

(1) FORMÜLLE

VARIANCE

$$\sigma^2 = \frac{\sum (x - \mu)^2}{n}$$

x : observation

μ : mean

n : sample size

`np.var(array)`

STANDARD DEVIATION

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

Maryansın
Karekökü

`np.std(array)`

Veya

`np.sqrt(\sigma^2)`

STANDARD ERROR

$$SEM = \frac{std}{\sqrt{n}}$$

`df[column_name].std() / np.sqrt(len(column_name))`

STANDARD SCORE - Z TEST

$$z = \frac{x - \text{mean}}{\text{std}}$$

x 'den Olasılığı:



`stats.norm.cdf(z)`

Olasılıkta x 'e:



`stats.norm.ppf(q)`

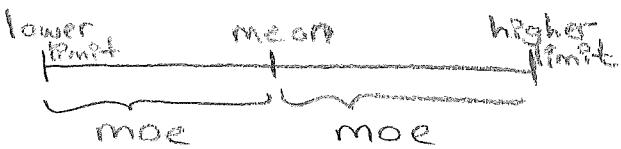


`1 - stats.norm.cdf(z)`

MARGIN OF ERROR

$$moe = 2 \cdot \frac{std}{\sqrt{n}} \cdot sem$$

confidence interval = c_i



$$moe = 2 \cdot sem$$

stats.norm.interval(alpha, loc=0, scale=1)

c_i mean sem

$df.column_name.mean() + moe \rightarrow$ Higher Limit

$df.column_name.mean() - moe \rightarrow$ Lower Limit

T TEST

"What is degrees of freedom?"

(dof)

$dof = n - 1$

X' der Olsiliga:

$stats.t.cdf(X, dof) \Leftrightarrow p\text{-value}$

Olsilikton X' e:

$stats.t.ppf(q, dof)$

olsilk

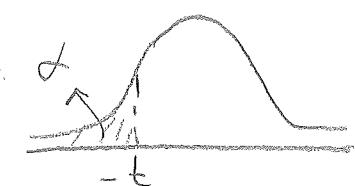
P-Value

(2)

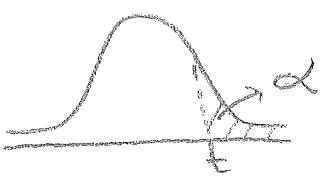
Ne kadar küçükse H_0' , reddetmek için o kadar çok kanıt toplar, H_a' , kabul ederiz.

α = Significance Level

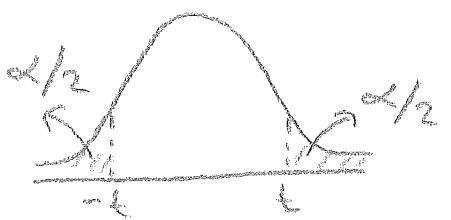
$P\text{-value} < \alpha$ ise H_0 reddedilir.



(Left tail test)



(Right tail test)



(Two tail test)

DEPENDENT SAMPLE T TEST

One group \rightarrow Before
 \downarrow After

$t\text{-dep} = \text{stats.ttest_rel}(x_1, x_2, \text{alternative}=\text{"two-sided"})$

✓
array-like

$t\text{-dep. statistic} \rightarrow t$ degerini verir.

$t\text{-dep. pvalue} \rightarrow pvalue$ degerini verir.

$\sim \sim \sim \sim \sim$
P-value iin 2. gol:

$$p\text{-value} = 2 * (1 - \text{stats.t.cdf}(t, dof))$$

INDEPENDENT SAMPLE T TEST

Independent 2 groups.

```
t_ind = stats.ttest_ind(x1, x2, equal_var=True,  
array-like) alternative="two-sided")
```

Independent + Test 'te önce varyanslar esit mi değil mi diye bakılır.



Equal Variances NOT assumed



Equal Variances assumed

Bunların formüller var ama `scipy Levene Test` ile de yapılabilir.

Levene Test: `stats.levene(array1, array2)`

Burdan bir pvalue değeri gelecek. Eğer $pvalue < \alpha$ ise "NOT equal" formülü kullanılır ve sayfa basında yardımımıza koddı `equal_var=False` olarak alınır.

! Levene test yapıldıktan sonra t-test hesabına gelir.

One-way ANOVA

(3)

3 veya daha fazla grubun ortalaması karşılaştırıldığında F-testi kullanılır.

$$f\text{-test} = \text{stats.f_oneway}(x_1, x_2, x_3, \dots)$$

array-lık

$f\text{-test.statistic}$ \Rightarrow f değerini verir.

$f\text{-test.pvalue}$ \Rightarrow p-value değerini verir.

veya p-value hesabi için:

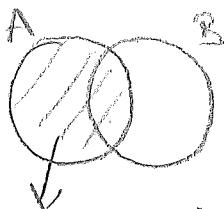
$$p\text{-value} = 1 - \text{stats.f.cdf}(f, df_1, df_2)$$

$$df_1 = \text{grup sayıısı} - 1$$

$$df_2 = \text{"örnek sayıısı} - \text{grup sayıısı}$$

CONDITIONAL PROBABILITY

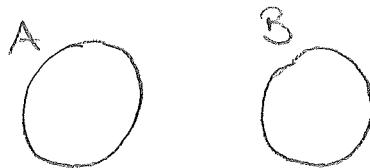
Dependent



A olmuş, B 'nin olma
İhtimali

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

Independent



$$P(A|B) = P(A)$$

$$P(B|A) = P(B)$$

BAYES THEOREM

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

BINOMIAL DISTRIBUTION

Two possible outcomes Success
 Failure

stats. binom. pmf(x, n, p) 2 zar atıldığında 2'nci üstüne
gelen sayıların toplamının n
success trial probability olma olasılığı

stats. binom. cdf(x, n, p) Üst yere gelen sayılar toplamının
 n veya n 'den büyük olma olasılığı
(Cumulative Toplam)

POISSON EXPERIMENT

Nadir meydana gelen bağımsız olaylar

stats. poisson. pmf(x, λ)

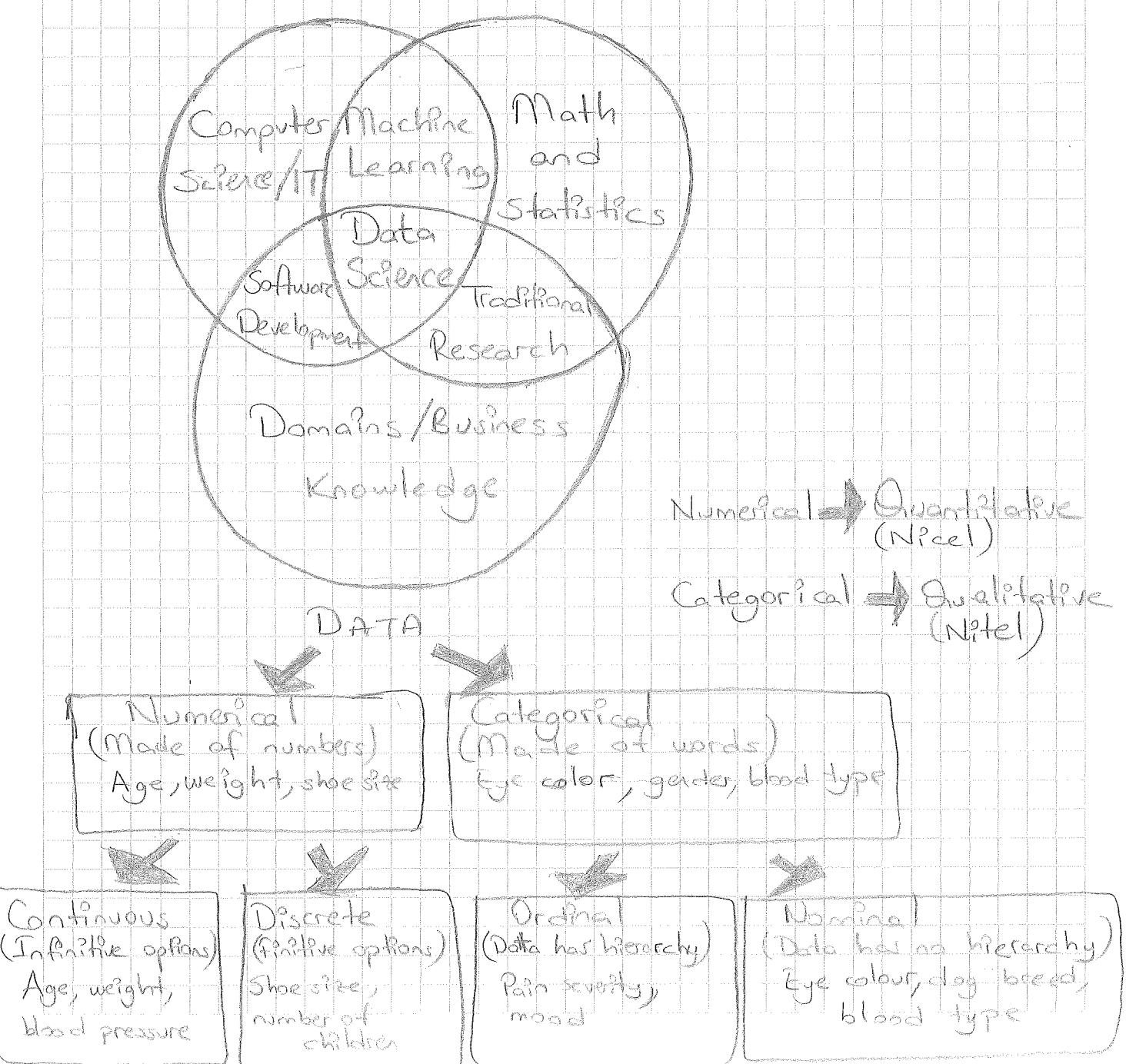
x : Aranan olay sayısı

stats. poisson. cdf(x, λ)

λ : Başarı ortalaması

* Statistik is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data.

* Statistic is a branch of mathematics dealing with the collection, analysis, interpretation and presentation of masses of numerical data.



NUMERICAL DATA

Discrete Data

It contains only **finite** (saylı) values.

The number of students in a school } Sonuç olamaz. Örneğin

The number of building in a city } 3m-30k arasında olamaz.

If its possible values form a set of separate numbers, such as 0, 1, 2, 3, ... \Rightarrow it's a discrete.

Continuous Data (Sürekli Ver)

It can have an **infinitive** number of values between any two values.

Height or weight of a person

House price

The time assigned to a special task

Sonuç sayıda

değer sahip

olabilir

10, 11, 10.6655

olabilir.

If its possible values form an **interval** \Rightarrow it's a

continuous

CATEGORICAL DATA

Subject:

Date:

(*) Female / Male genders

Automatic / Semi-automatic / manual gearboxes

} Nominal data

You can assign numbers to categorical data, for example you can use 0 for male and 1 for female. But even in this case, numbers do not have actual numerical meaning. They represent the categories.

Ordinal Data

(*) You might rank the taste of the meal as "1" for bad, "2" for average and "3" for good. The survey data is a good example of ordinal data. Here we use numbers but they do not have actual numerical meaning.

Ordinal Data \Rightarrow It is a categorical one for which the categories are ordered from low to high in some sense.

Nominal Data \Rightarrow It is a categorical data that does not have a natural order or ranking

1 Statistic

It is the study of how best to collect, analyse and draw conclusions from data.

Data

Data are characteristics or information, usually numerical, that are collected through observation.

Understanding different types of data crucial prerequisite for doing Exploratory Data Analysis (EDA) because you can only use certain statistical measurement for certain types of data.

Variable

It is any characteristic observed in a study.

1 Level of Measurements

- (1) Nominal } Kategorik veriler
 } Sıra Kullanılmaz
- (2) Ordinal }
- (3) Interval } Numerik veriler
 } Sıra Kullanılabilir
- (4) Ratio }

Here the important point is that there is a hierarchy between these levels, and this hierarchy increases from nominal to ratio level.

① Nominal Level of Measurement,

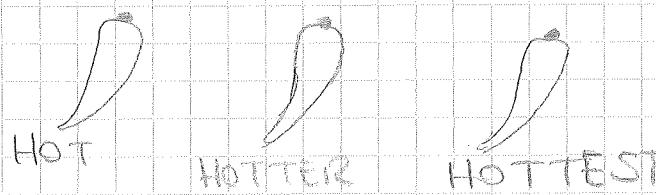
It simply names something without assigning an order.

for male \rightarrow 0 In this level, the numbers in the variable are used only to classify
for female \rightarrow 1 the data.

Red / Blue / Yellow Examples
Success / fail

② Ordinal Level of Measurement,

In this level, the attributes are ordered, however, distances between attributes have no meaning.


HOT HOTTER HOTTEST

Degrees are not parallel to quantitatives!

Small / Medium / Large \rightarrow Example for: Ordinal

③ Interval Level of Measurement

Degerler arasındaki anlam yorumlanabilir.
(Ordinal Level'in aksine)

Mesela sıcaklık ölçerken 40°C ile 60°C arasındaki mesafe aynıdır.

Ancak ORAN interval level'da bir anlam ifade etmez.

- ✓ Nitelikler arası mesafeli anlamı var.
- ✗ Oran bir anlam ifade etmiyor.

④ Ratio Level of Measurement

Gözlemler interval level ile aynı aralıklara sahip olacak şekilde sıfır değeri de sahiptir.

Length

Height

Weight

Cell phone charge capacity

} Examples

- ✓ Interval'dan farklı olarak oranlar anlamlıdır. Bir cisimin yüksekliğinin 2 katı olduğunu söylemek mümkün.
- ✗ Gökku sıfır onu anlamaz.

- ✓ Oranlar mantıktır.

- ✗ Sıfır notlaşılmaz.

Subject:

RATIO

Named
+
Ordered
+
Proportionate Interval between variables
+
Can accommodate Absolute Zero

INTERVAL

Named
+
Ordered
+

Proportionate Interval between variables

ORDINAL

Named
+

Ordered Variables

NOMINAL

Named Variables

Nominal = Categorical = Qualitative

→ Sex (0, 1)
Color (1, 2, 3)

Ortalama hesaplanamazdır.

%'da olarak verilebilir.

Ordinal → Rank

Satisfaction

Fondness

Meski bir tarafe daha

amaascalılar eşittir.

olmazdır.

%'da olarak verilebilir.

Mesela insanların davranışlarıyla ilgili araştırmalarde ort. kullanılır.

Interval / Ratio = Scale = Quantitative = Parametric

Discrete

Continuous

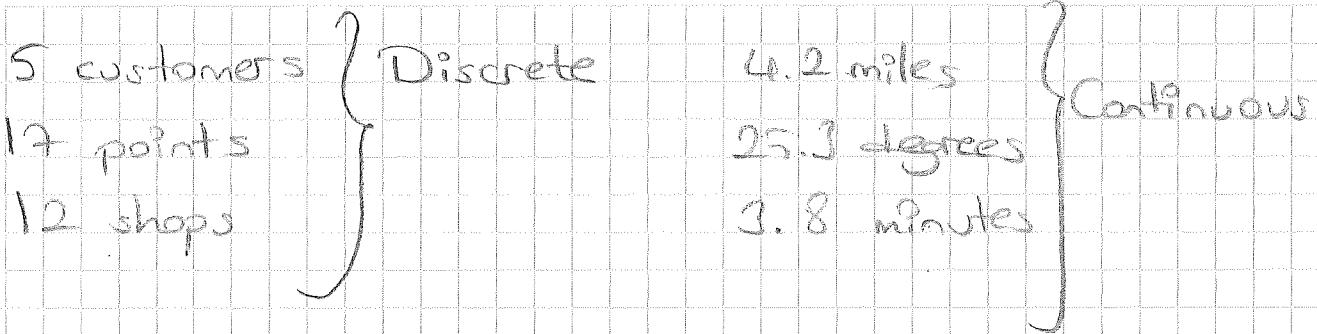
Ağırlık

Yas

Başka gibi sınıflandırılan değil de
“başka” degerler için.

Standard Deviation
hesaplanabilir.

Mean
Median



NOMINAL \Rightarrow En sadece grafik sütunları kullanılır.
 Ana nepsini kullanılır.

ORDINAL \Rightarrow Dairesel grafikte gösterilememeli.
 Sütun veya cubuk grafikinde gösterilebilir.

\Rightarrow Cubuk grafisi veya histogramda gösterilebilir.

Cizgi grafisi ve kutu grafisi da kullanılır.

Type of chocolate \rightarrow Nominal Data

Satisfaction and Likelihood \rightarrow Ordinal Level Data

Age, number of chocolate bars \rightarrow Interval / Ratio data

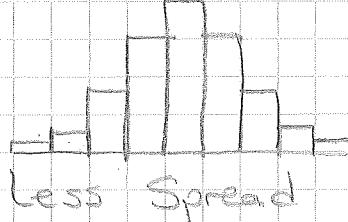
Data Patterns in Statistics

1. Center

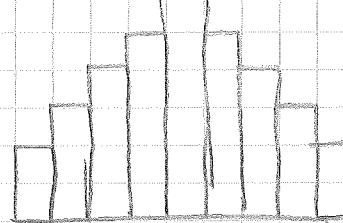


Half Half

2. Spread

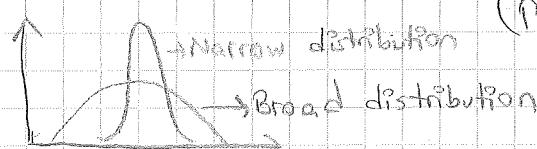


Less Spread



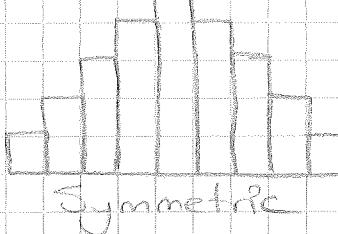
Wide spread

The spread of a distribution refers to the variation of the data.
(Max, min)



3. Shape

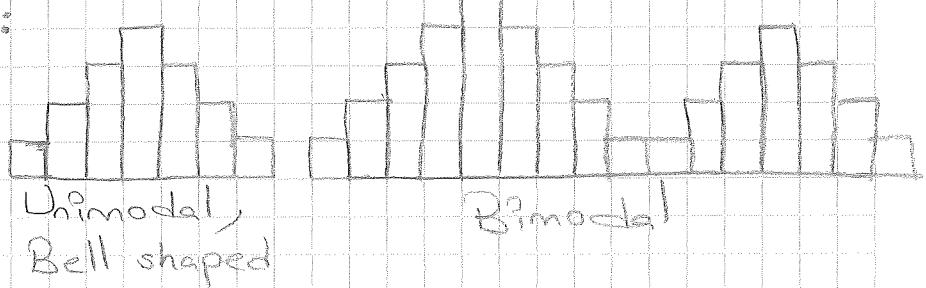
a) Symmetry



Symmetric

b) Number of Peaks:

Bic veya birden fazla tipe aittir.
Yerle.

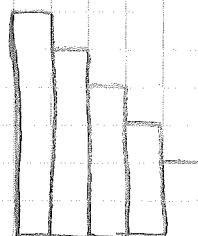


Subject :

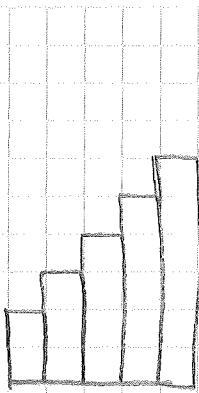
Date :

↳ Unimodal

c) Skewness :

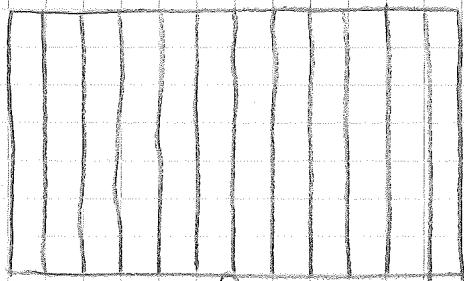


Skewed Right



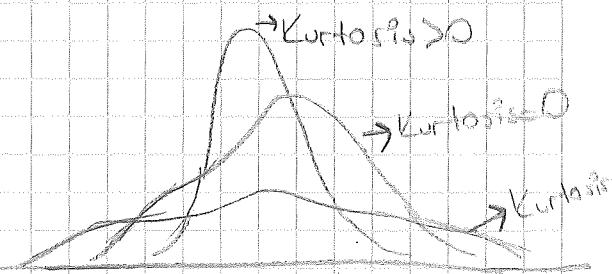
Skewed Left

d) Uniform :



Uniform
(Equally spread)

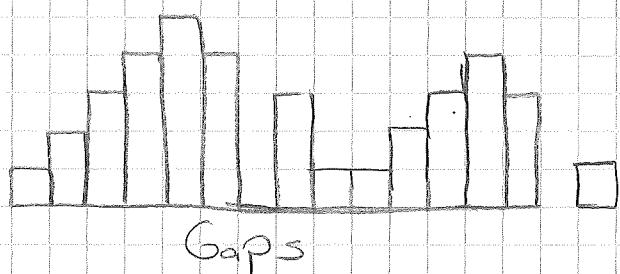
e) Kurtosis : (Bassliklik)



Unusual Features

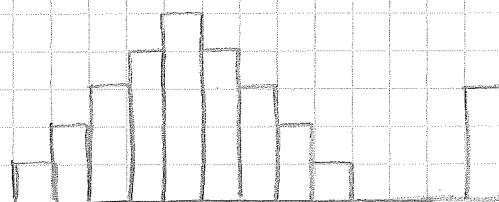
↳ Boleklik

a) Gaps :



Gaps

b) Outliers : (Mean'ı ve median'ı etkeler)



FREQUENCY TABLE, → Siklik tablosu

A frequency table is a listing of possible values for a variable, together with the number of observations for each value.

