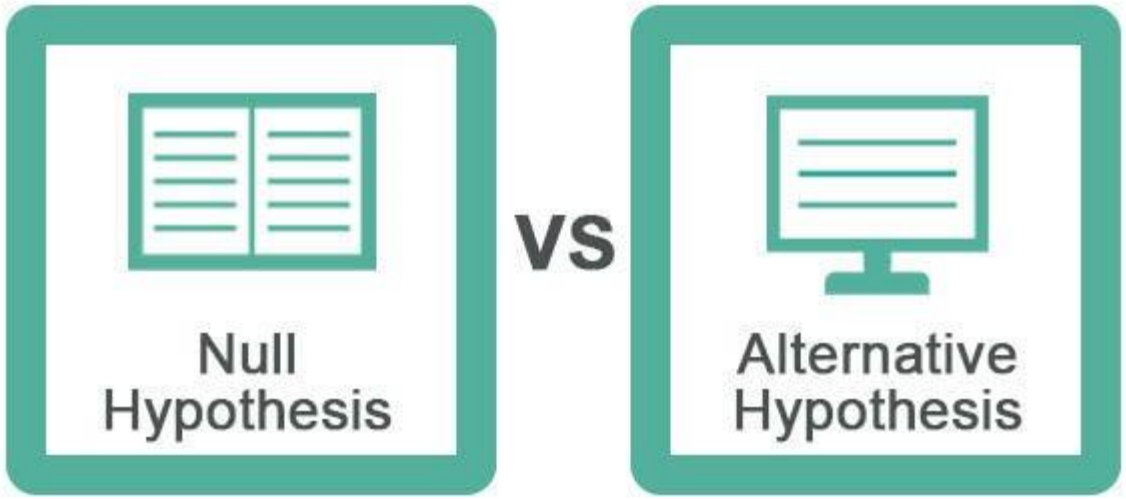
Hipotez Testleri

- Hipotez Nedir?
- Örnekleme ile test edilmeye çalışılan bir popülasyonun ilgili parametresi hakkında ortaya sunulan iddiadır.

- Örneğin;
- – A dersi için vize ortalaması 50'nin altındadır
- – A ve B lastik firmalarının ürettikleri lastiklerin kaliteleri aynıdır.
- Görüldüğü gibi bir konu hakkında öne sürülen ve doğruluğu henüz ispatlanmamış görüşler hipotezlerdir. Hipotezler üzerinde çeşitli işlemler yapılarak ifadenin "doğruluğu/yanlışlığı" araştırılır.

Hipotez Testi ve Aşamaları

- Popülasyonu incelemeye yönelik yapılan çalışmalar ve bunların raporlanması ile hipotezin kabul edilip edilmeyeceğinin belirlenmesi işlemine hipotez testi denir.
- Hipotez testi aslında bir nevi karşılaştırma ve seçim işlemi olduğu için birden fazla hipoteze ihtiyaç duyulur. Bu hipotezlere ise alternatif hipotez denir



Hipotez Testlerinin Aşamaları

- Bir hipotezi test ederken 5 aşamadan geçeriz bunlar,

1) Hipotezin Belirlenmesi

- Popülasyon parametresine genellikle belli bir değer atanır ve bu öne sürülen temel iddia sıfır veya farksızlık (null) hipotezidir. Bu hipotez sıfır/başlangıç hipotezi olarak da bilinir. H0 ile gösterilir.
- Mevcut veriler sıfır hipotezinin doğruluğu hakkında şüphe uyandırdığında kıyas yapmak için ortaya sunulan ikinci hipotez alternatif hipotezdir. Yapılan işlemler eğer H0'ı yanlış çıkarırsa bu Ha'nın kabulü anlamına gelir.

2) Önem veya Risk Derecesinin Belirlenmesi

- Genellikle risk derecesi olarak %5=0,05 ve %1=0,01kullanılmakla birlikte bu tercihi bir durumdur. Risk derecesi temelde doğru olan null hipotezinin reddedilme olasılığını gösterir.
- Risk derecesini belirleyerek hipotez testi sırasında yapılabilecek hataları minimuma indirmek isteriz. Bir hipotez testi sırasında null hipotezinin doğruluk/yanlışlık ve kabul/reddedilme durumlarına göre 2 tip hata yapılabilir (1.tip ve 2.tip hata).
- Null hipotezi doğru iken reddedilirse 1. tip hata,
- Yanlış iken kabul edilirse 2.tip hata yapılmış olur.

Alınan karar	**Null hipotezi doğru**	**Null hipotezi yanlış**
Null hipotezi kabul etme	*Doğru karar*	*2. Tip hata*
Null hipotezi reddetme	*1. Tip hata*	*yorumsuz*

3) İstatistiksel Test Metodunun Belirlenmesi

- Örneğin F, t, ki kare istatistiksel testleri kullanılaraknull hipotezi ile ilgili değerin bulunması işlemidir

4) Null Hipotezinin Kabul/Red Durumunun Belirlenmesi

- Yukarıdaki maddede (3) bulunacak değerin durumuna göre null hipotezinin kabul/red koşullarının belirlenmesidir.

5)Null Hipotezi İçin Karar Verme

- Yapılan işlem sonuçlarına göre null hipotezinin kabul edilip edilmeyeceği belirlenir.

1) Hipotezleri Belirlemek

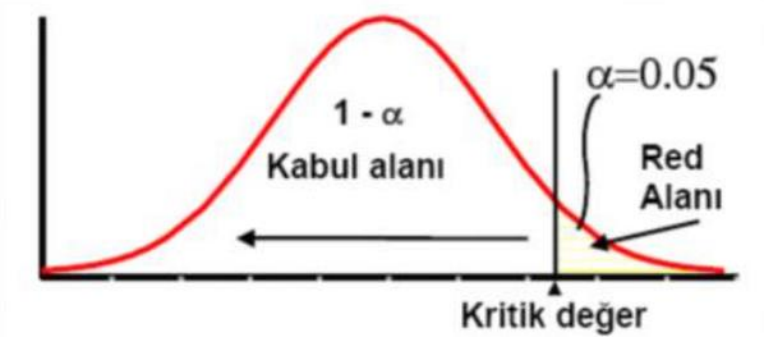
- Ders geçmek için gerekli minimum notun ortalama 60 olduğu bir sınıftan seçilen 40 öğrencinin aldığı notların ortalaması 64 olsun. Bu durumda popülasyonun(sınıfın) gerçek ortalaması 60'ın üzerinde midir?

$H_0: \mu = 60$

$H_A: \mu > 60$

2) Önem Derecesini Belirlemek

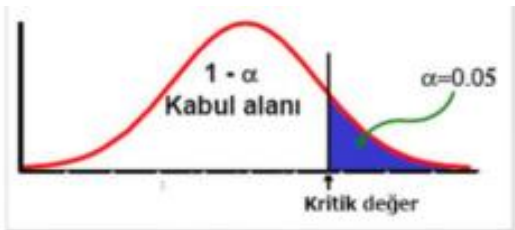
- Sıfır hipotezini gerçekten doğru iken reddetme olasılığının yani önem derecesinin α=0,05olduğunu kabul edelim. Bu durumda grafiksel bir açıklama yapacak olursak;



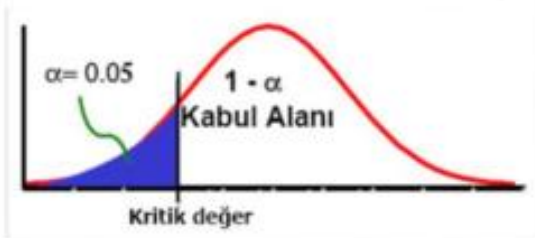
3) Hipotez Testinin Yönünü Belirlemek

- Alternatif hipotez için yazılan duruma göre hipotez testi tek yönlü ya da iki yönlü olabilir.
- Tek yönlü hipotez testi için α direk alınır iken iki yönlü hipotez testinde alan belirlenirken α yerine α/2 değeri ile işlem yapılır.

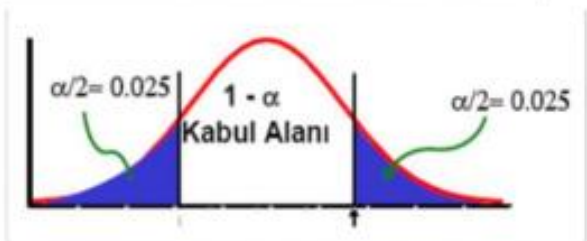
$H_A: \mu \geq 60$ (tek yönlü)



$H_A: \mu \leq 60$ (tek yönlü)



$H_A: \mu \neq 60$ (çift yönlü)



4) Kritik Değerleri Belirlemek

- Null hipotezinin doğru olduğu varsayımı ile olasılığı 1-α olan değer aranan kritik değerdir. İlgili istatistik testi için değişmekle birlikte kritik değer standart normal dağılımlar için z* ile gösterilir.
- Eğer popülasyon için standart sapma değeri biliniyor ise ya da gözlem sayısı n≥30 ise 0,5-α değerine karşılık gelen z değeri tablodan bulunur ve aranan z*değeri odur.

5) Test İstatistiğini Belirlemek ve Kritik Değer ile Karşılaştırmak

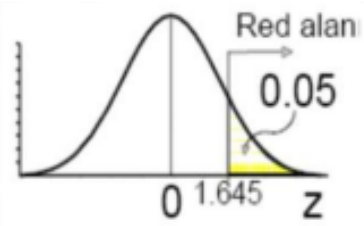
- μ = popülasyonun ortalaması
- σ = popülasyonun standart sapması
- s = örneklemin standart sapması
- $\bar{x}$ = örneklemin ortalaması
- Z = kritik değer olmak üzere;
- Popülasyona ait standart sapma biliniyor ise; $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$
- Popülasyonun standart sapması bilinmiyor ve n≥30 ise σ yerine s alınarak z değeri bulunur. Daha sonra z ile z* değeri karşılaştırılarak karara varılır.

Özet olarak Hipotez Testi Adımları

1. H0 ve Ha hipotezleri belirlenir
2. α tespit edilir
3. Hipotez testinin yönü belirlenir
4. Kritik değer z* bulunur
5. Test istatistiği yapılarak z değeri bulunur ve karşılaştırma ile karar verilir

Z testi ile bir örnek yapalım,

- Rasgele seçilen 25 kutu mısır gevreğinin ortalaması 372.5gr. ve üretici firmanın belirlemelerine göre standart sapma 15gr'dır. Bu durumda 0.05 önem derecesi ile bir kutu mısır gevreğinin 368gr üzerinde olmasını test ediniz.
- Verilenlere göre, H0 : $\mu \leq 368$ ve Ha : $\mu > 368$ ve $\alpha = 0.05$ $n = 25$ $\sigma = 15$ $X = 372.5$
- Veriler değerler formülde yerine yazarsak,
- $z = (372.5-368)/(15/5) = 1.5$ olur.
- Diğer taraftan
- $\alpha = 0.05$ için standart normal dağılım tablosundan 0.5-0.05=0.45 değerine karşılık gelen $z^* = 1.645$ dir.



- z=1.5 değeri taralı alanın dışında olduğundan null hipotezi kabul edilir.

- Öyleyse yorumumuz, Mısır gevreklerinin kutularının ortalama 368gr'ın üzerinde olduğuna dair yeterli bir bilgi yoktur

T testi

- T testine geçmeden önce t testini yapabilmek için t dağılımını bilmemiz gerekir. T dağılımı normal dağılımın özel bir halidir. Özellik olarak t dağılımı,
- t ortalaması 0 olan bir dağılımdır.
- t ortalamaya göre simetrik dağılır
- t varyansı 1'den büyük olan bir dağılımdır; ancak örneklem büyüklüğü arttıkça, varyans 1'e yaklaşır.

Ve 3 çeşittir.

