# Probability and Statistics

1. Discreptive Statistics
2. Correlation and Regression
3. Probability Distributions
4. Sampling Theory
5. Tests of Hypothesis

Data Science

Statistics

Population

Sample

collection of data
- primary data
- secondary data

| C.I | f | c.f |
|---|---|---|
| 20-30 | 1 | 1 |
| 30-40 | 3 | 4 |
| 40-50 | 2 | 6 |
| 50-60 | 3 | 9 |
| 60-70 | 5 | 14 |
| 70-80 | 4 | 18 |
| 80-90 | 2 | 20 |
| 90-100 | 1 | 21 |

Line diagram

Bar diagram

Rectangular diagram

Histogram

Pie diagram

# Measures of Central Tendency

(i) Mean / Arithmetic Mean / Agrigrade

(ii) Median

(iii) Mode

(iv) Geometric Mean

(v) Harmonic Mean

## ① Arithmetic Mean

(i) **Individual Series** : If $x_1, x_2, x_3, \ldots x_n$ are n observations, then

$$A.M = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Eg: 44 65 70 68 64 85

$$A.M = \frac{44+65+70+68+64+85}{6} = 66$$

## ② Discrete Frequency Distribution

If $x_1, x_2, x_3, \ldots x_n$ are n observations with corresponding frequencies $f_1, f_2, f_3, \ldots f_n$

| $x_i$ | $f_i$ | $x_i f_i$ |
|---|---|---|
| $x_1$ | $f_1$ | $f_1 x_1$ |
| $x_2$ | $f_2$ | $f_2 x_2$ |
| $x_3$ | $f_3$ | $f_3 x_3$ |
| ⋮ | ⋮ | ⋮ |
| $x_n$ | $f_n$ | $f_n x_n$ |
| | $N$ | $\sum f_i x_i$ |

$$A.M = \frac{1}{N}\sum f_i x_i$$

where $n = \sum f_i$

Eg:

| $x_i$ | $f_i$ | $x_i f_i$ |
|---|---|---|
| 1 | 5 | 5 |
| 2 | 9 | 18 |
| 3 | 12 | 36 |
| 4 | 17 | 68 |
| 5 | 14 | 70 |
| 6 | 10 | 60 |
| 7 | 6 | 42 |
| | 73 | 299 |

$$A.M = \frac{1}{N}\sum f_i x_i$$

$$= \frac{1}{73} \times 299 = 4.09$$

| C.I | $f_i$ | Mid points $x_i$ | $f_i x_i$ |
|---|---|---|---|
| $l_0-l_1$ | $f_1$ | $x_1$ | $f_1x_1$ |
| $l_1-l_2$ | $f_2$ | $x_2$ | $f_2x_2$ |
| $l_2-l_3$ | $f_3$ | $x_3$ | $f_3x_3$ |
| ⋮ | ⋮ | ⋮ | ⋮ |
| $l_{n-1}-l_n$ | $f_n$ | $x_n$ | $f_nx_n$ |
| | $\Sigma f_i$ | | $\Sigma f_i x_i$ |

$$A.M = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

→ Find the mean of the following frequency distribution

| Marks | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 |
|---|---|---|---|---|---|---|
| No. of Students | 12 | 18 | 27 | 20 | 17 | 6 |

| C.I | $f_i$ | $x_i$ | $f_i x_i$ |
|---|---|---|---|
| 0-10 | 12 | 5 | 60 |
| 10-20 | 18 | 15 | 270 |
| 20-30 | 27 | 25 | 675 |
| 30-40 | 20 | 35 | 700 |
| 40-50 | 17 | 45 | 765 |
| 50-60 | 6 | 55 | 330 |

$\Sigma f_i = 100$ $\quad$ $\Sigma f_i x_i = 2800$

$$A.M = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{2800}{100} = 28$$

## Deviation Method

case-(i): $d_i = x_i - A$ where 'A' is Assumed Mean

$$A.M = A + \frac{\Sigma f_i d_i}{\Sigma f_i}$$

case-(ii): $d_i = \frac{x_i - A}{h}$

$$A.M = A + \frac{\Sigma f_i d_i}{N} \times h$$

where $N = \Sigma f_i$

$h$ = length of the class

→ Find the A.M of the following distribution by using Deviation Methods

| Marks | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 |
|---|---|---|---|---|---|---|
| No. of students | 12 | 18 | 27 | 20 | 17 | 6 |

| C.I | $f_i$ | $x_i$ | $d_i = \frac{x_i - A}{h}$ | $f_i d_i$ |
|---|---|---|---|---|
| 0-10 | 12 | 5 | -3 | -36 |
| 10-20 | 18 | 15 | -2 | -36 |
| 20-30 | 27 | 25 | -1 | -27 |
| 30-40 | 20 | 35 | 0 | 0 |
| 40-50 | 17 | 45 | 1 | 17 |
| 50-60 | 6 | 55 | 2 | 12 |

$\Sigma f_i = N = 100$ $\Sigma f_i d_i = -70$

$$A.M = A + \frac{\Sigma f_i d_i}{N} \times h$$

$$A.M = 35 + \frac{-70}{100} \times 10$$

$$A.M = 28$$

## Median

(i) Individual Series

'n' no. of observations

If n is odd,

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{term}$$

If n is even,

$$\text{Median} = \text{Avg of } \left(\frac{n}{2}\right)^{th} \text{term and } \left(\frac{n}{2}+1\right)^{th} \text{term}$$

① Find the median of the following data

4 7 12 15 13 5 4

Ascending Order : 4 4 5 7 12 13 15

$n = 7$ (no. of observations)

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{term}$$

$$= \frac{7+1^{th}}{2} \text{ term}$$

$$= 4^{th} \text{ term}$$

$$= 15$$

② Find the median of the following data

4 6 12 15 20 24 2 10 35 40

Ascending order : 2 4 6 10 12 15 20 24 35 40

$n = 10$ (no. of observations)

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$$

$$= \frac{5^{th} + 6^{th}}{2}$$

$$= \frac{12+15}{2}$$

$$= \frac{27}{2}$$

$$= 13.5$$

(ii) **Discrete Frequency Distribution**

| x | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| f | 3 | 5 | 12 | 45 | 61 | 22 | 29 | 51 | 23 | 3 |

| x | f | lcf |
|---|---|---|
| 0 | 3 | 3 |
| 1 | 5 | 8 |
| 2 | 12 | 20 |
| 3 | 45 | 65 |
| 4 | 61 | 126 ← |
| 5 | 22 | 148 |
| 6 | 29 | 177 |
| 7 | 51 | 228 |
| 8 | 23 | 251 |
| 9 | 3 | 254 |

