06.03.2023
Monday

Statistics → Descriptive → organize, summarize, simplify, Presentation (so that an ordinary or layman can understand)
Statistics → Influential → Population, Sample

Variable / Parameter / Features.

Catagorial variable (Location) → Nominal, Ordinal
Numerical variable (Salary)

we convert all catagorial variable into numerical variable when using in machine learning.

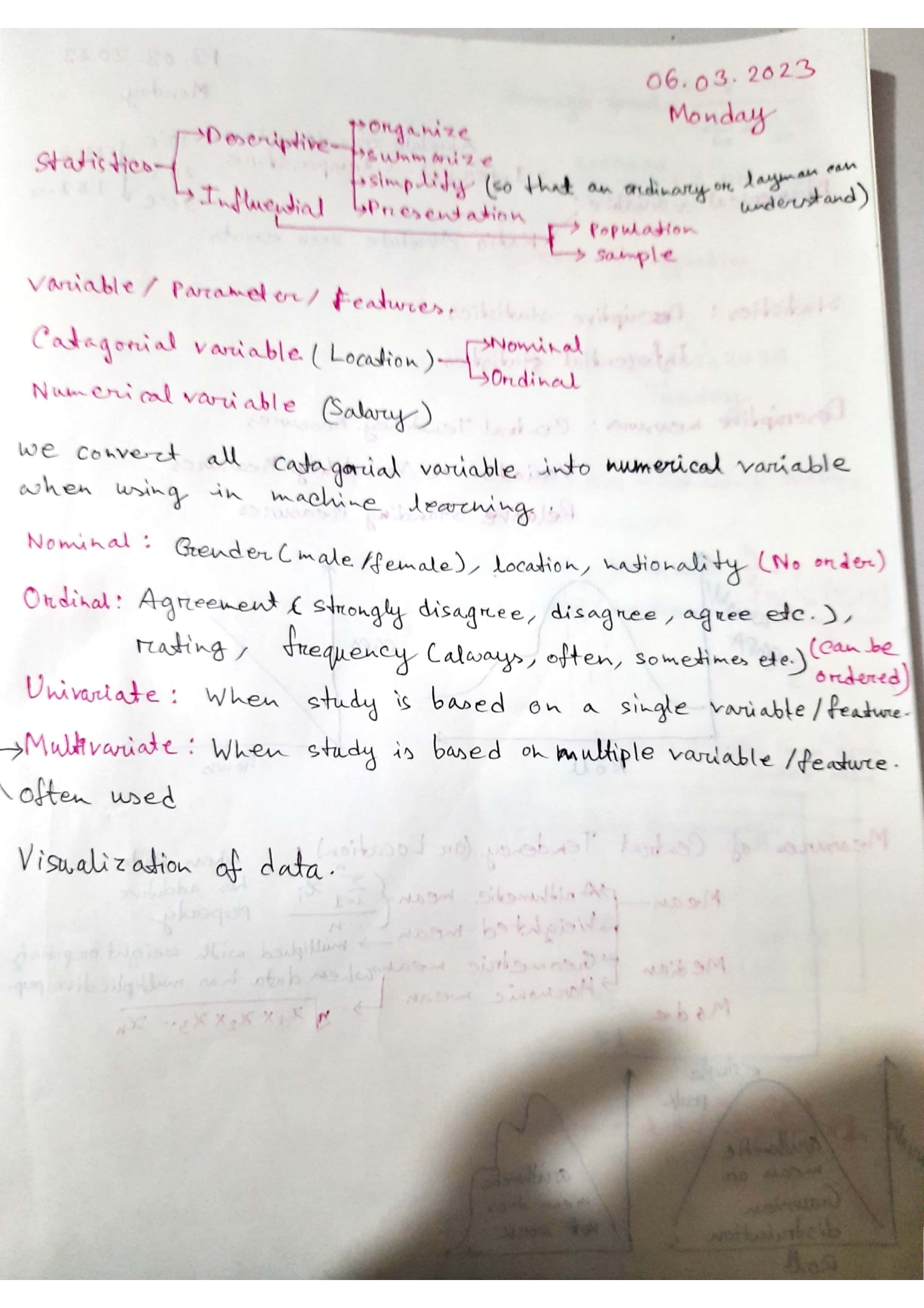
Nominal : Gender (male/female), location, nationality (No order)

Ordinal : Agreement (strongly disagree, disagree, agree etc.), rating, frequency (always, often, sometimes etc.) (can be ordered)

Univariate : When study is based on a single variable / feature.

→ Multivariate : When study is based on multiple variable / feature.
often used

Visualization of data.

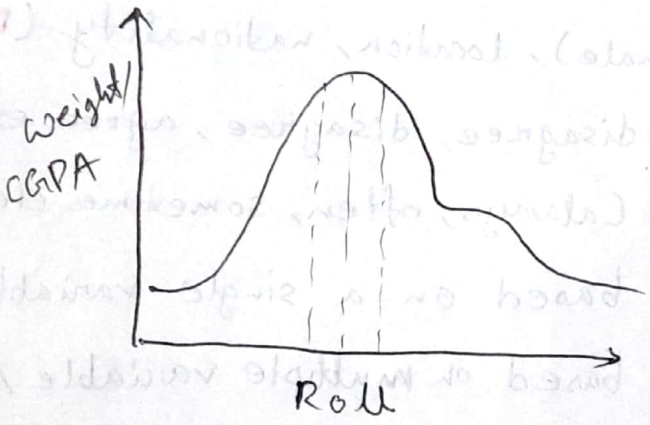


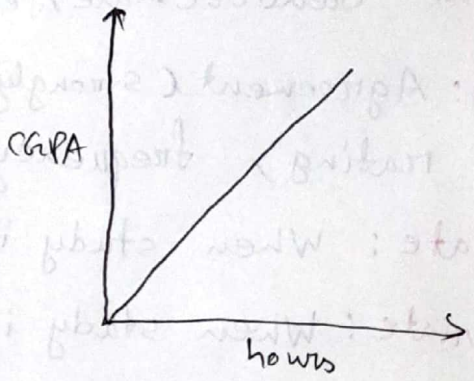
13.03.2023
Monday

Numerical variable → Interval → temparature (Absolute zero doesn't exist)
10°C ↓ 1.5 times
20°C ↓ 1.5 times
30°C
→ Ratio (Absolute zero exists)

Statistics: Descriptive statistics
Inferential statistics

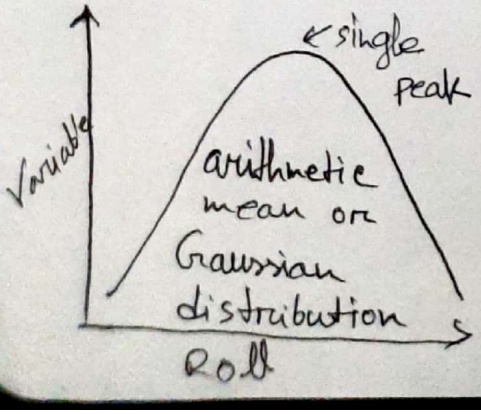
Descriptive measures: Central Tendency Measures
Variation or Variability Measures
Relative Standing Measures

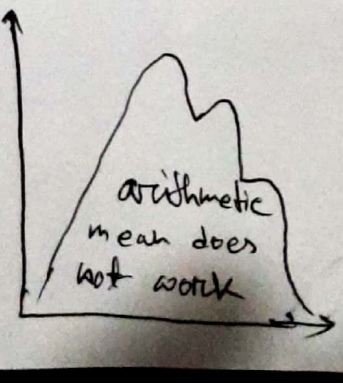


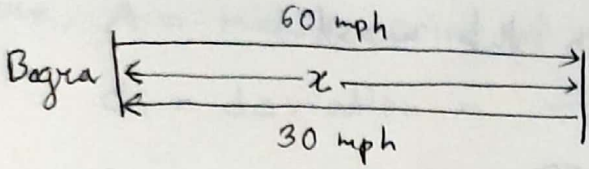




Measures of Central Tendency (or Location):

Mean → Arithmetic mean $\left\{\frac{\sum_{i=1}^{n} x_i}{n}\right.$ when data has additive property
→ Weighted mean → multiplied with weight or priority
Median → Geometric mean → when data has multiplicative property → $\sqrt[n]{x_1 \times x_2 \times x_3 \cdots x_n}$
→ Harmonic mean
Mode






Average speed $= \frac{2}{\frac{1}{60}+\frac{1}{30}}$

$= \frac{2\times60\times30}{60+30} = 40$ mph

$x = v_1 t_1 = 60 t_1 \; ; t_1 = \frac{x}{60}$ $\therefore 2x = v(t_1+t_2)$

$x = v_2 t_2 = 30 t_2 \; ; t_2 = \frac{x}{30}$ $v = \frac{2x}{\frac{x}{60}+\frac{x}{30}} = 40$ mph

Harmonic mean $= \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots\frac{1}{x_n}}$ ; $n$ = number of variables

14.03.2023
Tuesday

Book: Introduction to probability and statistics - Schaum's outline.

The median: number of element / odd, median $= \frac{n}{2}+1$

data needs to be sorted

number of element / even, median $= \frac{(n/2)+(n/2+1)}{2}$

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

It can be computed for
Ratio-level, Interval-level, Ordinal-level,
open ended frequency distribution if the median does not lie in an open-ended class.

| Height (cm) | Number of girls | |
|---|---|---|
| | frequency | cummulative frequency |
| <145 | 4 | 4 |
| 140 - 145 | 7 | 11 |
| 145 - 150 | 18 | 29 |
| 150 - 155 | 11 | 40 |
| 155 - 160 | 6 | 46 |
| 160 - 165 | 5 | 51 |

$n = 51$ $\quad C_f = 11$

$n/2 = 25.6$ $\quad f = 18$

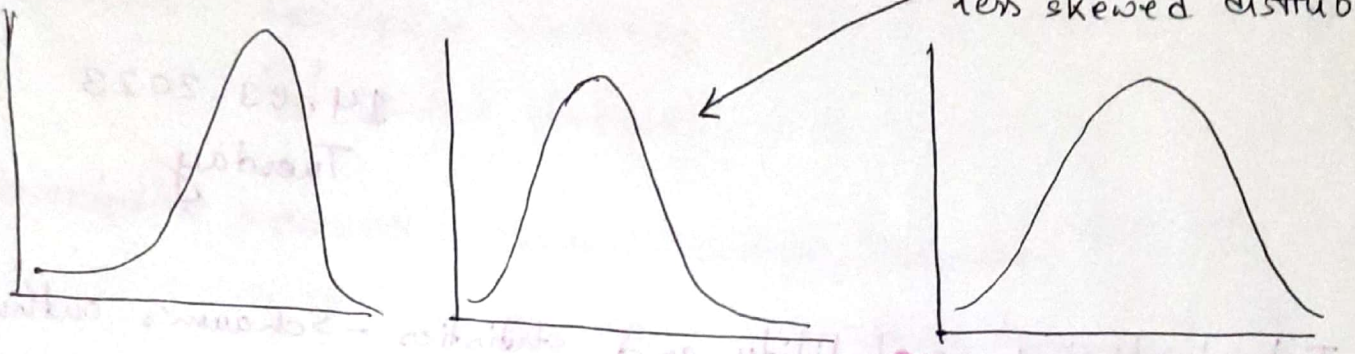
$\therefore$ median $= l + \left(\frac{n/2 - C_f}{f}\right) \times h$

