Statistics for Data Science

2023

By

Santosh Lohar

IBM Data Science Professional Certified

LinkedIn

Mo Tu We Th Fr Sa Su Date:

Statistics for

Data scientist / Data Analyst / Business Intelligence tools /

What is Statistics

⇒ Statistics is the science of collecting, Organizing and analyzing data for better decision making

Data ⇒ facts or piece of information that can be measured in a quantitative way.

Eg: The IQ of the class:
{92, 91, 90, 76, 75, 60}

Eg: Ages of students of class.
{30, 21, 21, 25, 29, 16, 15}

Types of Statistics

Descriptive Statistics

- It Consists of organizing and summarizing data.

Inferential Statistics

Technique where in we used the data that we have measured to form conclusions.

Eg: classroom of maths students 20 marks of the 1st Sem
{84, 72, 94, 59, 89, 90, 91}

Data science LinkedIn: Santosh-lohar

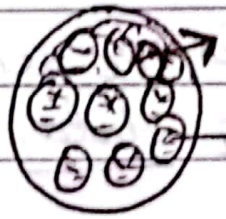
Mo Tu We Th Fr Sa Su Date:

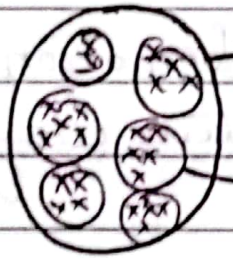
Descriptive statistics

→ What is the average marks of the students in the class? (from previous example data.)

Inferential Statistics:

Are the marks of the students of this classroom similar to the age of the maths classroom in the College?



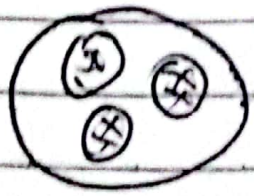


like election exit Poll.

Population (N) Sample (n)

Sampling Techniques:

① → Simple Random Sampling: Every member of the population (N) has an equal chance of being selected for your sample (n)



DATA SCIENCE

Ln: Santosh-lohar

Mo Tu We Th Fr Sa Su

Date:

② Stratified Sampling :- Where the Population (N) is split into non-overlapping groups (strata)

eg:

Gender : male
: female

Age-group : (0-10) (10-20) (20-40) (40-100)

Programming : Dot net PHP Python

Not-overlapped

③ Systematic Sampling (Thanos)

(N) ⟶ nth individual

eg: Mall ⇒ survey (covid)

↳ 8th person ⇒ survey.

④ Convenience Sampling (only those People)

↳ Survey → Data Scienie

RBI (House-hold survey)

↳ women

eg: Drug Tested

Mo Tu We Th Fr Sa Su

Date :

Variables :

A variable is a property that can take on any value : { 182, 175, 172, 170, 165 } Height
eg : Weight : { 50, 55, 60, 70, 90, 100 - }

types of variables

Quantitative variable

↓

measured numerically {add, sub, mul, div}

Qualitative variable (Categorical variable)

→ Based on some character istics we can derive categorical variables }

eg : Gender : male
female

Eg : IQ

0-10 → less IQ

10-50 → medium IQ

50-100 → good IQ

Blood Group

A +ve O-

T-Shirt-Size :

L,
XL
XXL
M
S

www.linkedin/in/santosh-lohar

Mo Tu We Th Fr Sa Su Date:

Quantitative

Discrete Variable
Eg: Whole number
No. of Bank A/c
eg: {2, 3, 4, 5, 6, 7 ...}
Total number of children
Total No. of family.

Continuous Variables.
eg: Height : 172.5, 170.5, 166.2
Weight : 110kg, 77.5kg, 98.75
Rainfall : 1.1, 2.3, 1.25, 1.35

Eg: What kind of variable Gender is? Categorical
" " " " marital status? : Categorical
River length : Continuous
Pin Code : { Discrete or categorical }

Variable measurement scales:

4 types of measured variable

1. Nominal data = { Categorical data : colors }
2. Ordinal data : order of the data matters value not
3. Interval : order matters value also matters (zero not A)
4. Ratio :

eg: Students (marks)	Ranks	
100	1.	ordinal data
96	2	
57	4	
85	3	

Mo Tu We Th Fr Sa Su Date:

Frequency Distribution:

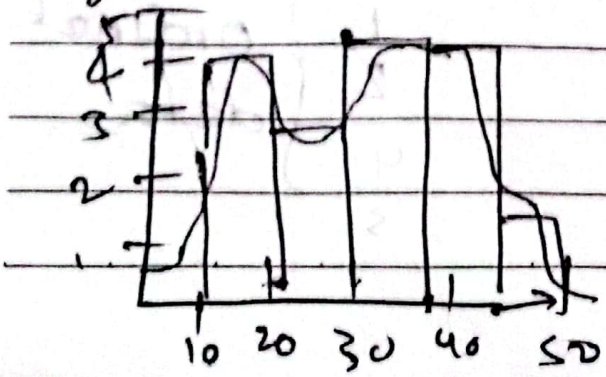
Sample dataset: Rose, lily, Sunflower, Rose, Lily, Sunflower, Rose, lily, Rose.

Flowers	frequency	Cummulative freq
Rose	4	4
lily	3	7
Sunflower	2	2

① BAR GRAPH:

② HISTOGRAM - Continuous

Ages: {10, 12, 15, 16, 18, 24, 25, 35, 36, 37, 40...



Bins

default bins = 10

DATA SCIENCE — SANTOSH-LOHAR — LinkedIN

Mo Tu We Th Fr Sa Su — Date:

Measure of Central Tendency:

* Arithmetic mean for Population & Sample.

Population (N) — Sample (n)

Average

eg: (N) (n)

$X = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$

$\mu = \sum_{i=1}^{N} \frac{(x_i)}{N}$ $\quad \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$

$\mu = \frac{1+1+2+2+3+3+4+5+5+6}{10}$ $\quad \bar{x} = 32/10$

$\bar{x} = 3.2$

$\mu = 32/10 = 3.2$

* Central Tendency:

① Mean ② Median ③ Mode

↳ Refers to the measures used to determine the center of the distribution of data.

eg: {1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100} → outlier.

Mean after outlier

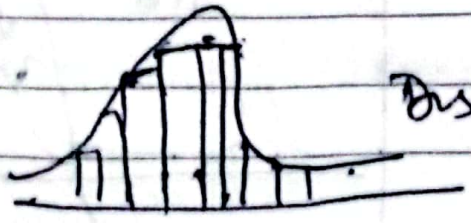
$\text{Mean} = \frac{32+100}{11}$

$= 12$

$\mu = 3.2$ ← Previous

After outlier 100

$\mu = 12$



Distribution

Mo Tu We Th Fr Sa Su Date :

Median :

{1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100, 112}

Step-1 : Sort the numbers.

if odd number mid number will result
if even number take average between two numbers.

Measure of Dispersion
↓
(Spread)

↳ Variance (σ^2)
↳ Standard deviation (σ)

*** Variance :**

Population Variance

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

Sample Variance

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

eg:

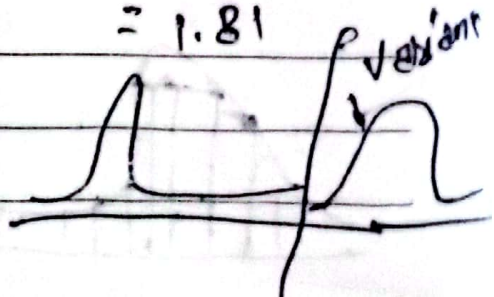
x	μ	$x-\mu$	$(x-\mu)^2$
1	2.83	-1.83	3.34
2	2.83	-0.83	0.6889
2	2.83	-0.83	0.6889
3	2.83	0.17	0.03
4	2.83	1.17	1.37
5	2.83	2.17	4.71
			10.84

$\mu = 2.83$

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

$$= \frac{10.84}{6}$$

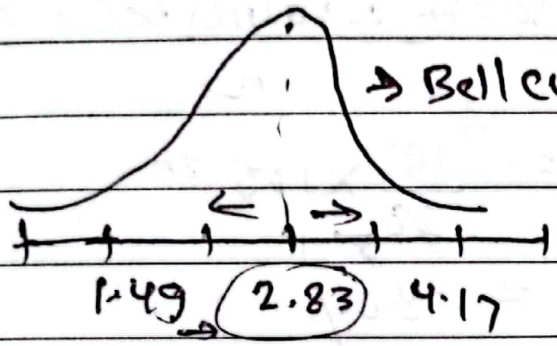
$$= 1.81$$



Mo Tu We Th Fr Sa Su Date :