$N = 254$

$\frac{N}{2} = \frac{254}{2}$

$= 127$

Median = 5

(iii) Continuous Frequency Distribution

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times h$$

where $l$ = lower limit of the median class

$f$ = frequency of the median class

$N = \Sigma f$

$m$ = cumulative frequency of the class preceeding the median class

$h$ = length of the class interval

→ Find the median of the following frequency distribution.

| C.I | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 |
|---|---|---|---|---|---|---|
| f | 12 | 18 | 27 | 20 | 17 | 6 |

| C.I | f | cf |
|---|---|---|
| 0-10 | 12 | 12 |
| 10-20 | 18 | 30 |
| 20-30 | 27 | 57 |
| 30-40 | 20 | 77 |
| 40-50 | 17 | 94 |
| 50-60 | 6 | 100 |

$N = 100$

$\frac{N}{2} = 50$

Here, $l = 20$

$f = 27$

$m = 30$

$= 10$

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times h$$

$$= 20 + \frac{50-30}{27} \times 10$$

$$= 20 + \frac{200}{27}$$

$$= 27.407$$

**Mode:** Most frequently occuring value in a given series

(i) **Individual Series**

3, 4, 3, 5, 3, 7, 7, 7, 9, 5, 3, 10, 3

**Mode = 3**

(ii) **Discrete ~~Continuous~~ Frequency Distribution**

| x | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| f | 5 | 15 | 25 | 15 | 20 | 6 |

**Mode = 3**

(iii) **Continuous Frequency Distribution**

$$\text{Mode} = l + \frac{f_1 - f_0}{(f_1 - f_0) - (f_2 - f_1)} \times h$$

$$\text{Mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

where $l$ = lower limit of the modal class

$f_1$ = frequency of the modal class

$f_0$ = frequency of the class preceeding the modal class

$f_2$ = frequency of the class suceeding the modal class

$h$ = length of the class interval

→ Find the mode of the following frequency distribution.

| C.I | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 |
|---|---|---|---|---|---|---|---|
| f | 5 | 6 | 8 | 12 | 15 | 5 | 3 |

Here $l = 40$

$f_1 = 15$

$f_0 = 12$

$f_2 = 5$

$h = 10$

$$\text{Mode} = l + \frac{f_1 - f_0}{(f_1 - f_0) - (f_2 - f_1)} \times h$$

$$= 40 + \frac{3}{3+10} \times 10$$

$$= 40 + \frac{30}{13}$$

$$= 42.307$$

# Geometric Mean

## (i) Individual Series

If $x_1, x_2, x_3, \ldots x_n$ are 'n' observations, then

$$G.M = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n}$$

$$\log G.M = \log(x_1 \cdot x_2 \cdots x_n)^{1/n}$$

$$= \frac{1}{n} \log(x_1 \cdot x_2 \cdots x_n)$$

$$= (\log x_1 +$$

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

$$G.M = \text{ntilog}\left(\frac{1}{n} \sum \log x_i\right)$$

For frequency distribution,

$$G.M = \text{Antilog}\left(\frac{1}{N} \sum_{i=1}^{N} f_i \log x_i\right)$$

→

**Harmonic Mean:** It is defined as "the reciprocal of avg of reciprocals of the given set of observations"

(i) **Individual Series:**

If $x_1, x_2, x_3, \ldots x_n$ are 'n' observations, then their reciprocals are

$$\frac{1}{x_1}, \frac{1}{x_2}, \ldots \frac{1}{x_n}$$

$$A.M = \frac{\frac{1}{x_1} + \frac{1}{x_2} + \ldots \frac{1}{x_n}}{n} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}$$

$$H.M = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}}$$

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

For a frequency distribution,

$$H.M = \frac{N}{\sum_{i=1}^{n} f_i\left(\frac{1}{x_i}\right)} \quad \left(\text{where } N = \sum_{i=1}^{n} f_i\right)$$

→ Find the H.M of the set of observations: 2, 4, 8, 12, 16, 24

| $x_i$ | $\frac{1}{x_i}$ |
|---|---|
| 2 | 0.5 |
| 4 | 0.25 |
| 8 | 0.125 |
| 12 | 0.083 |
| 16 | 0.062 |
| 24 | 0.042 |
| | 1.062 |

$n = 6$

$$H.M = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)}$$

$$= \frac{6}{1.062}$$

$$= 5.649$$

→ Find G.M and H.M of the following frequency distribution.

| C.I | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 |
|---|---|---|---|---|---|---|---|
| f | 4 | 7 | 13 | 21 | 15 | 25 | 20 |

| C.I | $f_i$ | $x_i$ | $\log x_i$ | $f_i \log x_i$ | $\frac{1}{x_i}$ | $f_i\left(\frac{1}{x_i}\right)$ |
|---|---|---|---|---|---|---|
| 0-10 | 4 | 5 | 0.6989 | 2.7956 | 0.200 | 0.800 |
| 10-20 | 7 | 15 | 1.1760 | 8.2320 | 0.066 | 0.462 |
| 20-30 | 13 | 25 | 1.3979 | 18.17 | 0.040 | 0.520 |
| 30-40 | 21 | 35 | 1.5440 | 32.4240 | 0.028 | 0.588 |
| 40-50 | 15 | 45 | 1.6532 | 24.7980 | 0.022 | 0.330 |
| 50-60 | 25 | 55 | 1.7403 | 43.5075 | 0.018 | 0.450 |
| 60-70 | 20 | 65 | 1.8129 | 36.2580 | 0.015 | 0.300 |
| | 105 | | 10.0232 | 166.1878 | | 3.450 |

