

Statistics for Data Science

- Örnek Teorisi
- Betimsel İstatistikler
- Güven Aralıkları
- Olasılık Dağılımları
- Hipotez Testleri
- Varyans Analizi
- Korelasyon Analizi

■ Örnek Teorisi

Merkezi Limit Teoremi: Bağımsız ve aynı dağılıma sahip rassal değişkenlerin toplamı ya da aritmetik ortalaması yaklaşık olarak normal dağılmaktadır. "Ana kitleden alacağımız örneklerin ortalamasını aratırın etme."

■ Betimsel İstatistikler

- Kovaryans
- Ortalama
- Medyan
- Mod
- Kartiller
- Degr̄im Aralığı
- Standart Sapma
- Kovaryans
- Korelasyon

→ Kovaryans = İki değişken arasındaki ilişkinin dēişkenlik ölçesidir.

$\text{cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$
iki rastgele değişkenin kendi ortalamalarından olan sapmalarının çarpımının beklenen değeridir.

→ Korelasyon = İki değişken arasındaki ilişkiyi, ilişkinin anlamlı olup olmadığını ilişkinin şiddetini ve yönünü ifade eden istatistiksel bir tekniktir.

$$r_{xy} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sqrt{(\sum x_i^2 - n \bar{x}^2)} \sqrt{(\sum y_i^2 - n \bar{y}^2)}}$$

■ Güven Aralıkları

Güven Aralığı = Anakölle parametresinin tahmini değerini kapsayabilecek iki sayıdan oluşan bir aralık bulunmasıdır.

"Ölçümün hassasiyetinin bir göstergesidir. Ayrıca yapmış olduğumuz tahminlerin ne kadar güvenilir olduğunuyla ilgili bir değer sunar."

Güven Aralığı Nasıl Hesaplanır?

Adım 1: n , ortalama ve standart sapmayı bul.

$n=100$, $\text{ortalama}=180$, $\text{standart sapma}=40$

Adım 2 = Güven aralığına karar ver 95 mi? (Varsayılan her yerde %95'tir)

Z tablo değerini hesapla (%95'e göre 1,96-2,57)

Adım 3 = Yukarıdaki değerleri kullanarak güven aralığını hesapla

$$\bar{x} \pm Z \frac{s}{\sqrt{n}} = 180 \pm 1,96 \times \frac{40}{\sqrt{100}}$$

Ortalamanın etrafına standart sapmayı ve gözlem sayısını ve Z tablo değerini göz önünde bulundurarak bir aralık hesapla.

Sonuç = $180 \pm 7,84$ yani 172 ile 188 arasıdır.

"İstatistikte olarat %95 güvenilirlik ile değerlerimiz bu aralıktadır"

"100 denemeden 5'i bu aralığın dışında olabilir"

■ Olasılık Dağılımları

Rassal Değişken= Değerlerini bir deneyin sonuçlarından alan değişken rassal değişken denir.

Dağılım Nedir= Evrende gerçekleşen olaylar ya da durumların Sayısal karşılıklarının ortaya gitirdiği yapıya dağılım denir.

Olasılık Dağılımı= Bir rassal olaya ait değerler ve bu değerlerin gerçekleşme olasılıklarının bir arada ifade edilmesine denir.

Olasılık Fonksiyonu= Bir değişkenin herhangi bir değeri alması olasılığını hesaplamaya yarayan fonksiyondur.

Olasılık= Olayların olabilirliğinin sayısal ifadesidir.