Ogeleri listeleyelim

Kac kez olduguunu gosterir.

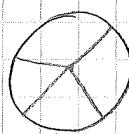
Example :

Developer Type	Frequency	Relative Frequency
Front-end Developer	25	0.25
Backend Developer	15	0.15
Full-Stack Developer	20	0.20
Data Scientist	40	0.40

A category or a numerical value \rightarrow Relative Frequency
 Total number of data

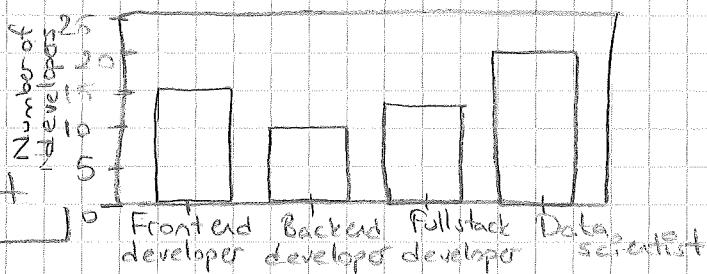
Kategori
verilerin
değerlerini
ideali

① Pie Chart



Kategorik veriler için uygun

- * Az sayıdaki (Kategoriler) görsütlendirmek için etkilidir.
- * Çok sayıda kategori için önebilmez.
- * İki farklı anket veya deneyin sonuçlarını karşılaştırma için önebilmez.
- * Dilimler küçükse anlaşılır olmasının renklere veya oklara dayanırmalıdır.



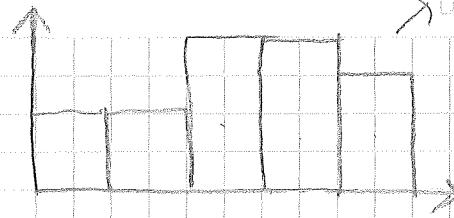
② Bar Chart

- * X ve Y ekseninden oluşur.
- * Y eksenli gözlemlerin yerlerinden eğrilede her kategori deki gözlem sayısını gösterir. (Pie chart gibi)
- * Cubuklar yatay veya dikey çizilebilir.
- * Dikey cubuklu bazen süzgeç grafiği denir.

⚠ Eksenin sıfırın üzerinde bir değerde başlama. Bunu yaparsan cubuklar karışık ve karma karıştırıcı bir görsel oluştur.

⚠ Rainbow effect kullanma. Bu kulaga hoş gelse de, genelde tablolardan anlaşılmaması lastirici.

(3) Histogram



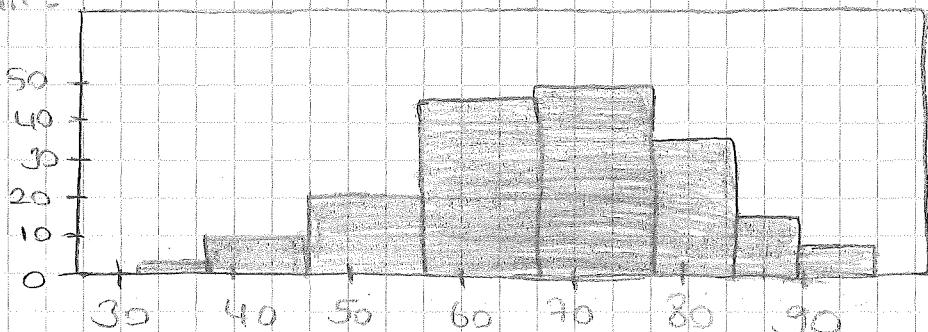
- * Çok sayıda gözleme olduguanda kullanılır.
- * Siklik tablosunda 100'den fazla veri kullanmak mümkün değildir. Bu yoldan veriler gruplanabilir

Interval's Lower Limit	Interval's Upper Limit	Class Frequency
32	38.4	1
38.4	44.8	4
44.8	51.2	19
51.2	57.6	22
57.6	64	49
64	70.4	50
70.4	76.8	38
76.8	83.2	48
83.2	89.6	13
89.6	96	6

- * Bu tabloyu oluşturmak için, 1. quan aralığı / sınıf aralığının boyundaki ilk aralık 32'den 38.4'e, ikinci 38.4'ten 44.8'e ...
 - * Sonra sınıf frekanslarını elde etmek için her aralığın içeren quantların sayısını sayıldı. (Birinci aralıkta 1, ikinci aralıkta 49'un)
 - * Sınıf aralıklarının genişliği "bin widths" olarak adlandırılır.
 - * Bin genişliği secimi, sınıf aralıklarının sayısını belirler.
 - * Bin sayısına göre veriler gruplanır.
- Bin = Çubuk Sayısı!

* Bir histogramda sınıf frekansları cubuklarla temsil edilir.

* Her cubugun yüksekliği, sınıf frekansına karşılık gelir.



Histogram of scores on a Statistics test

* The histogram makes it plain that most of the scores are in the middle of the distribution, with fewer scores in the extremes. You can see from the chart that frequency is greatest in the 60 to 70.4 score group.

Y) Bar chart da x eksenindeki etiketler \Rightarrow Categorical
Histogram'da \Rightarrow Quantitative (Nice!)

Population attributes → Parameters
Sample attributes → Statistic

Subject:

Date:/..../..

Population

A population is the total group about whom you want to make conclusions. It can be people, animals, objects, buildings, cars --.

All people living in the USA }
All building in a country }
Disabled children in India }
Diesel cars in Europe }

For example

Sample → Populationun alt kümlesi

A sample is a subset of the population for whom you actually have data.

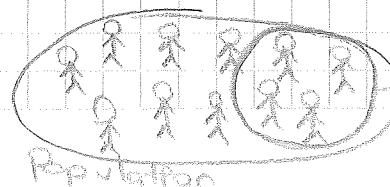
100 milyon kişiden oluşan arastirma yapılmakta,
onlar içinde seçilen 2000 kişi ile arastirma yapılmak
otta kalaydı.

The elements under study → Sampling units

The number of units in sample → Sample size

Population → Whole members with certain features.

Sample → Limited subset of the population, selected by a special process.



→ Sample

The mean $\rightarrow \bar{x}$
 (ortalaması)

The standard deviation $\rightarrow s$
 (Standart sapma)

Popülasyon ortalaması \rightarrow Popülasyondaki tüm gözlemlerin ortalaması.

Popülasyon standart sapması \rightarrow Popülasyon ortalamasıyla ilgili popülasyon dağılımlarının değişkenliğini tanımlar. Bu da genelde bilinmez.

Inferential statistical methods \rightarrow Örnek istatistiklere dayalı olarak popülasyon parametreleri hakkında karar ve tahminler yapmamta yardımcı olur.

! A parameter \rightarrow Popülasyonun sayısal bir öznidiği.

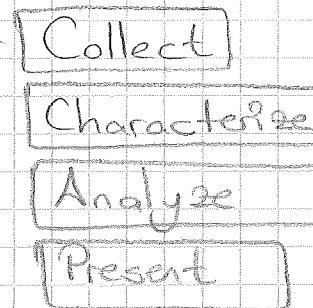
A statistic \rightarrow Bir örneğin sayısal bir öznidiği.

Subject :

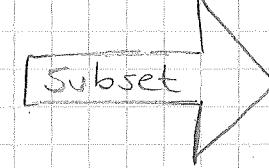
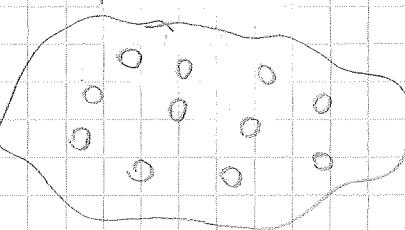
Date :/...../.....

Statistics

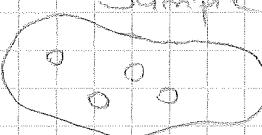
Most fundamentally, statistics is all about data.



Population



Sample



* Populations have Parameters (like $\mu, \sigma^2, \theta, \rho$)

* Samples have Statistics, functions of observed data
(like $\bar{X}, \tilde{X}, S^2, \hat{\theta}, \hat{p}$)

Statistic

Descriptive Statistic
(Tanımlayıcı İstatistikler)

Veriyi summarize eder,
basitleştirir.

Inferential statistikle
altıyaş yapar.

Inferential Statistic
(Uzakumsal İstatistikler)

Drawing conclusions about
a population based on
data observed in a sample

What is Data?

Data are characteristics or information, usually numerical, that are collected through observation.
(Gözlem yoluyla toplanan genellikle sayısal olan,
bazen kategorik bilgilerdir.)

DATA

Numerical

Continuous (float) height
weight
age

Discrete (Int) number of pets
number of children

Categorical

Ordinal (Data has a hierarchy)
S-M-L, good-average-poor

Nominal (Data has no hierarchy)
employees status, color, race

LEVEL OF MEASUREMENT

Subject :

Date :

Type	Measure property	Mathematical operations
Nominal (Religion, gender, mood, color)	Classification, membership	=, ≠
Ordinal (Size, quality, Likert scale)	Comparison, level	>, <
Interval (Oranlama yapılımaz) Gerek bir sıfır yok. Fark alnabilir.	Difference, affinity	+/-
Ratio (Gerek bir sıfır var (Oranlama yapabılır))	Magnitude, amount	×, /

Ordinal → Ordered

(Size
Quality
Likert scale)

Not actual value
Comparison Level

Nominal → No order

(Religion
Gender
Mood
Color)

Classification
Membership

Interval → Ranked

(Not öngörlü
yok
Temperature
Cartesian location)

Measured
Arbitrary
Difference

Ratio → Ranked

(Weight
Height
Age)

Measured
True zero
Magnitude

* Two basic divisions of statistics are:

Inferential and descriptive

(Sıklık Sayısı)

FREQUENCY

Görelebilir frekans

① Relative frequency → How often something happens

Her bir verinin sıklık sayısı divided by all outcomes

Toplam sıklık sayısı

② Cumulative Frequency → The accumulation of the previous relative frequencies

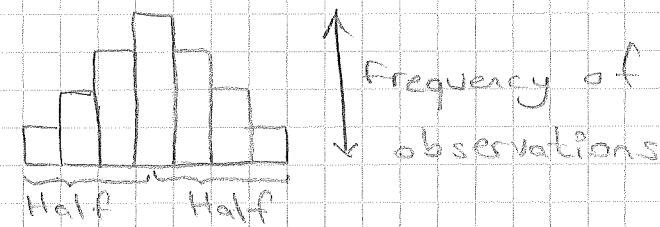
Daha önceki relative frekansları toplaya toplaya gider.

Data Value	Frequency	Relative Frequency → $\frac{\text{Sıklık}}{20}$	Cumulative Relative Frequency
2	3	$\frac{3}{20} = 0.15$	0.15
3	5	$\frac{5}{20} = 0.25$	$0.15 + 0.25 = 0.40$
4	3	$\frac{3}{20} = 0.15$	$0.40 + 0.15 = 0.55$
5	6	$\frac{6}{20} = 0.30$	$0.55 + 0.30 = 0.85$
6	2	$\frac{2}{20} = 0.10$	$0.85 + 0.10 = 0.95$
7	1	$\frac{1}{20} = 0.05$	$0.95 + 0.05 = 1.00$
t	20		

En son 16
karasılık gelirler.

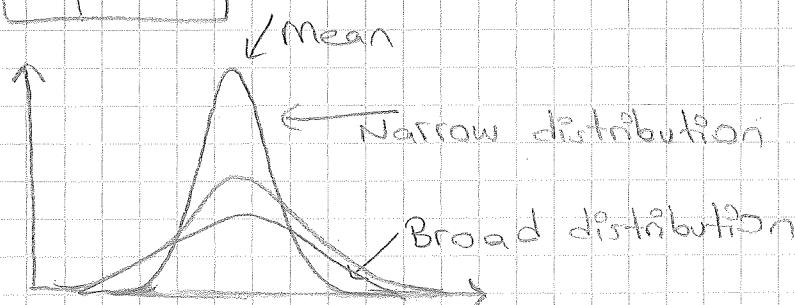
DATA PATTERNS IN STATISTICS

① Center



The center of a distribution, graphically, is located at the median of the distribution.

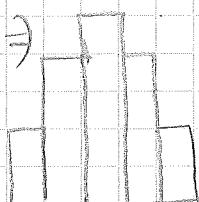
② Spread



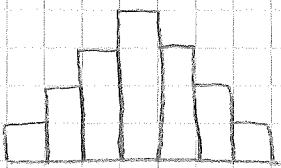
The spread of a distribution refers to the variation of the data.

③ Shape

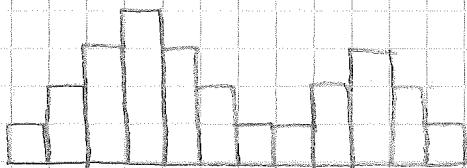
a) Symmetry



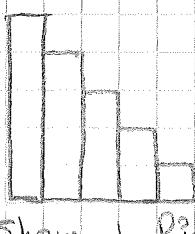
→ Gerçek hayatı çok az

b) Number of peaks

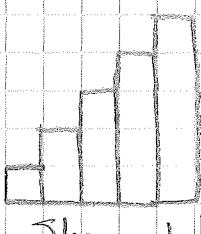
Unimodal,
Bell shaped



Bimodal

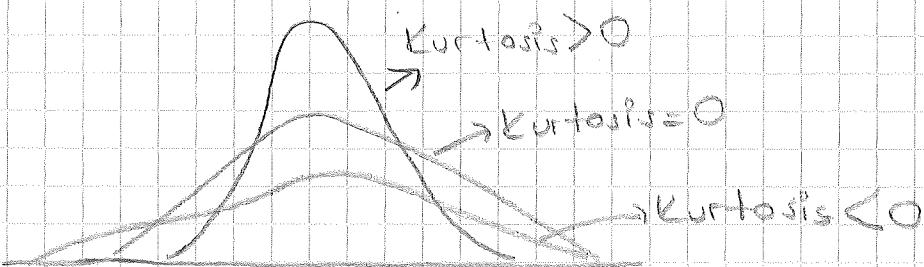
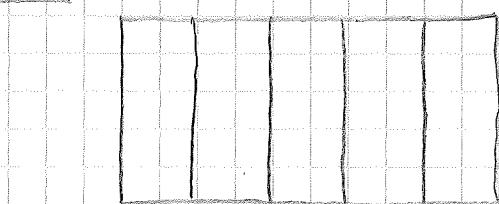
c) Skewness (Görüklik)

Skewed Right



Skewed Left

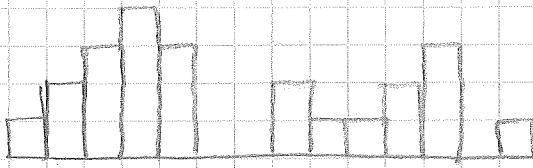
→ Kuyruk nerdeye o
yön) soylu yön)

d) Kurtosis (Basıklık)e) Uniform

→ Her bir cuburgun sevinci sensiz,
sikligi ayin.

(4) Unusual features

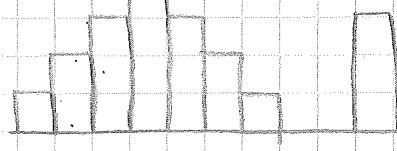
a) Gaps



→ Distribution 'in bell'
bir kısmında veri yok
(Missing value)

Eğer kayıp olmayaçaksa kayıp veri satırı tamamen silinebilir.

b) Outliers (Aşırı değerler)



→ Yanlış veri girisi olmuş olabilir.
Mesela yaş 150 girdilmiştir
yanlışlıkla.
Outliers mean'i kendine
doğru getirir. Medyan etkiler
mez.

Bar Graph

Bars have equal space

Histogram

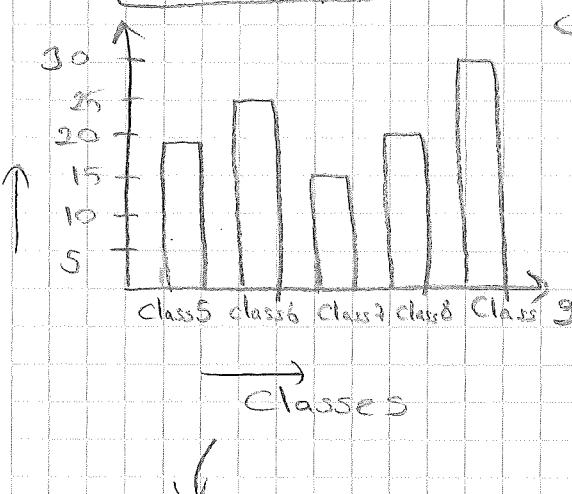
Bars are fixed

On the y-axis, we have numbers & on the x-axis, we have data which can be anything.

On the y-axis, we have numbers & on the x-axis, we have data which is continuous & will always be numbers.

Continuous
yazılı
yaş
yaz
6-7
a-

Bar Graph

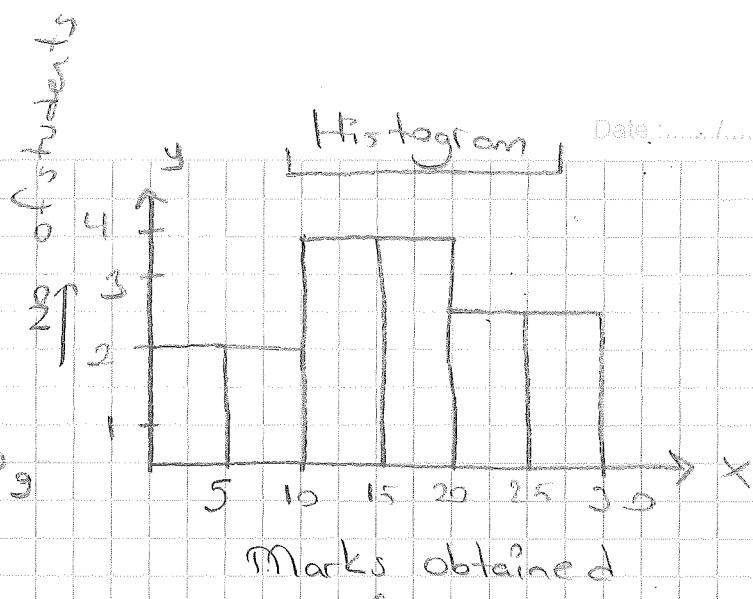


Eşit aralıklarla sıralır.

X → Kategoriler

Y → Numerik

Histogram



Birlikte çizilir.

X ve Y numeric

Parameters and Statistics

Population attributes → Parameters

Sample attributes → Statistics

Sample statistics are often used to estimate population parameters. Populasyon vesilesine ulaşmak çok zordur.

Sampling Techniques

① Probability sampling → Olasılıkla yararlanılır. Her bir bireyin eşit seçilme şansı var.

② Non-probability sampling → Olasılıkla yararlanılmaz. Veya toplama teknikleri.