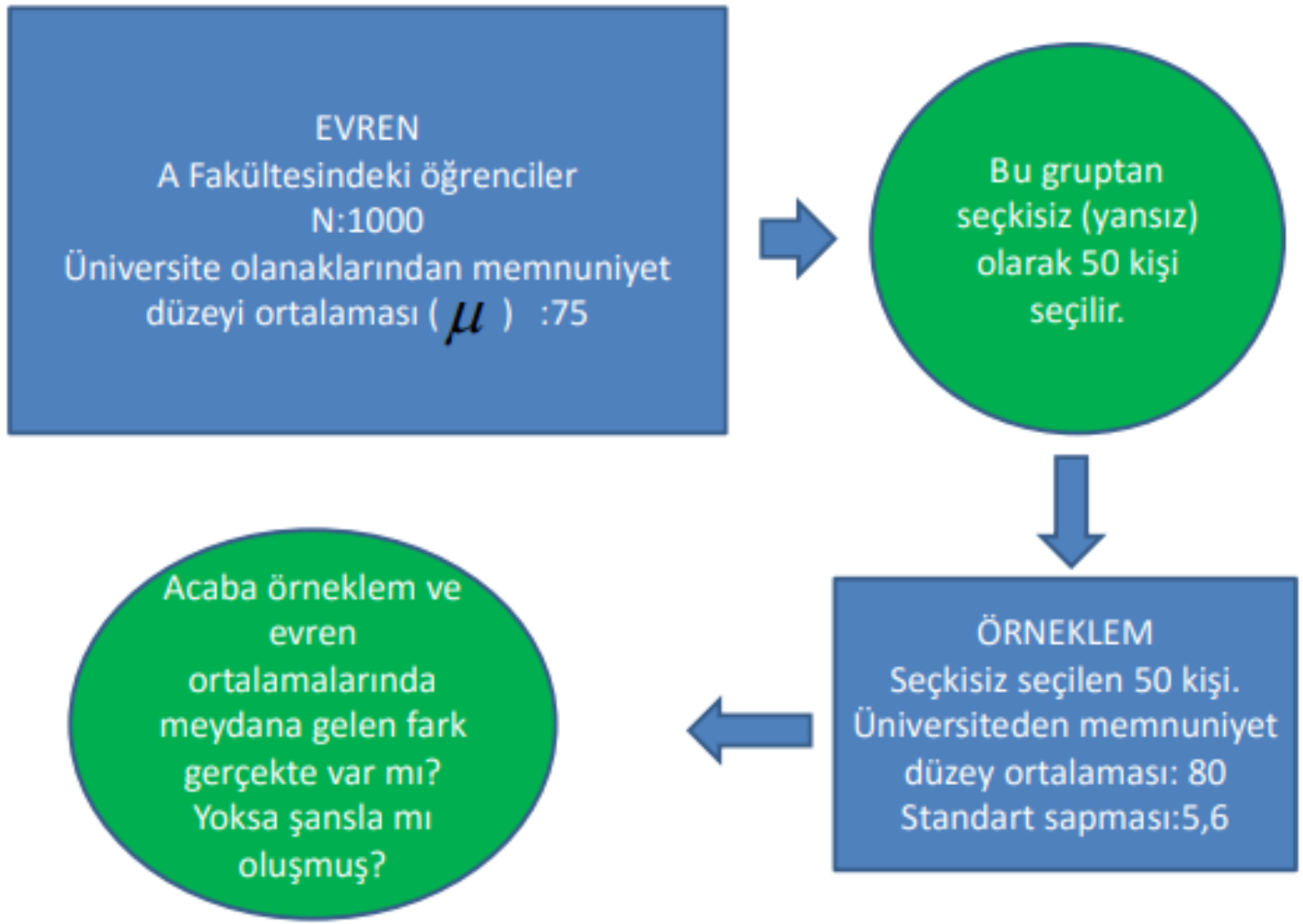
- 1. Tek örneklem için t testi
- 2. Bağımsız örneklem için t testi
- 3. Bağımlı örneklem için t testi

Tek örneklem için t testi

- Tek örneklem t testinde hipotezler, örneklemden elde edilen ortalama ile evren ortalaması arasında fark olup olmamasına göre oluşturulur.
- H0 : $\mu - \bar{x} = 0$ vs Ha : $\mu - \bar{x} = 0$

- Örneklem ortalamasının anlamlılığını test etmek üzere kullanılan parametrik bir tekniktir. 2 varsayımı mevcuttur;
- Bağımlı değişkene ait puanlar eşit aralıklı ya da eşit oranlı ölçek düzeyindedir.
- Bağımlı değişkene ait puanlar evrende normal dağılım gösterir

- $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$


EVREN
A Fakültesindeki öğrenciler
N:1000
Üniversite olanaklarından memnuniyet düzeyi ortalaması (μ) :75
Bu gruptan seçkisiz (yansız) olarak 50 kişi seçilir.
ÖRNEKLEM
Seçkisiz seçilen 50 kişi.
Üniversiteden memnuniyet düzey ortalaması: 80
Standart sapması:5,6
Acaba örneklem ve evren ortalamalarında meydana gelen fark gerçekte var mı? Yoksa şansla mı oluşmuş?

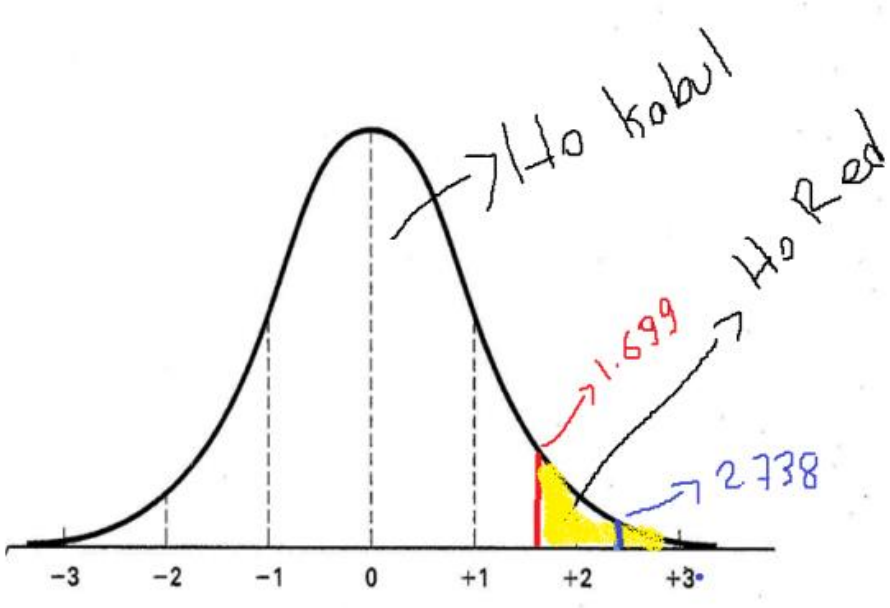
Örnek:

- X üniversitesindeki öğrencilerin IQ (zeka) puan ortalaması 100'dür. Yeni geliştirilen sıvıyı içen öğrencilerin zeka seviyelerinde bir farklılık oluşacaktır. Daha sonra iksir içen öğrencilerden 30'u yansız olarak seçiliyor ve zeka düzeyleri ölçülüyor. Ölçümler sonucunda örneklem ortalaması: 110 ve standart sapması 20 olarak hesaplanıyor. Evren ortalaması ve örneklem ortalaması arasındaki bu farkın gerçekten var olduğunu nerden bilebilirim? Ya bu fark şans eseri ortaya çıkmışsa?
- H0 : $\mu = 100$ vs Ha : $\mu < 100$
- α : 0.05
- Evren ortalaması: 100
- Örneklem Ortalaması: 110
- Örneklem standart sapması: 20
- n: 30
- Serbestlik derecesi: 30-1= 29

- $t = \frac{\bar{X}-\mu}{s/\sqrt{n}} = \frac{110-100}{20/\sqrt{30}} = \frac{10}{3.651} = 2.738,$

- 29 serbestlik derecesi için 0,05 düzeyinde kritik değer 1.699

- Bulduğumuz t değeri tablo değerinden yüksek olduğu için H0 hipotezini reddeder yani örneklemin ortalamasının grup ortalamasından farklı olduğunu belirtiriz.



Varyans Analizi (Anova)

- Varyans analizi =ANOVA (Analysis of Variance)
- İki ya da daha çok evrene ait ortalamaların karşılaştırılması
- Bağımlı değişkene işaret eden örneklemlerin bağımsız değişken açısından karşılaştırılması üzerine kuruludur.
- Tek yönlü ANOVA: Bir çalışmada ya da deneyde bağımlı değişken üzerindeki etkisi araştırılan tek bir bağımsız değişken olduğunda yapılır (Bağımsız değişken = faktör).
- Dört farklı eğitim uygulanmış grubun sosyal beceri düzeyleri karşılaştırılmak istensin (Uygulanan eğitime göre sosyal beceri değişiyor mu?)
- Bağımsız değişken (faktör): Eğitim Türü
- Bağımlı değişken: Sosyal Beceri
- Bağımsız değişkenin (faktörün) her bir durumu: Düzey (k) ya da işlem
- Bağımsız değişkenin farklı durumları, puanlarda manidar farklılık yaratıyorsa, buna faktörün işlem etkisi denir.
- T testi ile benzer mantığa dayanır; t testinde iki örneklem ortalamaları arasındaki farkı, iki ortalama arasındaki standart hatayla karşılaştırarak değerlendiriyorduk. ANOVA'da ise t testinin bir uzantısı olarak iki ya da daha çok ortalama karşılaştırılır

Anova Varsayımları

1. Bağımlı değişkenin ölçüldüğü ölçek en az eşit aralık ölçeği düzeyindedir.
2. k örneklem bağımsızdır ve evrenlerden yansız olarak seçilir.
3. Örneklemlerin seçildiği evrenlere ait puanların dağılımı normaldir.
4. k evrene ait varyanslar homojendir.
5. Bağımsız değişken kategoriktir.

- Grupların varyansının birbirine eşit olması (homojenlik) şartının sağlanıp sağlanmadığını kontrol etmek için kullanılan yöntemlerden biri: F maksimum testi.
- F değeri, varyansı en büyük olan grubun varyansının, varyansı en küçük olan grubun varyansına bölünmesi ile elde edilir.
- ANOVA'nın temeli F oranıdır.
- İki ya da daha fazla örneklem ortalamasının farklı μ'leri temsil edip etmediğini belirlemek için tüm örneklem ortalamalarının aynı anda karşılaştırılmasını sağlar.
- Varyans analizi (ANOVA), varyansları analiz eder; varyansı kısımlara ayırarak puanlar arasındaki farkları ölçer.
- Puanlardaki toplam varyansın iki kaynağı vardır:
  - Gruplararası (açıklanan varyans):Bağımlı değişkendeki toplam varyansın bağımsız değişken tarafından açıklanan miktarıdır
  - Gruplariçi (açıklanamayan varyans): Faktörün her düzeyinin kendi içindeki değişkenliğini gösterir

- Eğer $H0$ doğru ise KO_GA ve KO_Gİ 'nin eşit ya da benzer olacağı tahmin edilebilir. Bu durumda, grup ortalamaları arasındaki değişkenlik sadece örneklem hatasına bağlı olacaktır (istendiği gibi bağımlı değişken (deneysel uygulamadan ya da incelenen kategorik değişkenden) değil.
- Eğer $H0$ yanlış ise KO_GA'nın, KOGİ den büyük olacaktır. Bu durumda da bazı evren ortalamaları birbirinden farklı olacaktır

F Dağılımı

- ANOVA analizi yapmamız için gerek ve yeter koşullardan biri de F dağılımını bilmemizdir.
- ANOVA'nın temel aldığı dağılımdır.
- F dağılımı, sürekli bir dağılımdır.
- Varyanslar bilindiğinde, büyük olan varyans değeri küçük olana bölünür (F= KO_GA/KO_Gİ)
- $H0$ doğru ise F oranının 1'e eşit olacağı tahmin edilir
- $H0$ yanlış ise F oranı 1'i aşacaktır
- Her F dağılımında, pay ve paydaya karşılık gelen iki farklı serbestlik derecesi vardır. ANOVA'da,
  - KO_GA (pay) için serbestlik derecesi grup sayısından bir çıkarılması (sd=k-1)
  - KO_Gİ(payda) için serbestlik derecesi bütün gruplardaki gözlem sayısından grup sayısının çıkarılması (sd= N-k)

F Oranının Hesaplanması

ANOVA'da varyans (S_X^2) kareler toplamının (KT) serbestlik derecesine (sd) bölünmesi ile elde edilir. Varyans, ANOVA'da kareler ortalaması (KO) olarak bilinir:

$$S_X^2 = \frac{\sum (X - \overline{X})^2}{N - 1} = \frac{KT}{sd} = KO;$$

Burada serbestlik derecesi N-1'dir.

Gruplararası kareler ortalaması (KO_{GA}), gruplararası kareler toplamının (KT_{GA}) gruplararası serbestlik derecesine (sd_{GA}) bölünmesi ile hesaplanır. Bunun gibi, gruplariçi kareler ortalaması ($KO_{Gİ}$) da gruplariçi kareler toplamının ($KT_{Gİ}$), gruplariçi serbestlik derecesine ($sd_{Gİ}$) bölünmesi ile bulunur. KO_{GA}'nın, $KO_{Gİ}$'na oranlanması ile de F değeri bulunur. Aşağıda tek yönlü ANOVA için bir özet tablo verilmektedir.