Kesikli Olasılık Dağılımları

- Bernoulli
- Binom
- Poisson
- Bernoulli = Başarılı-Başarısız, olumlu-olumsuz şeklindeki iki sonuçlu olaylar ile ilgilenildiğinde kullanılan kesikli olasılık dağılımıdır.

$$f(x; p) = p^x (1-p)^{1-x}, x \in \{0, 1\}$$

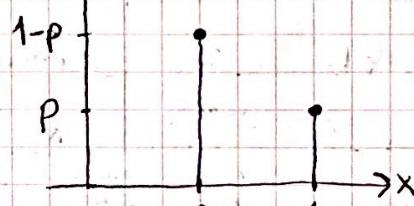
$$E(X) = p \quad \text{var}(X) = pq = p(1-p)$$

p =olasılık
 x =kesikli değişkenin
alacağı değer (0 veya 1)

Sürekli Olasılık Dağılımları

- Normal Dağılım
- Üniform Dağılım
- Üstel Dağılım

$f_X(x) = p_X(x) \quad X \sim \text{Bernoulli}(p)$



- Binom Dağılımı: Bağımsız n deneme sonucu k başarılı olma olasılığı ile ilgilenildiğinde kullanılan dağılımdir.

$$f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, 2, \dots, n$$

→ Bir madeni para 4 kez atılıyor. 2 kere yazı gelme olasılığı. Gözümü:

$$f(2; 4, 0.5) = \binom{4}{2} 0.5^2 (1-0.5)^{4-2} = 0.375$$

- Poisson Dağılımı: Belirli bir zaman aralığında belirli bir alanda nadiren rastlanan olayların olasılıklarını hesaplamak için kullanılır.

$$f(x, \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, 2, \dots, n \quad E(X) = \lambda \quad \text{Var}(X) = \lambda$$

λ =lambda (galışmadan bulduğumuz
veya daha önce hesaplanmış değer)
Beklenen sonucun ortalaması
gerçekleşme sayısı"

#Poisson dağılımında ort ve varyans birbirine eşittir. Lambda değeridir.
 e =Sabit
 x =İlgilendiğimiz olayın ortaya çıkma sayısı

n : büyük (gözlem sayısı)
 p : küçük (gerçekleşme olasılığı)

$n > 50$
 $n * p < 5$ beklenir

#Rassal denemeler iki sonuçlu olmalıdırını koşullar altında gerçekleştirmeli.
→ Rassal denemeler birbirinden bağımsız olmalıdır.

→ Bir Üniversitede 5000 not girişinde 5 tane notun yanlış girilmesi olasılığı?
Dağılımin Poisson olduğu biliniyor ve Lambda = 0,2.

$$f(5; 0.2) = \frac{0.2^5 e^{-0.2}}{5!} = 0.0000218328201$$

- Normal Dağılım: Normal dağıldığı bilinen sürekli rassal değişkenler için olasılık hesaplaması yapmak amacıyla kullanılır.

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$\mu = \text{mü (ortakma)}$
 $x = \text{rassal değişkenin değeri}$
 $\sigma^2 = \text{varyans}$
 $\pi = \text{sabit}$ $e = \text{sabit}$

Sırekli değişkenlerin olasılığı hesaplandığı için bir aralığı alabildiğimizden dolayı aralık olasılığında integral alınır.

- Hipotez Testleri (Şans eseri gerçekleşmesini reddetme/reddedememe durumu)
 Bir inanış (bir savi, birtahmini vs.) test etmek için kullanılan istatistiksel bir tekniktir.

Hipotezler ve Türleri

Sıfır Hipotez: $H_0: \mu = 50$ $H_0: \mu \leq 50$ $H_0: \mu \geq 50$

Alternatif Hipotez $H_1: \mu \neq 50$ $H_1: \mu > 50$ $H_1: \mu < 50$

Hata Tipleri:

		Hipotez Testi Sonucu Verilen Karar	
		H_0 reddedilmedi	H_0 reddedildi
Gercek	H_0 doğru	Doğru Karar $(1-\alpha) \rightarrow$ Güven Düzeyi	I. Tip Hata α
	H_0 yanlış	II. Tip Hata β	Doğru Karar $(1-\beta) \rightarrow$ Testin Gücü

P-value: $p < 0.05$ (varsayılan kabul edilebilir hata miktarı)

Hipotez testlerinin sonuçlarını değerlendirmek üzere programlar tarafından p-value değeri verilir. Bu değer üzerinden kolayca yorum yapılabilir.

p-value değeri 0.05'ten küçükse ilgili H_0 hipotezini genellikle reddettiğimiz sonucuna varız.