$\therefore$ 145 - 150 group → median class

$h = 5$ $\quad l = 145$

**The Mode**: Value that appears the most
used when data is not sorted.
not affected by extreme values
There may be several modes or no mode (when unique)

Mean − Mode = 3 (Mean − Median) used for average less skewed distribution



Skewed distribution

Normal distribution

15.03.2023
Wednesday

Mean : Grouped data:

| Range | frequency ($f_i$) | class middlepoint ($x_i$) | $f_i x_i$ | $d_i$ |
|---|---|---|---|---|
| 0−10 | 7 | 5 | = | −2 |
| 10−20 | 8 | 15 | = | −1 |
| 20−30 | 20 | 25 A | = | 0 |
| 30−40 | 10 | 35 | = | 1 |
| 40−50 | 5 | 45 | = | 2 |

$$\text{Mean, } \bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} \leftarrow \text{Direct method}$$

$$\text{Mean, } \bar{x} = A + \frac{\Sigma f_i d_i}{N} + h \leftarrow \text{Assumed mean method}$$

$$= A + \frac{\Sigma f_i d_i}{\Sigma f_i} + h$$

Here, A = middle point of middle class (assumed mean)

$d_i$ = deviation = $\frac{x_i - A}{h}$

h = class size

## Mode: Grouped data:

| Range | frequency |
|---|---|
| 30 - 40 | 18 |
| 40 - 50 | 37 |
| 50 - 60 | 45 |
| 60 - 70 | 27 |
| 70 - 80 | 15 |
| 80 - 90 | 8 |

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h = 50 + \frac{8}{8+18} \times 10 = 53.08$$

Here, L = lower limit of mode class

$\Delta_1$ = frequency of mode class - frequency of previous class

$\Delta_2$ = " " " - " " next "

h = class size

## Measure of disperdion or measure of spread:

I) Range (maximum data - minimum data)

II) Standard deviation / $\sqrt{\text{variance}}$

$$\text{Variance} = \frac{\Sigma(x-\bar{x})^2}{n} \leftarrow \text{for population}$$

$$\text{Variance} = \frac{\Sigma(x-\bar{x})^2}{n-1} \leftarrow \text{for sample}$$

| | |
|---|---|
| S1 | 0 $x_1$ |
| S2 | 100 $x_2$ |

$\bar{x} = 50$

variance (population) = $\frac{(0-50)^2 + (100-50)^2}{2} = 50^2 = 2500$

or variance (sample) = $\frac{50^2 + 50^2}{1} = 5000$

Standard deviation, $SD = \sqrt{\text{Variance}}$

for normal distribution. SD = +1 — +3 range

20.03.2023

Monday

## Measure of disperdion:

1. Range

$R = L - S$ L = Largest value, S = Smallest value

Absolute measure of variation

Relative measure of variation

Limitation: cannot tell us anything about the character in the distribution within two extreme observations.

2. Interquartile Range / Quartile deviation

3. Mean deviation

Ungrouped data: $M.D = \frac{\Sigma|x_i - \bar{x}|}{N}$

Grouped data: $M.D = \frac{\Sigma f_i|x_i - \bar{x}|}{N}$ ; $\bar{x} = A + \frac{\Sigma f_i d_i}{N} \times h$

every data is scattered from mean in (mean deviation) range.

4. Variance: Ungrouped data: $\sigma^2 = \frac{\Sigma(x_i - \bar{x})^2}{N}$

(derivation) Prove ⟶ $= \frac{\overline{x_i^2} - (\bar{x})^2}{N}$

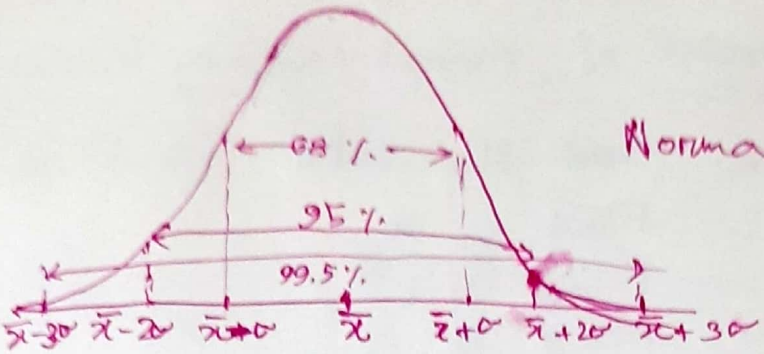
Grouped data: $\sigma^2 = \frac{\Sigma f_i (x_i - \bar{x})^2}{N}$

(derivation) Prove ⟶ $= \left[\frac{\Sigma f_i d_i^2}{N} - \left(\frac{\Sigma f_i d_i}{N}\right)^2\right] \times h^2$

5. Standard deviation: $\sigma = \sqrt{\text{Variance}}$

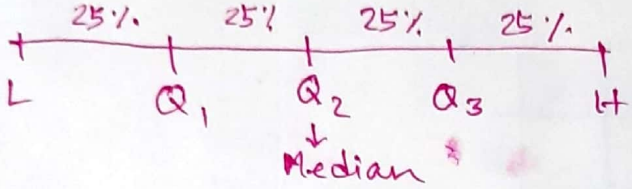
Ungrouped data: $\sigma = \sqrt{\frac{\Sigma(x_9 - \bar{x})^2}{N}}$

Grouped data: $\sigma = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{n}}$




Normal distribution

Interquartile range:




Example:

odd: 2 3 5 7 (8) 9 10 12 15

$Q_1$ $Q_2$ $Q_3$

even: ~~2 3 5 7 8~~ 10 12 14 15 14 16 17 18 10 19 17 17

10 10 (12 14) 14 (15 16) 17 (17 18) 18 19

$Q_1 = 13$ $Q_2 = 15.5$ $Q_3 = 17$

$$Q_i = L + \frac{\frac{i \times N}{4} - \text{P.c.f}}{f} \times h \quad ; \quad i = 1, 2, 3$$

Deciles: $D_i = L + \frac{\frac{i \times N}{10} - \text{P.c.f}}{f} \times h$

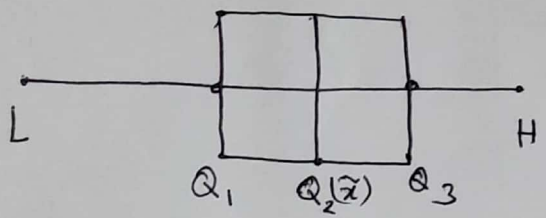
Percentile: $P_i = L + \frac{\frac{i \times N}{100} - \text{P.c.f}}{f} \times h$

22.03.2023
Wednesday

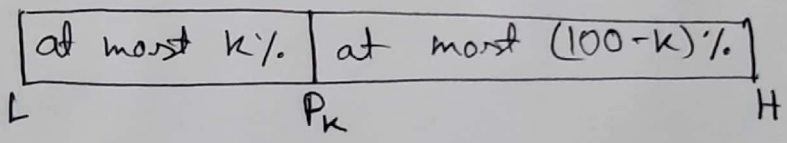
**Midquartile**: Parameter of central tendency measure.

**Five Number Summary**: Sort the data
Find L, H
Find $Q_2$ ($\tilde{x}$)
Find $Q_1$, $Q_3$

**Box whisker plot**:
Graphical representation:




**Percentile**:




Ungrouped data: $I = n \frac{P}{100}$ → $p^{th}$ percentile

I: position of the percentile; n: sample size

Sort the data : 11 11 14 15 16 16 17 19 22 25 26 27 31 34 36

$n = 15$

$P = 30$

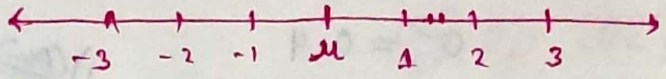
$$I = 15 \frac{30}{100} = 4.5 \approx 5$$ → round it up to next integer

∴ Percentile at 5th position → 16

When I = n (integer) Percentile will be $\frac{n^{th} + (n+1)^{th}}{2}$

## Z-score: (standard score)

(Z = −3 to +3)



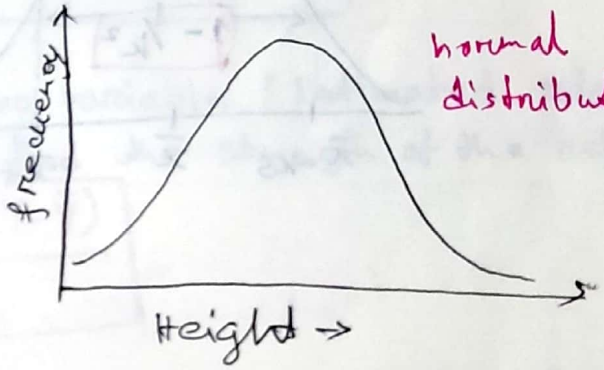

Metric
Metrix
to describe data

$$Z = \frac{\text{value} - \text{mean}}{\text{st. dev.}} = \frac{x - \bar{x}}{s} = \frac{x - \mu}{\sigma}$$

| Class A | Class B |
|---|---|
| $\mu = 80$ | $\mu = 73$ |
| $\sigma = 5$ | $\sigma = 6$ |
| Marks = 87 | Marks = 82 |
| $Z = \frac{87 - 80}{5}$ | $Z = \frac{82 - 73}{6}$ |
| $= 1.4$ | $= 1.5$ (better performance) |

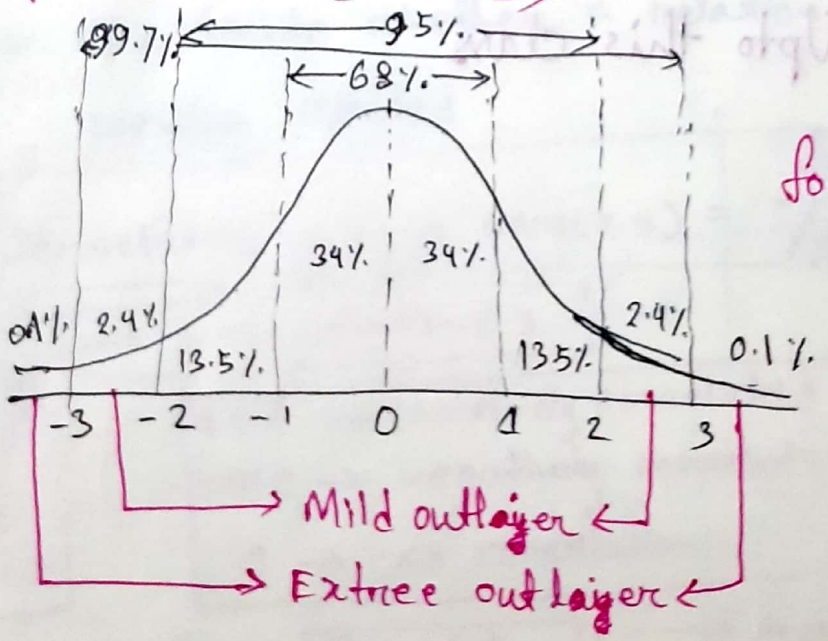
Frequency distribution:

Distribution of male height:




normal distribution

## The Empirical Rule (z-score)




for normal distribution

Outlayer: data with peculier characteristics

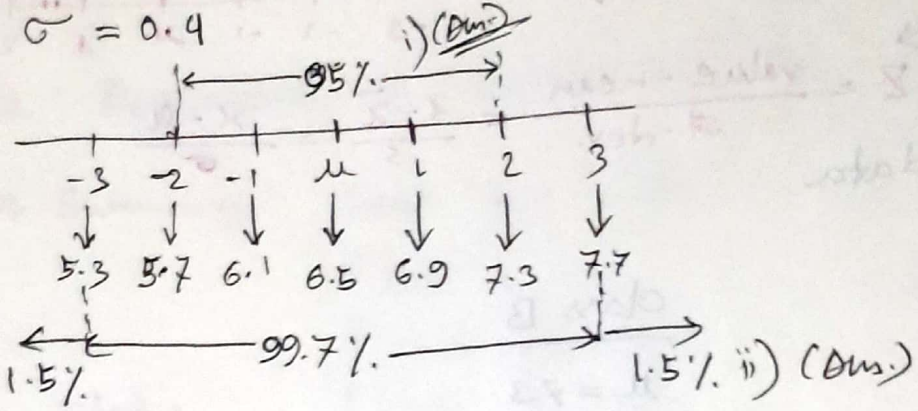
~~Standization~~ Standardization

Standardized data

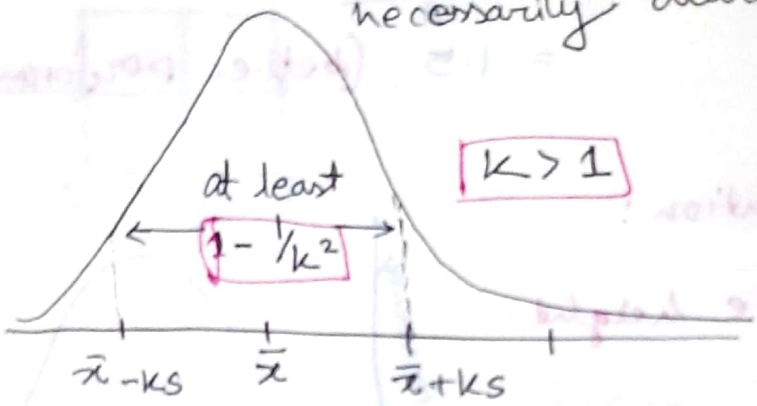
Example: $\mu = 6.5$

$\sigma = 0.4$

i), ii):




Chebysher's Theorem: When the distribution is not necessarily always normal.




CT syllabus: Upto this class

Wednesday
03.05.2023

Moment, Skewness, Kurtosis

Univariate → Range, Standard deviation, Variance
Univariate → Mean, Median, Mode

Multivariate → Covariance, Correlation, Regression

Stock price of company X & Y:

| DAY | X | Y | $x-\bar{x}$ | $y-\bar{y}$ | $(x-\bar{x})(y-\bar{y})$ |
|---|---|---|---|---|---|
| 1 | 30 | 5 | −3 ↓ | −1 ↓ | 3 |
| 2 | 35 | 8 | +2 ↑ | +2 ↑ | 4 |
| 3 | 40 | 8 | +7 ↑ | +2 ↑ | 14 |
| 4 | 25 | 4 | −8 ↓ | −2 ↓ | 16 |
| 5 | 35 | 5 | +2 ↑ | −1 ↓ | −2 |
| Mean | 33 | 6 | | | Sum = 35 |

Covariance = $\frac{35}{4}$

**Covariance**: Relation between two variable (Estimated relation is not ideal) (cannot find the strength of the relation)

$$\text{covariance}_{(x,y)} = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{n-1}$$

**Correlation**: Can decide whether a relation is strongly or weakly related.

works on simple relation

P = Karl Pearson Coefficient → works only on linear relation

$$\text{Correlation}(x,y) = \text{corr}(x,y) = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y}$$

standard deviation

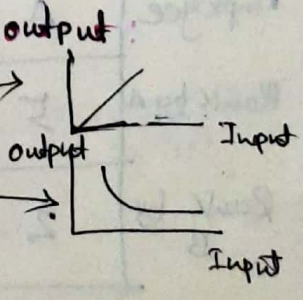
$-1 \leq \text{correlation} \leq 1$

+ve → positively correlated
−ve → negatively correlated
0 → no correlation

$$\sigma = SD = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n}}$$

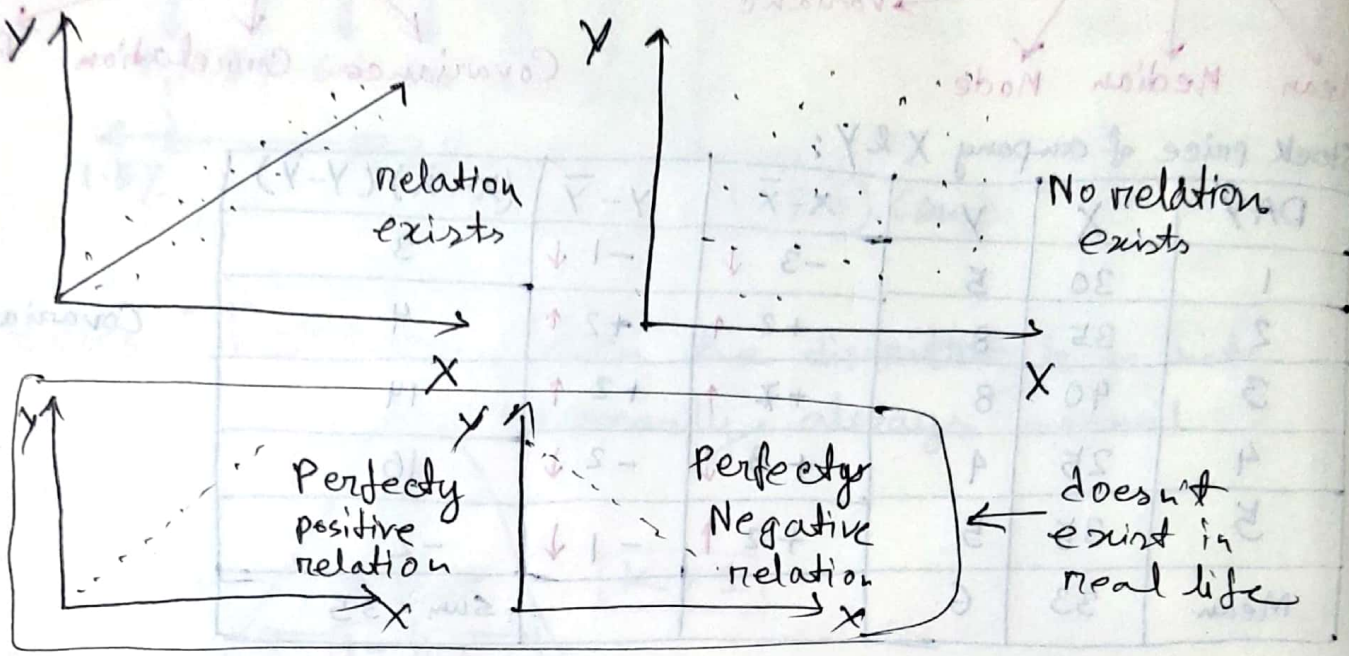
P = ~~Karl Pearson~~ Coefficient

Correlation → Positive, Negative
Correlation → Linear, Non linear
Correlation → Simple (two variable depend on eachother), Multiple (one variable depend on multiple variables)



# Simple - Linear Correlation: i) Scatter plot
## ii) Karl pearson coefficient

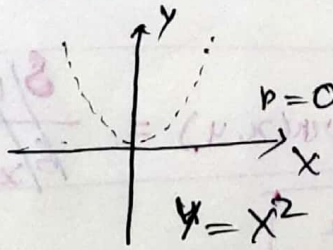
### Scatter plot :




### Karl pearson cofficient :

$$r = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma (x-\bar{x})^2}\sqrt{\Sigma (y-\bar{y})^2}}$$

if, $r = 0.1$ → there is no simple- linear correlation, but correlation exists and it may be polynomial correlation




| Employee | A | B | C | D | E | F | |
|---|---|---|---|---|---|---|---|
| Rank by A | 5 | 4 | 3 | 2 | 1 | 6 | $R_1$ |
| Rank by B | 2 | 4 | 5 | 6 | 7 | 8 | $R_2$ |

Special case for Karl Pearson coefficient for ordinal data :

Spearman's Rank Algorithm

Spearman's Rank Algorithm formula!

For unique Rank values →

$$r_s = 1 - \frac{6 \Sigma D^2}{N(N^2-1)}$$

$$D^2 = (R_1 - R_2)^2$$

Monday
08.05.2023

Example: Ranking relation between Manager 1 & Manager 2

| Employee | $R_1$ | $R_2$ | $D^2 = (R_1 - R_2)^2$ |
|---|---|---|---|
| A | 10 | | |
| B | 2 | | |
| C | 1 | | |
| D | 4 | | |
| E | 3 | | |
| F | 6 | | |
| G | 5 | | |
| H | 8 | | |
| I | 7 | | |
| J | 9 | | |
| N = 10 | | | |

$r_s = ?$

Example: Compute rank correlation coefficient

to rank: sort the preliminary test and rank in ascending order.

