STATISTICS AND PROBABILITY

Random Variables:

Dice: Six sides ⟶ [1,2,3,4,5,6]

Roll a dice ⟶ Any one outcome

X ⟶ Outcome of Experiment ⟶ Will be a Random Variable

↘ Can take any one value out of [1,2,3,4,5,6]

P(X=1) =Probability of Random variable taking 1 as the value.

= 1/6

Random Variable can take one value from a set of finite elements then it is known as DISCRETE RANDOM VARIABLE.

Height of Randomly picked student

Y=162.23,170.12121,156.3……

Random Variable Y which can take any value between a range is known as CONTINOUS RANDOM VARIABLE.

Measure of Central Tendency

Mean: Tells us about the average behaviour of observations.

$$\mu = \frac{sum\ of\ all\ observations}{No.of\ Observations}$$

-Even one Outlier can disturb the mean

Eg: Let's say you are observing salaries.

You will observe values [10k,20k,30k,22k,50k,10k……50000k]

We observe a potential outlier in our observation (salary of a CEO). This observation will change our mean also by a significant amount.

Median: Centre value of sorted observations

-Median is not affected by outliers

Eg: X= [1,1.1,1.2,1.4,1.6,1.6,1.8] ⟶ Here 1.4 is the median

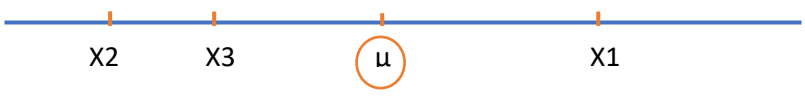
X'= [1,1.1,1.2,1. 4,1.6,1.6,1.8,56] ⟶ Here 56 is an Outlier and $\frac{1.4+1.6}{2}$ = 1.5

We can clearly observe that the median is not affected by significant level when there is an outlier.

Note: If more than 50% data (observations) are outliers then Median is affected.

Variance:

-How far are my points from mean (X1, X2 and X3 from μ) .



-Spread is variance

$$\text{Variance} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

-Squared- Because distance is -ve for one side.

-Therefore, average squared distance of each point from mean is Variance.

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

Standard deviation- What is the average deviation of points from mean value.

Percentile:

The pth percentile is a value such that at least p percent of the observations are less than or equal to this value and at least (100- p) percent of the observations are greater than or equal to this value.

CALCULATING THE pTH PERCENTILE

$$i=(\frac{p}{100})n$$

n=No. of data points i=Index p= pth percentile

If i is not an integer, round up. The next integer greater than i denotes the position of the pth percentile.

If i is an integer, the pth percentile is the average of the values in positions i and i 1.

COVARIANCE:

Covariance is a measure of how much two Random Variables vary together

(Is there any relation between X and Y).

X=Height	Y=Weight
160	62
150	54

$$\text{Cov(X,Y)} = \frac{1}{n}\sum_{1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

Where μ_x =Mean of X

μ_y=Mean of Y

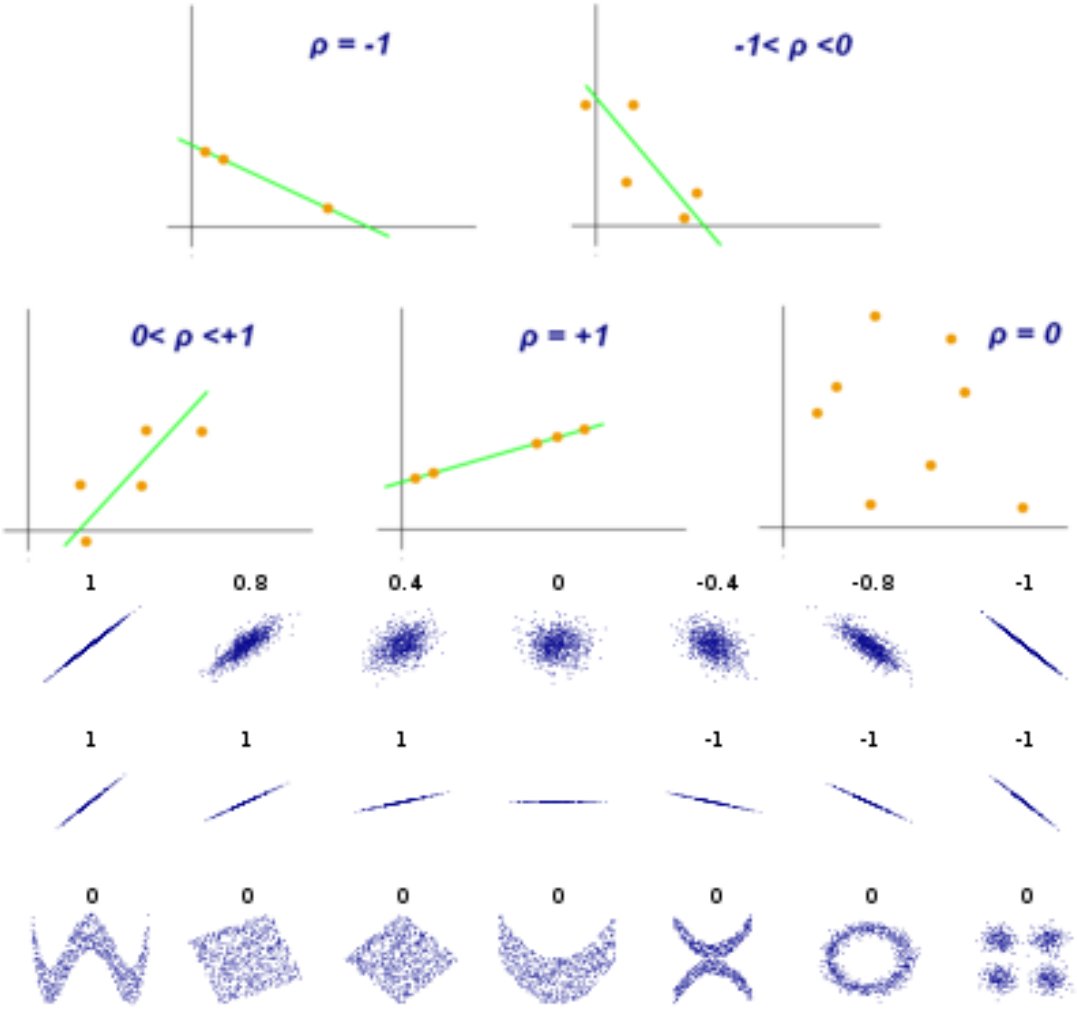
-Covariance is biased towards linear data

