Date: / /

Introduction to Statistics

Descriptive Stats

↓

1. Measure of central tendancy
2. Measure of dispersion

Anything related to summarizing the data.

Mean, Median, mode, SD, Variance

3. Gaussian Distribution
4. Normal "
5. Binomial "

Inferential Stats

↓

1. Z test → Python
   t test → Python
   ANOVA
   CHISQUARE
   Hypothesis testing
   (P Value)
   Z table, t table

Statistics

↳ collecting, organising & analyzing data

↳ {Better Decesion Making}

Data → Facts or pieces of information that can be measured.

Types of Statistics

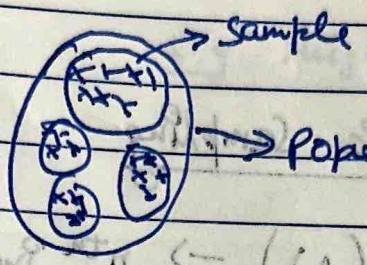
1. Discriptive → consists of organizing & summarizing data
2. Inferential → technique where we used the data that we have measured to form conclusion

Date: / /

Population & Sample

Elections → Goa, UP

Exit Pole

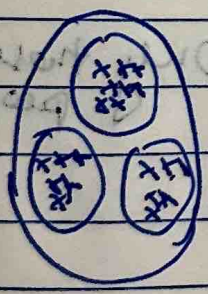


Population → (N)
Sample → (n)

Sampling Techniques

① Simple Random Sampling

↳ Every member of the population has an equal chance of being selected for your sample (n).



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

E.g. Gender → Male
→ Female

Date: / /

Age

(0-10) (10-20) (20-40) (40-100)

Profession

③ Systematic Sampling

(N) → nth individual

Eg Mall → Survey (covid)

↳ 8th person → survey

↓

It means every 8th person

④ Convenience Sampling

↳ Survey

↳ Data science

Only those people

Variables

A variable is a property that can take on any values.

a) Quantitative Variables → Measured Numerically

b) Qualitative / Categorical Variables.

Date: / /

Quantitative

↓ Discrete Variables

E.g: Whole Number

Total Child in family

2, 3, 4, 5

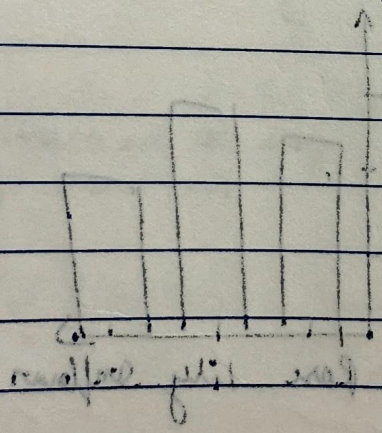
↓ Continous Variables

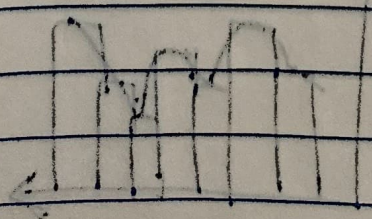
E.g Height: 172.5, 162.2cm

Variable: Measurment ~~of~~ Scales →

4 types of Measured Variable

1. Nominal → Class → Gender, Colours, type of flowers
2. Ordinal → Order of the ~~data matters~~, not values.
3. Interval → ~~Order matters~~, values also matters, natural zero is not present
4. Ratio





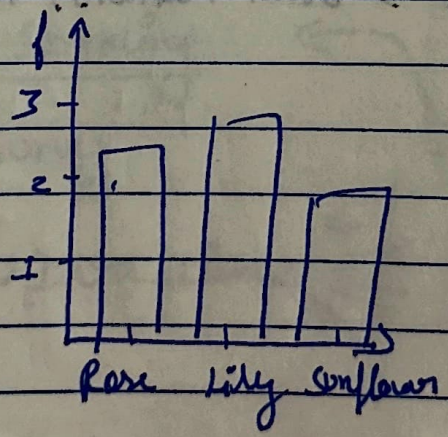
Date: / /

Frequency Distribution

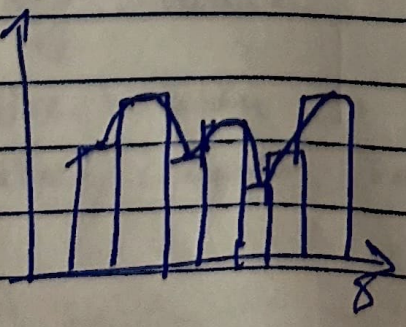
Sample Dataset :- Rose, Lilly, Sunflower, Rose, Lilly

flower	F	CF
Rose	3	3
Lilly	5	7
Sunflower	2	9

Bar Graph



Histogram :- Continous



Arithmetic Mean for population & Sample

Mean (Average)