Tie in rank (in the same column):

$$r_s = 1 - \frac{6\{\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \cdots\}}{(N^3 - N)}$$

Example:

| Marks in Bangla | $R_1$ | Marks in English | $R_2$ | $D^2 = (R_1^2 - R_2^2)$ |
|---|---|---|---|---|
| 15 | 2 | 40 | 6 | |
| 20 | 3.5 | 30 | 4 | |
| 28 | 5 | 50 | 7 | |
| 12 | 1 | 30 | 4 | |
| 40 | 6 | 20 | 2 | |
| 60 | 7 | 10 | 1 | |
| 20 | 3.5 | 30 | 4 | |
| 80 | 8 | 60 | 8 | |
| N = 8 | | | | |

12 15 20 20 28 40 60 80

$R_1 \rightarrow$ 1 2 3 4 5 6 7 8

$R_1 \rightarrow$ 1 2 3.5 3.5 5 6 7 8

$3+4$ ÷ 3.5

$m_1 \rightarrow$ (2)

10 20 30 30 30 40 50 60

$R_2 \rightarrow$ 1 2 3 4 5 6 7 8

$R_2 \rightarrow$ 1 2 4 4 4 6 7 8

$3+4+5 = 4$

$m_2 \rightarrow$ (3)

$r_s =$ (0) → No correlation between Marks in Bangla & Marks in English

Statistics

- Descriptive (Correlation) — relationship exists mean descriptive
- Inferential (regression) (classification) — Predict Continuous value

# Regression Analysis:

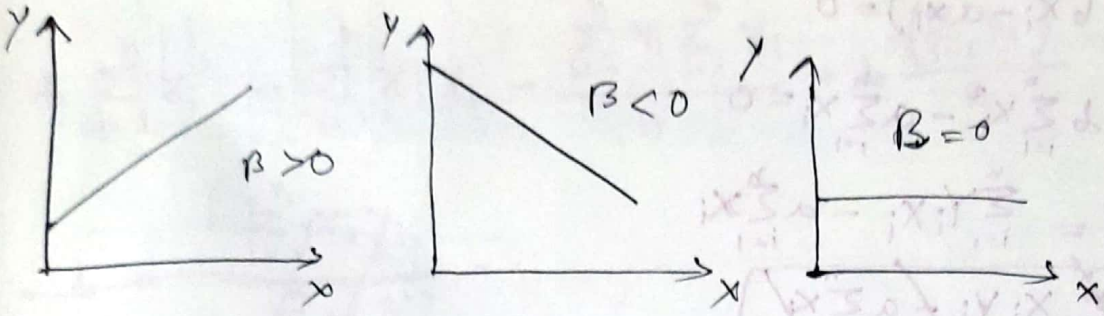
Dependent variable

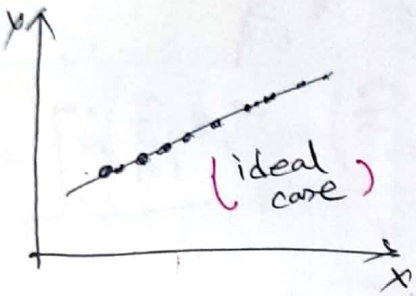
Independent variable

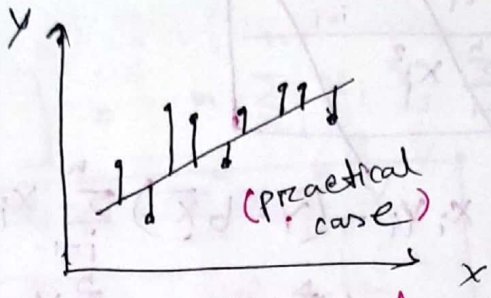
$Y = f(x)$

# Simple linear dependency

$Y = \beta X + \alpha$




**Deterministic Relationship**
$Y = \beta X + \alpha$

Can predict accurate point when $\beta, \alpha$ is known

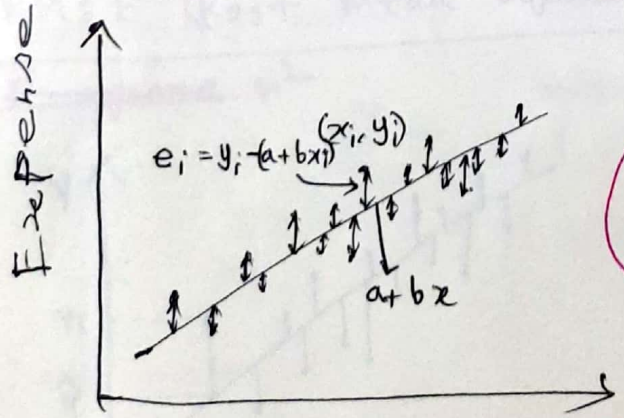
**Non-deterministic Relationship**
$Y = \beta X + \alpha + \epsilon$

Can not predict accurate point with only $\beta$ & $\alpha$ – slight randomness

## Simple - Linear Regression Model :

- Simple → No. of independent variable 1 (single)
- Linear → relationship between independent & dependent variable is linear (straight line)

predicted value → $\hat{Y} = bx + a$




$$\sum (Y_i - \hat{Y})^2$$
$$= \sum (Y_i - bx_i - a)^2$$

} error calculation

Sum of square error,

$$SSE = \sum (Y_i - bx_i - a)^2$$

$$\frac{\delta SSE}{\delta a} = -2 \sum_{i=1}^{n} (Y_i - bx_i - a) = 0$$

$$\Rightarrow \sum_{i=1}^{n} Y_i - b \sum_{i=1}^{n} x_i - na = 0$$

$$\Rightarrow na = \sum_{i=1}^{n} Y_i - b \sum_{i=1}^{n} x_i$$

$$\therefore a = \frac{\sum_{i=1}^{n} Y_i}{n} - b \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\therefore a = \bar{Y} - b\bar{X}$$

$$\frac{\delta SSE}{\delta b} = -2\sum_{i=1}^{n}(Y_i - bx_i - a)\,x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n}(Y_i x_i - b x_i^2 - a x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} Y_i x_i - b\sum_{i=1}^{n} x_i^2 - a\sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} Y_i x_i - a\sum_{i=1}^{n} x_i$$

$$\Rightarrow \boxed{b = \frac{\sum_{i=1}^{n} x_i Y_i - a\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}}$$

$$\Rightarrow b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i Y_i - (\bar{Y} - b\bar{x})\sum_{i=1}^{n} x_i$$

$$\Rightarrow b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i Y_i - \left(\frac{\sum_{i=1}^{n} Y_i}{n} - b\frac{\sum_{i=1}^{n} x_i}{n}\right)\sum_{i=1}^{n} x_i$$

$$\boxed{b = \frac{n\sum x_i Y_i - \sum x_i \sum Y_i}{n\sum x_i^2 - (\sum x_i)^2}}$$

10.05.2023
Wednesday

$$b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} + b\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$b = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

$$b\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right] = n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

$$\therefore b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Example:

Find $a, b$

Find $\hat{Y} = bx + a$

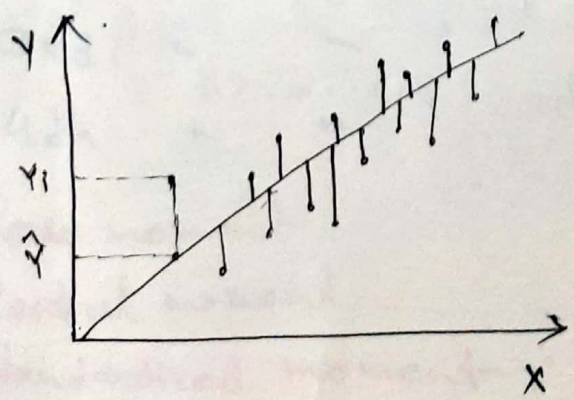
Find $\hat{Y}$ for a given $X$

When $X$ dependent & $Y$ independent variable, $X = f(Y) = dY + c$

MSE (Mean square error) $= \frac{\Sigma(Y_i - \hat{Y})^2}{n}$

RMSE (Root mean square error) $= \sqrt{\frac{\Sigma(Y_i - \hat{Y})^2}{n}}$

~~R-square~~ $r^2$



SST (Total sum of square) $= \sum_{i=1}^{n} (Y_i - \bar{y})^2$

SSR (sum of square of regression) $= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

SSE (sum of square ~~exple~~ error) $= \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$

$$r^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y})^2}{\sum_{i=1}^{n} (\hat{Y}_i - \bar{y})^2} = 0.5$$

can explain 50% variance of the dependent variable

- When $r^2 \approx 0$ it doesn't hold linear regression

Linear
Multiple Regression Model:

$$\hat{Y} = b_1 x_1 + b_2 x_2 + a$$

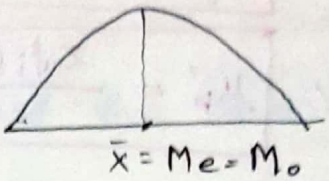
Polynomial Regression Model:

$$\hat{Y} = a + b_1 X + b_2 X^2$$

22.05.2023
Monday

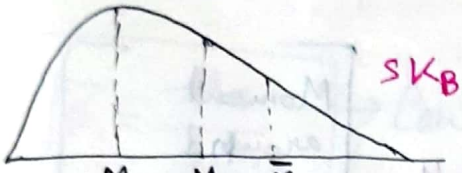
## Skewness:

For perfect normal distribution $\bar{x} = Me = Mo$




$SK_B = 0$
$SK_P = 0$

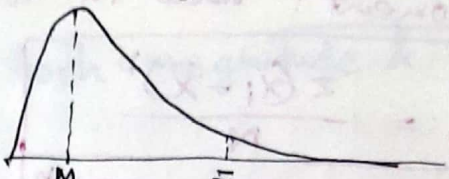
Symmetrical Distribution



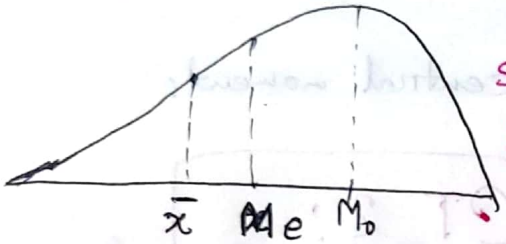

$SK_B > 0$
$SK_P > 0$

$\bar{x} > Me > M_o$

Positively ~~symmet~~ Skewed Distribution




If skewness increases the distance of Mode and Mean increases.




$SK_B < 0$
$SK_P < 0$

$\bar{x} < Me < M_o$

Negatively Skewed Distribution

Karl Pearson's coefficient

$$\therefore \text{Skewness, } SK_p = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

so that translation & scaling invariance doesn't affect skewness.

## Moments: Distance & Direction

provide unique characterization of a distribution.
unification of all measures (central tendency, variation).

