COĞRAFYADA Olasılık ve İstatistik

Ders Notları

Doç. Dr. Hasan

ÇOMÜ, Fef, Coğrafya Bölümü, Çanakkale

e-posta:tatli@comu.edu.tr

Giriş

- Doğa bilimleri ve/veya sosyal olaylarda karşılaştığımız problemlerin birçoğunda, gözlemlenen veya ölçülen değişkenlerin değerleri bilindiğinde probleme kesin ve tek çözüm bulunabilir. Örneğin, bir noktasal katı cismin kütle merkezine etki eden kuvvet ve kütlesi bilindiğinde kolayca cismin ivmesini hesaplayabiliriz.

$$F = m\ a \Rightarrow a = F/m$$

Oysa, bazı durumlarda sonucu kesin olarak bilmek mümkün olmayabilir. Örneğin, bir ay sonra bir bölgede yağış olup olamayacağı, olacaksa ne kadar miktarda yağış olacağını kesin bilmek mümkün değildir.

→ Sonucu kesin olarak bilinmeyen olayları belirsizlik içeren olaylar şeklinde, geniş anlamda tanımlayabiliriz.

> Belirsizlik içeren olaylar; olasılık teorisi ve istatistik bilimi yardımı ile haklarında bir yargıya varmamıza imkan verirler.

2-TEMEL YAKLAŞIM

1. **Sonucunu kesin hesaplayabildiğimiz olay (veya süreç) DETERMİNİSTİK**

2. **Sonucunu belli bir olasılık (belirsizlik) altında hesaplayabildiğimiz olay STOKASTİK**

Temel Yaklaşım

Olay

Deterministik

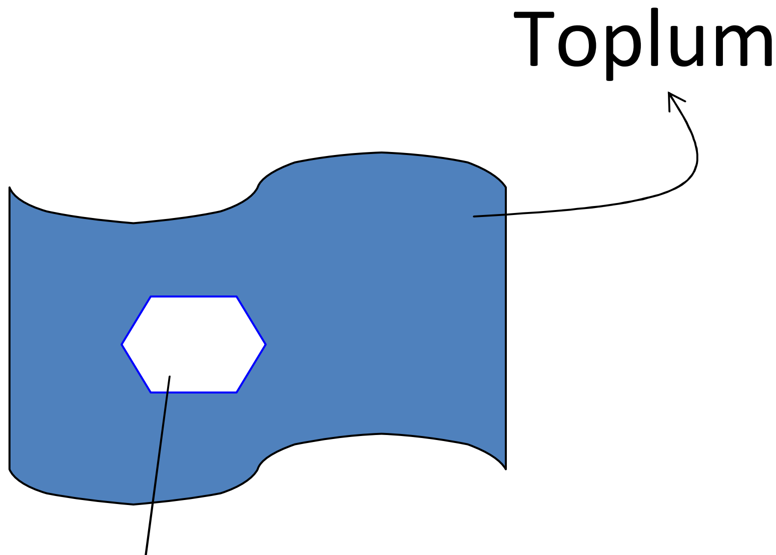
Stokastik

Lineer Non-lineer Kaotik

İstatistik Kaotik Rassal

BU DERSİN KONUSU

2. VERİLERİN SINIFLANDIRILMASI



Örnek

- R: Dağılım genişliği (ingilizcesi. Range)
- S: Sınıf sayısı
- C: Sınıf aralığı
- As: Sınıfın alt sınırı
- Üs: Sınıfın üst sınırı

Verilerin Sınıflandırılması (devamı)

R = En büyük değer – En küçük değer

S = 1+3.3 log (n): Sınıf sayısı

n: verisi sayısı

C = R/S : Sınıf aralığı

AS_1= En küçük değer	$ÜS_1=AS2-B$
$AS_2= AS_1 +C$	$ÜS_2=ÜS_1+C$
$AS_3 =AS_2+C$	$ÜS_3=ÜS_2+C$
....	...
$AS_m= AS_{m-1}+C$	$ÜS_m=ÜS_{m-1}+C$

B= 1 Değerlerin hepsi tam sayı ise

B= 0.1 Değerlerden en az birinin virgülden sonra tek basamaklı ondalıklı bir sayı olması

B = 0.01 Değerlerden en az birinin virgülden sonra iki basamaklı ondalıklı bir sayı olması

B = $1/10^k$ Değerlerden en az birinin virgülden sonra k basamaklı ondalıklı bir sayı olması

Verilerin Sınıflandırılması (Devamı)

Sınıf aralığı belirlenirken, değerlerin hepsi tam sayı ise == > C de tam sayıdır.

Aksi durumda, örneğin değerlerin en az bir tanesi ondalık sayı ise == > C de aynı türden ondalık sayı olmalıdır. Bu işlemlerden sonra verilerin sınıflandırılmasına geçilir.

Frekans Dağılım Tabloları

Örnek 1: Aşağıdaki tabloda verilen 60 adet veriyi, sınıf sayısı 10 olacak şekilde sınıflayınız

8	9	9.6	10.1	10.4	10.7	11.2	11.5	11.9	12.8
8.2	9.1	9.7	10.1	10.5	10.8	11.3	11.6	12.1	12.8
8.4	9.2	9.9	10.3	10.6	10.8	11.3	11.6	12.2	12.9
8.6	9.3	9.9	10.3	10.6	10.9	11.4	11.6	12.3	13.4
8.7	9.3	10	10.4	10.7	11	11.4	11.7	12.3	13.6
8.8	9.4	10.1	10.4	10.7	11	11.4	11.8	12.4	13.9

Örnek 1 (devamı)

n = 60 (veri sayısı)

S = 10 (sınıf sayısı)

max = 13.9 (en büyük değer)

min = 8 (en küçük değer)

R = max – min = 13.9 – 8 = 5.9 (değişim aralığı)

C = R /S = 5.9 /10 = 0.59 (sınıf aralığı)

$$C \approx 0.6$$

Çünkü veriler virgülden sonra tek basamaklı olduğundan sınıf aralığı da tek basamaklı olmalıdır. Bu nedenle C = 0.59 yerine C = 0.6 olarak, benzer şekilde B = 0.1 seçilmesi gerekir. Bu durumda sınıflara ilişkin alt ve üst sınırlar,

$As_1 = 8$	$Üs_1 = As_2 - B = 8.6 - 0.1 = 8.5$
$As_2 = 8 + 0.6 = 8.6$	$Üs_2 = Üs_1 + C = 8.5 + 0.6 = 9.1$
$As_3 = 8.6 + 0.6 = 9.2$	$Üs_3 = 9.1 + 0.6 = 9.7$
...	...
$As_{10} = 12.8 + 0.6 = 9.2$	$Üs_{10} = 9.1 + 0.6 = 9.7$

Örnek 1 (devamı)

Biraz önce hesaplanan sınıflara ilişkin alt ve üst sınıflar dikkate alınarak frekans dağılım tablosu elde edilir.

Sınıf No	**As**	**Üs**	**f**
1	8	8.5	3
2	8.6	9.1	5
3	9.2	9.7	6
4	9.8	10.3	8
5	10.4	10.9	12
6	11	11.5	9
7	11.6	12.1	7
8	12.2	12.7	4
9	12.8	13.3	3
10	13.4	13.9	3

Örnek 1 (devamı)

Biraz önce elde edilen frekans dağılım tablosundan, dağılımı tanımlayıcı ölçütler hesaplanırken SINIF ORTA NOKTASI (SON) ve SINIF ARA DEĞER (d) aşağıdaki gibi bulunur.

SON = (i. sınıf alt sınırı + aynı sınıfın üst sınırı)/2

d_i = (i-1. sınıf üst sınırı + i. sınıfın alt sınırı)/2

3. DAĞILIMI BETİMLEYİCİ ÖLÇÜTLER (Descriptive measures)

Bir dağılıma ilişkin gözlemler genelde ait oldukları toplumun özelliklerini göstermeleri beklenir. Bu özelliklerden yararlanarak dağılıma ilişkin betimleyici bilgiler elde edilir. Bu ölçütler 2 gruba ayrılabilir:

- Yer gösteren ölçütler
- Yaygınlık gösteren ölçütler

3.1 Yer Gösteren Ölçütler

Bu ölçütler gözlemlerin aldıkları değerlere göre dağılım içindeki konumlarını belirlemede ve dağılıma bağlı miktarsal bilgi edinmede kullanılırlar. Bu ölçütler kendi içinde 2 gruba ayrılırlar.

1. Merkezi ölçütler (beklenen değerler, Medyan, Mod)
2. Merkezi olmayan ölçütler (çeyrekler ve yüzdeler)

Merkezi ölçütler

ARİTMETİK ORTALAMA: Merkezi ölçütlerden en sık kullanılanı ve bir dağılımın diğer bir dağılımla karşılaştırılmasını sağladığı gibi; bir çok sayısal değerleri de özetler. Olasılığı az olan (tesadüfi, rassal: random) nedenlerden az etkilenir. Ancak, uç değerlerden çok etkilenir.

$$\overline{X} = \frac{\sum_{i}^{n} x_i}{n} = \frac{\text{Gözlemlerin toplamı}}{\text{Gözlem sayısı}}$$

Not: Ortalamadan büyük değerlerin toplamı, ortalamdan küçük değerlerin toplamına eşittir. Yani seriyi 2 eşit parçaya ayırır.

Aritmetik Ortalama (devamı)

Eğer veriler sınıflandırılmışsa, bu durumda ortalama, frekans dağılım tablosu kullanılarak bulunur.

$$\bar{x} = \frac{\sum_{i}^{n} f_i \cdot SON_i}{n}$$

SON_i: i. sınıf orta noktası

f_i : i. Sınıf frekansı

Örnek 2:

Bir iş yerinde çalışan 382 kişilik çalışanın bir yılda hastalık nedeniyle almış oldukları sağlık raporları sayısına göre frekans dağılım tablosu yanda verilmiştir. Ortalama rapor sayısını bulunuz.

Rapor sayısı	**Çalışan kişi (f)**	**$f_i \cdot x_i$**
0	92	0
1	105	105
2	67	134
3	43	129
4	32	128
5	14	70
6	12	72
7	9	63
8	5	40
9	3	27
Toplam:	n = 382	**768**

$$\overline{X} = \frac{768}{382} = 2\,\text{rapor}$$

Ağırlıklı ortalama

Ortalama hesaplamaları frekanslar üzerinde yapıldığı gibi hızların yüzdelerin de ortalaması alınabilir. Bu durumda hızların toplamları alınıp hız sayısına bölerek ortalama bulunduğunda yanlışlık yapılmış olur. Bu gibi durumlarda ağırlıklı (tartılı) ortalama almak gerekir.

Örnek 3: Öğrencilerin Matematik, İstatistik ve Bilgisayar dersindekli başarı yüzdeleri aşağıda verilmiştir. Buna göre öğrenci başarı ortalamasını hesaplayınız.

Örnek 3 (devamı)

Ders ismi	Öğrenci sayısı (f)	Başarı yüzdesi (%)
Matematik	74	56.7
İstatistik	61	24.6
Bilgisayar	75	28
TOPLAM	210	**37.1**

Eğer üç dersin aritmetik
Ortalaması alınsa = % 36.4
Bulunur ki bu sonuç yanlış
olur. Oysa, ağırlıklı ortalaması
Alınırsa:

$$P = \frac{74 \cdot 0.567 + 61 \cdot 0.246 + 75 \cdot 0.28}{210} = 0.371 = \%37.1$$

Geometrik Ortalama

Geometrik ortalamadada ortaloamalara giren bütün değerlerin katılmasıyla elde edilir. Daima aritmetik ortalamadan küçüktür. Veriler, bir geometrik diziyi takip ediyorsa gerçek ortalama, geometrik ortalama ile bulunması gerekir. Aritmetik ortalamaya göre, uç değerlerden daha az etkilenir.

$$\overline{X}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

Her zaman n'inci kuvvetten kök almak kolay olamayacağından, aynı işlem verilerin logaritmalarının aritmek ortalaması alınarak yapılır.

$$\ln \overline{X}_G = \frac{\ln x_1 + \ln x_2 + \ldots + \ln x_n}{n} \Rightarrow \overline{X}_G = \exp(\ln \overline{X}_G)$$

Harmonik ortalama

Bu ortalama, ortalamaya giren verilerin evrik değerlerinin (Reciprocal values) ortalamasının evrik değeridir.

$$\frac{1}{\overline{X}_H} = \frac{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}{n}$$

$$\Rightarrow \overline{X}_H = \frac{1}{\frac{1}{n}\sum_{i}^{n}\frac{1}{x_i}}$$

Örnek 4:

Bir sürücü bir şehirden diğerine saatte ortalama 75 km hızla gitmiş olsun. Dönüşte aynı yolu kötü hava koşulları nedeniyle saatte ortalama 60 km hızla aldığını varsayalım. Bu sürücünün gidiş-dönüş bütün yol boyunca ortalama hızı kaç km/sa olur?

Çözüm: Dikkat etmemiz gereken gidiş ve dönüş yolunun aynı uzunlukta olduğudur. Ayrıca, hızın yol/zaman şeklinde bir oran olduğunu düşünürsek çözüm için harmonik ortalama kullanılabilir mi?

Yolun uzunluğuna L km diyelim. Gidiş-dönüş yol uzunluğu 2L km olacaktır. Yolda geçen süreler gidişte L/75 ve dönüşte L/60 saattir. Öyleyse, ortalama hız:

$$\frac{2L}{\frac{L}{75}+\frac{L}{60}}=\frac{2}{\frac{1}{75}+\frac{1}{60}}=66.6\,km/sa$$

Ortalamaların genel formülü

$$\overline{X}_r = \left[\frac{\sum_{i}^{n} x_i^r}{n} \right]^{1/r}$$

X_i: gözlem değerleri
n: örneklemin gözlem sayısı
$r : -\infty \text{ ile } +\infty \text{ arasında bir tamsayı}$

r = -1 : Ters (harmonik) ortalama
r = 0 : Geometrik ortalama
r = 1: Aritmetik ortalama
r = 2: Kareli ortalama

Eğer r $-\infty$ yakın bir değer alırsa, bu durumda ortalamanın değeri gözlem değerleri arasındaki en küçük gözlem değerine yaklaşır. Bunun tersine r + sonsuza yakın bir değer alırsa, bu durumda ortalama da gözlem değerinin en büyüğüne yaklaşır. Ortalamanın gözlem değerlerinin en küçük ve en büyüğü arasında bir değer alması zorunluluğu burada da görülmektedir.

Örnek:

Ortalamaların genel formülünde r = 0 için geometrik ortalama elde edilir. Formülde r yerine 0 konulduğunda $\overline{X}_0 = 1^{\infty}$ elde edilir. Bu durumdan kurtulmak için r'ye doğrudan sıfır değeri vermek yerine r -> 0 için limit değerini kullanmak gerekir. Her iki tarafın logaritması alınırsa;

$$Ln\overline{X}_r = \left[\ln\left(\sum x_i^r\right) - \ln n\right] / r$$

$$r = 0 \text{ için } \ln\overline{\mathrm{X}}_0 = (\ln n - \ln n)/0 = \frac{0}{0}$$

Karşılaşılan bu belirsizlik durumunda Hospital kuralı uygulanır ve r -> 0 için limit alınırsa

$$Ln\overline{X}_r = \sum x_i^r \ln x_i / \sum x_i^r$$

$$\ln\overline{\mathrm{X}}_0 = \left(\sum \ln x_i\right)/n$$

Bu eşitliğin antilogaritması alınırsa

$$\overline{\mathrm{X}}_0 = \left(\prod_i^n x_i\right)^{1/n} = \sqrt[n]{\prod_i^n x_i}$$

Bütün Gözlem Değerlerinden Etkilenmeyen Ortalamalar

Burada göreceğimiz ortalamalar, veri kümesindeki gözlem değerlerinin bazıları önemli ölçüde değişse bile bundan etkilenmeyebilir. Bu özellikleri nedeniyle aykırı (uç) değerlerin ortalamayı saptırmasını engelleyebilir. Bu tür ortalamaların en çok kullanıldıkları aşama, verileri ilk incelemeye aldığımız sorgulayıcı aşamadır. Bu çözümleme aşamasında veriler, bir sonraki aşama olan doğrulayıcı çözümleme aşamasında daha ayrıntılı incelemek, bazı ip uçlarını bulunması, araştırmaya değer bazı özelliklerin ortaya çıkartılması, çabuk ama kaba bilgiler elde edilmesi v.b. Amaçlara yönelik olarak hazırlanır.

Bu bölümde, bu tür ortalamalardan uygulamada sık kullanılan MOD, MEDYAN ve daha az kullanılan ama özellikle 1970'lerin ikinci yarısından bu yana önem kazanan dördebilenler (çeyrekler veya kartiller) ve Yüzdeler (sentiller) ve kırpık ortalama (trimmed mean) konusu işlenecektir.

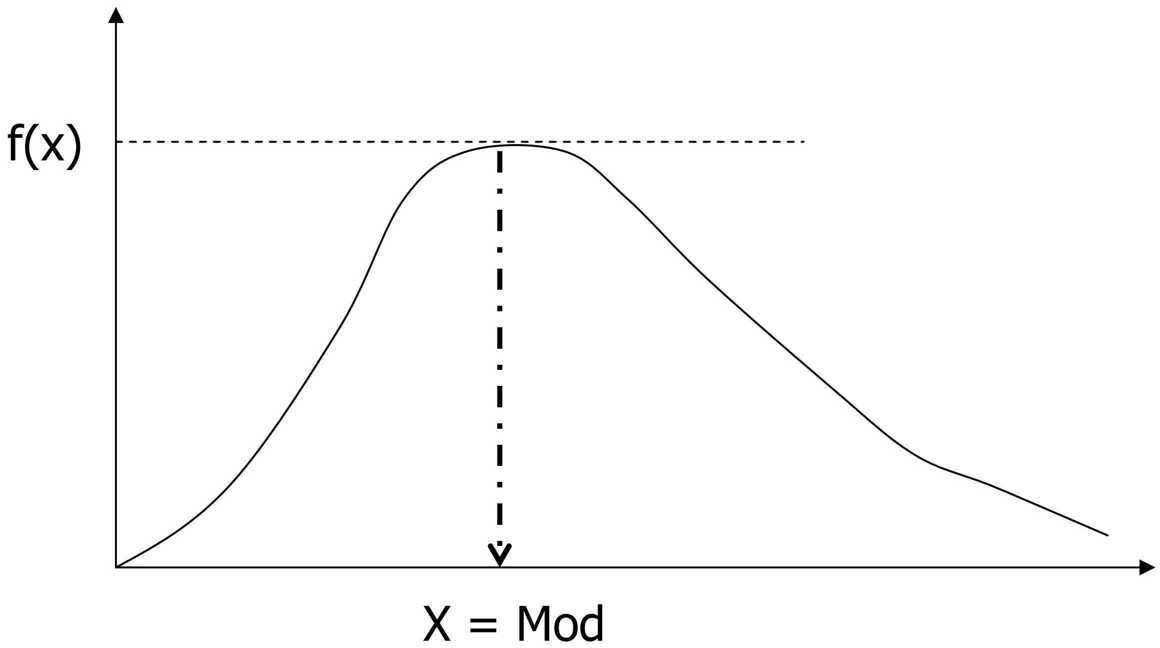
MOD (TEPE DEĞERİ)

Verilerin en yoğun olarak hangi gözlem değeri düzeyinde toplandığını gösterir. Mod aykırı değerlerden pek etkilenmez.

Mod'u diğer ortalamalardan ayıran önemlli özelliği, gözlem değerlerinin sayısal bir değere ilişkin olma zorunluluğunun bulunmamasıdır. Diğer taraftan mod yalnız tek tepeli dağılımlar için elverişlidir. Eğer bir dağılımda birçok tepe varsa, diğer bir deyişle, gözlem değerleri birden çok yerde yoğunlaşırsa mod analizi anlamlı olmayacaktır.

Mod sürekli bir rastgele değişkenin olasılık yoğunluk fonksiyonun maksimumdan geçtiği noktanın absisidir.

MOD (Devamı)



$$\left.\frac{df(x)}{dx}\right|_{x=\mathrm{mod}\,x} = 0$$

Örnek: Üç ayrı yere ait sisli günlerin aylık sayıları aşağıdaki Tablo'da verilmiştir. Bu çizelgeden yararlanarak her bir yerin en sık sisli gün sayısını bulunuz.

	O	Ş	M	N	M	H	T	A	E	E	K	A
A	4	12	6	10	14	4	10	12	20	6	14	20
B	4	8	14	6	8	6	8	6	14	8	18	4
C	4	10	6	4	6	20	6	12	16	20	12	4

Örnek (devamı)

Çözüm:

Tablodan A yeri için her bir sisli gün sayısını ikişer kere ortaya çıktığı görülür. Her bir sisli gün sayısının diğerine üstünlüğü olmadığından, A yeri için Mod'un olmadığı anlaşılır.

B yeri için en fazla tekrarlanan sisli gün sayısı 8 olduğundan modu 8'e eşittir.

C yeri sisli günlerine göz atıldığında 4 ve 12 sisli günlerin 3'er kez ortaya çıktığı görülmektedir. Dolayısıyla, bu yerde bir değil iki tane mod değerinin olduğunu söyleyebiliriz.

MEDYAN (ORTANCA DEĞER)

Gözlem değerlerinin hepsinden etkilenmeyen ortalamalardan biri de medyandır. Medyan, gözlem değerleri büyüklük sırasındayken tam ortaya düşen değerdir. Herhangi bir gözlem değerinin medyandan büyük ya da küçük kalması olayının olasılığı 0.5'dir. Veri kümesinin eleman sayısının tek veya çift olması durumunda medyan hesabı:

$$q_{0.5} = \begin{cases} x_{(n+1)/2} & \text{n tek sayı ise} \\ \dfrac{\mathrm{x}_{n+1} + x_{(n/2)+1}}{2} & \text{n çift sayı ise} \end{cases}$$

Frekans Dağılım Tablolarından Medyanın Hesabı

Medyan (M) sınıflandırılmış verilerde aşağıdaki şekilde hesaplanır.

$$M = D + \frac{\frac{n}{2} - F_1}{f} \times S.A.$$

D: Medyanın bulunduğu sınıfın başlangıç noktası
F_1: Medyanın bulunduğu sınıfa kadar olan Kümülatif Frekans
S.A.:Sınıf Aralığı

Örnek

Yanda verilen çizelgede 382 kişinin bir yılda aldıkları rapor sayılarının Medyanını bulunuz.

$$M = 1 + \frac{\frac{382}{2} - 92}{105} \times 1 = 1 + \frac{99}{105} \times 1 = 1.94$$

Rapor sayısı (x)	Çalışan kişi (f)	Kümülatif frekans (F)
0	92	92
1	105	197
2	67	264
3	43	307
4	32	339
5	14	353
6	12	365
7	9	374
8	5	379
9	3	**382**
TOPLAM	**382**	

BÖLENLER

Bir veri kümesini eşit sayıda parçalara ayıran değerlerin genel adı bölenlerdir. Kantiller de denen bu değerlerden biri medyanı gösterir. Bilindiği gibi medyan gözlem sayılarını 2 eşit parçaya ayırır. Diğer bölenler de aynı ilkeye göre bulunur.