Here $N = \Sigma f_i = 105$

$$G.M = \text{Antilog}\left(\frac{1}{N}\sum_{i=1}^{n} f_i \log x_i\right)$$

$$= \text{Antilog}\left(\frac{1}{105} \times 166.1878\right)$$

$$= \text{Antilog}(1.5827)$$

$$= 38.2596$$

$$H.M = \frac{N}{\sum_{i=1}^{n} f_i\left(\frac{1}{x_i}\right)}$$

$$= \frac{105}{3.450}$$

$$= 30.4347$$

# Measures of Dispersion :

(i) Range

(ii) Quartile Deviation

(iii) Mean Deviation

(iv) Standard Deviation

## Range :

$$\text{Range} = \text{Max value} - \text{Min value}$$

Let $X \rightarrow x_1, x_2, x_3 \cdots x_n$

$$\text{Range} = X_{max} - X_{min}$$

$$\text{Coeff of Range} = \frac{X_{max} - X_{min}}{X_{max} + X_{min}}$$

→ Find the range of the given set of observations : 48, 82, 64, 72, 66, 68

$$\text{Range} = X_{max} - X_{min}$$

$$= 82 - 48$$

$$= 34$$

$$\text{Coeff of Range} = \frac{82-48}{82+48} = \frac{34}{130}$$

$$= 0.2615$$

## Quartile Deviation (Q.D)

$$Q.D = \frac{Q_3 - Q_1}{2}$$

$Q_1 \rightarrow$ First Quartile

$Q_3 \rightarrow$ Third Quartile

where, $Q_1 = l + \frac{\frac{N}{4} - m}{f} \times c$

$l$ = lower limit of the first quartile class

$f$ = frequency of the first quartile class

$m$ = cf of the class preceeding the first quartile class

$c$ = length of the class interval.

$$Q_3 = l + \frac{\frac{3N}{4} - m}{f} \times c$$

$l$ = lower limit of the third quartile class

$f$ = frequency of the third quartile class

$m$ = cf of the class preceeding the third quartile class

$c$ = length of the class interval

→ Find the quartile deviation of the following frequency distribution.

| C.I | 0-15 | 15-30 | 30-45 | 45-60 | 60-75 | 75-90 | 90-105 |
|---|---|---|---|---|---|---|---|
| f | 8 | 26 | 30 | 45 | 20 | 17 | 4 |

| C.I | f | cf |
|---|---|---|
| 0-15 | 8 | 8 |
| 15-30 | 26 | 34 |
| 30-45 | 30 | 64 |
| 45-60 | 45 | 109 |
| 60-75 | 20 | 129 |
| 75-90 | 17 | 146 |
| 90-105 | 4 | 150 |

$N = 150$

$\frac{N}{4} = \frac{150}{4} = 37.5$

$l_1 = 30,\ f_1 = 30,\ m_1 = 34,\ c = 15$

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c$$

$$= 30 + \frac{37.5 - 34}{30} \times 15$$

$$= 31.75$$

$N = 150$

$\frac{3N}{4} = \frac{450}{4} = 112.5$

$l_2 = 60,\ f_2 = 20,\ m_2 = 109,\ c = 15$

$$Q_3 = l_2 + \frac{\frac{3N}{2} - m_2}{f_2} \times c$$

$$= 60 + \frac{112.5 - 109}{20} \times 15$$

$$= 62.625$$

$$Q.D = \frac{Q_3 - Q_1}{2}$$

$$= \frac{62.625 - 31.75}{2}$$

$$= 15.4375$$

# Mean Deviation

(i) Individual Series

If $x_1, x_2, x_3 \cdots x_n$ are 'n' observations, then

$$M.D = \frac{1}{n}\sum_{i=1}^{n}|(x_i - A)|$$

where 'A' is any aggregate

For frequency distribution,

$$M.D = \frac{1}{n}\sum_{i=1}^{n} f_i|x_i - A|$$

→ Find the mean deviation ~~from~~ about the mean to the following data.

| Marks less than | 80 | 70 | 60 | 50 | 40 | 30 | 20 | 10 |
|---|---|---|---|---|---|---|---|---|
| No. of students | 100 | 80 | 70 | 60 | 32 | 20 | 13 | 5 |

| C.I | $f_i$ | $x_i$ | $f_i x_i$ | $\lvert x_i - A\rvert$ (A = 47) | $f_i\lvert x_i - A\rvert$ |
|---|---|---|---|---|---|
| 0-10 | 5 | 5 | 25 | 42 | |
| 10-20 | 8 | 15 | 120 | 32 | |
| 20-30 | 7 | 25 | 175 | 22 | |
| 30-40 | 12 | 35 | 420 | 12 | |
| 40-50 | 28 | 45 | 1260 | 2 | |
| 50-60 | 10 | 55 | 550 | 8 | |
| 60-70 | 10 | 65 | 650 | 18 | |
| 70-80 | 20 | 75 | 1500 | 28 | |
| | 100 | | 4700 | | |

$\Sigma f_i = 100$

$\Sigma f_i x_i = 4700$

Mean $= \frac{1}{N}$

$\Sigma f_i = 100$, $\Sigma f_i x_i = 4700$

$$\text{Mean} = \frac{1}{N}\Sigma f_i x_i = \frac{4700}{100} = 47$$

$\Sigma f_i|(x_i - A)| = 1640$

$$M.D = \frac{1}{n}\sum_{i=1}^{n} f_i|x_i - A|$$

$$= \frac{1640}{100}$$

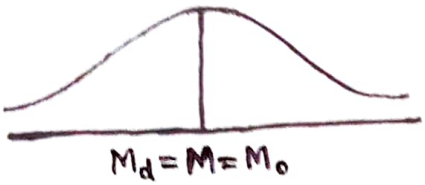
$$= 16.4$$

## Skewness

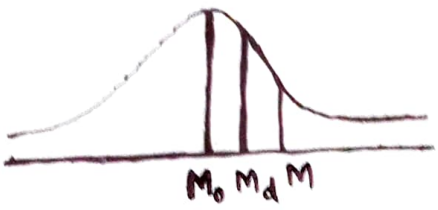
Lack of Symmetry

Unsymmetric curves

Normal curve

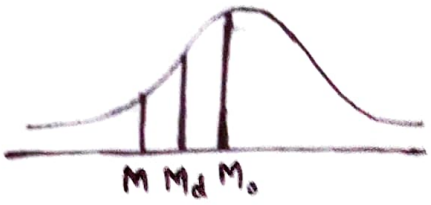



Symmetric curve




Skewed to the right

(Positively skewed)




Skewed to the left

(Negatively skewed)

## Measures

$S_K$ = Mean − Median

= Mean − Mode

$S_K = (Q_3 - M_d) - (M_d - Q_1)$

$= Q_3 + Q_1 - 2M_d$

Absolute value of skewness

Mode = 3 Median − 2 Mean

$M_o = 3M_d - 2M$

Karl Pearson's

Coeff of skewness = $\frac{M - M_o}{S.D}$

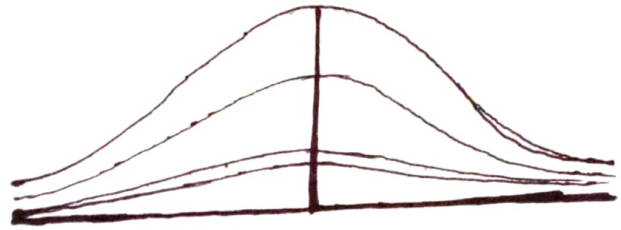
$= \frac{M - (3M_d - 2M)}{\sigma} = \frac{3(M - M_d)}{\sigma}$