Probability Sampling

- ① Simple random sample (Rastgele seçim)
- ② Systematic sample (Sıralama yapılır, belli aralıklarla seçim yapılır.)
- ③ Stratified sample (Alt populasyonlar bölünür. Mevcut durumu yüksük - orta - düşük gibi sınıflara göre)
- ④ Cluster sample (Birbirine benzeyen kümeler seçilir.)
(Küme örnekleme)

Non-Probability Sampling

- ① Convenience sample (En ulaşılabilir kişiler)
- ② Voluntary sample (Örnekler size gelir)
- ③ Purposive sample (Araştırmaya en uygun kişiler bulunur.)
- ④ Snowball sample (İki kişiye ulaşın, onlar başkalarına ulaşır.)

Extreme values = Outliers

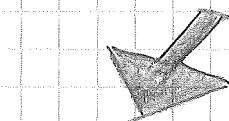
Subject:

25.09.2021

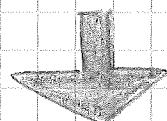
Date:

Central Tendency (Merkezi Eğilim)

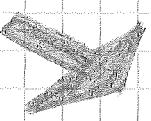
Bir dağılımin merkezi noktası, ortası



Mean



Median



Mode

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

Population Mean

Sample Mean

$$\mu = \frac{\sum x}{N}$$

Population mean

$$\bar{x} = \frac{\sum x}{n}$$

Sample mean

! Frekans ile mean bulma :

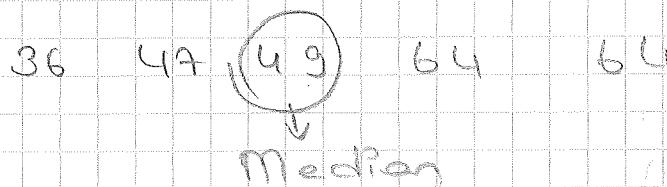
x	frequency
10	3
12	5
15	2
17	6

$$\frac{10 \times 3 + 12 \times 5 + 15 \times 2 + 17 \times 6}{3+5+2+6} = 13.875$$

5

Median = Ortanca Değer

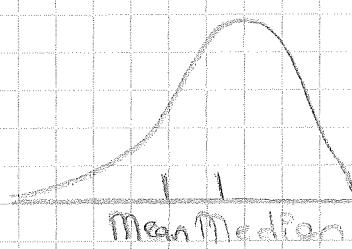
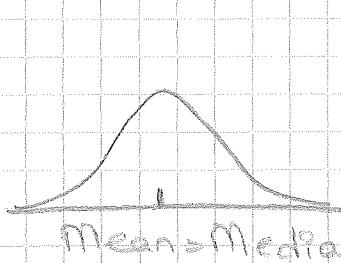
Median outlier' dan etkilenmez



Median outlier veri setlerinde medyanın daha kullanıldığı. Bu şekilde mean daha kullanılır. Küçük ama outlier'ı çok olan veri setlerinde de median tercih edilir.

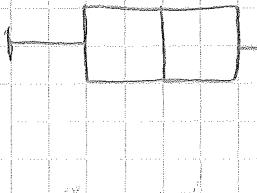
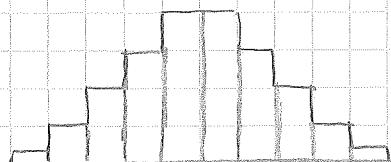


MEDIAN is resistant to outliers

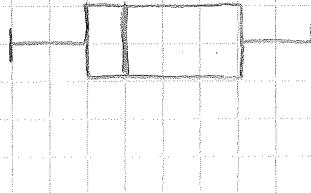
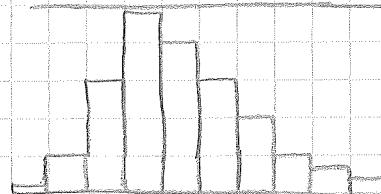


Extreme değerler mean'ı kendine çeker.

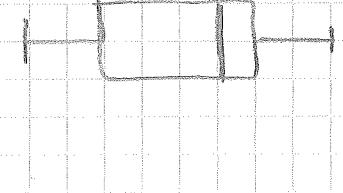
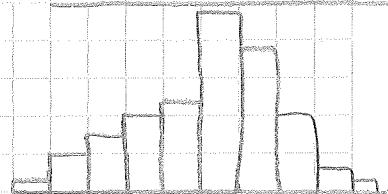
Symmetric



Skewed right (positive)



Skewed left (negative)



Mode 1 = Tepe Nokta

En çok tekrar eden değer.

Frekans Tablosuyla Mode Bulma

Classes	Frequency
0-10	7
10-20	16
20-30	18
30-40	9
Total	50

$$\text{Mode} = I + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot h$$

$$\text{Mode} = 20 + \frac{2}{2+9} \cdot 10 = 21.82$$

Modu yüksek olan
bu sınıftedenden bunu
seçti.

$I = 20$ oluyor. 0 sınıfındaki
en yüksek değer.

\rightarrow sınıflar arası mesafe

* Mode kategorik veriler için daha uygundur.

* Outlier ve extreme değerlerden etkilenmez.

* Nominal verilerde de kullanılır ama kategorikler için daha uygun.

* Tüm değerlerden biri içermektedir. Bu dezavantajı (Median gibi)

* Uniform dağılımla mode yoktur.

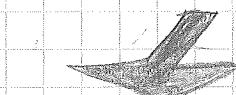


Subject :

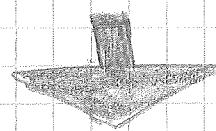
Date :/.....

Von ne Vardar
yolmis?

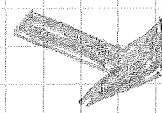
Dispersion (Measure of Spread)



Range



Standard Deviation - Variance



Interquartile Range (IQR)

Variance \rightarrow $(\text{Standard deviation})^2$

Outliers' dan etkileşti.

① Range \rightarrow En yüksek değerle en düşük arasındaki fark

Daihlik

Q_1 Q_2 (Median) Q_3

25% 25% 25% 25%

② IQR \rightarrow

$$\boxed{\text{IQR} = Q_3 - Q_1}$$

0 1 2 3 4

5 6 7 8 9

$$Q_1 = 2$$

$$Q_2 = 4.5$$

$$Q_3 = 7$$

$$\text{IQR} = 7 - 2 = 5$$

! IQR helps us
to make a technical
description of
outliers.

! Yani istiklal 999 girdim
değilim median ve IQR
değerler.

Mean \leftrightarrow Variance

Median \longleftrightarrow IQR

Subject :

Date :

① VARIANCE

Mean based bir spread ölçüsü

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

element
mean
number of elements

Büyük N olduğu için bu bir population varyansı

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

observation
mean
number of observations

Sample varyansı

x_i (veriler)	\bar{x} (mean)	Deviation ($x_i - \bar{x}$)	$(x_i - \bar{x})^2$
3	5	-2	4
4	5	-1	1
5	5	0	0
6	5	1	1
7	5	2	4
<u>+ 25</u>		<u>$\Sigma = 10$</u>	

Outlier is, any data point more than 1.5 IQR below the Q1 or above the Q3

Subject: EXAMPLE:

$$\text{Outliers} = (Q_1 - 1.5 * \text{IQR}) \text{ or } (Q_3 + 1.5 * \text{IQR})$$

0 1 5 6

$$\text{Mean} = M = \frac{\sum x}{N} = \frac{0+1+5+6}{4} = \frac{12}{4} = 3$$

$$SS = \sum (x - \mu)^2$$

$$SS = (0-3)^2 + (1-3)^2 + (5-3)^2 + (6-3)^2$$

$$SS = 9+4+4+9 = 26 \quad (\text{Korelo Ziplanı})$$

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

$\underbrace{\qquad\qquad\qquad}_{\text{Population variance}}$

$$\frac{26}{4-1} = \frac{26}{3}$$

$\underbrace{\qquad\qquad\qquad}_{\text{Sample variance}}$

Payda ikinci oldugu
için daha büyük
olması beklenir.

④ STANDARD DEVIATION, → Outliers'ın etkileri

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

element mean

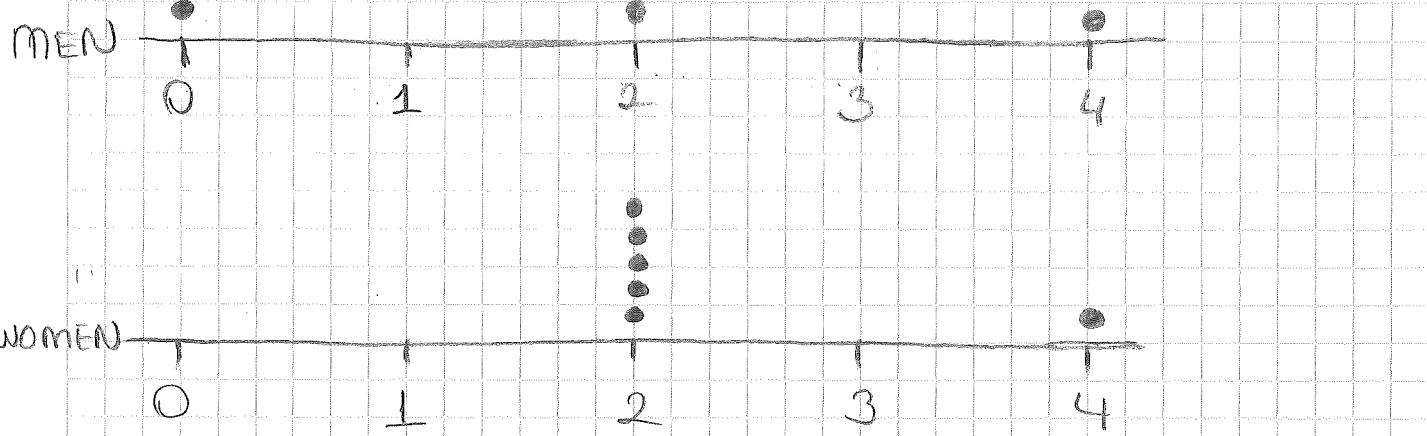
Variance formülünden varlığından
elde edilirse standart sapma
bulunur.

number of elements

Subject:

Mean = 2

Date:/..../.....



$$\text{Men} \Rightarrow s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{24}{6}} = 2.0$$

$$\text{Women} \Rightarrow s = 1.2$$

Kadınlar daha çok hem fikir. Bu yuzeden standart sapma az.

	X	Add 10
Data	0	10
	20	30
	40	50
	40	50
	50	60
Mean	30	40
Median	40	50
Mode	40	50
Range	50	50
IQR	20	20
Std Dev	17.9	17.9

ARTAR

DEĞİŞMELİ

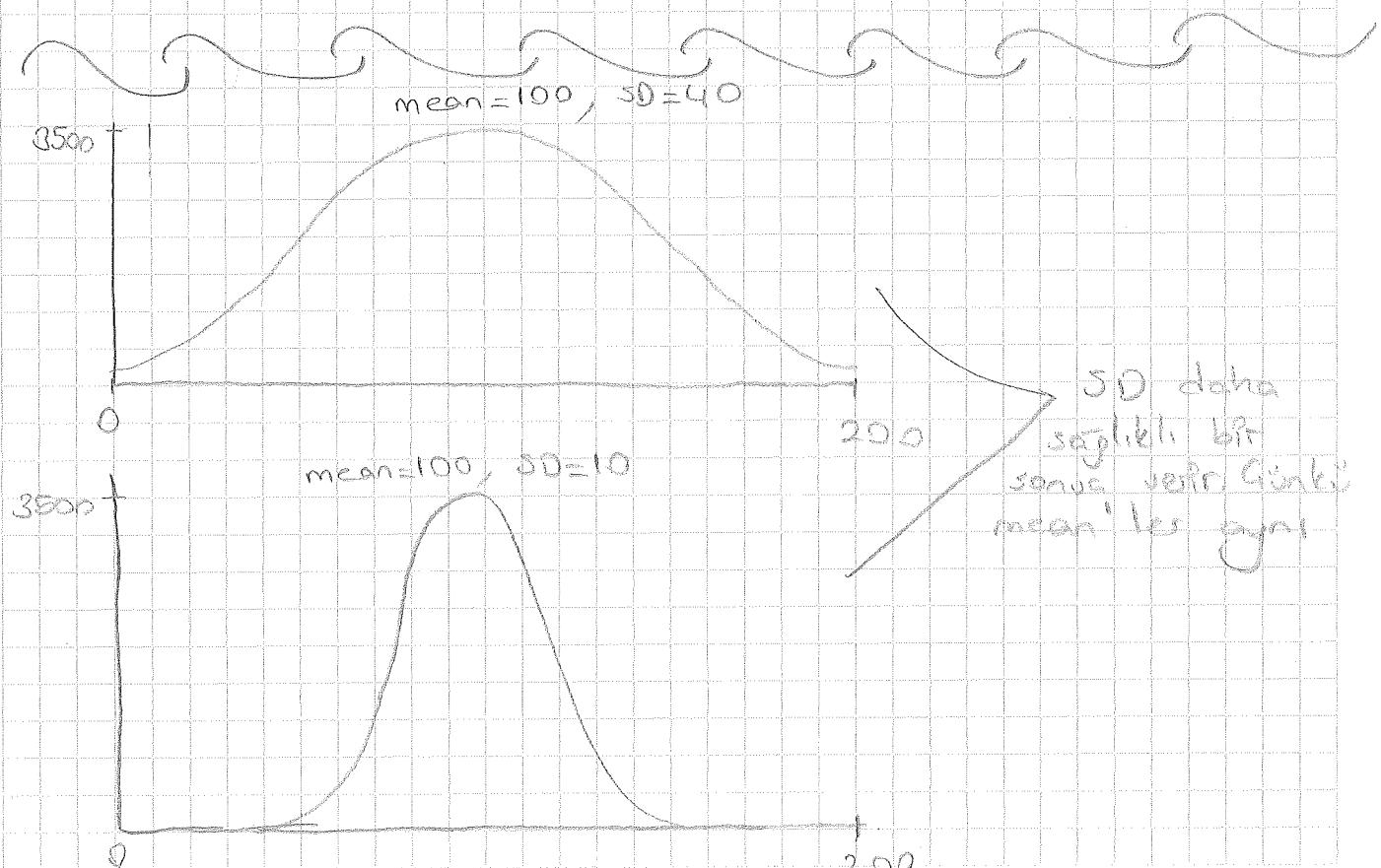
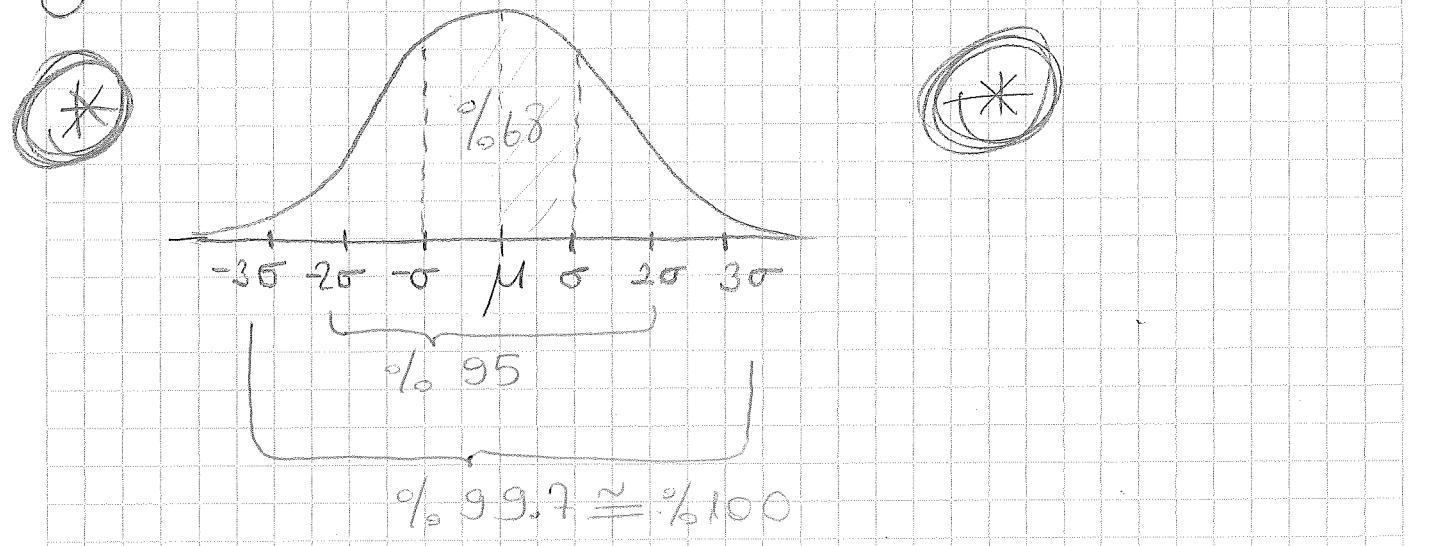
	X	Multiply by 10
Data	0	0
	20	200
	40	400
	40	400
	50	500
Mean	30	300
Median	40	400
Mode	40	400
Range	50	500
IQR	20	200
Std Dev	17.9	179

Artar

Ayn
oran
da
artar

The Empirical Rule

Normal bir dağılımda ort. 'dan 1σ sağa ve
1σ sola gidersek arada kalan alan kism analizler
üzer %68'lik kisimdir.



import numpy as np

a = [1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 1, 4, 4, 5, 5, 6, 7, 8, 9, 11]

(*) len(a)

→ 21

(*) np.mean(a)

→ 3.7619...

(*) np.median(a)

→ 3

from scipy import stats # mad hevabi için bunu

import ettiğim.

(*) stats.mode(a)

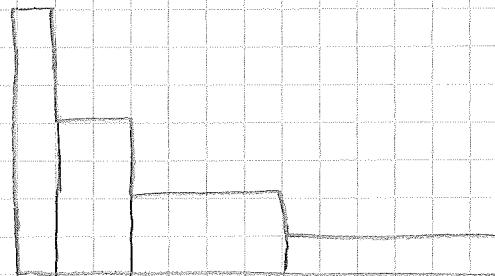
(*) stats.skew(a)

→ 0.988...

Skewed right

import matplotlib.pyplot as plt # Görselleştirme için
bunu import etmek

(*) plt.hist(a)



Right skewed

(*) np.var(a) # Varyans hesabi
→ 8.371882...

(*) np.sqrt(8.371882...) # Varyansın Korelasyon Standart Sapma
→ 2.893420...

(*) np.std(a)
→ 2.893420...

Subject :

Date :

import numpy as np

salary = [102, 33, 26, 27, 30, 25, 33, 33, 24]

print ("Range : ", (np.max(salary)-np.min(salary)))

print ("Variance : ", (np.var(salary)))

print ("Std : ", (np.std(salary)))

→ Range: 78

Variance: 539.555...

Std: 23.22833...

AYKLEİ DEĞER BULMA

number-list = [1, 5, 10, 15, 40]

min number = 1

max number = 40

median = 10

 $Q_1 = 5$ $Q_3 = 15$

$$\text{IQR} = Q_3 - Q_1 = 15 - 5 = 10$$

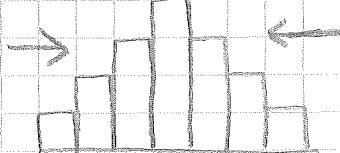
$$1.5 * \text{IQR} = 1.5 \cdot 10 = 15$$

$$Q_1 - (1.5 * \text{IQR}) = 5 - 15 = -10$$

$$Q_3 + (1.5 * \text{IQR}) = 15 + 15 = 30$$

-10 ile
30 arası
değerler

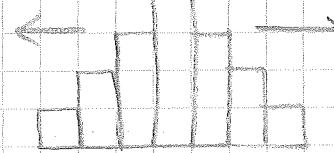
Bu aralığın dışındaki kalın
10 değeri outlier dir.

Measures of Centre

Mean

Median

Mode

Measure of Spread

Range

Standard Deviation

Variance

IQR

Quartiles, \rightarrow Veri kaca bölgüne bölü olarak ökülmeli degisir.

quartiles \rightarrow 4 groups

deciles \rightarrow 10 groups

percentiles \rightarrow 100 groups

Percentiles, \rightarrow For data, the p th percentile is the value of x such that $p\%$ of the data is less than or equal to x .

① Special percentiles: Min \rightarrow 0th percentile

Median \rightarrow 50th percentile

Max \rightarrow 100th percentile

② Quartiles: 25th and 75th percentiles

(Lower fourth = Upper fourth)

③ Interquartile Range (IQR):

$IQR = 75^{\text{th}} \text{ percentile} - 25^{\text{th}} \text{ percentile}$

Sometimes IQR is known as the "fourth spread"

Frequency Tables: Discrete and continuous data

Bar Charts: Discrete data

Histograms: Continuous data

- Location, variation, skewness, bimodality, outliers.
- Plots in two dimensions.