Dağılımı 4 eşit parçaya bölen değerlere dördebölen (kartil veya çeyrek), 10 eşit parçaya bölenlere desil ve 100 eşit parçaya bölenlere de yüzebölen (persantil) denir.

İstatistikte çok az kullanılmalarına karşın, eğer elde veri kümesinin küçük olması, rastgele değişkenin dağılımının çarpık olması ve örnekte aykırı değerlerin bulunması halinde bu tipten parametrelerin değerleri büyük değişmeler gösterebilir. Bu ve benzeri durumlarda bölenlerin kullanılması yerinde olur.

Dördebölenler (Kartiller)

q_1 (Alt kartil veya birinci çeyrek): Küçükten büyüğe sıralanmış veya gruplandırılmış verilerde ¼ cü gözlemin değeridir.

q_2 (İkinci kartil veya medyan): Verilerde ½ cü gözlemin değeridir.

q_3 (Üst kartil veya üçüncü çeyrek): Verilerde ¾ cü gözlemin değeridir.

Görüldüğü üzere q_2 medyan olduğundan, kartil denince q_1 ve q_3 anlaşılır.

Sınıflandırılmış Verilerde Kartillerin Hesabı

Kartiller, sınıflandırılmamış verilerde, veriler küçükten büyüğe doğru sıralandığındaaşağıdaki hesaplanır.

$$q_1 = \frac{n+1}{4} \qquad q_3 = \frac{3(n+1)}{4}$$

Kartiller, sınıflandırılmış verilerden (frekans dağılım tablolarından) aşağıdaki formüller ile hesaplanır.

$$q_1 = L_1 + \frac{n/4 - F_1}{f_1} \times S.A. \qquad \mathrm{q}_3 = L_3 + \frac{3(n/4) - F_3}{f_3} \times S.A.$$

L_1: Alt kartilin bulunduğu sınıfın başlangıç noktası
n/4: Alt kartil olarak bulunacak değer
F_1: Alt kartil sınıfına kadar olan kümülatif frekans
f1: Alt kartil sınıfının frekansı

Kartiller (devamı)

Üst kartil ile alt kartil arasındaki farka (q_3-q_1) kartiller (çeyrekler) arası genişlik denir ve gözlemlerin yarısı bu iki değer arasında toplanmıştır.

Genellikle bir dağılımın yayılmasını ölçmek için Kartiller arası farkın yarısı alınarak kullanılır. Bu değere Kartiller arası yarı genişlik (semi inter quartil)denir; q harfi ile gösterilir.

$$q = \frac{q_3 - q_1}{2}$$

Kartiller ile Medyan arasındaki farklardan yararlanarak bir dağılımın simetrik olup olmadığı ölçülebilir.

Kartiller (devamı)

Medyan – Alt Kartil = a = $q_2 - q_1$
Üst Kartil – Medyan = b = $q_3 - q_2$

Bu fark ne kadar biribirine yakınsa dağılımın simetrik olduğuna güvenle karar verilebilir. $a = b$ ise dağılım simetriktir denir.

$a > b$ ise dağılım soldan çarpıktır,

$a < b$ ise dağılım sağdan çarpıktır denir.

Örnek

Aşağıdaki tabloda, **A** ve **B** gibi iki farklı bölgenin yıl boyunca yağışlı günlerini gösteren frekans dağılımları verilmiştir. Bu tablodan yararlanarak, Alt-Kartil (q_1), Üst-Kartil (q_3) ve Kartiller arası yarı genişliği (q) bulunuz.

	A - Bölgesi		**B-Bölgesi**	
Yağış (mm)	Yağışlı gün sayısı Frekans (f)	F	Yağışlı gün sayısı Frekans (f)	F
2 <	80	80	150	150
2- 50	50	130	90	240
51 -100	180	310	70	310
101 -150	25	335	30	340
151-200	25	360	10	350
200 >	5	365	15	365
TOPLAM	365		365	

Örnek (devamı)

A- bölgesinde q_1 = 365/4 = 91.25'inci değerdir.

B –bölgesinde de veri sayı eşit olduğundan aynı çıkar.

A-Bölgesinde, kümülatif frekansın(F) 91.25'inci değeri 2-50 sınıfında ve 130 frekanstan biri olacaktır. Formülde değerler yerine konulduğunda;

$$q_1 = 2 + \frac{365/4 - 80}{50} \times 50 = 13.25$$

$$q_3 = 51 + \frac{3(365/4) - 130}{180} \times 50 = 90.93$$

$$q = \frac{q_3 - q_1}{2} = \frac{90.93 - 13.25}{2} = 38.8$$

Örnek (devamı)

B- bölgesindeki q_1 = 365/4 = 91.25'inci değerdir, çünkü veri sayıları eşittir.

B-Bölgesinde, kümülatif frekansın(F) 91.25'inci değeri 2< sınıfında ve 150 frekanstan biri olacaktır. Formülde değerler yerine konulduğunda;

$$q_1 = 0 + \frac{365\ /\ 4 - 0}{150} \times 2 = 1.22$$

$$q_3 = 51 + \frac{3(365/4) - 240}{70} \times 50 = 75.1$$

$$q = \frac{q_3 - q_1}{2} = \frac{75.1 - 1.22}{2} = 36.9$$

Persantiller (Yüzdeler)

Kartiller veriyi 4 eşit parçayı ayırırken; persantiller, gözlemleri %10, %20, %25, %30, %40, %50, %60, %70, %75, %80, %90 gibi yüzde bölümlerine ayırır. Ve genellikle P (%n) şeklinde gösterilir. Örneğin;

$$P(10) = L_{10} + \frac{10n/100 - F_{10}}{f_10} \times S.A.$$

Bu formülde:
10n/100: %10'luk persantil değeri
L10: %10'luk persantil değerinin bulunduğu sınıfın başlangıç noktası
F10: %10'luk persantil sınıfına kadar olan kümülatif frekans
F10: %10'luk persantil sınıfının frekansı
S.A.: Sınıf aralığı

Persantiller arası genişlik kartillerdeki gibi bulunur.

4. YAYIKLIK, ÇARPIKLIK, BASIKLIK ÖLÇÜLERİ

Bundan önceki bölümde verilerin bazı özelliklerinin incelemesi gerektiğinden bahsedildi. Bu özelliklerden biri olan ortalama düzeyin nasıl özetlenebileceği üzerinde duruldu. Bu ortalamalardan uygun birini ya da bir kaçını kullanarak elimizdeki veri kümesinin hangi düzeyde olduğunu kısa, öz bir biçimde belirtilmesinin yanısıra birden çok veri kümesinin ortalama düzeylerinin birbirine göre konumlarını da karşılaştırabildik. Ancak, bir veri kümesinin yalnızca ortalama düzeyi bakımından özetlenmesi çoğu zaman yetersiz olur. Bunun yanında verinin ortalamdan ne kadar uzaklaşabildiği, bu uzaklaşmanın ortalamaya göre simetrik olup olmadığı, tepe değerleri, sivrilikleri ve basıklıklarını da bilmek gerekir.

(4. devamı)

Bu bölümde verilerin yayıklık (değişkenlik), çarpıklık ve basıklık gibi özelliklerinin nasıl ölçüleceği, ölçüm sonuçlarının özetlenebileceği konular ele alınacaktır.

4.1 YAYIKLIK (DEĞİŞKENLİK) ÖLÇÜLERİ

Bir veri kümesinin yayıklık ya da değişkenlik özelliği denilince anlaşılması gereken, gözlem değerlerinin sayı ekseni üzerinde ne kadar toplu ya da yayılmış biçimde bulunduklarıdır. Bu özelliğin ölçülmesi çeşitli yollarla yapılabilir. Aşağıda bu ölçülerden önemli görülenleri açıklanacaktır.

Aralık (Range): Bir veri kümesinin yayıklığının en kolay hesaplanabilen ölçüsü en büyük değerle en küçük değerin farkıdır. Bu fark aralık olarak adlandırılır. Aralığı, **R**, en büyük ve en küçük değerleri Xmax ve Xmin ile gösterirsek

$$R = X_{max} - X_{min}$$

Dördebölenler Aralığı (Interquartile Range, IQR) ve Dördebölenler Sapması (Quartile Deviation, QD):

IQR = $q_{0.75} - q_{0.25}$ = Üst kartil – Alt kartil

QD=IQR/2 = Kartiller Arası Yarı Açıklık

Ortalama Mutlak Sapma (Mean Absolute DEviation, MAD):

$$MAD = \frac{\sum_{i}^{n} \left| x_i - \overline{x} \right|}{n} \quad \text{veya}$$

$$MAD = Medyan \left| x_i - q_{0.5} \right|$$

Varyans ve Standart Sapma:

Toplum için→

$$\sigma^2 = \frac{\sum_{i}^{n}(x_i - \overline{x})^2}{n-1}$$

$$\sigma = \sqrt{\frac{\sum_{i}^{n}(x_i - \overline{x})^2}{n-1}}$$

Varyans ve Standart Sapma:

Örnek için→

$$s^2 = \frac{\sum_{i}^{n}(x_i - \overline{x})^2}{n-1}$$

$$s = \sqrt{\frac{\sum_{i}^{n}(x_i - \overline{x})^2}{n-1}}$$

Örnek: Aşağıdaki eşitliğin geçerli olduğunu gösteriniz.

$$\sigma^2 = \overline{X}_2^2 - \overline{X}_1^2 = \text{Kareli ortalama} - \text{Aritmetik ortalamanın karesi}$$

Örnek: a) Bir gözlem setindeki her bir eleman *c* gibi bir sabit sayı ile çarpılırsa,

b) Bir gözlem setindeki her bir elemana c sabit sayısı eklenirse standart sapma nasıl etkilenir?

Örnek : Aşağıdaki tabloda verilen 60 adet verinin

a) Varyans,

b) Standart sapmasını bulunuz.

8	-6	15	21	-5	10	10	18	10	5
9	-9	17	24	-15	11	37	0	0	6
11	5	9	45	-20	12	64	7	4	7
-5	4	2	56	-23	15	-47	3	21	8
9	3	0	8	43	16	-11	26	22	18
16	1	4	0	13	19	25	9	40	21

$$s^2 = \frac{\sum_{i}^{60} (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i}^{60} (x_i - 10.43)^2}{60-1} \approx 318$$

$$s \approx \sqrt{318} \approx 17.83$$

Örnek: Aşağıda frekans dağılım tablosu şeklinde verilen verinin varyans ve standart sapmasını bulunuz.

sınıf	As	Üs	f	D	fD
1	8	8.5	9	-4	-36
2	8.6	9.1	8	-3	-24
3	9.2	9.7	7	-2	-14
4	9.8	10.3	8	-1	-8
5	10.4	10.9	12	0	0
6	11	11.5	9	1	9
7	11.6	12.1	7	2	14
8	12.2	12.7	4	3	12
9	12.8	13.3	6	4	24
10	13.4	13.9	4	5	20
				Top:	-3

$$\overline{X} = Ç.O. + \overline{D} \cdot S.A.$$

$$\overline{X} = Ç.O. + \frac{\sum_{i}^{S} f_i \cdot D_i}{n}$$

$$s = \sqrt{\frac{\sum (f_i \cdot D_i{}^2) - \frac{\left(\sum f_i D_i\right)^2}{n}}{n-1}} \cdot S.A.$$

Örnek (devamı)

sınıf	As	Üs	f	D	f.D	f.D²
1	8	8.5	9	-4	-36	144
2	8.6	9.1	8	-3	-24	72
3	9.2	9.7	7	-2	-14	28
4	9.8	10.3	8	-1	-8	8
5	10.4	10.9	12	0	0	0
6	11	11.5	9	1	9	9
7	11.6	12.1	7	2	14	28
8	12.2	12.7	4	3	12	36
9	12.8	13.3	6	4	24	96
10	13.4	13.9	4	5	20	100
				Top:	-3	521

S.A. = 0.6

S=1.6 bulunur.

Değişme (Varyasyon) Katsayısı (Cv %) :

Frekans dağılım tablolarında ölçülerin birimleri verinin cinsine göre değişmektedir. Örneğin, boy için cm, hız için km/sa, sıcaklık için C^o gibi.

Böyle ölçü birimlerinin değişik olduğu dağılımları birbiriyle karşılaştırmada standart sapmalarda verilerin ölçü birimleri cinsinden olduğundan daima zorlukla karşılaşılır. Bunu önlemek için standart sapmanın ortalamaya oranın 100 ile çarpımı olan ve hiç bir ölçü birimi kullanılmayan değişme katsayısı kullanılır. Cv% olarak gösterilir.

$$Cv = \frac{s}{\bar{x}} \cdot 100$$

Örnek: Aşağıda verilen, ortalamaları farklı, fakat standart sapmaları farklı 2 örneğin varyasyon katsayılarını karşılaştırınız.

$$\bar{x}_1 = 250 \qquad s_1 = 5 \qquad \bar{x}_2 = 50 \qquad s_2 = 5$$

$$Cv_1 = \frac{5}{250} \cdot 100 = \%\,2$$

$$Cv_2 = \frac{5}{50} \cdot 100 = \%\,10$$

Varyasyon katsayısı ne kadar küçük ise yayılma o kadar az demektir. Yukarıdaki örnekte, birinci değişkenin yayılması, ikinci değişkenin değişmesine göre daha azdır.

Standart Hata ve Örnek Ortalamasının Toplumda Geçerliği

Standart Hata: Örnekte elde edilen standart sapmanın örnek hacmi *n*'nin kareköküne bölünmesiyle ile bulunan bir değerdir.

$$SH\overline{x} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$$

a)Örnekten bulunan bulunan ortalamanın toplum ortalamasına ne kadar yaklaştığının ölçüsüdür. Diğer bir deyişle, doğruluk ölçüsüdür. Genellikle örneklerden alınacak bilginin doğruluk derecesi:

Örnekleme yöntemlerinin hatasız uygulanmasına,

Örnek hacmi (*n*) büyüklüğüne,

Araştırılan değişkenin (özellik) yayılışına bağlıdır.

Standart sapma örnek hacmi ile doğrudan değil de örnek hacminin karekökü ile ters orantılıdır. Örneğin, bir örnek hacmi 9'dan 16 katı olan 144'e çıkartılırsa standart sapma 144/9=16 değil de Karekök(144/9)= 4 kat azalır.

Eğer toplumda çok sayıda *n* hacimli örnekler alınıp örneklerin ortalamalarının ortalaması alınırsa bu toplumun ortalaması μ'ye çok küçük bir hata ile yaklaşılır. Alınan örneklerin ortalamalarının toplum ortalamasından sapmalarının ortalama değeri, standart hata ile gösterilir.

$$\overline{x} \mp 1.96\, SH \overline{x}$$

Şeklinde gösterilir. Bu formül n hacimli örnekten elde edilen bir ortalama yardımı ile X-değişkenin toplum ortalaması μ'nün hangi sınırlar içinde bulunabileceğini tahmine yardım eder.

%95 olasılıkla μ'nün içinde bulunacağı değerler

$\overline{x} \mp 1.96\, SH\, \overline{x}$

%99 olasılıkla μ'nün içinde bulunacağı değerler

$\overline{x} \mp 2.56\, SH\, \overline{x}$ olur.

Momentler İle Çarpıklık Ve Sivrilik

Not yazdırıldı

5. OLASILIK


Bİ GERGEDANLA AYNI ISSIZ ADAYA DÜŞME OLASILIĞIN NEDİR?
TRİLYONDA BİR FİLANDIR HERHALDE...
PEKİ O GERGEDANIN MANYAK OLMA OLASILIĞI?..

Olasılığın Tanımları

- Klasik (A Priori) Olasılık
- Frekans (A Posteriori) Olasılığı
- Aksiyom Olasılığı

NOT:Bu sıralama olasılık teorisinin tarihsel gelişimini tanımlamaktadır.

Klasik Olasılık

Eğer bir örnek uzayı n(S) adet ayrık ve eşit olasılıkla ortaya çıkan basit olaylardan oluşuyor ve örnek uzayındaki basit olaylardan n(A) adedi A olayının özelliğine sahip ise A'nın olasılığı:

$$P(A) = n(A) / n(S)$$

kesri ile elde edilir

Klasik olasılık TÜMDENGELİME dayanan çıkarımlar yaparak olasılığı bulur.

Örnek: Bir kapta 5 sarı, 5 lacivert ve 5 adet yeşil bilye bulunmaktadır. Çekilen bir bilyenin sarı olma olasılığı nedir?

A: Çekilen bir bilyenin sarı olması

n(S): Örnek uzayı eleman sayısı = 15

n(A): Örnek uzayındaki A elemanı sayısı = 5

$$P(A) = \frac{n(A)}{n(S)} = \frac{5}{15} = \frac{1}{3}$$

Klasik Olasılık Niçin Yetersizdir

- Örnek uzayının eleman sayısı sonsuz olduğu durumlarda,
- Eşit olasılıklı olay varsayımı yapılamadığı durumlarda,
- Tümdengelim çıkarımları yapılamadığında klasik olasılık ile hesaplama yapılamayacağından dolayı yetersizdir.

Ne Yapılabilir?

- Araştırılan anakütle üzerinde tekrarlı deneyler gerçekleştirilerek sonuçlar analiz edilmek üzere kayıt edilmelidir.

Frekans Olasılığı (Bağıl Sıklık Kavramı)

Araştırılan anakütle üzerinde *n* adet deney uygulanır. Yapılan bu deneylerde ilgilenilen A olayı m(A) defa gözlenmiş ise A olayının bağıl frekansı (veya yaklaşık olasılığı):

P(A) = m(A) / n

olarak bulunur.

- P (olay) = m / n
- m = İstenen olayın oluşma sayısı
- n = Mümkün tüm olayların sayısı

Arızalı olma olasılığı = 2/100 bulunur.

Frekans Olasılığının Kararlılık Özelliği

Gerçekleştirilen deney sayısı arttıkça P(A) olasılık değerindeki değişkenlik azalacak ve giderek bir sabit değere yaklaşacaktır. Bu duruma ***kararlılık özelliği*** adı verilir.

Bir olayın olasılığı deneyin tekrarlama sayısı sonsuza yaklaşırken o olayın göreli frekansının alacağı limit değer olarak tanımlanır:

$$p = P(A) = \lim_{n \to \infty} m(A) / n$$

Frekans Olasılığı Niçin Yetersizdir?

Olasılığın kararlılık değerine ulaştığı deneme sayısı kaçtır?

1. Sonsuz adet deneme yapmak mümkün değildir.

2. Aynı deney iki defa aynı tekrar sayısı ile gerçekleştirildiğinde elde edilen olasılıklardan hangisi olayın gerçek olasılığı olarak kabul görecektir?

Aksiyom (Teorik) Olasılık Nedir?

1. Olasılığın matematiksel teorisini tanımlar.
2. Bu teorinin oluşturduğu **ideal modeller** yaşadığımız dünyanın problemlerini çözmede kullanılır.
3. Olasılığın iki genel tipinin sahip olduğu önemli ortak nokta: Her ikisinin de, benzer koşullarda (teorik olarak aynı koşullarda) uygulanan deneylere gereksinim duymasıdır.
4. Bununla birlikte benzer koşullarda tekrarlı olarak uygulanamayan durumlarda olasılıkların hesaplanmasında AKSİYOM OLASILIĞI yardımcı olur.

Benzer Koşullarda Tekrarlı Olarak Uygulanamayan Durumlara Örnekler:

- İlk aldığınızda İstatistik dersinden başarılı olma olasılığı?
- Önümüzdeki 1 yıl içinde İzmir'de en az 6 büyüklüğünde deprem olması olasılığı nedir?
- Fenerbahçe - Galatasaray maçının 6-0 bitmesi olasılığı nedir?

Aksiyomlar

- **Aksiyom 1:**
  - P(A) örnek uzayı S'deki her A olayı için $P(A) \geq 0$ olan bir gerçel sayıdır.
- **Aksiyom 2:**
  - $P(S)=1$ $\{ P(\varnothing)=0 \}$
- **Aksiyom 3:**
  - Eğer $S_1, S_2, \ldots$ Olaylarının her biri S'deki ayrık olaylar ise, diğer bir deyişle $S_i \cap S_j = \varnothing$ tüm $i \neq j$ için ise,

$$P(S_1 \cup S_2 \cup \ldots) = P(S_1) + P(S_2) + \ldots$$

Sadece Aksiyomlar Yeterli mi?

HAYIR

- Bu aksiyomların ve onlara bağlı teoremlerin faydalı bir model geliştirilmesinde bize yardımcı olabilmesi için, S örnek uzayındaki her bir A olayı için olasılığın hesaplanmasında kullanılacak bir FONKSİYONA ya da bir KURALA gereksinim vardır

- Bu fonksiyonlar İlgilenilen anakütlenin Tanımladığı ÖRNEK UZAYINA Göre Farklılık Gösterir.

Sık karşılaşılan üç farklı örnek uzayı;

- **Sonlu elemanlı kesikli örnek uzayı (sayılabilir sonlu)**
- **Genel kesikli örnek uzayı (sayılabilir sonsuz)**
- **Sürekli örnek uzayı (sayılamaz sonsuz)**

olarak ifade edilir.

- x : herhangi bir gün içinde yağmur yağması

  $x = 0$ (yağmur yağmaz)

  $x = 1$ (yağmur yağar)

Örnek Uzayı;

$S = \{ x / 0, 1 \}$

veya

$S = \{ x$ / Yağmursuz , Yağmurlu $\}$

olarak belirlenir ve sayılabilir sonlu bir örnek uzayıdır.

- x : bir zar için 6 gelinceye kadar yapılan atış sayısı

Örnek Uzayı;

S = { x / 1,2,3,……….. }

olarak belirlenir ve sayılabilir sonsuz bir örnek uzayıdır.
(kesikli şans değişkeni)

- x : öğrencilerin boyları

Örnek Uzayı;

S = { x / 150 < x < 200 }

olarak belirlenir ve sayılamaz sonsuz bir örnek uzayıdır.
(sürekli şans değişkeni)

Örnek Uzayı ve Olay Sayısını Belirleyen Sayma Yöntemleri

- Klasik olasılığın diğer bir ifade ile eşit olasılıklı olayların geçerli olduğu durumlarda:
  - Örnek uzayının eleman sayısı,
  - İlgilenilen olayın eleman sayısının

  belirlenmesi gereklidir.

Kullanılan iki temel prensip;

1) Toplama Yöntemi

2) Çarpma Yöntemi

Toplama Yöntemi

- Bir A olayı *m* farklı şekilde, başka bir B olayı da *n* farklı şekilde oluşabilen **ayrık olaylar** ise;

A veya B olayı $n + m$ farklı şekilde oluşabilir.