-Cov (x1, y1) ≠Cov (x2, y2) x1, y1 units are in cm and kg

X2, y2 units are in ft. and lbs

PEARSON CORRELATION COEFFICIENT:

$$PCC(p)=\frac{COV(X,Y)}{stdDev(X)*stdDev(Y)}$$



In the last row we can see that Pearson correlation coefficient gives a value of 0.

But we can see there is a relation between X and Y.

We can say that Pearson coeff, is also biased towards linear data.

SPEARMAN RANK CORRELATION COEFFICIENT (r):

Student	X	Y	r_x	r_y
S1	160	52	4	3
S2	150	66	2	4
S3	170	68	5	5
S4	140	46	1	1
S5	152	51	3	3

$$r = PCC\ (r_x, r_y)$$

-If X increases and Y increases (Linear or non-Linear)

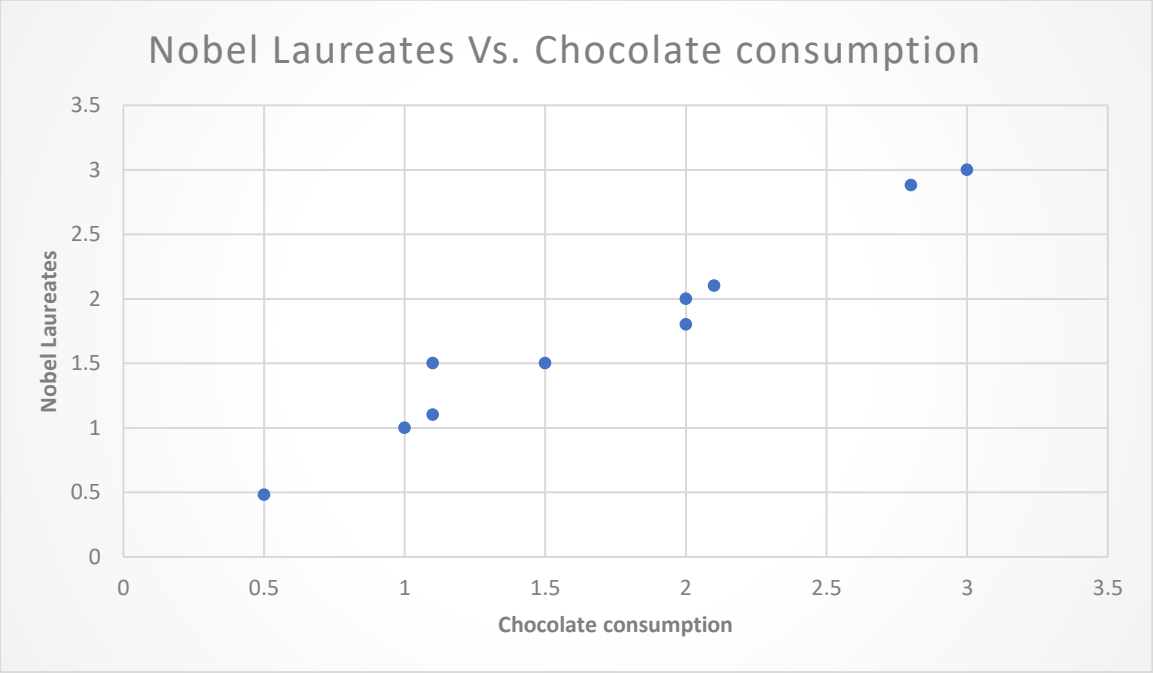
r=+1

-If X increases and Y decreases (Linear or non-Linear)

r=-1

NOTE:

Correlation does not imply causation.



In above example when X increases Y also increases but chocolate consumption is not at all related to No. of Laureates.