Standard deviation(σ) = $\sqrt{\text{variance}} = \sqrt{\sigma^2}$

$= \sqrt{1.81} = 1.345$



2.83	2.83
1.34	− 1.34
4.17	1.49

⇒ 68% 1.5 σ from the μ

Percentiles and Quartiles (find author)

Percentage: 1, 2, 3, 4, 5

% of the numbers that are odd?

$$\% = \frac{\#\text{ of numbers that are odd}}{\text{Total numbers}}$$

$= 3/5 = 0.6 = 60\%$

Percentile: A Percentile is a value below which a certain percentage of observations lie.

eg: Dataset: 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

What is the Percentile ranking of 10?

here : $n = 20$

Mo Tu We Th Fr Sa Su Date:

from the given Problem:

$x = 10$

Now, Percentage rank of x = $\frac{\#\text{ of values below } x}{n} \times 100$

$= \frac{16}{20} \times 100$ (4, 20, 5)

$= 80\%$.

(X)

80% of the entire distribution is less than 10

find the percentile rank of 11 ?

here, $x_1 = 11$

$n = 20$

Now! Percentile Rank of x_1 = $\frac{\#\text{ of values below } x_1}{n} \times 100$

$= \frac{17}{20} \times 100$ (5)

$= 85\%$.

Hence, 85% of the entire distribution is less than 11

* what value exists at percentile ranking of 25% ?

Value = $\frac{\text{Percentile}}{100} \times (n+1)$

$= \frac{25}{100} \times (21) = 5.25$ → Index Position.

→ (5) value.

Locate value of the distribution

Mo Tu We Th Fr Sa Su Date:

✱ Five Number Summary ✱

↳ minimum
↳ first quartile (Q1)
↳ Median
↳ Third quartile (Q3)
↳ maximum.

Removing Outlier

eg: {1, 2, 2, 2, 3, 3, 4, 5, 55, 6, 6, 6, 7, 8, 8, 9, 27} → outlier

(5 index) (15 index)

[Lower fence ⟷ higher fence]

lower fence = $Q_1 - 1.5(IQR)$
Higher fence = $Q_3 + 1.5(IQR)$

$Q_3 = 75\%$
$Q_1 = 25\%$

Interquartile Range (IQR) = $Q_3 - Q_1$

Value of 25% = $\frac{25}{100}(n+1) = \frac{25}{100}(19+1) = 5$ (index)

$= 3$

Value of 75% = $\frac{75}{100}(19+1) = \frac{75}{100} \times 20^{15} = 15$ (index)

$= 7$

Mo Tu We Th Fr Sa Su

Date :

Now,

Interquartile Range (IQR) = $Q_3 - Q_1$

$= 7 - 3$

$= 4$

Again, lower fence = $Q_1 - 1.5 \times (IQR)$

$= 3 - 1.5(4)$

$= 3 - 6$

$= -3$ //

Higher fence = $Q_3 + 1.5(IQR)$

$= 7 + 1.5(4)$

$= 13$

[lower fence $\longleftrightarrow$ Higher fence]

[$-3 \longrightarrow 13$

* From the above distribution remove the elements as per lower & higher fence calculation value. So, distribution elements lies which is between -3 & 13 fence.

New distribution

Remaining data { 1, 2, 2, 2, 3, 3, ~~4~~, 5, 5, 5, 6, 6, 6, 6, 7, 6, 8, 9, ~~27~~ } Removed

median = $\frac{5+5}{2} = 10/2 = 5$

IN : SANTOSH - LOHAR

Mo Tu We Th Fr Sa Su

Date :

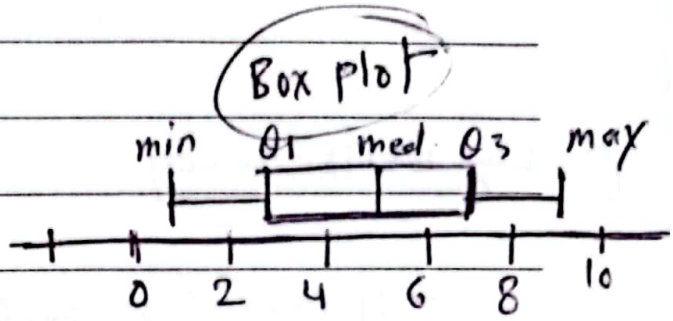
Minimum = 1

Q_1 = 3

median = 5

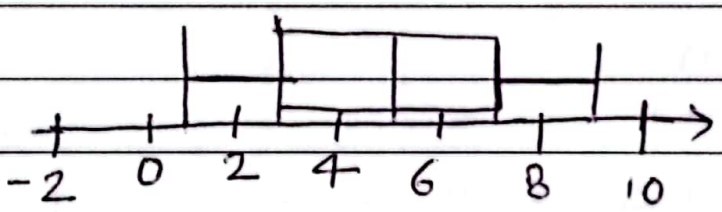
Q_3 = 7

maximum = 9



Box Plot

Box plot is used to determine Outlier



Application : Data Visualization. (Boxplot)

Why Sample variance is n-1 ?

{

linked in : Santosh-Lohar

Mo Tu We Th Fr Sa Su

Date :

Advance statistics

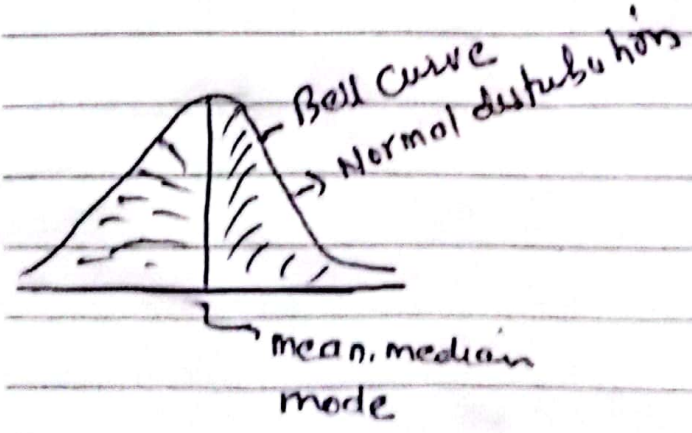
① ↳ Distribution

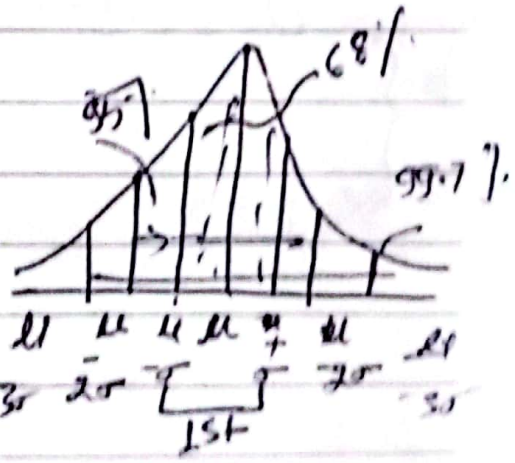
- Normal distribution
- Standard Normal distribution.
- Z score
- Log normal distribution.
- Bernoullis distribution.
- Binomial distribution.

✳ Distribution

eg: Ages = { 24, 26, 28, 29, 30, 32 ..... }

* Gaussian / Normal Distribution





68% dataset distribution lies in 1st standard deviation

IN : SANTOSH-LOHAR

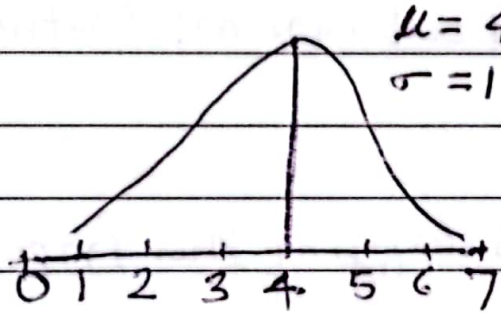
Mo Tu We Th Fr Sa Su Date:

Emperical formula.