"
Boxplots: Continuous data

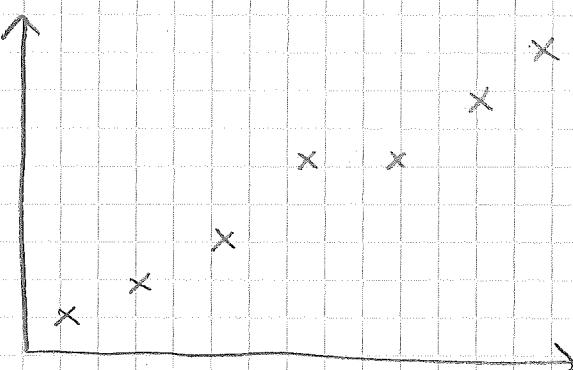
- Location, variation, skewness, outliers
- Plots in one dimension.

Scatter plots: 2 continuous variables

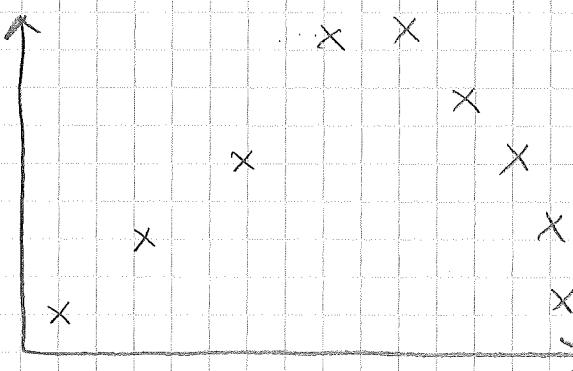
- Shows how variables relate (or not)

Patterns of Data in Scatter Plot,

① Linearity :



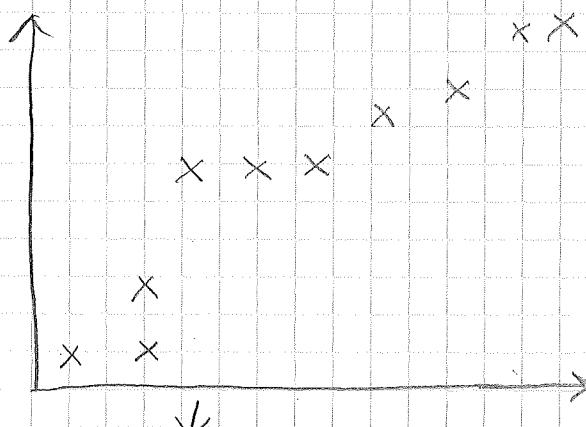
Linear (Diagonal)



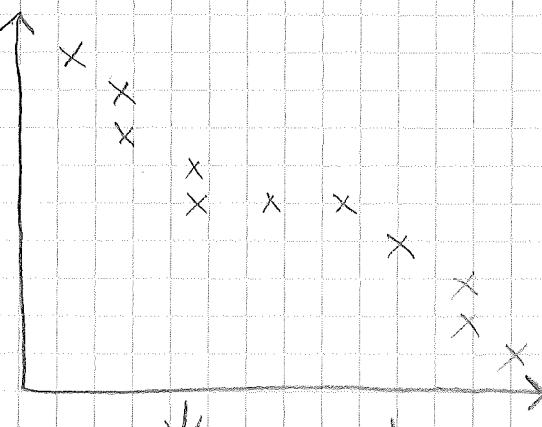
Non Linear

② Slope : x büyürken y büyürse $\rightarrow (+)$ eğim

x büyürken y küçülsse $\rightarrow (-)$ eğim



Positive slope

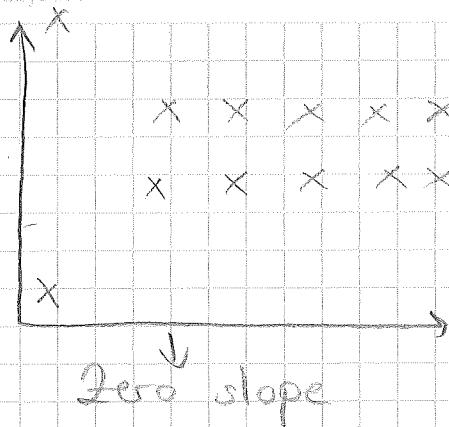


Negative slope

$\frac{\partial^2 f}{\partial x^2} = 0$ zero slope

Subject :

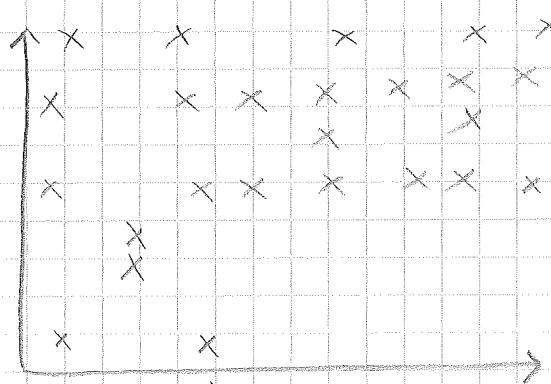
Date :



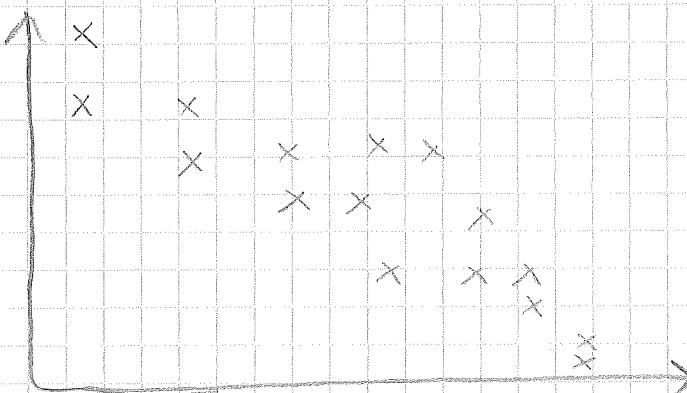
X ile arasındak eğim yok

③

Strength :

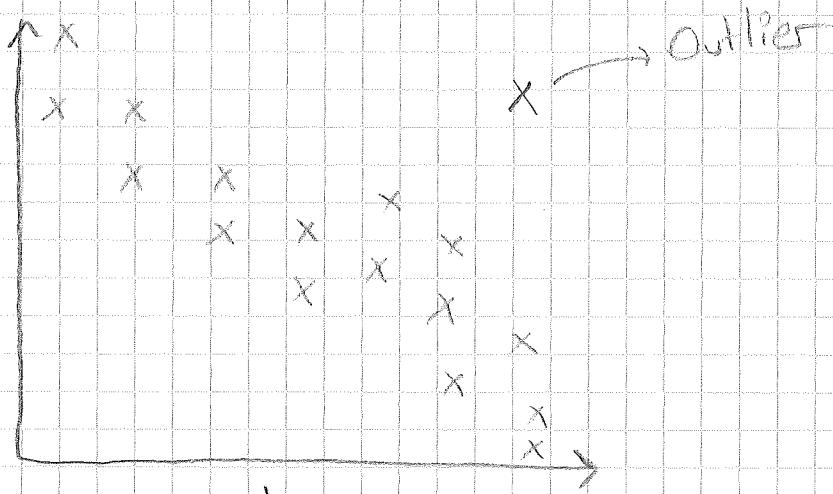
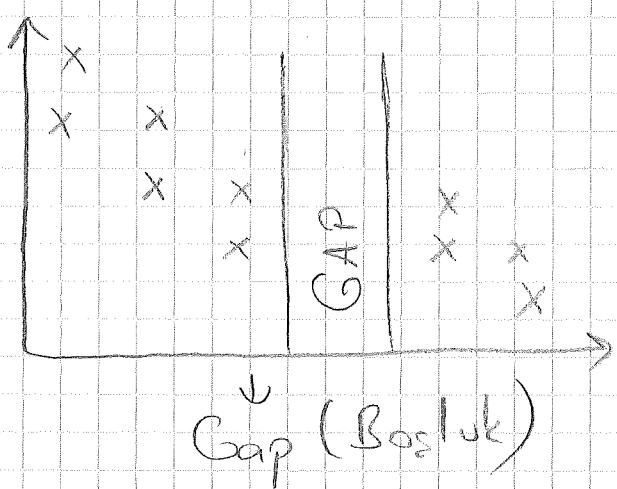
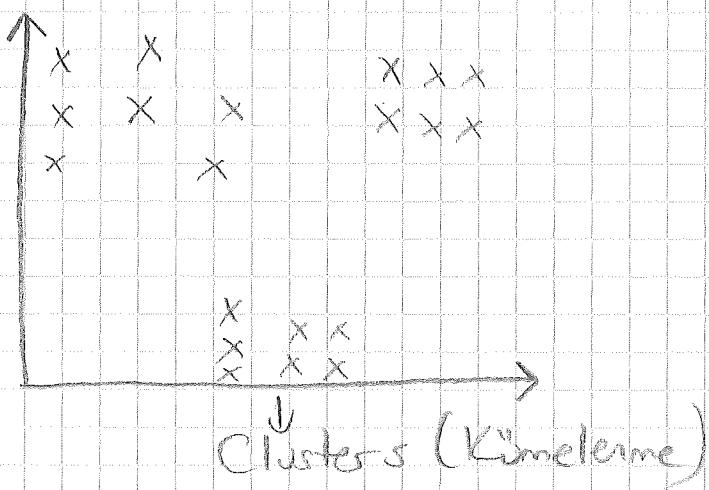


Weak relationship (Nesiller ayıktır)



Strong relationship

4 Unusual Features: (Clusters, Gaps, Outliers)



Outlier (Veriler arasında çok farklılık gösteren.)

Box and Whisker Plots.

Subject :

Date :/..../.....

Box Plot (Kutu grafigi)

Median based yaklasim. (Bunlari göstermenin

1st quartile $\rightarrow Q_1$

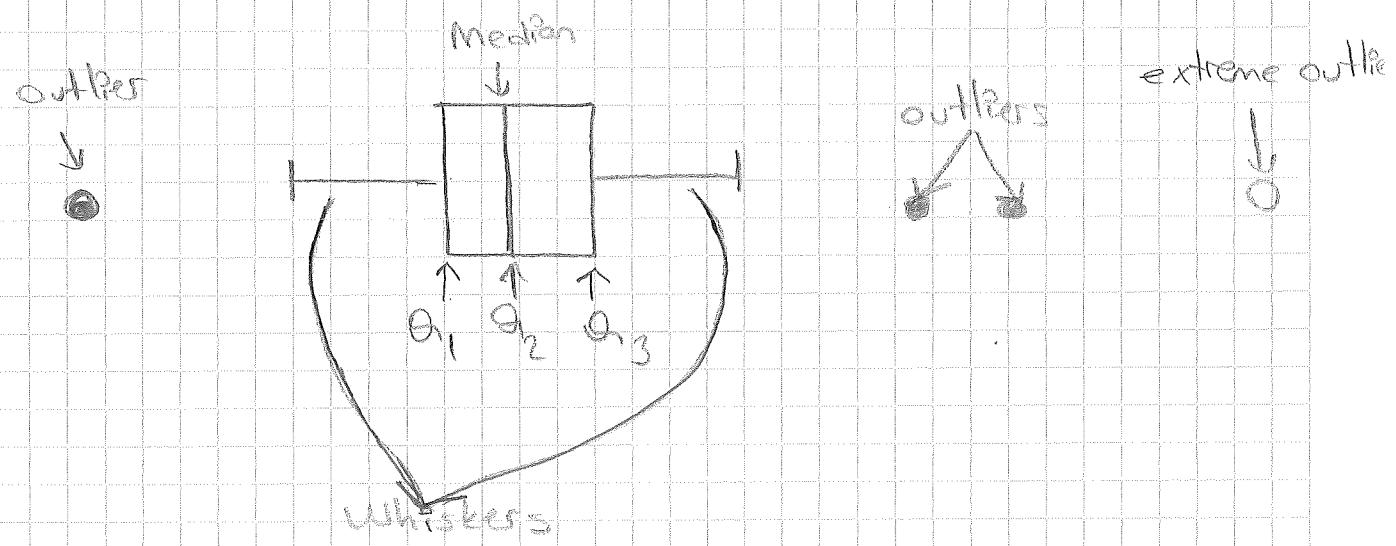
görsel yolu.

3rd quartile $\rightarrow Q_3$

(Outlier tespit etmenin
yollarından biri)

Outliers

Min - Max



* Only useful for continuous variables

* 2 grubu karşılastırmak için uygundur.
(Erkek ve kadınların bayan gibi)

Subject :

Date :

Weight, kg
38
25
37
28
35
29
35
29
34
30

Step 1 → ~~Verilen~~ sırala

Step 2 → Find the median

$$\text{Median} = 32$$

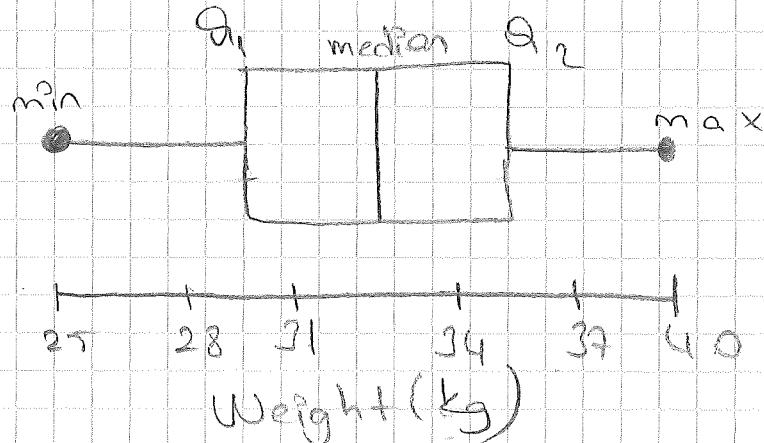
Step 3 → Find the quartiles

$$Q_1 = 29 \quad Q_3 = 35$$

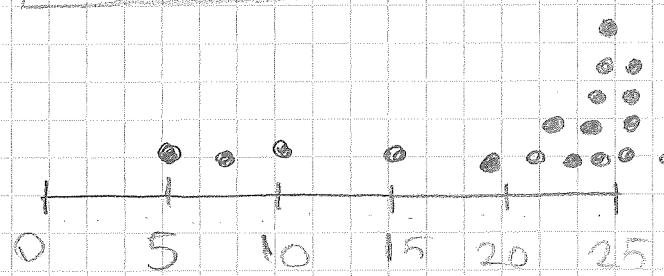
Step 4 → $\text{Min} = 25$

$$\text{Max} = 38$$

Min	25
Q_1	29
Median	32
Q_3	35
Max	38



Box Plot ile Outlier Tespiti

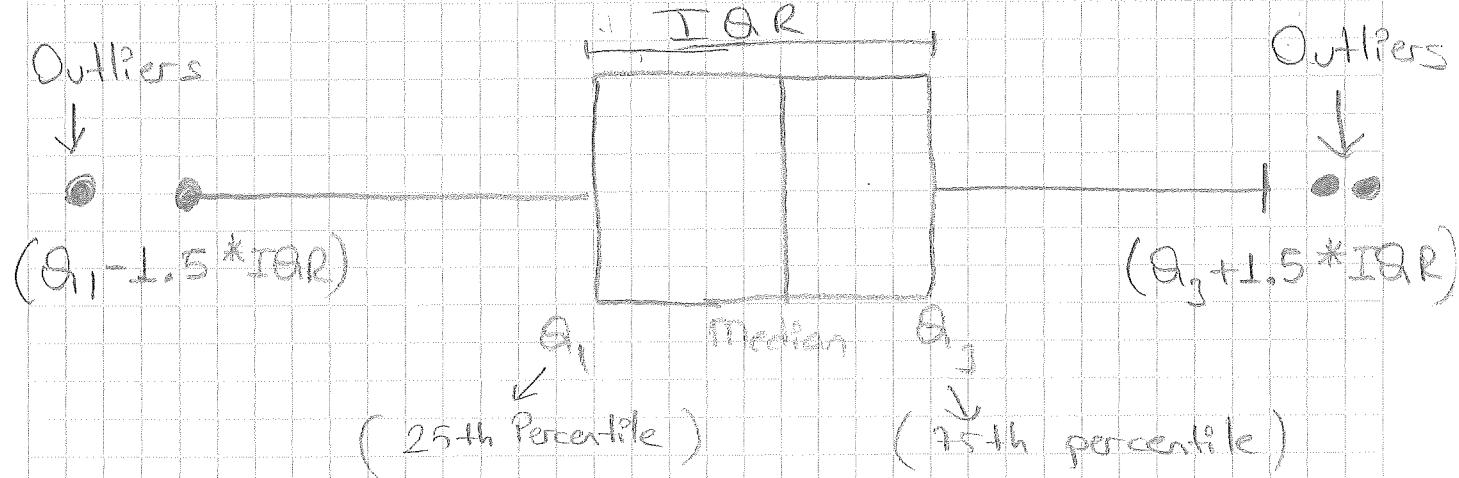


3-4 tane outlier var. Bilir yarına bosphorus

Outlier kavramına John Tukey açıklık getirmis :

1 IQR kutusundan uzaklasısanı discards çok az.)

2 IQR kutusundan uzaklasısanı çok fazla denir ve 1.5 IQR kuralları kaymaz

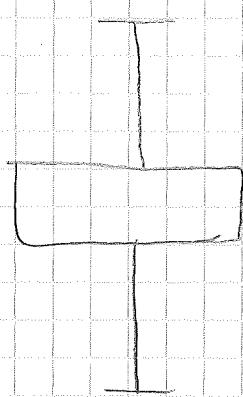


Kutudan sağ ve soldan 1.5 IQR uzaklaşanız ve discarda kalınlara outliers denir.

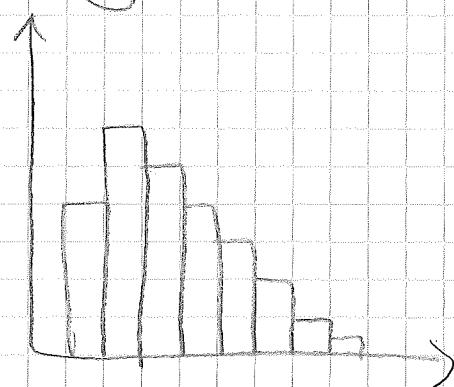
Subject :

Date :/..../.....

Box Plot



Histogram



EXAMPLE

3 6 8 9 9 10 12 14 19

Median = 9

$$Q_1 = 8 \quad 8 * 1.5 = 12$$

$$Q_3 = 12 \quad 12 * 1.5 = 18$$

Min = 3

Max = 19

Covariance

Covariance ve Corelasyon bağlantılı konular.

X ekseniindeki değişiklik y ekseninde ne kadar konsant biriyor. Bu hesaplanabilir.

$\text{Cov}(x,y) > 0 \Rightarrow$ Doğru orantılılar

$\text{Cov}(x,y) < 0 \Rightarrow$ Ters orantılılar

$\text{Cov}(x,y) = 0 \Rightarrow$ Birbirinden bağımsız

Correlation

$-1 < \text{Correlation} < +1$

Riskinin yanı kütü şenliği.

(Boy ortalaması kütüne artmam beklenir. İlişki pozitiftir.)

(Sigara içme sayı akciğer kapasitesi negatif bir ilişki)

Direction

- Positive : Move in same direction
- Negative : Move in opposite directions

Strength

Weak : Widely spread

Strong : Concentrated around a line

Karl Pearson \Rightarrow Pearson correlation : İleri seviyeler

Corelasyon kavramı -1 ile 1 arasında
şekilidir.