Population (N)

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

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$$

$$= \frac{1+1+2+2+3+3+4+5+5+6}{10}$$

$$= \frac{28}{10} = 2.8 = \frac{32}{10} = 3.2$$

Sample (n)

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

$$= 3.2$$

Central Tendency → refers to the measure used to determine the centre of the distribution of data.

- Mean
- Median
- Mode

$\{1, 1, 2, 2, 3, 3, 4, 5, 5, 6, \boxed{100}\}$ → outliers

$\frac{12}{2} = \frac{6+7}{2}$ $\frac{12}{26}$ $\frac{6+7}{2} =$ $\frac{11}{2} =$ 5th.

11 Date: / 15+1

Median { 1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100 } $\frac{11}{2} =$ 5-5

① Sort the Numbers

In case Odd Number = $\frac{N}{2} = \frac{11}{2} =$ 5th element

$= 3$

In case Even Number = $\left(\frac{N}{2}\right)$th element + $\left(\frac{N}{2}+1\right)$th element

$\frac{3+4}{2} = 3.5$

{Median works well with Outliers}

Mode = { 1, 2, 2, 3, 4, 5, 6, 6, 6, 7, 8, 100, 100, 100, 100 }

= { Most Frequent Element }

[Mode works in categorial data]

Type of flowers : Petal length, petal width

Rose
Lily
Sunflower

{ Missing Value → Most frequent occuring element }

80% → 1, 2, 3, 5 → 80%

Date: / /

Measure of Dispersion → spread

a) Variance

b) Standard Deviation

Population Variance

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

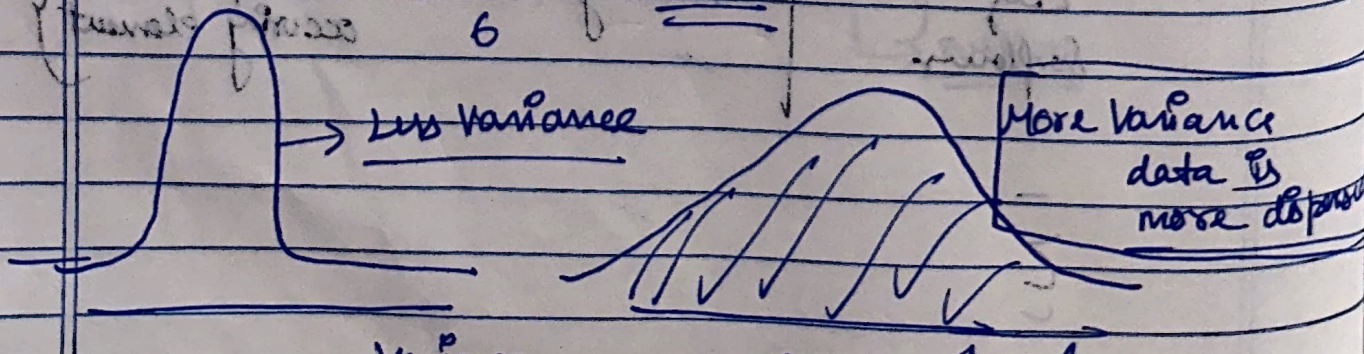
Sample Variance

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

x	μ	x − μ	(x − μ)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	4.711
			10.84

