STA 1201 : BASIC STATISTICS

STATISTICS: Is defined as the scientific method of collecting organizing, summarizing, presenting and analysing data as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis.

Sometimes statistics is used to mean the data themselves or numbers derived from the data, such as averages. Thus we speak of employment statistics, accident statistics e.t.c.

Statistics can be denided into two

(1) Descriptive statistics and
(2) Inductive or inferential or analytical statistics.

⟹ Descriptive statistics deals with method of describing large masses of data. It involve tabulation, graphical presentation and calculation of summary measures of the set of data. e.g Calculating averages.

⟹ Inductive or inferential or analytical statistics deals with method that enables a conclusion to be drawn from the given data. It is based on probability theory which enables us to make generalization and prediction beyond the collected set of data.

Statistics can only deal with numerical data, often however, data which are of a qualitative nature can be put into a quantitative form.

IMPORTANCE OF STATISTICAL INVESTIGATIONS

It is very important to study statistics for many reasons which are social and economical in nature

For example we need to know the number of student to be admitted so that proper arrangement will be made for their classes, Lecturers, accomodation e.t.c. We need to know the number of infected or sick persons so that proper arrangement for treating them will be made. problems like the one's mentioned above require statistical investigation, therefore it is necessary to study Statistics.

The scope of statistics is very large. It Covers all aspect of human endeavor.

DATA

DATA: These are raw facts and figures collected using a statistical method and techniques

COLLECTION OF DATA

VARIABLES: are those facts that we can describe numerically. They are called quantitative Variates e.g weight, age, height e.t.c.

Variables are classified into two sub-heading namely:

(1) Discrete variables and

(2) Continuous variables

⟹ Discrete variables are variable that take specific values. e.g number of student in a class.

⟹ Continuous variables are variables that take values within a given interval. e.g height, weight e.t.c.

TYPES OF DATA

There are two types of data

(1) primary data and
(2) Secondary data.

⟹ primary data are data collected for a specific purpose and the collector is fully aware of how it has been collected and the processing it has undergone. For example, list of part one student for the purpose of taking attendance.

The sources of primary data are Census and Sample surveys.

⟹ Secondary data are data collected from primary data. These are data extracted from primary data. They are obtained from published sources these include reports, journals, magazines e.t.c. Secondary data can also be obtained from unpublished sources. e.g police report, Hospital records e.t.c.

CENSUS

Census is defined as 100% count of the population of a particular place at a particular time. It is the complete investigation of the whole population items, objects.

SAMPLE SURVEY

Sample Survey is the study of part or fraction of the population under study.

The objectives of Sample Survey are as follows:

(1) It gives hints about the population under study.

(2) Generalization about the population can be made based on the information obtained from the sample.

(3) Some features such as mean, variance etc of the population can be estimated from the Sample Survey results.

PLANNING A CENSUS

When planning a census, the following should be considered

(1) Capital (2) Logistics (3) Feasibility study

(4) Man power.

⇒ Capital: A large sum of money is required in carrying out a census since the whole population under study must be covered. Money is needed for paying the enumerator and other workers

⇒ Logistics: All the necessary materials such as transport facilities, stationaries etc must be provided

⇒ Feasibility study: The total area to be covered by the exercise should be studied and if necessary demarcated into smaller units for easier coverage by enumerators.

A trial census should be conducted to ascertain the problem associated with the exercise so as to correct it before the actual exercise.

⇒ Man power: This include both permanent and temporary staff. These are enumerators, supervisor, facilitators, etc.

TYPES OF SAMPLING

Taking a sample is not simply a matter of taking the nearest items. If a good conclusion relating to the whole population are to be made from the sample, it is essential to ensure as far as possible that the sample is free from bias. i.e free from influence that will affect the reliability of the result

There are two types of sampling:

(1) Random sampling and
(2) Non random sampling.

⟹ A random sampling is a type of sampling in which each item in the population has an equal chance of being included.

Usually all units in the population are numbered then selection are made. The selection can be done by an individual or a computer.

Using a computer to make selection may be the easiest way of selecting a genuine random sample, in practice a table of random numbers is often used. Such table is simply published as a long sequence of random digits set out in blocks for ease of reading.

Example: A sequence of published random numbers runs as follows:

54261, 90067 02374 82816 39210 73829

The sample frame for the survey involved shows a total of 642 items and a random sample of 6 items is required.

This sample can be selected by dividing the random digit into sets of three (3) as

and selecting the first six (6) item thus indicated, ignoring any sets with values above 642, i.e
542/619 006/374 628/163 921/073
Thus numbers 702, 828 and 921 would be ignored leaving items 542. 619 006 374 163 & 073 comprising the random sample. The numbers on the table of random number can be read either vertically, horizontally, diagonally e.t.c.

Example of random sampling are:
(1) Simple random Sampling
(2) Systematic Sampling
(3) Stratified Sampling
(4) Cluster Sampling e.t.c.