Bowley's coeff of skewness

$$= \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)}$$

## Kurtosis

Convexity of the curve



Measures the flatness of the frequency distribution

A ⟶ Leptokurtic (sharp)

B ⟶ Mesokurtic (Normal)

C ⟶ platykurtic (flat)

Coeff of kurtosis whose measures are represented by $\beta_2$ and $\gamma_2$

$$\beta_2 = \frac{\mu_4}{\mu_2} = \frac{\mu_4}{\sigma^4} \qquad \gamma_2 = \beta_2 - 3$$

where $\mu_r = r^{th}$ moment about mean

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

# Correlation and Regression

## Correlation

Positive Correlation

Negative Correlation

No Correlation

Simple Correlation

Multiple Correlation

Partial Correlation

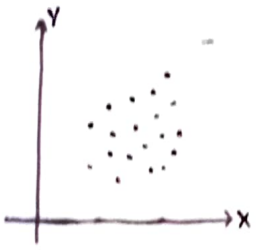
Linear Correlation

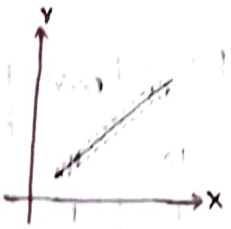
Curvi-linear Correlation

| | | | | | |
|---|---|---|---|---|---|
| x : | 3 | 5 | 7 | 9 | 11 |
| y : | 7 | 11 | 15 | 19 | 23 |

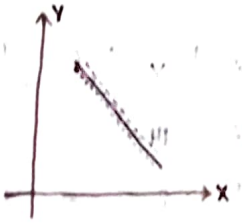
Scatter diagram

$\frac{x}{y} = \frac{2}{4} = \frac{2}{4} = \frac{2}{4} = \frac{2}{4}$




Positive Correlation




Negative Correlation

## Karl Pearson's Coeff of Correlation

$r(X,Y),\ r_{x,y}$

$$r = \frac{\text{cov}(x,y)}{\sigma_x \cdot \sigma_y}$$ ← co-variance

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

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

## Properties

$r = r(X, Y)$

(i) $-1 \le r \le 1$

If $0 < r < 1$, then X, Y are positively correlated

If $-1 < r < 0$, then X, Y are negatively correlated

If $r = 0$, then there is no correlation

If $r = 1$, then $\exists$ perfect positive correlation b/w X and Y

If $r = -1$, then $\exists$ perfect negative correlation b/w X and Y

(ii) Correlation coeff is independent of change of scale and origin

→ Find the correlation coeff b/w two variables X and Y follows

| X | 12 | 9 | 8 | 10 | 11 | 13 | 7 |
|---|---|---|---|---|---|---|---|
| Y | 14 | 8 | 6 | 9 | 11 | 12 | 3 |

| X | Y | $X-\bar{X}$ $=X-10$ | $Y-\bar{Y}$ $=Y-9$ | $(X-\bar{X})$ $(Y-\bar{Y})$ | $(X-\bar{X})^2$ | $(Y-\bar{Y})^2$ |
|---|---|---|---|---|---|---|
| 12 | 14 | 2 | 5 | 10 | 4 | 25 |
| 9 | 8 | -1 | -1 | 1 | 1 | 1 |
| 8 | 6 | -2 | -3 | 6 | 4 | 9 |
| 10 | 9 | 0 | 0 | 0 | 0 | 0 |
| 11 | 11 | 1 | 2 | 2 | 1 | 4 |
| 13 | 12 | 3 | 3 | 9 | 9 | 9 |
| 7 | 3 | -3 | -6 | 18 | 9 | 36 |
| 70 | 63 | | | 46 | 28 | 84 |

$\Sigma X = 70$, $\Sigma Y = 63$, $n = 7$

$\bar{X} = \frac{1}{n}\Sigma X = \frac{70}{7} = 10$

$\bar{Y} = \frac{1}{n}\Sigma Y = \frac{63}{7} = 9$

$\Sigma (X-\bar{X})(Y-\bar{Y}) = 46$

$\Sigma (X-\bar{X})^2 = 28$

$\Sigma (Y-\bar{Y})^2 = 84$

Coeff of correlation

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

$$r = \frac{\sum_i (X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma (X-\bar{X})^2}\sqrt{\Sigma (Y-\bar{Y})^2}}$$

$$r = \frac{46}{\sqrt{28}\sqrt{84}}$$

$$r = \frac{46}{48.498} = 0.948 \cong 0.95$$

∴ The two variables X and Y are positively correlated

→ Find if there is any significant correlation b/w the heights & weights given below.

| Heights (in inches) | 57 | 59 | 62 | 63 | 64 | 65 | 55 | 58 | 57 |
|---|---|---|---|---|---|---|---|---|---|
| Weights (in pounds) | 113 | 117 | 126 | 126 | 130 | 129 | 111 | 116 | 112 |

| X | Y | $(x-\bar{x})$ | $(y-\bar{y})$ | $(x-\bar{x})(y-\bar{y})$ | $(x-\bar{x})^2$ | $(y-\bar{y})^2$ |
|---|---|---|---|---|---|---|
| | | $=x-60$ | $=Y-120$ | | | |
| 57 | 113 | -3 | -7 | 21 | 9 | 49 |
| 59 | 117 | -1 | -3 | 3 | 1 | 9 |
| 62 | 126 | 2 | 6 | 12 | 4 | 36 |
| 63 | 126 | 3 | 6 | | 9 | 36 |
| 64 | 130 | 4 | 10 | 4 | 16 | 100 |
| 65 | 129 | 5 | 9 | 45 | 25 | 81 |
| 55 | 111 | -5 | -9 | | 25 | 81 |
| 58 | 116 | -2 | -4 | 8 | 4 | 16 |
| 57 | 112 | -3 | -8 | 24 | 9 | 64 |
| 540 | 1080 | | | 216 | 102 | 472 |

