Biostatistics: Introduction and applications

Dr. B. K. Jain
Principal
M. G. Science Institute
Ahmedabad

What is Statistics ?

Statistics is the science which deals with collection, analysis and interpretation of data obtained by conducting a survey or an experimental study.

What is Biostatistics ?

The application of statistics in biology is known as biostatistics or biometry.

Topics to be studied under biostatistics:

- Sample and sampling
- Collection and representation of data
- Measures of central tendency
- Measures of dispersion
- Distribution patterns
- Test of significance
- The chi-square test
- Correlation
- Regression Analysis
- Analysis of variance (ANOVA)

Methods of Data collection

- Data can be collected in two ways ;
- 1. Census method 2. Sampling method

1. **Census method** : Counting of Number of people/individuals of same species living in an area, region or country .

- **Advantages :**
  - All the items of analysis are studied
  - The analysis of data becomes more representative and true
  - Characteristics of population is maintained
  - Highest degree of accuracy is maintained

- **Disadvantages :**
  - Required large amount of time, energy and money
  - Large number of enumerations may require
  - In a changing situation the information may change with the change of time.

Sampling Method

2. **Sampling Method** : A part of population is taken into the consideration for analysis.

- A small group is chosen deliberately or at random from a large population..
- **Sampling methods may be :**

  A) Random sampling method **B) Non random sampling method**

(A). Random sampling method; (Probability Sampling):

-Random is not used in the sense of haphazard

-Random sampling suggests that selection should be made without deliberate discrimination .

It can be classified into:

i) Simple Random Sampling method

ii) Stratified Random Sampling method

iii) Systematic Random Sampling method

i) **Simple Random Sampling method**: A sample is selected in such a way that each item of the population has an equal and independence chance of being included in the sample. For Example :

a) Lottery method : Slips are made on each individual items. Somebody who is neutral or unbiased select the items from polpulation.

b) Random Sampling Number : First assign the serial number to each item. Now consult the random table given by L.H.C. Tippett.

ii) **Stratified Random Sampling :** This method is recommended when population is heterogenous.

Population is divided into strata or sub groups possessing the similar characteristics. Samples are selected by taking equal proportion of items from each group.

iii) **Systematic Random Sampling : (Quasi Method)**

Items are arranged in either temporal (temp., time), spatial (size, shape) or alphabetical order. Items are selected at fixed intervals.

B)Non Random Sampling Method:

Data are collected on the basis of expert judgment or convenience. It is of following types :

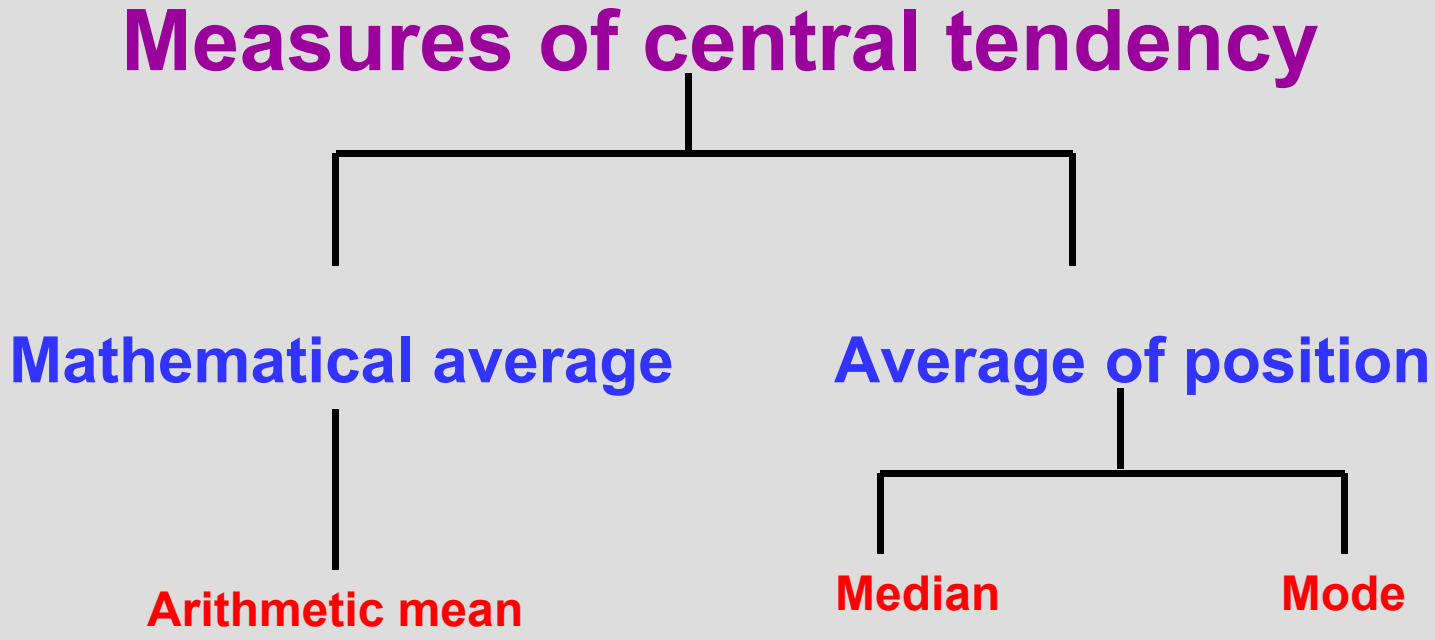
i) Purposive or judgment sampling method : No systematic planning is required. Investigator has the power of discretion and can deliberately select or reject any item. All the items do not have the same chance of being selected.

ii) Quota Sampling Method : Quota are set up for specific characteristics such as age, religion, urban area, rural area or salary groups. Items are selected non randomly from the groups.

iii) Convenience sampling method : fraction of population is being investigated. Selection is neither based on random nor on judgment but on convenience.