68% - 95% - 99.7% Rule

1st standard deviation values | 2nd standard deviation values | 3rd standard deviation values.

eg:



4.5 → standard deviation

+0.5 SD.

4.75 ??

$$\text{Z-score} = \frac{x_i - \mu}{\sigma} = \frac{4.75 - 4}{1} = 0.75 \text{ sd}$$

Dataset = {1, 2, 3, 4, 5, 6, 7}

find z-score, $z = \frac{x_i - \mu}{\sigma}$

Here,

$\mu = 4, \sigma = 1$

Now, $z(1) = \frac{1-4}{1} = -3$

$z(2) = \frac{2-4}{1} = -2$

$z(3) = \frac{3-4}{1} = -1$

$z(4) = \frac{4-4}{1} = 0$

$z(5) = \frac{5-4}{1} = 1$

$z(6) = \frac{6-4}{1} = 2$

$z(7) = \frac{7-4}{1} = 3$

Now, After z-score new data set will be

$z = \{-3, -2, -1, 0, 1, 2, 3\}$

linked in : Santosh-lohar

Mo Tu We Th Fr Sa Su Date

dataset = {1, 2, 3, 4, 5, 6, 7} — Normal Distribution

Where, $\mu = 4$

$\sigma = 1$

After Z-score calculation

dataset_z-score = {-3, -2, -1, 0, 1, 2, 3} — Standard Normal Distribution

Where, $\mu = 0$

$\sigma = 1$

Satisfying this Property.

Practical Application:

DATASET: ← unit (standardization) zscore

Age (yr)	Salary (Rs.)	Weight (kg)
24	40K	60
25	50K	70
26	60K	49
27	45K	59

Normalization {$\mu = 0, \sigma = 1$}

↙ (-1 to 1) | (0 to 1)

IN = Santosh · Lohar

Mo Tu We Th Fr Sa Su Date :

Practical eg :-

{ IND Vs SA } {ODI Series}

ODI Series 2021 (CRICKET)

Scores Average 2021 = 250
Standard deviation = 10
Team final score = 240

ODI Series 2020 (CRICKET)

Series Avg 2020 = 260
standard deviation = 12
Team final score = 245

Q1

Compare the both the series in which year Rishabh final score was better ?

Now. 2021

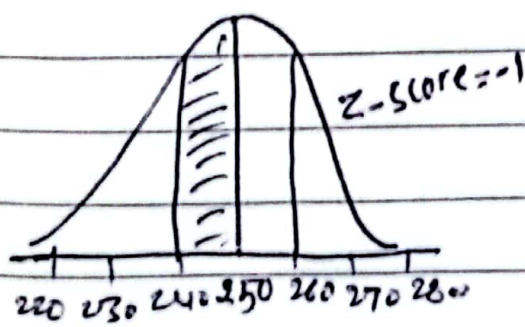
$$z\text{-score} = \frac{x_1 - \mu}{\sigma} = \frac{240 - 250}{10} = -1$$

2020

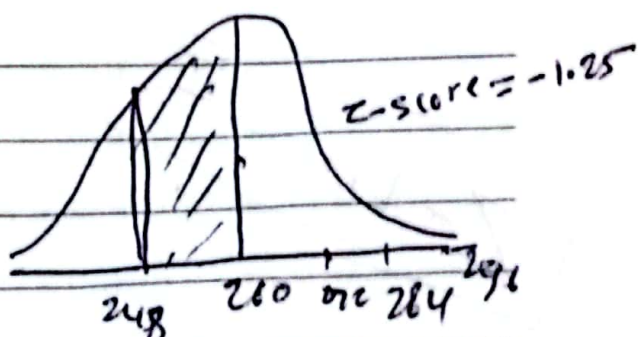
$$z\text{-score} = \frac{x_1 - \mu}{\sigma} = \frac{245 - 260}{12} = 1.25$$

In 2021 :

$\mu = 250, x_1 = 240, \sigma = 10$



$\mu = 260, x_1 = 245, \sigma = 12$

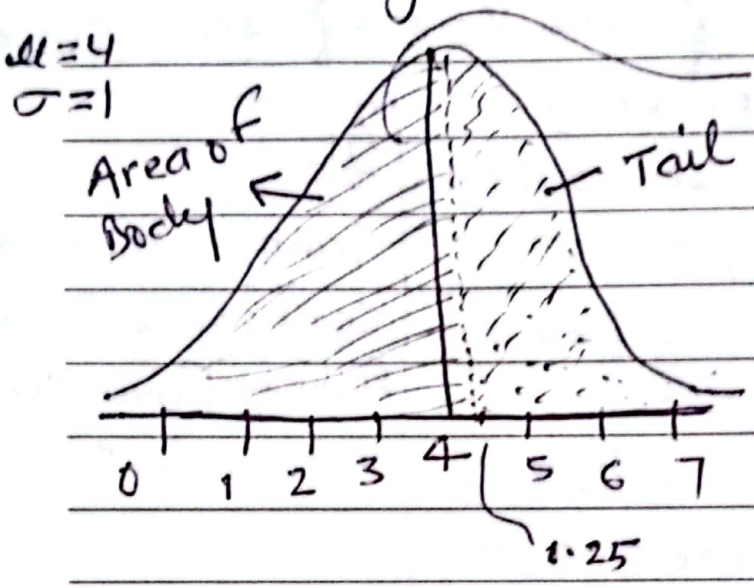


iN Santosh - Lohar

Mo Tu We Th Fr Sa Su Date

Assignment

What Percentage of scores falls above 4.25 ?



1-Left Area

$$\text{z-score} = \frac{x_1 - \mu}{\sigma} = \frac{4.25 - 4}{1} = 0.25$$

Locate z-table

Now,

Right Area (Tail) = 1 - left Area

= 1 - 0.59871

= 0.4013.

= 40%

Z-TABLE

→ PRACTICAL

LINKEDIN SANTOSH LOHAR

Mo Tu We Th Fr Sa Su Date:

PRACTICAL

①
```
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```
} import Python Library

②
```
import statistics
```

③
```
df = sns.load_dataset('tips')
```

④
```
df.head()
```

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	female	No	Sun	Dinner	2
1	10.34	1.66	male	No	Sun	Dinner	3
2	21.01	3.50	male	No	Sun	Dinner	3
3	23.68	3.31	male	No	Sun	Dinner	2
4	24.59	3.61	female	No	Sun	Dinner	4

⑤
```
np.mean(df['total_bill'])
```
⇒ 19.78594...

⑥
```
np.median(df['total_bill'])
```
⇒ 17.795

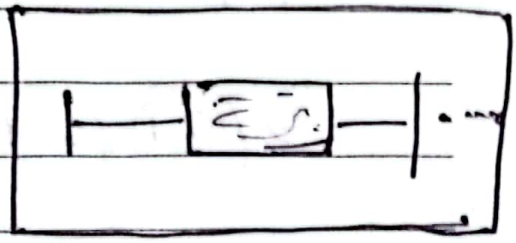
⑦
```
statistics.mode(df['total_bill'])
```
⇒ 13.42

IN SANTOSH-LOHAR

Mo Tu We Th Fr Sa Su Date:

Box plot (import seaborn as sns)

```
sns.boxplot(df['total_bill'])
```



```
sns.histplot(df['total_bill'])
sns.histplot(df1['sepal_length'], kde=True)
```

↓

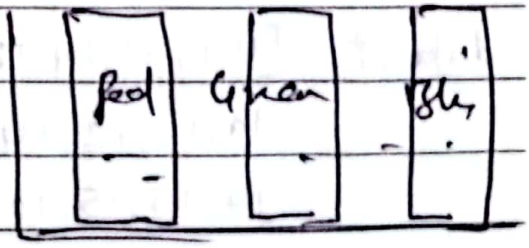
sepal_width

Before 'sepal_length' load data.