$\Sigma x = 540$, $\Sigma Y = 1080$, $n = 9$

$\bar{x} = \frac{1}{n}\Sigma x = \frac{540}{9} = 60$

$\bar{y} = \frac{1}{n}\Sigma Y = \frac{1080}{9} = 120$

$\Sigma(x-\bar{x})(y-\bar{y}) = 216$

$\Sigma(x-\bar{x})^2 = 102$

$\Sigma(y-\bar{y})^2 = 472$

Coeff of correlation

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

$$r = \frac{216}{\sqrt{102}\sqrt{472}}$$

$r = 0.9\ 5$

→ When deviations are taken from Assumed Mean

$$r = \frac{\quad}{\sqrt{\quad}}$$

→ Find the correlation coeff for the following paired data

| X | 28 | 41 | 40 | 38 | 35 | 33 | 40 | 32 | 36 | 33 |
|---|---|---|---|---|---|---|---|---|---|---|
| Y | 23 | 34 | 33 | 31 | 30 | 26 | 28 | 31 | 36 | 38 |

| X | Y | $x = X - \bar{X}$ | $y = Y - \bar{Y}$ | xy | $x^2$ | $y^2$ |
|---|---|---|---|---|---|---|
| 28 | 23 | -7 | -8 | 56 | 49 | 64 |
| 41 | 34 | 6 | 3 | 18 | 36 | 9 |
| 40 | 33 | 5 | 2 | 10 | 25 | 4 |
| 38 | 34 | 3 | 3 | 9 | 9 | 9 |
| 35 | 30 | 0 | -1 | 0 | 0 | 1 |
| 33 | 26 | -2 | -5 | 10 | 4 | 25 |
| 40 | 28 | 5 | -3 | 15 | 25 | 9 |
| 32 | 31 | -3 | 0 | -0 | 9 | 0 |
| 36 | 36 | 1 | 5 | 5 | 1 | 25 |
| 33 | 38 | -2 | 7 | -14 | 4 | 49 |
| 356 | 313 | 6 | 3 | 79 | 162 | 195 |

$\Sigma X = 356$, $\Sigma Y = 313$, $n = 10$

$\bar{X} = \frac{1}{n}\Sigma X = \frac{356}{10} = 35.6$

$\bar{Y} = \frac{1}{n}\Sigma Y = \frac{313}{10} = 31.3$

$\Sigma x = 6$, $\Sigma y = 3$, $\Sigma xy =$

$\Sigma x^2 = 162$, $\Sigma y^2 = 195$

Coeff of correlation

$$r = \frac{\Sigma xy - \frac{1}{n}\Sigma x \Sigma y}{\sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}\sqrt{\Sigma y^2 - \frac{1}{n}(\Sigma y)^2}}$$

$$r = \frac{79 - \frac{1}{10}(6)(3)}{\sqrt{162 - \frac{1}{10}(\ )}\sqrt{195 - \frac{1}{10}(9)}}$$

$$r = \frac{79 - 1.8}{\sqrt{162 - 3.6}\sqrt{195 - 0.9}}$$

$$r = \frac{77.2}{\sqrt{158.4}\sqrt{194.31}}$$

$$r = \frac{77.2}{(12.58)(13.93)}$$

$$r = \frac{77.2}{175.23}$$

$$r = 0.44$$

# Spearman's Rank Correlation Coefficient

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} \quad (-1 \le \rho \le 1)$$

where $d_i = x_i - y_i$

$x_i$ denotes ranks of x

$y_i$ denotes ranks of y

For repeated ranks, $\rho = 1 - \frac{6\ (\Sigma d_i^2 + C.F)}{n(n^2-1)}$

where C.F = correction factor

$$C.F = \frac{m(m^2-1)}{12}$$ (m = no. of times the obs is repeated)

If any obs of a given series is repeated 'n' times, then

$$C.F = \sum_{i=1}^{n} \frac{m_i(m_i^2-1)}{12}$$

→ random sample of five college students is selected & their marks in Mathematics & Statistics are found to be as follows;

| | | | | | |
|---|---|---|---|---|---|
| Math | 85 | 60 | 73 | 40 | 90 |
| Stat | 3 | 75 | 65 | 50 | 80 |

| x | y | $x_i$ | $y_i$ | $d_i = x_i - y_i$ | $d_i^2$ |
|---|---|---|---|---|---|
| 85 | 93 | 2 | 1 | 1 | 1 |
| 60 | 75 | 4 | 3 | 1 | 1 |
| 73 | 65 | 3 | 4 | -1 | 1 |
| 40 | 50 | 5 | 5 | 0 | 0 |
| 90 | 80 | 1 | 2 | -1 | 1 |
| | | | | | 4 |

Here n = 5

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 4}{5(24)} = 1 - \frac{1}{5} = 1 - 0.2$$

$$= 0.8$$

→ Find the rank correlation coefficient for the following data.

| X | 68 | 64 | 75 | 50 | 64 | 80 | 75 | 4 | 55 | 64 |
|---|---|---|---|---|---|---|---|---|---|---|
| Y | 62 | 58 | 68 | 45 | 81 | 60 | 68 | 48 | 50 | 70 |

| X | Y | $x_i$ | $y_i$ | $d_i = x_i - y_i$ | $d_i^2$ |
|---|---|---|---|---|---|
| 68 | 62 | 4 | 5 | -1 | 1 |
| 64 | 58 | 6 | 7 | -1 | 1 |
| 75 | 68 | 2.5 | 3.5 | -1 | 1 |
| 50 | 45 | 9 | 10 | -1 | 1 |
| 64 | 81 | 6 | 1 | 5 | 25 |
| 80 | 60 | 1 | 6 | -5 | 25 |
| 75 | 68 | 2.5 | 3.5 | -1 | 1 |
| 40 | 48 | 10 | 9 | 1 | 1 |
| 55 | 60 | 8 | 8 | 0 | 0 |
| 64 | 70 | 6 | 2 | 4 | 16 |
| | | | | | 72 |