Örnek: İstanbul'dan İzmir'e 2 farklı tren seferi, 4 farklı havayolu firması, 40 farklı otobüs firması ve 1 adet denizyolu firması ile gidilebildiğine göre İstanbul'dan İzmir'e kaç farklı şekilde gidilir?

$$2 + 4 + 40 + 1 = 47$$

Çarpma Yöntemi

- Bir A olayı *m* farklı şekilde, başka bir B olayı da *n* farklı şekilde oluşabilen ve **aynı anda oluşmaları mümkün olaylar** ise;

  A ve B olayı *n* * *m* farklı şekilde oluşabilir.

Örnek: Bir iskambil destesinden çekilen iki kartın birinin Kupa diğerinin Maça olması kaç farklı şekilde gerçekleşebilir?

$13 * 13 = 169$

NOT: Çarpma yöntemi bağımsız olaylar için kullanılır.

k farklı sonuç veren bir deney r kez tekrar edilirse ortaya çıkan tüm durumların sayısı;

k^r olarak hesaplanır.

Örnek: Bir zarı 3 kez attığımızda ortaya çıkabilecek tüm mümkün durumların sayısı sayısı;

$6^3 = 216$ adettir.

- ***Örnek uzayının eleman sayısı 216'dır.***

Örnek Uzayı ve Olay Sayısının Büyük Olduğu Durumlar

Örnek uzayı ve olay sayısının büyük olduğu durumlarda kullanılan sayma yöntemleri;

- **Permütasyon**
- **Kombinasyon**

Permütasyon

- Sıraya konulacak *n* adet nesne olsun ve her biri sadece bir kez kullanılmak üzere kaç farklı sıralama yapılabilir?

n	n-1	n-2	....	2	1

n nesnenin mümkün sıralamalarının sayısı:

$n(n\text{-}1)(n\text{-}2)\ldots(2)(1)=n!$ $\quad\quad {}_nP_n = n!$

- n tane nesne arasından seçilmiş x tane nesnenin permütasyon sayısı $_{n}P_{x}$ olarak ifade edilir.
- Toplam n tane nesne arasından x tane nesne seçilir ve bunlar sıraya konulursa ortaya çıkabilecek sıralamaların sayısıdır ve şu şekilde hesaplanır:

$$_{n}P_{x} = \frac{n!}{(n-x)!}$$

Kullanıldığı durumlar

- İadesiz örnekleme
- Örneğe çıkış sırası önemli

Örnek: 8 atletin katıldığı 100 metre yarışmasında ilk üç dereceye girenler kaç farklı şekilde belirlenir ?

$$_{8}P_{3} = \frac{8!}{(8-3)!} = 8*7*6 = 336$$

Örnek: 2,3,5,6,7 ve 9 sayılarını kullanarak 4 basamaklı rakamları birbirinden farklı kaç sayı oluşturulur?

$$_{6}P_{4} = \frac{6!}{(6-4)!} = 6*5*4*3 = 360$$

6	5	4	3

=360

Kombinasyon

- n adet nesne arasından seçilen x tanesinin kombinasyon sayısı $_nC_x$ ile gösterilir. Sıralama önemli olmaksızın tüm durumların sayısı olarak ifade edilir. Bu sayı şu şekilde hesaplanır:

$$_nC_x = \frac{n!}{(n-x)!x!}$$

- Kullanıldığı durumlar;
  - İadesiz örnekleme
  - Örneğe çıkış sırası önemsiz

Örnek: Beş kişilik bir topluluktan üç kişilik bir komisyon kaç farklı şekilde seçilir ?

$$ {}_5C_3 = \frac{5!}{(5-3)!3!} = \frac{5*4*3*2}{2*3*2} = 10 $$

Örnek: 10 bay ve 5 bayan arasından 2 bay ve 1 bayan üye içeren bir kurul kaç farklı şekilde oluşturulur?

$$ {}_{10}C_2 = \frac{10!}{(10-2)!2!} = \frac{10*9}{2} = 45 $$ (10 bay arasından 2 bay)

$$ {}_5C_1 = \frac{5!}{(5-1)!1!} = 5 $$ (5 bayan arasından 1 bayan)

Çarpım kuralı uygulanarak 45 * 5 =225 farklı şekilde

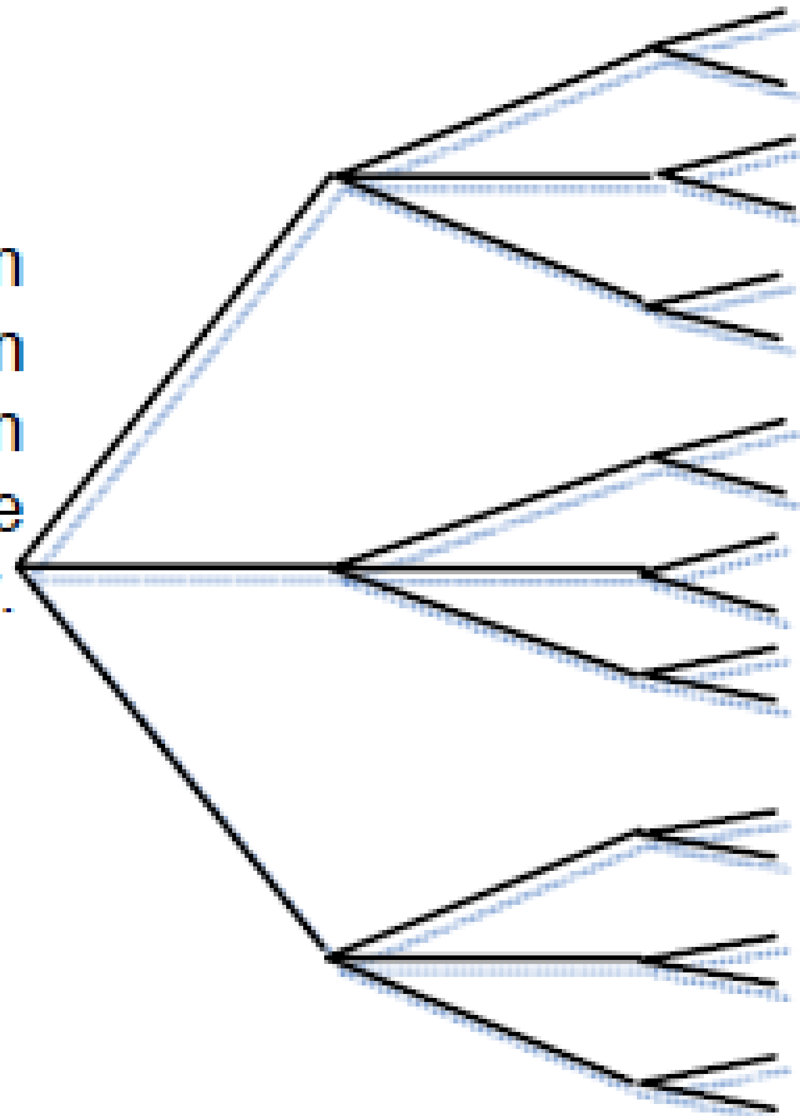
Örnek: 10 Coğrafya ve 8 Jeoloji öğrencisi arasından 5 kişilik bir komisyon oluşturulacaktır. Rasgele bir seçim yapıldığında komisyonda çoğunlukla Coğrafya öğrencisi olma olasılığı nedir?

5 Coğrafya 0 Jeoloji, 4 Coğrafya 1 Jeoloji, 3 Coğrafya 2 Jeoloji

$$\frac{{}_{10}C_5 \; {}_8C_0}{{}_{18}C_5} + \frac{{}_{10}C_4 \; {}_8C_1}{{}_{18}C_5} + \frac{{}_{10}C_3 \; {}_8C_2}{{}_{18}C_5} = \frac{5292}{8568} \approx 0{,}62$$

Ağaç Diyagramı

- Her birinin sonucunun sonlu sayıda olduğu birden fazla deneyin tüm mümkün sonuçlarını görsel bir şekilde ortaya koymak için kullanılır.

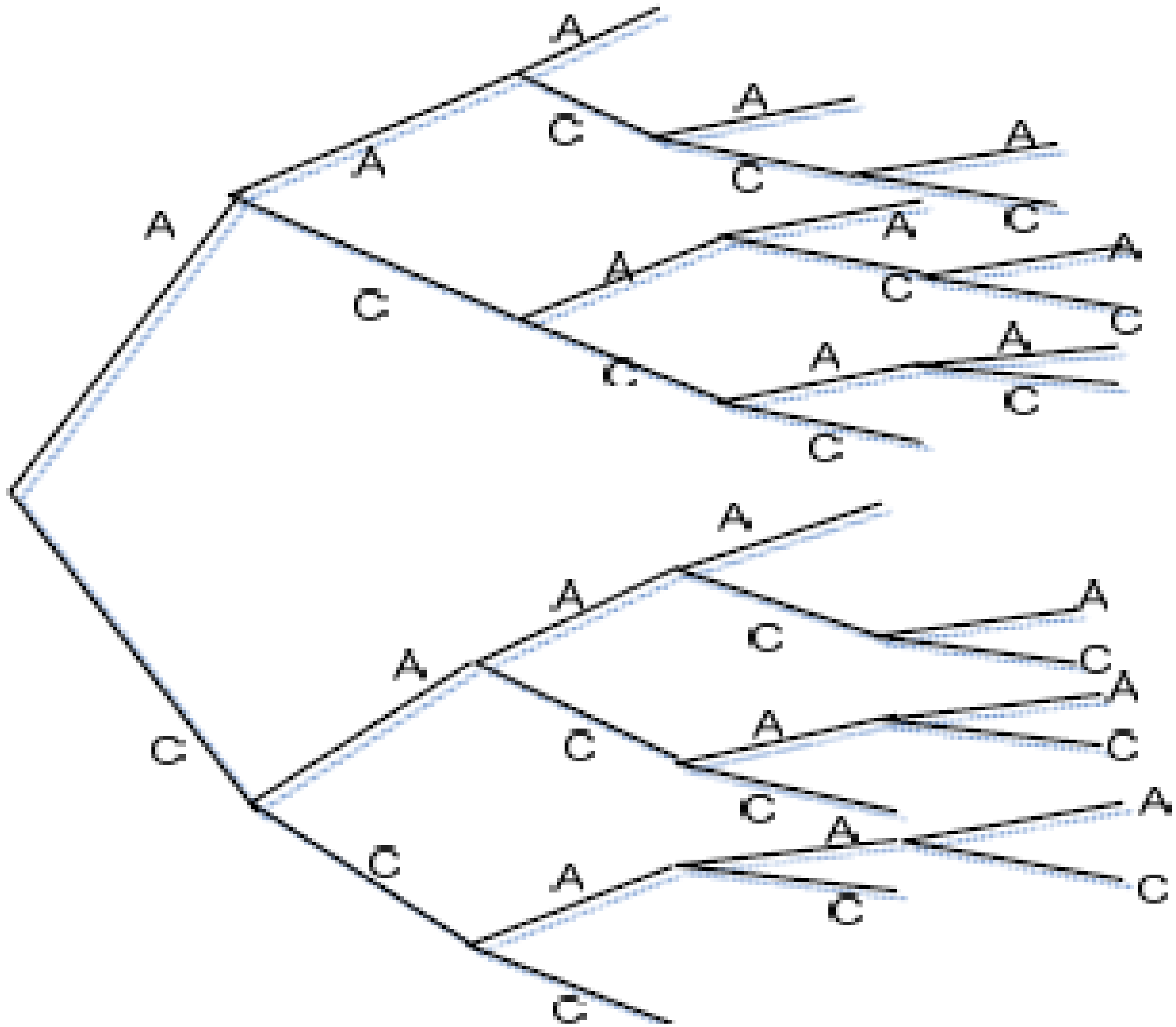


Örnek: Ali ile Can masa tenisi oynamaktadırlar. 3 set kazananın galip geleceği maçın ortaya çıkabilecek tüm mümkün sonuçlarını gösteren ağaç diyagramını oluşturunuz.

Olası Durumlar;

AAA,CCC
AACA,CCAC
ACAA,CACC
ACCC,CAAA
ACACA,CACAC
AACCA,CCAAC
AACCC,CCAAA
ACACC,CACAA
ACCAA,CAACC
ACCAC,CAACA

20 ADET



Olay Tipleri

Basit Olay (Elementer Olay):

Tek bir karakteristikle belirlenen olaylar

A: Bayan

B: 20 yaşın altında

C: Bir deste karttan kırmızı kart çekilmesi

D: Bir deste karttan bir as çekilmesi

Kesişen Olay**:** Aynı anda gerçekleşen olaylar

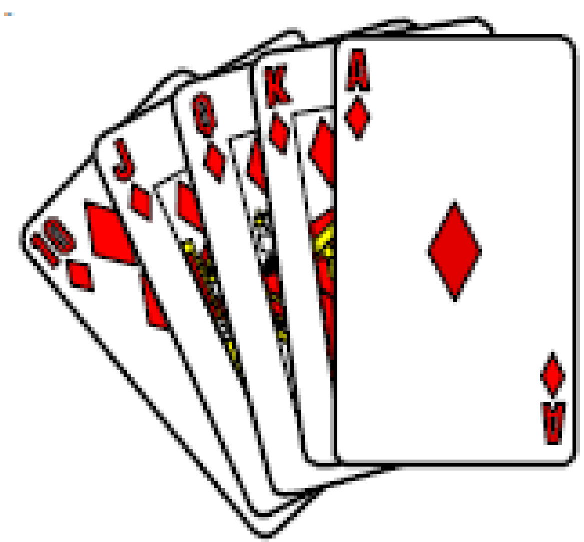
A ve B = (A∩B): Bayan ve 20 yaşın altında

C ve D = (C∩D): Kart destesinden kırmızı bir as çekilmesi

Bileşik olaylar (Birbirini Engelleyen Olaylar):

Olaylardan biri yada diğeri gerçekleşir, birden çok sonuçtan oluşur.

C yada **D** = (C U D): Bir deste karttan kırmızı veya as çekme



- **Bağımlı ve Bağımsız Olaylar**

  - **Bağımsız Olaylar (Independent Events) :** Eğer bir olayın ortaya çıkması (occurrence) öteki olayın ortaya çıkma olasılığını etkilemiyorsa, olaylar bağımsız olaylardır.

E_1 = Madeni bir para atma deneyinde tura gelmesi E_2 = Aynı paranın 2. atışında yazı gelmesi

İkinci atığın sonucu önceki atışın sonucuna bağlı değildir.

  - **Bağımlı Olaylar (Dependent Events):** Bir olayın ortaya çıkması diğerinin ortaya çıkması olasılığını etkiliyorsa bağımlı olaylardır.

E_1 = Meteorolojiden yağmur tahmini yapılması E_2 = Evden çıkarken şemsiye alınması

İkinci olayın sonucu 1.olayın sonucuna bağlıdır.

Basit Olaylar için Toplama Kuralı

- Bir E_i olayını olasılığı E_i olayını oluşturan çıktıların olasılıklarının toplamına eşittir.
- Şöyle ki;

$$\boxed{E_i = \{e_1, e_2, e_3\}}$$

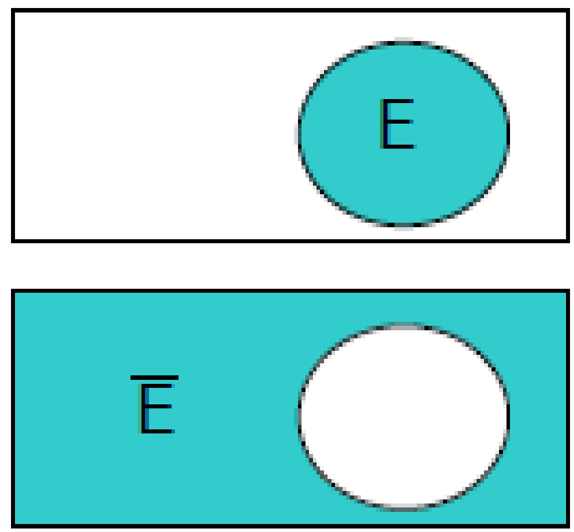
dolayısıyla:

$$\boxed{P(E_i) = P(e_1) + P(e_2) + P(e_3)}$$

Tamamlayıcı (Bütünleyici - Complement) Olay

İki olay kesinlikle aynı anda olamaz.
Para atımında aynı anda hem yazı hem de tura gelemez.

- Bir E olayının tamamlayıcısı E olayını içermeyen mümkün tüm basit olaylar kümesidir. Tamamlayıcı olay $\bar{E}$ ile gösterilir.
- **Tamamlayıcı Kural**



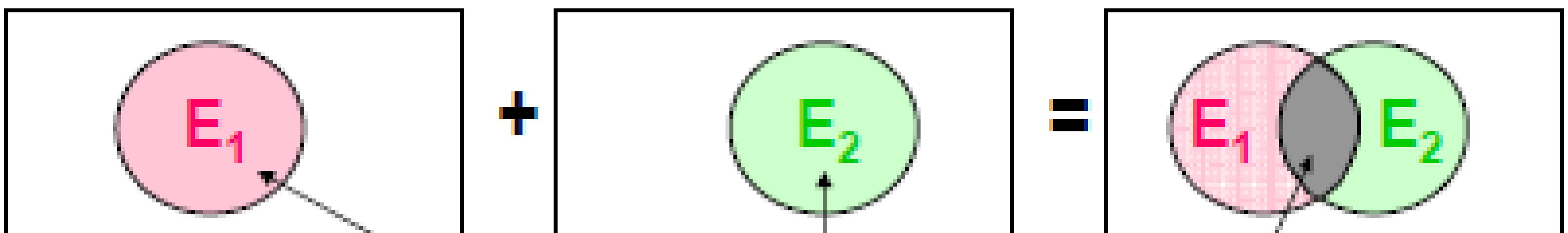
$$P(\bar{E}) = 1 - P(E)$$

veya, $P(E) + P(\bar{E}) = 1$

İki Olay İçin Toplama Kuralı

- **Toplama Kuralı:**

$$P(E_1 \text{ veya } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ and } E_2)$$



$$P(E_1 \text{ veya } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ and } E_2)$$

Kesişimi iki kere sayma!

Toplama Kuralı Örneği

P(Kırmızı veya As)=P(Kırmızı)+P(As)-P(Kırmızı ve As)

$$= 26/52 + 4/52 - \boxed{2/52} = 28/52$$

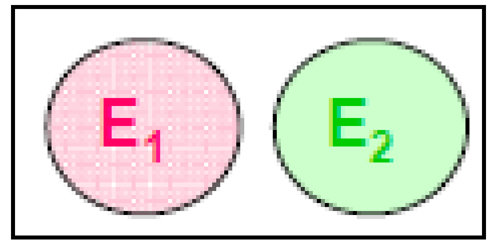
Tip	Renk		Toplam
	Kırmızı	Siyah	
As	2	2	4
As Değil	24	24	48
Toplam	26	26	52

Kesişimi iki kere sayma!

Ayrık Olaylar İçin Toplama Kuralı

- Eğer E_1 ve E_2 ayrık olaylarsa,

$$P(E_1 \text{ ve } E_2) = 0$$



= 0 ayrık olaylarsa

$$P(E_1 \text{ veya } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ ve } E_2)$$

Bu yüzden,

$$\boxed{P(E_1 \text{ veya } E_2) = P(E_1) + P(E_2)}$$

Koşullu Olasılık

Bir olayın gerçekleştiği bilindiği durumlarda diğer bir olayın gerçekleşme olasılığıdır.

$$\underbrace{P(A \mid \mathbf{B})}_{} = \frac{P(A \text{ ve } B)}{P(\mathbf{B})}$$

B olayının gerçekleştiği bilindiğine göre A olayının gerçekleşme olasılığı

$P(A \mid \mathbf{B}) = P(A)$ ise, A ve B birbirinden bağımsız olaylardır.

Kontenjans Tablosu yardımıyla koşullu olasılık hesabı:

Bir desteden çekilen bir kartın siyah olduğu bilindiğine göre as olma olasılığı nedir?

Tip	Renk		Top.
	Kırmızı	Siyah	
As	2	2	4
As değil	24	24	48
Toplam	26	26	52

$$P(\text{As} \mid \text{Siyah}) = \frac{P(\text{As VE Siyah})}{P(\text{Siyah})} = \frac{2/52}{26/52} = \frac{2}{26}$$

Koşullu Olasılık Örneği

- İkinci el araba pazarındaki arabaların 70% klima (KL) ve 40% CD çalar (CD) ve 20%'sinin ise her ikisine de sahip olduğu tespit edilmiştir.

- Kliması olan bir arabanın CD çalarının olması olasılığı nedir?

  $P(CD \mid AC) = ?$

- İkinci el araba pazarındaki arabaların 70% klima (KL) ve 40% CD çalar (CD) ve 20%'sinin ise her ikisine de sahip olduğu…

	CD	CD Yok	Toplam
KL	.2	.5	.7
KL Yok	.2	.1	.3
Toplam	.4	.6	1.0

$$P(CD \mid AC) = \frac{P(CD \text{ ve } AC)}{P(AC)} = \frac{.2}{.7} = .2857$$

Bağımsız olaylar İçin Koşullu Olasılık

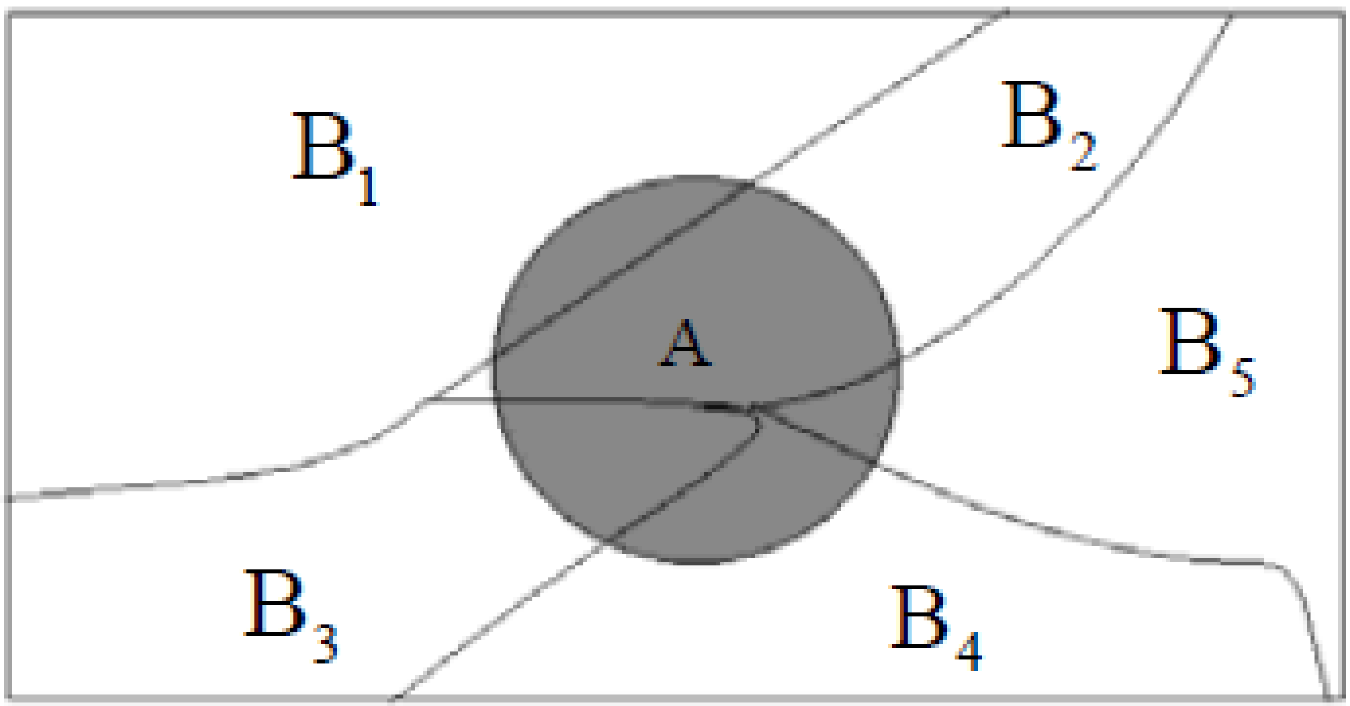
- Bağımsız olaylar E_1 , E_2 için koşullu olasılık:

$$\boxed{P(E_1 \mid E_2) = P(E_1)} \qquad P(E_2) > 0 \text{ şartı ile}$$

$$\boxed{P(E_2 \mid E_1) = P(E_2)} \qquad P(E_1) > 0 \text{ şartı ile}$$

Şartlı Olasılıkların Bilindiği Durumlarda Tek Bir Olayın Olasılığının Bulunması

Aşağıdaki şekilde A olayının birbiriyle ayrık olan 5 farklı olayın birleşiminden meydana geldiği görülür.