Measures of central tendency

- Generally it is found that values of the variable tend to concentrate around some central value of observation of an investigation , which can be taken as a representative for the whole data. This tendency of distribution is known as central tendency and the measures devised to consider this tendency are known as measures of central tendency.


Measures of central tendency
Mathematical average
Average of position
Arithmetic mean
Median
Mode

Mean

- The average obtained by adding together all the given values and by dividing this total value by the number of values.
- Simple mean can be calculated by using following formula :

$$\bar{X} = \frac{\Sigma X}{N}$$

If data are grouped :

(1) $$\bar{X} = \frac{\Sigma fX}{\Sigma f}$$

f = Frequency, X = value of variable

(2) $$\bar{X} = \frac{\Sigma fm}{\Sigma f}$$

m = mid value of a group

Example : 1

Find out the mean from given data:

5, 4, 6, 3,2

5 + 4 + 6 + 3 + 2 = 20

$$\bar{X} = \frac{\Sigma X}{N} = \frac{20}{5} = 4$$

Mean is 4

Example : 2

Find out the mean from given data:

Variable (x)	Frequency (f)	fx
2	10	20
3	08	24
4	12	48
5	20	100
Total	50	192

$$X = \frac{\Sigma fX}{\Sigma f} = \frac{192}{50} = 3.8$$

Example : 3

.

Calculate the mean from given data:

1,3, 5,2,5,8,7,4,6,4,3,3,1,2,9,7,8,2,2,1,

Class interval	Mid value(m)	Frequency(f)	mf
1-3	2	10	20
4-6	5	5	25
7-9	8	5	40
		20	85

$$X = \frac{\Sigma fm}{\Sigma f} = \frac{90}{20} = 4.2$$

Merits :

- It covers all the observations and is easy to calculate.
- It is affected least by fluctuation of sampling.
- It provides base for many other methods of statistics.

Demerits :

- ❖ It can not be determined by inspection.

- ❖ Obtained mean in a series may not be represented by any observation.
  2.6+2.6+3.2+3.2+3.4 = 15/5 = 3
- It is very much affected by extreme observations.
  3+5+50+62 = 120/4 =30
- Sometimes the value of mean will not be acceptable . Exam. Average no. of children in three families 2 + 2 + 3 = 7/3 = 2.3

Median:

A median of a distribution is defined as the value of that variable which divides the total frequency into two equal parts when the series is arranged in either ascending or descending order of magnitude.

Example : 3,5,2,4,1,7,8 (Odd number)

Arrange – 1,2,3,4,5,7,8

2,1,6,4,7,9,8,3

Arrange – 1,2,3,4,6,7,8,9 (Even number)

4+6 =10/2 =5

Merits :

It can be obtained directly.
It eliminates the effects of extreme values.
Easy to calculate.

Demerits :

It can not be found easily when data are in group.
It does not include extreme values in calculation
It is not very useful in further analysis

Mode:

Mode of a frequency distribution is defined as “that value of the variable for which the frequency is maximum”.

Example:

2,3,4,5,5,6,7,8, 5 is mode (Unimodel class)

2,3,4,5,5,6,6,7,8, 5&6 are mode (Bimodel)

Formula to calculate mode from grouped data

$$Mo = lo + F\text{-}1/ (F\text{-}1) + (F+1) \times I$$

lo =lower end value of model class

F-1 Freq. of class just prior to model class

F+1 Freq. of class just after to model class

I = class interval

Find out mode from given data (dry weight of plants in grams)

Class interval	frequency	mid value
161-170	04	165
171-180	07	176
181-190	09	187
191-200	12	198
201-210	16	209
211-220	21	220
221-230	18	231

$$Mo = lo + F\text{-}1/ (F\text{-}1) + (F+1) \times I$$

$$211 + 16/16+18 \times 10 = 215.70$$

Merits :

It can be ascertained by inspection.

It avoids the effects of extreme values

Demerits :

Arithmetic explanation of mode is not possible

It is difficult in multi model distribution

It is not based on all the observation of a series.

Measures of dispersion

3, 4, 5 Mean from these readings will be
3+4+5 =12/3 =4

4 does not indicate from which values it has been calculated 3+4+5 =12/3 =4
S.D. will be 4±1

This indicates that the minimum value in the series is 3 and maximum value is 5.

Dispersion can be calculated by using either variance or Standard deviation.

Variance (Measures of dispersion)

- Variance, also called mean square variance, is denoted by S^2.
- It is the sum of squared deviations of individual values from the mean, divided by the size of the sample less one.
- It can be calculated by the following formula:
- $S^2 = (X - X^-)^2 / (N) \text{ or } (N - 1)$
- $S^2 = \dfrac{\text{Sum total of } X^2}{N}$
- X = Individual value
- X^- = Mean value
- N = size

Calculate the variance from given data :
Length of fish (in cm) : 6, 7, 4, 5, 8

Length of fish	Mean	Deviation	X^2
6		6-6=0	00
7		7-6=1	01
4	30/5= 6	4-6= -2	04
5		5-6= -1	01
8		8-6= -2	04
---			---
30			10

S^2 = Sum total of X^2 / N

= 10/5 = 2 **Variance is 2**

Standard Deviation

- In biology most of the characteristics can not be depicted in square . For example height, weight, length etc can not ne depicted in square. Therefore, standard deviation is used to find out the deviation or variation from the mean value

- It may be defined as the square root of the arithmetic mean of the squares of deviations from the arithmetic mean.