```
df1= load_dataset('iris')
```

Bar Graph → countplot

```
sns.countplot(df1['species'])
```

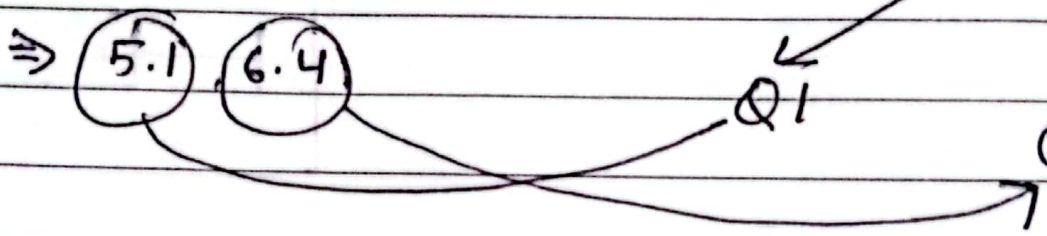


linked in : Santosh Lohar

Mo Tu We Th Fr Sa Su Date :

Percentile

↳ np.percentile(df1['sepal-length'], [25, 75])



Note : we can change parameter as we need.

IQR

* Outliers

① import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline.
] import Python Library.

① dataset = [11,10,12,14,12,15,14,13,15,102,12,14,17,19,107,
10,13,12,14,12,108,12,11,14,13,15,10,15,12,10,
14,13,15,10]
] → Define our Dataset

IN SANTOSH-LOHAR

Mo Tu We Th Fr Sa Su Date :

Find outlier | z-score

③

```
Outliers = []
def outlier_find(data):
    threshold = 3
    mean = np.mean(data)
    std  = np.std(data)

    for i in data:
        z-score = (i-mean)/std
        if np.abs(z-score) > threshold:
            outliers.append(i)

    return outliers
```

④

```
Outlier_find(dataset)
↳[102, 107, 108]
```

Python Code

linked iN : santosh-lohar

Mo Tu We Th Fr Sa Su Date

IQR → Interquartile Range

1. ↳ sort the particular dataset
2. ↳ Calculate Q1 & Q3
3. ↳ IQR (Q3 - Q1)
4. ↳ Find lower fence (Q1 - 1.5 (IQR))
5. ↳ find higher fence (Q3 + 1.5 (IQR))

```
# dataset = sorted(dataset)
  dataset
```

[Display Dataset]

```
# Q1, Q3 = np.percentile(dataset, [25, 75])
# Print(Q1, Q3)
```

⇒ 12.0 , 15.0

Q1 Q3

find lower & higher fence