μ = 2.83

$$\frac{10.84}{6} = 1.81$$



Variance means more spread

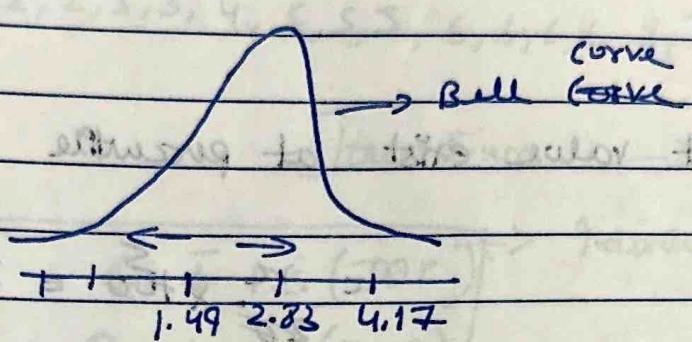
1+2+3+4 = 10/4 = 2.5

Date:

SD

$\sigma = \sqrt{\text{variance}} = \sqrt{1.81} = 1.345$

Mean = 2.83



2.83
+1.34
4.17

→ Variance tells how the data spread

→ SD tells the range of the data

Percentils & Quartiles [Find Outliers]

A percentile is a value below which a certain percentage of observation lies.

Eg

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

Q What is the percentile ranking of 10?

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

n = sample size

$= \frac{16}{20} \times 100 = 80\%$

80% of the dentists distribution is less than 10

$\frac{5+6}{2} = \frac{11}{2} = 5.5$

Date: / /

$= \frac{17}{20} \times \frac{5}{100} = 85\%$

Q What value exist at percentile ranking of 25%?

$= \frac{x}{20} \times \frac{5}{100} = 25$

$5x = 25$

$x = 5$

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

$= \frac{25}{100} \times 21 = 5.25$ → index position

5^{th} index + 6^{th} index / 2

$\frac{75}{100} \times 21 = 15.75 \rightarrow 9$

Five Number Summary

1. Minimum
2. 25% Q_1 (quantile 1)
3. Median
4. Q_3 (quantile 3)
5. Max

Date: / /

Removing the Outliers

{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 9, 27} (27 → outlier)

[Lower Fence ⟷ Higher Fence]

Lower Fence = $Q_1 - 1.5(IQR)$
Upper fence = $Q_3 + 1.5(IQR)$ → Interquartile range

$IQR = Q_3 - Q_1$

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

$IQR = 7 - 3 = 4$

$\frac{25}{100} \times 20 = 5 \rightarrow$ Index $= 3$

$\frac{75}{100} \times 20 = 15 \rightarrow$ Index $= 7$

Lower fence $= 3 - 1.5(4)$
$= 3 - 6 = \boxed{-3}$

Higher Fence $= 7 + 1.5(4)$
$= \boxed{13}$

[−3 ⟷ 13]

anything lesser than 3

anything greater than 13.

$\frac{18}{2} = 9^{th}$ Date: / /

5

Remaining Data

1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, ~~27~~

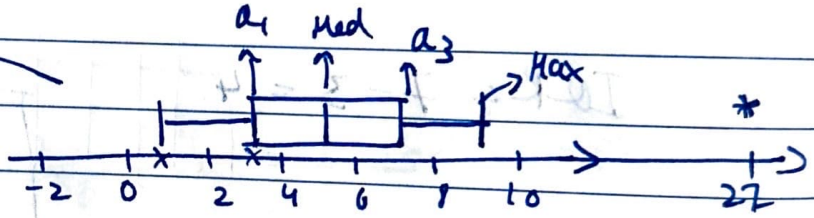
Min = 1
Q_1 = 3
Median = 5
Q_3 = 7
Max = 9

→ 5 Number Series

⇓

Box Plot

Data Visualization ←



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

→ Basal correction

↳ Degree of freedom

Why Sample variance is n-1?

Population (N)

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$$

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

Sample (n)

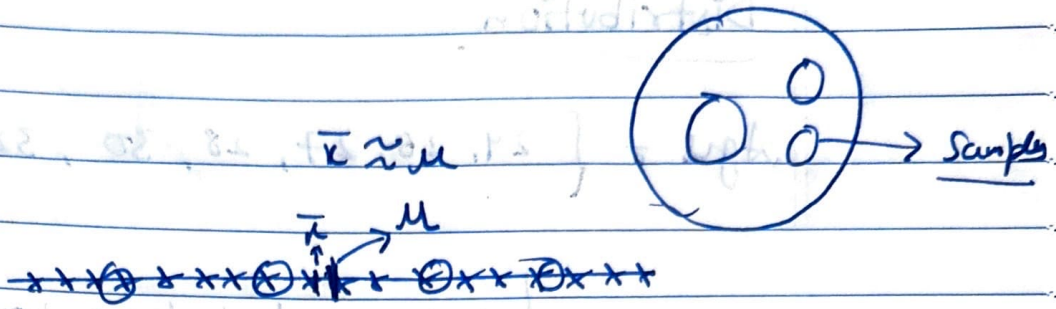
$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