It can be calculated using following formulae:

$$\sqrt{\frac{\sum d2}{N}}$$ where d is deviation

$$\sqrt{\frac{\sum fd2}{N}}$$ where f is frequency

Example

X value	mean	d= mean-X	d2
7		+1.6	2.56
6		+0.6	0.36
5	27/5	-0.4	0.16
3	=5.4	-2.4	5.76
6		-0.6	0.36
			9.20

$$\sqrt{\frac{\sum d2}{N}} = \sqrt{9.\frac{20}{5}} = 1.36$$

5.4±1.36

Example :

X	Freq.	Mean	d=mean-X	d2	fd2
2.6	8		+0.4	0.16	1.28
2.8	22		+0.2	0.04	0.88
3.0	40		0.0	0.00	0.00
3.2	18	15/3	+0.2	0.04	0.72
3.4	12	=3.0	+0.4	0.16	1.92
	100				4.80

$$\sqrt{\frac{\sum fd2}{N}} = \sqrt{4.\frac{80}{100}} = 0.22$$

S.d. = 3.00±0.22

C.V. = S/mean X100 0.22/3.0 X 100

= 7.3 %

Probability

The term probability is a vague concept which can not be defined mathematically.

Probability is the ratio of number of favorable cases to the total number of equally likely cases.

$$P = \frac{\text{Number of favorable cases}}{\text{Total number of equally likely cases}}$$

Example ; When we toss a coin, there are two equally likely results i.e. Head or Tail. Probability of any event will be always less than 1.

Suppose the result of a toss of coin 100 times is 60 times head and 40 times tail. The probability of head will be 60/100 = 0.60

Basic concept

- **An event :** An event is said to be collection of possible outcomes, when an experiment is conducted. Exam. In tossing a coin head and tail are the events.

- **Independent event :** Two events are said to be independent when occurrence of one does not affect the occurrence of other.

  Exam.; When two coins are tossed , the result of the first toss does not affect the result of second toss.

- **Dependent event :** Two events are said to be dependent if the occurrence of one affect the occurrence of other. Exam.: If one coin is tossed appearance of head affects the appearance of tail or vice-versa.

Theories of Probability

Additional theory: (Dependent event)

When two events , say A and B are mutually exclusive (that the events can not occur simultaneously) the chance of occurrence or probability of occurrence of A and B is a total of occurrence of A and B.

Exam.: 50 times head and 50 times tail, total probability of head and tail is 50 + 50 = 100

Multiple theory : (Independent event)

Probability of two or more independent events occurring together is the product of the probabilities of individual events.

Exam.; Result of Monohybrid cross – ¾ red flower , ¼ white flower

- ¾ tall plants , ¼ dwarf plants

Probability of tall plant with red flower is ¾ X ¾ = 9/16

THE Chi Square Test

- **Statistical method of determining whether the deviation from an expected result is significant or**
- **When we use a statistical test to determine how an observed ratio deviated from an expected ratio , we say we are determining "Goodness of Fit".**
- The test was developed by A.R. Fisher in 1870 and later on used by Karl Pearson in 1900.
- **Formula :**
- $$X2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} \cdots + \frac{(o_n - e_n)^2}{e_n}$$
- **O = Observed frequency e = Expected frequency**

Example : Determine the validity of monohybrid cross

Shape of seed	No. of seeds Observed	No. of seeds expected
Round	5474	5493 (3)
Wrinkled	1850	1831 (1)

$$X2 = \frac{(o - e)^2}{e}$$

- $$X^2 = \frac{(5474 - 5493)^2}{5493} + \frac{(1850 - 1831)^2}{1831}$$
- $X^2 = 0.06 + 0.1971$
- **$X^2 = 0.2571$**

Method to draw inferences :

- First find out Degree of Freedom

- Now find out the table value at either 1% (i.e 0.01)or 5% (0.05) level in chi square table using degree of freedom.

How to refer X^2 table

- Degree of freedom : It can be calculated by using following formula :

  DF = (r – 1) (c – 1) r = row; c = column

  (2 -1) (2 -1) = (1) (1) = 1

At degree of freedom 1 find out the table value at **either 1%(i.e. 0.01) or 5% (0.05) level .**

- At DF 1 at 5% level the tablulated value is 3.84

While calculated value is 0.2571

Method to draw inferences :

- If calculated value of X^2 is higher than tabulated value then result is considered as significant (expected and observed frequencies are different).

- If calculated value of X^2 is less than tabulated value then result is considered as insignificant (expected and observed frequencies are almost in agreement with each other).

Frequency distribution: **In most cases we take large number of observations and as the observation increases it becomes increasingly impractical to digest and understand them all in tabular form. The same data , however, can be grouped in categories or classes and the number of observations falling in each category is counted. Presentation of such condensed information of data is known as frequency distribution. The number occurring in each class is termed the frequency of that class.**

Types of Frequency distributed :