```
# lower_fence = Q1 - 1.5(IQR) = 7.5
# higher_fence = Q3 + 1.5(IQR) = 19.5
```

where, IQR = Q3 - Q1 ⇒ 3

Mo Tu We Th Fr Sa Su

Date

import seaborn as sns

sns.boxplot(dataset)

[Box Plot here]

Mo Tu We Th Fr Sa Su Date

Probability

Probability is a measure of the likelihood of an Event.

eg. Roll a dice {1, 2, 3, 4, 5, 6}

$Pr.(6) = 1/6 = \frac{\#\text{ of way an event can occur}}{\#\text{ of Possible outcome}}$

eg. Toss a coin {H, T}

$Pr.(H) = 1/2$

Addition Rule (Probability, "or")

* Mutual Exclusive Event *

→ Two events are mutual exclusive if they can not occur at the same time

eg. Rolling a dice {1, 2, 3, 4, 5, 6}

{1}

eg. Tossing a Coin {H, T}

{H}

* Non-Mutual Exclusive Event *

→ Multiple events can occur at the same time

eg. Deck of cards {Q, ♡}

↓red color

[IN] [SANTOSH-LOHAR]

eg. If I toss a coin, what is the probability of the coin landing on heads or tails?

→ Mutual exclusive (Addition Rule we called)

$Pr(A \text{ or } B) = P(A) + P(B)$

$= 1/2 + 1/2$

$= 1$

$\therefore Pr(A \text{ or } B) = 1$

eg. Roll a dice {1, 2, 3, 4, 5, 6}

$Pr(1 \text{ or } 3 \text{ or } 6) = Pr(1) + Pr(3) + Pr(6)$

$= 1/6 + 1/6 + 1/6$

$= 3/6 = 1/2 = 0.5$

eg. Non-mutual exclusive

→ you are picking a card randomly from a deck what is the probability of chosing a card that is queen or a heart?

⇒ Non-mutual exclusive

$P(Q) = 4/52$, $P(♡) = 13/52$, $Pr(Q \text{ and } ♡) = 1/52$ → multiplication

Addition Rule for non-mutual exclusive events.

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

$= \frac{4}{52} + 13/52 - 1/52$

$= 16/52 = 4/13$

LINKED IN : SANTOSH-LOHAR

Mo Tu We Th Fr Sa Su Date

Multiplication Rule

↳ Independent Events

eg. Rolling a dice {1,2,3,4,5,6}
Each & Every are independent
It may be {1,1,2 ...}

↳ Dependent Events

eg.

● ● ● ●
○ ○ ○ ○ ○

Pr(●) = 4/9
Pr(○) = 5/9

Now, After Pickup/taken out a ball (○) from bag

● ● ● ●
○ ○ ○ ○

Pr(●) = 4/8
Pr(○) = 4/8

Independent Events

eg. What is the probability of rolling a "5" and then a "4" in a dice?

Ans ⇒ Independent event

Multiplication Rule

Pr(A and B) = P(A) × P(B)
= 1/6 × 1/6
= 1/36

Mo Tu We Th Fr Sa Su Date :

eg. What is the probability of drawing a Queen and then a Aces from a deck of cards?

Ans: Dependent

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

conditional Probability.

$$= 4/52 * 4/51$$

Permutation and Combination

eg: school trip {chocolate Factory} → {Dairy milk, Jems, silk: dairy, milk, Jucy}

→ $6 \times 5 \times 4 = 120$

formula:

$$nP_r = \frac{n!}{(n-r)!} = \frac{6!}{(6-3)!} = \frac{6 \times 5 \times 4 \times 3!}{3!}$$

$$= 120$$

Dairy Gems Eclair

Combination: $^nC_r = \frac{n!}{r!(n-r)!} = \frac{6!}{3!(6-3)!}$

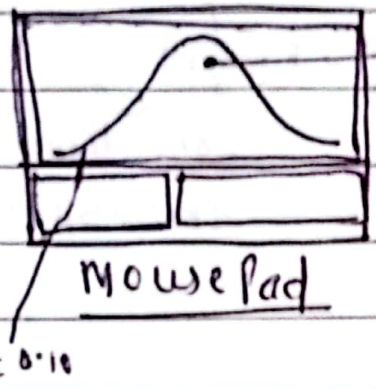
formula:

$$= \frac{6 \times 5 \times 4 \times 3!}{3 \times 2 \times 1 \times 3!}$$

$$= 20$$

Mo Tu We Th Fr Sa Su Date

P value



Every 100 times I touch the mousepad 80 times I touch this specific region

Hypothesis testing

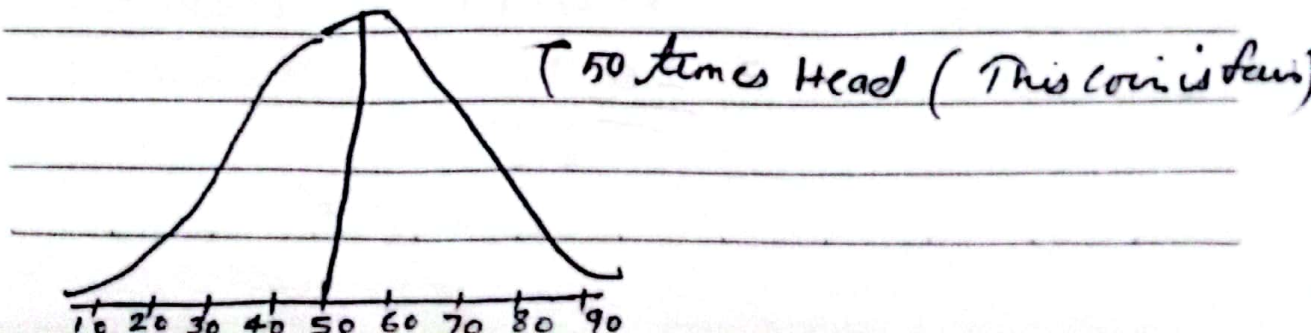
Coin → Test whether this coin is a fair coin or not by perform 100 tosses

$P(H) = 0.5$, $P(T) = 0.5$

50 times head (The coin is fair)

Hypothesis Testing.

① → Null Hypothesis Testing : Coin is fair
② → Alternative Hypothesis : Coin is unfair
③ → Experiment
④ → Reject or Accept the Null Hypothesis.



Mo Tu We Th Fr Sa Su Date :

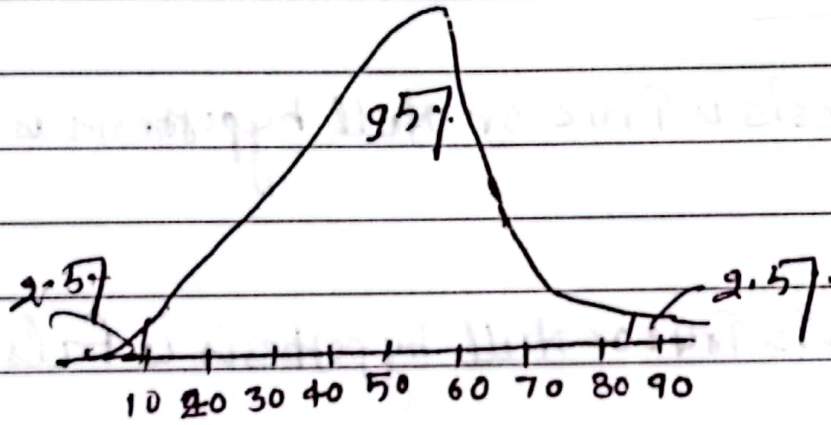
Significance Value : α (Alpha) α

$\alpha = 0.05$ (This is defined by Domain Expert)

Here, $1 - 0.05 = 0.95 = 95\%$

Which means,

Confidenie Interval (CI) = 95%.