Here $n = 10$

$$\rho = 1 - \frac{6(\Sigma d_i^2 + C.F)}{n(n^2-1)}$$

$$\rho = \frac{6(72+3)}{10 \times 99}$$

$$\rho = 1 - 0.4545$$

$$\rho = 0.545$$

In X-series: 75 repeated 2 times

64 repeated 3 times

In Y-series: 68 repeated 2 times

$$C.F = \frac{2(2^2-1)}{12} + \frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12}$$

$$C.F = \frac{1}{2} + 2 + \frac{1}{2}$$

$$C.F = 3$$

# Regression :

Linear Regression

## Regression line of Y on X

$$(Y-\bar{Y}) = b_{YX}(X-\bar{X})$$

$b_{YX}$ → regression coeff of Y on X

$$b_{YX} = r\cdot\frac{\sigma_Y}{\sigma_X}$$

$$b_{YX} = \frac{cov(X,Y)}{\sigma_X\cdot\sigma_Y}\cdot\frac{\sigma_Y}{\sigma_X}$$

$$b_{YX} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

## Regression line of X on Y

$$(x-\bar{x}) = b_{XY}(y-\bar{y})$$

$b_{XY}$ → regression coeff of X on Y

$$b_{XY} = r\cdot\frac{\sigma_X}{\sigma_Y}$$

$$b_{XY} = \frac{cov(Y,X)}{\sigma_Y\cdot\sigma_X}\cdot\frac{\sigma_X}{\sigma_Y}$$

$$b_{XY} = \frac{\Sigma(y-\bar{y})(x-\bar{x})}{\Sigma(y-\bar{y})^2}$$

→ Find the equations of the regression lines.

| X | 46 | 42 | 44 | 40 | 43 | 41 | 45 |
|---|---|---|---|---|---|---|---|
| Y | 40 | 38 | 36 | 35 | 39 | 37 | 41 |

| X | Y | $x-\bar{x}$ = x-43 | $y-\bar{y}$ = y-38 | $(x-\bar{x})(y-\bar{y})$ | $(x-\bar{x})^2$ | $(y-\bar{y})^2$ |
|---|---|---|---|---|---|---|
| 46 | 40 | 3 | 2 | 6 | 9 | 4 |
| 42 | 38 | -1 | 0 | 0 | 1 | 0 |
| 44 | 36 | 1 | -2 | -2 | 1 | 4 |
| 40 | 35 | -3 | -3 | 9 | 9 | 9 |
| 43 | 39 | 0 | 1 | 0 | 0 | 1 |
| 41 | 37 | -2 | -1 | 2 | 4 | 1 |
| 45 | 41 | 2 | 3 | 6 | 4 | 9 |
| 301 | 266 | | | 21 | 28 | 28 |

$\Sigma X = 301$, $\Sigma Y = 266$, $n = 7$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{301}{7} = 43$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{266}{7} = 38$$

$\Sigma(x-\bar{x})(y-\bar{y}) = 21$

$\Sigma(x-\bar{x})^2 = 28$

$\Sigma(y-\bar{y})^2 = 28$

Eqn to the regression line of Y on X is

$$(Y-\bar{Y}) = b_{YX}(X-\bar{X})$$

$$b_{YX} = r\cdot\frac{\sigma_Y}{\sigma_X} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{21}{28} = \frac{3}{4} = 0.75$$

Regression line of Y on X is

$$(Y-38) = (0.75)(X-43)$$

$$Y = (0.75)(38-32.5)$$

$$Y = (0.75)(\text{illegible})$$

→ Find the regression equations b/w the price indices of cotton & wool which are given below for the 12 months of a year.

| Cotton(x) | 78 | 77 | 85 | 88 | 87 | 82 | 81 | 77 | 76 | 83 | 97 | 93 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Wool(Y) | 84 | 82 | 82 | 85 | 89 | 90 | 88 | 92 | 83 | 89 | 98 | 99 |

$\Sigma X = 1004, \Sigma Y = 1061, n = 12$

$\Sigma(x-\bar{x})(y-\bar{y}) = 288.66$

$\bar{X} = \frac{1}{n}\Sigma X = \frac{1004}{12} = 83.7$

$\Sigma(x-\bar{x})^2 = 486.68$

$\bar{Y} = \frac{1}{n}\Sigma Y = \frac{1061}{12} = 88.4$

$\Sigma(Y-\bar{Y})^2 = 362.92$

| $X$ | $Y$ | $X-\bar{X}$ $=X-83.7$ | $Y-\bar{Y}$ $=Y-88.4$ | $(x-\bar{x})(y-\bar{y})$ | $(x-\bar{x})^2$ | $(y-\bar{y})^2$ |
|---|---|---|---|---|---|---|
| 78 | 84 | -5.7 | -4.4 | 25.08 | 32.49 | 19.36 |
| 77 | 82 | -6.7 | -6.4 | 42.88 | 44.89 | 40.96 |
| 85 | 82 | 1.3 | -6.4 | -8.32 | 1.69 | 40.96 |
| 88 | 85 | 4.3 | -3.4 | -14.62 | 18.49 | 11.56 |
| 87 | 89 | 3.3 | | 1.98 | 10.89 | 0.36 |
| 82 | 90 | -1.7 | | -2.72 | 2.89 | 2.56 |
| 81 | 88 | -2.7 | -0.4 | 1.08 | 7.29 | 0.16 |
| 77 | 92 | -6.7 | 3.6 | -24.12 | 44.89 | 12.96 |
| 76 | 83 | -7.7 | -5.4 | 41.58 | 59.29 | 29.16 |
| 83 | 89 | -0.7 | 0.6 | -0.42 | .49 | 0.36 |
| 91 | 98 | 18.3 | 9.6 | 127.68 | 176.89 | 92.16 |
| 93 | 99 | 9.3 | 10.6 | 98.58 | 86.49 | 112.36 |
| 1004 | 1061 | | | 288.66 | 486.68 | 362.92 |

Eqn to the regression line of $Y$ on $X$ is

$$(Y-\bar{Y}) = b_{YX}(X-\bar{X})$$

$$b_{YX} = r\cdot\frac{\sigma_Y}{\sigma_X} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{288.66}{486.68} = 0.593$$

Regression line of $Y$ on $X$ is

$$(Y-88.4) = (0.593)(X-83.7)$$

$$Y = (0.593)X + (88.4-49.634)$$

$$Y = (0.593)X + 38.7$$

Eqn to the regression line of $X$ on $Y$ is

$$(X-\bar{X}) = b_{XY}(Y-\bar{Y})$$

$$b_{XY} = r\cdot\frac{\sigma_X}{\sigma_Y} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2} = \frac{288.66}{362.92} = 0.795$$

Regression line of $X$ on $Y$ is

$$(X-83.7) = (0.795)(Y-88.4)$$

$$X = (0.795)Y + (83.7-70.278)$$

$$X = (0.795)Y + 13.422$$

→ Estimate the price index of cotton corresponding to price index of wool when

=94, = ?