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

Mean - Mean

Date: / /

Ages



① Distribution

- Normal (Gaussian)
- Standard Normal
- Z score
- Log Normal Distribution
- Bernoulli Distribution
- Binomial Distribution

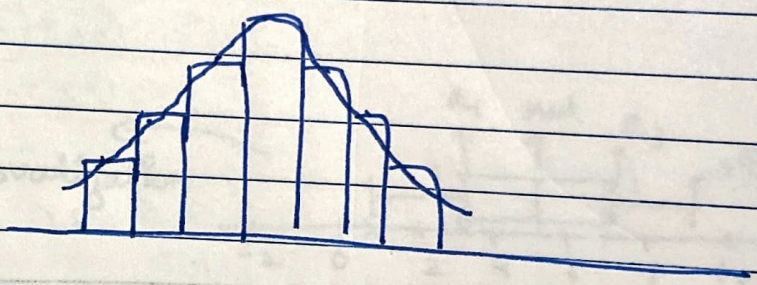
Date: / /

Distribution

Ages = { 24, 26, 27, 28, 30, 32 ------ }

↳ we need to focus how we see the data in visualize way

↳ Our objective is to visualize the data.

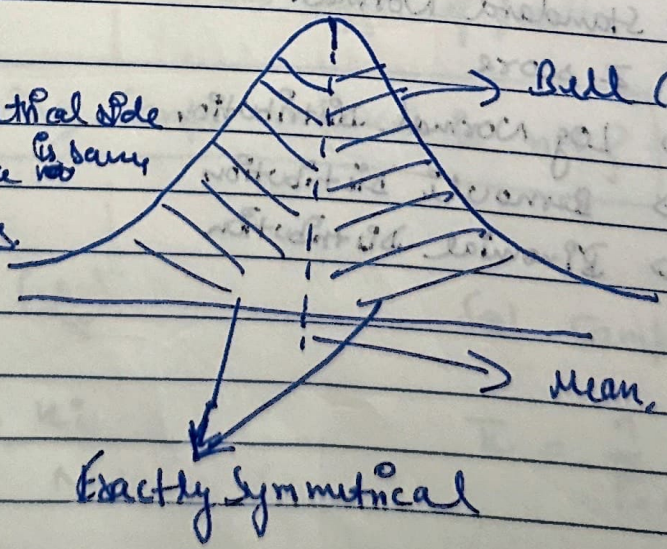


Gaussian / Normal Distribution

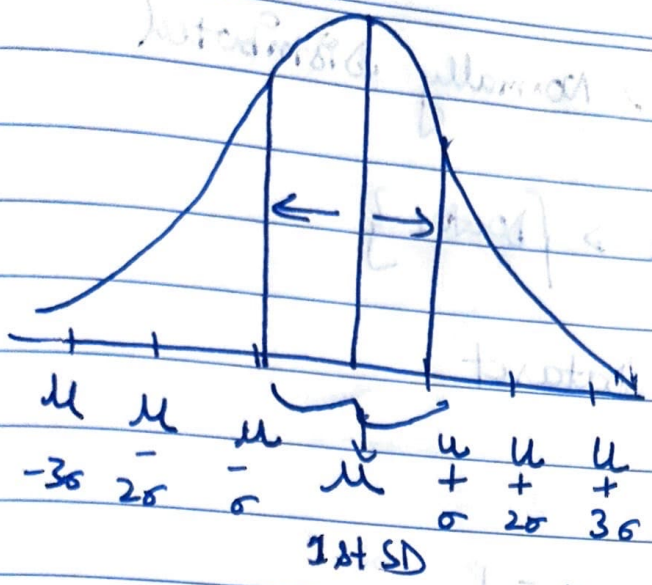
Property

-> same symetrical side.

-> Amount of data is same in both parts.