Tablo 9.1: *Tek Yönlü ANOVA Özet Tablosu*

Varyansın Kaynağı	Kareler Toplamı	Sd	Kareler Ortalaması	F
Gruplararası	KT_{GA}	Sd_{GA}	KO_{GA}	F_{Hesap}
Grup içi	$KT_{Gİ}$	$Sd_{Gİ}$	$KO_{Gİ}$	
Toplam	KT_{Top}	Sd_{Top}		

Tablo 4.4. İlişkisiz Ölçümler İçin Bir Yönlü ANOVA (Formüller)

Varyansın Kaynağı	Kareler Toplamı (KT)	Serbestlik Dereccesi (sd)	Kareler Ortalaması (KO)	F-Oranı
Gruplararası	KT_A	A–1	$[KT_A/A–1] = KO_A$	KO_A/KO_e
Gruplariçi	KA_e	n–A	$[KT_e/n–A]=KO_e$	
Toplam	KT_T	n–1		

Örnek:

Yaşın yaratıcılık üzerindeki etkisinin incelendiği bir çalışmadan elde edilen veriler tabloda verilmektedir:

Yaş 4	Yaş 7	Yaş 10	Yaş 13
3	9	9	7
5	11	12	7
7	14	9	6
4	10	8	4
3	10	9	5

a. F değerini hesaplayın. ANOVA özet tablosunu hazırlayın.

b. Manidarlık düzeyi .05'de, hesaplanan F değeri için karar verin.

c. Yaratıcılık puanlarında açıklanabilen varyans miktarını hesaplayarak yorumlayın?

	Yaş 4	Yaş 7	Yaş 10	Yaş 13	
	3	9	9	7	
	5	11	12	7	
	7	14	9	6	
	4	10	8	4	
	3	10	9	5	
∑X	22	54	47	29	152
$\sum X^2$	108	598	451	175	1332
n	5	5	5	5	20
$\bar{X}$	4.4	10.8	9.4	5.8	30.4

Kareler Toplamlarının Hesaplanması

1. Toplam kareler toplamının hesaplanması (KT_{Top}) için şu formül kullanılır:

$$KT_{Top} = \sum X_{Top}^2 - \left(\frac{\left(\sum X_{Top}\right)^2}{N} \right)$$

Bu formüle tabloda hesaplanan değerler yerleştirilerek KT_{Top} hesaplanır.

$$KT_{Top} = 1332 - \frac{152^2}{20} = 1332 - 1155.2 = 176.8$$

2. Gruplararası kareler toplamının (KT_{GA}) hesaplanmasından da şu formül kullanılır:

$$KT_{GA} = \sum \left(\frac{(sütündakipuanlartoplamı)}{sütundakipuanadedi} \right) - \left(\frac{\left(\sum X_{Top}\right)^2}{N} \right)$$

Bu formüle tablodaki her sütun için hesaplanan değerler ve genel X toplamının karesi yerleştirilerek KT_{GA} hesaplanır.

$$KT_{GA} = \left(\frac{(22)^2}{5} + \frac{(54)^2}{5} + \frac{(47)^2}{5} + \frac{(29)^2}{5}\right) - \frac{152^2}{20} = 134.8$$

$$KT_{Gİ} = KT_{GT} - KT_{GA} = 176.8 - 134.8 = 42.00$$

Kareler Ortalamasının Hesaplanması

Daha önce hesaplanan gruplararası ve gruplariçi kareler toplamının kendi serbestlik derecelerine bölünmesi ile gruplararası kareler ortalaması (KO_{GA}) ve gruplariçi kareler ortalaması ($KO_{Gİ}$) elde edilir.

$$KO_{GA} = \frac{KT_{GA}}{sd_{GA}} \qquad \textbf{Formül 9.7} \qquad KO_{Gİ} = \frac{KT_{Gİ}}{sd_{Gİ}}$$

Gruplararası ve gruplariçi kareler toplamı ile serbestlik dereceleri formüldeki yerlerine konulduğunda, kareler ortalaması aşağıdaki gibi bulunur. Serbestlik derecesi, gruplariçi için N-K=20-4=16, gruplarası için k-1=4-1=3'tür.

$KO_{GA} = \frac{KT_{GA}}{sd_{GA}} = \frac{134.8}{3} = 44.93$ ve $KO_{Gİ} = \frac{KT_{Gİ}}{sd_{Gİ}} = \frac{42}{16} = 2.63$

F Oranının Hesaplanması

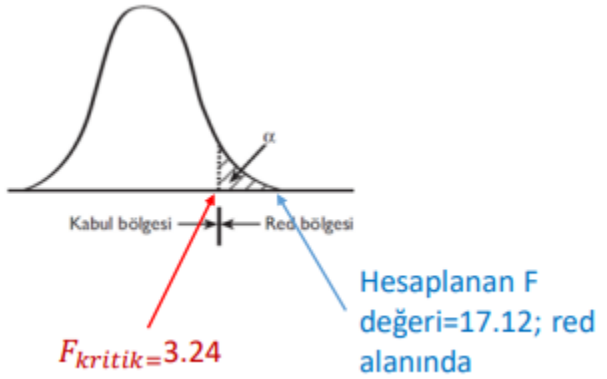
F değerinin hesaplanması için yukarıda bulunan KO_{GA} ve $KO_{Gİ}$'nin birbirine oranı alınır. Aşağıda F için hesaplama formülü verilmektedir:

$$F = \frac{KO_{GA}}{KO_{Gİ}}$$

Bu çalışmada, $KO_{Gİ} = 2.63$ ve $KO_{GA} = 44.93$

Buna göre $F = \frac{44.93}{2.63} = 17.117$

Varyansın Kaynağı	Kareler Toplamı	sd	Kareler Ortalaması	F
Gruplararası	134.80	3	44.93	17.12
Gruplariçi	42.00	16	2.63	
Toplam	176.80	19		



$F_{(3,16)}$= 17.12 ve F_{kritik}=3.24'tür.

Hesaplanan F değeri, kritik F değerinden büyüktür, null hipotezi reddetme alanındadır. Bu nedenle, F değeri manidardır.

Pratik anlamda, ortalamalar arasında fark yoktur, diyen null hipotez yanlışlanır. Karşılaştırılan grup ortalamalarından en az ikisi arasında manidar fark vardır.

Regresyon Analizi

- Regresyon analizi değişkenler arasındaki ilişkiyi incelemek ve modellemek için kullanılan istatistiksel bir yöntemdir.
- Regresyon analizinin en önemli alanlarından biri değişkenler arasındaki ilişkiyi açıklayabilecek doğru modele karar vermektir. Modele karar verebilmek için değişkenler arasındaki ilişkinin ön analizinin yapılmasıdır. Bu analize dayalı olarak model belirlenmelidir
- Regresyonun, mühendislik, fizik ve kimya bilimler, iktisat, yönetim, yaşam ve biyoloji bilimleri ve sosyal bilimler gibi hemen hemen tüm alanlarda farklı amaçlarla kullanılmaktadır.
- Bu amaçlardan bazıları şunlardır:
- Verinin tanımlanması ve özetlenmesi,
- Parametre kestirimleri,
- Kestirim ve önkestirimler,
- Denetleme.

- Doğrusal regresyon analizi, bağımlı değişken ile bir veya daha fazla bağımsız değişken arasında bir ilişki kurar.
- Doğrusal model, bağımlı değişkeni bağımsız değişkenin aldığı değerin doğrudan oranı olarak gösterir.
- Basit Doğrusal regresyon analizinde sadece bir bağımsız değişken bulunur.

- Şimdi bakalım, Bağımlı ve Bağımsız değişken nedir?
- **Bağımlı değişken (y)**
- Regresyon modelinde açıklanan veya tahmin edilecek olan değişkendir. Bu değişkenin bağımsız değişkenle fonksiyonel bir ilişkide olduğu varsayılır.
- **Bağımsız değişken (x)**
- Regresyon modelinde bağımlı değişken ile ilişkili değişkendir. Bağımsız değişken, regresyon modelinde bağımlı değişkenin değerini tahmin etmek için kullanılır.

Basit Doğrusal Regresyon Modeli

- **$y = \alpha + \beta x + \varepsilon$**

- y= bağımlı değişken
- x= bağımsız değişken
- α = sabit (y-eksenini kestiği nokta)
- β= regresyon doğrusunun eğimi
- ε= hata terimi veya artık

Regresyon Parametreleri

α = sabit

- doğrunun y eksenini kestiği nokta.
- Bağımsız değişkenin değerinin = 0 olduğu durumda bağımlı değişkenin aldığı değerdir.

β = eğim

- Bağımsız değişkendeki değişime dayalı olarak bağımlı değişkende görülen değişimdir.
- Eğimin alacağı katsayının işareti iki değişken arasındaki ilişkiye bağlı olarak pozitif veya negatif olabilir

EKK ile tahminlenen Regresyon Modeli

$$\hat{y} = a + bx$$

$\hat{y}$ = Tahmin edilen y değeri (bağımlı değişken)

a = regresyon sabit değerinin yansız tahmini

b = regresyon eğiminin yansız tahmini

x = bağımsız değişken değeri

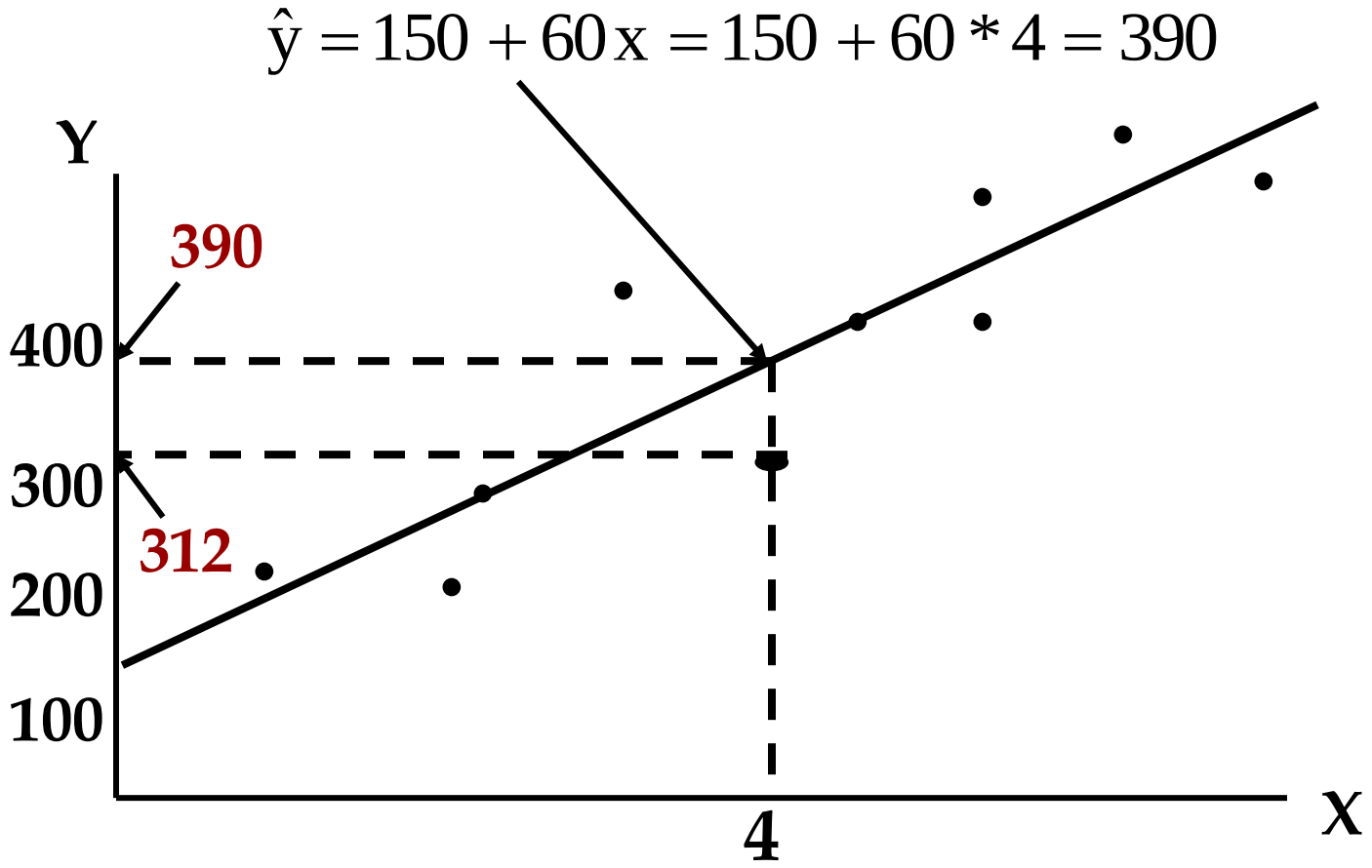
Basit doğrusal regresyon modelin bazı varsayımları bulunmaktadır:

- ε hata terimlerinin her biri istatistiksel olarak bir diğerinden bağımsızdır.
- ε hata terimlerinin aldığı değerler normal dağılım özelliği göstermelidir.
- Hata varyansı sabittir ve veriler arasında hiç değişmediği varsayılır. Buna otokorelasyon veya serisel korelasyon bulunmaması varsayımı adı verilir.
- Bağımsız değişken hatasızdır. Eğer bağımsız değişkende hata bulunduğu varsayılırsa özel bir yöntem şekli olan değişkenler-içinde-hata modeli teknikler kullanılarak model kurulmalıdır.

