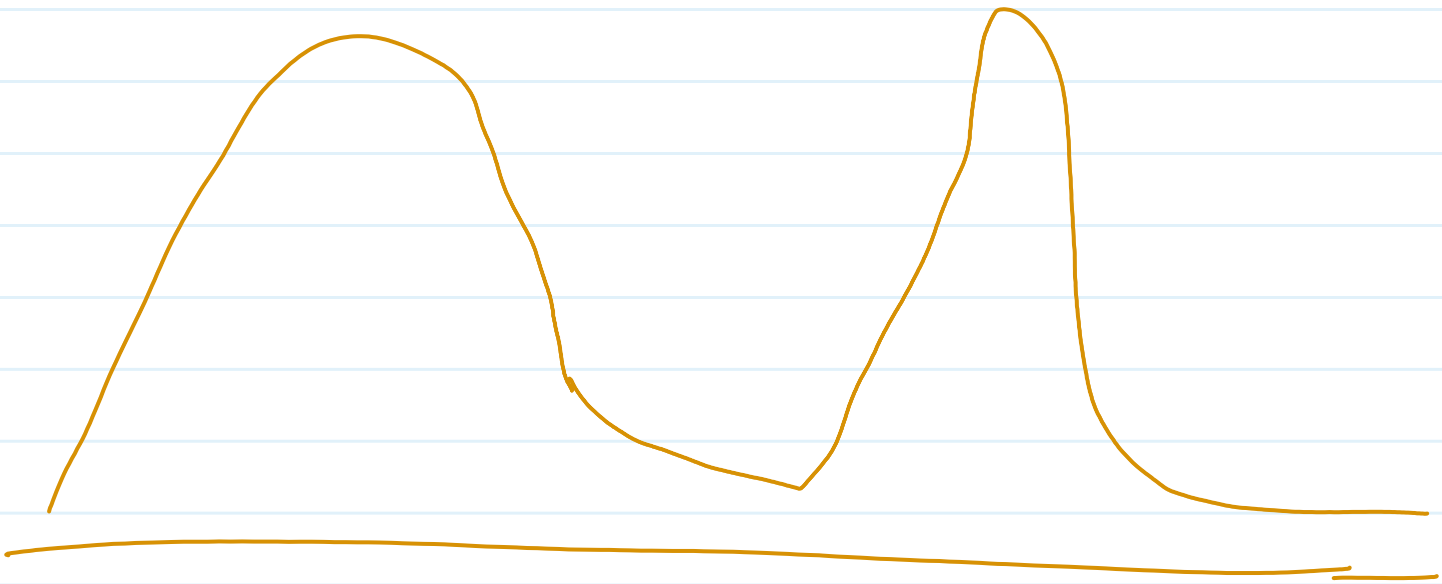
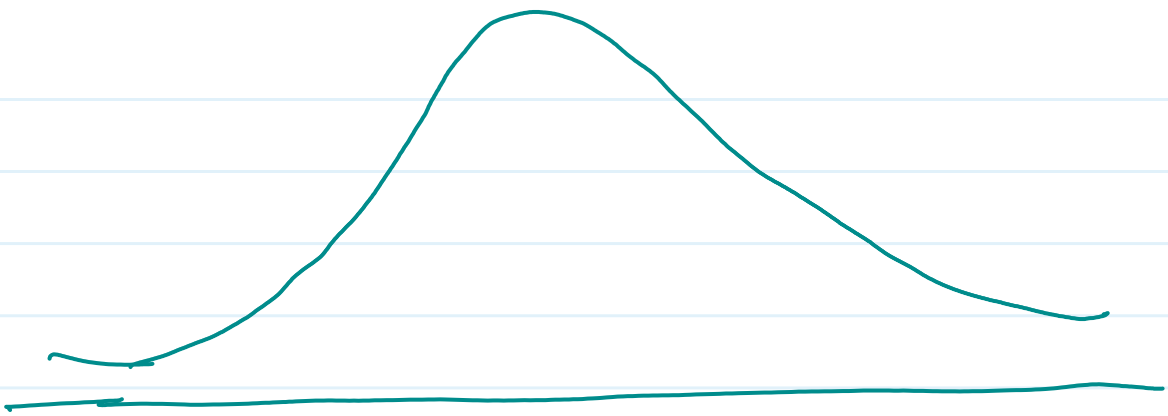
* Center limit theorem

The CLT states that regarding of the shape of population dist. the dist. of sample means will be approximately normal



$n_1 - \bar{x}_1$
$n_2 - \bar{x}_2$
$n_3 - \bar{x}_3$
$\vdots \quad \vdots$
$n_{50} - \bar{x}_{50}$

New sample data $[\bar{x}_1\ \bar{x}_2\ \bar{x}_3 - - - - - \bar{x}_{50}]$



Eg:- population size - 10k

Sample size $n \geq 30$

more the sample you will take
more the plot will normally Dist.