Note! The experiment is falls between 95%. then it should be assumed fair (coin)

Note! Outside 2.5%. it is unfair (coin)

Mo Tu We Th Fr Sa Su Date :

Type 1 and Type 2 Error

Null Hypothesis (H_0) = Coin is fair
Alternative Hypothesis (H_1) = Coin is not fair

Reality Check

↳ Null hypothesis is True or Null hypothesis is false

Decision

Null hypothesis is True or Null hypothesis is false.

Outcome 1 We reject the Null Hypothesis, When in reality it is false → yes

Output 2 : We reject the Null Hypothesis, When in reality it is true → Type 1 Error

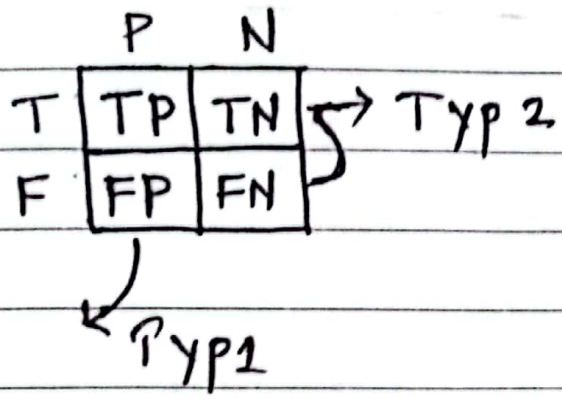
Output 3 : We accept the Null Hypothesis When in reality it is false → Type 2 Error

Output 4 : We accept the Null Hypothesis When in reality it is true → Good

Mo Tu We Th Fr Sa Su

Date :

Confusion matrix



Mo Tu We Th Fr Sa Su Date :

Confidence Interval

Point estimate: The value of any statistic that estimates the value of a parameter

In Inferential Stats.

SAMPLE → POPULATION

Mean	Mean
$\bar{x}$ →	μ
$\bar{x} = 2.9$	$\mu = 3$

Note: Estimating the μ value → Population.

↓

Point estimate.

Confidence Intervals

Point Estimate ± margin of Error

eg: In the quant test of CMAT Exam, the standard deviation is known to be 100. A sample of 25 test takers has a mean of 520 score. Construct a 95% CI about the mean?

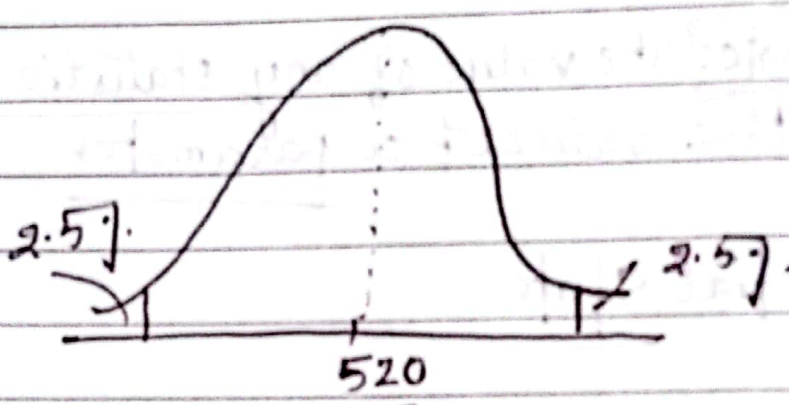
Ans> $\sigma = 100$, $n = 25$, $\alpha = 0.05$, $\bar{x} = 520$

→

Sol?

here,

$\alpha = 1 - 0.95 = 0.05$



Condition

① Population std is given } z-Test
② $n \geq 30$

- Point Estimate ± margin of error

$$\bar{x} \pm z_{\alpha/2}\left[\frac{\sigma}{\sqrt{n}}\right] \rightarrow \text{standard error}$$

Upper bound $= \bar{x} + z_{\frac{0.05}{2}} \frac{100}{\sqrt{25}}$

$= 559.2$ //

lower bound $= \bar{x} - z_{\frac{0.05}{2}} \frac{100}{\sqrt{25}}$

$= 480.8$

IN SANTOSH LOHAR

Mo Tu We Th Fr Sa Su

Date

eg(2) ⇒ On the quant test of CAT exam, a sample of 25 test takers has a mean of 520 with a standard deviations of 80. Construct 95% CI about the mean?

⇒

Condition : | $n = 25$, $\bar{x} = 520$, $s/d = 80$, $\alpha = 0.05$

Here, Population sd is not given. → t-test

Point Estimate ± margin of Error

$\bar{x} \pm t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}}\right)$ → standard error

upper bond = $\bar{x} + t_{\frac{\alpha}{2}} (s/\sqrt{n})$

Degree of freedom = $n - 1 = 25 - 1 = 24$

$= 520 + 2.064 (80/5)$

$= 553.024$ //

lower bond = $\bar{x} - t_{0.05/2} (s/\sqrt{n})$

$= 520 - 2.064 (80/5)$

$= 486.97$

[486.97 → 553.024]

Mo Tu We Th Fr Sa Su

Date :

(1) One Sample Z-Test

↳ Population sd is given
↳ Sample size $n \geq 30$

*

In the Population, the average IQ is 100 with a sd of 15. · wants to test a new medication to see if there is position or negative effort on intelligence, or no effect at all. A Sample of 30 Participants who have taken the medication has a mean of 140. Did the medication affect the intelligence?

$\alpha = 0.05$, (CI = 95%)

Ans//

1) Define Null Hypothesis

$H_0 = \mu = 100$

2) Alternative Hypothesis H_1 $\mu \neq 100$

3) State Alpha

$\alpha = 0.05$

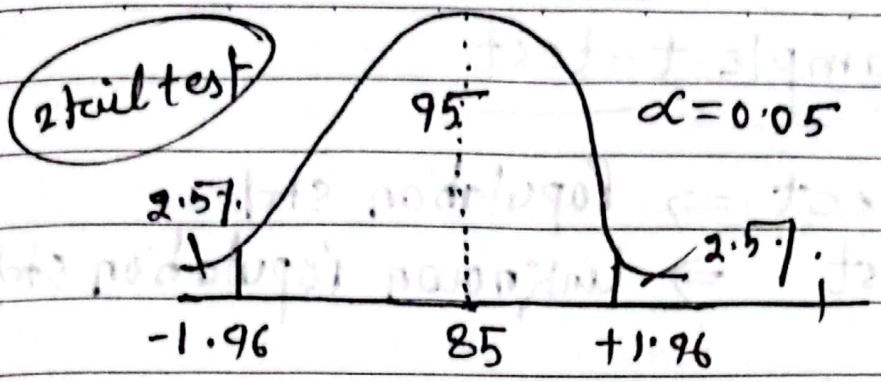
4) State Decision Rule

DATA SCIENCE

SANTOSH LOHAR

Mo Tu We Th Fr Sa Su

Date:



$P \leq 0.05$

5) Calculate Test statistics

$$Z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$ → standard error

$$= \frac{140 - 100}{\frac{15}{\sqrt{30}}} = \frac{40}{15} \times \sqrt{30} = 14.60 //$$

State Our Decision:

$14.60 > 1.96$

{ z = -0.2
{ z = -14.60 }