Dağılıma uygunluk testlerinde buna bakıp reddedemeyiz.

Bu tür durumlarda farklı şekilde test ederiz.

H_0 "örnek dağılım ile teorik dağılım arasında fark yoktur" der.

Hipotez Testi Adımları

Adım 1: Hipotezlerin kurulması ve yönlerinin belirlenmesi

Adım 2: Anlamlılık düzeyinin ve tablo değerin belirlenmesi

(Bir olayın olma olasılığının 0 ile 1 arası olması mantığına dayanır)

Adım 3: Test istatistiğinin belirlenmesi ve test istatistiğinin hesaplanması

Adım 4: Hesaplanan test istatistiği ile alfaya karşılık gelen tablo değerinin karşılaştırılması

Test istatistiği (Z_h) $>$ Tablo Değeri (Z_t) ise H_0 Red

Adım 5: Yorum.

■ Tek Örneklem T Testi

Popülasyon ortalaması ile varyansısal bir değer arasında istatistiksel olarak anlamlı bir farklılık olup olmadığını test etmek için kullanılan parametrik bir testtir.
"Elimizde tek bir örnekleme ilişkin test yapma ihtiyacı olduğunda kullanılan testtir!"

Test istatistiği Varsayımlar: Normal (Kad ile uygulandı)
Dagılım

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \quad z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- 1. Anakütle standart sapması biliniyorsa
z istatistiği kullanılır.
- 2. Anakütle standart sapması bilinmiyorsa
ve $n > 30$ ise z istatistiği kullanılır.
- 3. Anakütle standart sapması bilinmiyor
ve $n < 30$ ise t istatistiği kullanılır.

Veri bilimi kapsamındaki genelde $n > 30$ olacağınından dolayı ve örnek sayısı arttıkça t dağılımı normal dağılıma yaklaşacaktır olduğundan dolayı genellikle "t" tercih edilebilir. R ve Python genelde "t" kullanır.

Problem: Sepete ürün ekleme işlemi sonrasında ödeme ekranında 5 adımda vardır ve bu adımların birisi sorulmaktadır.

- Her adının 20'ser sn. olması hedefi var. 4.adım soruluyor.
- Bu durumu test etmek için 100 örnek alınıyor.
- örnek standart sapması 5 saniyedir. örnek ort. ise 19sn.

Adım 1: Hipotezlerin kurulması ve yönlerinin belirlenmesi

$$H_0: \mu = 20$$

$$H_1: \mu \neq 20$$

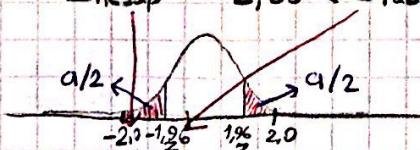
Adım 2: Anlamlılık düzeyinin ve tablo değerinin belirlenmesi
 $\alpha = 0,05 \quad \frac{\alpha}{2} = 0,025$ α = hatalı bir kabul edilebilir

Z tablo olasılık değeri: $0,5 - 0,025 = 0,475$ α = hata miktarı
 Z tablo kritik değer = $-1,96 + 1,96$ z tablosundaki karşılığına bakılır.

Adım 3: Test istatistiğinin belirlenmesi ve hesaplanması

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \quad z_{\text{hesap}} = \frac{19 - 20}{\frac{5}{\sqrt{100}}} = -2,00 \quad n = 100 \quad s. \text{ sapma} = 5 \text{ sn} \quad \text{örnek ort.} = 19 \text{ sn}$$

Adım 4: Z tablo Z hesap karşılaştırması $Z_h > Z_t$ yada $-Z_h < -Z_t$ ise H_0 Red
 $Z_{\text{hesap}} = -2,00 < Z_{\text{tablo}} = -1,96$ olduğundan H_0 reddedilir.



Adım 5: Yorum = 4.adımda geçirilen sürenin 20 saniye olduğunu iddia eden H_0 hipotezi reddedilmiştir. Buna göre kullanıcılar istatistiksel olarak yüzde 95 güvenilirlik ile 4. adımda 20 saniyeden farklı süre geçirmektedir.