Pearson Correlation \Rightarrow r

Subject:

Date:

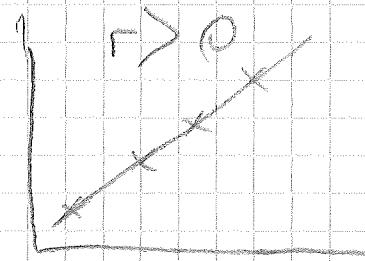
- Pearson correlation \Rightarrow Sınırları -1 ile 1 arasında
- r ile gösterilir,
 - 1 veya 1 'e ne kadar yakınsa
o kadar güçlü,
 - 0 'a ne kadar yakınsa o
kadar zayıf

Sample Correlation $\Rightarrow r = \frac{\text{Cov}(x, y)}{s_x \cdot s_y}$

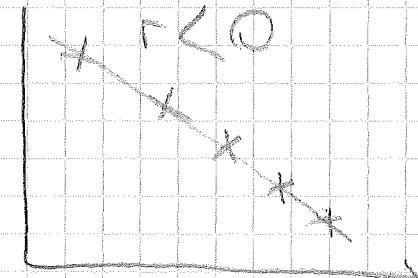
Population Correlation $\Rightarrow \rho = \frac{\text{Cov}(x, y)}{\sigma_x \cdot \sigma_y}$

Slope:

* Slope yukarı bakıyorsa $\Rightarrow r > 0$

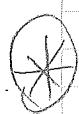


* Slope aşağı bakıyorsa $\Rightarrow r < 0$



Subject :

Date :/.....



$$r=1$$

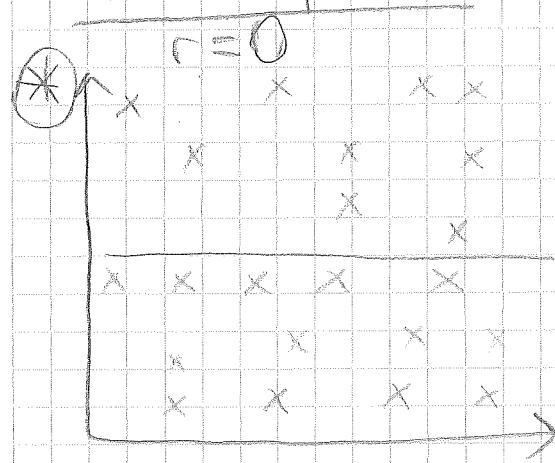
⇒ Max positive correlation

$$r=-1$$

⇒ Max negative correlation

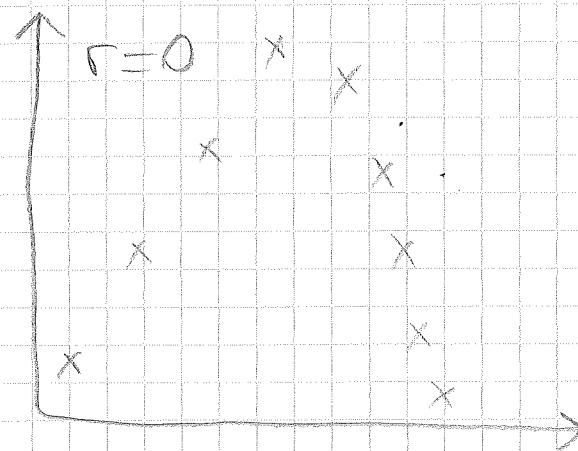
Tam çizgili
Sıradanı打破.

Random pattern



↓
Eğim yok
Correlation = 0

Curvilinear pattern



✓ $r \rightarrow (+)$ ise pozitif bir ilişkii var
• $r \rightarrow (-)$ ise negatif bir ilişkii yok (x ve y deşikten birbirinden bağımsız)

$r=0$ ise correlation yoktur.

Correlasyondan farklı, arada neden sonus iliskisi vardır.

Correlation \Rightarrow iki değişkenin
ilişkinin yönü
ilişkinin gücü

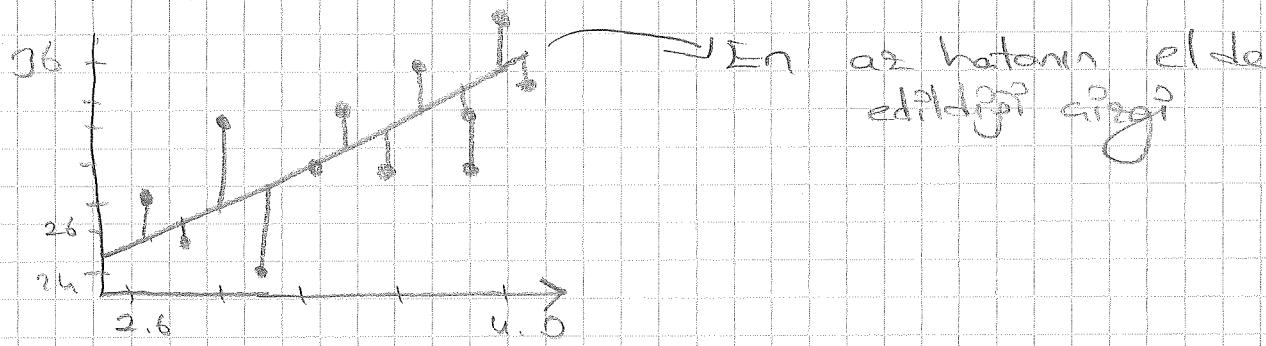
2 veri arasında kiyaslama yapmak için:

① Correlation

② Regression

1 Least squares linear regression

Least squares linear regression is a method for predicting the value of a dependent variable y , based on the value of an independent variable x .



Example

UNEDEN
Life expectancy \rightarrow GNP Per Capita

Earning in \$ \rightarrow Happiness Index score

$$\text{Regression} \Rightarrow y = \beta_0 + \beta_1 x$$

dependent variable

constant
(y'i kesen nokta)

Regression coefficient
(Eğim)

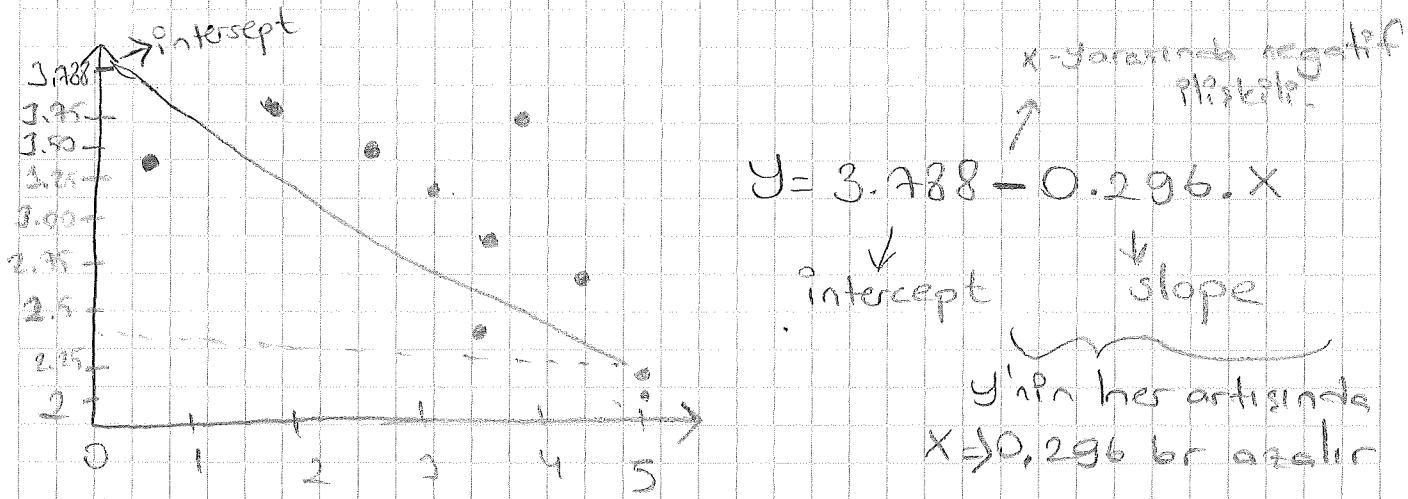
independent variable

Simple Lineer Regression

Date:

- Aralarda lineer ilişkisi var.
- 2 değışken arasında olur.
- Sayısal değişkenler.

! 2 independent 1 dependent variable olursa
multi-regression olur.



INTERSEPT \Rightarrow $X=0$ oldunda $Y=$ Intercept

0 saat TV izleyen bir kişi 3.788 not ort. getirebilir.

$$\text{Slope} = 5\text{km}$$

PYTHON İLE HESAPLAMA

```
tv_hours = np.array([3.5, 2.0, 5.0, 3.2, 4.0, 3.0])
```

```
GPA = np.array([2.9, 2.1, 3.3, 3.4, 2.0, 3.0, 3.6, 2.8, 3.5, 2.6])
```

```
res = stats.linregress(tv_hours, GPA)
```

```
print("b0:", res.intercept)  $\rightarrow 3.7874 \dots$ 
```

```
print("b1:", res.slope)  $\rightarrow -0.2958 \dots$ 
```

Subject:

Date:

The regression line \rightarrow Minimizes the sum of squared differences between the actual scores and the estimated scores

How To Calculate Coefficients

X	$X - \bar{X}$	$(X - \bar{X})^2$	Y	$Y - \bar{Y}$	$(Y - \bar{Y})^2$	$(X - \bar{X}) \cdot (Y - \bar{Y})$
---	---------------	-------------------	---	---------------	-------------------	-------------------------------------

ssx

ssy

$$\bar{X} \rightarrow X^{\text{ortalı}} \text{ ort.}$$

$$\bar{Y} \rightarrow Y^{\text{ortalı}} \text{ ort.}$$

$$\beta_1 = \frac{SP}{SS_x} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Bölle bir tablo

al, bu formülü yugula.
k: öğren

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

Avoid Extrapolation

Range for x : $[0:5] \rightarrow$ Range'in dışına çıkma

Bir aralık belirtmek gerekiyor.

Extrapolation'dan kaçının

Mesela 12 saat TV izleyen bir çocuğun not ort. "a" bu formülden bilmek doğru olmaz.

ERROR TERM

Respondent	TV viewing hours	GPA
A	3	2.90
B	5	2.10
:		
:		

Residual = Gerçek değer - Tahmin değer

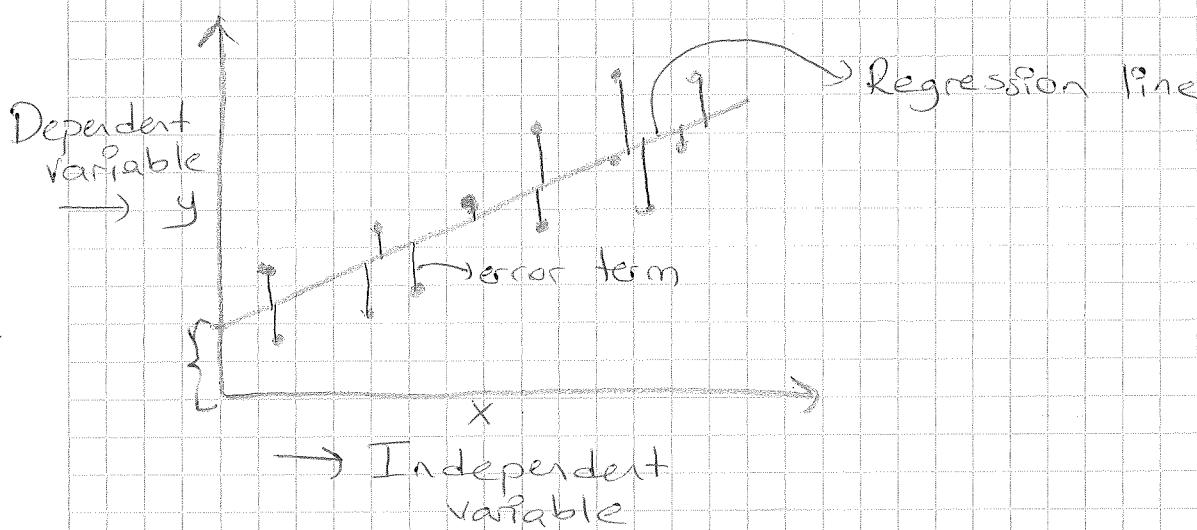
?

Regression Model

Formülümze hata degerini de ekler.

$$y_i = b_0 + b_1 x_i + \epsilon_i$$

Error term



Coefficient of Determination $\Rightarrow R^2$ = Verimlilik
Katsayısı

Modelim ne kadar iyi bir model?

Y'deki varyansın ne kadarını açıklıyoruz?

$R^2 = 0 - 1$ arasında

$R^2 = 0$ ise Y değişkeni (bağımlı değişken) açıklanamaz.

$R^2 = 1$ ise Y değişkeni açıklanabilir.

$R^2 \rightarrow$ Y'deki varyansın açıklanabilirliği

Simple Linear Regression $\Rightarrow r_{xy}^2$

EXAMPLE,

Correlation katsayısi = 0.64 ise :

$$R^2 = (0.64)^2 = 0.40$$

\rightarrow %40 oranında açıklanabilir. Orta seviyede bir ilişkisi var. Yani y_i 'yi tahmin ederken x yete \rightarrow kalınamaz. Daha fazla bilgi (bağımlılık değişken) gerekir.

R-squared,

Total variation in $y \rightarrow SST = \sum_{i=1}^n (y_i - \bar{y})^2$

Total Sum of squares

Explained variation $\rightarrow SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

Sum of squares regression

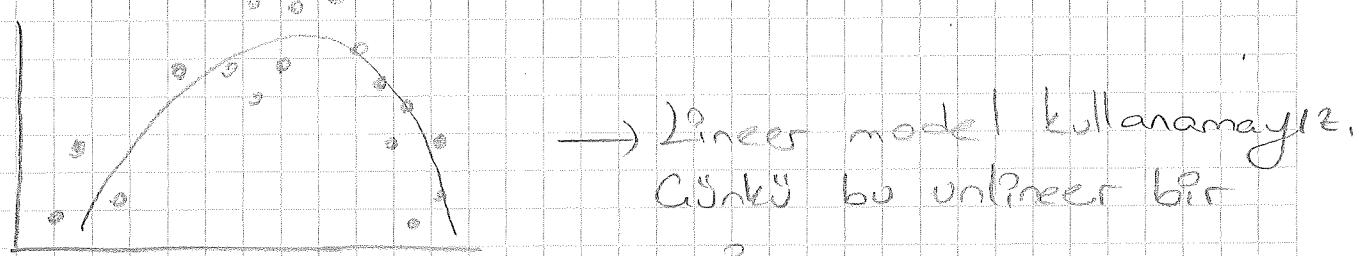
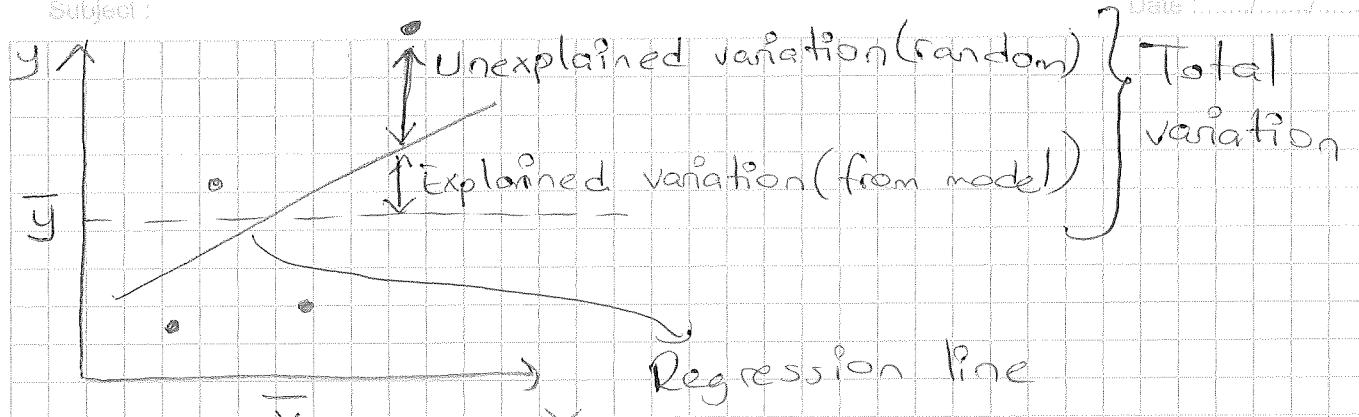
Unexplained variation $\rightarrow SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

Sum of squares error

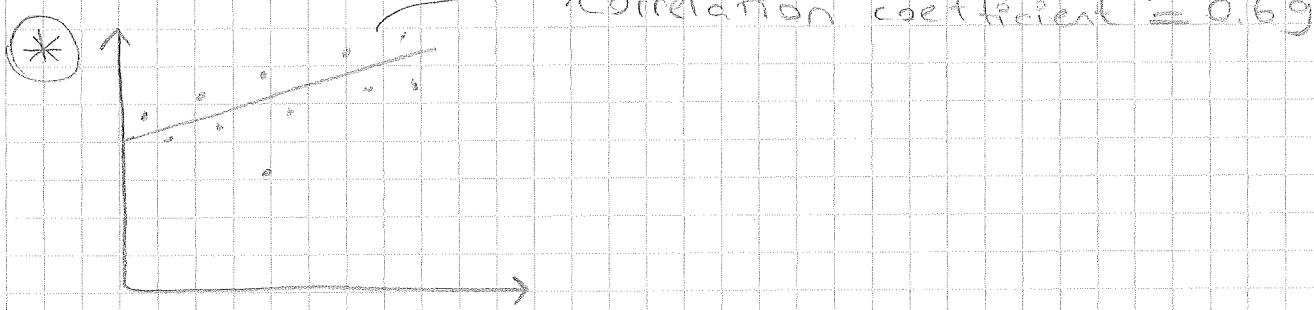
Coefficient of Determination $\Rightarrow R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$

Subject :

Date :



KAHOOIT



$R^2 \Rightarrow$ Coefficient of determination

EXAMPLE

x in ort. $\Rightarrow \bar{x} = 3$ } The regression line that
predict y from x ;
 y in ort. $\Rightarrow \bar{y} = 7$ } $(3, 7)$ noktasından
geçer;

PRE-CLASS

Regression

Dövizsel bir plaka? Kullanarak X (bağımsız) deşirkeinden y (bağımlı) deşirkenin tahmin etmek.

A functional relationship between two or more correlated variables that is often empirically determined from data and is used especially to predict values of one variable when given values of the others.

<u>Variable X</u> Independent Variable	<u>Variable Y</u> Dependent Variable
---	---

$$y = B_0 + B_1 \cdot x$$

↗
 Bağımlı Sabit
 deş.
 ↘
 regresyon katsayıısı

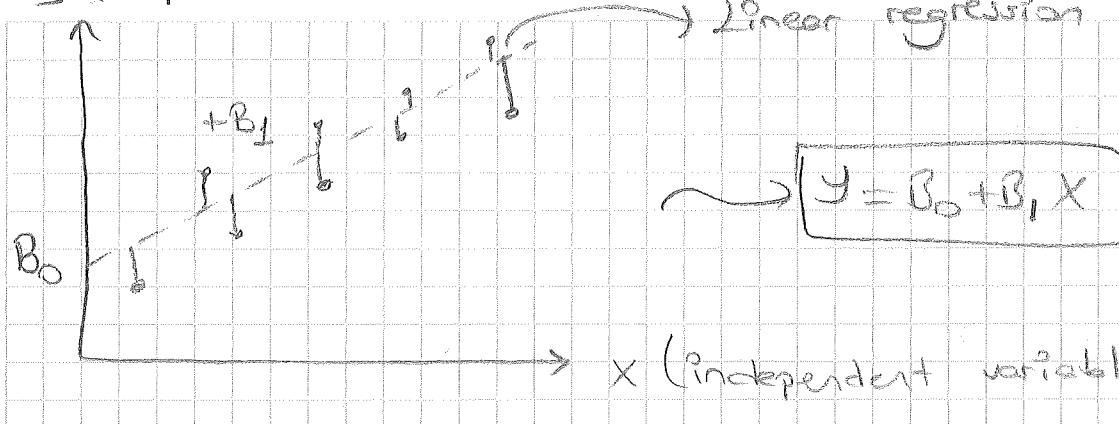
sum of products

$$B_1 = \frac{SP}{SS_x} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

sum of squares for the independent variable

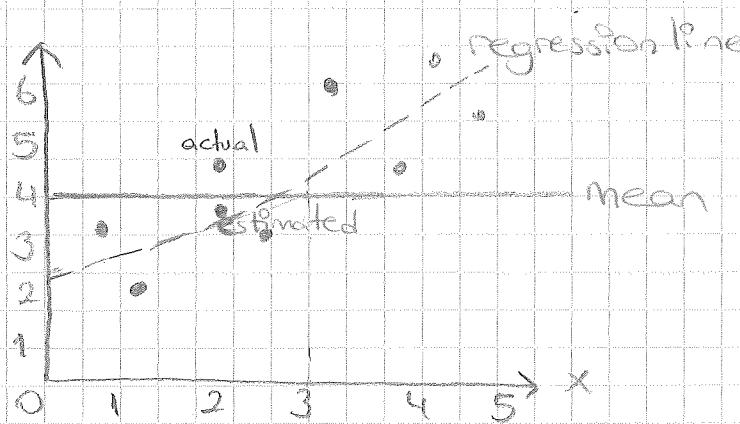
$\bar{x} \rightarrow x$ ların ortalaması,
 $\bar{y} \rightarrow y$ ların ortalaması

y (dependent variable)



$$R^2$$

(R^2 Squared)



$$R^2$$

Degiskenler arasindaki iliskileri kisa ve anlasılır bir sekilde anlatmak için kullanılır. $R^2 = 1$ ve -1 'e yaklaştıkça verimlik artar.