Date: / /



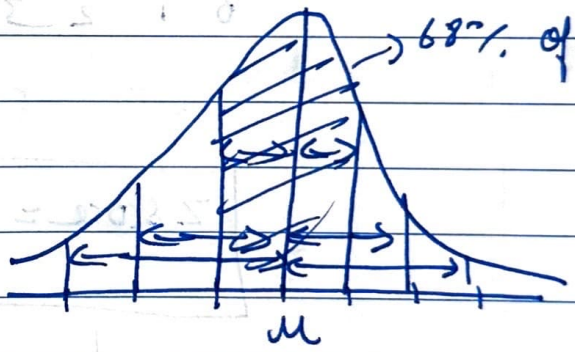
Empirical Formula

~~68 - 95 - 95~~

68

95 - 99.7% Rule

Dataset [100 Data points]



95% of distribution

99.7% of distribution.

Date: / /

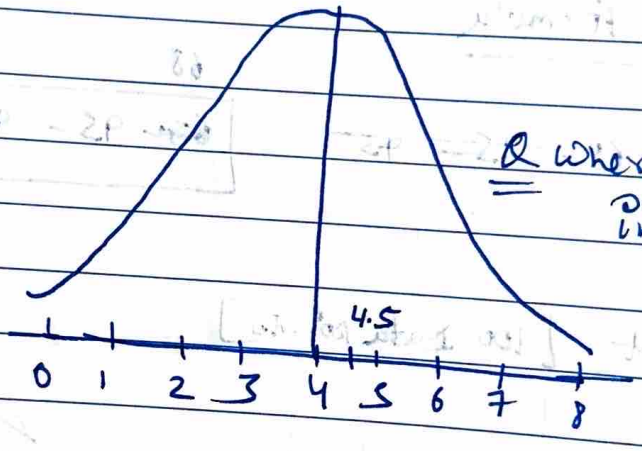
Eg. Height → Normally Distributed

↓

Domain Expert → {Doctor}

Weight, Iris Dataset

Eg $\mu = 4$, $\sigma = 1$



4.5 → SD

Q Where does 4.5 fall in terms of SD?

~~0.5~~

+0.5 sd

~~+0.7~~

$$\text{ZScore} = \frac{x_i - \mu}{\sigma}$$

↳ with basically helps to find out how much standard deviation a value is away from its mean.

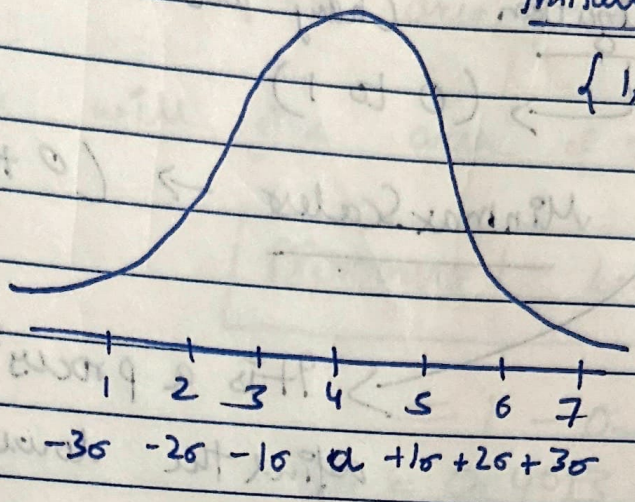
$\frac{1-4}{1} = -3$ $\frac{-3-0}{1} = -3$

Date: / /

Initially distribution

$\mu = 4$

$\sigma = 1$



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

↑ Apply

$Z\ score = \frac{x_i - \mu}{\sigma}$

{-3, -2, -1, 0, 1, 2, 3}

{1, 2, 3, 4, 5, 6, 7} → Normal distribution

⇓

Z score

⇓

{-3, -2, -1, 0, 1, 2, 3} → Standard Normal distribution ($\mu = 0, \sigma = 1$)