Histogram

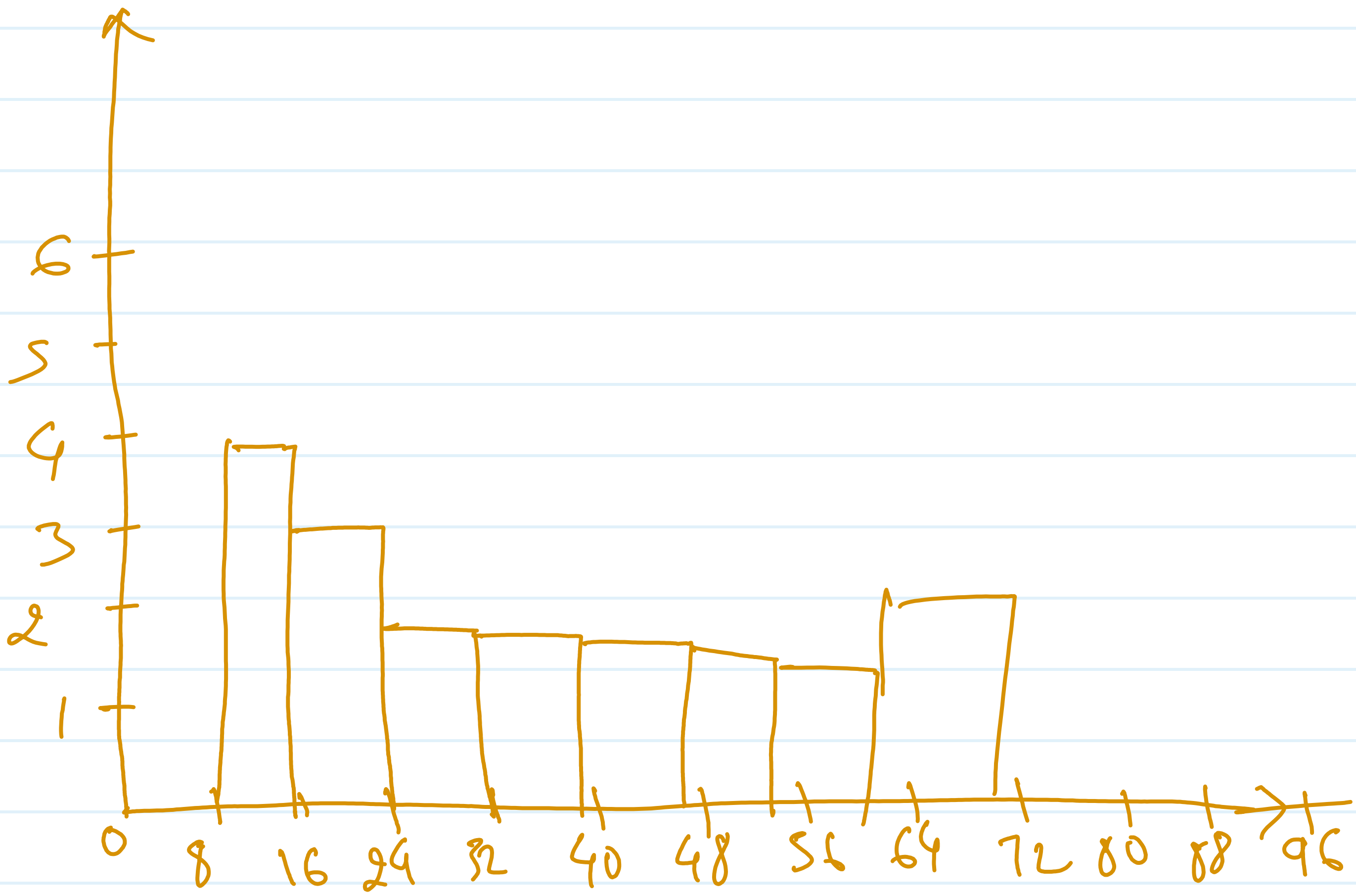
Dataset = [10, 12, 13, 14, 20, 22, 24, 25, 26, 35, 38, 42, 47, 55, 56, 68, 69, 82, 87, 92,]