Subject:

30.10.2021

Date:

PROBABILITY

Relative frequency:

$$RF = \frac{\text{Number of times event occurs}}{\text{Number of trials}}$$

Sample space \rightarrow Gelebilecek tüm olasılıklar

$$P(A') = 1 - P(A)$$

PERMUTATION, \rightarrow Order

Arrange

$$\frac{n!}{(n-r)!}$$

COMBINATION, \rightarrow Not order

Select

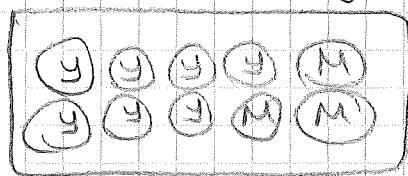
$$\frac{n!}{r!(n-r)!}$$

Depend Events, (Bağımlı Olay)

Desteden 2 kart çektiğinde AS gelme olasılığı.

Firtına çıktığında veysaların iptal edilme olasılığı

Renkli topların öklüğü kutudan 2 top çekmek. (Gez bırakmadan)



2 kere Yeşil çekme olasılığı:

$$\frac{7}{10} \cdot \frac{6}{9} = 0.46$$

$$1 \text{ yeşil} - 1 \text{ mavi} \Rightarrow \frac{7}{10} \cdot \frac{3}{9} = 0.23$$

Independent Events (Bağımsız Olay)

Bir zar ve bir paraşagi aynı anda atmak

Ard arda 5 kez paraşanın havaya atılması

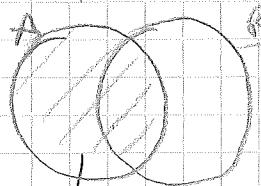
$$P(A \text{ and } B) = P(A) \cdot P(B)$$

$$\left(\frac{1}{2}\right)^5 = \frac{1}{32}$$

Conditional Probability

$$P(B|A)$$

A olmesi, buna bağlı olarak B'nin olasılığını hesaplıyoruz.



$$A \text{ oldu, } B \text{ 'nın olasılığı} \Rightarrow \frac{A \cap B}{A}$$

$$B|A = \frac{A \cap B}{A}$$

Paydaya kenisin olasılığı yazılır.



A ve B bağımsız olaysa $P(B|A) = P(B)$

$$P(A|B) = P(A)$$

BAYES THEOREM

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \rightarrow P(A|B) \cdot P(B) = P(A \cap B)$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \rightarrow P(B|A) \cdot P(A) = P(A \cap B)$$

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

Q What is the probability of two girls given at least one girl?

$$(2G|1G) = ?$$

$$(2g|1g) = \frac{(1g|2g) \cdot (2g)}{(1g)} = \frac{1 \cdot 1/4}{3/4} = 0.33$$

2 girls
one girl
same
opp.

GG, GB, BG, BB

GG, GB, BG, BB
3/4

Random Variable

Ordinary Variable

People, places, things

Values can vary

Random Variable

People, places, things

Values can vary

Value based on chance

(Number, count derive discrete objects)

① Discrete (Kesikli) \Rightarrow Birinci türdeki belirli değerler alabilir. (Ayakkabı no gibi)

② Continuous \Rightarrow Her türlü değer alabilir. (Yas gibi)

$P(X=x)$ \rightarrow x 'in olasılığının değerleri.

Random variable'in

(Büyük varfle gösterilir.)

$P(X=2) = \alpha$ \rightarrow Values

2 deka tara
gelebilme
olasılığı

③ Discrete 'de Tablo oluşturabilirimiz

Discrete \Rightarrow Probability Mass Func. (PMF)

Continuous \Rightarrow Probability Density Func. (PDF)

Subject: _____

Date: _____

Probability distribution

Values of X		Probability
0	X	0.25
1	X	0.5
2	X	0.25

→ Kesten

Discrete Probability Distributions

Binomial

Geometric

Hypergeometric

Poisson

Negativ Binomial

Trial
Parameters: $\rightarrow (n, p)$ \rightarrow success
(Discrete)

Subject:

Keskiö birc

Date:/.....

BINOMIAL DISTRIBUTION

7 dagitim

- (*) 2 sonuc olmali (Parçalı 2 kere atmak)
- (*) Two possible outcomes \rightarrow Success (Mesela tıra gelmesi) \rightarrow Failure
- (*) Probability of success is constant
- (*) Trials are independent

Hastanın iyilesip iyileşmemesi

Binomial random variable

The number of successes X in
n repeated trials of a binomial
experiment.

Binomial distribution

The probability distribution
of a binomial random
variable

- (*) 2 yaşı tıra atılımlı
 - 0 Tıra }
1 Tıra } gelebilir
2 Tıra }

Outcomes	Probability
0 heads	$P(X=0)$
1 head	$P(X=1)$
2 heads	$P(X=2)$



- The performance of a machine learning model
- Number of patients responding to a treatment

Binomial

dist'a
örnek

Scipy binomial $\rightarrow \text{pmf}(x, n, p, \text{loc}=0)$

Subject:



`stats.binom.pmf(x, n, p)`

x = success

n = trials (Durchsayisi)

p = probability

$$P(x=x) = \frac{n!}{x!(n-x)!} \cdot p^x \cdot (1-p)^{n-x}$$

\hookrightarrow Bernoulli distribution

$T=0$ classif.:

$$P(x=0) = \frac{2!}{0!(2-0)!} \cdot 0,5^0 \cdot (1-0,5)^2 = 0,25$$

Mean

$$\mu = np \quad (\text{Expected value})$$

Standard Deviation

$$\sigma = \sqrt{np(1-p)}$$

`stats.binom.pmf(x, n, p)`

`stats.binom.cdf(x, n, p)` \Rightarrow Cumulative

Subject:

Binom dopilim
ölçelik

Date:

Bernoulli Experiment

Binom dan farklı \rightarrow 1 deneme yapılıyor.

$$\mu = p$$

$$\sigma = \sqrt{p(1-p)}$$

Meni doğan bir çocuk kız mı erkek mi?

Sınavda baserili-başarılıdır.

İnsan bir sezi sever veya sevmeyez.

Poisson Experiment

Parametresi λ

Nadiren olan olaylar

Bütün olaylar birbirinden bağımsızdır.

İki olay aynı anda meydana gelmez.

λ = The mean of number of success

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

$$\begin{cases} \mu = 2 \\ \sigma^2 = 2 \end{cases}$$

e = Sabit sayı

X = Aranan olay sayısı

Poisson

Bir kilitin kaçta hata sayısı

Gecikmeli kalkan usak sayısı

Kredi kartı dolandırıcılığı tespiti

stats.poisson.pmf(x, λ) \Rightarrow Poisson
.cdf

Normal Distribution = Bell curve

Subject :

Date :

Continuous Probability Distribution

Olasılık hesabı aralık içinde yapılabilir.

PDF \rightarrow Olasılık Topluk Fonk.
(Probability Density Func.)

- ① y is a function of random variable
- ② y is greater than or equal to zero for all x
- ③ Area under the curve is equal to one.

Tek bir sayı üzerinde olasılık hesabı yapamıyoruz.
Aralık belirleyip alan hesabı yapıyoruz.

CDF \rightarrow Cumulative Distribution Func.

\hookrightarrow Toplaya toplaya gitiyor. 0'ından 1'e doğru
üzerinde.

! With continuous probability distribution:
 $P(X=a) = 0$

PDF Szerinden alan hesabi

$$\text{Find } P(0.6 < x < 1.0)$$

P = Area under PDF curve

$$\text{Area} = 0.4 \times 1.0 = 0.4$$

$$P(0.6 < x < 1.0) = 0.4$$

Continuous Distributions Grafikleri

Uniform

Normal

Exponential

Gamma

Chi-square

Uniform Distribution

Parametreleri $\Rightarrow a \text{ ve } b$

$U(a, b)$ or $\text{unif}(a, b)$

\rightarrow Sınırlı bir dağılım

a ile b arasında tek düz dağılım.

$a \rightarrow$ Location parameter

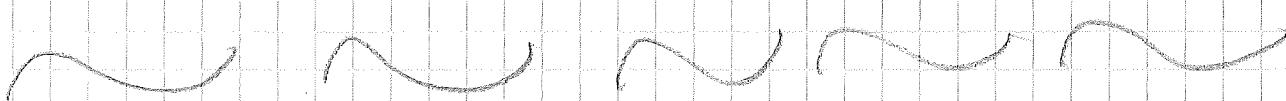
$b \rightarrow$ Scale parameter

Mean

$$\frac{1}{2} (a+b)$$

Variance

$$\frac{1}{12} (b-a)^2$$



Mean = 0 } Normal Distribution = \mathcal{Z}
 $SD = 1$



FINDING STANDARD SCORE (Z -score)

$$Z = \frac{X - \text{mean}}{\text{Standard deviation}}$$

Ort ve SD 'i bilmen bir dağılım, Z tablosuna uygunlabilir bir hale getiriyoruz.

Her var ki meseli iken bir normal dağılım eğrisi değiştirmek yerine, Z skoruna göre bir standart normal dağılım eğrisi kullanabiliyoruz.

Described by two parameters: Mean, SD

Subject:

Date:

Normal Distribution

Weights of people

Blood pressures

Scores on a test

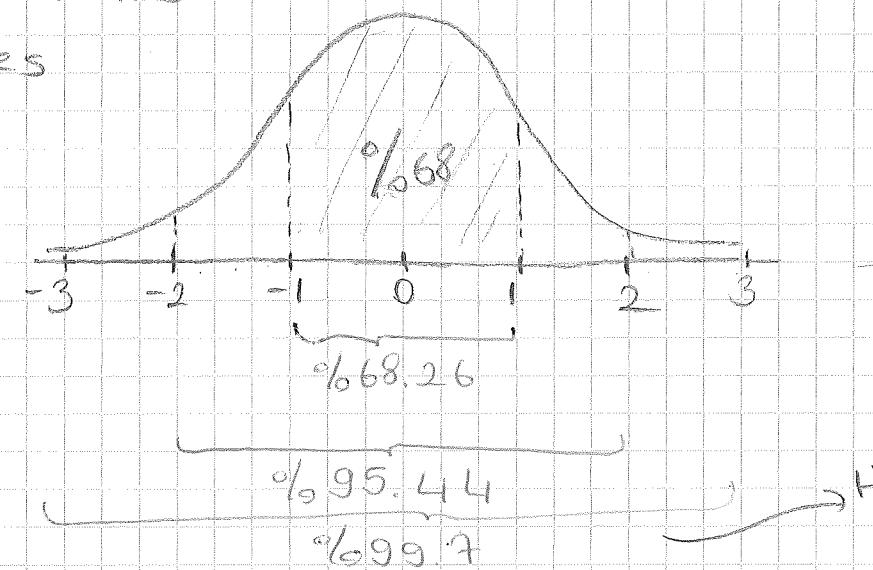
IQ scores

Wages

→ Bell curve

Mean = 1

Mean = Median = Mode



Hablic domain
% 100 olma.

Parameter → A number that describes the data from
a population

Statistic → A number that describes the data from
a sample

2 SCORE

Test score = 205 (x)

Mean = 180 (μ)

Standard Dev = 20 (σ)

$$Z = \frac{x - \mu}{\sigma} = \frac{205 - 180}{20} = \frac{25}{20} = 1.25$$

t Distribution

→ Student's t-Distribution

"Serbestlik derecesi ne olan dağılım?"
sonusuna soracaktır.

Kıymakla daşıführündür.

Sample size daşı k^{th} oldupnata kılınır.

{ Serbestlik Derecesi = $n - 1$ }

Sample Distribution

Örneklem dağılımı → Population distribution } Bütçə
 Sample size }
 Method of sampling } bəzi

"Örneklem dağılımının standart sapması" Standard Error
 den.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

↓
Sample Standard Error

$$\text{Var} = \text{Std}$$

11.11.2021

Subject:

Date:

The Steps of Significance Test:

- ① Assumptions
- ② Hypotheses (H_0, H_a)
- ③ Test Statistic (Asıl hesaplama burda basılır.)
- ④ P-value (Test istatistikini gecenin P-value'sını hesaplanır.)

① Assumptions

- Rastgele örnek seçilireli
- Gözlemler birbirinden bağımsız olmalı
- Population σ bilinir ve 30'un üzerinde örneklem var.

↓
Sigma bilgi mi var? Sample 30 dan büyük mü?
ikisi zamansa bir sonraki adıma geçer

② HYPOTESES

* Sıfır hipotezi $= H_0 \Rightarrow$ Başlangıçta doğru olduğunu varsayılar
iddia

* Alternatif hipotezi $= H_a \Rightarrow$ Bu iddia doğru olabileceğini.
Birbirinle kesişmez. Birbirinden birek veya kesişikler.

! Alternatif hipotezi kabul eden şəxsiyən,

"Fail H_0 reject H_a " diyənə.

↓
Başlangıçta kabul edilen seyi red etməkdən
başarılı olmak.

! Hipotezi, populasiyan üzərinə təqdim. (Ömeklem üzərinə deyil)

Oran \Rightarrow Kategorik değişkenlerde kullanılan parametre.

Mean \Rightarrow Sayısal değişkenlerde. " " "

Subject:

Date: 11.11.2023

$H_0 \rightarrow$ "Süçlüz patet edilenler, kadar -sanık suçsuzdur."

$H_a \rightarrow$ "Sanık suçludur." der ve kanıt toplar.

EXAMPLE

$H_0 \rightarrow M = 52$

$H_a \rightarrow M \neq 52$ yani $M < 52$ veya $M > 52$

! H_0 hipotezi eşitlik üzerine kurulur.

! H_0 ve H_a hibet zaman KESİŞMEZ. Birbirlerini dışlarlar.

(3) TEST STATISTIC

Popülasyon standart sapmasını biliyor musun?

Evet

Hayır

Sample 30'dan büyük mi?

Evet

Hayır

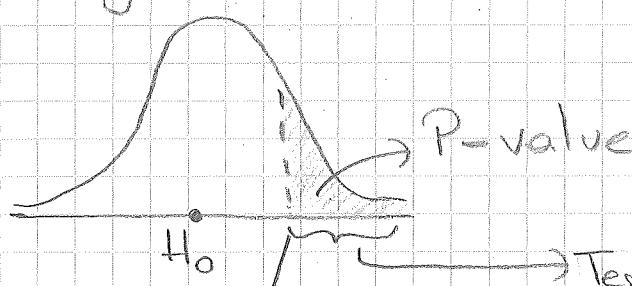
Z testi

t testi

! Gelen sıklığı hesabında olduğunu gibi hipotezde de parametreleriniz ; mean, var, χ^2 grubun ort. farklı olabilir.

Step 4 : P-Value

Test istatistikinin ötesinde kalan alan



$$\Rightarrow P\text{value} = 0.05 \text{ (default)}$$

H_0 → Test istatistikinin ötesinde kalan alan = P value
Sample value of test statistic (Extreme değer elde etme ihtimalı)

! P value ne kadar küçükse, H_0 'a karşı o kadar güclü bir kanıt bulmuş oluruz.

Sample test istatistikini 'z' tabloları içindeki hipotezi reddedebilebilim.

Step 5 : Conclusion

P-value, α 'dan daha küçük olduğunda H_0 'i reddedebilirsem, böylece reddedemem.

NULL HYPOTHESIS \rightarrow Fail to Reject the Null
 $P\text{-value} > \alpha$

: Reject the Null
 $P\text{-value} < \alpha$

α = Anlamılık Seviyesi = Significance Level

$P\text{-value} < \alpha$ } oldugunda H_0 hipotezi reddedilebilir.

$P\text{-value} \leq 0.05$ ise Reject

$0.05 < P\text{-value} < 1$ ise Do Not Reject

↓
Significance level = α

↓
Type I Error (yapma olasılığının) = α

Type I Error → Doğru kabul etme

H_0 hipotezi doğrular, elindeki kanıtlarla onu reddediyor.
(Sıkı sıkı bir inşaatın cezasına tutturulması)

Type II Error → Yanlış kabul etme.

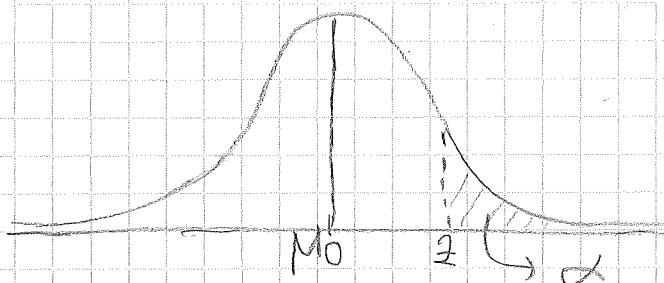
Reddedilmesi gereken bir H_0 hipotezinin kabul ediyor. Fail to reject. (H_0' , reddetmeye yeteneğ kalmıyor)

⚠ Type I Error daha kritik bir hatadır.

Correct Conclusion	Type I Error (Doğru olan H_0' 'i reddetme)	α = Type I hata yapma olasılığı
Type II Error (Yanlış olan H_0' 'i kabul etme)	Correct Conclusion	

↓
 β = Type II hata yapma olasılığı

Tek Kuyruk Hipotez Testleri



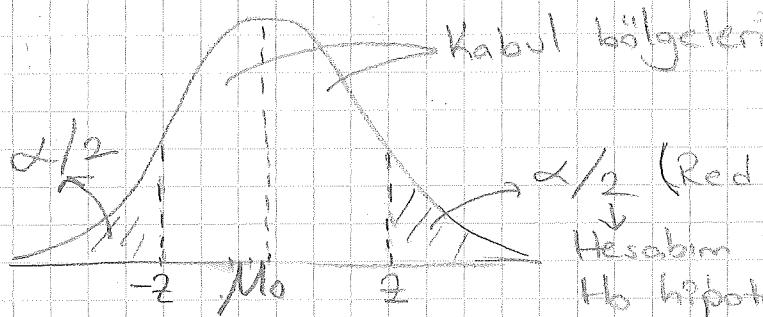
$$H_0: \mu = \mu_0 \quad H_a: \mu > \mu_0$$

Alternatif hipotezde
“>” işaretli varsa
tek kuyruklu
hipotez olur.

↳ Default $\alpha = 0.05$

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Cift Kuyruk Hipotez Testleri



$\alpha/2$ (Red bölge)

Hesabim bu bölgeye düşerse
 H_0 hipotezin reddeceğim.

Hesaplayacağımız test istatistikler red bölgelerine düşerse
 H_0 hipotezi reddecektir.

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

σ / \sqrt{n} → Standard error

$$H_0: \mu = \mu_0$$

Həm Mətbət

- Həm Mətbət Pseçəft kuyruklu test dövri və dəhər 8k? kuyruq paylaşıdır.

Example

Suppose that a beach is safe to swim if the mean level of lead in the water is 10.0 (Mo) parts /million.

We assume $X_i \sim N(\mu, \sigma^2 = 1.5)$

Water safety is going to be determined by taking 40 water samples and using the test statistic.

$$\text{Sample mean} = 10.5 = \bar{X}$$

$$\alpha = 0.05$$

Bir sahildeki ort. kurşun seviyəsi milyonda 10.0 parça olduğunu biliyor mus. Bu bir varsayımdır. Bu nəzərdən buna H_0 düşüldür. Bu nəzərdən normal bir dağılım olduğunu, ortalamanın bilinmediğini ama $\text{std} = 1.5$ olduğunu varsayıyoruz. 40 farklı su əməğini alıyoruz. (Kanıt sayımları $n=40$.) Bu kanıtlardan yola çıkarak ort. kurşun seviyəni milyonda 10.0 parça bulmuş.

$$\alpha = 0.05 \text{ (Significant level)}$$

Bu sahil yuzmek için güvenli mi?