A olayı her bir B olayı ile kesişimleri cinsinden ifade edildiğinde;(birbirini engelleyen olayların birleşiminin olasılığı toplama kuralına göre)

$$P(A)=P(A\cap B_1)+P(A\cap B_2)+....+P(A\cap B_5)$$

$$P(A\cap B_i)=P(A/B_i).P(B_i)$$

$$P(A)=P(A/B_1)P(B_1)+P(A/B_2)P(B_2)+P(A/B_3)P(B_3)$$
$$+P(A/B_4)P(B_4)+P(A/B_5)P(B_5)$$

Ornek: Bir ilaç üç fabrika tarafından üretilmektedir. 1. Fabrikanın üretimi 2. ve 3. fabrikaların üretiminin 2 katıdır. Ayrıca 1. ve 2. fabrikalar % 2, 3. fabrika % 4 oranında bozuk ilaç üretmektedir. Üretilen tüm ilaçlar aynı depoda saklandığına göre bu depodan rast gele seçilen bir ilacın bozuk olma olasılığı nedir.

A = Seçilen ilacın bozuk olma olasılığı P (A) = ?

B_i= Seçilen ilacın i nci fabrikada üretilmesi

$P(B_1) = P(B_2) + P(B_3)$

$P(B_1) + P(B_2) + P(B_3) = 1$ olduğundan;

$P(B_1) = 0{,}50$ $P(B_2) = P(B_3) = 0{,}25$ olarak elde edilir.

$$P(A) = P(A/B_1)P(B_1) + P(A/B_2)P(B_2) + P(A/B_3)P(B_3)$$

$$P(A)=(0.02)(0.5)+(0.02)(0.25)+(0.04)(0.25)=0{,}025$$

Depodan seçilen 1000 ürünün 25 tanesinin hatalıdır.

Çarpma Kuralı

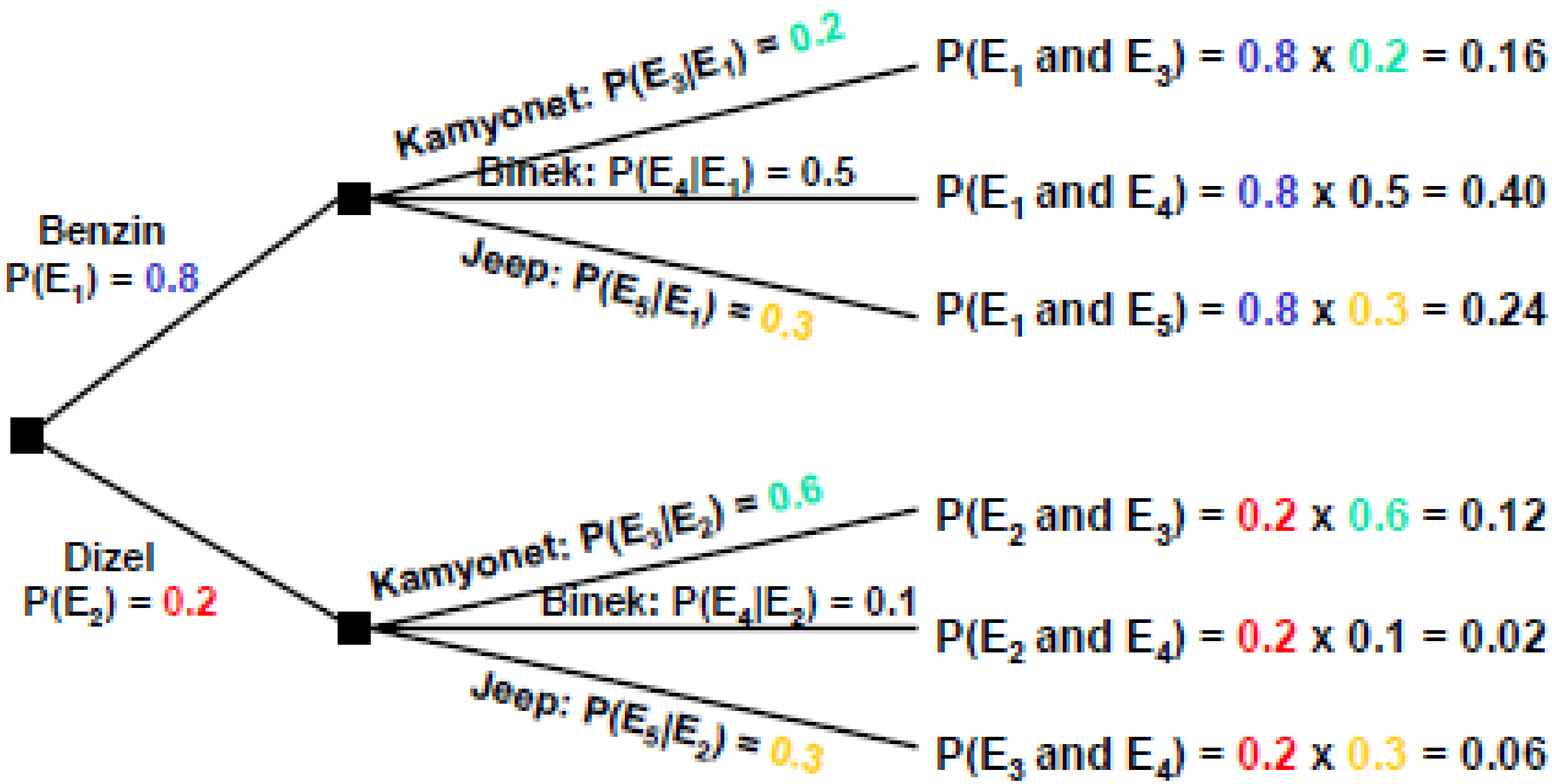
- İki olay E_1 ve E_2 için çarpma kuralı:

$$\boxed{P(E_1 \text{ VE } E_2) = P(E_1)P(E_2 \mid E_1)}$$

Not: Eğer **E_1** ve **E_2** bağımsız olaylar ise, yani $P(E_2 \mid E_1) = P(E_2)$
Çarpma kuralı basit çarpma olarak oluşur:

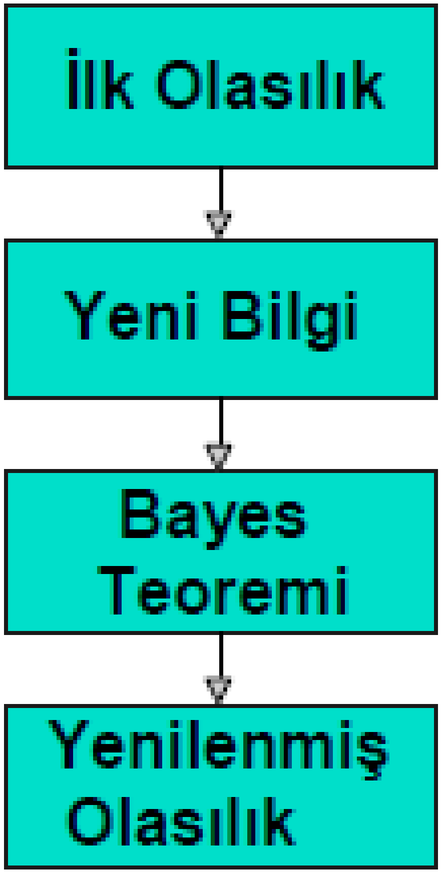
$$P(E_1 \text{ VE } E_2) = P(E_1)P(E_2)$$

Ağaç Diyagram Örneği



Bayes Teoremi

1. Eski olasılıkların yeni bilgiler ışığında güncellenmesi için kullanılır.
2. Koşullu olasılığın bir çeşididir.
3. Tamamen ayrık olaylar için uygulanır.



Bayes Teoreminin Formülü

$$P(B_i \mid A) = \frac{P(A \mid B_i)\cdot P(B_i)}{P(A \mid B_1)\cdot P(B_1) + \cdots + P(A \mid B_k)\cdot P(B_k)}$$

$$= \frac{P(B_i \cap A)}{P(A)}$$

Aynı
olay

Bayes Teoremi

• Sonucun bilindiği durumda sebebin hangi olasılıkla hangi olaydan meydana geldiği ile ilgilenir.

• Ele alınan örnekte depodan rast gele seçilen bir ilacın bozuk çıkması halinde 1.fabrikadan gelmesinin olasılığı araştırıldığında Bayes Teoremine ihtiyaç duyulmaktadır.

$$P(B_i / A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(A / B_i)P(B_i)}{\sum_{i=1}^{k} P(A / B_i)P(B_i)}$$

Depodan rasgele seçilen bir ilacın bozuk olduğu bilindiğine göre 1 nci fabrikadan gelmiş olma olasılığı;

$$P(B_1/A) = \frac{P(A/B_1)P(B_1)}{P(A/B_1)P(B_1) + P(A/B_2)P(B_2) + P(A/B_3)P(B_3)}$$

$$P(B_1/A) = \frac{(0.02)(0.5)}{(0.02)(0.5) + (0.02)(0.25) + (0.04)(0.25)} = 0{,}40$$

6. OLASILIK FONKSİYONLARI

Kesikli X rastgele değişkeni ile ilgili iki fonksiyon tanımlanabilir. Birincisi olasılık kütle fonksiyonu, ikincisi de kümülatif (birikimli) dağılım fonksiyonudur. Olasılık kütle fonksiyonu f ile gösterilirse X'in belirli bir a değerine sahip olması olasılığı,

$$P(X=a) = f(X)$$

İle tanımlanır. Kümülatif fonksiyonu, X ratgele değişkenin belirli bir değere eşit ya da o değerden küçük olması olasılığını verir.

$$F(a) = P(x \leq a)$$

Biçiminde tanımlanır.ve

Benzer şekilde eğer X sürekli bir değişkeni için de iki fonksiyon tanımlanabilir. Olasılık yoğunluk fonksiyonu f(X) ile gösterilir. [a,b] aralığında tanımlı X sürekli rastgele değişkeninin, a≤ < c ≤ b aralığında olması olasılığı,

$$P(X = c) = f(X)$$

Kümülatif dağılım fonksiyonu,

$$F(c)=\int P(X \leq c)dx= \int f(x)dx$$

Dağılımlar

- **Binom**
- **Poisson**
- Geometrik
- Hipergeometrik
- Üniform
- **Normal**
- Lognormal
- Gamma
- Gumbel (Ekstrem değer)
- Weibull

6.1 Normal Dağılım

Çoğu sürekli değişken normal dağılım yardımıyla istenen olasılıkların hesaplanabilmesi bu dağılıma istatistik kuramı içerisinde çok önemli bir yer sağlar. Normal Olasılık yoğunluk fonksiyonu:

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

σ: Verinin standart sapması

μ: Verinin ortalamasıdır.

Standart Normal Dağılım

Veriler,

Not: Çeşitli Z değerlerine ilişkin olasıkların bulunması için Z-Tablolarından yararlanılır

$Z = (x-\mu)/\sigma$

dönüşümü yardımıyla standartlaştırılır. Yani,

$\mu = 0$

$\sigma^2 = 1$

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

Örnek: X rastgele değişkenin ortalaması 100 ve standart sapması 15 olan normal dağıldığı varsayılmaktadır.

a) x'lerin 120'de büyük ve eşit olma olasılığı?

b) x'lerin %90 ının hangi değerlerden küçük olması beklenir?

c) X'lerin %95 inin hangi değerlerden büyük olması beklenir?

a) $P(X \geq 120)$?

$Z = (120 - 100)/15 = 1.33$

Tablodan $P(z \geq 1.33) = 0.0918$

b) $P(Z \leq 1.282) = 0.9$

$1.282 = (x - 100)/15 \Rightarrow x = 119.23$

c) $P(Z \geq -1.645) = 0.95$

$-1.645 = (x - 100)/150 \Rightarrow x = 75.325$

6.2. Binom Dağılımı

- iki halden birinin söz konusu olduğu sayımlı deneylerle ilgili olasılık hesaplarında kullanılır.
- Tekrarlanan deneyleri her bakımdan birbirinin aynı olması gerekli. (Yazı-Tura deneyinde yazı yada tura gelme olasılığı her deneyde eşit olmalı)

Binom Dağılımı

TTT=ppp=p^3

TTY=ppq=p^2q

TYT=pqp=p^2q

YTT=qpp=p^2q

TYY=pqq=pq^2

YTY=qpq=pq^2

YYT=qqp=pq^2

YYY=qqq=q^3

$p^3+3p^2q+3pq^2+q^3=1$

$p+q=1$

p: başarı olasılığı

q: başarısızlık olasılığı

Binom Dağılımı

$$P(x) = C_n^x . p^x . q^{n-x}$$

Binom Dağılımı

$$C_n^x = \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Binom Dağılımı

- $p = 0.5$ olduğunda Dağılım simetrik olur.
- Dağılımın ortalaması: $n \cdot p$
- Varyans sapması: $n \cdot p \cdot q$

Binom Dağılımı

Örnek: 0.20’si kusurlu, 0.80’i kusursuz olduğu bilinen bir ana kütleden 5 parça seçilirse bu seçilen parçalar içinde 3 tane kusurlu olma olasılığı nedir?

Çözüm:

n=5

X=3

p= 0.2

q= 1- p= 1-0.2 = 0.8

$$P(x=3)=\frac{5!}{3!(5-3)!}0.2^{3}0.8^{5-3}=0.0512$$

6.3 Poisson Dağılımı

- $p \leq 0.1$
- $N > 20$
- $n \cdot p \leq 5$

ise Poisson Dağılımı kullanılabilir.

Poisson Dağılımı

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Poisson Dağılımı

- Ortalaması:$\lambda = n \cdot p$
- Varyansı: λ

Poisson Dağılımı

Örnek: Bir hava alanına saat 14-15 arasında her 15 dakikada ortalama 3 uçak inmektedir. 14-15 arasında herhangi bir 15 dakikada alana 5 uçak inmesi olasılığı nedir?

Çözüm:

$$P(x=5)=\frac{3^{5}e^{-3}}{5!}=0.099225$$

6.4 Risk (exponansiyel dağılım)

Hidrolojik, Tarımsal v.b. çalışmalarda yaygındır

P_0: T yılda bir veya daha fazla ortaya çıkmama olasılığı olarak tanımlanır. (T dönüş periyodu)

$$P_0 = p(x = 0) = e^{-\lambda} = 1 - P_n$$

Risk (exponansiyel dağılım)

Örnek: 7 gün içinde bir kavşakta ortalama kaza sayısı 3 ise 2 gün içinde hiç kaza görülmemesi olasılığı nedir?

$$p = 3/7$$

$$\lambda = n \cdot p = 2 \cdot 3/7$$

$$P_2 = e^{-2 \cdot 3/7} = 0.424$$

7. HİPOTEZ TESTLERİ

Hipotez Nedir?

•HİPOTEZ→ parametre hakkındaki bir inanıştır.

Parametre hakkındaki inanışı test etmek için hipotez testi yapılır.

Hipotez testleri sayesinde örneklerden elde edilen istatistikler aracılığıyla, toplum (anakütle) parametreleri hakkında karar verilir.

Anakütle parametreleri hakkında karar verirken doğru ya da yanlış olması muhtemel yargılardan hareket edilir.

Örnek: “Bu sınıfın not ortalamasının 75 olduğuna inanıyorum”

Bir hipotez testinde iki hipotez yer alır:

H_0 : Boş hipotez, sıfır hipotezi

H_1 ya da H_a : Alternatif hipotez

Daha önce doğru olduğu ispatlanan veya ortak kabul görmüş yargılara sıfır hipotezi(H_0) denir. İnandığımız durum H_0 hipotezinde yer alır. Aksi ispat edilemedikçe H_0 hipotezi doğru kabul edilir. İddia edilen durum H_1 hipotezinde ele alınır. Sıfır hipotezinde belirtilen yargının tersi bir yargıyı içinde bulunduran hipoteze alternatif hipotez (H_1) denir.

Kendini kanıtlama zorunluluğu H_1 hipotezine aittir. H_1 hipotezi daima H_0 hipotezinin tersi olarak ifade edilir.

Problemlerdeki hipotezleri belirlemek:

Populasyon ortalamasının 75 olduğunu test ediniz.

Adımlar:

Soruyu istatistiksel olarak belirtin (H_0: $\mu = 75$)

Zıddını istatistiksel olarak belirtin (H_1: $\mu \neq 75$)

Hipotezler birbirinden tamamen ayrıktır.

Hipotezlerin belirlenmesi alıştırmaları:

Aşağıdaki durumlarda hipotezleri oluşturunuz:

1. Populasyonun günde TV seyretme süresinin ortalaması 12 midir?

   H_0: $\mu = 12$

   H_1: $\mu \neq 12$

2. Populasyonun günde TV seyretme süresinin ortalaması 12 den farklı mıdır?

   H_0: $\mu = 12$

   H_1: $\mu \neq 12$

3. Bir şapkanın ortalama maliyetinin 2 TL’den büyük olduğu iddia edilmektedir, araştırınız.

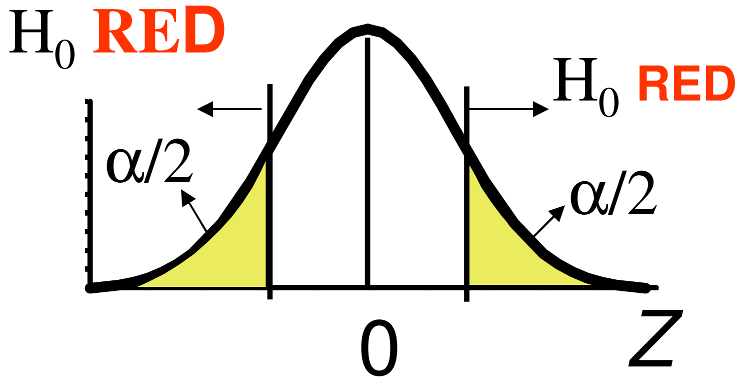
H_0: $\mu \leq 2$ H_1: $\mu > 2$

4. Kitapçıda harcanan paranın 25 TL’den küçük olduğu iddia edilmektedir, araştırınız.

H_0: $\mu > 25$ H_1: $\mu \leq 25$

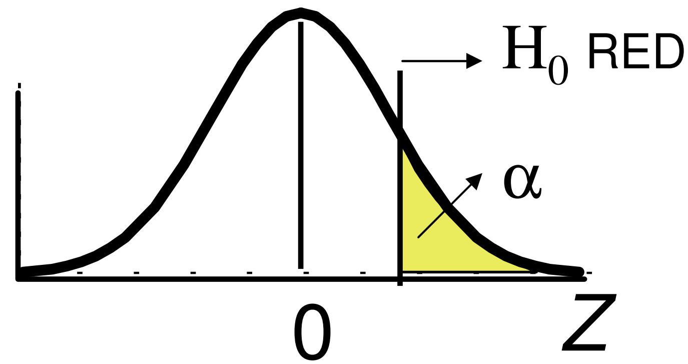
Hipotez Çiftleri:

$$\left.\begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{array}\right\}$$ **Çift kuyruklu test**



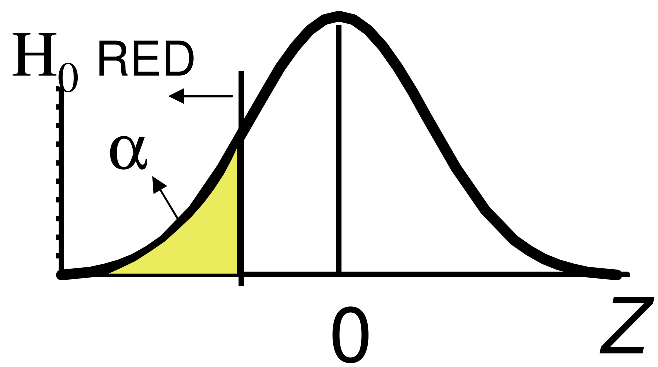
$$\left.\begin{array}{l} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{array}\right\}$$ **Tek kuyruklu test**

(Sağ taraf testi)



$$\left.\begin{array}{l} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{array}\right\}$$ **Tek kuyruklu test**

(Sol taraf testi)



Yanılma Olasılığı Ve Anlamlılık Seviyesi: α, 1- α

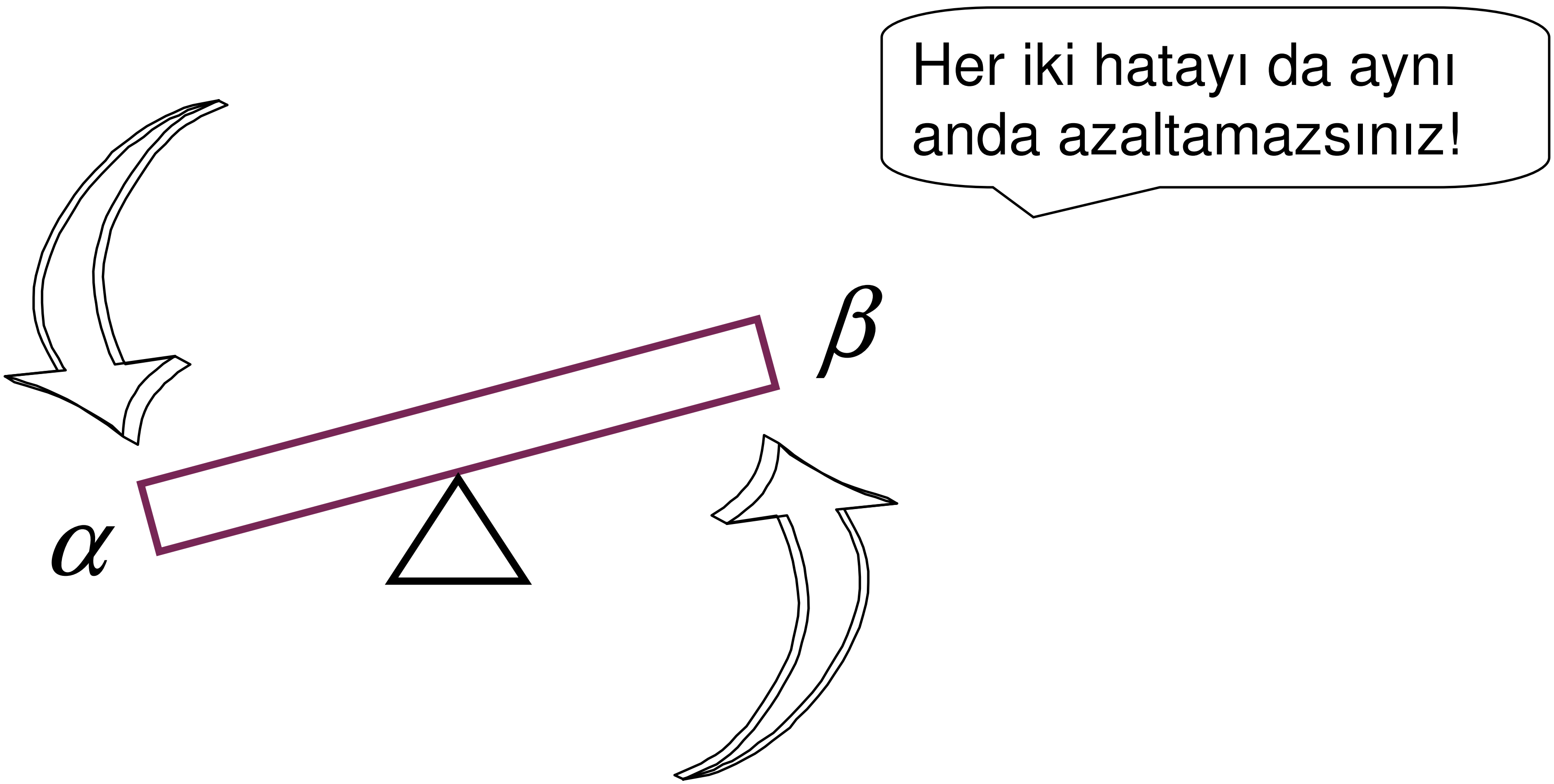
- Örnekleme dağılımının RED bölgesinin büyüklüğünü gösterir.
- Tipik değerleri: 0.01, 0.05, 0.10
- Araştırmanın başında araştırmacı tarafından seçilir.