If z is less than -1.96 or grether than 1.96, reject the Null Hypothesis //

Meditation improve the intelligence or, decrease?

⇒ Improve the intelligence //

IN SANTOSH-LOHAR

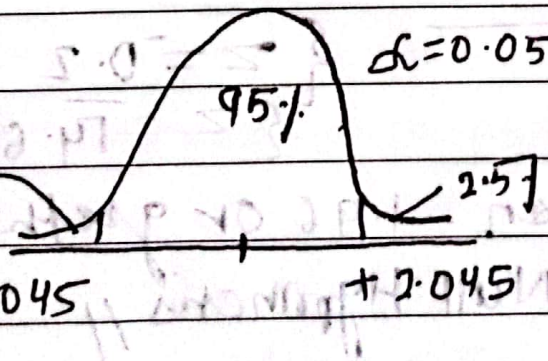
Mo Tu We Th Fr Sa Su Date :

One Sample t-test

Z-Test ⇒ Population std
t-test ⇒ Unknown Population std.

Q) Population the average IQ = 100
$n = 30$, $\bar{x} = 140$, $s = 20$
Did the meditation affect intelligence?
$\alpha = 0.05$

Ans) 1) $H_0 = \mu = 100$
2) $H_1 : \mu \neq 100$
3) Calculate the degree of freedom
$n - 1 = 30 - 1 = 29$
4) State Decision Rule:



$\alpha = 0.05$, $t = 10.96 > 2.045$
Reject Null Hypothesis
$P \leq$ significance value
⇓

5) T-Test

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

$\bar{x} = 140$
$\mu = 100$
$s = 20$
$n = 30$

$= 10.96$

{ Increase the Intelligence }

└ Reject the Null Hypothesis //

IN ⇒ SANTOSH LOHAR

Mo Tu We Th Fr Sa Su Date :

CHI SQUARE TEST

1) Chi-square test ~~clears~~ claims about population ~~proportional~~ proportions

It is a non parametric test that is performed on categorical (normal or ordinal) data.

Q.) In the 2000 Indian census, the age of the individu in a small town were found to be the following

Less than 18	18-35	>35
20%	30%	50%

In 2010, age of $n = 500$ individuals were sampled Below are the results.

~~<18~~	<18	18-35	>35
	121	288	91

Now, Using $\alpha = 0.05$, would you conclude the population distribution of ages has changed in the last 10 yrs.

Sol)

<18	18-35	>35	
20%	30%	50%	⇒ Expected.

<18	18-35	>35	$n = 500$
121	288	91	⇒ Observed
100 = 500×0.20	500×30% = 150	500×50% = 250	⇒ Expected

iN SANTOSH-LOHAR

distribution = 3

Mo Tu We Th Fr Sa Su Date:

< 18	18-35	> 35	
121	288	91	Observation
100	150	250	- Expected

Follow: Chi square Table.

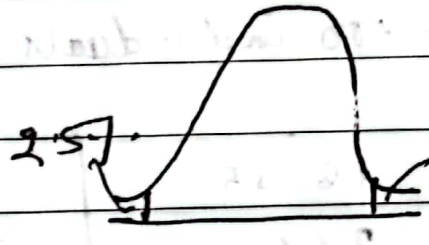
Now ① H_0 = The data meets the distribution 2011 Census

H_1 = The data does not meet " " "

② $\alpha = 0.05$ (95% CI)

③ Degree of freedom = $(n-1) = (3-1) = 2$

④ Decision boundry:



If χ^2 is greater than 5.991 reject H_0

See Chi square table.

(Join the help of α & df

⑤ Calculate test statistics.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150} + \frac{(91-250)^2}{250}$$

$$= 232.494$$

IN : SANTOSH - LOHAR

Mo Tu We Th Fr Sa Su Date :

Here,

$x^2 = 232.494 > 5.99$

$\Rightarrow$ Reject the Null Hypothesis

$0.11 > 0.05$

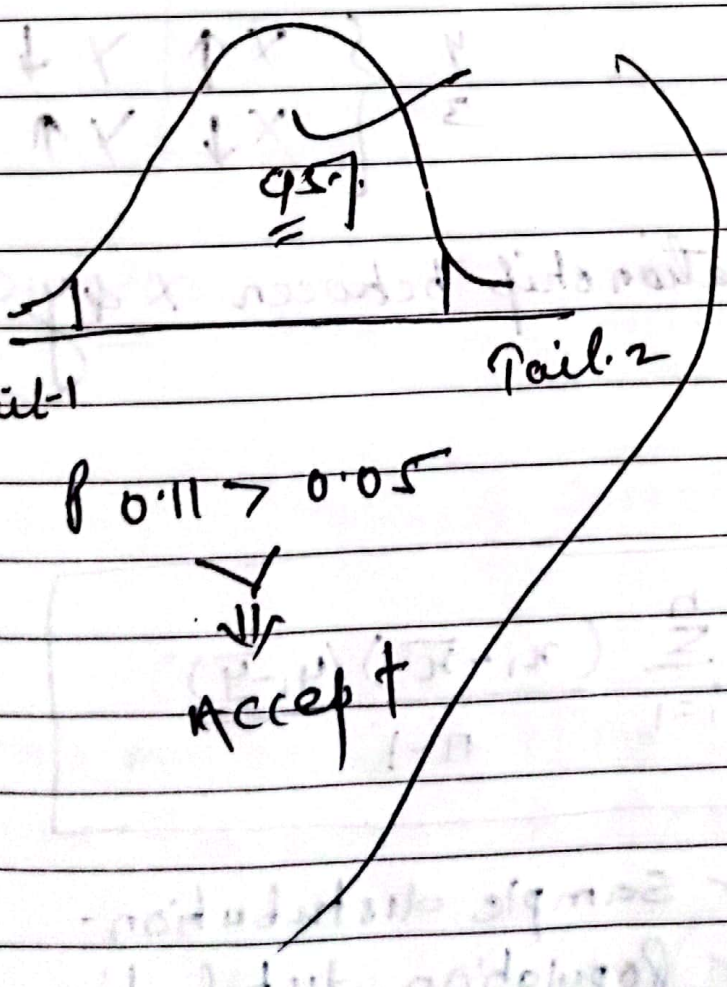
↳ Accept the Null.

P-value < significance

↳ Reject the Null hypothesis

OR

Accept the Null Hypothesis



$P\ 0.11 > 0.05$

⇓

Accept

$P = 0.002 < 0.05$

⇓

Reject the Null Hypothesis

linked in : Santosh-lohar

Mo Tu We Th Fr Sa Su Date :

Covariance

X	Y
weight	Height
50	160
60	170
70	180
75	181

We can say that
$X \uparrow \rightarrow Y \uparrow$
$X \downarrow \rightarrow Y \uparrow$

No. of hour study	play.
2	6
3	4
4	3

$X \uparrow \; Y \downarrow$
$X \downarrow \; Y \uparrow$

Now, quantify relationship between X & Y.

Covariance

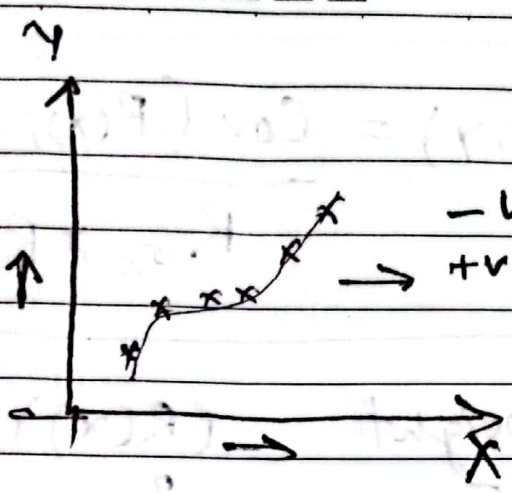
$$Cov(X,Y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$n-1$ ⇒ For Sample distribution.

N ⇒ For Population distribution.

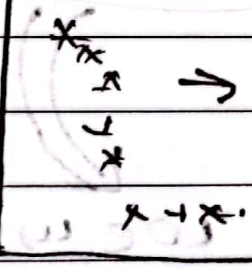
linked in : Santosh-lohar

Mo Tu We Th Fr Sa Su Date :

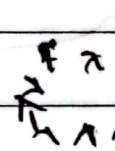


-ve - covariance.
→ +ve correlation

X ↓ → Y ↑
X ↑ → Y ↓
⟹ Negative -ve Correlation,
-ve ↓ Covariance



→ -ve correlation
↓
-ve covariance



Covariance = 0

Disadvantage of covariance

Pearson Correlation Coefficient

(-1 to +1)

⇒ The more towards +1 move positively Correlation
The more towards -1 move negative Correlated.

