REGRESSION ANALYSIS

Regression measures the nature of relationship between two variables.It provides us with a functional relationship between two variables X and y in the form of an equation.Regression analysis is used for prediction/ estimation.That is once the functional relationship between X and y is known we can predict the unknown value of one variable given the known value of the other variable.In regression analysis we come across two types of variables - independent variable and dependent variable.The variable whose value is to be predicted is called dependent variable and the variable that is used for prediction is called independent variable.

Linear regression

When we plot the points in a bivariate data in a scatter diagram ,it concentrates around a straight line then we say that there is a linear regression between x and y and the straight line around which the points of the data concentrate is called the regression line or line of best fit.

Properties of regression coefficients

• byx ≠ bxy (Regression coefficients are not symmetric) • byx is the coefficient of x in the regression equation of y on x • boxy is the coefficient of y in the regression equation of x on y. • byx * bxy ≤ 1 • Correlation coefficient $=\sqrt{byx*bxy}$. That is correlation coefficient is the geometric mean of the regression coefficients. • r , byx and bxy have the same sign . Either all the three coefficients are positive or all the three are negative. • If the correlation coefficient r =0 , the two regression lines are perpendicular to each other. • If the correlation coefficient r =1, then the two regression lines coincide. • The point of intersection of two regression lines is ($\overline{x}$, $\overline{y}$). So by solving the equations of two regression lines we can find the means ($\overline{x}$, $\overline{y}$).

Identifying the two regression lines

Suppose we are given the equations of two regression lines. To identify which one is the regression line of y on x and which one is the regression line of x on y , we first assume that one of them as the regression line of y on x and the other as the regression line of x on y. From the assumed regression line of y on x , we calculate the regression coefficient byx and from the regression line of x on y , we calculate bxy . Then we will find the product byx* bxy. If this product is less than or equal to 1 then our assumption is true otherwise our assumption is false and we have to take the assumed regression line of y on x as x on y and regression line of x on y as y on x.

PROBABILITY

The word probability means chance. In statistics probability refers to a real number between 0 and 1 which represents the chance of occurrence of an event. **Random experiment** : A random experiment is an experiment having several possible outcomes but we cannot predict which outcome will turn up in a particular trial.
Example :- Tossing a coin, Throwing a die , selecting a card from a pack of cards etc.
Sample space : Sample space is the set of all possible outcomes of a random experiment. It is denoted by the letter S or Ω
Example :- • In the coin tossing experiment sample space S = { H , T } • In the die throwing experiment sample space S = { 1,2,3,4,5,6 }
• In the random experiment of tossing two coins sample space S ={ HH,HT,TH,TT }
Sample points : The elements of the sample space are called samples points. Trial : Trial is an attempt to produce an outcome of the random experiment. Events : Out of all the outcomes of a sample space certain outcomes satisfy a particular condition. Set of such outcomes are called events. Events are subsets of the sample space. They are denoted by the letters A,B,C …

Equally likely events : Two or more events are said to be equally likely if they have the same chance of occurrence in a trial.

Mutually exclusive events : Two or more events are said to be mutually exclusive if they cannot occur simultaneously in the same trial.

Favourable cases of an event : The outcomes which results in the happening of the event is called favourable cases of the event.

Exhaustive cases of a random experiment : The totality of outcomes of a random experiment are called its exhaustive cases.

Factorial notation , Permutations and Combinations Factorial : The factorial of a number is the product of integers from 1 to that number.
n! = 1x2x3x….n
Eg : 5! = 1x2x3x4x5

Permutations : The number of arrangements of n things taken r at a time is given by nPr = n(n-1)(n-2) …….. [r terms] Eg : 7P3 = 7 x 6 x 5
The number of arrangements of n things taken n at a time is given by nPn = n(n-1)(n-2)….3.2.1 = n!
Note : In probability permutation is used in problems of arranging the letters of
a word , arranging people on seats etc.

Combinations : The number of ways in which n things can be combined taking r at a time is given by nCr $= n(n-1)(n-2)/.....1.2.3.....r$

Limitations of classical definition of probability

• If the events of a random experiment are not equally likely classical definition cannot be applied.
• If the total number of outcomes of a random experiment becomes infinite classical definition fails to give a measure of probability. • Classical definition of probability can be applied mainly to games of chance like tossing a coin ,throwing a die etc.