a) Normal distribution

b) Binomial distribution

c) Poisson distribution

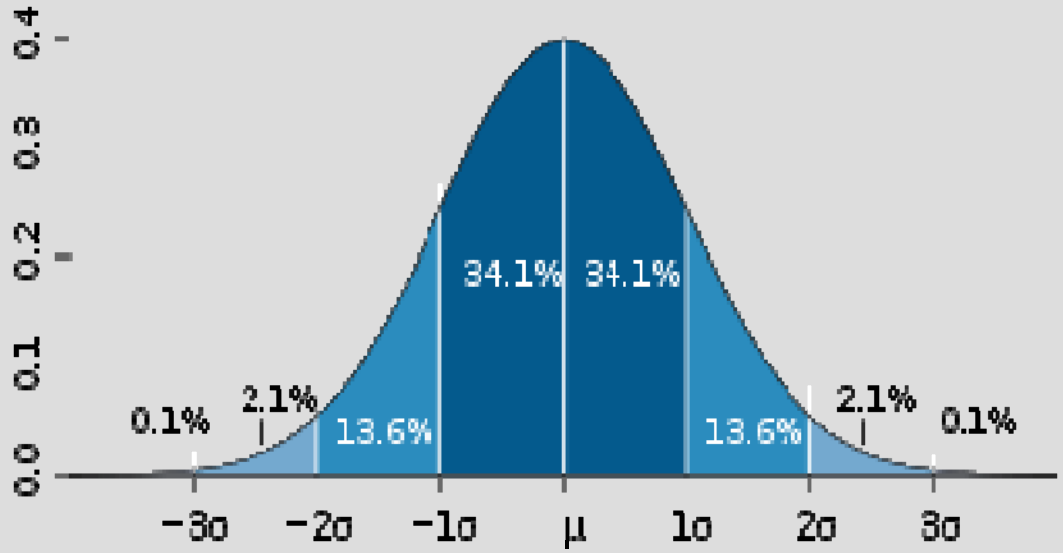
a) Normal distribution:

- **The normal distribution is a continuous curve and it stretches to infinity on both direction. The curve is bell shaped.**
- **The mean value of the variables is in the exact centre of the curve and the largest number of data lie at this point. There are relatively few observations at extremes.**
- **The area between –1S and +1S will include 68.0% of the total area and indicates that 68.0% of the observation lie within a distance equal to the -1S and +1S on both the sides of mean.**
-

-The area between -2S and +2S includes 95% of the observation

- Area from -3S to +3S includes 99.7% of the observation.

- The normal distribution is also known as Gaussian distribution..



Binomial Distribution : The binomial distribution having only two possible outcomes ,each with a known probability is called binomial distribution.

- Many problem in genetics concern not only with the probability that a certain event will occur but also with the probability that a certain combination of events will occur.
- For example it might be of value to determine with what probabilities two offsprings of a mating of Aa X aa will have particular genetic constitutions i.e. both with Aa, both with aa or one with Aa and other with aa ?
- Since the occurrence of any particular genotype in a single offspring is not influenced by the genotype of the other offspring, these are independent events. The probability that 2Aa offsprings will be formed from this mating is, therefore, equal to the product of their separate probabilities.
- Aa = ½ X ½ = ¼ or 25 %

- Thus probabilities for each sequence of two children are as follow:

First child	Second child	Probabilities
Aa	Aa	½ X ½ = ¼
Aa	aa	½ X ½ = ¼
aa	Aa	½ X ½ = ¼
aa	aa	½ X ½ = ¼

Thus the probability that both offsprings are Aa is ¼, that one is Aa and other is aa is 2/4 and that both are aa is ¼.

Both offsprings Aa = ¼

One Aa and other aa = ¼ + ¼ = 2/4

Both aa = ¼

In other words the pattern for this distribution is 1 : 2: 1. This also represents the coefficients of raising two values, bionomial - P and q to the power of square.

$(P + q)^2 = P^2 + 2Pq + q^2$

Or if we substitute Aa for P and aa for q then

$(Aa + aa)^2 = (Aa)^2 + 2\,(Aa)\,(aa) + (aa)^2$

or 1 (Aa) (Aa) + 2 (Aa) (aa) + 1 (aa) (aa)

Poisson distribution:

Binomial expansion will produce a symmetrical distribution around a central value when the two genes or genotypes involved in the expansion are in equal proportion.

Example : If the probability of A =a=1/2 the binomial $(A + a)^2$ will produce

1 AA + 2 Aa + 1 aa

As "n" in the expansion $(A + a)^n$ is increased, more terms are added but the most frequent values are those occupied by genotypes in which equal number of "A" and "a" alleles are present. Other genotypes such as AA or aa are less frequent but are nevertheless "normally" distributed , since their frequencies fall off equally on both sides of the central genotypes.

