Q1. Differentiate between correlation and regression. Explain with suitable example using data.

Ans

Correlation	Regression
Correlation examines the relationship between two variables using a standardized unit. However, most applications use raw units as an input.	Regression examines the relationship between one dependent variables and one or more independent variables. Calculations may use either raw unit values, or standardized units as input.
The calculation is symmetrical, meaning that the order of comparison does NOT change the result.	The calculation is NOT symmetrical. So one variable is assigned the dependent role (the values being predicted) and one or more the independent role (the values hypothesized to impact the dependent variable).
Correlation coefficients indicate the strength of a relationship.	Regression shows the effect of one unit change in an independent variable on the dependent variable.
Correlation removes the effect of different measurement scales. Therefore, comparison between different models is possible since the rho coefficient is in standardized units.	Linear regression using raw unit measurement scales can be used to predict outcomes. For example, if a model shows that spending more money on advertising will increases sales, then one can say that for every added $ in advertising our sales will increase by β.

Correlation :

Example :

From the following data, compute the coefficient of correlation between X and Y :

	X series	X series
Number of Items	15	15
Arithmetic Mean	25	18
Sum of Square Deviations	136	138

Summation of products of deviations of X
and X series from their arithmetic means = 122.

Solution :

Here $n=15$, $\bar{X}=25$, $\bar{Y}=18$,

$\sum(X-\bar{X})2=\sum(Y-\bar{Y})2=138$

$\sum(X-\bar{X})2(Y-\bar{Y})2 = 122$ and hence

$R = \sum(X-\bar{X})(Y-\bar{Y}) / \sqrt{\sum(X-\bar{X})2\sum(Y-\bar{Y})2}$

$=122 / \sqrt{(136)(138)} = 122 / 137 = 0.89$.

Regression :

Example :

To find the Simple/Linear Regression of

X Values	Y Values
60	3.1
61	3.6
62	3.8
63	4
65	4.1

To find regression equation, we will first find slope, intercept and use it to form regression equation.

Number of values.   N = 5

Find XY, X^2 See the below table

X Value	Y Value	X*Y	X*X
60	3.1	60 * 3.1 = 186	60 * 60 = 3600
61	3.6	61 * 3.6 = 219.6	61 * 61 = 3721
62	3.8	62 * 3.8 = 235.6	62 * 62 = 3844
63	4	63 * 4 = 252	63 * 63 = 3969
65	4.1	65 * 4.1 = 266.5	65 * 65 = 4225

Find ΣX, ΣY, ΣXY, ΣX^2.

$\Sigma X = 311$ $\Sigma Y = 18.6$ $\Sigma XY = 1159.7$ $\Sigma X^2 = 19359$

Substitute in the above slope formula given.

Slope(b) =

$(N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$

 $= ((5)*(1159.7)-(311)*(18.6)) / ((5)*(19359)-(311)^2)$

$= (5798.5 - 5784.6)/(96795 - 96721) = 13.9 / 74 = 0.19$

Now, again substitute in the above intercept formula given.

Intercept(a) =

$(\Sigma Y - b(\Sigma X)) / N = (18.6 - 0.19(311))/5$

$= (18.6 - 59.09)/5$

$= -40.49/5 = -8.098$

Then substitute these values in regression equation formula

Regression Equation(y) = a + bx = -8.098 + 0.19x.

Suppose if we want to know the approximate y value for the variable x = 64.

Then we can substitute the value in the above equation.

Regression Equation(y) = a + bx = -8.098 + 0.19(64).
= -8.098 + 12.16
= 4.06

Q2. You want to find a measure of central tendency for income of persons who have boarded a particular train on a particular date at the originating station. What measure will you use and why? What measure will you choose for dispersion and why?

Ans.

Central tendency : It is a measure of central value for measuring Probability distribution.

It can be measured in three different ways :
1. Mean
2. Median
3. Mode

In this particular case we will use mode to measure the central tendency as :

1. There is no middle value to extracted from an incremental series ie Median.
2. No multiple data to calculate mean.
3. I will calculate mode as that is the price of the ticket is the only factor that reflects person income.

For Dispersion :
I will choose Mean Absolute Deviation

Mean Absolute Deviation :
The mean absolute deviation is the mean (average) of the absolute value of the difference between the individual values in the data set and the mean. The method tries to measure the average distances between the values in the data set and the mean

$$\text{Population MAD} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

Q.3
A) Distinguish between independence of events and mutual exclusivity of events with the help of example.

Ans

Independence of events :
Suppose there are two events A and B. If the occurrence of event A is not affecting the probability of occurring of event B, events are called independent event.

Ex : A person tossed a coin

CASE 1 :
Head occurs : Here's the probability of occurrence of head is independent of tail of coin. Hence it is an independent event.

CASE 2 :
Tail occurs : Here's the probability of occurrence of tail is independent of head of coin. Hence it is an independent event.

Mutual exclusivity :
It states that there will be one occurrence of one single event.
Suppose there are two events A and B. Now, either A will occur or B will occur.

CASE 1 :
Head will occur : If head occurs there's no probability that tail will occur.

CASE 2 :
Tail will occur : If head occurs there's no probability that head will occur.

B) In how many ways can the letters of the word "MADAM" be rearranged?

Ans

No of M = 2
No of A = 2

Total permutations = 5! / 2!2! = 30.