Nonparametrik Tek Örneklem T Testi
 Önceki bölümde yapmış olduğumuz Tek Örneklem T testiydi.
 Parametrik bir testti.
 Parametrik test = Geçitli Varsayımların sağlanıldığı durumda uygulanabilen testler.
 → Önceki bölümdeki teste ilişkin varsayımların normallik varsayılmıyordu. Bu varsayımlar sağlanmadığında Nonparametrik Tek Örneklem Testi kullanılır.

Tek Örneklem Oran Testi
 Oransal bir ifade test adımet istenildiğinde kullanılır.

$$H_0: P = P_0 \quad H_0: P \leq P_0 \quad H_0: P \geq P_0 \\ H_1: P \neq P_0 \quad H_1: P > P_0 \quad H_1: P < P_0$$

Test istatistiği:

$$z = \frac{\hat{P} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$$

\hat{P} = örnektenden P_0 = sınamak üzere elde ettiğimiz odaaklılığımız değer.

P_0 = sınamak üzere odaaklılığımız değer.

Varsayımlımız: $n > 30$, n : Topladığımız örneklem
 Bağımsız İki Örneklem T Testi (AB Testi)
 İki grup ortalaması arasında karşılaştırma yapılmak istenildiğinde kullanılır.

Elimize gerçekte değerlerini bilmemişiz iki tane farklı analityk parametresi var. Bu iki parametrenin birbirinden farklılığını inceliyoruz.
 Elde ettiğimiz örnekler üzerinden yapacağımız karşılaştırma işlemidir.

$$H_0: \mu_1 = \mu_2 \quad H_0: \mu_1 \leq \mu_2 \quad H_0: \mu_1 \geq \mu_2 \\ H_1: \mu_1 \neq \mu_2 \quad H_1: \mu_1 > \mu_2 \quad H_1: \mu_1 < \mu_2$$

+ Örnek sayıları aynı, varyanslar homojen ise:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p \sqrt{\frac{2}{n}}} \quad S_p = \sqrt{\frac{s^2_{x_1} + s^2_{x_2}}{2}}$$

+ Örnek sayısı farklı, varyanslar homojen ise:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad S_p = \sqrt{\frac{(n_1-1)s^2_{x_1} + (n_2-1)s^2_{x_2}}{n_1 + n_2 - 2}}$$

+ Örnek sayıları farklı, varyanslar homojen değil ise:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{\Delta}}} \quad S_{\bar{\Delta}} = \sqrt{\frac{s^2_{x_1}}{n_1} + \frac{s^2_{x_2}}{n_2}}$$

Son 2 hesaplamada örnek sayıları eşit olsa da hesaplama yapılabilir.
 Varsayımlarımız:

- Normallik

- Varyans Homojentliği

Nonparametrik Bağımsız İki Örneklem Testi

Varsayımlarımızın ikisi de sağlanmazsa kullanılır.

- Bağımlı İki Örneklem T Testi
Bağımlı İki grup ortalaması arasında karşılaştırma yapmak istenildiğinde kullanılır.

$$H_0: \mu_o = \mu_s \quad H_0: \mu_o \leq \mu_s \quad H_0: \mu_o \geq \mu_s \quad ; \quad o = \text{öncesi} \\ H_1: \mu_o \neq \mu_s \quad H_1: \mu_o > \mu_s \quad H_1: \mu_o < \mu_s \quad ; \quad s = \text{sonrası}$$

Test İstatistiği

$$t = \frac{\bar{x}_D - \mu_0}{\frac{s_D}{\sqrt{n}}} \quad D = \text{fark (sonrası - öncesi)}$$

Varsayımlar = Normallik ve Varyans Homojenliği

- İki Örneklem Oran Testi

İki oran arasında karşılaştırma yapmak için kullanılır.

$$H_0: P_1 = P_2 \quad H_0: P_1 \leq P_2 \quad H_0: P_1 \geq P_2 \\ H_1: P_1 \neq P_2 \quad H_1: P_1 > P_2 \quad H_1: P_1 < P_2$$

Test İstatistiği

$$z_h = \frac{(P_1 - P_2)}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Varsayımlar

$$\rightarrow n_1 > 30 \\ \rightarrow n_2 > 30$$

■ Varyans Analizi:

İkiden fazla grup olduğunda kullanacağımız doğrudan hipotez testi yaklaşımıdır.