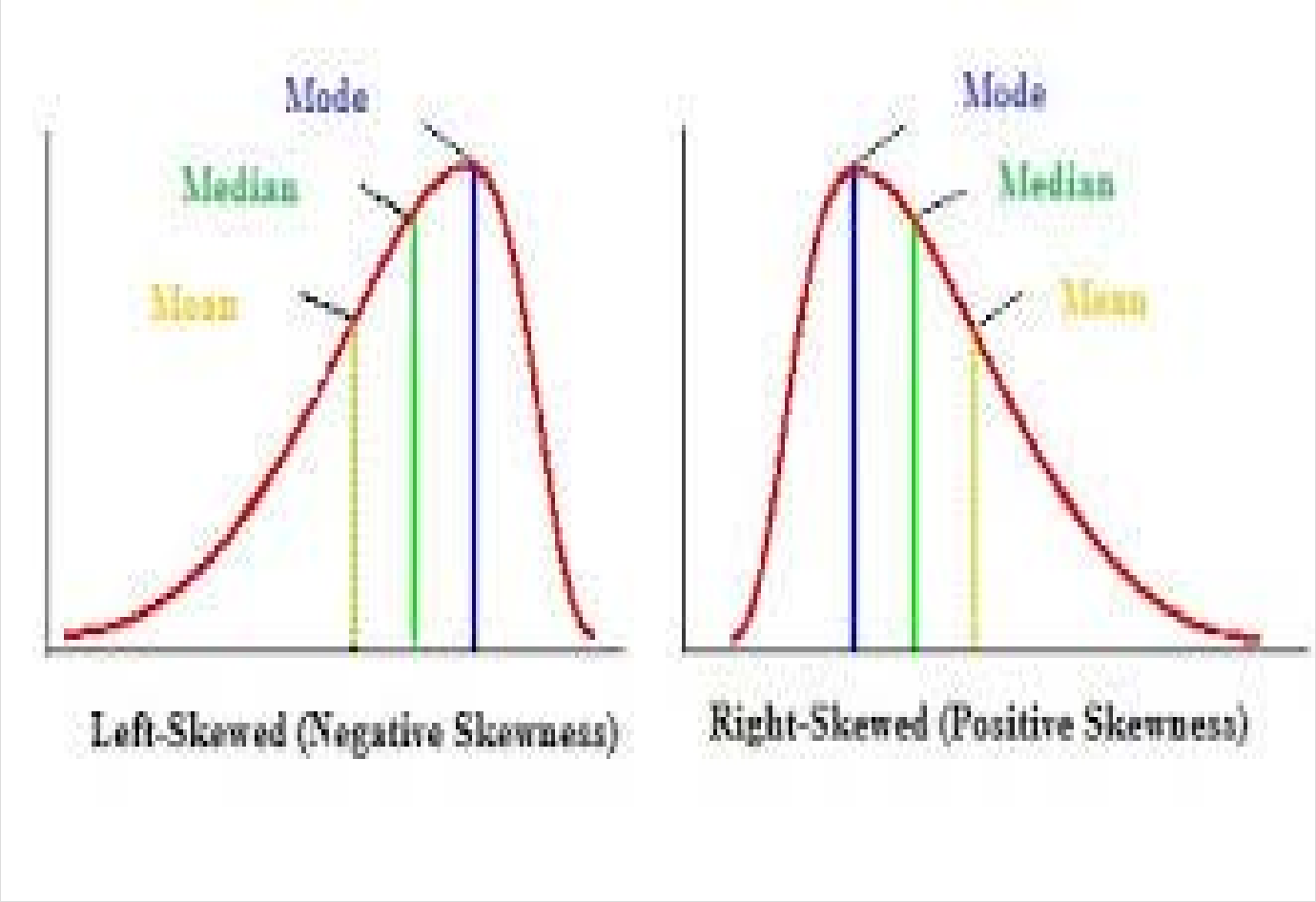
If frequency of "A" is not equal to "a", the normal bell shape of this distribution becomes distorted or skewed with the most common genotypes giving to one side or the other.

Example: If the proportion of "A" is 0.75 and that of "a" is 0.25, the binomial expansion $(0.75 + 0.25)^2$ will produce three genotypes in the following ratio :

1 (0.75) (0.75) + 2 (0.75) (0.25) + 1 (0.25) (0.25)

0.5625 AA + 0.3750 Aa + 0.0625 aa

Skewness


Mode
Median
Mean
Left-Skewed (Negative Skewness)
Mode
Median
Mean
Right-Skewed (Positive Skewness)

Correlation :

Tendency of simultaneous variation between two variables is called correlation.

Methods to study correlation :

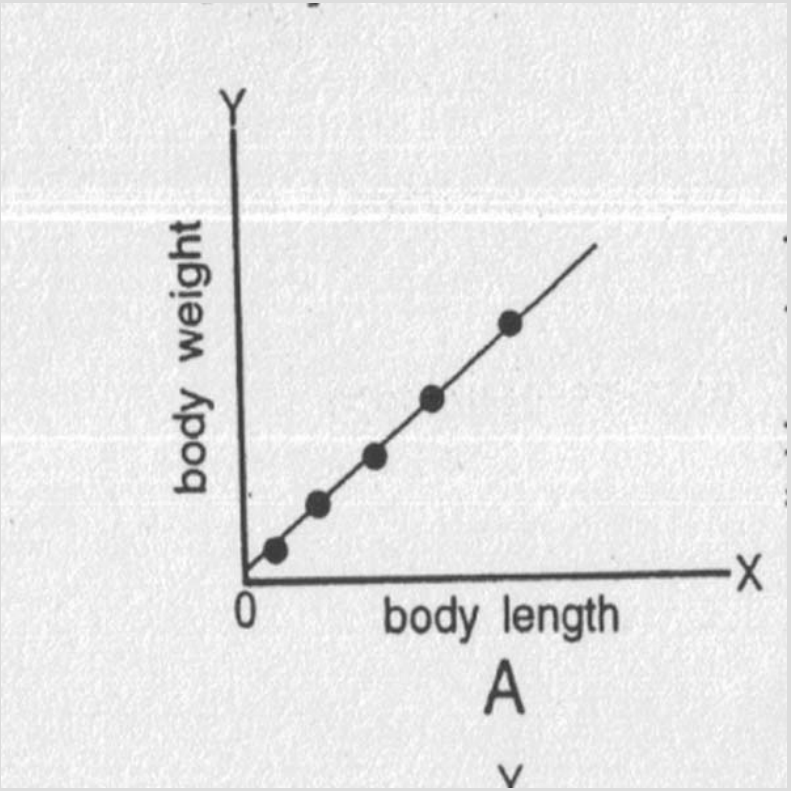
- **Scatter diagram method**
- **Pearson's product moment method**

1.Scatter diagram method:

a. Perfect positive correlation:

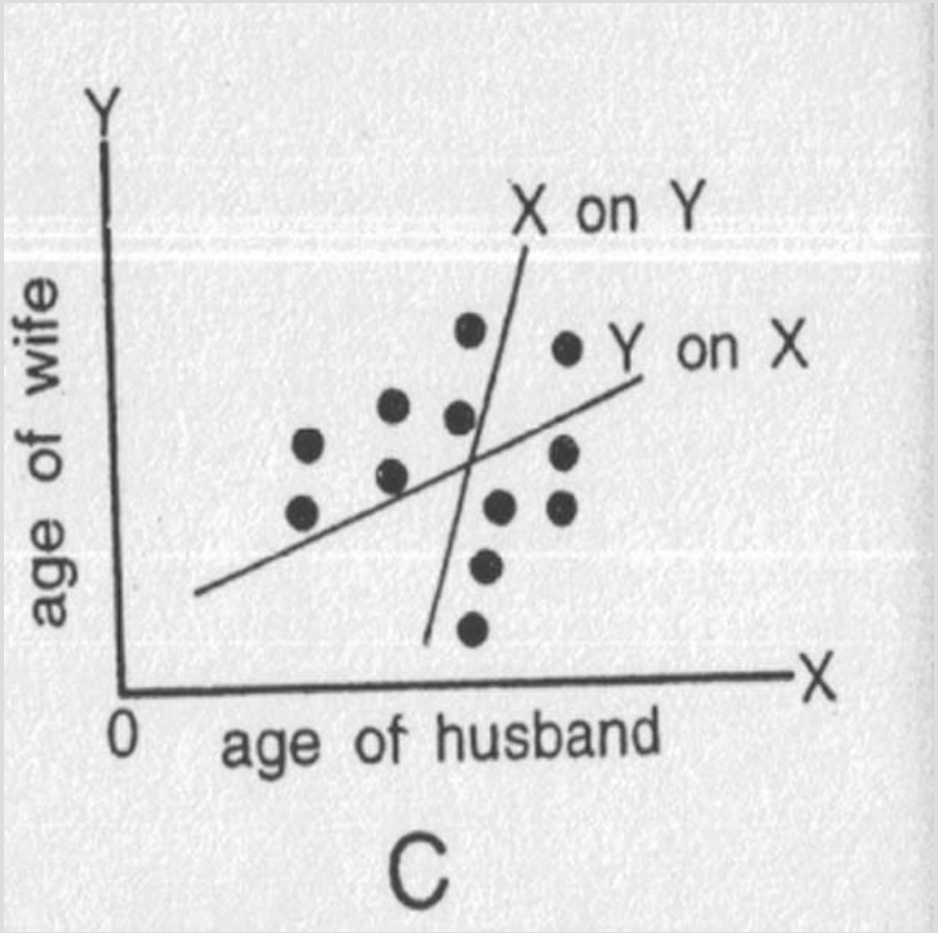
Body length and body weight

Rain and Humidity



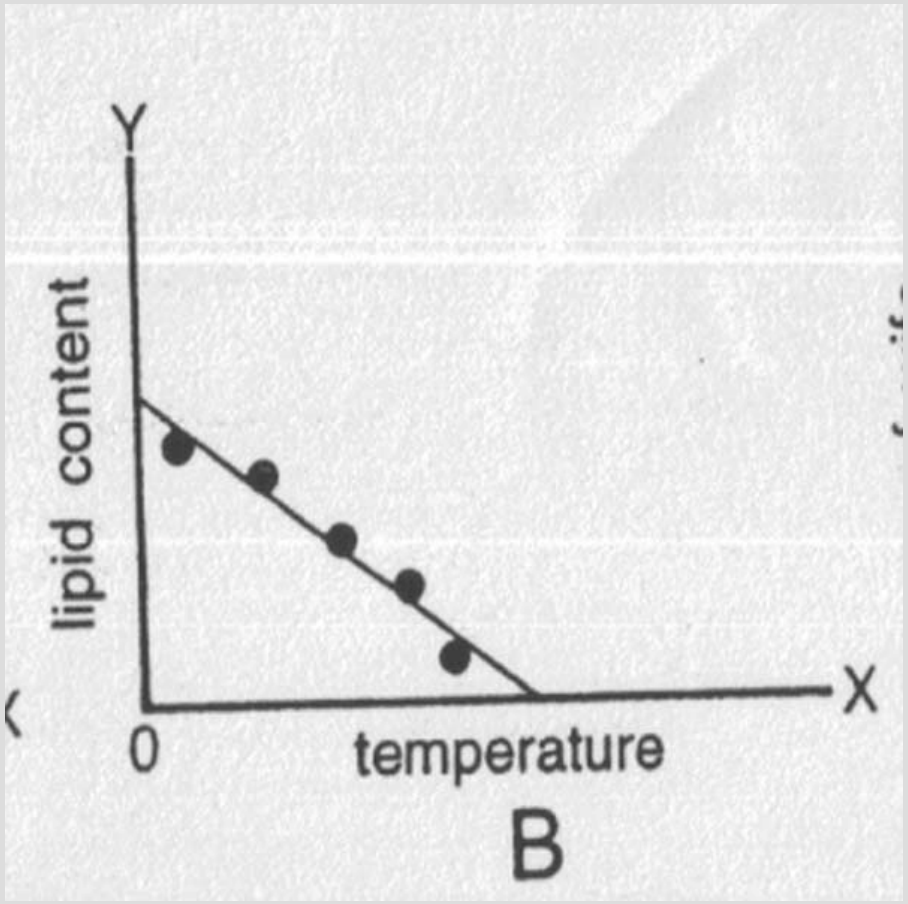
b. Moderate positive correlation:

Age of husband and age of wife



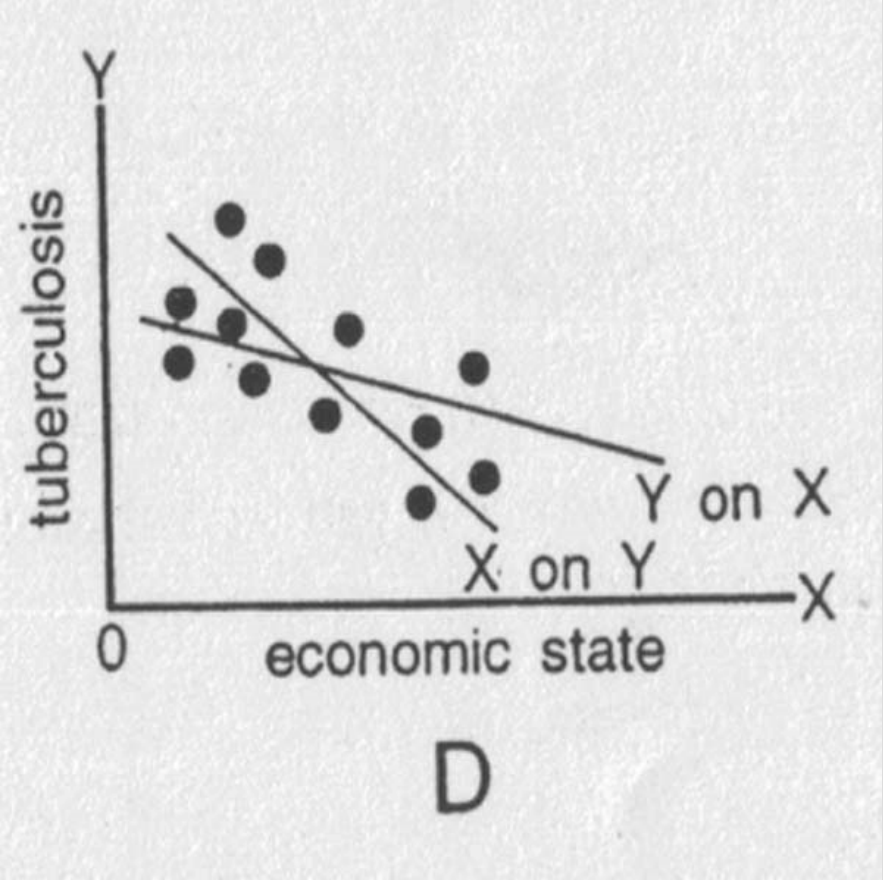
c. Perfect negative correlation:

Lipid content and temperature



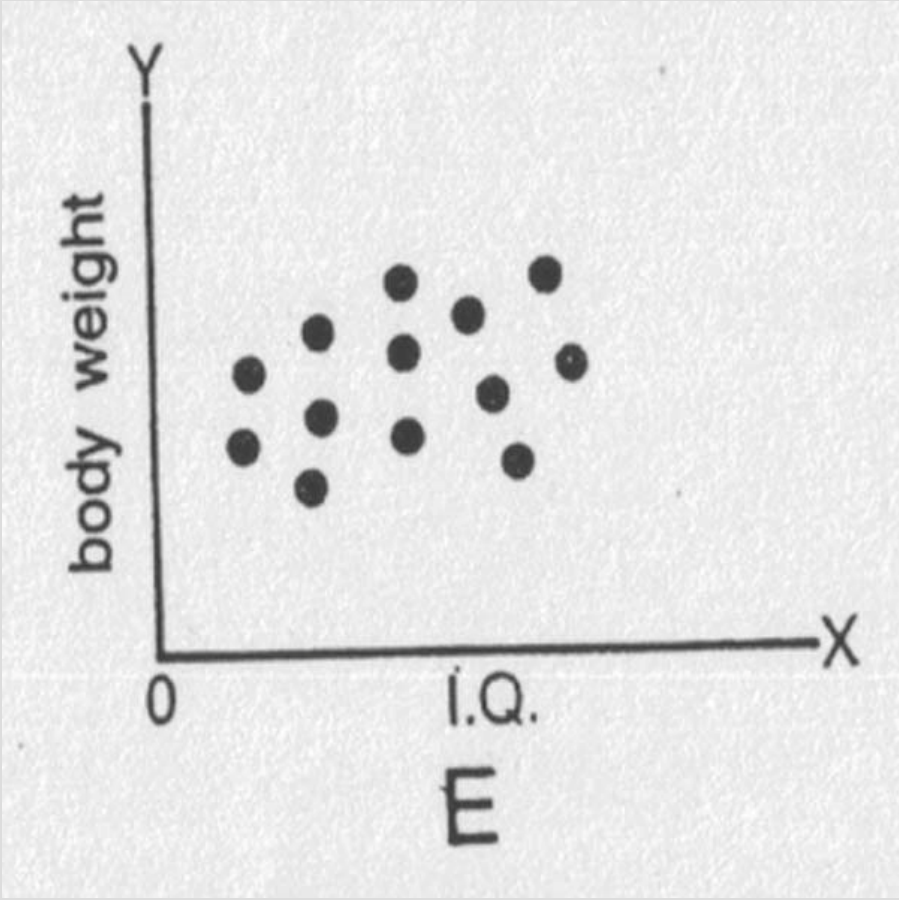
d. Moderate negative correlation

income and mortality rate



e. No correlation

I.Q. and body weight



2: **Pearson's product moment method**

Numerical expression of correlation is called coefficient of correlation.

It can be calculate by using following formula:

X and Y are variables, E indicates sum total

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - (E(X))^2}\sqrt{E(Y^2) - (E(Y))^2}}.$$

Degree of correlation	Positive	Negative
Perfect corr.	+ 1	- 1
Very high degree of corr.	+0.9 or more	-0.9 or more
Sufficient high	+0.75 to 0.9	-0.75 to 0.9
Moderate degree	+0.6 to 0.75	-0.6 to 0.75
Only possibility	+0.3 to 0.6	-0.3 to 0.6
Possibly no corr.	+ 0.3	-0.3
Absence of corr.	00	00

Regression

- Literally meaning – Stepping back (Sir Francis Galton)
- In later half of 19th century Galton studied relationship between height of fathers and their sons and arrived at interesting conclusion :
- 1. Tall fathers have tall sons and short fathers have short sons.
- 2. The mean height of sons of tall fathers is less than mean height of tall fathers
- 3. The mean height of sons of short fathers is more than the mean heights of their fathers.
- Galton concluded that when the height of fathers move above or below the mean height, the height of sons tended to go back or regress.