10'dan küçük olduğunu bulursak güvenlidir, 10'dan büyük olmasa güvenli değildir elbette.

$$M_0 = 10.0$$

$$\sigma = 1.5$$

$$\alpha = 0.05$$

$$n = 40$$

$$\text{Sample mean} = 10.5 = \bar{x}$$

① Assumptions:

Rastgele 30'lu sealmış (40'ın serisini)

Örnekler bağımsız olmalı

6 bilinmeli, sample 30'dan büyük olmalı.

İlk asamada
ki koşullar
sağlanmış

② Hypothesis:

The null hypothesis $\Rightarrow H_0: M = 10.0$

The alternative hypothesis $\Rightarrow H_a: M > 10.0$

! Büyüğük durumu olduğunu Pan
tek taraflı hipotezi

③ Test Statistic:

$$Z = \frac{\bar{x} - M_0}{\sigma / \sqrt{n}} = \frac{10.5 - 10.0}{1.5 / \sqrt{40}} = 2.1$$

(Test Pstatistikti)

Buradan P-value değeri
bulacağınız.

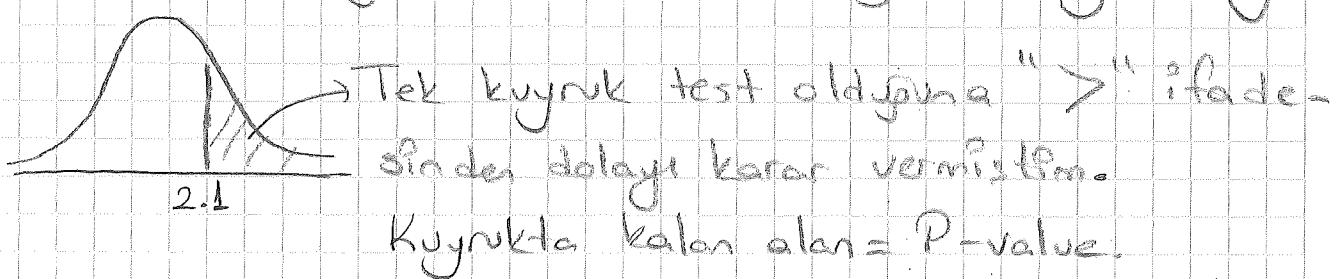
! import scipy.stats as stats } P-value kodu
1 - stats.norm.cdf(2.1)

Subject:

Date: 1.1.

④ P-value:

2 testi istatistiklerinden P-value değerini gitmeliiz.



Sadece sağda kalan vermissem.

Kuyruktaki kalan alanı = P-value.

2 tablosunda;

1	0.0					
1	1					
2.1	-0.079					



2.1 değeri 2 tablosunda 0.0179'a karşılık gelir.

$$P(Z > 2.1 \mid H_0 \text{ True}) = 0.0179$$

Başlangıçta eşik değerim $\alpha = 0.05$ verilmiştir.

0.0179 eşik değerinden küçük çıktı. Yani red bölge sine düşülmüştür.

Pvalue $< \alpha$ olduğu için H_0 hipotezini reddettik.



Reject the Null

"Saklı kapatmak için yeterli kanıt var. Yani şimdiki genelik değil."

Veya

" $\alpha = 0.05$ aralığında seviyesinde H_0 hipotezini reddetmek için yeterli kanıt sağlanmıştır."

EXAMPLE

A department store manager determines that a new billing system will be cost-effective only if the mean monthly account is more than \$170.

A random sample of 400 monthly accounts is drawn, for which the sample mean is \$178. The accounts are approximately normally distributed with a standard deviation of \$65.

Can we conclude that the new system will be cost-effective?

Yeni bir fiseleme sistemi var. Bu sistemin effective olup olmadığını department store manager belirliyor.

Eğer aylık ort. hesabımız \$170'dan fazla çıkarsa "cost-effective" sonucuna ulaşacağız. Kanıt olarak 400 farklı aylık hesaplarını topluyoruz.

$$\text{Sample mean} = 178 = \bar{x}$$

$$\text{Standart sapma} = 65 = S \quad (\Sigma \text{sigma demiyorum çünkü populasyon değil, sample'in standart sapması})$$

$$M_0 = 170$$

$$n = 400$$

"Yeni sistem cost-effective mi?" diye soruyor. Yani \$170'dan büyük çıkarsa cost-effective deneliyor.

① Assumptions:

Rastgele örnek seçimi (100 örnek seçildi) } Bu koşullar
 Örnekler bağımsız.
 σ bilinmeli, sample 30'dan büyük olmalı. } sağlanır.

② Hypothesis:

$$H_0: \mu = 170$$

$H_0: \mu > 170$ → Büyüklük ifadesi olduğu için "one-tail"

! $H_0 \rightarrow$ Eşitlik yerine kurulur.

$H_a \rightarrow$ Büyüklük yerine kurulur.

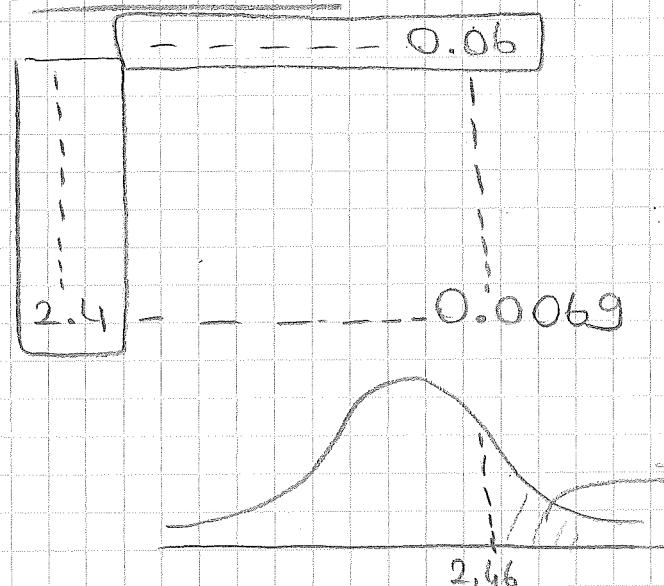
③ Test Statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{178 - 170}{65 / \sqrt{400}} = 2.46$$

Sigma yerine
 s kullanacağız.

2.4 0.06
 z tablosuna gidiyoruz.

④ P-value:



2.46 'dan büyük olma olasılığı 0.0069 mus. Ya bir değer bulduk. P-value oldukça küçük. Güçlü bir kanıt elde ettik. Fazlaıyla kanıt sunmuş olduk.

$$\text{P-value} < \alpha \\ 0.0069 < 0.05$$

soruda vermediği için default

* Scipy Kodu:

```
import scipy.stats as stats
1 - stats.norm.cdf(2.46)
```

⑤ Conclusion:

$\text{P-value} < \alpha$ olduğu için "Statistical Significant." H_0 hipotezinin reddedebiliriz. Elbette de yetenice kanıt olduğu için "Yeni sistem cost-effective."



Biz simdiye kadar Z test hesabi yapıp

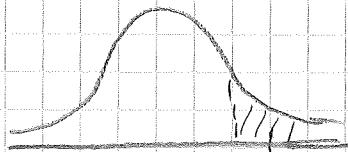
P -value bulduk ve

$P < \alpha$ karşılastırmasına göre sonucunu

bulduk.

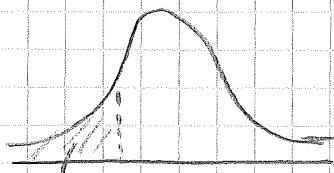
İkinci yol olarak :

α' dan Z test hesabi yapıp kritik Z (Z_c) sonucu ile karşılaştırıp da bulabiliriz.



Right tail test

$Z > Z_c$ ise red bölge sine düşer.



Left tail test

$Z < Z_c$ ise red bölge sine düşer

Z_c değeri α' dan hesaplanıyor mu? ama görmedik.

SMALL SAMPLE - EXAMPLES)

Bon Air ELEM has 1000 students. The principal of the school thinks that the average IQ of students at Bon Air is at least 110.

To prove her point, she administers an IQ test to 20 randomly selected students.

Among the sampled students, the average IQ is 108 with a standard deviation of 10.

Based on these results, should the principal accept or reject her original hypothesis?

$$\alpha = 0.01$$

Bir okulun 1000 öğrencisi var. Okul müdürü, öğrencilerin IQ'larının en az 110 olduğunu düşündür. Buğu kanıtlamak için rastgele 20 öğrenciyi IQ testi yapılıyor.

$$\text{Sample mean} = 108 = \bar{x}$$

$$\text{Standard sopma} = 10 = s \quad (\text{Sample std})$$

$$H_0 = 110$$

$$n = 20$$

① Assumptions:

Small sample ve populasyon standart偏差ını bilmiyoruz.

İnden t-distribütöünü kullanabiliyoruz.

② Hypothesis:

$$H_0: M = 110$$

$$H_a: M < 110$$

Müşteri, H_0 'nın en az 110 olduğunu söylemiştir.
110'dan küçük bulursak müsterinin iddiasını
çürüteceğiz.

③ Test Statistic:

$$t = \frac{\bar{x} - M}{S/\sqrt{n}} = \frac{108 - 110}{10/\sqrt{20}} = \frac{-2}{2.236} = -0.894$$

t değerini kritik t değerini karşılastırma yolunu
seviyoruz. t kritik değeri, α ile bulunur:

$$\text{stats.t.ppf}(0.01, df=19)$$

Olasılıkta ($\alpha=0.01$)
değer hesabi yapılmıştır

$n-1$ 'den $20-1=19$

$df = \text{degrees of freedom}$

Output : -2.539

$$t = -0.894 \quad t\text{-kritik} = -2.539$$

t değeri, t -kritik değerin ötesinde kaldı. Yani kabul bölgesinde kaldı. "Fail to reject hypothesis."
 H_0 'ı reddedemedik.

(4) P-Value :

Import scipy.stats as stats

stats.t.cdf(-0.894, 19)

$$t_c = -2.539 \quad t = -0.894$$

↓
 Sol tarafa
 geçemedi!
 (Fail to reject)

t değerinde olağanipa gitmediğimiz
 için cdf kullandık.

Output : 0.1913

↓
 P-value küçük değil. 0.01 seviyesinin çok ötesinde
 "Fail to reject"

↓
 t-kritik yöntemiyle da P-value yöntemiyle de aynı
 sonucu bulduk

Significant Test Özet

- (*) Hangi testi yapacağımızı seciniz.
- (*) Hipotezlerini olusturuyoruz. (H_0 , H_a)
- (*) Test istatistikini hesaplıyoruz (t or z distribution)
- (*) P-value hesaplıyoruz.
- (*) P-value ile α yi kıyaslıyoruz.
- (*) Sonuca göre H_0 red veya H_0 red değil gibi bir sonuc buluyoruz.

$H_0 \rightarrow$ Esitlik Yerine kurul.

Daha once kabul edilen bir sey varsa onun Yerine kurul.

(No difference - No correlation)

$H_a \rightarrow H_0$ hipotezinin dislayan şekilde kurulur.

$H_0 \neq H_a \rightarrow$ Two tail

$H_0 > H_a \rightarrow$ Right tail }
 $H_0 < H_a \rightarrow$ Left tail } One tail

EXAMPLE

Muzzle velocities of eight shells tested with a new gunpowder, along with the sample mean and sample standard deviation, $\bar{y} = 2959$ and $s = 39.1$

The manufacturer claims that the new gunpowder produces an average velocity of not less than 3000 feet per second.

Do the sample data provide sufficient evidence to contradict the mean manufacturer's claim at the .025 level of significance? $\alpha = 0.025$

8 mermiin nolu ağız histan yeni bir barut ile test edilmiştir.

$$\text{Sample mean} = 2959 = \bar{x}$$

$$\text{Sample std} = 39.1 = s$$

"Bu yeni barut saniyede 3000 feet' den daha az bir hızı sahip değildir." Bu kanıtlanmak için 8 mermi toplamı ve bu bir significance level = $0.025 = \alpha$ belirlenmiş.

$$\bar{x} = 2959$$

$$s = 2959$$

$$\alpha = 0.025$$

① Assumptions:

Değişkenler sayısal ✓

Örnekler rastsal seçilmiştir ✓

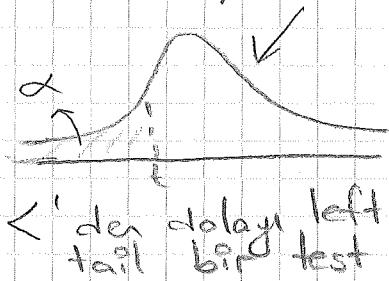
Populasyon dağılımı normal dağılmış ✓

} t-test koşulları sağlanmış.

② Hypothesis:

$$H_0: \mu = 3000$$

$H_a: \mu < 3000 \rightarrow$ Dediğimizdeki 3000'den küçük olamaz denmiş. Küçük bir değer bulursak H_0 hipotezini reddedip



$\mu <$ den dolayı left tail bir test

Aşında sample ort = 2959 olarak vermiş. Ortalama zaten 3000'den küçük ama elmine deki sample sayısı çok az ve bir std sapma var. Burda örnekleme hatası da yapılmış olabilir. Bu yüzden pesin hükmüle bir sonucu varmaya test esnasına geçiyorsunuz.

③ Test Statistic :

$$t = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{2959 - 3000}{39.1 / \sqrt{87}} = -2.966$$

Acaba bu değer yeterince küçük mü? Red bölge sine düşüyor mu?

Serbestlik derecesi $n-1 = 8-1 = 7$ olan bir t dağılımı incelessek bu sorunun cevabını bulabiliriz.

$$t = -2.966$$

④ P-Value :

```
from scipy import stats
stats.t.cdf(-2.966, 7)
```

Output : 0.014 → P-value

$$\alpha = 0.025 \text{ Pdi} \quad P\text{-value} = 0.014$$

⑤ Conclusion :

$$P\text{-value} < \alpha$$

$0.014 < 0.025$ Elpmizde yeteri kadar konut var.
"Reject the Null" dedik.

"0.025 antamlik olasılığında merminin sonijedeki hızı 3000 feet'ten azdır." H_0 hipotesini reddediyoruz.

Independent Sample T-Tests

İki grupta toplanıyor ve bunlar birbirinden bağımsızdır.
Mesela erkekler ve kadınlar grubu toplayırsınız ve
değişken isterinden test yapıyorsunuz.

Salary	Gender
35.000	Male
45.000	Male
38.000	Female
50.000	Female
28.000	Male

Bağımlı
değişken

Bağımsız değişkenler
(Male - Female)

M_1 : Male
 M_2 : Female

İlk yıl mezunları için, cinsiyete göre maaş差别 var mı?

① Assumptions :

İki grubun da normal dağılımını varsayıyoruz.

Rastgele örnek seçimi

Her iki grupta da sayılabilir değer var.

② Hypothesis :

$$H_0: M_1 = M_2$$

(No difference) → "Erkeklerin maaşı, kadınlarinkine eşittir" deyiğ.

$$H_a: M_1 \neq M_2$$

Equal Variance Assumed:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{sp \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad sp = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

$$df = n_1 + n_2 - 2$$

③ Test Statistic :

Equal Variances NOT Assumed:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1-1} \cdot \left(\frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2-1} \cdot \left(\frac{s_2^2}{n_2} \right)^2}$$

EXAMPLE

Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process.

Specifically, catalyst 1 is currently in use, but catalyst 2 is acceptable. Since catalyst 2 is cheaper, it should be adopted, provided it does not change the process yield.

A test is run in the pilot plant and results in the data shown in Table. Is there any difference between the mean yields?

Use 0.05, and assume equal variances.

<u>Observation Num</u>	<u>Catalyst 1</u>	<u>Catalyst 2</u>
1	91.50	89.19
2	90.18	90.95
3	92.18	90.46
4	95.39	93.21
5	91.79	99.19
6	89.02	97.04
7	94.32	91.07
8	89.21	92.75
		$\bar{x}_2 = 92.933$
		$s_2 = 2.98$
$\bar{x}_1 = 92.255$		
$s_1 = 2.39$		

Peki farklı katalizör var. Cat - 1 kullanında, Cat - 2 kabul edilebilir seviyede. Cat - 2'ye girmek istiyoruz, daha ucuz. Acaba H_0 'yi aracılıkla farklı var mı? Cat - 2, "Cat - 1 ile ortalama aynı verimliliği sağlıyor"sa kabul edilebiliriz." Düşünelim. Bir sonraki plot test yapılım 8'er tane.

① Assumptions:

Sayısal değişkenler var.

Gözlemler bağımsız.

Populasyonda normal dağılımlar

} Independent Sample
T-test kullanabiliriz

② Hypothesis:

$$H_0: M_1 = M_2$$

$$H_a: M_1 \neq M_2 \rightarrow \text{Eşitlik olmadığı için two tail test}$$



③ Test Statistic:

$$n_1 = 8 \quad s_1 = 2.39$$

$$n_2 = 8 \quad s_2 = 2.98$$

"Equal variance formülü" (3'ü) kullan" diye soruda belirtildi:

$$S_p = \sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{4 \cdot (2.39)^2 + 4 \cdot (2.98)^2}{8 + 8 - 2}} = 2.70$$

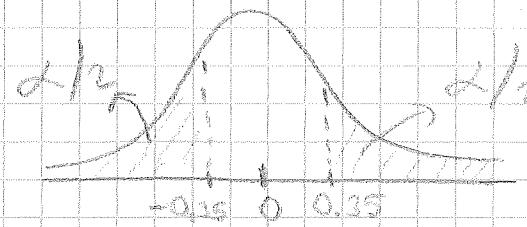
$$\bar{x}_1 = 92.255 \quad \bar{x}_2 = 92.733 \quad S_p = 2.70$$

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{92.255 - 92.733 - 0}{2.70 \sqrt{\frac{1}{8} + \frac{1}{8}}} = -0.35$$

Degrees of freedom (df) kaan olan bir t dağılımı kullanacağım?

$$\begin{aligned} n_1 - 1 &= 8 - 1 = 7 \\ n_2 - 1 &= 8 - 1 = 7 \\ df &= 14 \end{aligned}$$

$t = -0.35$ bulduk



P-value hesabını yapmadan da burdan karar verebiliriz.

Büyük bir değer çıktı. Yani "no difference". Elimize de "no difference", reddetmek için kanıt yok. Yani Cat-1 ile Cat-2 arasında fark yoktur, Cat-2'ye geçilebilir diyoruz.

④ P-values :

import scipy.stats as stats

2 * stats.t.cdf(-0.35, 14)

↙

2 ile çarpıyoruz, çünkü "two tail" bir test. ($\frac{\alpha}{2} + \frac{\alpha}{2}$)

Output : 0.731 → P-value

⑤ P-value > α

$$0.731 > 0.05$$

" ↴ "

"Fail to Reject the Null". Cat-1, Cat-2'de farklılığı
diyebildiğim yetki kalmam yok. Aynı verimlilikte sahip.
Cat-2'ye geçiş yapılmamalı.

ttest_ind(a, b, axis=0, equal_var=True)

equal not formülüm
kullanacağım False
yazarım Date.....

Subject:

```
from scipy import stats  
import pandas as pd
```

X sayısal değişken
(ortalamaları karşılaştıracağımız
değişken)

X = [91.50, 94.18, 92.18, 95.39, 91.79, 89.07, 94.72, 89.21
89.19, 90.95, 90.46, 93.21, 97.19, 97.04, 91.07, 92.75]

group = [1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2]
~~~~~  
        ↓                    ~~~~~  
        Cat - 1            Cat - 2

Scipy.org → Statistical tests → ttest\_ind kodunu  
kullanacağım.

df = pd.DataFrame ({ "yield": X, "group": group })

df oluşturduk. X'deki ilk 8 değer  
1. grubu, son 8 değer 2. grubu gibi.

test.stats.ttest\_ind(df[df.group==1]["yield"],  
df[df.group==2]["yield"] )

output: statistic = -0.353, pvalue = 0.728

test.statistic

output: -0.353

test.pvalue

output: 0.728

Aynı ayrı da bu değerleri yazdırabilirim.

## Decision

$$\alpha = 0.05$$

```
if test.pvalue < alpha:  
    print("Reject the null")
```

```
else:  
    print("Fail to reject")
```

Output: fail to reject  $\Rightarrow$  Formül kullanmadan scipy kizim P-Value hesapladı.

2.90 L

Diyelim ki elimde DataFrame yok ancak Sample means ve std sapmalar ve observation numbers ( $n$ ) var ise descriptive istatistikler de hesaplanır yapılabilir.