- ε : Hata Terimi (artık)
- Regresyon modelleri tam (%100) doğru tahmin yapma özeliğine sahip değillerdir. Hata terimi (artık), gözlenen değer ile model tarafından tahmin edilen değer arasındaki farktır.

$$\varepsilon = y - \hat{y}$$

Artık terminin (hata) grafiksel gösterimi



ε= Artık = 312 - 390 = -78

Regresyon Parametrelerinin Tahmini

- b ve a katsayıları aşağıdaki eşitlikler kullanılarak hesaplanır :

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$\mathrm{a} = \bar{\mathrm{y}} - \mathrm{b}\bar{\mathrm{x}}$$

- Örnek: Gebelik haftası ile hemoglobin düzeyi arasında anlamlı bir ilişki bulunmakta mıdır?

No	Hafta	Hemoglobin	No	Hafta	Hemoglobin
1	33	10.8	11	33	10.5
2	33	9.5	12	30	11.0
3	23	14.2	13	35	10.9
4	34	9.7	14	25	14.0
5	32	11.2	15	22	13.8
6	35	9.7	16	28	12.9
7	30	12.1	17	27	12.8
8	23	13.0	18	29	11.0
9	28	12.0	19	24	13.5
10	26	13.2	20	31	10.8

No	Hafta (x)	Hemo. (y)	x^2	xy
1	33	10.8	1089	356.4
2	33	9.5	1089	313.5
3	23	14.2	529	326.6
4	34	9.7	1156	329.8
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
17	27	12.8	729	345.6
18	29	11.0	841	319.0
19	24	13.5	576	324.0
20	31	10.8	961	334.8
Total	**581**	**236.6**	**17215**	**6761.6**

- Regresyon parametrelerinin tahmini

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{6761.6 - \frac{581 * 236.6}{20}}{17215 - \frac{(581)^2}{20}} = -0.331$$

$$a = \bar{y} - b\bar{x} = \frac{236.6}{20} - (-0.331)\frac{581}{20} = 21.4$$

$$y = 21.4 - 0.331x$$

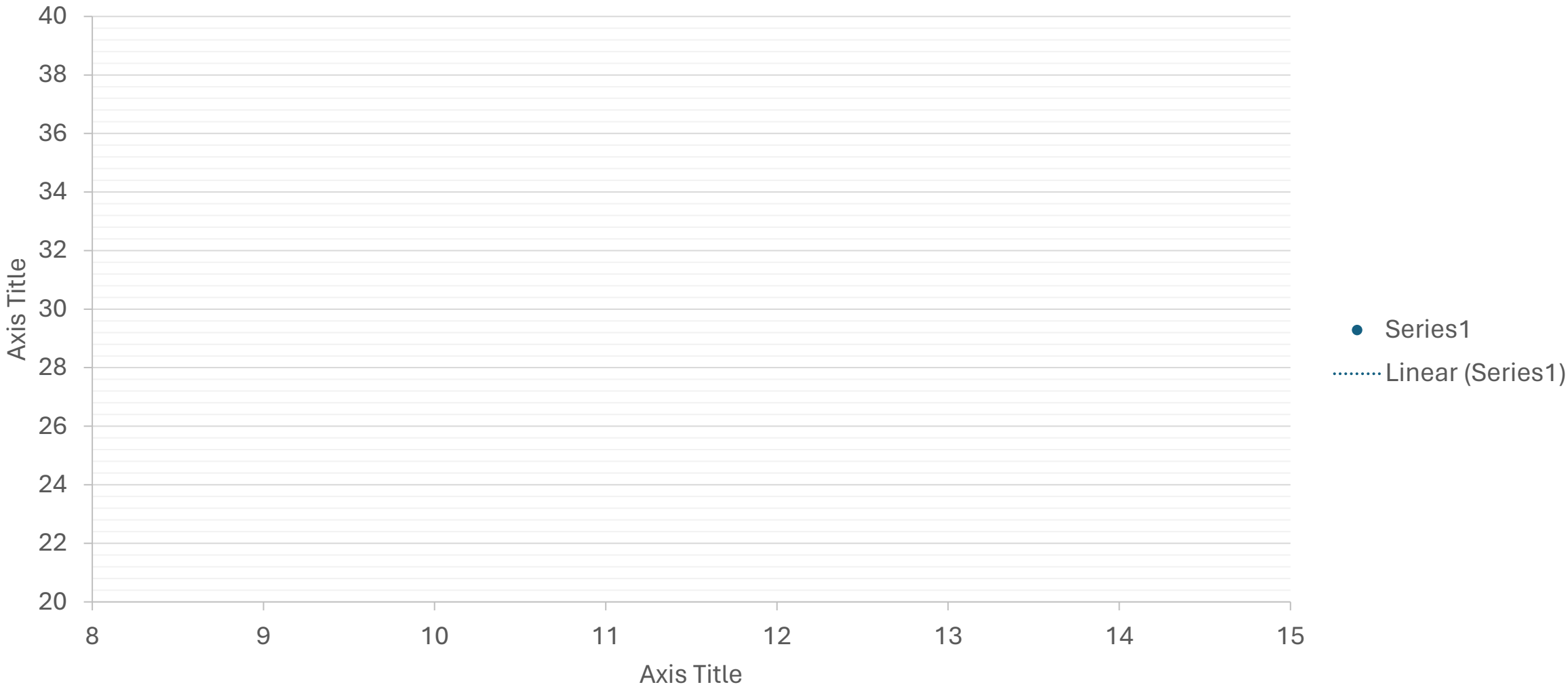
- Eğim parametresinin (b) anlamlılığının testi

$H_0 : \beta = 0$

$H_1 : \beta \neq 0$

$$S_b = 0.033 \qquad t = \frac{b}{S_b} = \frac{-0.331}{0.033} = -10.1$$

- $t\alpha$, n-(p+1)= t(0.05, 18)= 2.1, t= 10.1 > t(0.05, 18)= 2.1, H0; eğim sıfır değildir. Red.
- (n= örneklem genişliği, p= bağımsız değişken sayısı)


40
38
36
34
32
30
28
26
24
22
20
Axis Title
8
9
10
11
12
13
14
15
Axis Title
Series1
Linear (Series1)

Determinasyon Katsayısı (R^2)

- Deneysel verilerin doğrusal bir eğriye ne kadar iyi uyduğunun en iyi ölçütü, regresyon analiz işleminde hesaplanmış “determinasyon katsayısıdır.
- R^2 = 1 olması, deneysel verilerin kusursuz bir doğrusal eğri sağlandığının kanıtıdır.
- Ne kadar çok veri noktası varsa, R^2’nin güvenirliği o kadar yüksektir.
- R^2=0.85 ise, y değişkenindeki toplam değişimin %85’i bağımlı değişken x tarafından açıklanabilirken, %15’i açıklanamaz olarak anlamlandırılır.

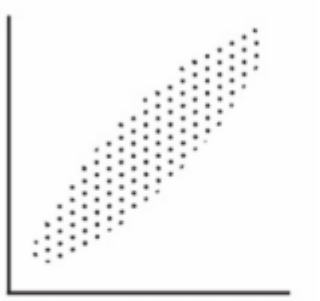
$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

Kolerasyon Katsayısı (r)

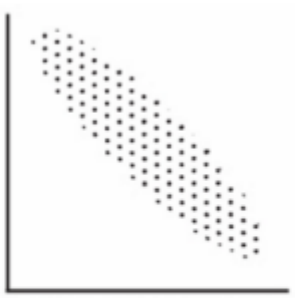
- Korelasyon iki ya da daha fazla değişken arasındaki doğrusal ilişkiyi gösterir.
- İki değişken arasındaki ilişki miktarı, ikili ya da basit korelasyon denen korelasyon teknikleriyle hesaplanır.
- Örneğin;
- Öğrencilerin okul öncesi eğitime başlama yaşları (ay olarak) ile birinci sınıf başarısı arasındaki ilişki,
- Öğrencilerin istatistik başarı puanları ile istatistiğe yönelik tutumları arasındaki ilişki,

Basit Korelasyon

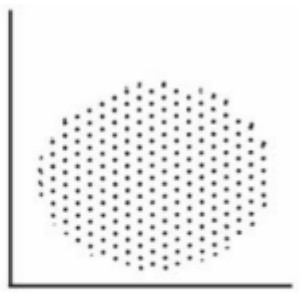
- Bir bireye ait iki ölçüm olduğunda bu iki değişken arasındaki ilişkiyi belirler.
- Korelasyon analizi sonucunda, doğrusal ilişki olup olmadığı ve varsa bu ilişkinin derecesi korelasyon katsayısı ile hesaplanır.
- Korelasyon katsayısı “r” ile gösterilir ve -1 ile +1 arasında değerler alır.



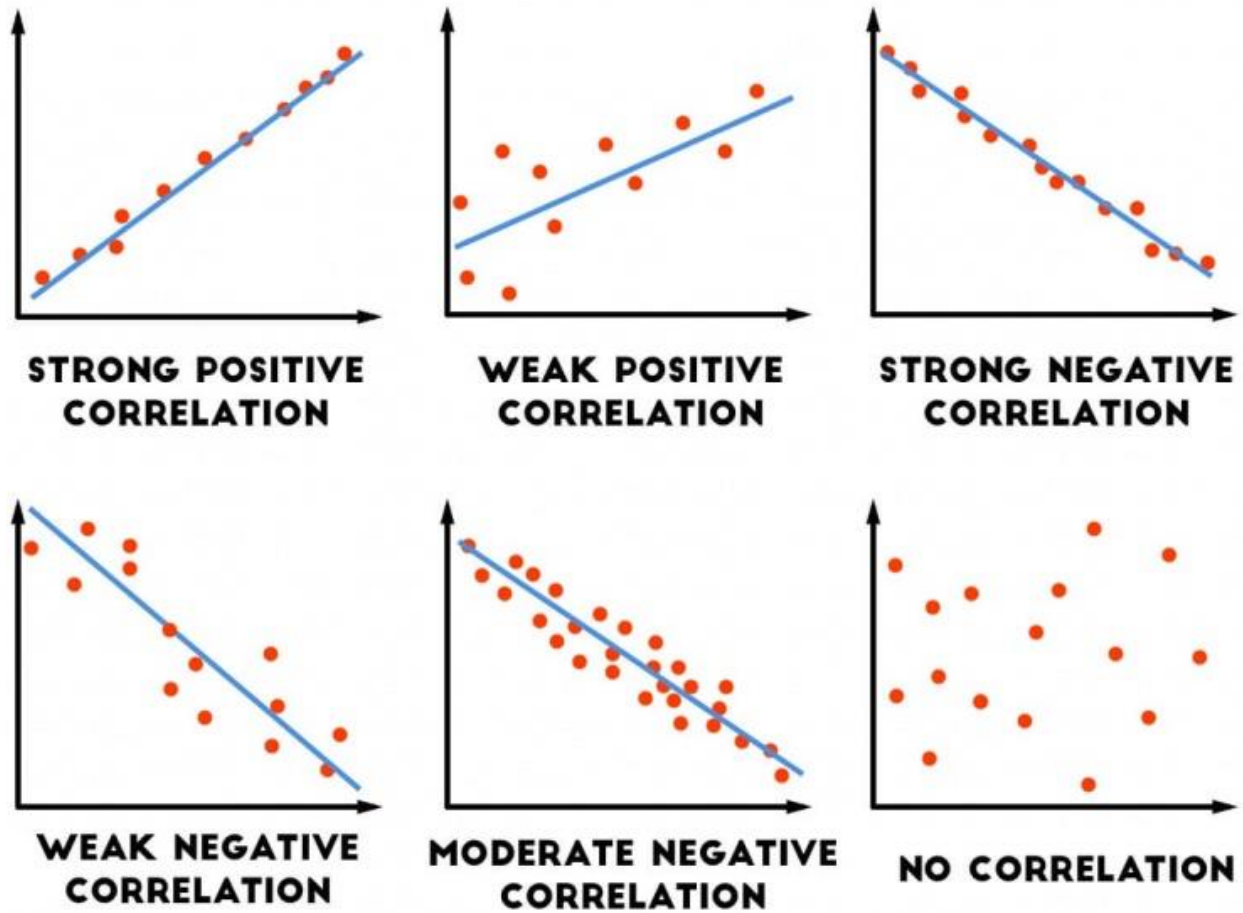
Pozitif yönlü ilişki



Negatif yönlü ilişki



İlişki yok


STRONG POSITIVE CORRELATION
WEAK POSITIVE CORRELATION
STRONG NEGATIVE CORRELATION
WEAK NEGATIVE CORRELATION
MODERATE NEGATIVE CORRELATION
NO CORRELATION

Pearson Korelasyon

- En az eşit aralıklı ölçek düzeyinde ölçülen iki sürekli değişken arasındaki doğrusal ilişkinin derecesinin belirlenmesinde kullanılır.
- A ve B değişkenleri arasında anlamlı bir ilişki var mıdır? sorusunun cevabı aranır.