Bin size

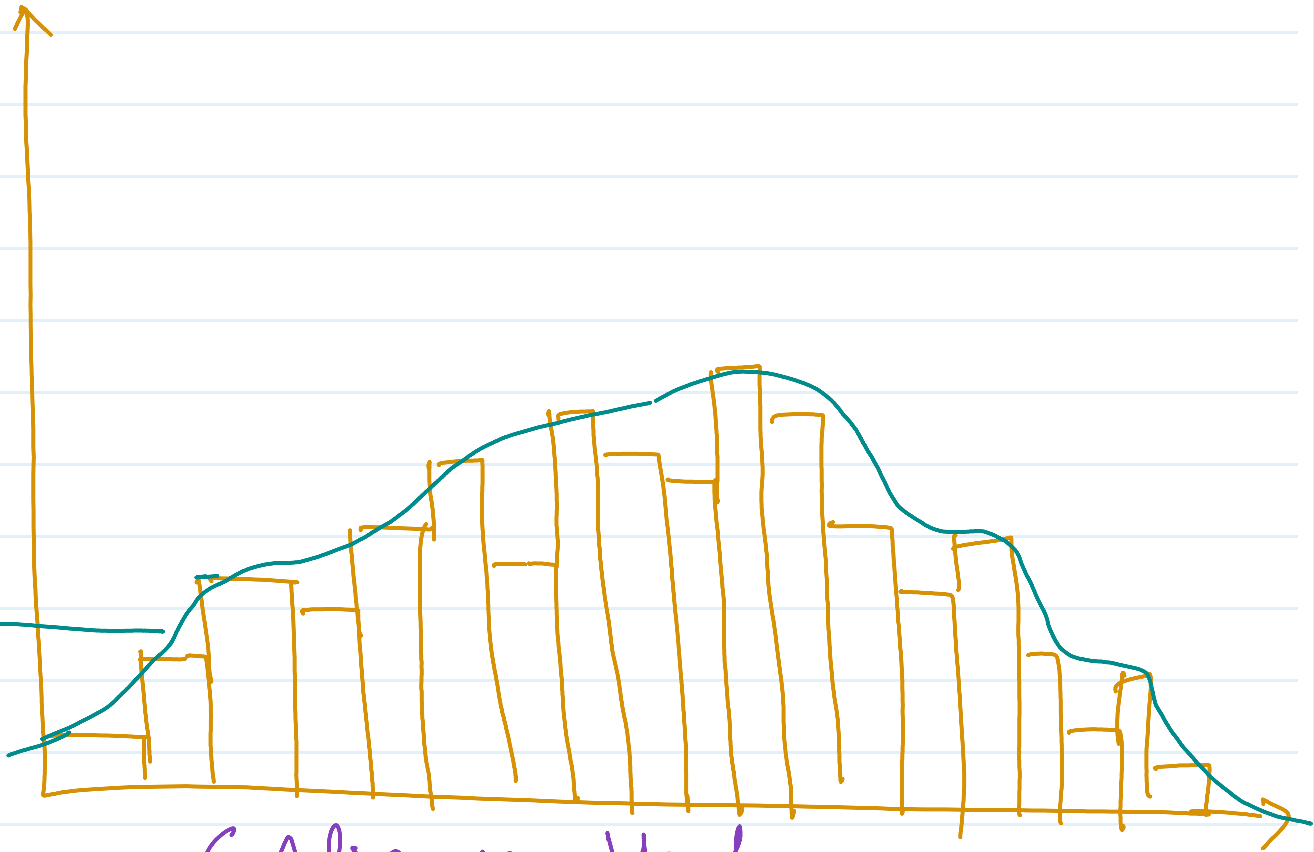
$$\text{Bin size} = \frac{\text{last ele} - \text{first ele.}}{\text{bin no.}}$$