Frequency definition of probability

Consider a random experiment which is repeated n times. Let an event A occurs in f out of n repetitions. Then f is called frequency and   is called the frequency ratio or relative frequency of the event A. When n becomes very large , the frequency ratio becomes more regular and approaches a constant. This constant value is called Probability of event A and the process of frequency ratio approaching a constant when n becomes very large is called Statistical regularity. That is by frequency definition , $\mathrm{P(A)} = \boldsymbol{fn}$, when n is very large **Limitations of frequency definition of probability** • If a random experiment is repeated a large number of times , the experimental conditions may not remain identical.
• The limit may not attain a unique value, however large n may be

Algebra of events

For an event A
1) Occurrence of an event A → A
2) Non occurrence of an event A (not A)→ Ac or A'
For any two events A and B
1) Occurrence of both (A and B)→ AՈ B
2) Occurrence of at least one of A, B (A or B) → AUB 3) Occurrence of only A (A and not B) → AՈ Bc 4) Occurrence of only B (not A and B) → AcՈ B 5) Occurrence of exactly one → (AՈ Bc)U(AcՈ B) 6) Occurrence of none (not A and not B) → AcՈ Bc

Addition theorem of probability

For any two events A and B , P(A or B) = P(A) + P(B) – P(A and B)
P(AUB) = P(A) + P(B) – P(AՈ B)
In particular if A and B are two mutually exclusive events then P(AUB) = P(A) + P(B)

FITTING OF CURVES/ CURVE FITTING

Fitting of curve to a given bivariate data refers to finding the most appropriate curve which fits the given data. Such a curve is called the curve of best fit. The curve of best fit is obtained using the principle of least squares.
The principle of least squares states that the sum of squares of errors between the observed values and the expected values should be minimum. Consider a bivariate data (x_1,y_1) ,(x_2,y_2),…… (x   ,y   ). We are interested in finding out an equation of a curve y = f(x) which shows the approximate relation between x and y values so that the sum of squares of errors is minimum. Here x is the independent variable and y is the dependent variable. In the data y_i represents the observed value of the variable y corresponding to the value of x , say x_i , i = 1,2…n .Let ye be the expected value of y .Then the error $e_i = y_i$-ye , i = 1,2…n The sum of squares of errors is given by $S = \boldsymbol{\Sigma}(y_i\text{- ye})^2$ By the method of calculus S is minimum if its first derivative is 0 and the second derivative is greater than 0. By equating the first derivative to 0 , we can find the unknown constants in the equation y = f(x) and hence find the equation of the curve of best fit.

Fitting of straight line Consider the problem of fitting a straight line of the form y= ax + b to a given bivariate data. Here we have to find out the constants a and b. The two normal equations for finding the unknown constants a and b is obtained using the principle of least squares. By equating dS/da= 0 and dS/db= 0 we get two normal equations ,The normal equations are
$\boldsymbol{\Sigma y_i = a\Sigma x_i + nb}$
$\boldsymbol{\Sigma x_iy_i = a\Sigma x_i2 + b\Sigma x_i}$
By solving these two normal equations we can find out a and b and hence the line of best fit to the given data.

CORRELATION ANALYSIS

• Statistical data : Statistical data refers to set of numbers collected for a predetermined purpose.
• Univariate data :Statistical data providing information about a single characteristic or variable (x) is called univariate data. Eg : The marks of students in a class. • Bivariate data : Statistical data providing information about two characteristics or variables (x,y) is called bivariate data. Eg : The heights and weights of students in a class Price and demand of different commodities
CORRELATION Correlation is a statistical device which measures the degree or strength of relationship between two variables. A high correlation means there is a strong relationship between the variables and a low correlation means the realationship between the variables are weak. **TYPES OF CORRELATION** • Positive correlation : If in a bivariate data , the values of two variables move in the same direction , the correlation is said to be positive. That is when the value of the variable x increases ,the value of the variable y also increases or when x decreases y also decreases , then the correlation is said to be positive. • Negative correlation : If in a bivariate data , the values of two variables move in the opposite directions , the correlation is said to be negative. That is when the value of the variable x increases , the value of y decreases or when x decreases y increases ,then the correlation is said to be negative. • Zero correlation or no correlation : When there is no association between the two variables x and y then there is no correlation or zero correlation between x and y . • Perfect correlation : If the values of one variable is proportional to the values of other variable , then the correlation is said to be perfect. If the values are directly proportional then the correlation is perfectly positive