1st order moment: refers the mean of dataset
2nd " " : ~~vari~~ " " variance " " "
3rd " " : " " skewness " " "
4th " " : " " kurtosis " " "

Raw moment
Central moment
Standardized moment

## Raw moment:

Assumed mean

ungrouped data, $\mu_r' = \frac{\sum (x_i - A)^r}{N}$ $r = 1, 2, 3, 4$

grouped data, $\mu_r' = \frac{\sum f_i (x_i - A)^r}{N}$ $r = 1, 2, 3, 4$

with respect to origin,

$\mu_r' = \frac{\sum x_i^r}{N}$

$\mu'r = \frac{\sum f_i x_i^r}{N}$

## Central moment: used in real life

ungrouped, $\mu_r = \frac{\sum (x_i - \bar{x})^r}{N}$ $r = 1, 2, 3, 4$

grouped, $\mu_r = \frac{\sum f_i (x_i - \bar{x})^r}{N}$ $r = 1, 2, 3, 4$

Moment around mean always zero.
$\therefore \mu_1 = 0$

Relation between raw moment & central moment:

$\mu_1 = 0$

$\mu_2 = \mu_2' - \mu_1'^2$

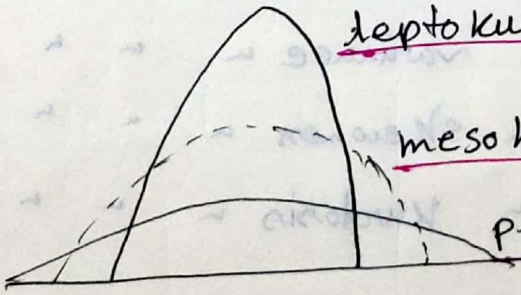
$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$

$\mu_4 = $ [illegible]

CT 2:
~~Relation~~
Correlation,
Regression,
Skewness,
Moment,
Kurtosis

## Kurtosis:

measure the peakedness



leptokurtic, kurtosis > 3

mesokurtic, kurtosis = 3

platykurtic, kurtosis < 3

Measures of skewness & kurtosis using moments:

Relative measure of skewness, $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$

always positive ← $\beta_1$ → variance

$\beta_1$ can determine the magnitude of the skewness but not the direction

$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}}$ → Can determine both magnitude & direction

Relative measure of kurtosis, $\beta_2 = \frac{\mu_4}{\mu_2^2}$

or $\gamma_2 = \beta_2 - 3$ ← excess kurtosis

$\beta_2 > 3 \rightarrow \gamma_2 > 0 \rightarrow$ leptokurtic

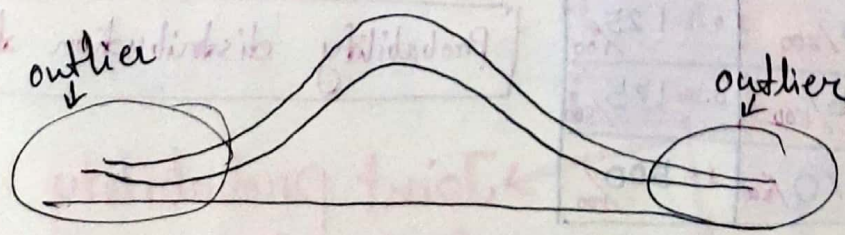
$\beta_2 = 3 \rightarrow \gamma = 0 \rightarrow$ mesokurtic

$\beta_2 < 3 \rightarrow \gamma < 0 \rightarrow$ platykurtic

To measure kurtosis we used to concentrate on peak of the data but nowadays we concentrate more on tail or outlier of the data.

amount of outlier ↑ kurtosis ↑

amount of outlier ↓ kurtosis ↓



28.05.23
Wednesday

## Descriptive Statistics

**Univariate**

- Mean
- Median
- Mode
- Variance
- Standard deviation
- Normal distribution / z-score
- Skewness
- Kurtosis
- Moments

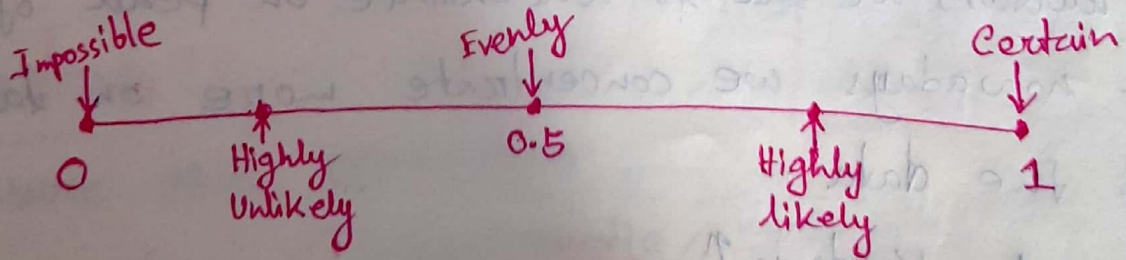
**Multivariate**

- Covariance
- Correlation
- Regression

## Probability :

- Joint probability
- Marginal probability
- Conditional probability
- Independence

## Event probability :




| # | Male | Female | Total |
|---|---|---|---|
| Friends | 0.16 = 80/500 | 0.24 = 120/500 | 0.4 = 200/500 |
| Big Bang Theory | 0.2 = 100/500 | 0.05 = 25/500 | 0.25 = 125/500 |
| Others | 0.1 = 50/500 | 0.25 = 125/500 | 0.35 = 175/500 |
| Total | 0.46 = 230/500 | 0.54 = 270/500 | 1 = 500/500 |

Probability distribution table.

Joint probability Distribution {P(A∩B)}

Marginal probability distribution {P(A)}

Marginal probability : single term is represented
or ~~single~~ probability
simple

* $P(G = M \text{ OR } F = \text{Big Bang Theory}) = 0.51 = 0.16 + 0.2 + 0.1 + 0.05$

* $\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)} = 0.46 + 0.25 - 0.2 = 0.51$

* We know that subscriber is female then what is the probability of first choice to be Big Bang theory ?

$$P(F = \text{"BBT"} \mid G = 'F') = \frac{0.05}{0.54} = 0.09$$

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

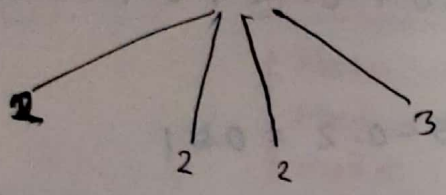
an event ~~does~~ already occured and another will happen w.r.t. it. : Conditional probability

Independent event : $\boxed{P(A|B) = P(A)}$ No relation between events A & B

If two events are independent $\Rightarrow \boxed{P(A \cap B) = P(A) \cdot P(B)}$

29.05.2023
Monday

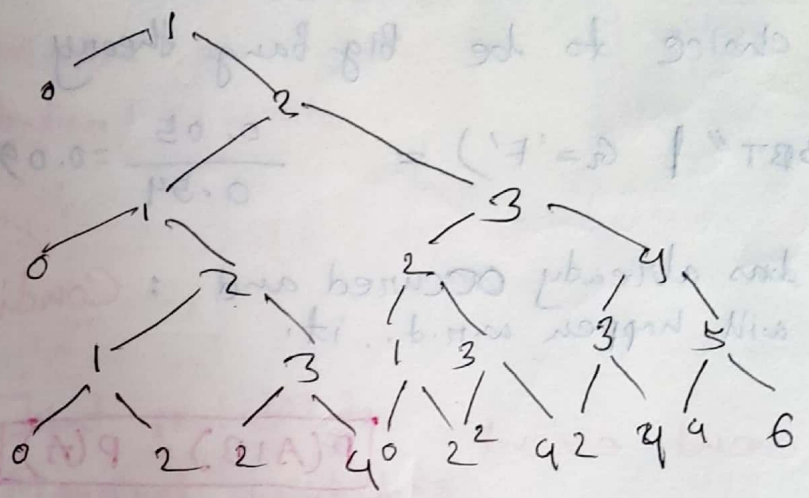
# 9 toys




$= {}^9C_3 * {}^6C_2 * {}^4C_2 * {}^2C_2$

$= 7560 = \frac{9!}{2!\,2!\,2!\,3!}$

2.3.3 ÷




a) 14 b) 2 c) 4

Poker:

Royal flush: 10 J Q K A → same suit (4 possible)

probability : $\frac{4}{{}^{52}C_5}$

Straight flush: 5 consecutive card → same suit

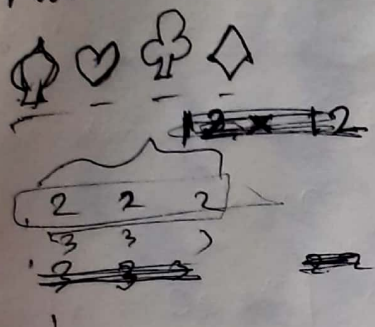
$\frac{9 \times 4}{{}^{52}C_5}$

2 3 4 5 6
3 4 5 6 7
4 5 6 7 8
5 6 7 8 9
6 7 8 9 10
10 J Q K A → Royal flush
A 2 3 4 5
K A 2 3 4
Q K A 2 3
J Q K A 2

Four of a kind : 4 card same number , 1 card different

$$\frac{13 \times (52-4)}{{}^{52}C_5}$$

Full house : 3 of a kind (3 card same number), 1 pair (2 card same)



$$\frac{13 \times {}^{4}C_3 \times 12 \times {}^{4}C_2}{{}^{52}C_5}$$

Flush : 5 card same suit

$$\frac{{}^{13}C_5 \times 4 - (4 + (9 \times 4))}{{}^{52}C_5}$$

Straight : 5 consecutive card → suit doesn't matter

$$\frac{(10 \times 4^5) - (40)}{{}^{52}C_5}$$

31.05.2023
Wednesday

# Conditional Probability
# Bayes Theorem
# Naive Bayes Theorem

## Conditional Probability:

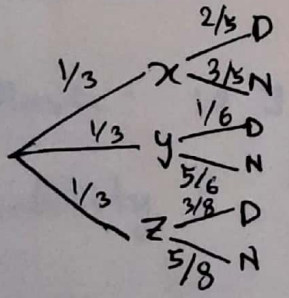
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Dependent

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

\# 

| 10 | 6 | 8 |
|---|---|---|
| D = 4 | D = 1 | D = 3 |
| ND = 6 | ND = 5 | ND = 5 |




$$P(D) = \frac{1}{3} \cdot \frac{2}{5} + \frac{1}{3} \cdot \frac{1}{6} + \frac{1}{3} \cdot \frac{3}{8}$$

$$P(ND) = 1 - P(D)$$

$$P(N|Z) \quad \frac{P(N \cap Z)}{P(Z)} = \frac{P(Z) \cdot P(N|Z)}{P(X) \cdot P(N|X) + P(Y) \cdot P(N|Y) + P(Z) \cdot P(N|Z)}$$

$P(N|Z)$ : $P(Z|N)$, $P(X|N)$, $P(Y|N)$

Bayes Theorem

$$P(Z|N) = \frac{P(N \cap Z)}{P(N)} = \frac{P(Z) \cdot P(N|Z)}{P(X) \cdot P(N|X) + P(Y) \cdot P(N|Y) + P(Z) \cdot P(N|Z)}$$

\# Example 4.8

1.

$P(f) = 30\%$.

$P(S) = 40\%$.

$P(J) = 20\%$.

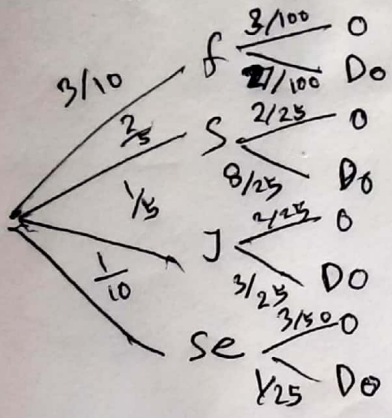
$P(Se) = 10\%$.

$P(O|f) = 10\%$.

$P(O|S) = 20\%$.

$P(O|J) = 40\%$.

$P(O|Se) = 60\%$.




$P(O) = 3/100 + 2/25 + 2/25 + 3/50$

$= 0.25$

$P(J|O) = \frac{P(J \cap O)}{P(O)} = \frac{P(J)\,P(O|J)}{P(O)}$

$= \frac{2/25}{0.25}$.