$$= \frac{92 - 10}{10}$$

$$= 82/10 = \underline{8.2}$$



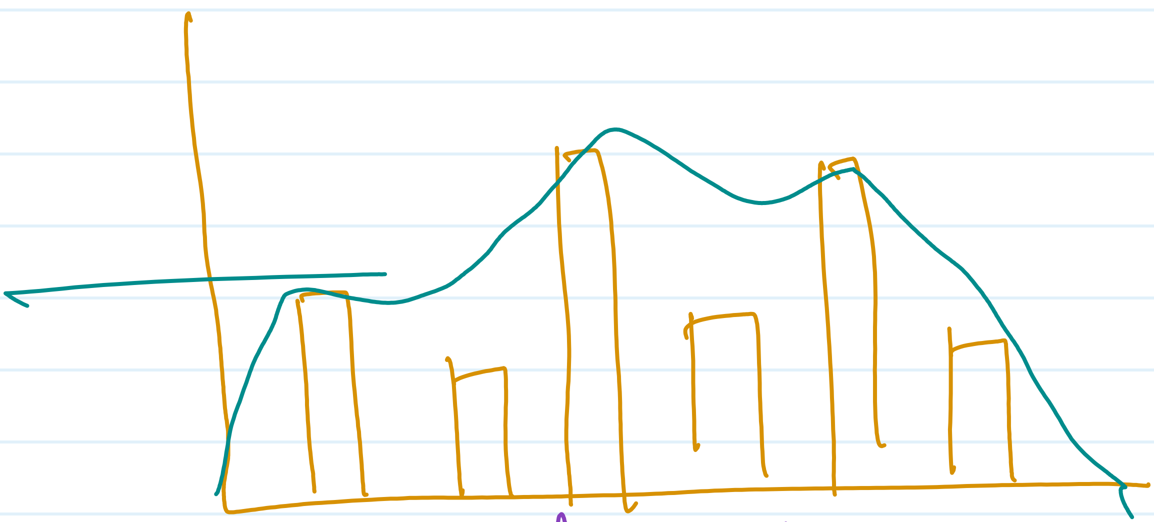
Pdf



Continues Hist.

Pdf = Probability density funct.

Pmf



Discrete Histo

Pmf - Probability mass function.

✯ Covariance

E.g.

X	Y
Age	Weight
14	40
15	45
18	51
20	68
25	74

$$\left.\begin{array}{ll} X\uparrow & Y\uparrow \\ X\downarrow & Y\downarrow \end{array}\right\} \text{+ve covariance}$$

$$\left.\begin{array}{ll} X\uparrow & Y\downarrow \\ X\downarrow & Y\uparrow \end{array}\right\} \text{−ve covariance}$$

$$\left.\begin{array}{ll} X\uparrow & Y\uparrow \\ X\uparrow & Y\downarrow \\ X\uparrow & Y \uparrow\downarrow \end{array}\right\} \text{Zero Covariance}$$

Age	weight	height	BMI
X_1	X_2	X_3	Y
.	.	.	9
—	—	—	.

X = feature column / Independent var.

Y = target column / Dependent vari.

Covariance :-

Quality the relationship b/w X & Y numeric value.

population $\text{COV}(X,Y) = \sum_{i=1}^{N} \frac{(X-\bar{X})(Y-\bar{Y})}{N}$

$$\text{Cov}(x,y) = \sum_{i=1}^{n} \frac{(x-\bar{x})(y-\bar{y})}{n-1}$$

Eg. Eco. growth   Nifty 50%

2.1	8
2.5	12
3.6	10
4.0	14

Covariance

x	y	$\bar{x}$	$\bar{y}$	$(x-\bar{x})$	$(y-\bar{y})$
2.1	8			-1	-3
2.5	12	3.05		-0.6	1
3.6	10		11	0.5	-1
4.0	14			0.9	3

$$\text{Cov}(x,y) = \sum_{i=1}^{n} \frac{(x-\bar{x})(y-\bar{y})}{n-1}$$

$$= \frac{(-1)(-3)+(-0.6)(1)+(0.5)(-1)+(0.9)(3)}{4-1}$$

$$= \frac{4.6}{3} = \boxed{1.533}$$

if value comes in +ve than there are positive covariance b/w x and y.