Hipotez Testi

	Karar vermedeki hatalar				
	Gerçek durum			Gerçek durum	
Karar	Masum	Suçlu	Karar	H_0 doğru	H_0 yanlış
Masum	Doğru	Hata	H_0 red edilemez	1-α	II.tip hata(β)
Suçlu	Hata	Doğru	H_0 red	I.tip hata(α)	1-β

H_0: Masumdur

α & β Ters yönlü ilişki içindedir



β'yı Etkileyen Faktörler:

- Populasyon parametresinin gerçek değeri

  Hipotezdeki parametre değeri ile parametrenin gerçek değeri arasındaki fark arttıkça β da artar.

- Yanılma olasılığı: α

  α azalırken β artar.
- Populasyon standart sapması: σ

  σ arttıkça β artar.
- Örnek hacmi: *n*

  n azaldıkça β artar

Hipotez testi adımları:

1. H_0'ı belirle.
2. H_1'i belirle.
3. α'yı seç.
4. n'i seç.
5. Test istatistiğini seç
6. Kritik değerleri hesapla.
7. Veri topla.
8. Test istatistiğini hesapla.
9. İstatistiksel kararı ver.
10. Kararı açıkla ve yorumla.

Hipotez Testinde Test İstatistiğinin Belirlenmesi

Ortalama yada iki ortalama farkı için

- σ^2 biliniyor → **Z İstatistiği**
- σ^2 bilinmiyor
  - $n \geq 30$ → **Z İstatistiği**
  - $n < 30$ → **t İstatistiği**

Oranlar yada iki oran farkı için

- → **Z İstatistiği**

Varyansların testi

- Bir varyans için → **χ^2 İstatistiği**
- İki varyans oranı için → **F İstatistiği**

ORTALAMALARLA İLGİLİ HİPOTEZ TESTLERİ

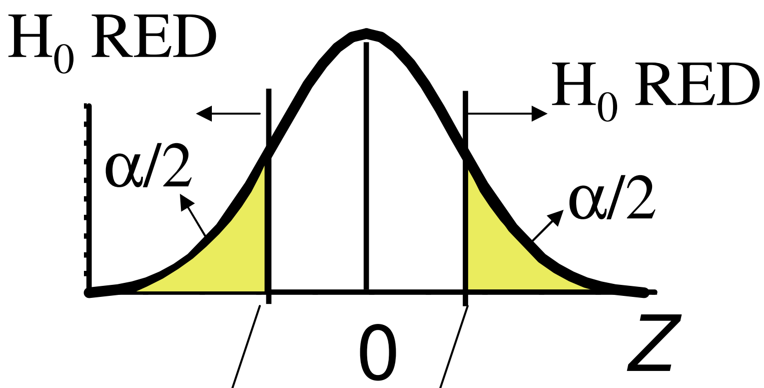
σ bilindiğinde Z test istatistiği

$$Z = \frac{\bar{x} - \mu_x}{\sigma_{\bar{X}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

σ bilinmediğinde fakat n≥30 olduğunda Z test istatistiği

$$Z = \frac{\bar{x} - \mu_x}{s_{\bar{X}}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Kabul ve Red Alanları: (Çift Kuyruklu Test)



$Z_{\alpha/2}$ kritik değerler tablodan bulunur.

$$H_0 : \mu = 45$$

$$H_1 : \mu \neq 45$$

Çift Kuyruklu Z Testine Örnek:

Bir fabrikada üretilmekte olan vidaların boylarının ortalaması 100 mm, ve standart sapması 2 mm olan normal dağılım gösterdikleri bilinmektedir. Makinalarda olan bir arıza giderildikten sonra üretilen vidalardan alınan 9 vidalık bir örneğin boy ortalaması 102 mm olarak bulunmuştur. **Makinalardaki arıza giderilirken vidaların boyunun ayarı bozulmuş mudur**? α=0.05 için test ediniz ve yorumlayınız.

1. Adım: Hipotezlerin belirlenmesi

$$H_0 : \mu = 100\, mm$$

$$H_1 : \mu \neq 100\, mm$$

2. Adım: Test istatistiğinin hesaplanması

$$Z_{hesap} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{102 - 100}{2/\sqrt{9}} = 3$$

μ=100mm

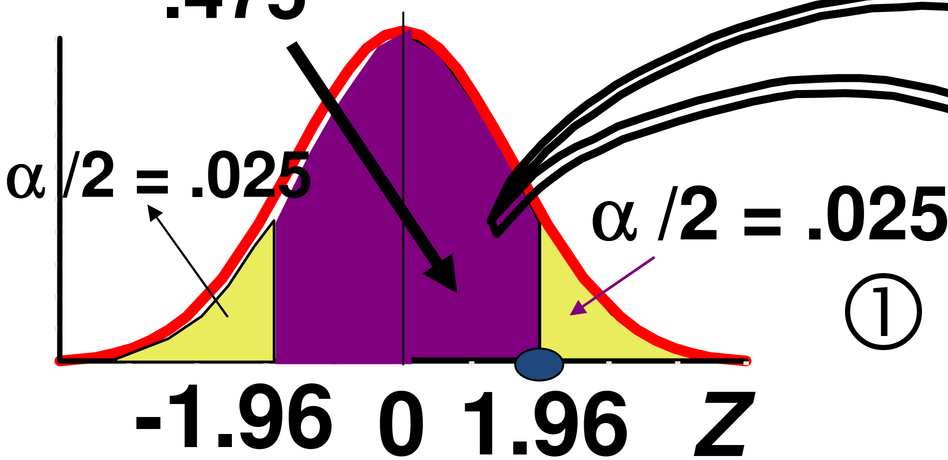
σ=2mm

n=9

$\bar{X} = 102 mm$

3. Adım: Kritik değerlerin belirlenmesi:

$$\begin{array}{r} .500 \\ -\ \underline{.025} \\ .475 \end{array}$$ ②



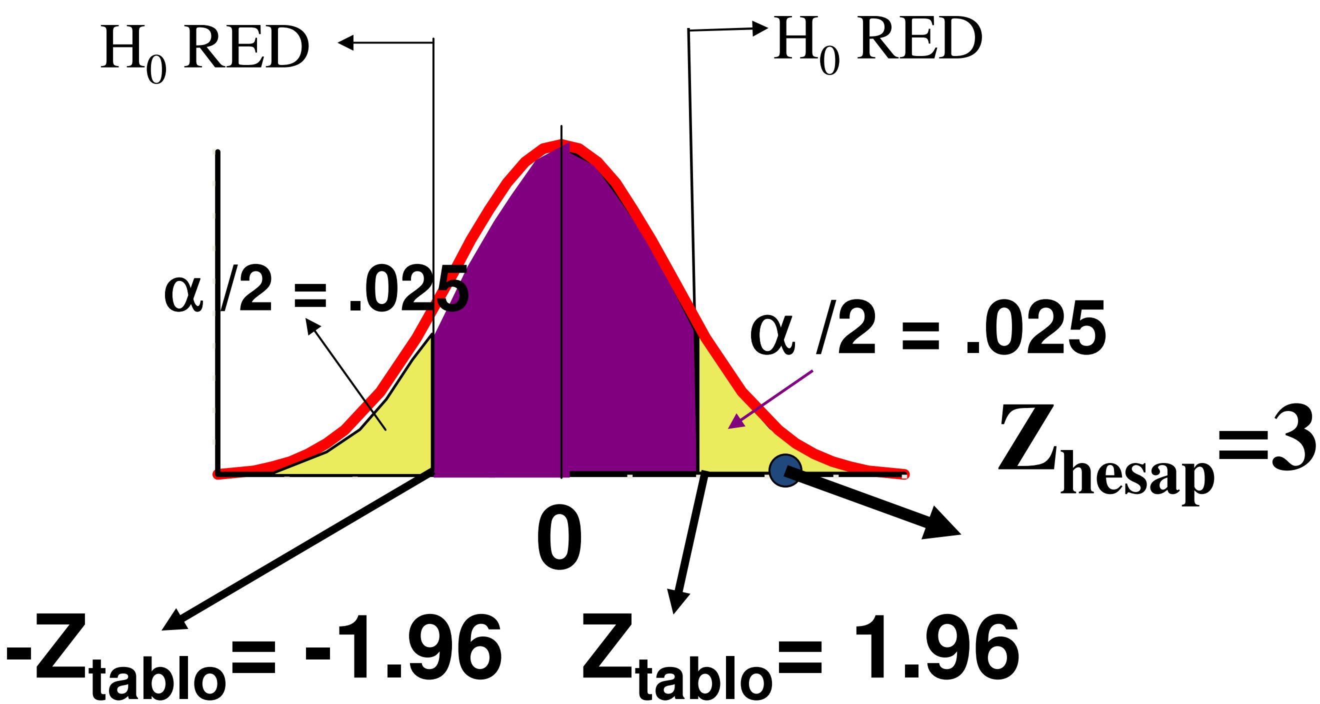
③ ① ④

Standart Normal Dağılım Tablosu

Z	.05	**.06**	.07
1.6	.4505	.4515	.4525
1.7	.4599	.4608	.4616
1.8	.4678	.4686	.4693
1.9	.4744	$\boxed{.4750}$	.4756

4. Adım: İstatistiksel karşılaştırmanın yapılması:

$$Z_{hesap} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{102 - 100}{2/\sqrt{9}} = 3$$ idi



5. Adım: Karar verme ve yorumlama:

Z_{hesap} değeri H_0 red bölgesine düştüğü için H_0 hipotezi reddedilir, yani vidaları boy ortalaması 100 mm'den farklıdır, makinanın ayarı bozulmuştur.

Tek Kuruklu Z Testi Örneği

- Bir kutu mısır gevreğinin ağırlığının **368 gr'dan fazla olduğu iddia edilmektedir**. Ayrıca σ = **15** gram olduğunu belirtilmiştir. n= **25** kutuluk bir örnek alınmış ve $\bar{X}$ **= 372.5** gr. olarak bulunmuştur. α= 0.05 yanılma seviyesinde test ediniz.

H_0: $\mu \leq 368$

H_1: $\mu > 368$

$\alpha = 0.05$

$n = 25$

Kritik değer:

Çözüm

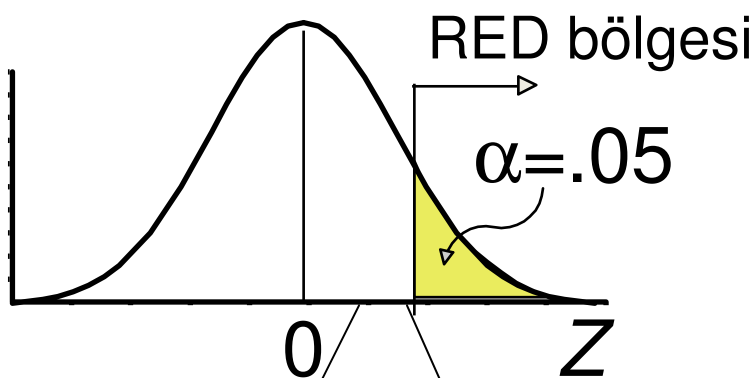
***H*0:** $\mu \leq 368$

***H*1:** $\mu > 368$

$\alpha = 0.05$

$n = 25$

- **Kritik değer:**



$Z_{hesap}=1.5$ $Z_{tablo}=1.645$

Test İstatistiği:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{372.5 - 368}{\frac{15}{\sqrt{25}}} = +1.50$$

Karar:

$\alpha = .05$ için H_0 hipotezi reddedilemez.

Yorum:

Ortalamanın 368 gr.dan fazla olduğuna dair yeterli kanıt yoktur.

Çift Kuyruklu Z testi örneği:

Bu sene ÇOMÜ, FEF, Coğrafya bölümünden mezun olacak öğrencilerin mezuniyet not ortalamalarının 70 olduğu iddia edilmektedir. Bu amaçla mezuniyet sonrası 36 öğrencilik bir örnek alınmış ve mezuniyet ortalamalarının 66, standart sapmasının 12 olduğu bulunmuştur. Bu veriler ışığında iddiayı α=0.01 için test ediniz.

$H_0 : \mu = 70$

$H_1 : \mu \neq 70$

$$Z = \frac{\bar{X} - \mu_{\bar{x}}}{s_{\bar{x}}} = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{66 - 70}{12/6} = \frac{-4}{2} = -2$$

$\alpha = 0.01$ için z tablo değeri 2.58

$|z_{hes}| < |z_{tab}|$ H_0 red edilemez.

ORANLARLA İLGİLİ HİPOTEZ TESTİ

Çift Kuyruk Testi

$$H_0 : P = P_0$$
$$H_1 : P \neq P_0$$

Sol Kuyruk Testi

$$H_0 : P \geq P_0$$
$$H_1 : P < P_0$$

Sağ Kuyruk Testi

$$H_0 : P \leq P_0$$
$$H_1 : P > P_0$$

Örnekten hesaplanan oran p ile gösterilirse oranlarla ilgili test istatistiği;

$$Z = \frac{p - P}{\sigma_p} = \frac{p - P}{\sqrt{\frac{P(1-P)}{n}}}$$

ÖRNEK

Bir süpermarketler zinciri sahibi müşterilerinin %95'ten fazlasının süpermarketlerindeki fiyatlardan memnun olduğunu söylemektedir. Tesadüfi olarak seçilen 200 müşteriden 184'ü fiyatlardan memnun olduğunu bildirmektedir. %99 önem düzeyinde, süpermarketteki fiyatlardan memnun olanların oranının %95'e eşit olmadığını söyleyebilir miyiz?

$$H_0 : P = P_0 \qquad \alpha = 0.01 \qquad p = 184/200 = 0.92$$

$$H_1 : P \neq P_0 \qquad Z_{tab} = \pm 2.58$$

$$Z = \frac{p-P}{\sigma_p} = \frac{p-P}{\sqrt{\dfrac{P(1-P)}{n}}} = \frac{0.92-0.95}{\sqrt{\dfrac{0.95(1-0.95)}{200}}} = -1.95$$

H_0 Kabul

ORTALAMALAR ARASI FARKLARLA İLGİLİ HİPOTEZ TESTLERİ

Çift Kuyruk Testi

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Sol Kuyruk Testi

$$H_0 : \mu_1 \geq \mu_2$$
$$H_1 : \mu_1 < \mu_2$$

Sağ Kuyruk Testi

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

İki Ortalama Farkı İçin Test İstatistiği

Ortalamalar arası farklarla ilgili hipotez testlerine ait test istatistiği

σ biliniyor ise:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\overline{X}_1 - \overline{X}_2}} = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Anakütle varyansları bilinmediğinde bunların yerine örnek varyansları kullanılır. Sıfır hipotezinin doğru olduğu varsayımı ile hareket edildiğinden μ_1-μ_2 farkı sıfır kabul edilir.

σ bilinmiyor fakat

örnek hacimleri ≥ 30 ise:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Sıfır hipotezi örneklerin aynı anakütleden alındığını belirttiği için tersi ispatlanmadığı sürece s_1 ve s_2 değerlerinin birbiriyle homojen olduğunu varsayılır ve ortak varyans hesaplanır.

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$$

$$Z_h = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

ÖRNEK

Aynı faaliyet kolunda üretim yapan fabrikaların birincisinden tesadüfi olarak seçilen 80 mamulün ortalama dayanma süresi 135 gün ve standart sapması 15 gün; ikincisinden alınan 95 mamulün ise ortalama dayanma süresi 130 gün ve standart sapması 18 gündür. %1 yanılma olasılığı seviyesinde, birinci fabrikada üretilen mamullerin ortalama dayanma süresinin daha fazla olduğunu söyleyebilir miyiz?

$$H_0 : \mu_1 \leq \mu_2 \qquad \alpha = 0.01 \qquad 0.5 - 0.01 = 0.49 \qquad Z_{tab} = 2.33$$

$$H_1 : \mu_1 > \mu_2$$

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} = \frac{80(15)^2 + 95(18)^2}{80+90} = 286.94$$

$$Z_h = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{(135-130)}{\sqrt{286.94\left(\frac{1}{80} + \frac{1}{95}\right)}} = 1.95$$

$$P(Z > 1.95) = 0.5 - 0.4744 = 0.0512$$

%1 yanılma seviyesinde sıfır hipotezi kabul edilerek birinci fabrikada üretilen mamullerin ortalama dayanma süresinin diğerlerinden daha fazla olmadığına karar verilir.

ORANLAR ARASI FARKLARLA İLGİLİ HİPOTEZ TESTLERİ

Çift Kuyruk Testi

$$H_0 : P_1 = P_2$$
$$H_1 : P_1 \neq P_2$$

Sol Kuyruk Testi

$$H_0 : P_1 \geq P_2$$
$$H_1 : P_1 < P_2$$

Sağ Kuyruk Testi

$$H_0 : P_1 \leq P_2$$
$$H_1 : P_1 > P_2$$

Oranlar arası farklarla ilgili hipotez testlerine ait test istatistiği

$$Z_h = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\dfrac{P_1(1 - P_1)}{n_1} + \dfrac{P_2(1 - P_2)}{n_2}}}$$

Anakütle oranları bilinmediğinde bunun yerine örnek oranları kullanılabilir. Sıfır hipotezinin doğru olabileceği varsayımıyla hareket edildiğinden test istatistiği formülündeki P_1-P_2 farkı sıfır kabul edilir. Test istatistiği aşağıdaki gibidir:

$$Z_h = \frac{(p_1 - p_2)}{\sqrt{\dfrac{p_1(1 - p_1)}{n_1} + \dfrac{p_2(1 - p_2)}{n_2}}}$$

Sıfır hipotezi örneklerin aynı anakütleden alındığını belirttiği için p_1 ve p_2 değerleri birbiriyle homojendir. Aşağıdaki ortak varyans hesaplanır.

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$Z_h = \frac{(p_1 - p_2)}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Örnek

Bir video kaset kiralayıcısı macera filmi kiralamanın yöredeki erkek ve kadınlar itibariyle farklılık gösterip göstermediğini merak etmektedir. Sözkonusu şahıs belli bir zaman dönemi içersinde dükkanına gelen 60 erkekten 51'nin ve 40 kadından 20'sinin macera filmi kiraladığını müşahede etmiştir. Bu verilere göre yöredeki erkeklerin kadınlardan daha fazla macera filmi kiraladığını % 95 önem seviyesinde söyleyebilir misiniz?

$$p_1 = \frac{51}{60} = 0.85 \qquad p_2 = \frac{20}{40} = 0.50$$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \qquad p = \frac{60(0.85) + 40(0.50)}{60 + 40} = 0.71$$

$$H_0 : P_1 \leq P_2$$

$$H_1 : P_1 > P_2$$

$$Z_{tablo} = 1.645$$

$$Z_h = \frac{(p_1 - p_2)}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$$Z_h = \frac{0.85 - 050}{\sqrt{0.71(1-0.71)(\frac{1}{60} + \frac{1}{40})}} = 3.78$$

H_0 RED

ALIŞTIRMALAR

1. Bir toplumda erkekler arasında akciğer hastalığı oranının %30 olduğu bilinmektedir. Sigara içenlerde akciğer hastalıklarına daha sık rastlanıp rastlanmadığı araştırılmak isteniyor. Bu amaçla sigara içen erkekler arasından rasgele seçilen 200 erkekten 80'inin bir akciğer hastalığı geçirdiği/geçirmekte olduğu saptanıyor. Sigara içenlerde akciğer hastalığına yakalanma oranının daha fazla olduğu söylenebilir mi?

2. Büyük bir alışveriş merkezinin kayıtlarına göre merkeze gelen 1000 erkekten 100'ü, 1000 bayandan ise 250'si oyuncak reyonundan alışveriş yapmıştır. Bayanların çocuklarına daha çok oyuncak alıp almadıklarını test ediniz.