KARL PEARSON'S COEFFICIENT OF CORRELATION

Karl Pearson's coefficient of correlation is a mathematical method for studing correlation. Karl Pearson's coefficient of correlation is a real number lying between -1 and +1 which tells us the degree or strength of relationship between two variables. It is denoted by the letter 'r' or 'rxy'.

$$\mathrm{r} = \frac{Cov(x,y)}{\sqrt{V(x)}\sqrt{V(y)}} = \frac{Cov(x,y)}{\sigma x\,\sigma y}$$

On simplification , the formula for Karl Pearson's correlation coefficient can be written as

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

Spearmans's rank correlation coefficient measures the correlation between two sets of ranks. Usually qualities like beauty , intelligence , sincerity etc cannot be measured directly. Instead they can be given ranks. In such cases we can use Spearman's rank correlation coefficient to measure their degree of relationship

RANDOM VARIABLES

A random variable is a real valued function defined over the sample space of a random experiment. Random variables are usually denoted by X ,Y, Z etc. The domain of the function random variable is the sample space and the range is the set of real numbers. **Discrete random variables :** A random variable is said to be discrete if it takes finite or countably infinite number of values.

Probability mass function (pmf)

Let X be a discrete random variable then the function f(x) = P(X = x) is called the probability mass function or pmf of X , if it satisfies the following conditions
1) f(x) ≥ 0 , for every x
2) $\Sigma f(x)x = 1$ (Total pmf = 1)

Cumulative distribution function (cdf) /Distribution function let X be a discrete random variable having the pmf f(x) , then the cdf of X is given by F(x) = P(X ≤ x) = $\Sigma f(x)x{-}\infty$
Properties of cdf
- F(x) ≥ 0
- F(-∞) = 0
- F(+∞) = 1
- The graph of the cdf F(x) is a step function.
- P(a < X ≤ b) = F(b) – F(a)

Continuous random variables
A random variable is said to be continuous if it can take uncountably infinite number of values.
Examples of continuous random variables
- Life time of an electric bulb.
- Temperature of a place.
- Time taken to finish a running rac
- Length of a film.

Probability density function (pdf)

The probability of a continuous random variable X is represented by a function f(x), This function f(x) is called the probability density function or pdf of X if it satisfies the following conditions
1) f(x) ≥ 0 , for every x
2) $\int f(x)dx = 1x$ (Total pdf =1)

MEASURES OF DISPERSION

Measures of dispersion are the statistical devices to measure the scatteredness or variation of observations in a set of data. They tell us the extent to which the values of a a data differ between each other or from their average.
Purpose of measuring variation
• To test the reliability of an average.• To compare the variability of two or more sets of data.• To exercise control over variability.Desirable properties of an ideal measure of dispersion• It should be rigidly defined.• It should be simple to understand and easy to calculate.• It should be based on all observations.• It should be capable of further mathematical treatment.• It should have sampling stability **Types of measures of dispersion**Measures of dispersion are classified as absolute measures of dispersion and relative measures of dispersion.Absolute measures of dispersion are used to find the variations among a single set of data.Relative measures of dispersion are used for comparing two or more sets of data for their variability.Important absolute measures of dispersion are
• Range• Quartile Deviation (QD)• Mean Deviation (MD)• Standard Deviation (SD)
Corresponding relative measures of dispersion are
• Coefficient of range• Coefficient of quartile deviation• Coefficient of mean deviation
• Coefficient of standard deviation and Coefficient of variation.Note : Relative measure of dispersion is the ratio of absolute measure of dispersion to an appropriate average from which the deviations are measured.

RANGE

Range is the difference between the highest and lowest values in a data. Range = H – L Coefficient of range $= H{-}LH{+}L$Where H = Highest value in the data andL = Lowest value in the data.
Merits of range• It is simple to understand and easy to calculate.• Range is a popular measure in the field of medicine and weather forecast.**Demerits of range**• It is not based on all observations.• It does not have sampling stability .
• It cannot be calculated for open end data.
Quartile deviation is defined as half the difference between third and first quartiles. QD $= Q3{-}\ Q12$
The only measure of dispersion that can be applied for an open end data is Quartile deviation
Merits of QD• It is rigidly defined.• It is simple to understand and easy to calculate.• It is not unduly affected by extreme values.• It can be calculated for open end data.
Demerits of QD• It is not based on all observations.• It is not capable of further mathematical treatment.• It doesn't have sampling stability.