- Korelasyon katsayısı -1 ile +1 arasında değerler alır.
- r = -1 ise tam negatif doğrusal bir ilişki vardır.
- r = +1 ise tam pozitif doğrusal bir ilişki vardır.
- r = 0 ise iki değişken arasında ilişki yoktur.

r	İlişki
0.00	ilişki yok
0.01 - 0.29	düşük düzeyde ilişki
0.30 - 0.70	orta düzeyde ilişki
0.71 - 0.99	yüksek düzeyde ilişki
1.00	mükemmel ilişki

- Kolerasyon hesaplamalarında,

- Popülasyon için

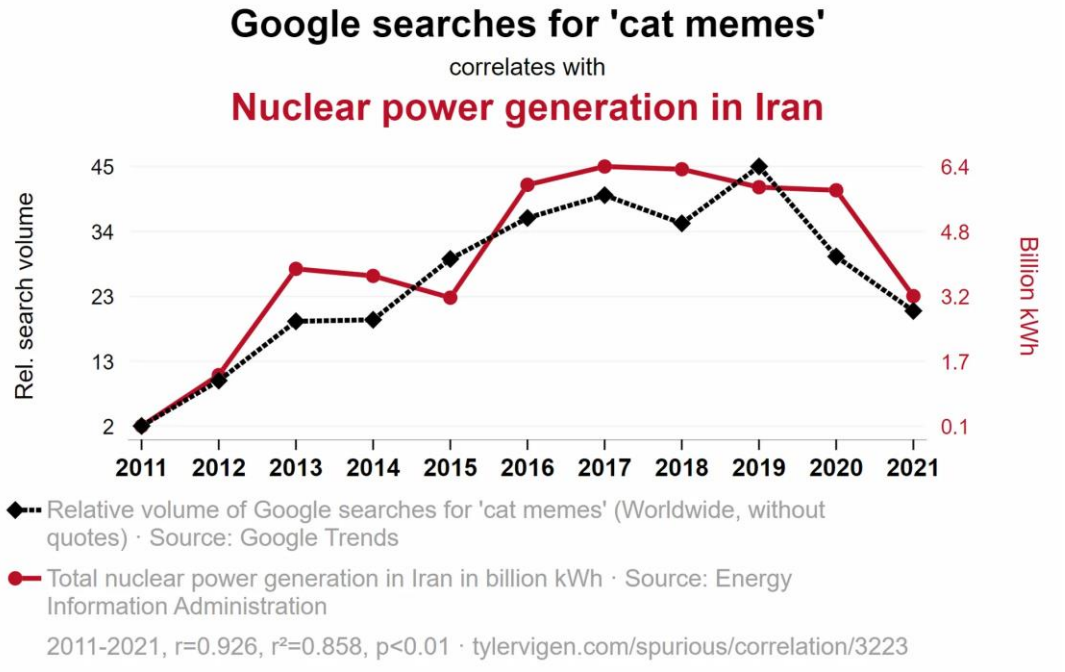
$$\rho = \frac{\sum (X_i - \mu_x)(Y_i - \mu_y)}{\sqrt{\sum (X_i - \mu_x)^2 \sum (Y_i - \mu_y)^2}}$$

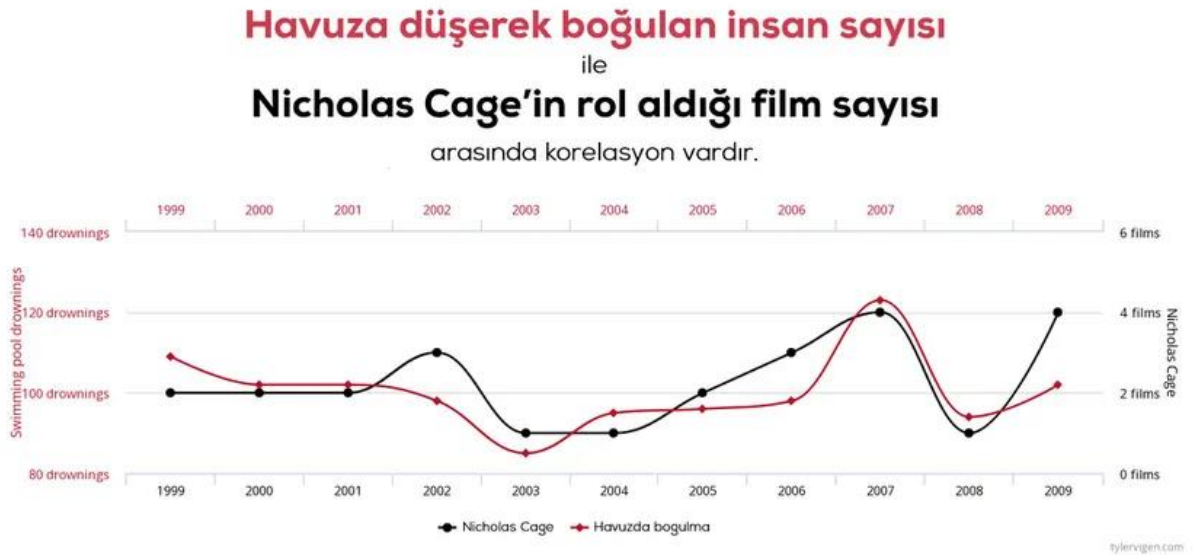
- Örneklem için

$$r = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}}$$

Korelasyon vs Nedensellik

- İki değişken arasında doğrusal bir ilişki olması (biri artarken veya azalırken, diğerinin de ilkiyle aynı yönde artması veya azalması), ikinci değişkende gördüğümüz değişimin sebebinin birinci değişkenin değişimi olduğu anlamına gelmez!
- Bu tür ilişkilere anlamsız ilişki (İng: "spurious correlation") adı verilmektedir. Bir unsurun bir sonucun nedeni olabilmesi için, o unsur ile o sonuç arasında takip edilebilir bir etkileşim olmalıdır.





Ve son olarak r mi R^2 mi

- Değişkenlerin birbirlerinde açıkladıkları varyans miktarı korelasyon katsayının karesine eşittir ve buna determinasyon katsayısı denir.

- Örneğin Türkçe başarısı ile okuma hızı arasındaki korelasyon r=0.80 olsun. Buna göre determinasyon katsayısı R^2 = 0.64’dür.

Veri Toplama ve Kalitesi Ölçme Yolları

- Veri toplama çeşitli yollarla yapılabilmektedir. Doküman/kayıt incelemesi, anket, soru formu, test, görüşme, odak grup, gözlem, kontrol listesi, etnografya, sözlü tarih, örnek olay incelemesi ve deney veri toplamak amacıyla daha yaygın olarak başvurulan yöntemlerdir
- Veri kalitesi yetersiz ise gerekli bilgiye ulaşmak çok maliyetli olabilir. Ayrıca doğru olmayan veriler muhasebe başta olmak üzere pek çok konuda maliyet kaybına sebep olabilir. Örneğin satış rakamlarının yanlış girilmesi, tamir ve bakım gerektiren cihazların onarım tarihlerinin geç yazılması. Bu hususta dikkate alınması gereken birçok nokta vardır.
- Ondalık sayıların kullanımından virgül veya nokta kullanılmasına, giriş alanlarına € veya $ gibi sembollerin girilmesine kadar tüm detaylar dikkat edilmesi gerekir. Özellikle standartlar veya geriye dönük dönüşümler yoksa satış rakamlarını doğru bir şekilde düzeltmek oldukça zor olabilir. Neyin doldurulması gerektiğini bulmak ve onu onarmak çok zaman ve para gerektirir.

- Veri kalitesi, onu kullanmak istediğiniz amaca ne ölçüde uygun olduğudur. Ancak bir veri doğruluk, zaman, tutarlı vb. gibi sekiz hususu karşılıyorsa kaliteli anlamına gelir. Verilerinizin kaliteli olduğundan eminseniz size pek çok fayda sağlar. Veri kalitesinin temel faydaları ise şöyle sıralanabilir:

- **Karar verme:** Veri kalitesi ne kadar iyi olursa, çalışanlar ürettikleri sonuçlara o kadar güvenir, sonuçlardaki riski azaltır ve verimliliği artırır. Sonuçlar güvenilir olduğunda, tahminde bulunma ve karar vermedeki risk azaltılabilir.
- **Üretkenlik:** Kaliteli veriler, çalışanları daha üretken kılar. Veri hatalarını doğrulamak ve onarmak için zaman harcamak yerine temel görevlerine odaklanabilirler.
- **Uyum:** Yasama organlarının müşterilerle ilişkilerin veya iş faaliyetlerinin nasıl gerçekleşeceğini belirlediği sektörlerde, iyi veri kalitesi önemli ölçüde mali kayıpları önleyebilir. Çünkü bu veriler arasında uyum çok önemlidir.
- **Pazarlama:** Daha iyi veriler, özellikle birçok kuruluşun faaliyet gösterdiği veya çalışmak istediği çok kanallı ortamlarda doğru hedefleme ve etkili müşteri iletişimi sağlar.
- **Rekabet avantajı:** İyi veri kalitesi rekabet avantajı sağlayabilir, çünkü bir kuruluş müşteriler, ürünler ve süreçler hakkında daha iyi öngörü kazanır ve pazar fırsatlarını daha hızlı belirleyebilir.
- **Mali kazanç:** Bir kuruluş veri kalitesini ciddiye alıyorsa getiriler de gözle görülür şekilde iyileşir ve bunun sonucunda kurum daha değerli hale gelir. Tabii ki daha fazla kar elde eder.

Veri Kalitesi Nedir?

- Veri kalitesinin net bir tanımını yapmak zordur. Gerçek şu ki, veriler kullanma amacına ulaşırsa veri kaliteniz iyidir. Örneğin, kuruluşa yön vermek için bir yönetim panosunda doğru değerlerin gösterilmesi, yönetimin de tutarlı olmasını ve sürecin doğru yönetilmesini sağlar.

- Eksiksiz olması: Tüm veriler girildi mi, eksik bilgi var mı?
- Benzersizlik: Verilerde yinelenen değerler var mı
- Zaman tutarlılığı: Doğru veriler belirli bir zaman noktası için mi talep ediliyor?
- Geçerlilik: Veriler belirlenen kurallara göre girildi mi?
- Doğruluk: Girilen veriler doğru mu? Veriler gerçeği doğru bir şekilde yansıtıyor mu?
- Tutarlılık: Veriler farklı depolama konumlarında da aynı mı?
- Netlik: Veriler yoruma açık mı, yoksa sadece bir şeyi mi ifade ediyor?
- Alaka düzeyi: Veriler, kullanıcı ve kullanım amacı ile tam olarak ilgili mi?

Veri Görselleştirme, Neden?

- Teknolojik gelişmelere paralel olarak depolama maliyetlerinin düşmesi ile artan
- veri hacmi ve bu veriden hızlı bir şekilde anlam çıkarma ihtiyacı,
- Görsel anlatımın diğer anlatım tekniklerinden daha etkili olması,
- Dolayısıyla sektörde artan veri üzerinden görsel araçlarla anlam çıkarma
- konusunda uzman iş gücü ihtiyacı

Neden “Görsel Anlatım” ?

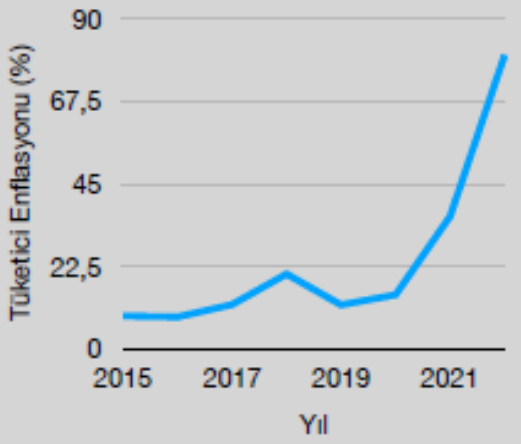
Metin

2015 yılından bu güne Türkiye’deki tüketici enflasyonu incelendiğinde, her yıl bir önceki yıla göre artış göstermiştir. 2015 yılında bir önceki yıla göre %8.81 artan enflasyon izleyen yıllarda sırasıyla, %8.53, %11.92, %20.3, %11.84, %14.6, %36.08 ve %80.21 oranında artış göstermiştir.