3. Bir çimento fabrikası ürettiği çimentodan yapılan beton blokların sağlamlığının standart sapmasının 10 kg/m^2 'den fazla olduğunu iddia etmektedir. İddiayı test etmek amacıyla 10 beton blok alınmış ve sağlamlık test yapılmıştır. Test sonucunda alınan örneğin sağlamlık ortalaması 312 kg/m^2, varyansı 195 kg^2/m^4 olarak bulunmuştur.

a) İddiayı %95 güvenle test ediniz.

b) Aynı veriler için populasyon varyansının 200'ün altında olduğu iddiasını test ediniz.

c) Aynı veriler için populasyon varyansının 100 olup olmadığını test ediniz.

t-dağılımı ve t-testi

156

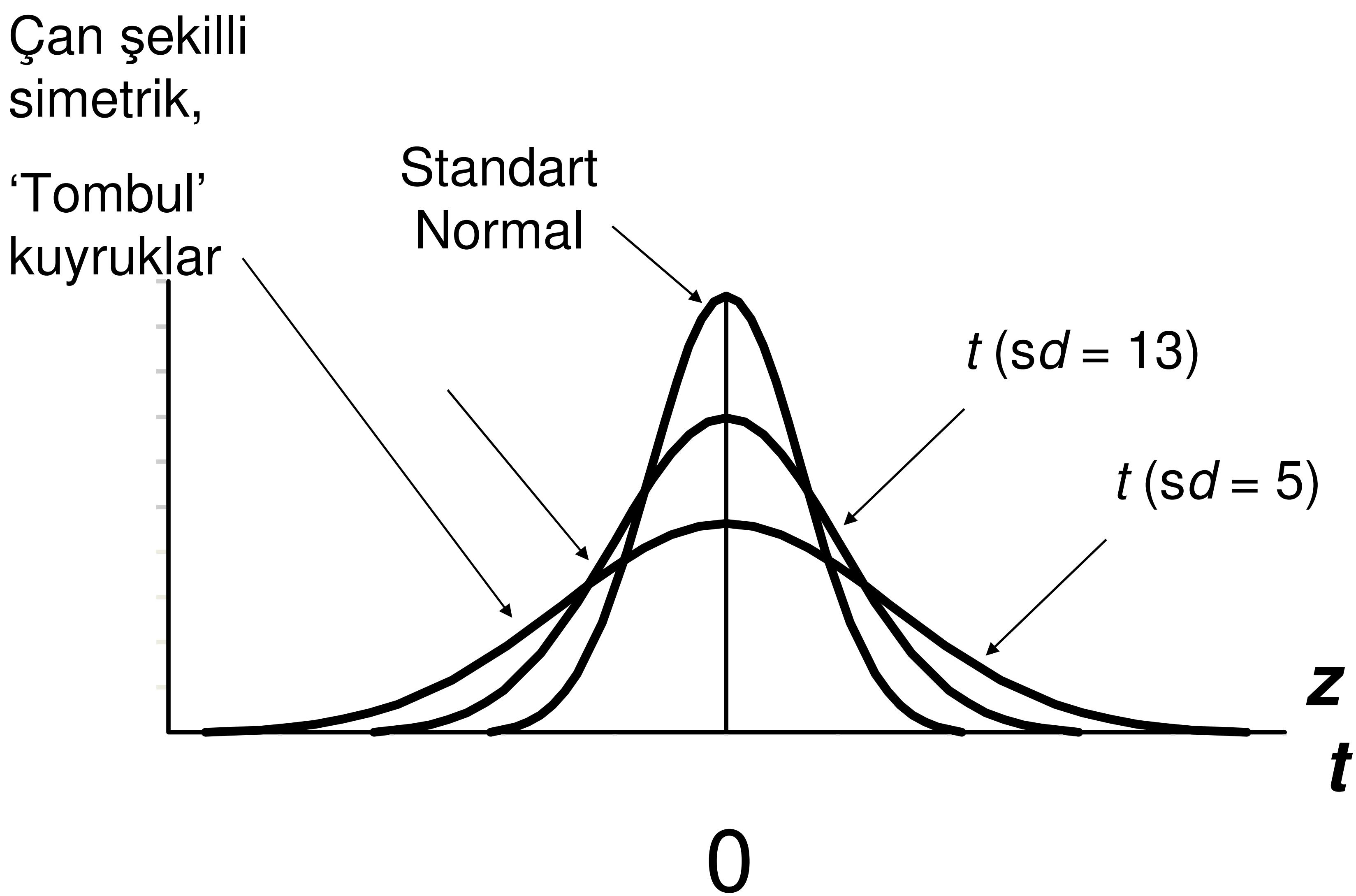
Student *t*-dağılımı

•Küçük örneklerden (n< 30) elde edilen istatistiklerin dağılımı Student t dağılımına uyar.

•Küçük örnek istatistiklerinin gösterdiği dağılım normal eğri gibi simetriktir. Normal eğriye göre daha basık ve yaygın bir şekil alır. Böylece eğrinin kuyruklarında daha büyük bir alan oluşur.

•Küçük örnekler için z cetveli yerine, çeşitli örnek büyüklükleri ve olasılık seviyeleri için ayrı ayrı hesaplanmış t cetvelleri kullanılır.

157


Çan şekilli
simetrik,
'Tombul'
kuyruklar
Standart
Normal
t (sd = 13)
t (sd = 5)
z
t
0

158

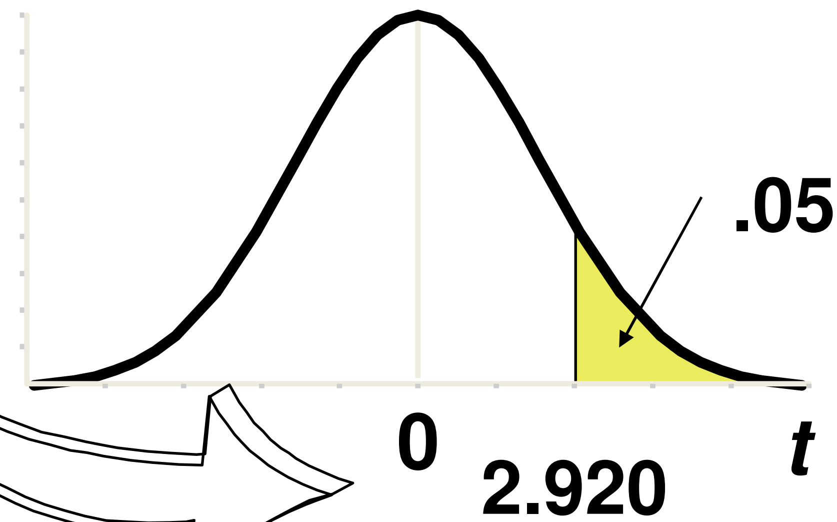
Student'ın *t* Tablosu

	Üst kuyruk alanı		
sd	.25	.10	**.05**
1	1.000	3.078	6.314
2	0.817	1.886	**2.920**
3	0.765	1.638	2.353

t değerleri

$n = 3$
sd $= n - 1 = 2$
$\alpha = .10$
$\alpha/2 = .05$

Olsun:



159

ORTALAMALARLA İLGİLİ HİPOTEZ TESTLERİ

Çift Kuyruk Testi

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Sol Kuyruk Testi

$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0$$

Sağ Kuyruk Testi

$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$

Ortalamalarla ilgili hipotez testlerine ait test istatistiği:

$$t_h = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n-1}}}$$

160

ÖRNEK

Bir konserve fabrikasının imal ettiği konservelerin üzerinde brüt 455 gr yazmaktadır. Bu konservelerin brüt ağırlıkları ile ilgili bir karar vermek üzere rasgele seçilen 17 kutunun ortalama ağırlığı 450 gr ve standart sapması 13 gr bulunmuştur. Brüt ağırlığın 455 gr olmadığını 0.05 yanılma seviyesinde söyleyebilir misiniz?

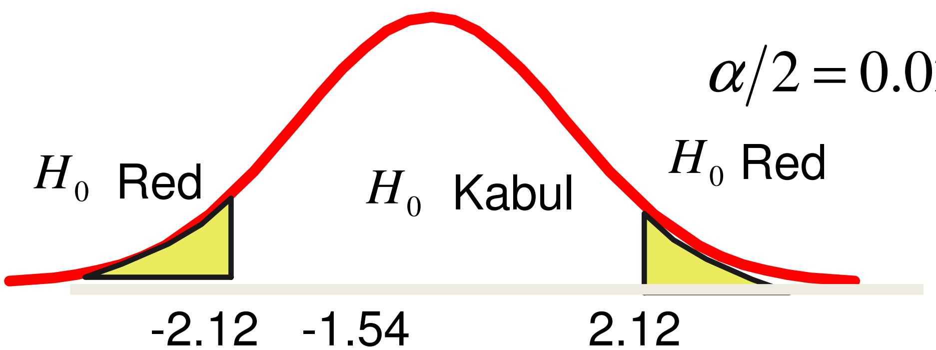
$$H_0 : \mu = 455$$
$$H_1 : \mu \neq 455$$

$$n = 17$$
$$\bar{X} = 450 gr.$$
$$s = 13 gr.$$

$$sd = n - 1 = 16$$

$$t_{tab} = \pm 2.12$$

$$\alpha/2 = 0.025$$



$$t_h = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n-1}}} = \frac{450 - 455}{\frac{13}{\sqrt{17-1}}} \cong 1.54$$

161

İKİ ANAKÜTLE ORTAMASINA İLİŞKİN HİPOTEZ TESTLERİ

Bağımsız ve İlişkili Populasyonlar

Bağımsız

1. Farklı veri kaynakları
   - İlişkisiz
   - Bağımsız
2. İki örnek ortalaması arasındaki farkın kullanılması

$$\overline{X}_1 - \overline{X}_2$$

İlişkili

1. Aynı veri kaynağı
   - Eşleştirilmiş
   - Tekrarlı ölçümler
2. Her gözlem çifti arasındaki farkın kullanılması

$$D_n = X_{1n} - X_{2n}$$

1- ÖRNEKLERİN BAĞIMSIZ OLMASI HALİ

Çift Kuyruk Testi

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Sol Kuyruk Testi

$$H_0 : \mu_1 \geq \mu_2$$

$$H_1 : \mu_1 < \mu_2$$

Sağ Kuyruk Testi

$$H_0 : \mu_1 \leq \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{s^2 \frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Ortalamalar arası farklarla ilgili hipotez testlerine ait test istatistiği: (σ bilinmiyor)

Örnek

- Verilen iki ayrı kesimhanenin et üretim kayıtlarıyla ilgili aşağıdaki verileri topladığınızı varsayınız:

	fab1	**fab2**
n	**21**	**25**
Ortalama	**3.27**	**2.53**
Std Sapma	**1.30**	**1.16**

Eşit varyans varsayımı altında, ortalama üretimde bir fark var mıdır (α = **0.05**)?

164

	fab1	fab2
n	**21**	**25**
Ortalama	**3.27**	**2.53**
Std Sapma	**1.30**	**1.16**

Test İstatistiğinin Hesaplanması

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{(21-1)\cdot 1.30^2 + (25-1)\cdot 1.16^2}{(21-1) + (25-1)} = 1.510$$

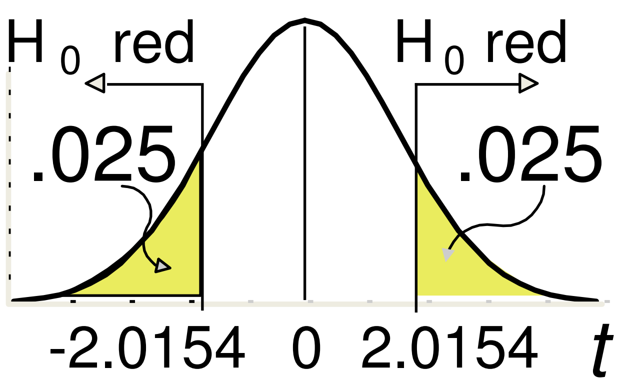
$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(3.27 - 2.53) - (0)}{\sqrt{1.510 \cdot \left(\frac{1}{21} + \frac{1}{25}\right)}} = +2.03$$

165

Çözüm

- H_0: $\mu_1 - \mu_2 = 0 \quad (\mu_1 = \mu_2)$
- H_1: $\mu_1 - \mu_2 \neq 0 \quad (\mu_1 \neq \mu_2)$

  $\alpha = 0.05$
- **sd** $= 21 + 25 - 2 = 44$
- **Kritik Değerler:**



$t_{hes} = 2.03$

$t_{hes} > t_{tab}$

$2.03 > 2.01$ H_0 red.

Ortalamalarda bir fark olabilir.

2-Eşleştirilmiş Örnek t Testi

1. İki ilişkili populasyonun ortalamasını test eder.
   - Çift ya da eşleştirilmiş
   - Tekrarlı gözlemler (önce/sonra)
2. Nesneler arasındaki varyasyonu ortadan kaldırır.
3. Varsayımları

   i. İki populasyon da normal dağılımlıdır.

   ii. Eğer normal değilse normale yaklaşmaktadır.

   iii. ($n_1 \geq 30$ ve $n_2 \geq 30$)

Eşleştirilmiş Örnek t Testi

İki komisyoncunun aynı evlere farklı fiyatlar verdiği iddia edilmektedir. İddiayı test etmek için 12 ev seçiliyor ve komisyonculardan bu evlere 1000$ bazında fiyat vermeleri isteniyor. Elde edilen sonuçlar aşağıdaki gibidir. **İki komisyoncunun aynı evlere farklı fiyatlar verip vermediğini test ediniz.**

Komisyoncular				
Evler	**A**	**B**	**D**	**D^2**
1	181.0	182.0	-1.0	1.00
2	179.9	180.0	-0.1	0.01
3	163.0	161.5	1.5	2.25
4	218.0	215.0	3.0	9.00
5	213.0	216.5	-3.5	12.25
6	175.0	175.0	0.0	0.00
7	217.9	219.5	-1.6	2.56
8	151.0	150.0	1.0	1.00
9	164.9	165.5	-0.6	0.36
10	192.5	195.0	-2.5	6.25
11	225.0	222.7	2.3	5.29
12	177.5	178.0	-0.5	0.25
Toplam			**-2.0**	**40.22**

168

Eşleştirilmiş Örnek t Testi

1.Adım: $H_0: \mu_D = 0$

$H_1: \mu_D \neq 0$

2.Adım: $\overline{D} = \frac{\sum D}{n} = \frac{-2}{12} = -0.167$ $\quad s_D = \sqrt{\frac{\sum D^2 - \frac{\left(\sum D\right)^2}{n}}{n-1}} = \sqrt{\frac{40.22 - \frac{(-2)^2}{12}}{12-1}} = 1.904$

$$t_{hes} = \frac{\overline{D}}{s_D / \sqrt{n}} = \frac{-0.167}{1.904 / \sqrt{12}} = -0.30$$

Sd=n-1=12-1=11

3.Adım: $t_{tab} : t_{11,0.05} = \pm\, 2.201$

4.Adım: $|t_{hes}| < |t_{tab}|$

H_0 reddedilemez. Çünkü %95 önem düzeyinde fiyatlandırma yönünden komisyoncuların birbirinden farklı olmadığına karar verebiliriz.

169

Alıştırma Soruları

1.Belli bir mesafeyi erkek yüzücülerin kız yüzücülerden daha kısa zamanda yüzdüğü iddia edilmektedir. Rassal olarak seçilen 200 erkek yüzücünün ortalama derecesi 60 ve standart sapması 10 dakika, 150 kız yüzücünün ortalama derecesi 70 ve standart sapması 15 dakika olarak bulunmuştur. % 1 yanılma payı düzeyinde karar veriniz.

2. A ve B marka ampullerin ömürlerinin farklı olduğu iddia edilmektedir. Rassal olarak seçilen A marka 10 ampulün ortalama ömrü 850 ve standart sapması 100 saat, B marka 1 ampulün ortalama ömrü 650 ve standart sapması 150 saat olarak bulunmuştur. %95 anlamlılık seviyesine göre karar veriniz.

170

ANAKÜTLE VARYANSI İÇİN HİPOTEZ TESTİ

Bir anakütle varyansının belirli bir değere eşit olup olmadığını veya büyük/küçük olup olmadığı test edilecektir.

Anakütle varyansına ilişkin testlerde karşılaşılabilecek muhtemel hipotez çiftleri aşağıdadır:

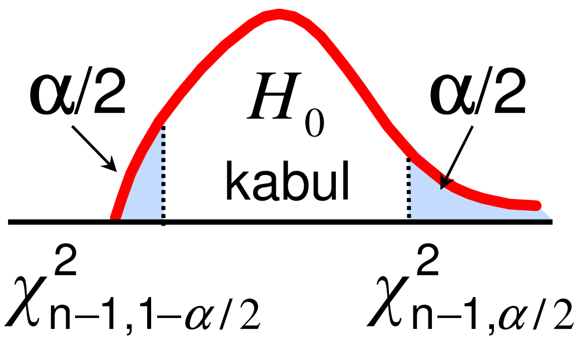
Çift kuyruk testi	Sol kuyruk testi	Sağ kuyruk testi
$H_0 : \sigma^2 = \sigma_0^2$	$H_0 : \sigma^2 \geq \sigma_0^2$	$H_0 : \sigma^2 \leq \sigma_0^2$
$H_1 : \sigma^2 \neq \sigma_0^2$	$H_1 : \sigma^2 < \sigma_0^2$	$H_1 : \sigma^2 > \sigma_0^2$

ANAKÜTLE VARYANSI İÇİN HİPOTEZ TESTİ

Çift Kuyruk Testi

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_1 : \sigma^2 \neq \sigma_0^2$$



Red H_0 eğer

$$\chi^2_{n-1} > \chi^2_{n-1,\alpha/2}$$

veya

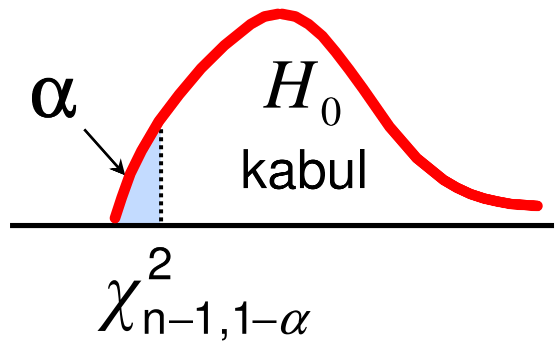
$$\chi^2_{n-1} < \chi^2_{n-1,1-\alpha/2}$$

Çift kuyruk testinde sd=n-1 ile $\alpha/2$ ve $1-\alpha/2$ önem seviyesi sütunlarının tarif ettiği iki değer kritik χ^2 değerleridir.

Test istatistiği bu değerler arasına düştüğünde H_0 hipotezi kabul edilir.

- Sol Kuyruk Testi

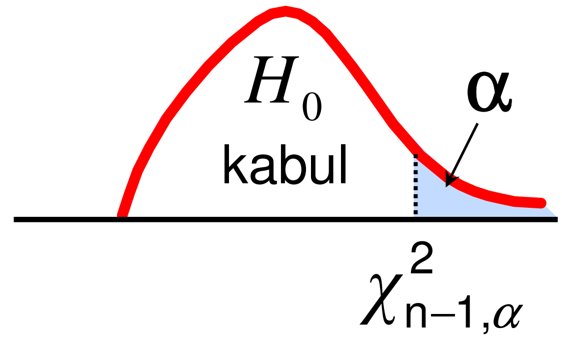
$$H_0 : \sigma^2 \geq \sigma_0^2$$
$$H_1 : \sigma^2 < \sigma_0^2$$



Red H_0 $\chi^2_{n-1} < \chi^2_{n-1,1-\alpha}$

Sol kuyruk testinde *ki-kare* cetvelinden sd=n-1 ve 1-α önem seviyesine göre kritik değer belirlenir. Test istatistiği kritik *ki-kare* değerinden küçük ise H_0 hipotezi red edilir.

$$H_0 : \sigma^2 \leq \sigma_0^2$$
$$H_1 : \sigma^2 > \sigma_0^2$$



Red H_0 $\quad \chi^2_{n-1} > \chi^2_{n-1,\alpha}$

Sağ kuyruk testinde sd=n-1 ve α yanılma payı seviyesine göre ki-kare cetvelinden kritik ki-kare değeri bulunur. Test istatistiği , bu değerden büyük olursa H_0 red edilir.

Test istatistiği; $\chi^2_{test} = \dfrac{(n-1)s^2}{\sigma_0^2}$

ÖRNEK

Bir ekmek fırınında üretilen ekmeklere ait gramajların ortalama etrafında normal dağıldığı ve standart sapmanın 9 gr olduğu iddia edilmektedir. İddiayı test etmek için tesadüfi olarak seçilen 20 ekmeğin standart sapması 10 gr bulunmuştur. **Standart sapmasının 9 gr'dan fazla olduğunu** %95 önem seviyesinde söyleyebilir misiniz?