Regression analysis

- **In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed.**

Regression equations

- $X = \bar{x} + bxy\,(y - \bar{y})$
- $Y = \bar{y} + byx\,(x - \bar{x})$

Regression is a statistical tool with the help of which we are in a position to estimate(predict) unknown value of one variable from known value of another variable.

Find out regression equations from given data

Conc. (mg)	40	80	120	160	200
O.D.	0.40	0.60	1.20	1.60	2.00

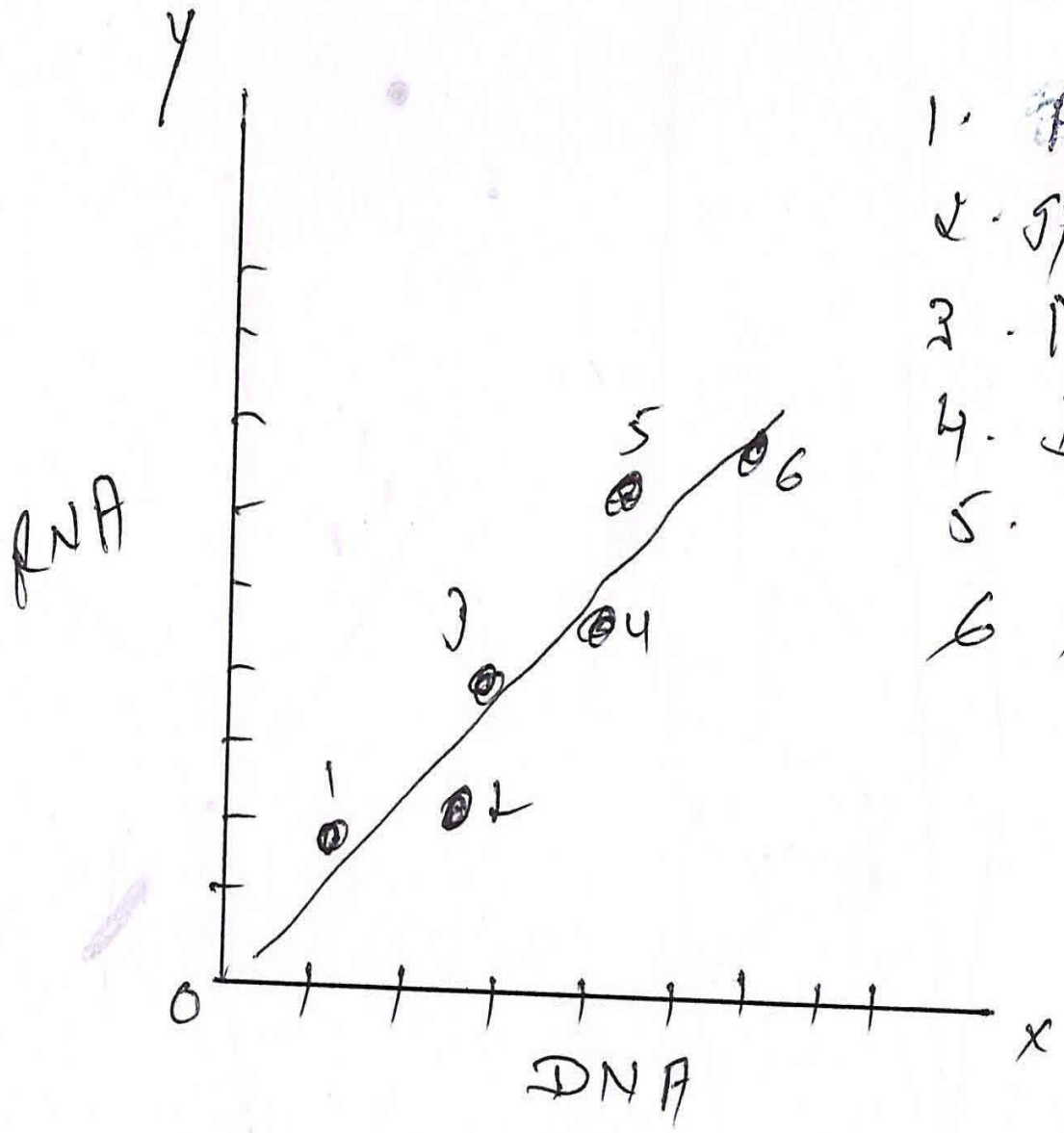
X	dx $(x - \bar{x})$	dx^2	y	dy $(y - \bar{y})$	dy^2	dxdy
40	−80	6400	0.40	−0.80	0.6400	64
80	−40	1600	0.80	−0.40	0.1600	16
120	00	00	1.20	0.00	0.00	00
160	+40	1600	1.60	+0.40	0.1600	16
200	+80	6400	2.00	+0.80	0.6400	64
$\frac{600}{5}$		16000	$\frac{6.00}{5}$		1.6	160

$\bar{x} = 120$

$\bar{y} = 1.20$

$X = 100y$

$y = 0.01X$



1. Archesporial
2. Sporogenous
3. MMC
4. Dyad
5. Tetrad
6. Pollen

Differences between Correlation and Regression

Correlation	Regression
It tests the closeness and direction of relationship between two phenomena It is the measure of co- variability between two variables It indicates the direction and quantity between two variables but do not indicate that the one variable is the cause of other	It measures the nature and extent of relationship, thus enabling us to make prediction It indicates the resultant relationship between independent and dependent variables. It indicates clearly the reason of relationship between two variables

Thank You