MEAN DEVIATION (MD)

Mean deviation is defined as the Arithmetic mean of absolute values of deviations of observations from an average. The average can be mean , median or mode.**Merits of mean deviation**
• Mean deviation is rigidly defined• It is based on all observations• It is less affected by extreme values**Demerits of mean deviation**• Mean deviation suffers from inaccuracy because signa of the deviations are ignored• It is not capable of further mathematical treatment
• Cannot be calculated for open end data

STANDARD DEVIATION (SD)

Standard deviation is defined as the square root of arithmetic mean of squares of deviations of observations from arithmetic mean.
It is denoted by the letter σ .
The square of standard deviation is called variance.
It is denoted by σ2 .
Standard deviation in raw data SD , $\sigma = \sqrt{[1n\Sigma x2-\ x\ 2]}$ Or $\sigma = \sqrt{1n\Sigma(x{-}x\ )2}$ Variance , $\sigma 2 = 1n\Sigma(x{-}x\ )2$ Coefficient of standard deviation $= SDAM = \sigma x$ Coefficient of variation (CV) $= \sigma x\ X\ 100$
Standard deviation in discrete series SD , $\sigma = \sqrt{1N\Sigma f(x{-}x\ )2}$, where N = Σf Variance , $\sigma 2 = 1N\Sigma f(x{-}x\ )2$ Coefficient of standard deviation $= SDAM = \sigma x$ Coefficient of variation (CV) $= \sigma x\ X\ 100$
Standard deviation in continuous series SD , $\sigma = \sqrt{1N\Sigma f(x{-}x\ )2}$, where N = Σf and x= mid value of classes Variance , $\sigma 2 = 1N\Sigma f(x{-}x\ )2$ Coefficient of standard deviation $= SDAM = \sigma x$ Coefficient of variation (CV) $= \sigma x\ X\ 100$

Merits of SD• It is rigidly defined
• It is based on all observations• It is capable of further mathematical treatment• It has sampling stability**Demerits of SD**• It is difficult to understand and calculate• It cannot be calculated for open end data

Definition of statistics as statistical data :

According to professor Horace Secrist "Statistics are aggregates of facts affected to a marked extent by multiplicity of causes, numerically expressed , enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation with each other. In short statistics are set of numbers collected for a predetermined purpose.

Definition of statistics as statistical methods :

According to Croxton and Cowden "Statistics may be defined as the science of collection, presentation, analysis and interpretation of numerical data

Statistical data

Statistical data may be classified into three types:
• Raw data. • Discrete series / discrete frequency distribution. • Continuous series / continuous frequency distribution.

Raw data is the unarranged data. Here all the numbers in the data are simply listed.

In discrete series the data values (x) are written along with their frequencies (f). **In continuous series** the data is divided into different classes and each class is written along with their frequencies.

MEASURES OF CENTRAL TENDENCY (AVERAGES)

Measures of Central tendency or averages are single representative values which represents a large number of numerical values. It can be considered as a central value around which all other values in the data cluster. Important measures of Central tendency /Averages are
• Arithmetic mean (AM) • Geometric mean (GM)
• Harmonic mean (HM) • Median • Mode

Here Arithmetic mean ,Geometric mean and Harmonic mean are called mathematical averages while Median is called a positional average.

ARITHMETIC MEAN (AM):

Arithmetic mean is defined as sum of observations divided by number of observations.

Arithmetic mean in raw data If there are n observations in the data then their arithmetic mean is given by $\bar{x}=\Sigma xn$

Arithmetic mean in discrete series :

If x denotes the data values and f denotes their frequencies then their arithmetic mean is given by $\bar{x}=\Sigma xfN$ Where N = Σf

Arithmetic mean in continuous series

If x denotes the mid values of the classes and and f denotes their frequencies, then their arithmetic mean is given by
$\bar{x}=\Sigma xfN$, where N = Σf

Merits of arithmetic mean • It is simple to understand and easy to calculate. • It is rigidly defined. • It is based on all observations. • It is capable of further mathematical treatment. • It has sampling stability. **Demerits of arithmetic mean** • It is effected by extreme values. • It cannot be calculated for open end data • It cannot be determined graphically

GEOMETRIC MEAN

Geometric mean of n observations is defined as the nth root of the product of the observations.

Geometric mean in raw data

Geometric mean , GM =Antilog(Σlogxn)

Geometric mean in discrete series

Geometric mean , GM =Antilog(ΣflogxN)
Where N= Σf . Here x represents the observations and f represents their frequencies.