Tablo

Yıl	Tüketici Enflasyonu (%)
2015	8.81
2016	8.53
2017	11.92
2018	20.3
2019	11.84
2020	14.6
2021	36.08
2022	80.21

Grafik



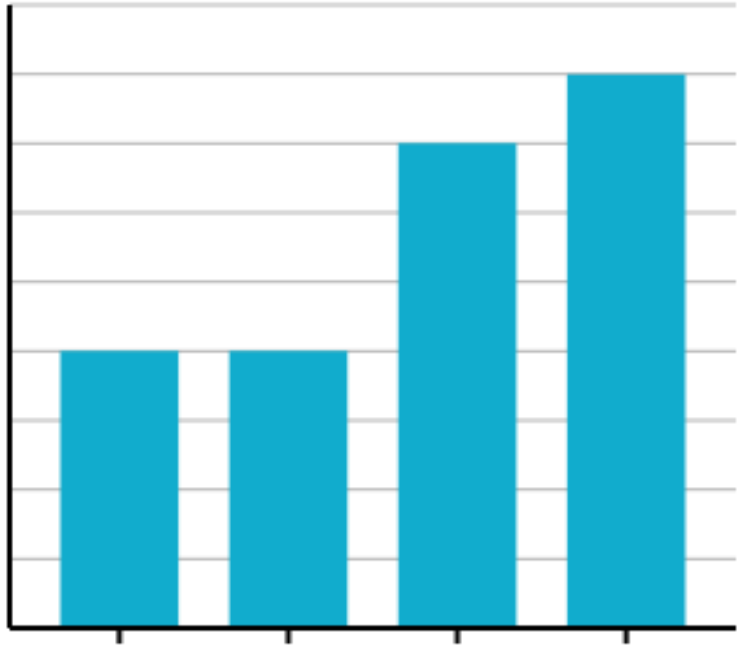
İnsan beyni görsel verileri, metin verilerinden 60.000 kat daha hızlı işler.*

Miktarların Görselleştirilmesi

- Miktar görselleştirme çalışmalarında:
- Çubuk grafiği (bar-plot)
- Gruplandırılmış çubuk grafiği (grouped bar-plot)
- Yığılmış çubuk grafiği (stacked bar-plot) kullanılabilir. Çubuk grafiklerine alternatif olarak, nokta grafikleri (dot-plot) ve ısı haritaları (heatmap) da kullanılabilir.

Çubuk Grafikleri

- Çubuk grafiklerinde sık karşılaşılan iki sorunlardan biri, etiketler arasında herhangi bir mantıksal sıralama düzeni olmadığı durumlarda çubukların rastgele sıralanmasıdır.
- Grafiğin daha kolay yorumlanabilmesi için çubukların artan ya da azalan sırada düzenlenmesi gerekmektedir.
- Diğer sorun ise, eksenlerde yer alan etiket isimlerinin yatayda çok fazla yer kapması, hatta üst üste binmesi nedeniyle okunamaz halde olabilmesidir. Bu durumda en etkili çözüm eksenlerin ters çevrilmesidir.



Gruplandırılmış Çubuk Grafikleri

- Miktar görselleştirirken veri setinde birden fazla kategorik değişken olması durumunda gruplandırılmış çubuk grafikleri kullanılabilir.

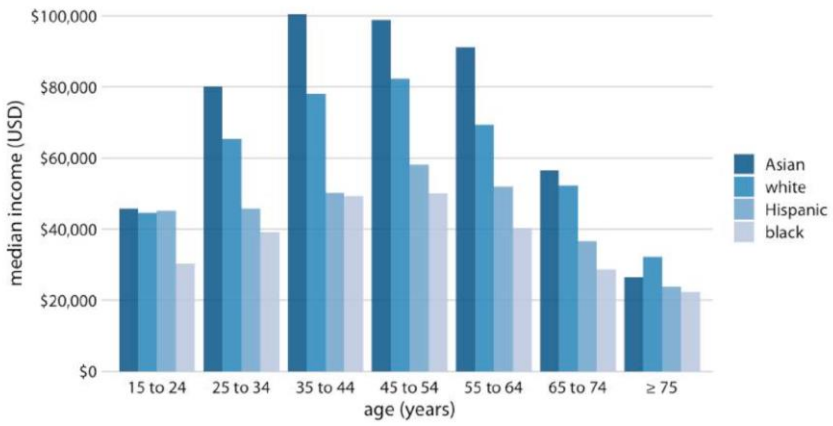


Figure 6-7. 2016 median US annual household income versus age group and race. Age groups are shown along the x axis, and for each age group there are four bars, corresponding to the median income of Asian, white, Hispanic, and black people, respectively. Data source: US Census Bureau.

Nokta Grafikleri

- Çubuk grafiklerinin en önemli dezavantajı, temsil ettikleri miktarlar ile orantılı olabilmeleri için sıfır noktasından başlamaları gerekliliğidir. Bu durum düzey sayısı ve miktar arttığında grafiklerin okunmasını zorlaştırabilir. Bu gibi durumlarda nokta grafikleri iyi bir alternatiftir

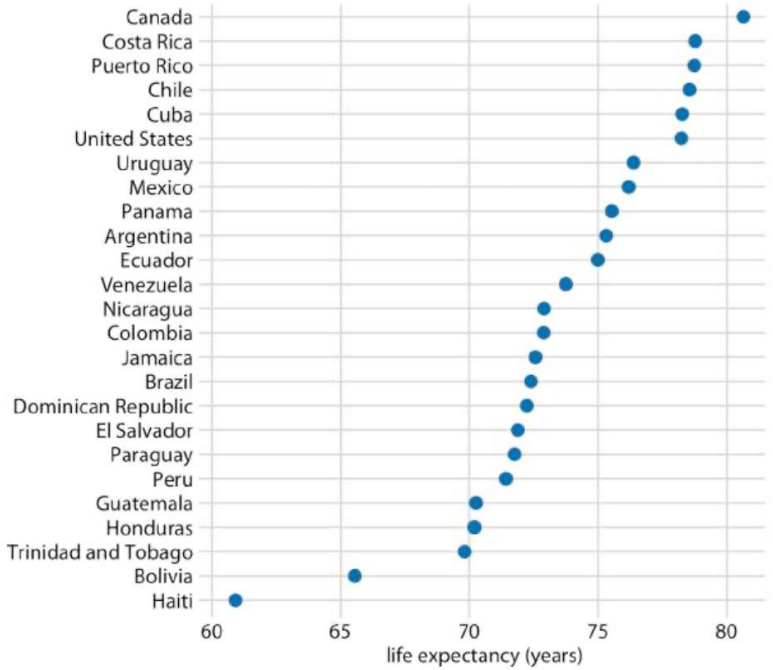


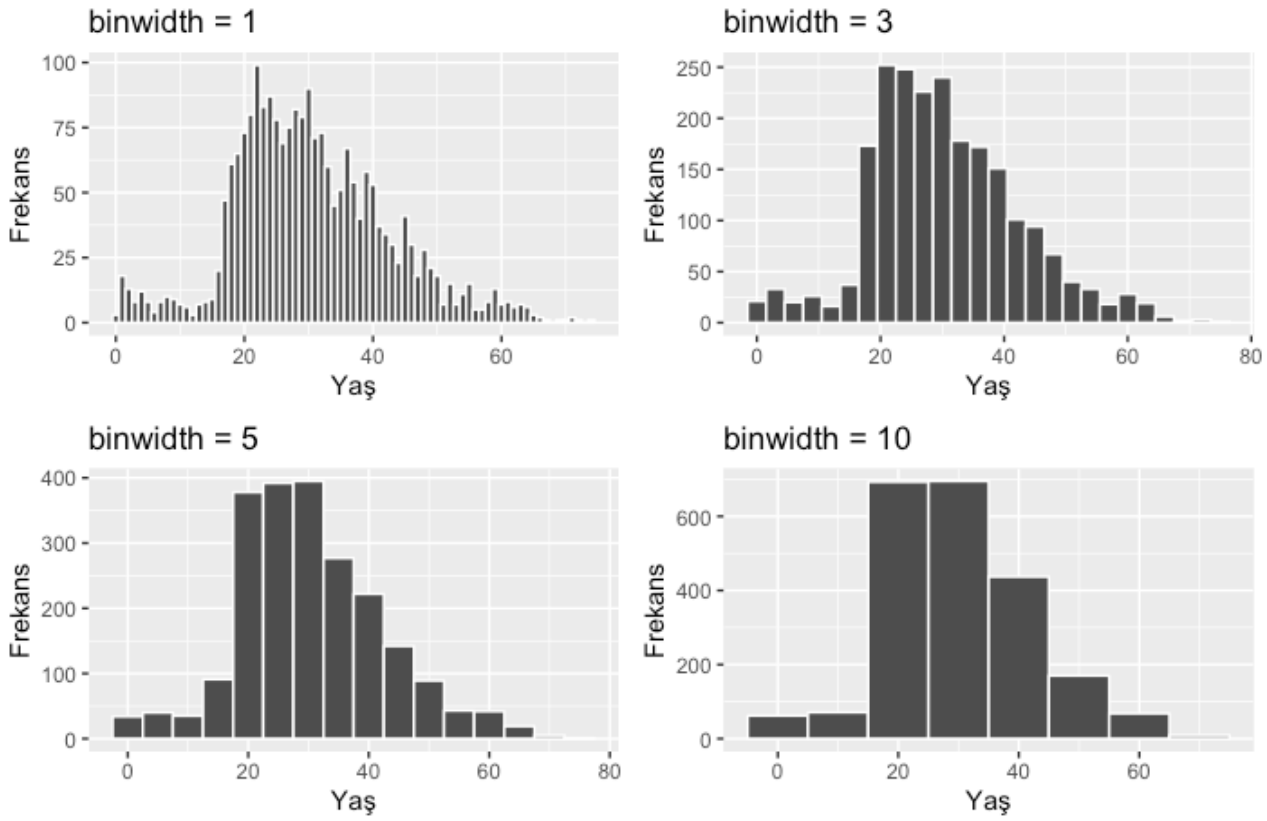
Figure 6-11. Life expectancies of countries in the Americas, for the year 2007. Data source: Gapminder.

Dağılımların Görselleştirilmesi

- Bir değişkenin dağılımının görselleştirilmesi için:
- Histogram
- Kernel yoğunluk tahmini kullanılır.

Histogram

- Gözlem değerlerinin sabit kutu genişliklerine göre gruplandırılarak görselleştirilmesi ile oluşturulur.
- Histogram oluşturulmasında en önemli sorun, görünümünün seçilen kutu genişliğine bağlı olmasıdır.
- Eğer kutu genişliği olması gerektiğinden daha küçük seçilirse, histogramda aşırı pik değerler gözlemlenir ve yorumlanması zorlaşır.
- Olması gerektiğinden daha geniş seçilirlerse, küçük aralıklardaki önemli değişimler histogramda kaybolur ve tespiti mümkün olmayabilir.



Kernel Yoğunluk Tahmini

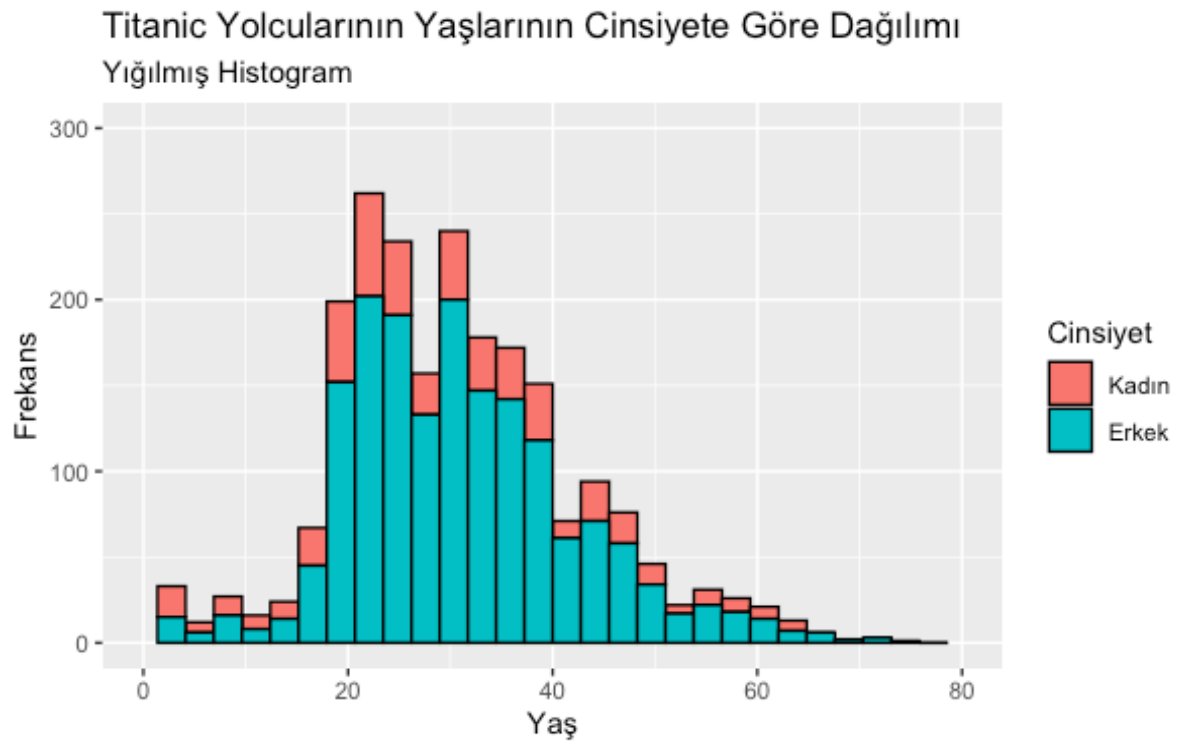
- Pratikte histogram daha sık tercih ediliyor olsa da, kernel yoğunluk tahmininin kullanımı son yıllarda artmıştır.
- Kernel yoğunluk tahminlerinin en önemli sorunu hiç bir gözlem bulunmayan noktalarda gözlem varmış gibi bir görsel ortaya çıkarabilmesidir.
- Örneğin yaş değişkeni gibi negatif değerler almayan bir değişkenin görselleştirilmesinde negatif bir yaş değeri ile karşılaşılabilir. Bu gibi durumlara karşı dikkatli olunması gerekmektedir.