ÇÖZÜM

İddia edilen anakütle varyansı; $\sigma^2 = 9^2 = 81$

$H_0 : \sigma^2 \leq 81$

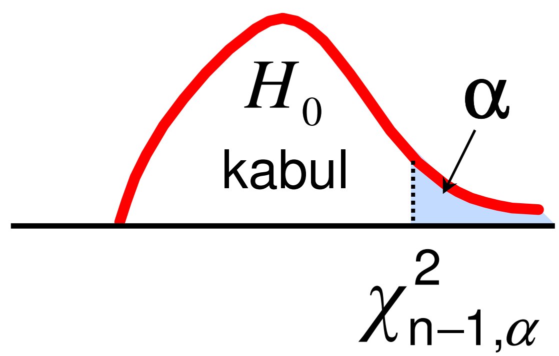
$H_1 : \sigma^2 > 81$

sd = n-1 = 20-1 = 19

$\alpha = 0.05$ $\longrightarrow$ $\chi^2_{tab} = 30.144$

$$\chi^2_{test} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(20-1)100}{81} = 23.457$$

$\chi^2_{tab} > \chi^2_{test}$ H_0 kabul



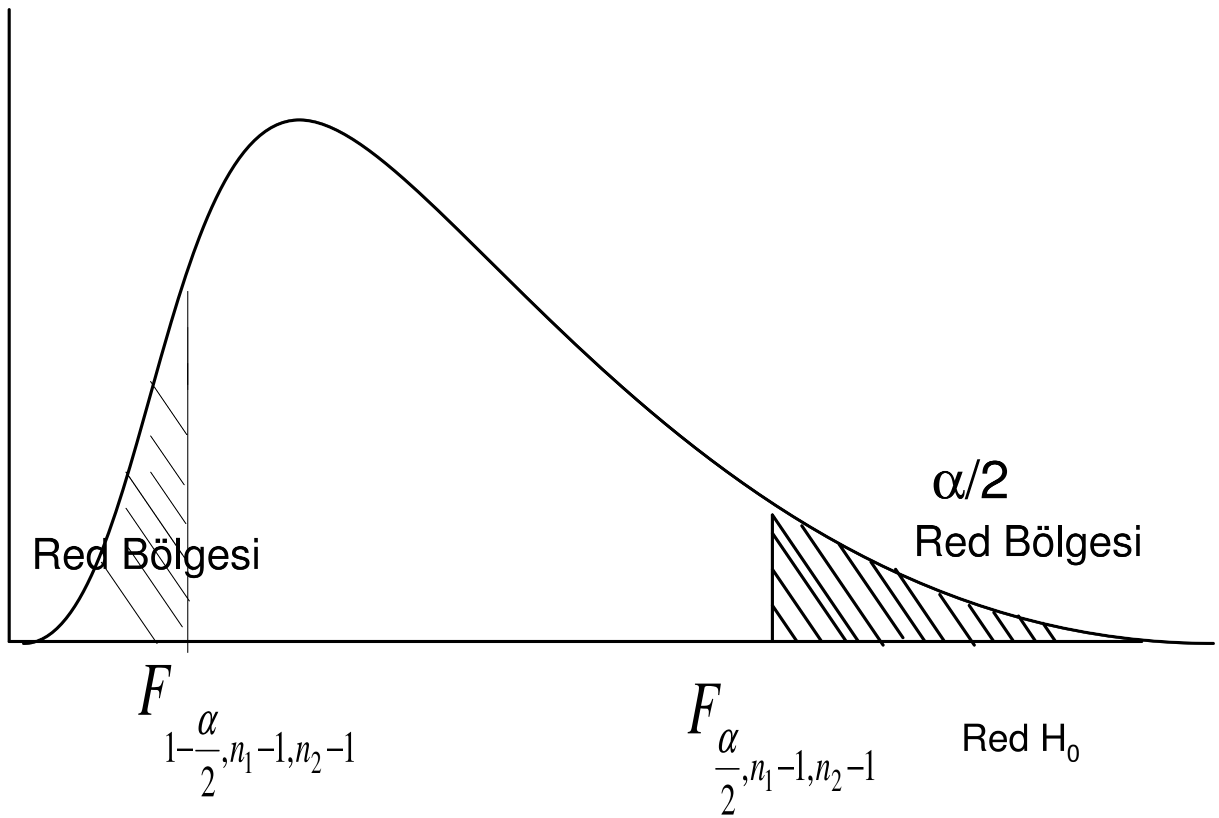
İKİ ANAKÜTLE VARYANSI İÇİN HİPOTEZ TESTİ

Varyansları σ_1^2 ve σ_2^2 olan normal dağılımlı iki anakütleden n_1 ve n_2 gözlemli bağımsız iki örneğin varyansları s_1^2 ve s_2^2 olsun. İki anakütle varyansının birbirine eşit olup olmadığını test etmek için:

Çift Kuyruk Testi

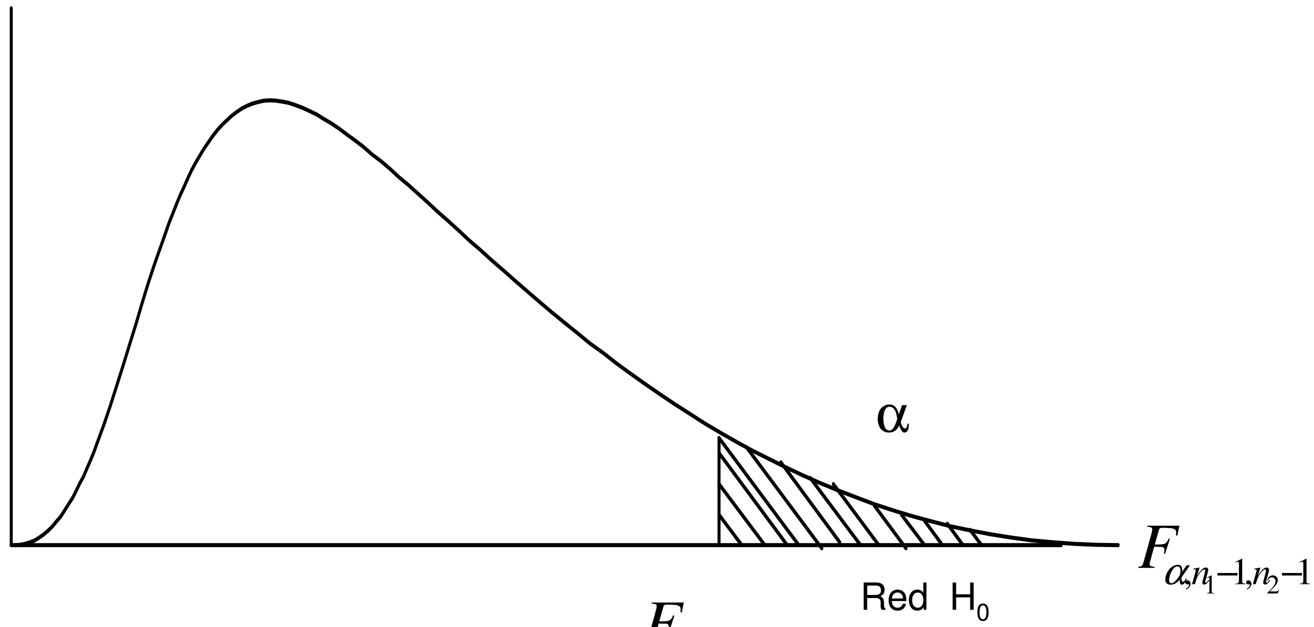
$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$



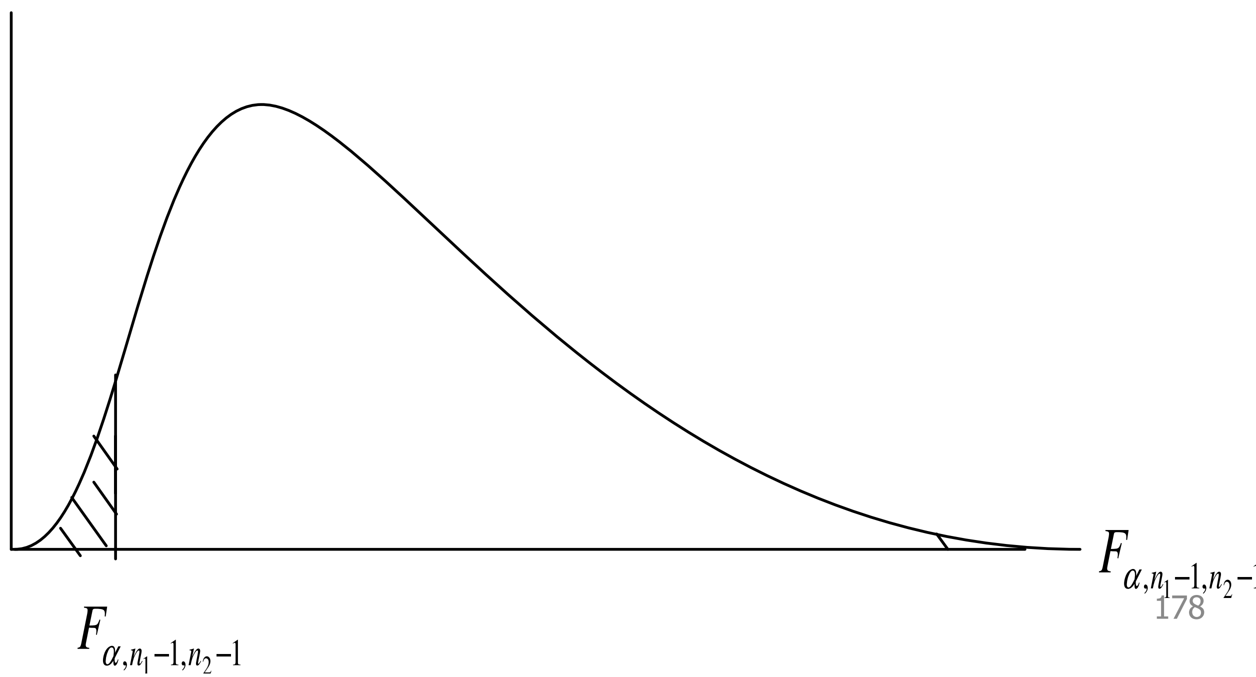
Sağ Kuyruk Testi

$$H_0 : \sigma_1^2 \leq \sigma_2^2$$
$$H_1 : \sigma_1^2 > \sigma_2^2$$



Sol Kuyruk Testi

$$H_0 : \sigma_1^2 \geq \sigma_2^2$$
$$H_1 : \sigma_1^2 < \sigma_2^2$$



$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ hipotezi altında test istatistiği;

$$F == \frac{\dfrac{\chi_1^2}{n_1 - 1}}{\dfrac{\chi_2^2}{n_2 - 1}} = \frac{\dfrac{s_1^2}{\sigma^2}}{\dfrac{s_2^2}{\sigma^2}} = \frac{s_1^2}{s_2^2}$$

$s_1^{\,2}$ iki örnek varyansının büyük olanıdır

ÖRNEK

Pazara yeni sürülmüş on yedi AAA dereceli sınai tahvilden oluşan rassal bir örneklemde vadelerin varyansı 123.35’dir. Onbir yeni CCC dereceli sınai tahvilden oluşan bağımsız bir rassal örneklemde vadelerin varyansı 8.02’dir. Bu iki tahvilin değişkenliklerinin eşit olup olmadığını test ediniz.

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

n_1-1=16 n_2-1=10 sd.

$\alpha = 0.02$ $\longrightarrow$ $F_{16,10,0.01} = 4.53$

$$n_1 = 17 \quad s_1^2 = 123.35$$

$$n_2 = 11 \quad s_2^2 = 8.02$$

$$F_{test} = \frac{s_1^2}{s_2^2} = \frac{123.35}{8.02} = 15.38$$

$F_{test} > F_{tab}$ H_0 RED

8. BAĞINTI (REGRESYON) VE KORELASYON (İLİŞKİ)

8.1 Korelasyon

$$\operatorname{cov}(x,y)=\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})$$

$$r_{xy}=\frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}$$

$$\sigma_x^2=\frac{1}{n}\sum(x_i-\bar{x})^2$$

$$\sigma_y^2=\frac{1}{n}\sum(y_i-\bar{y})^2$$

$\Rightarrow$

$$r_{xy}=\frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}$$

$$r_{xy}=\frac{\sum_i^n(x_i-\bar{x})(y_i-\bar{y})}{s_x\cdot s_y}=\sum_i^n\left(\frac{(x_i-\bar{x}}{s_x}\right)\left(\frac{y\quad-\bar{y}}{s_y}\right)$$

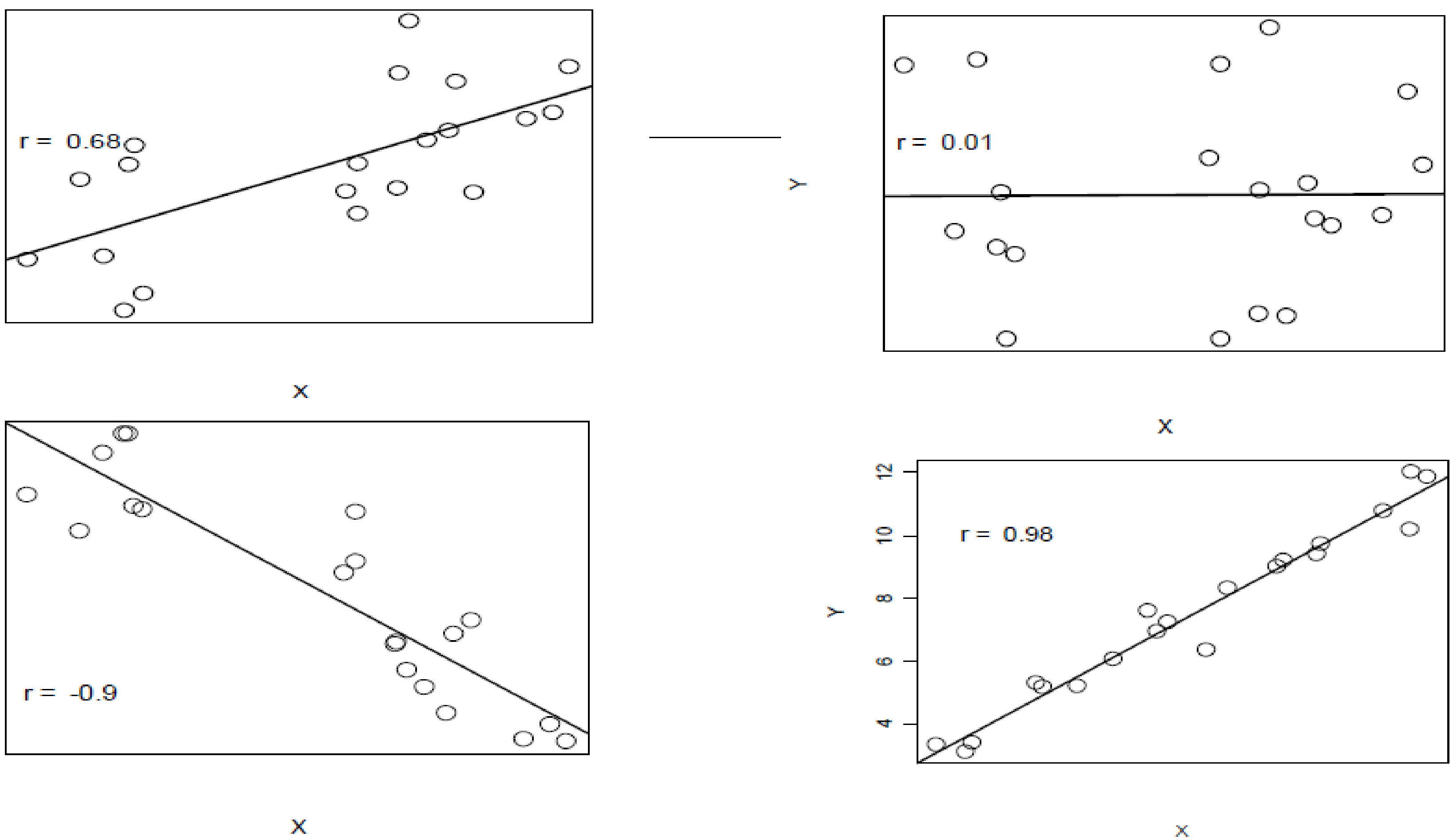
$$r_{xy}=\frac{1}{n}\sum_i^n Z_i^x\cdot Z_i^y$$

Parametrik Basit Korelasyon
Pearson Korelasyon Katsayısı

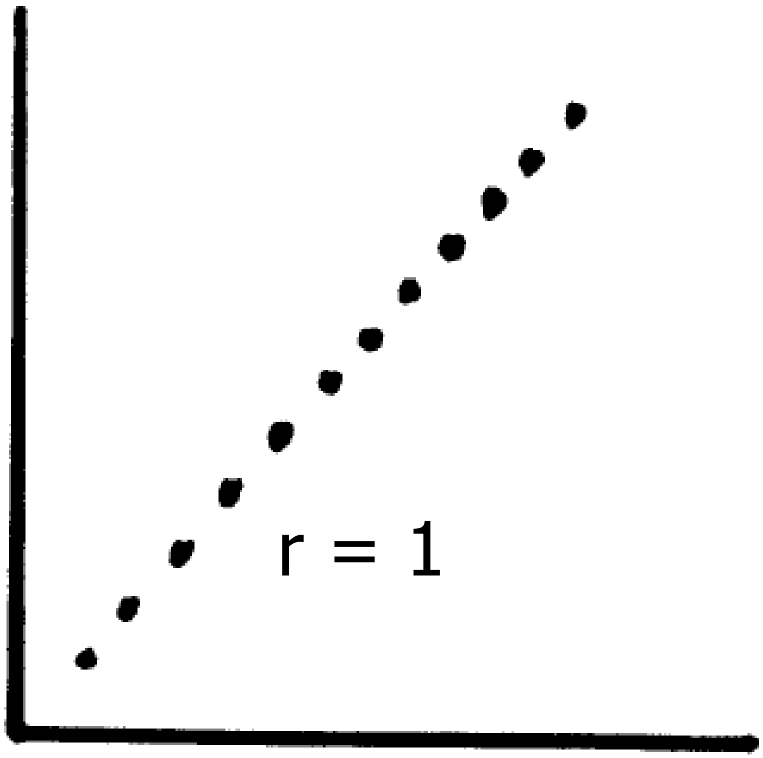
Korelasyon katsayısı iki özellik arasındaki ilişkinin anlamlılığını ölçer. Burada bahsedilecek olan korelasyon katsayısı sadece bir tane x değişkeni olduğunda x ile y arasındaki doğrusal ilişkinin derecesini vermektedir. +1 ile -1 arasında değerler alır örnekten hesaplanan istatistik değeri r ile gösterilir ve birimi yoktur.

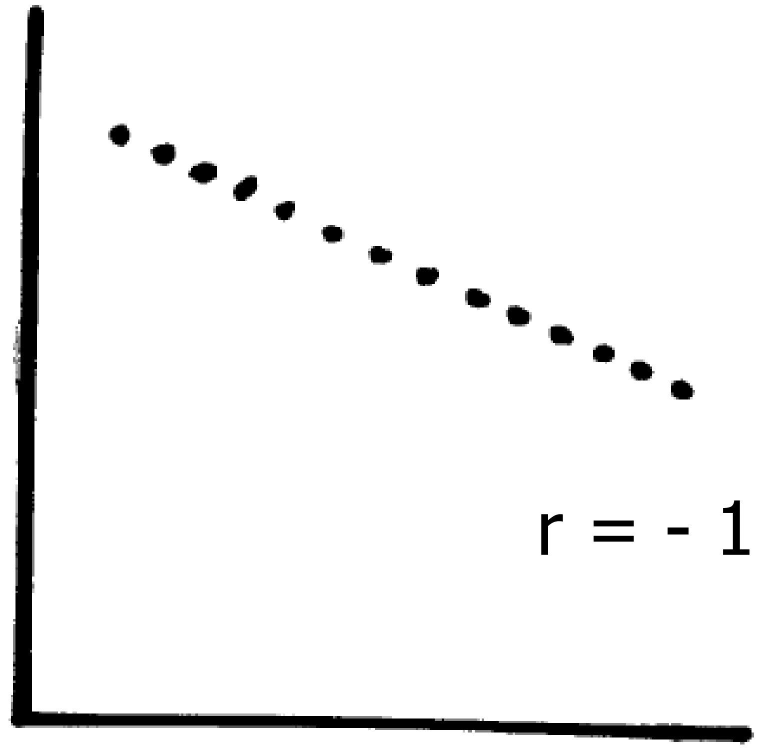
- $r > 0$ Pozitif ilişki
- $r < 0$ Negatif ilişki
- $r = 0$ Doğrusal ilişki yok

Saçılma Diyagramları
x-ekseni bağımsız değişken, düşey y-eksen bağımlı değişkendir.