Independent : $P(A \cap B) = P(A) \cdot P(B)$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Chapter 2, 3, 4, 5

09.07.2023
Sunday

$$P(C_i|X) = \frac{P(X|C_i)\,P(C_i)}{P(X)}$$

$C_i$ = Class we belong to (Yes / No) class

$X$ = feature vector

$P(Yes|X)$

$P(No|X)$

Need to find $P(X|C_i) * P(C_i)$

$$P(X|C_i) = \prod P(x_1|C_i)\,P(x_2|C_i) \ldots P(x_n|C_i)$$

Example:

$X = (\underbrace{age <= 30}_{x_1}, \underbrace{income = medium}_{x_2}, \underbrace{student = yes}_{x_3}, \underbrace{credit\_rating = fair}_{x_4})$

find each $P(X_n | C_i)$ → for both Yes & No class from dataset, then multiply.

the answer of $P(X|C_i) * P(C_i)$ shows X belong to which class. (it belongs to the class which has greater $P(X|C_i) * P(C_i)$ value)

## Random variable: (Important) (Chapter 5)(Book)

Always map to the value of the outcome of a Random Process.

Outcome cannot be predicted

represented with capital (X) letter always.

$$X = \begin{cases} 0, & \text{if outcome} \rightarrow \text{tail} \\ 1, & \text{if outcome} \rightarrow \text{head} \end{cases}$$

* Discrete random variable (countable)(finite)
* Continuous random variable (cannot be counted exactly) (e.g., predicting the amount of rain tomorrow)

Example 5.3: $X = max(a,b) =$ images (we can count it, thus it is discrete) $1, 2, 3, 4, 5, 6$ (all possible values)

$Y = sum(a,b) = 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$

X:

$f(1) = (1,1) \rightarrow 1/36$

$f(2) = (1,2), (2,2), (2,1) \rightarrow 3/36$

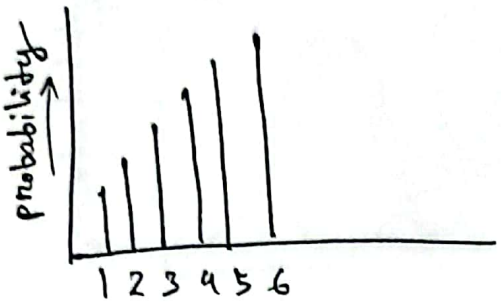
$f(3) = (1,3), (2,3), (3,3), (3,2), (3,1) \rightarrow 5/36$

$f(4) = 7/36$, $f(5) = 9/36$, $f(6) = 11/36$

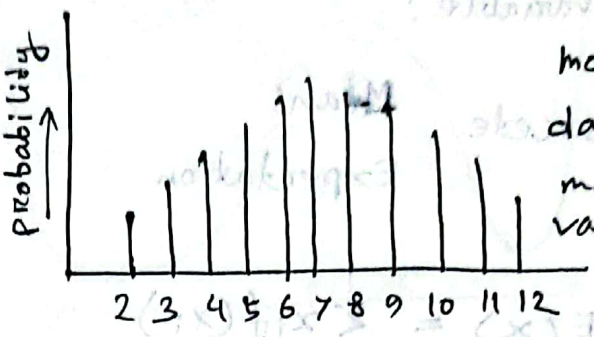
Y: same way find for Y

Distribution:

Distribution of X:




Distribution of Y:




more spread data = more variance

## Expectation of a finite random variable:

(Mean)

$$E = E(X) = x_1 f(x_1) + x_2 f(x_2) + \dots + x_n f(x_n) = \sum x_i f(x_i)$$

or, $E = E(X) = x_1 P_1 + x_2 P_2 + \dots + x_n P_n = \sum x_i P_i$,

where, $P_i = f(x_i)$

Example 5.8:

| $x_i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $P_i$ | $1/64$ | $6/64$ | $15/64$ | $30/64$ | $15/64$ | $6/64$ | $1/64$ |

Probability of expecting head when tossing 6 times.

(use $n_{cr}$)

Example 5.9: a dice is tossed. if value 2, 3, 5 → wins
if value 1, 4, 6 → losses

| $x$ | 2 | 3 | 5 | -1 | -4 | -6 |
|---|---|---|---|---|---|---|
| $f(x)$ | $1/6$ | $1/6$ | $1/6$ | $1/6$ | $1/6$ | $1/6$ |

$$E(x) = 2(1/6) + 3(1/6) + 5(1/6) - 1(1/6) - 4(1/6) - 6(1/6) = -1/6$$

fair if $E(x)$ is 0

not a favourable game as $E(x)$ is negative. the game would have been

10.07.2023
Monday

Random variable:

Discrete. Mean. Expectation.

$$\mu = E(x) = \sum x_i f(x_i)$$

$f(x_i)$ → Probability distribution function

$$f(x_i) \geq 0 \quad ; \quad \sum f(x_i) = 1$$

$$\begin{aligned} Var(x) &= (x_1 - \mu)^2 f(x_1) + (x_2 - \mu)^2 f(x_2) + \cdots + (x_n - \mu)^2 f(x_n) \\ &= \sum (x_i - \mu)^2 f(x_i) \\ &= E(x^2) - \mu^2 \\ &= \sum x_i^2 f(x_i) - \mu^2 \end{aligned}$$

$$\sigma = \sqrt{Var(x)}$$

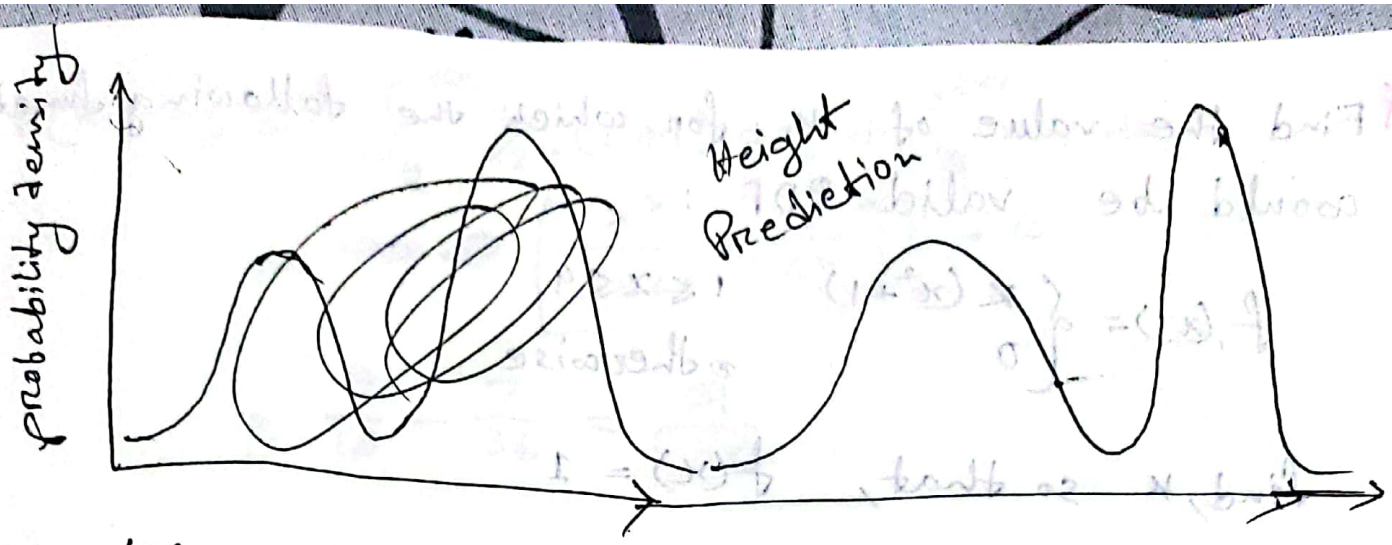
Example 5.11

Standardized Random variable (Later)

Joint random ---- (Skip)

Continuous Random Variable:

Here, $f(x_i)$ → Probability Density Function (PDF)




Here we work with intervals instead of exact point.

$$P(100 < X < 150) = \int_{100}^{150} f(x)\,dx$$

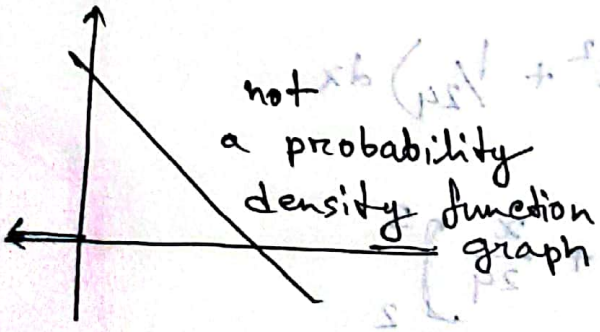
$$\boxed{P(a < x < b) = \int_a^b f(x)\,dx}$$

The total area under the curve, $\int_{-\infty}^{\infty} f(x)\,dx = 1$

Real life graph function:

$f(x)$ = very complecated

thus we use more theoradically made up graph.




to find wheather a graph is PDF or not:

1) Is the area > 0 ?
2) Is the area / probability = 1 ?

# Find the value of $k$ for which the following function could be valid PDF :

$$f(x) = \begin{cases} k(x^2+1) & 1 \le x \le 4 \\ 0 & \text{otherwise} \end{cases}$$

Find, $k$ so that, $f(x) = 1$

Ans. : $k = 1/24$

# $$f(x) = \begin{cases} 1/24\,(x^2+1) & ; \ 1 \le x \le 4 \\ 0 & ; \text{otherwise} \end{cases}$$

i) $P(x = 2)$

ii) $P(2 \le x \le 3)$

iii) $P(x \ge 2)$

i) $P(x = 2) \approx 0$

ii) $$P(2 \le x \le 3) = \int_2^3 1/24\,(x^2+1)\,dx$$

$$= \int_2^3 (1/24\,x^2 + 1/24)\,dx$$

$$= \left[\frac{x^3}{3 \times 24} + \frac{x}{24}\right]_2^3$$

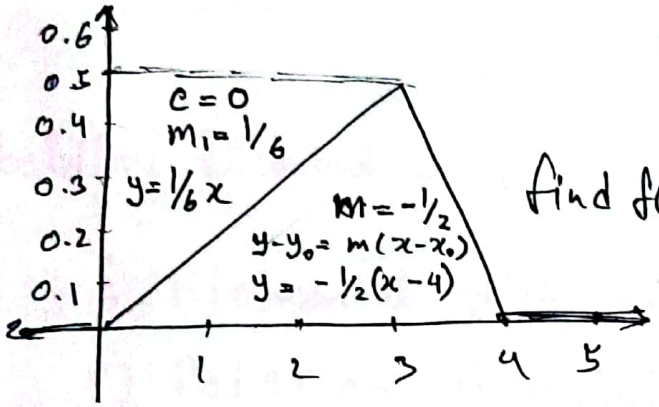
$$= \frac{11}{36}$$

iii) $P(x \geq 2) = \int_2^4 \frac{1}{24}(x^2+1)\,dx + \int_4^\infty \cdots dx$

$= \left[\frac{x^3}{3\times 24} + \frac{x}{24}\right]_2^4 + 0$

$= \frac{19}{18} - \frac{7}{36} = \frac{31}{36}$

#




find $f(x)$

$$f(x) = \begin{cases} 1/6\,x & ; 0 \leq x \leq 3 \\ -1/2(x-4) & ; 3 < x < 4 \end{cases}$$