Geometric mean in continuous series

Geometric mean , GM =Antilog(ΣflogxN)
Where N = Σ f . Here x represent midvalues of the classes and f represent their frequencies .

Note: Geometric mean is used to find the averages of ratios and percentages.

Merits of Geometric mean • It is rigidly defined.
• It is based on all observations. • It is capable of further mathematical treatment. • It has sampling stability. **Demerits of geometric mean**
• It is not simple to understand and is difficult to calculate. • If one or more of the values are zeros or negative then geometric means cannot be calculated. • It cannot be calculated for open end data. • It cannot be determined graphically

HARMONIC MEAN(HM)

Harmonic mean is defined as the reciprocal of the arithmetic mean of reciprocal of the observations.

Note :: Harmonic mean is used to find the average of speeds

Harmonic mean in raw data

If there are n observations in a data then their harmonic mean,
HM = $n\,\Sigma(1x)$

Harmonic mean in discrete series

If x denote the observations in the data and f represents their frequencies then
HM = $N\,\Sigma(fx)$
Where N = total frequency

Harmonic mean in continuous series

If x denote the mid values of the classes and f represents their frequencies then
HM = $N\,\Sigma(fx)$
Where N = total frequency

Merits of harmonic mean

• It is rigidly defined. • It is simple to understand and easy to calculate. • It is based on all observations. • It is capable of further mathematical treatment. • It has sampling stability. **Demerits of harmonic mean**
• It is affected by extreme values.
• If one of the values is zero, harmonic mean cannot be determined. • It cannot be determined graphically. • It cannot be calculated for open end data

Relation between AM,GM and HM

• AM ≥ GM ≥ HM
• GM2 = AM*HM or GM = √AM* HM

MEDIAN

Median is the middle most observation in a data which is arranged in ascending or descending order**.**

Median in raw data

Median = $n+12$ th observation, when the observations are arranged in ascending or descending order.

Median in discrete frequency distribution

Median =   +12 th observation where N = Σf
Steps for finding median in discrete frequency distribution
1. Calculate (N+1)/2
2. Find the cumulative frequency
3. Identify the cumulative frequency just greater than or equal to (N+1)/2
4. The observation corresponding to that cumulative frequency will be the median.

Median in continuous frequency distribution

Median = l + $(N2-m)cf$
Where l = lower limit of median class
Median class = class containing   +12 $\hbar$ observation. N = Σf ,total frequency
m = cumulative frequency of the class preceding the median class. c = class width of median class.
f = frequency of median class

Merits of median

• It is rigidly defined. • It is simple to understand and easy to calculate • It is not much affected by extreme values. • It can be calculated for open end data. • It can be determined graphically.
Demerits of median • It is not based on all observations. • It is not capable of further mathematical treatment. • It does not have sampling stability.

MODE : Mode is the most frequently occurring observation in a data.

Note : In some cases mode is ill defined. In such cases mode can be obtained using the empirical relation between mean median and mode, which is Mean – Mode = 3(Mean – Median)
Or Mode = 3Median – 2Mean.

Mode in raw data

Mode = The observation which repeats the most number of times

Mode in discrete series

Mode = observation having highest frequency.

Mode in continuous series

Mode = l +Δ1(Δ1+Δ2)x c
Where l = lower limit of the modal class.
Modal class = class having highest frequency.
f1 = frequency of the modal class.
f0 = frequency of the class preceding the modal class. f2 = frequency of the class succeeding the modal class.
Δ1 = f1 - f0
Δ2 = f1 - f2
c = class width of the model class.

Merits of mode

• Mode is simple to understand and easy to calculate. • It is not much affected by extreme values. • It can be calculated for open end data.
• It can be determined graphically.
• It is the most typical or representative value in a data since it has the greatest frequency.

Demerits of mode

• It is not rigidly defined. • It is not based on all observations. • It is not capable of of further mathematical treatment. • It does not have sampling stability.

AVERAGES	Raw data	Discrete series	Continuous series
Arithmetic mean	$\bar{x}=\Sigma xn$	$\bar{x}=\Sigma fxN$	$\bar{x}=\Sigma fxn$
Geometric mean	$GM=Antilog$ $(\Sigma logxn)$	$GM=A.ntilog$ $(\Sigma flogxN)$	$GM=Antilog$ $(\Sigma flogxN)$
Harmonic mean	**HM =** $n/\Sigma(1/x)$	**HM =** N/Σ(f/x)	**HM =** N/Σ(f/x)