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \Rightarrow (-1, \text{to } 1)$$

Mo Tu We Th Fr Sa Su Date:

$$Spear(X,Y) = \frac{Cov(R(x), R(y))}{R_{\sigma_x} \times R_{\sigma_y}}$$

Height	weight	(R(x))	(R(Y))
170	75	2	2
160	62	3	3
150	60	4	4
145	55	5	5
180	85	1	1

↳ this will be cancelled

We use distribution for Computing.

Non-linear Proportions.

$p \leq 0.05$ → Reject the Null Hypothesis

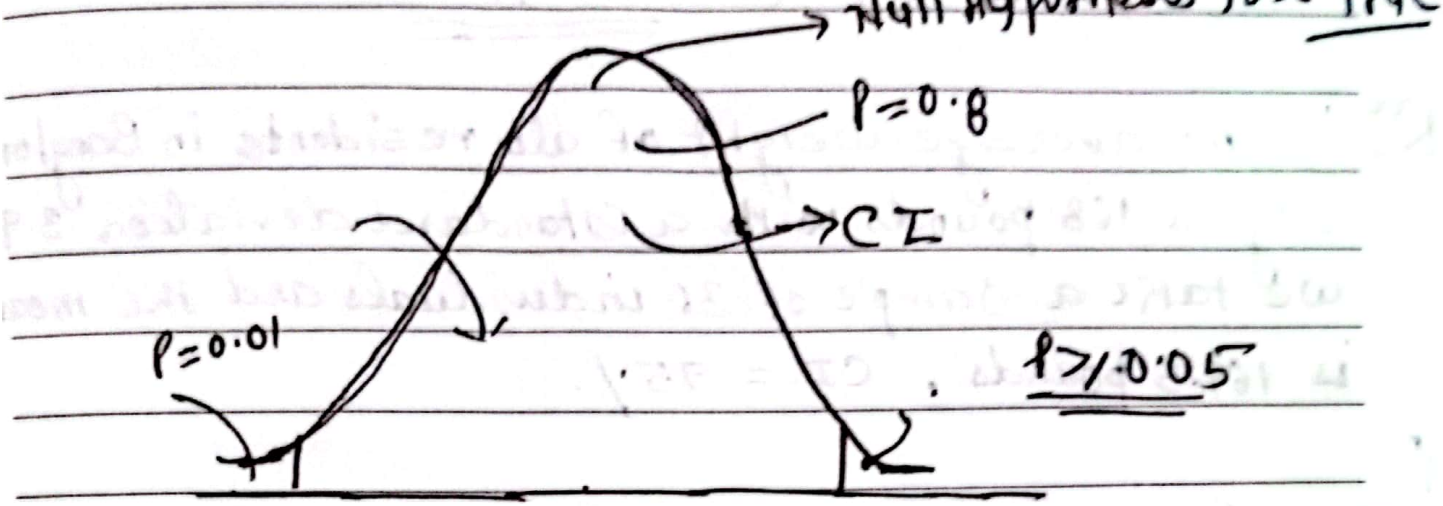
↳ $\alpha = 0.05$

⇒ 5% Probability the Null Hypothesis is Correct }

$p > 0.05$ → Accept the Null Hypothesis

IN ⇒ SANTOSH-LOHAR

Mo Tu We Th Fr Sa Su Date



Here, Null Hypothesis should be Accepted because of

$P >/ 0.05$

P-value <= C.I.

↳ Reject the Null Hypothesis

P-value >/ C.I

↳ Accepted Null H.

Mo Tu We Th Fr Sa Su Date

Additional P-value

Q> ∴ The average weight of all residents in Banglore city is 168 pounds with a Standard deviation 3.9 we take a Sample of 36 individuals and the mean is 169.5 pounds, C.I. = 95%.

Soln:

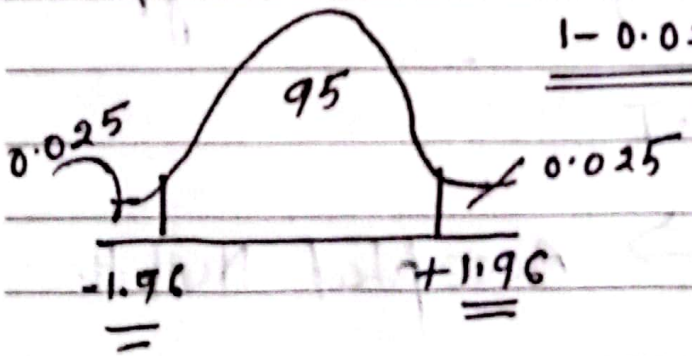
⇒ $\mu = 168$, $\sigma = 3.9$, $\bar{x} = 169.5$, $n = 36$, $\alpha = 0.05$

① $H_0 = \mu = 168$

$H_1 = \mu \neq 168$

② $\alpha = 0.05$

③ Decision Boundry:



$1 - 0.025 = 0.9750$

④ Z-test

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{169.5 - 168}{3.9/\sqrt{36}} = \frac{1.5}{3.9} \times 6$$

$$= 2.307$$

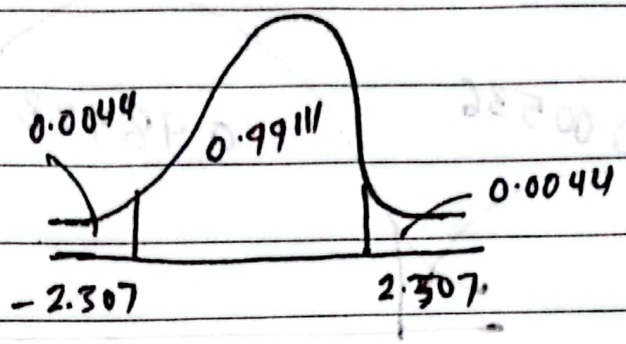
∴ $Z = 2.307 > 1.96$

⇒ Reject Null Hypothesis.

IN Santosh-Lohar

Mo Tu We Th Fr Sa Su Date :

P-value



1-0.9911

Pvalue = 0.0044+0.0044

= 0.0088

Here, Pvalue < 0.05

0.0088 < 0.05 ⇒ Reject the Null Hypothesis

* Pvalue ≤ significance value

⇓

Reject the Null Hypothesis

P-value > significance value

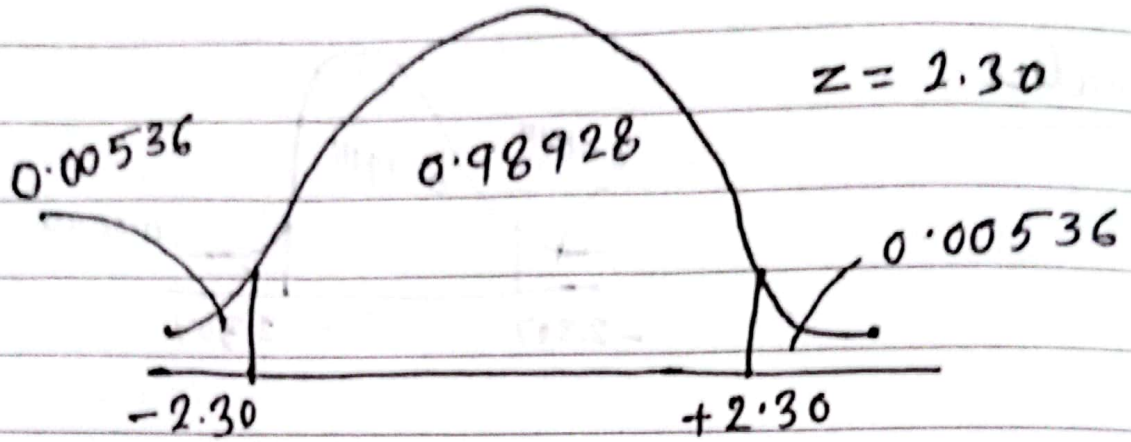
⇓

fail to Reject the Null Hypothesis

linked In : Santosh Sahoo

Mo Tu We Th Fr Sa Su

Date :



1 - 0.98928

Here, $2.30 > 1.96$ {Reject the Null Hypothesis}

Pvalue = 0.00536 + 0.00536

If less than α ⇒ Reject the hypothesis

eg:

Average age of a college is 24 years with a standard deviation 1.5. Sample of 36 students, students mean is 25 years with $\alpha = 0.05$, CI. = 95%. do the age vary?

⇒ ①

H_0 $\mu = 24$ $\sigma = 1.5$ $n = 36$ $\bar{x} = 25$ $\alpha = 0.05$

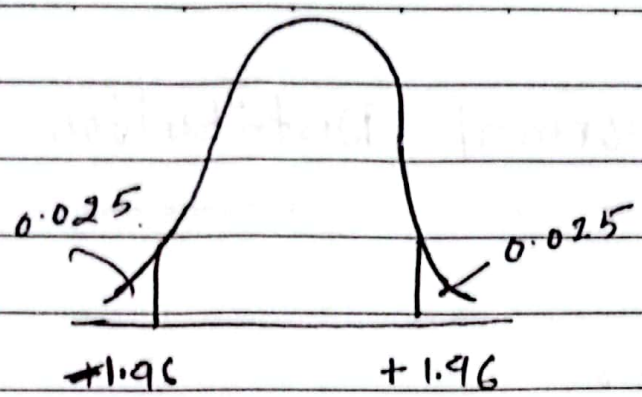
$H_1 = H_1, \mu \neq 24$

②

$\alpha = 0.05$

Mo Tu We Th Fr Sa Su Date :

(3)

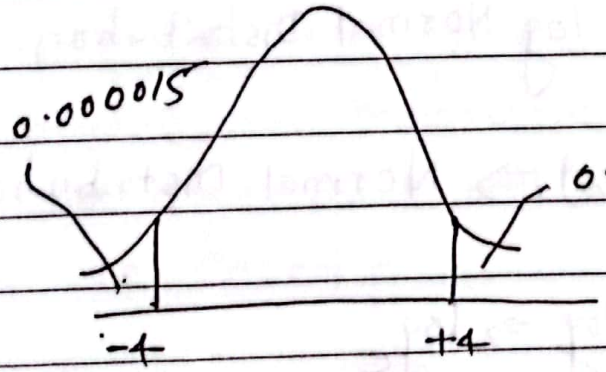


(4) $Z\text{-score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{25-24}{1.5} \times 6$

$= \frac{1 \times 6}{1.5}$

$= 4$

Now, $4 > 1.96$ → Reject Null Hypothesis.



$Pvalue = 0.000015 + 0.000015$

$= 0.00003$

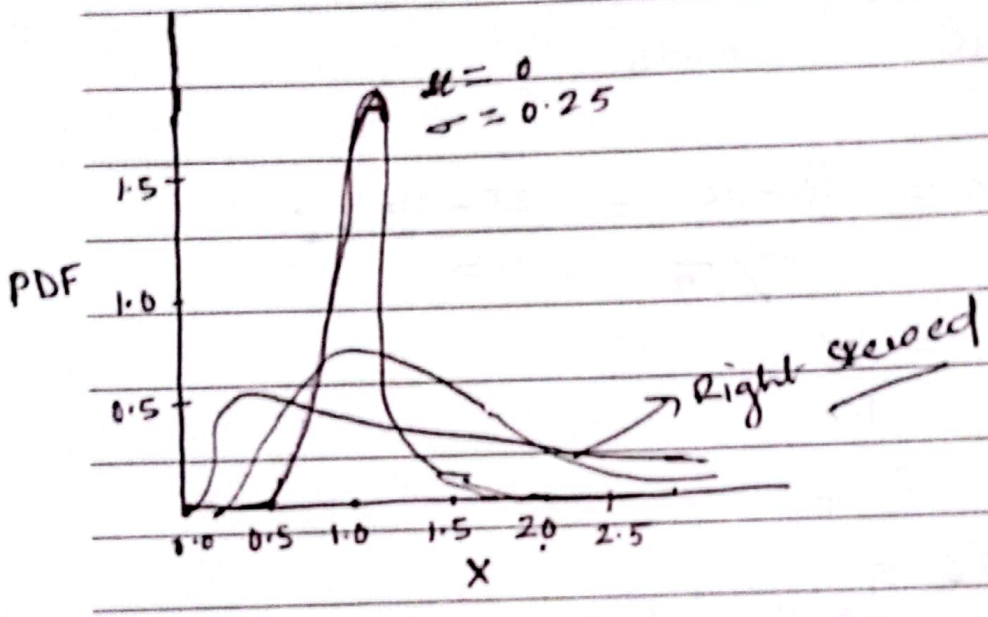
Here, $Pvalue \leq Significance$ → Reject the Null Hypothesis

linked In : Santosh Lohar

Mo Tu We Th Fr Sa Su

Date

Log Normal Distribution



if $X \approx$ log Normal Distribution.

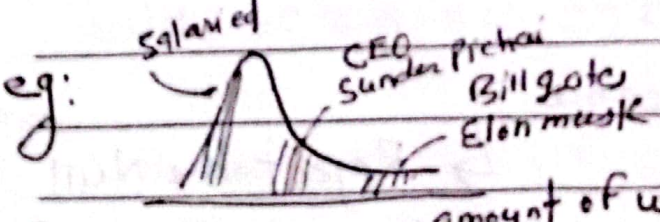
then,

$y \approx \ln(x) \Rightarrow$ Normal Distribution.

Natural log $\Rightarrow \log_e$

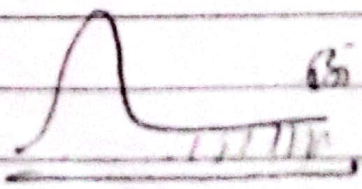
$X \approx \exp(y)$ (we can also write like this)

eg:



① Wealth distribution

② Comment in YT/FB



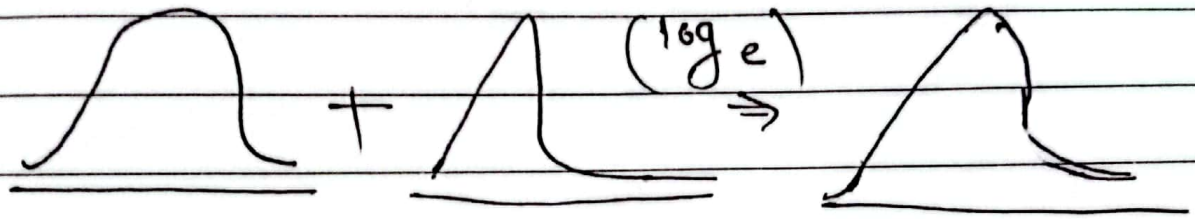
Mo Tu We Th Fr Sa Su Date :

Where we used (log Normal Distribution)

↳ Machine learning

⇓

Simple Linear Regression



⇓ Normal Distribution

⇓ log-Normal Distribution.

⇓ Normal Distribution

Now, We can use for the purpose of Training data model in machine learning and from this our algorithms will be efficiently Trained which we called

⇓

Data Transformation Technique