satisfying

Practical Application

To convert the dataset into this → $\mu = 0, \sigma = 1$

Dataset

Age (years)	Salary (₹)	Weight (kg)
24	40k	70
25	80k	80
26	60k	55
27	70k	45

Apply Z score & convert into SND

Standardization → Z score

-1.02

0.25 S.

0.26

0.2 Date: 0.06

Normalization Chang all the values in 0 to 1

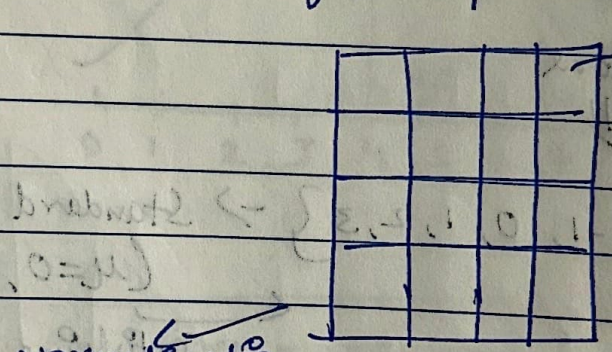
(-1 to 1) → (0 to 1)

MinMax Scaler → (0 to 1)

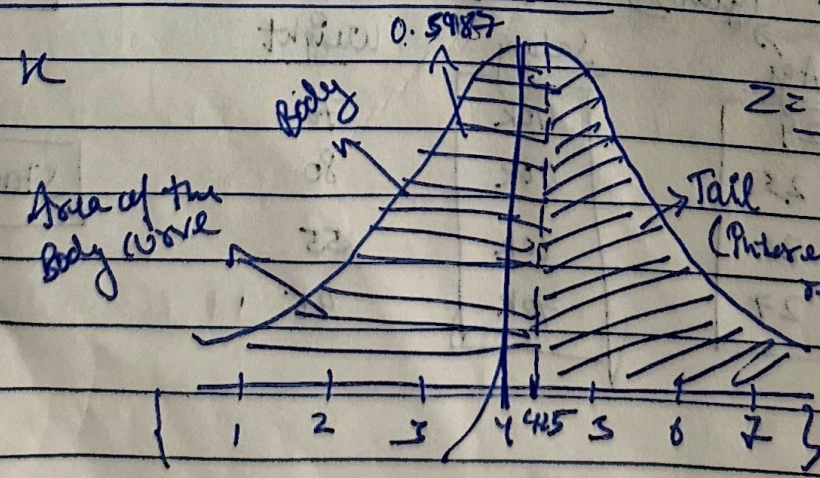
→ It is a process where you can define the lower bound & upper bound & you can convert the data between them

E.g

CNN → Image classification



Q. Stats Interview Question



$Z = \frac{x - \mu}{\sigma} = \frac{4.25 - 4}{1}$

$= 0.25$ std

Q What percentage of scores fall above 4.25?

$\frac{200}{225} = \frac{8}{9}$ $\quad \frac{3}{7} \times 100 = \frac{300}{7} = 42.8$