Cumulative Density function (CDF) → next class

17.07.2023

Monday

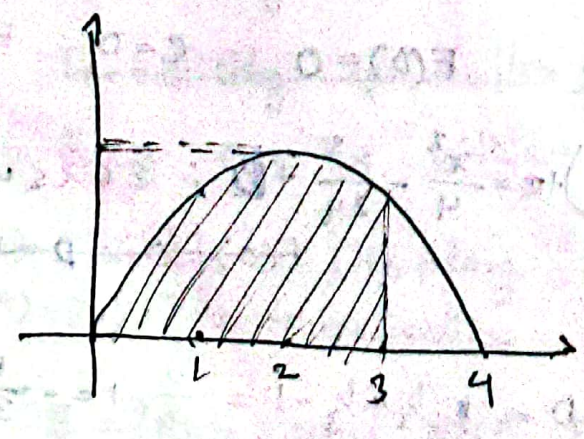
Cumulative Density Function

$F(x_0) = P(x \leq x_0)$

summation of density function

→ Discrete: $F(x_0) = \sum_{x=0}^{x_0} f(x_i)$

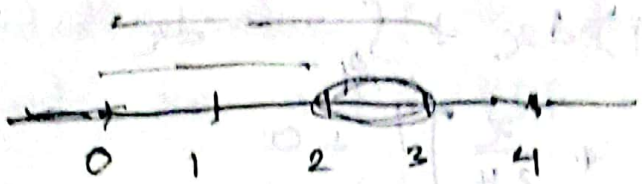
→ Continuous (integration): $F(x_0) = \int_0^{x_0} f(x)\,dx$




- Probability Density function (PDF)
- Cumulative " " (CDF)

identify first when doing math

#




$P(2<x<3) \longrightarrow F(3) - F(2)$

$P(x>2) \longrightarrow 1 - F(2)$

$$f(x) = \begin{cases} 0 & x<0 \\ \frac{1}{32}(6x^2 - x^3) & 0 \le x \le 4 \\ 1 & x>4 \end{cases}$$

CDF

Integrate

PDF $\longrightarrow$ CDF

Differentiate

CDF $\longrightarrow$ PDF

} Conversion

$$f(x) = \begin{cases} \frac{1}{8}x^2 & ; 0 \le x < 2 \\ \frac{1}{8}x(4-x) & ; 2 \le x \le 4 \\ 0 & ; \text{otherwise} \end{cases}$$

PDF

$$F(x) = \begin{cases} 0 & ; x<0 \\ \frac{1}{24}x^3 & ; 0 \le x \le 2 \\ \frac{x^2}{4} - \frac{x^3}{24} - \frac{1}{3} & ; 2 \le x \le 4 \\ 1 & ; x>4 \end{cases}$$

CDF

$F(x) = \int \frac{1}{8}x^2\,dx = \frac{1}{24}x^3 + c \; ; \; 0 \le x \le 2 \qquad F(0) = 0 \;\therefore\; c = 0$

$F(x) = \int \frac{1}{8}(4x - x^2)\,dx = \frac{1}{8}\left(2x^2 - \frac{x^3}{3}\right) + c = \frac{x^2}{4} - \frac{x^3}{24} + D \; ; \; 2 \le x \le 4$

~~$F(0) = 0 \;\therefore\; D = 0$~~

$F(2) = \frac{1}{24}2^3 = \frac{8}{24} = \frac{1}{3}$

$\therefore \frac{1}{3} = \frac{2^2}{4} - \frac{2^3}{24} + D \;\therefore\; D = \frac{1}{3} + \frac{1}{3} - 1 = \frac{2}{3} - 1 = -\frac{1}{3}$

# Mean & Variance of Continuous Random Variable

$$\mu = E(x) = \sum x_i \cdot f(x_i)$$
$$\sigma^2 = E(x^2) - \mu^2$$

→ Discrete random variable

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx$$
$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu^2$$

→ Continuous random variable

## Probability Distribution of Discrete Random Variable:

i) Binomial distribution
ii) Poission distribution

### Binomial Distribution:

can be applied to problems which have specific two outcomes
Success, Failure

### Bernoulli trial

Process of occurence → Bernoulli
Each trial is independent.
Probability of succes (P) → remain constant

# ~~Probability of an item~~ 3 items are picked are random, they can be either defective or non-defective.

success (P) failure (1-P) (q)

P(D) = 25% = 1/4 | 3 → Display
P(N) = 75% = 3/4 | 2 → defective

P (2 defective) = $3 \times \left(\frac{1}{4}\right)^2 \times \frac{3}{4}$

S = {NNN, NDN, NND, DNN, NDD, DND, DDN, DDD} = 9/64

$$P(x \text{ success}) = {}^{n}C_x \times p^x \times (1-p)^{n-x}$$

$$\text{or} = {}^{n}C_x \times p^x \times q^{n-x}$$

(p) = success

(1-p)/(q) = failure

# 10 → display → n

5 → non defective → n-x

$$P(x \text{ success}) = {}^{10}C_5 \times (1/4)^5 \times (3/4)^5$$

10 → display

at least 5 → defective

$$P(x \text{ success}) = P(5) + P(6) + P(7) + P(8) + P(9) + P(10)$$

$$= b(5, 10, 1/4) + b(6; 10, 1/4) + b(7) + \ldots$$

$$= \sum_{x=5}^{10} \left\{ {}^{10}C_x \times (1/4)^x \times (3/4)^{10-x} \right\}$$

$$= 1 - \sum_{x=0}^{4} \left\{ {}^{10}C_x \times (1/4)^x \times (3/4)^{10-x} \right\}$$

} Use calculator

10 → display

5 - 7 → defective

$$\sum_{x=0}^{7} - \sum_{x=0}^{4}$$

18.07.2023
Tuesday

## Discrete

Binomial distribution
Poisson distribution

## Continuous

Uniform distribution
Normal distribution

# a) $\sum_{x=1}^{20} {}^{20}C_x \left(\frac{3}{100}\right)^x \left(\frac{97}{100}\right)^{20-x}$

or $= 1 - {}^{20}C_0 \left(\frac{3}{100}\right)^0 \times \left(\frac{97}{100}\right)^{20}$

$= 0.4562$

b) ${}^{10}C_3 (0.4562)^3 (1-0.4562)^7 = 0.16023$

Mean & variance in binomial distribution:

$X = I_1 + I_2 + \cdots + I_n$

$\mu = E(X) = E(I_1) + E(I_2) + \cdots + E(I_n) = \underbrace{p + p + \cdots p}_{n \text{ terms}} = np$

$I \begin{cases} \to 0 \times q \\ \searrow 1 \times p \end{cases} \to$ mean $p$

$I_j \to E[I_j] = (0)(q) + (1)(p) = p$

$\sigma_{I_1}^2 = \sum x^2 f(x_i) - \mu^2$

$= 0^2 \times q + 1^2 \times p - p^2$

$= p - p^2 = p(1-p) = pq$

$\therefore \sigma_x^2 = \sigma_{I_1}^2 + \sigma_{I_2}^2 + \cdots + \sigma_{I_n}^2$

$= \underbrace{pq + pq + \cdots + pq}_{n \text{ terms}}$

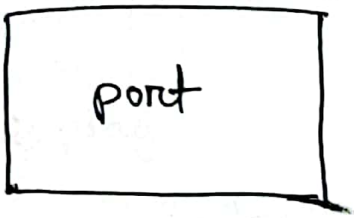
$= npq$

Poisson distribution: একটা Random event এর average rate জানা আছে, এখন নির্দিষ্ট সংখ্যক বার ঘটনা ঘটার probability বের করা যায় Poisson distribution দিয়ে,

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

#

Average = 10 oil tanker




Maximum capacity = 15 tankers.

What is the probability on a given day, A tanker has to be turned away?

$\therefore P(x>15) \to ?$

$$= 1 - \sum_{0}^{15} \frac{\lambda^{x} e^{-\lambda}}{x!} = 1 - \sum_{0}^{15} \frac{10^{x} e^{-10}}{x!}$$

Probability of possibility of 3 - 7 tanks to be available.

$$= \sum_{x=0}^{7} \frac{10^{x} e^{-10}}{x!} - \sum_{x=0}^{2} \frac{10^{x} e^{-10}}{x!}$$

$$\text{or} = \sum_{x=3}^{7} \frac{10^{x} e^{-10}}{x!}$$

# probability of accedent to occur on a day (average) = 0.005

In the next 400 days exactly once the accedent will occur → what is the probability of it?

$\lambda = 400 \times 0.005 = 2$

$\therefore \frac{2^{1} e^{-2}}{1!} = \frac{2}{e^{2}}$ (Answer)

# Relation between binomial & poisson distribution:

trial number of binomial distribution, $n \to \infty$

probability of success, $p \to 0$

$np \to$ constant

$${}^nC_x\, p^x q^{n-x} \Rightarrow \frac{\lambda^x e^{-\lambda}}{x!}$$ ← derivation

19.07.2023

Wednesday

$${}^nC_x\, p^x (1-p)^{n-x}$$

$$= \frac{n!}{x!\,(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{\lambda^x}{x!} \left(\frac{n!}{(n-x)!\,n^x}\right) \left(1-\frac{\lambda}{n}\right)^{n-x}$$

expanding:

$$1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x+1}{n}\right)$$

$$e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n$$

$$= \frac{\lambda^x}{x!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x+1}{n}\right)\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{\lambda^x}{x!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x+1}{n}\right)\left(1-\frac{\lambda}{n}\right)^{n}\left(1-\frac{\lambda}{n}\right)^{-x}$$

$$\lim_{n\to\infty} {}^nC_x\, p^x (1-p)^{n-x}$$

$$= \frac{\lambda^x}{x!} \lim_{n\to\infty}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x+1}{n}\right) \times \lim_{n\to\infty}\left(1+\frac{-\lambda}{n}\right)^n \lim_{n\to\infty}\left(1+\frac{-\lambda}{n}\right)^{-x}$$

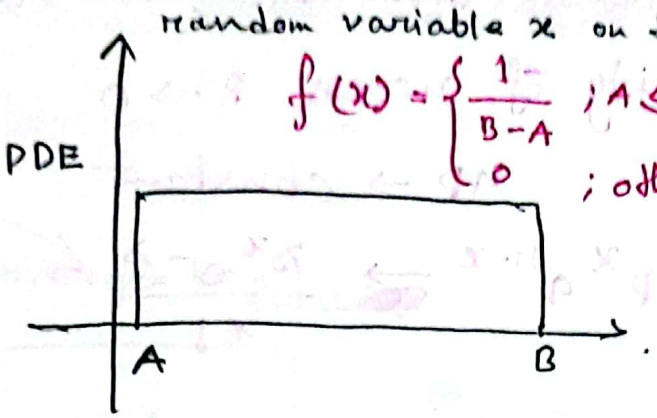
$$= \frac{\lambda^x}{x!} \cdot 1 \cdot e^{-\lambda} \cdot 1 = \frac{\lambda^x e^{-\lambda}}{x!}$$

Probability & statistics WALPOLE → distribution part.