$$X = (0.795)Y + 13.422$$

$$X = (0.795)(94) + 13.422$$

$$X = 74.73 + 13.422$$

$$X = 88.152$$

## Curv Fitting

### Method of Least Squares

$$y = f(x)$$

Normal equation

$$a_0x^0 + a_1x^1 + a_2x^2 + \cdots + a_nx^n, \quad a_n \neq 0$$

$$y = a + bx$$

$$y_i = f(x_i) \; ; \; 1 \leq x_i \leq n$$

$$i = y_i - f(x_i)$$

$$\boxed{S = \sum_{i=1}^{n} e_i^2 = \sum \{y_i - f(x_i)\}^2}$$

### Fitting of a Straight line

$$\boxed{y = a + bx} \longrightarrow I$$

Normal eqn are

$$\sum_i y_i = na + b\sum_i x_i$$

$$\sum_i y_i x_i = a\sum_i x_i + b\sum_i x_i^2$$

→ Fit a suitable straight line to the following data.

| x | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| y | 14 | 27 | 40 | 55 | 68 |

| x | y | xy | $x^2$ |
|---|---|---|---|
| 1 | 14 | 14 | 1 |
| 2 | 27 | 54 | 4 |
| 3 | 40 | 120 | 9 |
| 4 | 55 | 220 | 16 |
| 5 | 68 | 340 | 25 |
| 15 | 204 | 748 | 55 |

$n=5,\ \Sigma x_i=15,\ \Sigma y_i=204,\ \Sigma x_i y_i=748,\ \Sigma x_i^2=55$

Normal eqn are

$$204 = 5a + 15b \longrightarrow ①$$

$$748 = 15a + 55b \longrightarrow ②$$

$3 \times ① \Rightarrow 15a + 45b = 612$

$② \Rightarrow 15a + 55b = 748$

$-10b = -136$

$$\boxed{b = 13.6}$$

From eq①, $5a + 15(13.6) = 204$

$5a = 0$

$$\boxed{a = 0}$$

Required eqn of the straight line is

$$\boxed{= (13.6)x}$$

# Fitting of a power curve

$$y = ab^x$$

$$\log y = \log(ab^x)$$

$$\log y = \log a + \log b^x$$

$$\log y = \log a + x \log b$$

$$\boxed{Y = A + Bx}$$

where $A = \log a$

$B = \log b$

$Y = \log y$

Normal equations are

$$\sum_i Y_i = nA + B\sum_i x_i$$

$$\sum_i x_i Y_i = A\sum_i x_i + B\sum_i x_i^2$$

→ Fit a power curve to the following data

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| $y$ | 10 | 21 | 35 | 59 | 92 | 200 | 400 | 600 |

| $x$ | $y$ | $Y = \log_{10} y$ | $xY$ | $x^2$ |
|---|---|---|---|---|
| 0 | 10 | 1.0000 | 0.0000 | 0 |
| 1 | 21 | 1.3222 | 1.3222 | 1 |
| 2 | 35 | 1.5440 | 3.0880 | 4 |
| 3 | 59 | 1.7708 | 5.3124 | 9 |
| 4 | 92 | 1.9638 | 7.8552 | 16 |
| 5 | 200 | 2.3010 | 11.5050 | 25 |
| 6 | 400 | 2.6020 | 15.6120 | 36 |
| 7 | 600 | 2.7853 | 19.4971 | 49 |
| 28 | | 15.2891 | 64.1919 | 140 |

$n = 8$, $\sum_i x_i = 28$, $\sum_i Y_i = 15.2891$, $\sum x_i Y_i = 64.1919$, $\sum_i x_i^2 = 140$

Normal eqn are

$$\sum_i Y_i = nA + B\sum_i x_i$$

$$\sum_i x_i Y_i = A\sum_i x_i + B\sum_i x_i^2$$

$$15.2891 = 8A + 28B \longrightarrow ①$$

$$64.1919 = 28A + 140B$$

$$2.2926 = A + 5B \longrightarrow ②$$

Solving ① & ②

$$① \Longrightarrow 15.2891 = 8A + 28B$$

$$②\times 8 \Longrightarrow 18.3 = 8A + 40B$$

$$+12B = +3.051$$

$$B = \frac{3.0517}{12}$$

$$B = 0.2543$$

From eq ②,

$$A + 5$$

# 3. Random Variables

## Axioms

$0 \leq P(E) \leq 1$

$P(E_1 \cup E_2) = P(E_1) + P(E_2)$

$P(S) = 1$

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

## Addition Rule

$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

where $P(E_1 \cap E_2)$ = Joint probability

## Conditional Probability

$$P(E_1 | E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}; \quad P(E_2) \neq 0$$

$$P(E_2 | E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)}; \quad P(E_1) \neq 0$$

## Independent events

$P(E_1 | E_2) = P(E_1)$

$P(E_2 | E_1) = P(E_2)$

## Multiplication Rule/ Compound theorem on Probability

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2 | E_1)$$
$$= P(E_2) \cdot P(E_1 | E_2)$$

$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$ (for independent events)

→ Find the probability of getting a sum 10 or 11 when two dice are thrown.

$E_1 \rightarrow 10$

| die-I | die-II |
|---|---|
| 6 | 4 |
| 5 | 5 |
| 4 | 6 |

$P(E_1) = \frac{3}{36}$

$E_2 \rightarrow 11$

| die-I | die-II |
|---|---|
| 6 | 5 |
| 5 | 6 |

$P(E_2) = \frac{2}{36}$

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

$$= \frac{3}{36} + \frac{2}{36}$$

$$= \frac{5}{36}$$

→ A bag contains 4 white & 6 black balls. Two balls are drawn ~~as~~ out at random in succession. What is the probability that the two balls drawn out are white,

(i) the first ball drawn is replaced in the b before the second ball is drawn.

(ii) the first ball drawn is not replaced in the bag before the second ball is drawn

$E_1 \rightarrow$ first ball drawn

$E_2 \rightarrow$ second ball drawn

$P(E_1 \cap E_2) = ?$

$P(E_1) = \frac{4}{10}$

(i) with replacement

$P(E_2) = \frac{4}{10}$

$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$

$= \frac{4}{10} \times \frac{4}{10}$

(ii) without replacement

$P(E_2|E_1) = \frac{3}{9}$

$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2|E_1)$

$= \frac{4}{10} \times \frac{3}{9}$

## Discrete Probability Distribution

A discrete random variable X assumes $x_1, x_2, \ldots x_n$ with the corresponding probabilities $P_1, P_2, P_3, \ldots P_n$, then the following tabular form is known as Discrete Probability Distribution.

| $X$ : | $x_1$ | $x_2$ | $x_3$ | --- | $x_n$ |
|---|---|---|---|---|---|
| $P(X=x_i)$ : | $P_1$ | $P_2$ | $P_3$ | --- | $P_n$ |

$$P_i = P(X = x_i) \quad 1 \le i \le n$$

where $P_i$ = probability mass function (pmf)

↓

It must satisfy the following conditions

(i) $P_i \ge 0$,

(ii) $\sum_i P_i = 1$

→ Consider a discrete random variable X assumes the no.of heads turned up when two coins are tossed simultaneously.