```
stats.ttest_ind_from_stats(mean1, std1, nobs1,  
                           mean2, std2, nobs2,  
                           equal_var=True)  
alternative='two-sided')
```

$$X_1 = 92.255 \quad \text{std } l = 2.39 \quad n_1 = 8$$

$$X_2 = 92.733 \quad \text{std2} = 2.98 \quad n_2 = 8$$

stats.ttest\_2ind\_from\_stats(mean1=92.255, std1=2.39, nobs1=8,  
mean2=92.733, std2=2.98, nobs2=8)

## ( $\alpha$ Is Known Example)

A product developer is interested in reducing the drying time of a primer paint. Two formulations of the paint are tested; formulation 1 is the standard chemistry, and formulation 2 has a new drying ingredient that should reduce the drying time.

From experience, it is known that  $\sigma=8$  minutes, and this inherent variability should be unaffected by the addition of a new ingredient.

10 specimens are painted with formulation 1, and another 10 " " " " formulation 2; the 20 specimens are painted in random order.

The sample drying times are  $\bar{X}_1 = 121$  and  $\bar{X}_2 = 112$  minutes, respectively.

What conclusions can the product developer draw about the effectiveness of the new ingredient, using  $\alpha=0.05$ ?

Formulasyon-1 ve Formulasyon-2 olarak 2 boyas ve bunların kurumsa zamanları var. Form-1 standart bir kuryadan olusuyor. Form-2'de kurumsa şuresini azaltın yed bilesen kullanılmış. Onceki datalaradan std sapma=8 olduğu biliniyor.

Form-1 için 10, Form-2 için 10 örnek boyanmış.

Kurumsa zamanları ölçülümiş. ( $\bar{X}_1, \bar{X}_2$ )  $\alpha=0.05$ 'e göre

Form-2, Form-1'den daha mi iyi?

$$\sigma = 8$$

$$n_1 = 10$$

$$n_2 = 10$$

$$\bar{x}_1 = 121$$

$$\bar{x}_2 = 112$$

$$\alpha = 0.05$$

### ① Assumptions :

2 grup için sayılabilir değerler

Rastgele örnek seçimi?

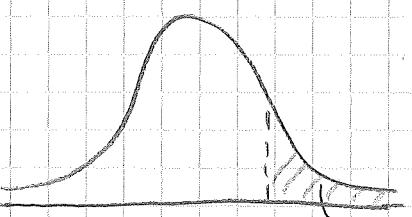
$\sigma$  biliniyor.

→  $\sigma$  bilindiği için 2-telli kullanacağız ama t-testi de kullanabiliyoruz.

### ② Hypothesis :

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 > \mu_2$$



Büyükük - küçükük var. One tail test

### ③ Test Statistic :

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{121 - 112 - 0}{\sqrt{\frac{8^2}{10} + \frac{8^2}{10}}} = 2.52$$

Z değeri 2 std sapmanın üzerindeyse extreme bir değerdir. Dolayısıyla 2 std sapmayı 2 std sola pidesen yaklaşık %95'lik bir alanı taranmış olurum.  $H_0$  reddedilebilir.

## ④ P-value :

Import scipy.stats as stats

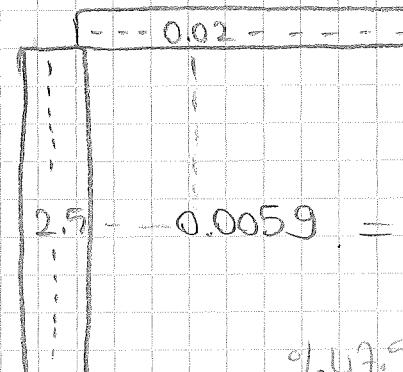
$1 - \text{stats.norm.cdf}(2.52)$

Output : 0.0059

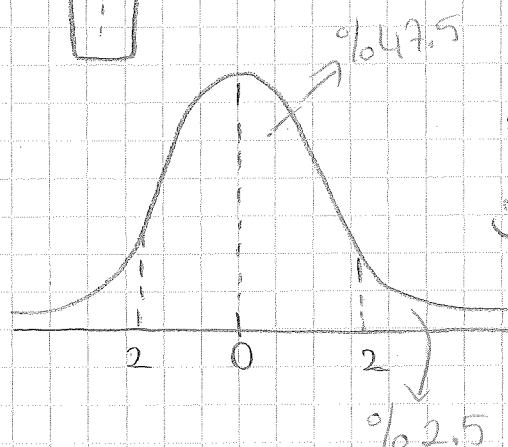
veya 2 tablosu ile

$$z = 2.52$$

$$\begin{array}{c} \swarrow \\ 2.5 \\ \searrow \\ 0.02 \end{array}$$



$$2.52 - 0.0059 = P\text{-value}$$



$z = 2.52$ 'ye göre 2'te sol, 2 std'da sağa  
giderken %95lik bir alan olur. Sağ tara-

fa  $\%0.0475$ , sol tarafa  $\%0.0475$

düzen. 0 zaman kırınlıklara

$\%50 - \%0.0475 = \%2.5$  kalır.

## ⑤ Conclusion :

$P\text{-value} < \alpha$

$0.0059 < 0.015 \rightarrow \text{"Reject the Null"}$

" $H_0$  kabul edilebilir. Form-2'deki gen bilgileri kullanma  
Zamanını aza getir." diye biliriz.

Subject:

Date: .....

## Large Sample, $\sigma$ is Unknown

A study was interested in determining if an exercise program had some effect on reduction of blood pressure in subjects with abnormally high blood pressure. For this purpose a sample of  $n_1 = 500$  patients with abnormally high blood pressure were required to adhere to the exercise regime.

A second sample  $n_2 = 400$  of patients with abnormally high blood pressure were not required to adhere to the exercise regime.

After a period of one year the reduction in blood pressure was measured for each patient in the study.

$$\bar{x}_1 = 10.67 \quad s_1 = 3.895$$

$$\bar{x}_2 = 7.83 \quad s_2 = 4.224$$

Anormal dördde yıldızla 500 hasta  $\rightarrow$  1. grup  
" " " " " 400 hasta  $\rightarrow$  2. grup

1. grup egzersiz programına uyumuyor  
2. grup " " " uyumuyor.

Independent samples.  
Çünkü farklı hastalar  
seçildi.

Eğzersiz yapan 2. grubun konusunda ortalaması 1. grubun  
sonunda düşük çıktı. Bu, onlarda mi değil mi?

### ① Assumptions:

Rasigale örneklər secilməz

Bəzi məsləhətlər " "

$\sigma$  is known, 30'dan fazla örnək

↳ 2-testi yapacaqız.

### ② Hypothesis:

$$H_0: M_1 - M_2 = 0$$

$$H_a: M_1 > M_2 \rightarrow \text{One tail test}$$

↳ 1.grubun kan basıncı ort. 'si 2.gruptan  
küçük mü?

### ③ Test Statistic:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{10.67 - 7.83 - 0}{\sqrt{\frac{3.895^2}{500} + \frac{4.224^2}{400}}} = 10.37$$

Z tablosunda çok  
ya bir değer. Elimizde  
kanıt var. P-value hesabına  
bile gerek yok.

### ④ P Value:

import scipy.stats as stats

1 = stats.norm.cdf(10.37)

Output  $\Rightarrow 0.0 \rightarrow P\text{-value}$

## (5) Conclusion :

P-value  $< \alpha$

0.0  $< 0.05 \rightarrow$  "Reject the null"

$H_0$  hipotezinin reddedilebiliriz. "Egzersiz yapanların kan basinci daha düşük olur." diyebiliriz.

Sample'ın çok olması (900 kişi) lüte çok küçük bir P-value getirdi.

## Scipy ile GÖZÜM

```
from scipy import stats
```

```
stats.ttest_ind.from_stats(mean1=10.67, std1=3.895, nobs1=500,  
                           mean2=7.83, std2=4.224, nobs2=400,  
                           Tek tarafta olduğunu  
                           değiştirdik.   ↪ alternative='greater')
```

Output: statistic=10.467, pvalue=1.420

Eldeki koddaki örneklem kümelerinin düşüncesinden

nobs1=5 } 9 kişi üzerinde hesap yapılmıştı

nobs2=4 } Sonuç:

statistic = 1.048, pvalue = 0.165 çıktı.

Elinde yeterli kanıt olmadığı için hata payını yükseltti ve sonucu beni yanılttı.

## Confidence Interval for $M_1 - M_2$

Ortalamlar arası farkın güven aralığı

$$\bar{X}_1 - \bar{X}_2 + 2\alpha/2 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$10.67 - 4.83 + (1.960)$$

$$\sqrt{\frac{3.895^2}{500} + \frac{4.224^2}{400}}$$

→ Kan basıncı  
soruşturmasından

$$2.84 \pm 1.96(0.233765)$$

$$2.84 \pm 0.537$$

$$2.303 \text{ to } 3.337$$

↓

Bu aralık sıfırı kapsamadığı için  $H_0'$ , reddedilebilir.

Mesela -1 to 3 olsaydı, %95 güven aralığının sıfırı kapsamadığı için  $H_0'$ , reddedilemezdi.

## Dependent t-Test

İki grup var. İki grupta da aynı kişilere ait veriler var.

Aynı grub içinde Before-After yapılır.

### ① Assumptions:

- (\*) Tek bir örneğ üzerinde iki farklı ölçüm yapılır.
- (\*) Yine sayısal değişken üzerinde hesaplanır yapılır ama iki farklı süren olur. (Independent takı gibi grouping variable yok.)
- (\*) Aradaki fark normal dağılım kabul edilir. (Before-After)

| ID | Before | After | Difference |
|----|--------|-------|------------|
| 1  | 12     | 10    | 2          |
| 2  | 18     | 7     | 11         |
| 3  | 21     | 22    | 1          |
| 4  | 10     | 12    | -2         |
| 5  | 8      | 4     | 4          |



One-sample t-Test'e çok benzeyen.  
Çünkü tek bir satır üzerinde t-Test yapılıyor. 0 satır da difference satırı

(2) Hypothesis:

$$H_0: \mu_0 = 0 \quad (\text{Difference ort.} = 0)$$

$$H_a: \mu_0 \neq 0 \quad (\text{Difference ort.} \neq 0)$$

!  $\mu_0 = \mu_1 - \mu_2$  (Her bir satır için)

↓      ↓  
Before    After

(3) Test Statistic:

$$t = \frac{\bar{x}_{\text{diff}} - 0}{s_{\bar{x}}}$$

$$s_{\bar{x}} = \frac{s_{\text{diff}}}{\sqrt{n}}$$

$\bar{x}_{\text{diff}}$  → Fark sütunuun ortalaması

$s_{\text{diff}}$  → Fark sütunuun std sapması

Subject :

Date : .....

## EXAMPLE ,

An article reports a comparison of several methods for predicting the shear strength for steel plate girders.

Data for two of these methods, the Karlsruhe and Lehigh procedures, when applied to nine specific girders, are shown in the table below.

Determine whether there is any difference (on the average) for the two methods. ( $\alpha = 0.05$ )

| Girder | Karlsruhe Method | Lehigh Method | Difference |
|--------|------------------|---------------|------------|
| S1/1   | 1.186            | 1.061         | 0.125      |
| S2/1   | 1.151            | 0.992         | 0.159      |
| S3/1   | 1.322            | 1.063         | 0.259      |
| S4/1   | 1.339            | 1.062         | 0.277      |
| S5/1   | 1.2              | 1.065         | 0.135      |
| S2/1   | 1.402            | 1.178         | 0.224      |
| S2/2   | 1.365            | 1.037         | 0.328      |
| S2/3   | 1.537            | 1.086         | 0.451      |
| S2/4   | 1.559            | 1.052         | 0.507      |

Cilek tabakaların bilmen nesini, farklı zamanlarda iki farklı method kullanarak aynı şmekler üzerinde ölçümeler ve bunların farklıları olusmus.

İki method arasında istatistiksel anlamlı bir fark var mı?  $\alpha = 0.05$

### ① Assumptions:

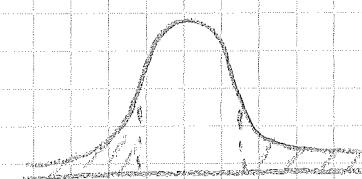
Bağımlı continuous değişkenler var ✓

Ölçümler bağımsız olmalı ✓

Fark istenilen normal dağılmali ✓

### ② Hypothesis:

$$H_0: \mu_D = 0$$



$$H_a: \mu_D \neq 0 \rightarrow \text{Güçlü kuyruk test olacak}$$

### ③ Test Statistic:

$$\bar{d} = 0.2469 \quad (\text{Difference'ların ortalaması})$$

$$s_d = 0.1350$$

$$n = 9$$

$$t_0 = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{0.2469}{0.1350 / \sqrt{9}} = 6.05$$

$t$  değeri oldukça extreme görülmüyor.

statsmodels → Scipy'nin alternatif görel bil  
kütiphaneyimis.

Subject :

Date : 2023/10/10

#### (4) P-value :

import scipy.stats as stats

$$2 * (1 - \text{stats.t.cdf}(6, 15, 8))$$

Output : 0.00027



Güft kuyruk olduğu  
için 2 ile çarpıldı.

$$\text{P-value} = 0.0003$$

Cok vs bir değer çıktı, elimde geterince kanıt var.

#### (5) Conclusions :

$$\text{P-value} < \alpha$$

$$0.0003 < 0.05$$



Antamli bir fark var.  $H_0$  hipotezi reddedebiliriz.

"Reject the Null"

Subject:

Grup ort. karşılaştırılır.

One-way ANOVA  $\rightarrow$  F istatistikî kullanılır.

Tek yönlü varyans analizi

Independence Sample t-Test'in gelişimi halidir.

Independence Sample t-test:

2 grubun ortalamaları karşılaştırılır. (Male - Female  
Yes - No)

One-way ANOVA:

3 veya daha fazla grubun ortalamaları karşılaştırılır.

| Sprint | Smoking |
|--------|---------|
| 5.1    | 0       |
| 7.8    | 2       |
| 7.1    | 1       |
| 8.6    | 2       |
| 4.9    | 0       |
| 7.7    | 1       |

Bapılı  
depikler

Baplısı  
grub

Nonsmoker  
(0)

Pastsmaiker  
(1)

Current  
smoker  
(2)

## ① Assumptions :

Continuous değişkenler.

Bağımsız gruplar 3 veya daha fazla.

Her bir bağımsız grup, kendisi içinde normal dağılmış.

Varyanslar homojen

## ② Hypothesis :

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k \quad (\text{Bütün ort. eşit})$$

$H_a$ : En azından bir  $\mu_i$  farklı

(Hepsinin birbirinden farklı olması gereklidir.)

En azından bir  $\mu_i$  farklı olsa yeterli.  $H_0$ 'yi reddederim.)

## ③ Test Statistic :

ANOVA TABLE

|                | Sum of squares    | df    | Mean Square       | F         |
|----------------|-------------------|-------|-------------------|-----------|
| Group(Between) | SSR               | $k-1$ | $MSR = SSR/(k-1)$ | $MSR/MSE$ |
| Error (Within) | SSE               | $n-k$ | $MSE = SSE/(n-k)$ |           |
| Total          | $SST = SSR + SSE$ | $n-1$ |                   |           |

$$F = \frac{\frac{MSR}{MSE}}{= \frac{\text{Gruplar arası varyans}}{\text{Grup içi varyans}}} \quad \begin{array}{l} \text{Tüm hesaplamalar} \\ \text{bu nü bulmaya yönelikis.} \end{array}$$

SSR  $\rightarrow$  Gruplar arası kareler toplamı.

$k-1 \rightarrow$  Serbestlik derecesi. (Grup sayısı - 1 = 2)

$$MSR = SSR / (k-1)$$

Mean squares (Gruplar arası varyans)

SSE  $\rightarrow$  Grup içi kareler toplamı.

$n-k \rightarrow$  Grup içi serbestlik derecesi. (Gözlen sayıısı - grup sayısı)

$$MSE = SSE / (n-k)$$

Mean square error (Grup içi varyans)

$$F = MSR / MSE$$

MSR ve MSE'yi program kullanarak hesaplıyor.

F istatistik değeri bulunuyor.

## One-way ANOVA nelerde kullanılır?

3 farklı tedavi yöntemine iliskin değerler,

3 farklı öğrenci grubu (Online - sınıf - bilgisayar destekli eğitim gibi)

### EXAMPLE

As an example, we analyze the Cushing's data set, which is available from the MASS package.

The type variable in the data set shows the underlying type of syndrome, which can be one of four categories:

adrenoma ( $n_1$ ),

bilateral hyperplasia ( $n_2$ ),

carcinoma ( $n_3$ ),

unknown ( $n_4$ ).

Our objective is to find whether the four groups are different with respect to urinary excretion rate of Tetrahydrocortisone.

$$\begin{aligned} n_1 &= 6 \\ n_2 &= 10 \\ n_3 &= 5 \\ n_4 &= 6 \end{aligned}$$

$$n = 27 \text{ (öğrenci sayısı)}$$

$$\left. \begin{aligned} M_1 &= 3.0 \\ M_2 &= 8.2 \\ M_3 &= 19.4 \\ M_4 &= 14.0 \end{aligned} \right\} \text{ort. 1ar}$$

F statistic' te 2 tane degrees of freedom var:

$$df_1 = 4 - 1 = 3$$

grup sayısı - 1  
(k-1)

$$df_2 = 24 - 4 = 20$$

toplam gözlem sayısı - grup sayısı  
(n-k)

### ① Assumptions:

Bağımlı continuous değişkenler

Bağımsız kategorik değişkenler

Her grup içinde normal dağılım

Varyanslar homojen

### ② Hypothesis:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$H_A$ : At least one  $\mu_i$  different

### ③ Test Statistic:

|                    | Sum of squares | df                 | Mean Square                     | F                                              |
|--------------------|----------------|--------------------|---------------------------------|------------------------------------------------|
| Group<br>(Between) | $SSR = 893.5$  | $k-1$<br>$4-1=3$   | $MSR = 893.5 / 3$<br>$= 297.8$  | $MSR / MSE =$<br>$297.8 / 92.3$<br>$= 3.226 *$ |
| Error<br>(Within)  | $SSE = 2123.6$ | $n-k$<br>$27-4=23$ | $MSE = 2123.6 / 23$<br>$= 92.3$ |                                                |
| Total              | $SST = 3017.1$ | $n-1 = 27$         |                                 |                                                |

#### ④ P-value :

$F = 3.226$  bulduk. Bu değerle karşılık gelen P-value bulacağınız.

import scipy.stats as stats

1 - stats.f.cdf( 3.226, dfn=3, dfd=23)

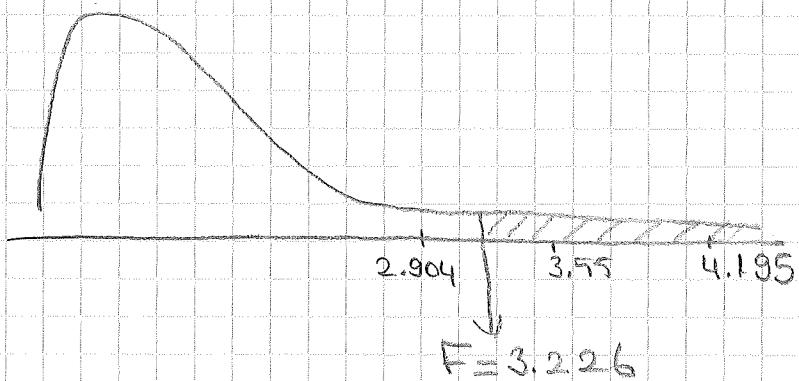
output : 0.041

! F dağılımında 96'i tane serbestlik derecesi var.

P-value <  $\alpha$

0.041 < 0.05  $\Rightarrow$  Reject the Null

P-value küçük çıktı. Anlamlı bir sonuc çıktı.  $H_0$  reddettik.



#### ⑤ Conclusions :

$P\text{-value} < \alpha$  bulduğumusak şire ; "bütün grupların aynı ortalamaya sahip oldukları hipotesini reddettik. Gruplar dan en az biri farklı." Hangisi hangisinden farklı olduğunu bilmiyoruz.