$-1.02 \quad -103$ $\quad \frac{4}{7} \times 100 = \frac{400}{7} = 57$

Date: / /

→ Z scores helps to find area of the body curve

→ Z tables will give area of the body curve

Right area = 1 − Left Area

$= 1 - 0.5987$

$= 0.4013$

$\downarrow$

40%

Q In India,

$\mu = 100$

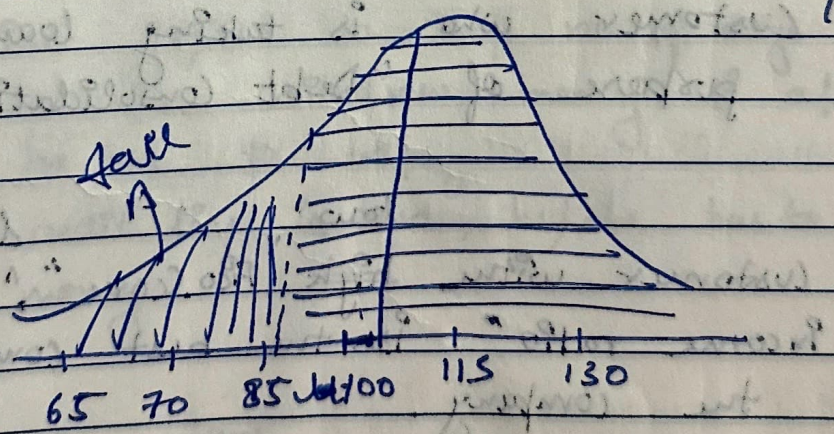
$\sigma = 15$

$x_i = 85$

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

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

$= \frac{85 - 100}{15} = -1$



Left = 1 − Right Area

Date: / /

Practical

Google Colab

```
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
import statistics

# mean, median, mode

df = sns.load_dataset('tips')

df.head()

np.mean(df['total_bill'])
np.median(df['total_bill'])
statistics.mode(df['total_bill'])

sns.boxplot(df['total_bill'])

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

histplot → histogram

kde = True → probability density function.

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

countplot → count the data grouping by category

Date: / /

np.percentile(df['petal_length'], [25, 75])

#outliers

dataset = [your own dataset]

outliers = []

def detect_outliers(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

IQR = $Q_3 - Q_1$

Date: / /

iqr

↳ sort the data

↓

Calculate Q_1 & Q_3

↓

IQR ($Q_3 - Q_1$)

↓

Lower fence = $Q_1 - 1.5(IQR)$

Upper fence = $Q_3 + 1.5(IQR)$

dataset = sorted(dataset)

q1, q3 = np.percentile(dataset, [25, 75])

Find the lower fence & higher fence

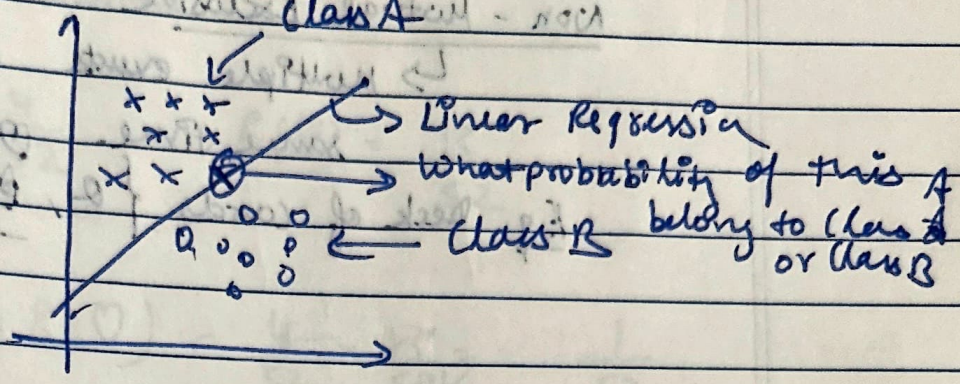
iqr = q3 - q1

lower_fence = q1 - (1.5 * iqr) # 7.5

higher_fence = q3 + (1.5 * iqr) # 19.5

Date: / /

Probability



It is a measure of the likelyhood of an Event

E.g. Flipping a dice $\{1, 2, 3, 4, 5, 6\}$

$$Pr(6) = \frac{1}{6} = \frac{\text{\# of way an event can occur}}{\text{\# of possible outcomes}}$$

Toss a coin $\{H, T\}$

$$Pr[H] = \frac{1}{2}$$

Addition Rule ('or')

Mutual Exclusive Events

Two events are mutual exclusive if they ~~are not~~ cannot occure at the same time

E.g. Rolling a dice $\{1, 2, 3, 4, 5, 6\}$

Date: / /

Non - Mutual Exclusive

↳ Multiple events can occure at the same time.

Eg Deck of cards {♠, ♥}

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

Ans: Mutual Exclusive

Pr (H or Tails)

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

$$= \frac{1}{2} + \frac{1}{2} = 1$$

Q Roll a dice

$$Pr(1 \text{ or } 3 \text{ or } 6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$$

Q Picking a card randomly from the deck. What is the probability of choosing a card that is queen or a heart?

Date: / /

Ans: Non-Mutual Exclusive

$$Pr(Q) = \frac{4}{52} \qquad Pr(♡) = \frac{13}{52}$$

$$Pr(Q \& ♡) = \frac{4^{1}}{52} \times \frac{13}{52_{4}} = \frac{1}{52}$$

Addition Rule for Non-Mutual Exclusive Events

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

$$P(Q \text{ or } ♡) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$$

$$= \frac{16}{52}$$

(3) Multiplication Rule

{Independent Events}

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

1, 1, 2, 2, 3, ... → one event is independent to another event.

{Dependent Events}