$X : S \to R$

$S : \{S_1, S_2, S_3, S_4\}$

$S = \{HH, HT, TH, TT\}$

$X(HH) = 2$

$X(HT) = 1$

$X(TH) = 1$

$X(TT) = 0$

| $X$ : | 0 | 1 | 2 |
|---|---|---|---|
| $P(X=x)$ : | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ |

$P(X=0) = P(TT)$

$= P(T \cap T)$

$= P(T) \cdot P(T)$

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

$= \frac{1}{4}$

$P(X=1) = P(HT \text{ or } TH)$

$= P(HT) + P(TH)$

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

$= \frac{1}{2}$

$P(X=2) = P(HH)$

$= P(H \cap H)$

$= P(H) \cdot P(H)$

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

$= \frac{1}{4}$

## Expectation of a Random Variable

$X \rightarrow$ discrete random variable

$$E(x) = \sum_i x_i \, P(x = x_i)$$

$$E(x^r) = \sum_i x_i^r \, P(x = x_i) \quad \text{where } r = 1, 2, 3, \ldots$$

$$E(x^2) = \sum_i x_i^2 \, P(x = x_i)$$

Mean and variance of discrete random variable

$$\text{Mean}(\mu) = E(x) = \sum_i x_i \, P(x = x_i)$$

$$\text{Variance}(\sigma^2) = E(x-\mu)^2 = E(x^2) - \mu^2$$

### Properties :

(i) $E(K) = K$, where 'K' is any constant

(ii) $E(X+Y) = E(X) + E(Y)$

(iii) $E(aX+bY) = a\,E(X) + b\,E(Y)$

(iv) $E(aX+b) = aE(X) + b$

(v) $V(K) = 0$ ; where V = Variance

(vi) $V(aX+b) = a^2 V(X)$

(vii) $E(XY) = E(X) \cdot E(Y)$ ; if X, Y are independent

→ For the following discrete probability distribution

| X | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| $P(x=x_i)$ | 0 | K | 2K | 2K | 3K | $K^2$ | $2K^2$ | $7K^2+K$ |

Find (i) K

(ii) $P(x<6)$

(iii) $P(x \geq 6)$

(iv) $P(0<x<5)$

(v) Mean and variance

X is a discrete random variable

$$\sum_i p_i = 1$$

$$0 + k + 2k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k = 1$$

$$10k^2 + 9k = 1$$

$$10k^2 + 9k - 1 = 0$$

$$10k^2 + 10k - k - 1 = 0$$

$$10k(k+1) - (k+1) = 0$$

$$(k+1)(10k-1) = 0$$

$$k = -1 \text{ (or) } k = \frac{1}{10}$$

$\because p_i > 0, \forall i$, then $k = \frac{1}{10}$ ✓

$P(X<6) = P(X=0 \text{ or } 1 \text{ or } 2 \text{ or } 3 \text{ or } 4 \text{ or } 5)$

$= P(0) + P(1) + P(2) + P(3) + P(4) + P(5)$

$= 0 + k + 2k + 2k + 3k + k^2$

$= k^2 + 8k$

$= \left(\frac{1}{10}\right)^2 + 8\left(\frac{1}{10}\right)$

$= \frac{81}{100}$

$= 0.81$

$P(X \geq 6) = P(X = 6 \text{ or } 7)$

$= P(6) + P(7)$

$= 2k^2 + 7k^2 + k$

$= 9k^2 + k$

$= 9\left(\frac{1}{100}\right) + \frac{1}{10}$

$= \frac{19}{100}$

$= 0.19$

$P(0<X<5) = P(X = \text{ or } \text{ or } \text{ or } )$

$= k + 2k + 2k + 3k$

$= 8k$

$= 8\left(\frac{1}{10}\right)$

$= 0.8$

$$\text{Mean}(\mu) = E(x)$$

$$= \Sigma x_i \, P(x = x_i)$$

$$\mu = 0(0) + 1(K) + 2(2K) + 3(2K) + 4(3K) + 5(K^2) + 6(2K^2) + 7(7K^2 + K)$$

$$\mu = 6K^2 + 30K$$

$$\mu = \frac{66}{100} + \frac{30}{10}$$

$$\mu = \frac{366}{100}$$

$$\boxed{\mu = 3.66}$$

$$\text{Variance}(\sigma^2) = E(x^2) - \mu^2$$

$$E(x^2) = \Sigma x_i^2 \, P(x = x_i) = 0^2(0) + 1^2(K) + 2^2(2K) + 3^2(2K) + 4^2(3K) + 5^2(K^2) + 6^2(2K^2) + 7^2(7K^2 + K)$$

$$= 440K^2 + 124K$$

$$= \frac{440}{100} + \frac{124}{10}$$

$$= 16.8$$

$$V(x) = E(x^2) - \mu^2$$

$$= 16.8 - (3.66)^2$$

$$= 3.4044$$

→ For the following discrete probability distribution.

| X | −3 | −2 | −1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| $P(x = x_i)$ | K | 0.1 | K | 0.2 | 2K | 0.4 | 2K |

Find (i) K

(ii) Mean

(iii) Variance

# Continuous Probability Distribution

$$\frac{d}{dx}(F(x)) = f(x)$$

$$P\left(x - \frac{\partial x}{2} \le X \le x + \frac{\partial x}{2}\right) = f(x)\,dx$$

$f(x) \rightarrow$ probability density function (pdf)

$f(x) \ge 0, \forall x$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\,dx$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx$$

$$E(X^r) = \int_{-\infty}^{\infty} x^r f(x)\,dx \quad \text{where } r = 1, 2, 3, \cdots$$

$\mu = E(X)$

$V(X) = E(X^2) - \mu^2$

① $\rightarrow$

$$f(x) = \begin{cases} K(1-x^2), & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$