X ↑ Y ↑

X ↓ Y ↓

+1000 to -1000

Co-relation ⇒

-1 to 1

Relation   gym / week

pearson correlation coffient

x	y
2.1	8
2.5	12
3.6	10
4.0	19

Formula

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

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

$$= \sqrt{\frac{(-1)^2 + (-0.6)^2 + (0.9)^2 + (0.5)^2}{3}}$$

$$= \sqrt{0.8060} = 0.8981$$

$$\sigma_y = \underline{\underline{2.58}}$$

$$Cov(x,y) = \underline{1.533}$$

$$\rho(x,y) = \frac{1.533}{(0.89) \cdot (2.58)}$$

$$= 0.66$$

$$= 66\%$$

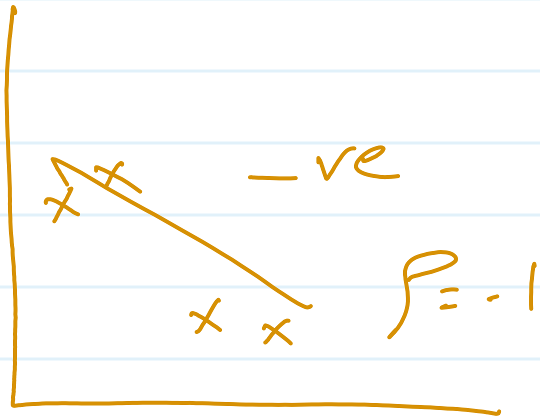
0.9 0.4 0.3 0.2 0.1 0 0.1 0.2 0.3 0.9

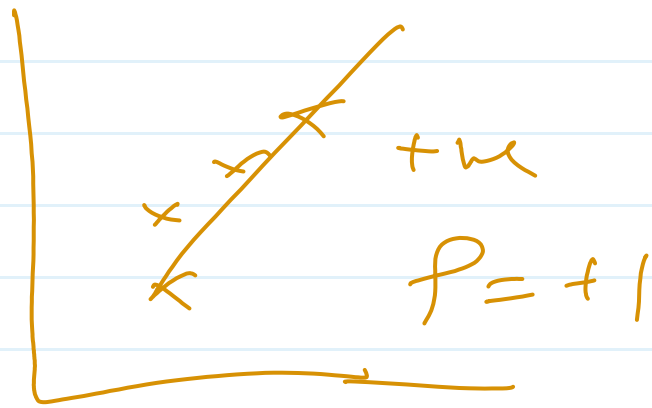
$\underbrace{-0.6}_{S}$ $\quad -0.1$ $\quad +0.1$ $\quad \underbrace{+0.7}_{S}$

W $\quad$ W.

$\underline{-0.01}$ $\quad \underline{+0.02}$

near to zero





Spearman Rank correlation

$$r_s = \frac{Cov(R_x, R_y)}{\sigma_{R_x} \sigma_{R_y}}$$

X	Y
2.1	8
2.5	12
3.6	10
4.0	14

Rank X $\qquad$ Rank Y

X		Y	
2.1	4	8	4
2.5	3	12	2
3.6	2	10	3
4.0	1	14	1

Mean $= R\bar{X}$ $\dfrac{4+3+2+1}{4} = 2.5$

Mean $R\bar{Y}$ $\dfrac{4+3+2+1}{4} = 2.5$

$(X-\bar{X})$		$(Y-\bar{Y})$	
$4-2.5$	$= 1.5$	$4-2.5$	$= 1.5$
$3-2.5$	$= 0.5$	$2-2.5$	$= -0.5$
$2-2.5$	$= -0.5$	$3-2.5$	$= 0.5$
$1-2.5$	$= -1.5$	$1-2.5$	$= -1.5$

$$Cov_{(x,y)} = \sum_{i=1}^{n} \frac{(x-\bar{x})(y-\bar{y})}{n-1}$$

$$= \frac{(1.5)\times(1.5) + (0.5)(-0.5) + (-0.5)(0.5) + (-1.5)(-1.5)}{4-1}$$

$$\Rightarrow \quad 1.33$$

$$SD\ \sigma_x = \sqrt{\frac{(1.5)^2 + (0.5)^2 + (-0.5)^2 + (-1.5)^2}{4-1}}$$

$$= 1.288$$

$$SD\ \sigma_y = 1.288$$

$$r_s = \frac{R_x R_y}{\sigma_{R_x}\ \sigma_{R_y}}$$

$$= \frac{1.33}{1.288 \times 1.288}$$

$= 0.81$

$$\boxed{= 81\%}$$

positive correlation is 81% out of It is quite strong.