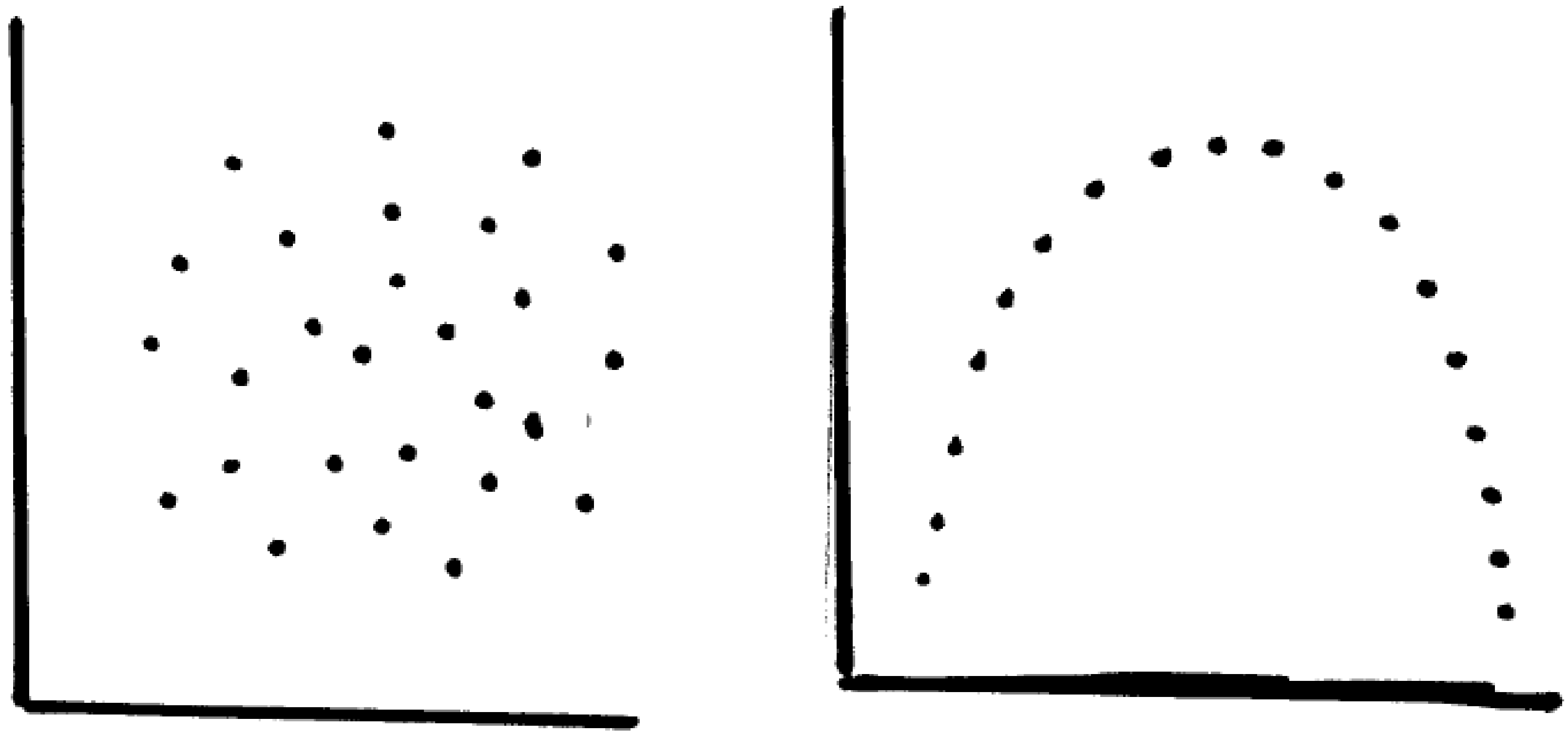


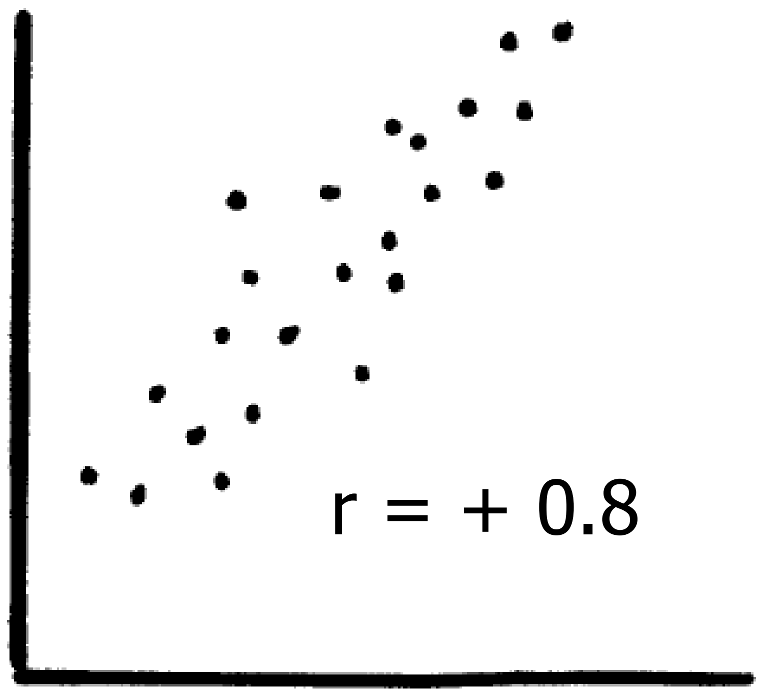
Örnekler

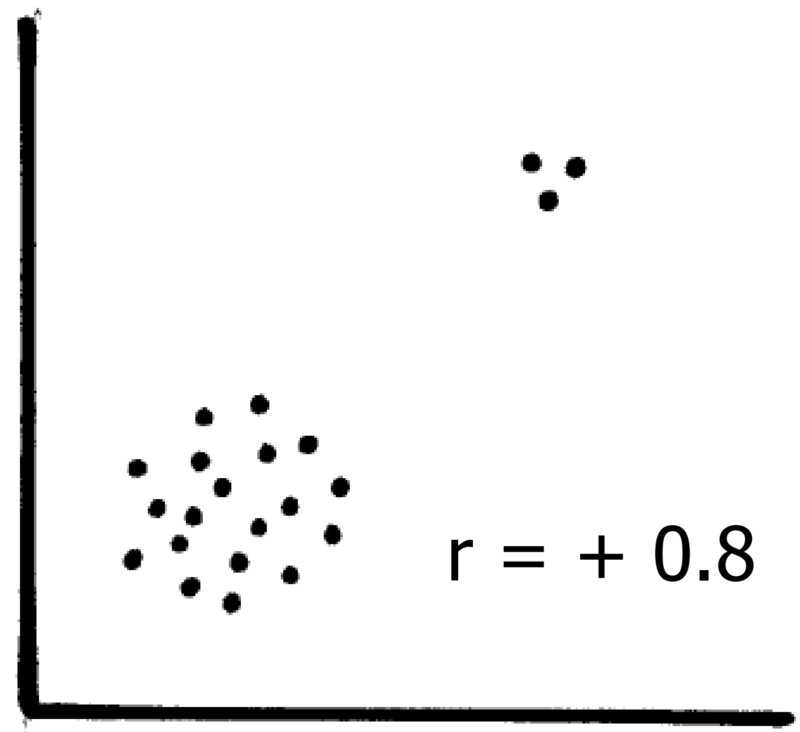




r = 0




r = + 0.8


r = + 0.8

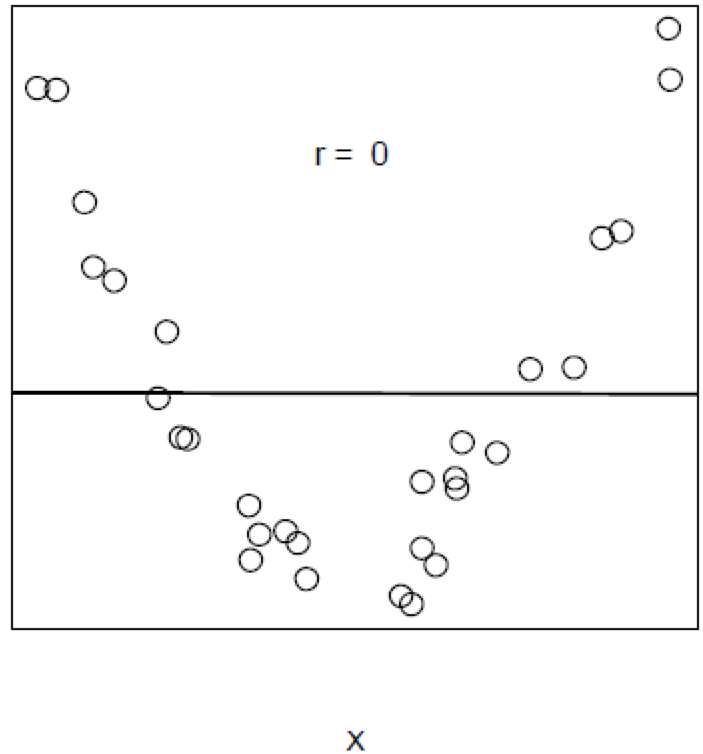
Korelasyon Katsayı Hesabının Uzun Formülü

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)\left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right)}}$$

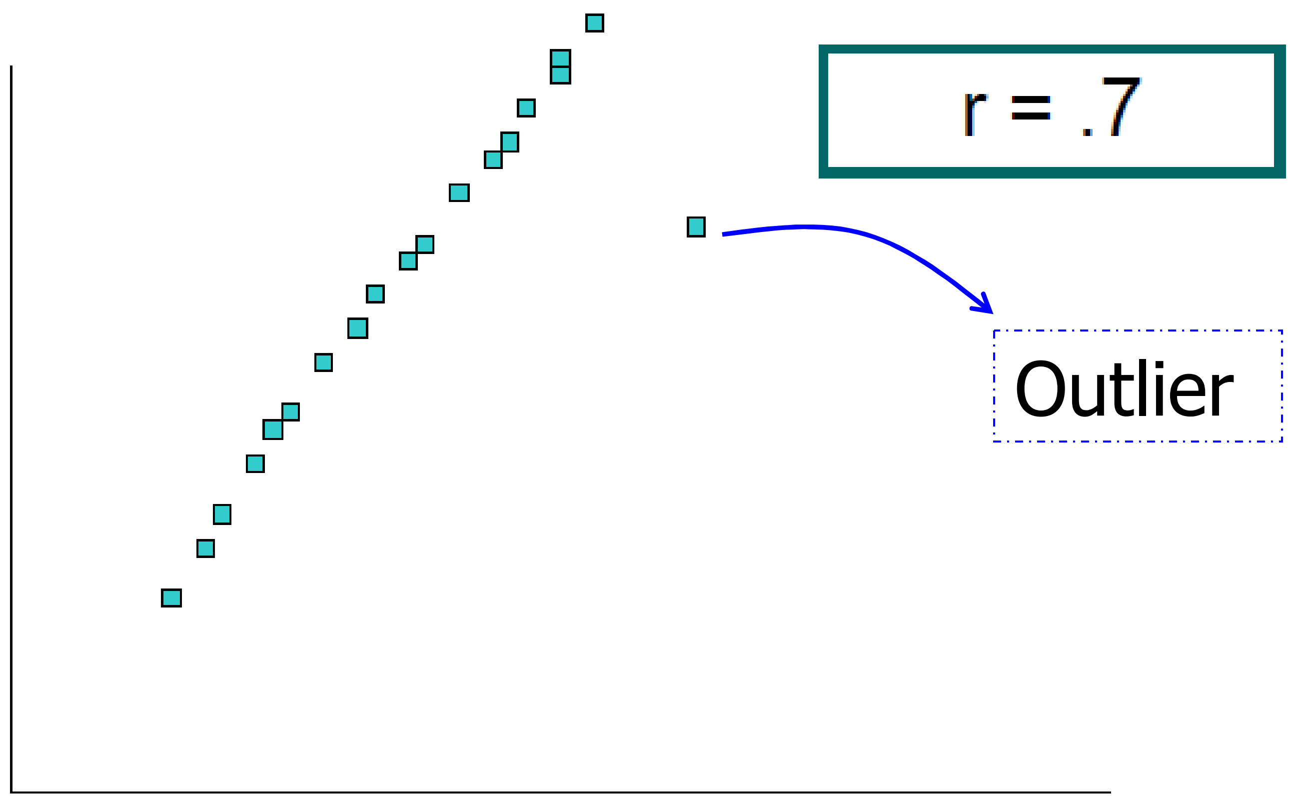
Hesaplama Tablosu

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
X_1	Y_1	X_1^2	Y_1^2	X_1Y_1
X_2	Y_2	X_2^2	Y_2^2	X_2Y_2
:	:	:	:	:
X_n	Y_n	X_n^2	Y_n^2	X_nY_n
ΣX_i	ΣY_i	ΣX_i^2	ΣY_i^2	ΣX_iY_i

Dikkat: r = 0 olması 2 değişkenin ilişkisi yoktur anlamına gelmez. Aralarında lineer ilişki yoktur demektir. Yandaki şekilde görüldüğü gibi non-lineer bir ilişki olabilir.



Aykırı (outlier) değerler korelasyon katsayısını etkileyebilir



Korelasyon Katsayısı r'nin anlamlılığının testi.

Var olduğu düşünülen iki dizideki gözlem değerlerinin bir ana kütleden gelen n sayıdaki gözlem değerleri olduğu düşünüldüğünde, ana kütle ile ilgili bir teorik korelasyon katsayısı (ρ) vardır ve örnek korelasyon katsayısı (r) ile tahmin edilmeye çalışılmaktadır denilebilmektedir.

Hipotez;

$$H_0 : \rho = 0$$

$$H_1 : \rho > 0$$

Korelasyon Katsayısının anlamlılığı t testi kullanılarak, s.d. = n-2 göre test edilebilir.

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

$t_{\text{hesap}} > t_{(1-\alpha, n-2)}$ Test RED edilir

8.2 Basit Doğrusal Regresyon

Bağımsız Değişken (Independent Variable)

Genellikle x ile gösterilir. Başka bir değişken tarafından etkilenmeyen ama y'nin nedeni olan yada onu etkilediği düşünülen (açıklayıcı) değişkendir.

Bağımlı Değişken (Dependent Variable)

Genellikle y ile gösterilir. x değişkenine bağlı olarak değişebilen yada ondan etkilenen **(açıklanan)** değişkendir.

- Bağımlı değişken sayısı tekdir. Ancak bağımsız değişken sayısı birden fazla olabilir. Eğer tek bağımsız değişken var ise “Basit Doğrusal Regresyon” iki ve daha fazla bağımsız değişken var ise “Çoklu Doğrusal Regresyon” adı verilmektedir.
- ***Bu derste sadece “Basit Doğrusal Regresyon Analizi” incelenecektir.***

- Regresyon Analizinde, değişkenler arasındaki ilişkiyi fonksiyonel olarak açıklamak ve bu ilişkiyi bir modelle tanımlayabilmek amaçlanmaktadır.
- Bir kitlede gözlenen X ve Y değişkenleri arasındaki doğrusal ilişki aşağıdaki **"Doğrusal Regresyon Modeli"** ile verilebilir;

$$Y=\beta_0+ \beta_1 X+\varepsilon$$

Burada;

X: Bağımsız (Açıklayıcı) Değişken

Y: Bağımlı (Açıklanan;Etkilenen;Cevap) Değişken

β_0: X=0 olduğunda bağımlı değişkenin alacağı değer (kesim noktası)

β_1: Regresyon Katsayısı

ε : Hata terimi (Ortalaması = 0 ve Varyansı = σ^2'dir)

Regresyon Katsayısı (β_1) :

Bağımsız değişkendeki bir birimlik değişimin, bağımlı değişkendeki yaratacağı ortalama değişimi göstermektedir.

Hata terimi (ε):

Her bir gözlem çiftindeki bağımlı değişkene ilişkin gerçek değer ile modelden tahmin edilen değer arasındaki farktır.

$$\varepsilon_i = \underbrace{(\beta_0 + \beta_1 X)}_{\hat{Y}_i} - Y_i$$

Tanımlanan Regresyon Modeli

1. Kitleden seçilen n gözlemli örnek için;

$$\hat{Y} = b_o + b_1 X$$

2. Yukarıdaki Doğrusal Regresyon Modeli Gözlemler için;

$$\hat{y}_i = b_0 + b_1 x_i + e_i;\ i = 1,2,\ldots,n$$

Kesim Noktası ve Regresyon Katsayısının Tahmin Yöntemi

- Doğru ve güvenilir bir regresyon modelinde amaç, gerçek gözlem değeri ile tahmin değeri arasında fark olmaması yada farkın minimum olmasıdır. Bunun için çeşitli tahmin yöntemleri geliştirilmiştir. Bu yöntemlerden biri **"En Küçük Kareler"** kriteridir.

$$\sum_{i=1}^{n} e_i^2 \quad_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Bu farkın en küçük olması amaçlanır

En Küçük Kareler Yöntemi ile Bulunan Tahminler Regresyon Katsayıları

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \left(n \bar{x} \bar{y}\right)}{\sum_{i=1}^{n} x_i^2 - \left(n \bar{x}^2\right)}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

- Değişkenler birlikte artıyor artıyor yada birlikte azalıyor ise **“b_1 pozitif değerli”**dir.

- Değişkenlerden biri artarken diğeri azalıyor ise **“b_1 negatif değerli”**dir.

Regresyon Katsayısının Önem Kontrolü

X bağımsız değişkeni ile Y bağımlı değişkeni arasında doğrusal bir ilişkinin varlığı, her bir bireyin / birimin x_i ve y_i değerlerinin koordinat düzlemi üzerinde oluşturdukları noktaların dağılımına bakılarak tahmin edilebilir. Ancak, bu tahminin tutarlı olup olmadığının araştırılması gerekir. Bunun için, regresyon katsayısının önem kontrolü, doğrusallıktan ayrılışın önem kontrolü yapılır.

Önem Kontrolü Yapabilmek için Kullanılacak Eşitlikler

X ortalamadan ayrılış kareler toplamı (XOAKT):

$$XOAKT = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Serbestlik derecesi = (n-1)

Y ortalamadan ayrılış kareler toplamı (YOAKT):

$$YOAKT = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

Serbestlik derecesi = (n-1)

XY Çarpımlar Toplamı (XYÇT):

$$XYÇT = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - (n\,\bar{\mathrm{x}}\,\bar{\mathrm{y}})$$

Regresyon Kareler Toplamı (RKT):

$$RKT = \sum_{i=1}^{n} (\hat{y}_i - \bar{Y})^2 = \frac{(XYÇT)^2}{XOAKT} = (b_1 XYÇT)$$

RKT'ye ilişkin serbestlik derecesi = 1'dir.

Hata Kareler Toplamı (HKT)

- Hata yada Artık Kareler Toplamı da denir -

$$HKT = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = YOAKT - RKT$$

HKT'na ilişkin serbestlik derecesi = (n – 2)'dir.

Regresyon Analizi için Varyans Analizi (ANOVA) Tablosu

Varyasyon (Değişim) Kaynağı	Serbestlik Derecesi (s.d.)	Kareler Toplamı (KT)	Kareler Ortalaması (KO)	F Hesap İstatistiği
Regresyon	1	RKT	RKO=RKT / 1	RKO /RKO
Hata (Artık)	(n-2)	HKT	HKO= HKT/(n-2)	
Toplam	(n-1)	YOAKT		

Basit Doğrusal Regresyon Analizinde İki Hipotez Test Edilir:

Doğrusallıktan Ayrılışın Önem Kontrolü

1. Hipotez Kurulur.

H_o: Gözlenen Noktaların Regresyon Doğrusuna Uyumu Önemsizdir (Model geçersizdir)

H_1 : Gözlenen Noktalar Regresyon Doğrusu ile tanımlanabilir (Model Geçerlidir)

2. Bu hipotezi test etmek için RKO ve RAKO varyanslarının oranı uygun test istatistiğidir. İki varyansın oranı F dağılımına yakınsayacağı için kullanılacak test dağılımı F'dir.

F_H=(RKO / HKO) değeri hesaplanır.

3. 1 ve (n-2) serbestlik dereceli ve belirlenen α anlamlılık düzeyinde $F_{(1;n-2;\alpha)}$ tablo değeri bulunur.

Eğer F_H=(RKO / HKO) < $F_{(1;n-2;\alpha)}$ ise

Ho Hpotezi KABUL Edilir.

İkinci Hipotez Testi
Regresyon Katsayısının Önem Kontrolü

1. Hipotez Kurulur

H_o: Regresyon Katsayısı Önemsizdir ($\beta_1 = 0$)

H_1: Regresyon Katsayısı Önemlidir ($\beta_1 \neq 0$)

Burada, regresyon katsayısının önemsiz olması demek; örneklemin çekildiği kitlede, bağımsız değişkende bir birimlik değişimin, bağımlı değişkende değişiklik yaratamayacağı anlamına gelir.

2. Test istatistiği hesaplanır;

$$t_h = \frac{b_1 - (\beta_1 = 0)}{S_{b1}}$$

$$S_{b1} = \sqrt{\frac{HKO}{XOAKT}}$$

3. Serbestlik derecesi (n-2) ve α anlamlılık düzeyinde, $t_{(n-2;\, \alpha)}$ tablo değeri bulunur.

Eğer $t_h > t_{(n-2;\, \alpha)}$ **ise Ho Hipotezi RED** edilir.

4. Regresyon katsayısının önemli olup olmadığına karar verilir.

Basit Doğrusal Regresyon Analizinde Özel Durum

Basit Doğrusal regresyonda tek bir **bağımsız değişken** olması nedeniyle t-dağılımı ve F-dağılımı arasında aşağıdaki matematiksel eşitlik söz konusudur :

$$t_h^2 = F_h$$

Belirtme (Determinasyon) Katsayısı) R^2

Determinasyon Katsayısı, regresyon analizinde önemlidir ve aşağıdaki gibi hesaplanır ;

$$R^2 = \frac{RKT}{YOAKT}$$

$$0 \leq R^2 \leq 1$$

Determinasyon Katsayısı 1 yakın bulunur ise, bağımlı değişkendeki değişimin büyük bir kısmı bağımsız değişken tarafından açıklanabilir yorumu yapılabilmektedir.

Örnek:

12-14 yaş grubu çocukların boy uzunluğu ile kulaç uzunluğu arasında ilişki olup olmadığını incelemek için 10 çocuk üzerinde bir araştırma planlanmıştır. Her çocuğun boy uzunluğu ile birlikte duvara yaslandırılarak ve kolları açtırılarak her iki ellerinin orta parmakları arasındaki mesafe (kulaç uzunlukları) ölçülmüştür.

- Burada amaç; çocukların kulaç uzunluğundan boy uzunluklarını tahmin etmek için bir model oluşturmaktır.
- Bu durumda;

Bağımlı Değişken (y): Boy uzunluğu

Bağımsız Değişken (x): Kulaç uzunluğu

Çocuk No	Boy uzunluğu (cm)	Kulaç uzunluğu (cm)
1	165	162
2	161	163
3	156	158
4	158	156
5	163	161
6	166	166
7	154	153
8	156	154
9	161	161
10	159	157

Test İstatistiklerini Hesaplamak İçin Gerekli İşlemler

$$\sum_{i=1}^{10} y_i = 1599$$

$$\sum_{i=1}^{10} x_i^2 = 253285$$

$$\sum_{i=1}^{10} x_i = 1591$$

$$\sum_{i=1}^{10} y_i^2 = 255825$$

$$\sum_{i=1}^{10} x_i y_i = 254538$$

$$\bar{y} = \frac{1599}{10} = 159.9$$

$$\bar{x} = \frac{1591}{10} = 159.1$$

$$XOAKT = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 253285 - (10 \cdot 159.1^2) = 156.9$$

$$YOAKT = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 255825 - (10 \cdot 159.9^2) = 144.9$$

$$XYÇT = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - (n\,\bar{\mathrm{x}}\,\bar{\mathrm{y}})$$

$$= 254538 - (10 \cdot 159.1 \cdot 159.9) = 137.1$$

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - (n\,\bar{x}\,\bar{y})}{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_{\mathrm{i}}^2 - (n\,\bar{\mathrm{x}}^2)} = \frac{137.1}{156.9} = 0.874$$

$$b_0 = \bar{y} - b_1\bar{x} = 159.9 - (0.874 \cdot 159.1) = 20.847$$

Boy Uzunluğu =20.874 + 0.874(kulaç uzunluğu)

Burada, kulaç uzunluğu 1 birim arttığında boy uzunluğunun ortalama 0.874 birim arttığını görmekteyiz.

Şimdi acaba bu regresyon katsayısı istatistiksel açıdan önemli midir? Sorusuna cevap vermemiz gerekiyor.

H_o: Regresyon Katsayısı Önemsizdir ($\beta_1=0$)

H_1: Regresyon Katsayısı Önemlidir ($\beta_1 \neq 0$)

$$RKT = \sum_{i=1}^{n} (\hat{y}_i - \overline{Y})^2 = \frac{(XYÇT)^2}{XOAKT} = (b_1 XYÇT) = (0.874 \cdot 137.1) = 119.8254$$

$$HKT = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = YOAKT - RKT = 144.9 - 119.83 = 25.07$$

$$RKO = \frac{RKT}{1} = \frac{119.83}{1} = 119.83$$

$$HKO = \frac{HKT}{n-2} = \frac{25.07}{8} = 3.13$$

$$S_{b1} = \sqrt{\frac{RHKO}{XOAKT}} = \sqrt{\frac{3.13}{156.9}} = 0.141$$

$$t_h = \frac{b_1 - (\beta_1 = 0)}{S_{b1}} = \frac{0.874 - 0}{0.141} = 6.19$$

t_h=6.29 > $t_{(8;\ 0.05)=}$2.306
Ho Hipotezi RED edilir

Yorum: %95 Güven seviyesinde, regresyon katsayısının sıfırdan farklı olduğunu, yani bulunan regresyon katsayısının istatistiksel açıdan önemli olduğunu söyleyebiliriz

Şimdi Modelin Geçerliliğini Test Edelim

H_o: Gözlenen Noktaların Regresyon Doğrusuna Uyumu Önemsizdir (Model geçersizdir)

H_1 : Gözlenen Noktalar Regresyon Doğrusu ile tanımlanabilir (Model Geçerlidir)

Varyasyon (Değişim) Kaynağı	Serb.Der. (sd)	Kareler Toplamı (KT)	Kareler Ortalaması (KO)	F Hesap İstatistiği
Regresyon	1	119.83 RKT	119.83 RKO	**38.28**
Hata (Artık)	8	25.07 HKT	3.13 HKO	
Toplam	9	144.9		

$R^2=119.83/144.9=0.83$

$F_H=(RKO / HKO) > F(1;n\text{-}2;\alpha)$ ise
Ho Hpotezi RED Edilir.
$F_H=38.28 > F_{(1;8;0.05)}=5.32$ olduğu için Ho hipotezi red edilir.

$t_h^2=(6.19)^2=38.3=F_h$ eşitliğinin sağlandığını da görebiliyoruz.

SONUÇ: %95 güven seviyesinde kulaç uzunluğundan boy uzunluğunu tahmin etmek için bulduğumuz modelin geçerli olduğunu söyleyebiliriz. Boy Uzunluğundaki değişimin %83'ünün (R^2) kulaç uzunluğu tarafından açıklanabildiğini, geriye kalan %17'lik kısım için başka değişkenlere ihtiyaç duyulduğunu söyleyebiliriz.

ÖNEMLİ NOT:

Bilimsel çalışmalarda herhangi bir modelleme çalışmasında genellikle çok değişkenli çalışılır. Burada anlatılan regresyon analizinin sadece tek değişkenli olduğu ve analizlerin burada bitmeyip modelin uygunluğuna ilişkin çok ileri yöntemler olduğu unutulmamalıdır.