Birden Fazla Değişkenin Dağılımının Görselleştirilmesi

- Sıklıkla birden fazla değişkenin dağılımının görselleştirilmesinin gerektiği durumlarla karşılaşabiliriz.
- Örneğin, Titanic yolcularının yaşlarının cinsiyete göre dağılımlarını incelememiz, aşağıdaki sorulara yanıt verme ihtiyacı duyabiliriz:
- Erkek ve kadın yolcuların ortalama yaşları benzer miydi?
- Cinsiyetlere göre yolcu yaşları arasında bir fark var mıydı?
- Bu gibi durumlarda iki cinsiyet grubu için ayrı ayrı histogramlar oluşturulabilir ya da yığılmış histogram kullanılabilir.

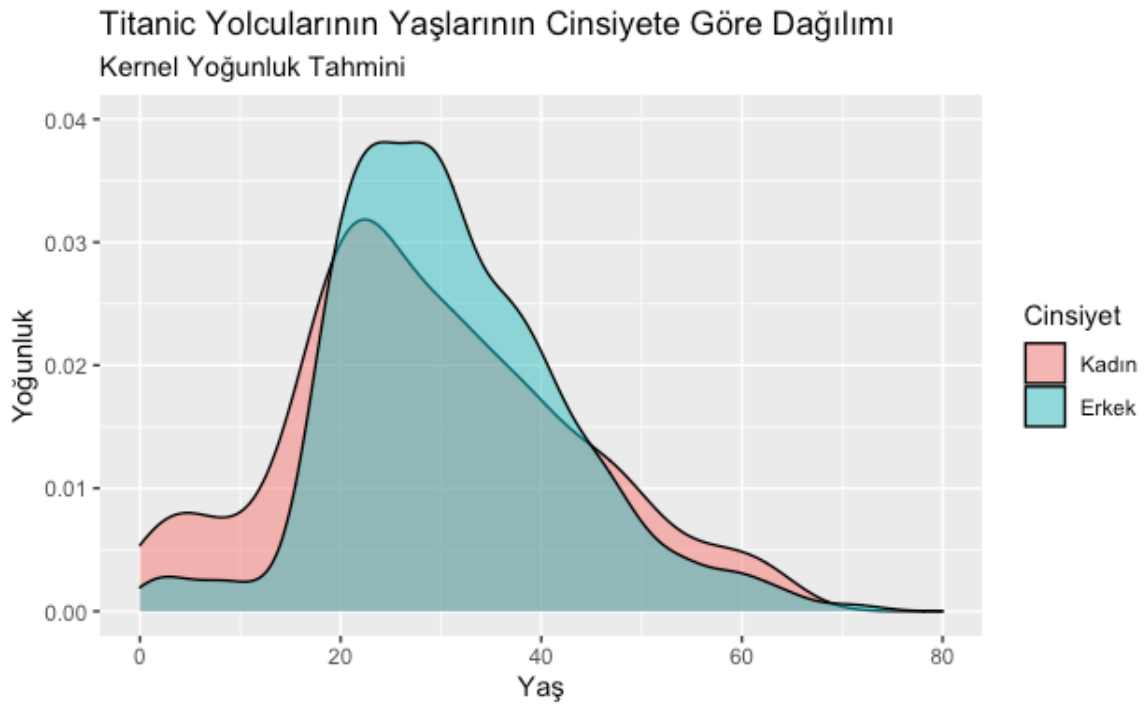
Yığılmış (stacked) Histogram

- Yığılmış histogram, grupları temsil eden çubukların farklı renkler ile üst üste çizilmesidir.



Kernel Yoğunluk Tahmini

- Yığılmış histogramın sınırlılıklarından dolayı birden fazla grup için Kernel yoğunluk tahminini kullanmak daha iyi bir çözümdür.

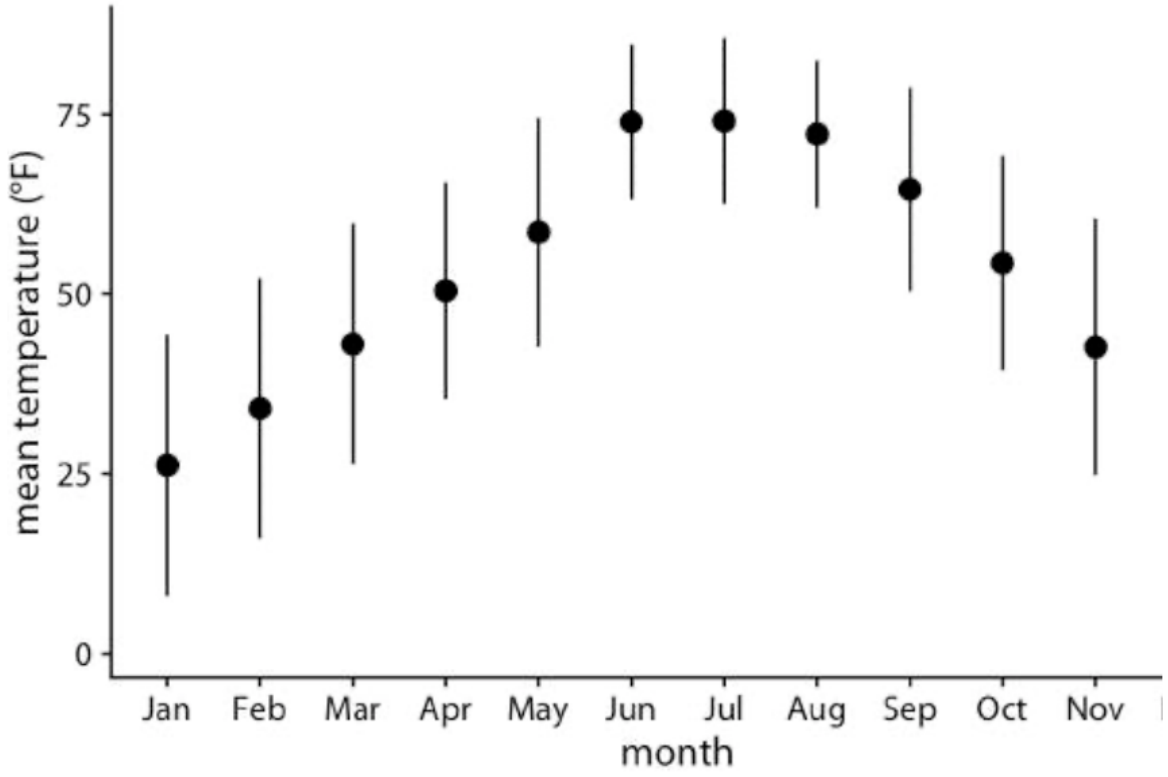


Birçok Değişkenin Dağılımının Görselleştirilmesi

- Aynı anda birçok değişkenin dağılımının görselleştirilmesi gereken durumlarla karşılaşılabilir:
- Aylık hava sıcaklıklarının dağılımı
- Ülkelerin kişi başına gelirlerinin dağılımı

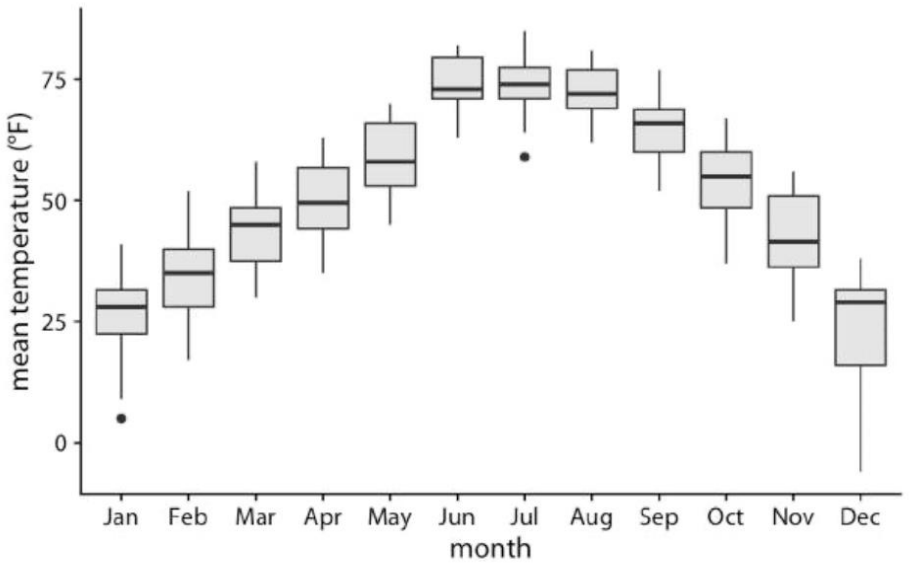
Hata Çubukları

- Bir çok dağılımı aynı anda görselleştirmenin en basit yolu hata çubuklarını kullanmaktır. Hata çubukları farklı şekillerde oluşturulabilir. Bu yollardan biri, medyanın nokta, medyanın bir standart sapma uzaklığını da çubuklar ile göstermektir.



Kutu-Bıyık (Box-and-whisker) Grafiği

- Veriyi 5 nokta (min, first quantile, median, third quantile, max



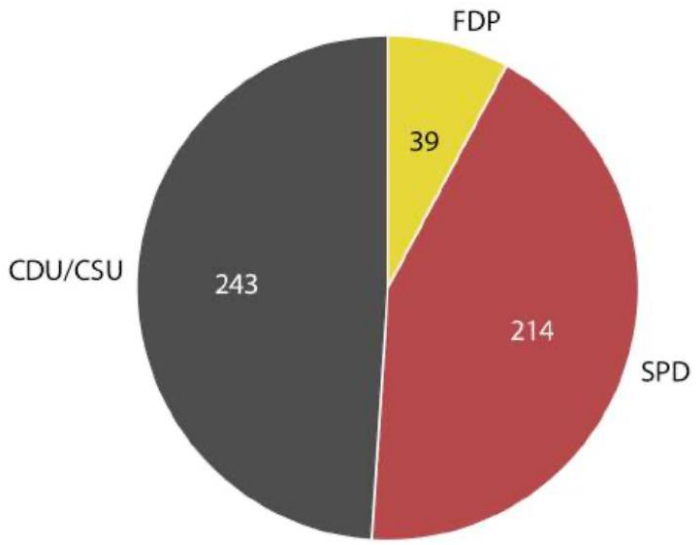
Oranların Görselleştirilmesi

- Pratikte bir grubun veya miktarın nasıl her biri bütünün bir oranını temsil eden ayrı ayrı parçalara ayrıldığını göstermek gerekebilir.
- Örneğin:
- bir topluluktaki kadın ve erkeklerin oranı,
- bir seçimde farklı siyasi partilere oy veren seçmenlerin oranı,
- şirketlerin pazar payları vb.

- Bu gibi oranların görselleştirilmesinde en sık kullanılan grafik türleri, pasta (piechart) ve çubuk (bar-plot) grafikleridir.

Pasta Grafiği (Pie Chart)

- Pasta grafiği, bir daireyi her dilimin alanı temsil ettiği ve genel toplama orantılı olacak şekilde dilimlere ayırır.
- Pasta grafikleri genellikle yarım, üçte bir ya da dörtte bir gibi basit oranları görselleştirirken daha iyi çalışır.
- Çok sayıda grup olduğunda iyi çalışmazlar. Bu gibi durumlarda çubuk grafikleri daha iyi bir alternatiftir.