Varyans analizi geniş bir konudur. Burada veri bilimi için bilgi verecektir.

İki ya da daha fazla grup ortalaması arasında istatistiksel olarak anlamlı farklılık olup olmadığı öğrenilmek istenildiğinde kullanılır.

$$H_0: \mu_1 = \mu_2 = \mu_3 \\ H_1: \text{Eşit değillerdir (en az birisi farklıdır)}$$

Test İstatistiği

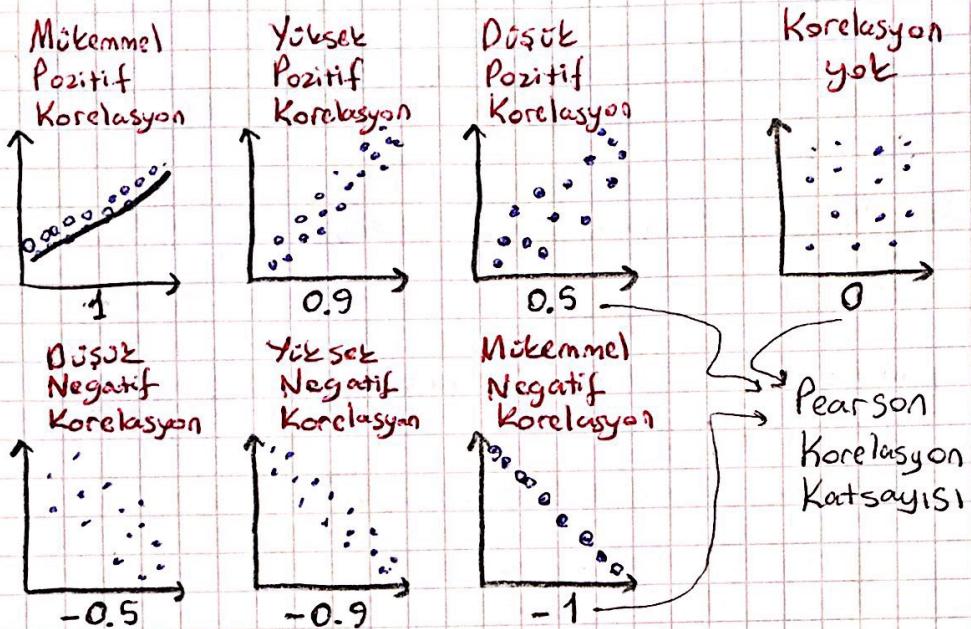
$$F_S = \frac{MS(\text{between})}{MS(\text{within})} \quad \begin{array}{l} (\text{Gruplar arası ort. hata}) \\ (\text{Grup içi ort. hata}) \end{array}$$

Varsayımlar

- Gözlemlerin birbirinden bağımsız olması (grupların)
- Normal dağılım
- Varyans homojenliği (mutlaka sağlanmalı)

Korelasyon Analizi (Korelasyon = İlişki)

Degişkenler arasındaki ilişki, bu ilişkinin yönü ve şiddeti ile ilgili bilgiler sağlayan istatistiksel bir yöntemdir.



Hipotezler (Korelasyonun Anlamılığının Testi)

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

Test istatistiği

$$r_{xy} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sqrt{(\sum x_i^2 - n \bar{x}^2)} \sqrt{(\sum y_i^2 - n \bar{y}^2)}}$$

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

x_i, y_i = gözlem birimleri
 r_{xy} = Korelasyon Katsayıısı
 n = gözlem sayısı

$t = t$ test istatistiği

Varsayımlar

- İki değişken içinde normalilik varsayımlı
- Varsayımlı sağlanıysa Pearson Korelasyon Katsayıısı
- Varsayımlı sağlanmıyorsa Spearman Korelasyon Katsayıısı