# Continuous Probability Distribution:

**Uniform Distribution:** the continuous function of the continuous random variable $x$ on the interval $[A, B]$ is,

$$f(x) = \begin{cases} \frac{1}{B-A} & ; A \leq x \leq B \\ 0 & ; \text{otherwise} \end{cases}$$



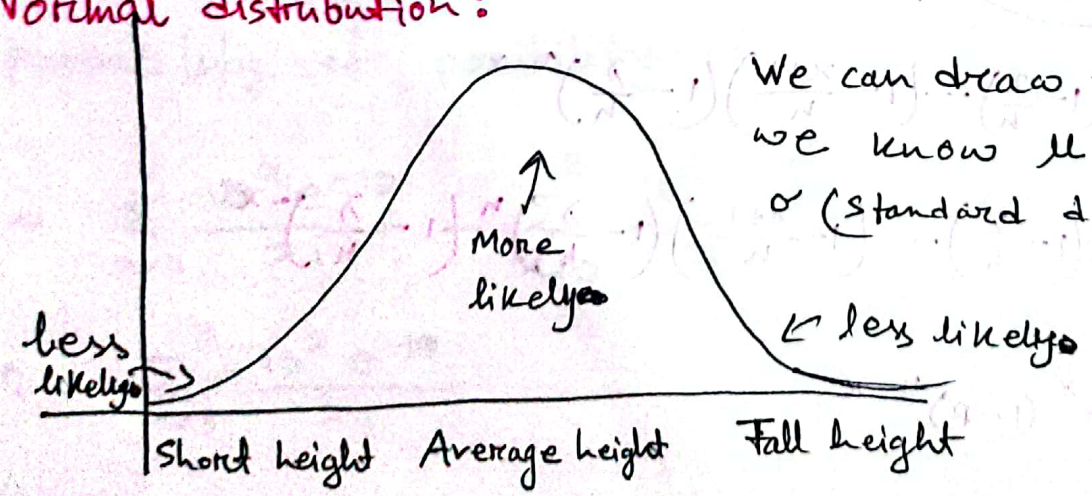

# a) $f(x) = \begin{cases} 1/4 & ; 0 \leq x \leq 4, \\ 0 & ; \text{elsewhere} \end{cases}$

b) Probability that any given conference lasts at least 3 hours:

$$P[x \geq 3] = \int_3^4 1/4 \, dx = 1/4$$

Real life uniform distribution is rare.

## Normal distribution:

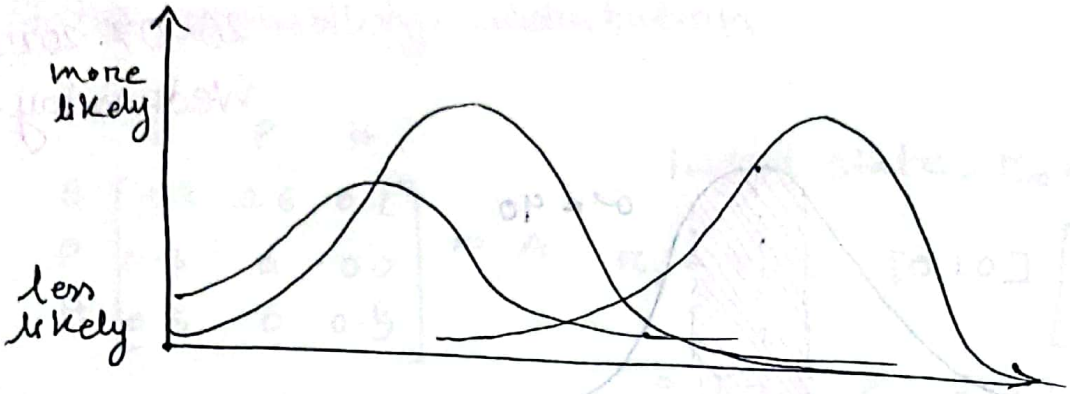



We can draw the curve if we know $\mu$ (mean) and $\sigma$ (standard deviation).

Maximum real life distribution is normal distribution due to "Central Limit Theorem"

Normal distribution function / Gaussian function / Bell shape curve:

$$n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad -\infty < x < \infty$$




Standardized normal distribution: $\mu = 0 \quad \sigma = 1$

convert using z-score: $= \frac{x-\mu}{\sigma}$

Theorem ~~Example~~ 6.2:

$$E(x) = E(x-\mu+\mu)$$
$$= E(x-\mu) + \mu$$
$$E(x-\mu) = \int_{-\infty}^{\infty} \frac{x-\mu}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$z = \frac{x-\mu}{\sigma}$ ; $dx = \sigma\, dz \longrightarrow E(x-\mu) = 0$

$\therefore E(x) = \mu$

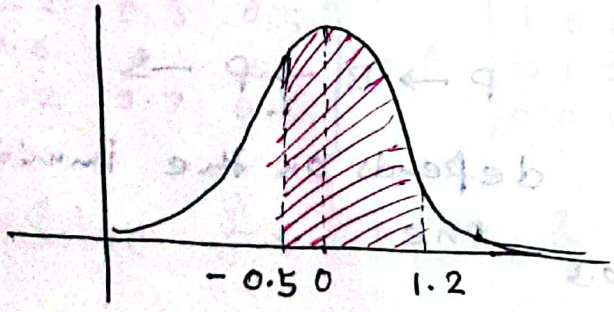
#

$\mu = 50$

$\sigma = 10$

$$z_1 = \frac{45-50}{10} = -0.5$$

$$z_2 = \frac{62-50}{10} = 1.2$$




$$P(45 < x < 62) = P(-0.5 < z < 1.2) = P(z < 1.2) - P(z < -0.5)$$

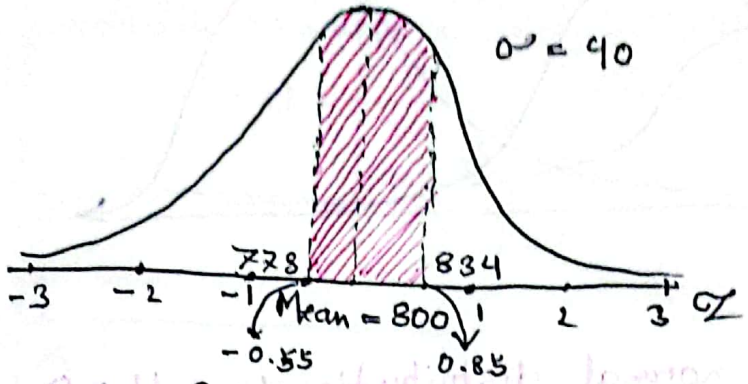
$$= 0.8849 - 0.3085 = 0.5764$$

we can find these values from z-table

Using calculator:

26.07.2023
Wednesday

Example:




$x_1 = 778 \quad x_2 = 834$

$$Z_1 = \frac{778 - 800}{40} \qquad Z_2 = \frac{834 - 800}{40}$$

$$= -0.55 \qquad = 0.85$$

$P(0.85) - P(-0.55)$ = Pink area

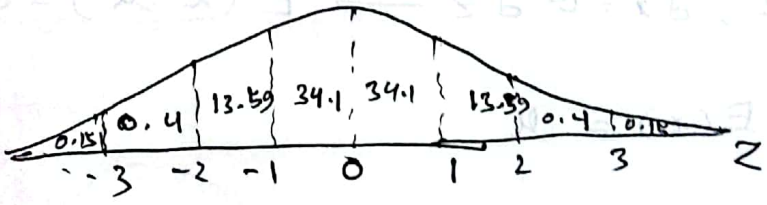
Example: Calculate z from given area (Probability):

$\mu = 40$

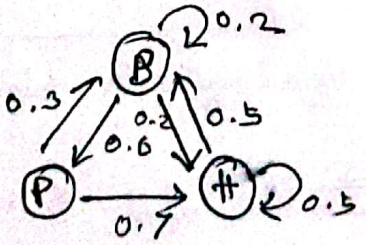
$\sigma = 6$

a) area 45% to the left

b) area 15% to the right




## MARCOV CHAIN




P → B → P →

depends on the immidiate previous one.

After 10 step

P(B) = 4/10 P(P) = 2/10 P(H) = 4/10

B → P → B → H → B → H → H → B → H → P

After ∞ steps → P(B) = 0.351911 P(P) = 0.231345 P(H) = 0.435644 → goes static

# Static probability distribution

$$A = \begin{array}{c} \\ B \\ P \\ H \end{array} \begin{array}{c} \begin{array}{ccc} B & P & H \end{array} \\ \begin{bmatrix} 0.2 & 0.6 & 0.2 \\ 0.3 & 0 & 0.7 \\ 0.5 & 0 & 0.5 \end{bmatrix} \end{array}$$

$$\begin{array}{ccc} B & P & H \end{array}$$

initial state, $\pi_0 = [0 \;\; 1 \;\; 0]$

$$\pi_1 = \pi_0 A = [0 \;\; 1 \;\; 0] \begin{bmatrix} 0.2 & 0.6 & 0.2 \\ 0.3 & 0 & 0.7 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

$$= [0.3 \;\; 0 \;\; 0.7]$$

$$\pi_2 = \pi_1 A = [0.3 \;\; 0 \;\; 0.7] \begin{bmatrix} 0.2 & 0.6 & 0.2 \\ 0.3 & 0 & 0.7 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

when static or stationary $\longrightarrow$ $\boxed{\pi_{\text{[redacted]}} = \pi_{\text{[redacted]}} A}$ $\rightarrow$ find using linear algebra as it is not possible to pursue brute-force way in a short time.

Eigen value, Eigen vector : $\boxed{Av = \lambda v}$

$$A\pi = 1 \cdot \pi$$

$\vec{V} \leftarrow \pi = [\pi_1 \;\; \pi_2 \;\; \pi_3]$

$\pi_1 + \pi_2 + \pi_3 = 1$

$\det(A - \lambda I)$

$(A - \lambda I) \cdot \vec{V} = 0 \Rightarrow (A^T - \lambda I)\vec{V}^T = 0$

Verification

$$\Rightarrow \left( \begin{bmatrix} 0.2 & 0.3 & 0.5 \\ 0.6 & 0 & 0 \\ 0.2 & 0.7 & 0.5 \end{bmatrix} - 1 \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right) \begin{bmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$\Rightarrow$ Solve for $\pi_1, \pi_2$ & $\pi_3$