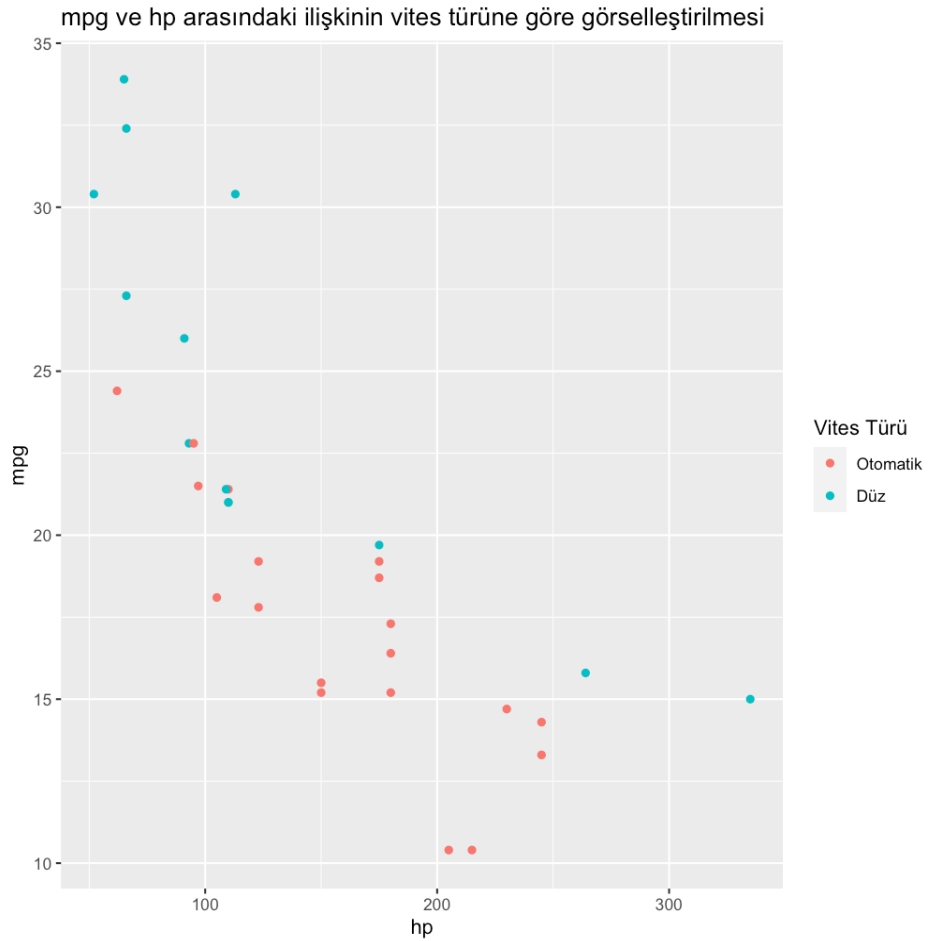


İlişkilerin Görselleştirilmesi

- Korelasyon, nedensellik belirtmez; ancak bir nedensellik aramak için iyi bir sebeptir!

1) Saçılım grafikleri (scatter-plot)

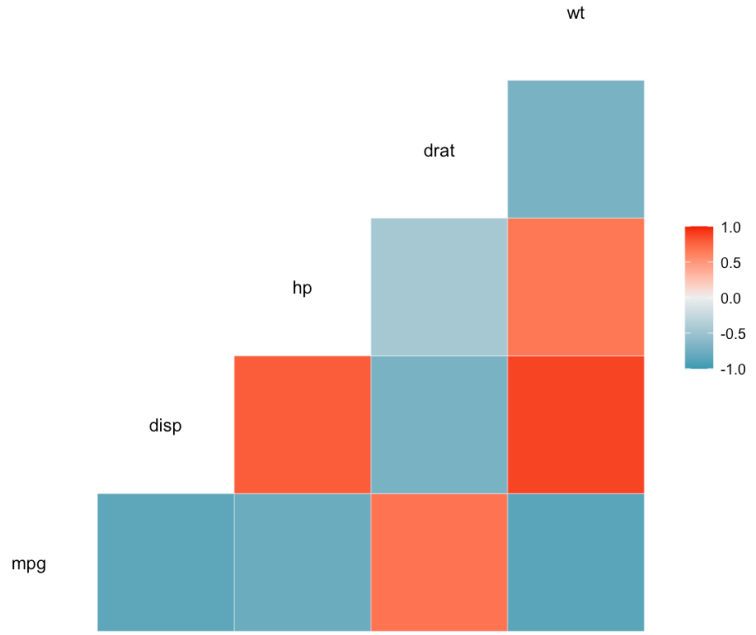
- İki değişken arasındaki ilişkinin araştırılması için kullanılan temel araçlardan biridir.
- İki boyutlu bir eksen üzerinde, her eksende bir değişkenin gözlem değerleri diğer eksendeki değişkenin gözlem değerleriyle eşleştirilerek oluşturulur.



Korelogram (correlogram)

- Değişken sayısı arttığında saçılım matrislerini okumak zorlaşmaktadır. Bu durumda ilişkiyi görselleştirmektense ilişki miktarını ölçmek ve bu ölçümleri görselleştirmek daha kullanışlıdır

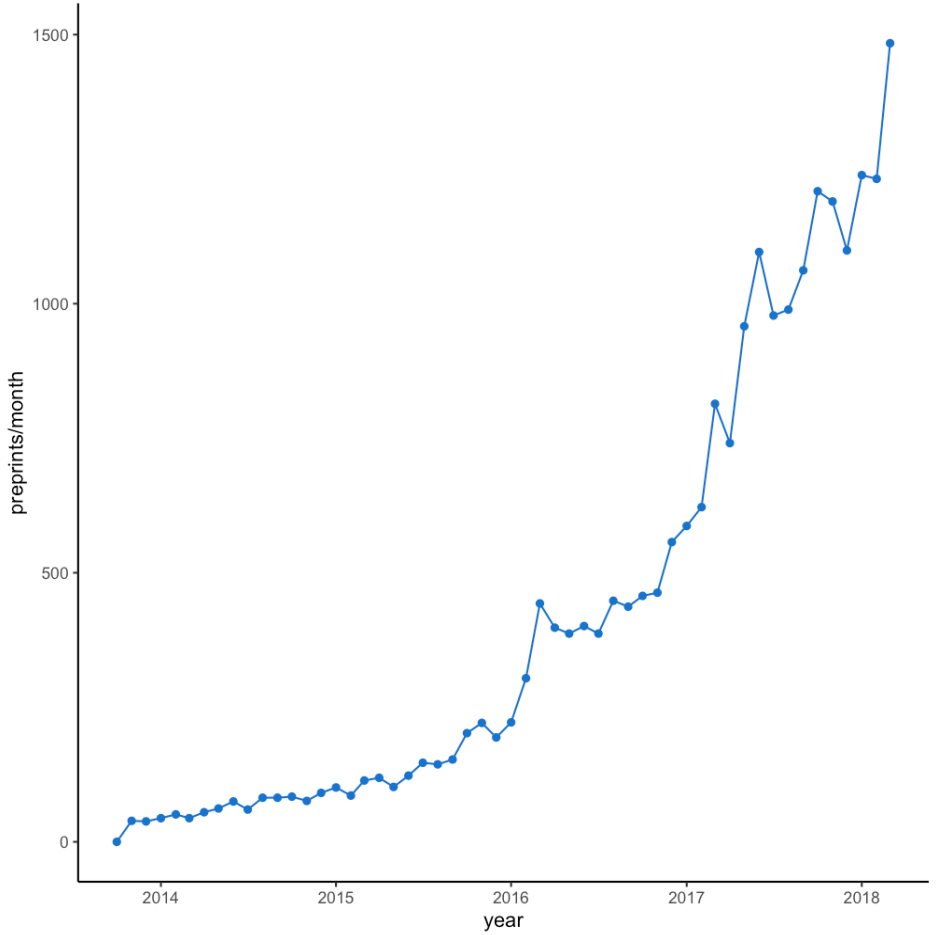
- Bunu yapabilmenin en kolay yolu, korelasyon katsayısını (r) hesaplamaktır. Korelasyon katsayısı -1 ve +1 arasında değerler alır.
- r = 0 değerini aldığında iki değişken arasında herhangi bir ilişki olmadığını, -1 ve ya +1 değerini aldığında ise yüksek düzeyde bir ilişki olduğunu gösterir.
- Korelasyon katsayısının işareti ise ilişkinin yönünü gösterir.
- Korelasyon katsayılarının görselleştirilmesi için korelogram kullanılır.



Zaman Serilerinin Görselleştirilmesi

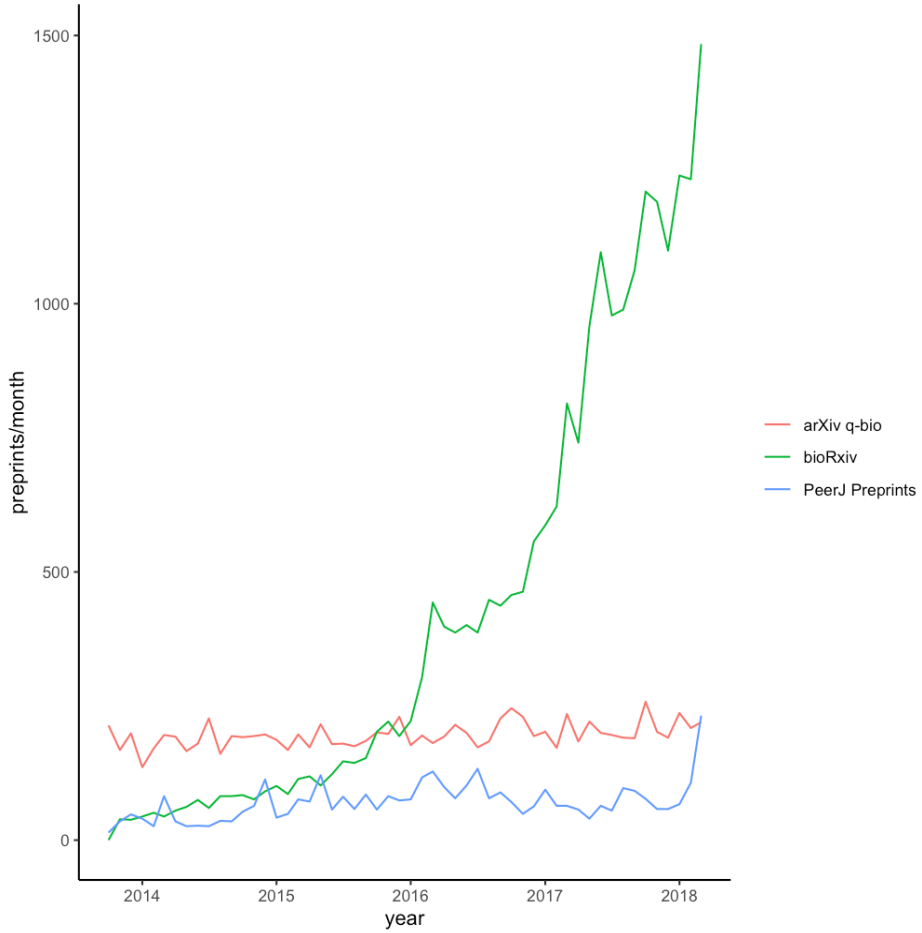
- Zaman serisi nedir?

- Zamana göre sıralanan veri noktalarının bütünüdür.
- Ardışık düzende gözlemlenirler.
- Sinyal işleme, ekonometri ve istatistik başta olmak üzere bir çok farklı
- disiplinde kullanım alanı bulunmaktadır.



Birden fazla zaman serisinin görselleştirilmesi

- Birden fazla zaman serisini karşılaştırmak ya da birlikte değişimlerini gözlemlemek amacıyla aynı grafik üzerinde görselleştirilmesi yaygın olarak kullanılan bir stratejidir



Veri Düzenleme Yöntemleri

- Veri düzenleme adımlarını yapmadan önce kontrol etmemiz gereken birkaç adım vardır, bunlardan biri, Aykırı değerleri tespit etmek,
- Aykırı değerlerin tespiti bizler için önemlidir.
- Birçok istatistiksel test ve makine öğrenmesi algoritması aykırı değerlere karşı hassastır. Bu nedenle, aykırı gözlemlerin tespit edilip, gözden geçirilmesi ve duruma göre müdahale edilmesi gerekmektedir. Aykırı değerler:
- Verilerin dağılımını ve ortalama, medyan vs. gibi veriyi temsil eden istatistikleri etkiler.
- Modellerden elde edilen sonuçlara etki eder.
- İstatistiksel testlerin gücünü düşürür.

Veri Düzenleme Yöntemleri

- Nasıl yapılır?
- Ne gibi karşımıza çıkar?
- Çözümleri nelerdir?