⟹ Non-random Sampling is a sample which are generated not on the principle of randomness and no probability is involved. Here the sample are at the discretion of the Sampler. e.g quota Sampling

UNITS

Units are the items of interest such as people household, Cars, book e.t.c.

METHOD OF DATA COLLECTION

Are the techniques used in collecting data. Such techniques includes:
(1) Interview method
(2) Mail questionnaire
(3) Experiment method
(4) Observation method

DESIGNING A QUESTIONNAIRE

Questionnaire is a list of questions and it is often in two parts:

(1) The first part is a classification section: This requires such details of the respondent as sex, age, marital status, name and occupation.

(2) The second part has the question relating to the subject matter of the survey.

In designing a questionnaire, the characteristics below should be considered:

THE CHARACTERISTICS OF A GOOD QUESTIONNAIRE

(1) The questions should be easily understood

(2) The questions should be presented in a systematic and well organised form.

(3) The questions should be capable of having a precise answer.

(4) Questions presented should not be tele-guiding ie should not suggest answers.

(5) The questionnair should not be too long.

POST ENUMERATION SURVEY

Is the carrying out a small survey after the real survey has been conducted.

ADVANTAGES

It serves as a very quick check on some of the short coming of the real survey.

FREQUENCY DISTRIBUTION

When summarizing large masses of the raw data, it is often useful to distribute the data into classes or categories and to determine the number of individual belonging to each class, called the class frequency.

A tabular arrangement of data by classes together with corresponding class frequencies is called a frequency distribution or frequency table. i.e It is a table showing various values of a certain variable together with number of time (frequency) each value of the variable occurs. Example

Table 1. Age of student in Math & stat. dept.

Age group	No of std
10 — 19	10
20 — 29	14
30 — 39	30
40 — 49	8
Total	62.

CLASS INTERVAL AND CLASS LIMITS

A symbol defining a class, such as 10 — 19 above is called a class interval. The numbers 10 and 19 are called class limits; the smaller number 10 is lower class limit and the larger number 19 is the upper class limit. Class interval like 50 and above or 50 and over is an open class interval

CLASS BOUNDARIES

The class boundaries are obtained by subtracting 0.5⁻ from the lower class limit and adding 0.5⁻ to the upper class limit. or true class limit.

For example:

In table 1: $10 - 0.5^- = 9.5^-$ and $19 + 0.5^- = 19.5^-$

$\therefore$ 9.5^- is the lower class boundaries and

19.5^- is the upper class boundaries

The size or width of a class interval is the difference between the lower and upper class boundaries and is also refered to as the class width, class size or class length.

CLASS MARK

Is the midpoint of the class interval and is obtained by adding the lower and upper class limit and dividing by 2. Thus the class mark of the interval 10 — 19 is $(10+19)/2 = 14.5^-$.

GENERAL RULES FOR FORMING FREQ. DISTR.

(1) Determine the largest and smallest numbers in the raw data and thus find the range (the difference between the largest and smallest number).

(2) Divide the range into a convenient number of class intervals having the same size. The number of class intervals is usually between 5 and 20, depending on the data.

#1420

(3) Determine the number of observations falling into each class interval; that is find the class frequency. This is best done by using tally, or score sheet.

Class should not overlap. e.g 0 — 4, 5 — 9, 10 — 14. should be use in preference to 0 — 4 4 — 9, 9 — 14.

When possible the class interval should have a uniform size.

The classes thus obtained are then placed in a column with the lowest class at the top and the rest of the classes following according to size.

Below are marks obtained by 40 students in an examination.

61	65	68	87	74	68	84	73	86
60	73	79	75	73	60	66	78	75
75	94	96	78	82	61	75	79	62
89	97	78	65	80	69	57	88	86
67	73	81	72					

Tabulate the marks in a frequency table.

Solution

(1) Range => 97 − 57 = 40

(2) Class size = 5

The first class interval must contain the least

Observation while the last interval must contain the highest observation.

Class mark	Tally	Frequency
55 — 59		
60 — 64		
65 — 69		
70 — 74		
75 — 79		
80 — 84		
85 — 89		
90 — 94		
95 — 99		
Total		

Example 2: The following are the marks scored by 30 students in a statistics test

7 9 12 21 19 32 23 45 29 35
6 43 15 27 21 33 45 38 34 25
8 26 46 23 42 46 39 22 17 18.

Show the marks in a grouped frequency table using the groups 1 — 10, 11 — 20, ...

UNGROUPED DATA

Example:

Below are number of Cement sold each week by Ashaka Cement plc in January 2018.

Weeks	Frequency (bags)
1	3000
2	2800
3	992
4	1125
	7,917.

RELATIVE FREQUENCY AND CUMULATIVE FREQUENCY DISTN

Relative frequencies are obtained by getting the proportion of a part from the whole. i.e the part is divided by the whole.

Example

Below are number of students offering various subjects.

Subject	No of students	Relative freq.
ENG	350	350/1300
MATHS	150	150/1300
PHY	215	215/1300
CHEM	180	180/1300
BIO	405	405/1300
TOTAL	1,300	

CUMULATIVE FREQUENCY DISTRIBUTION

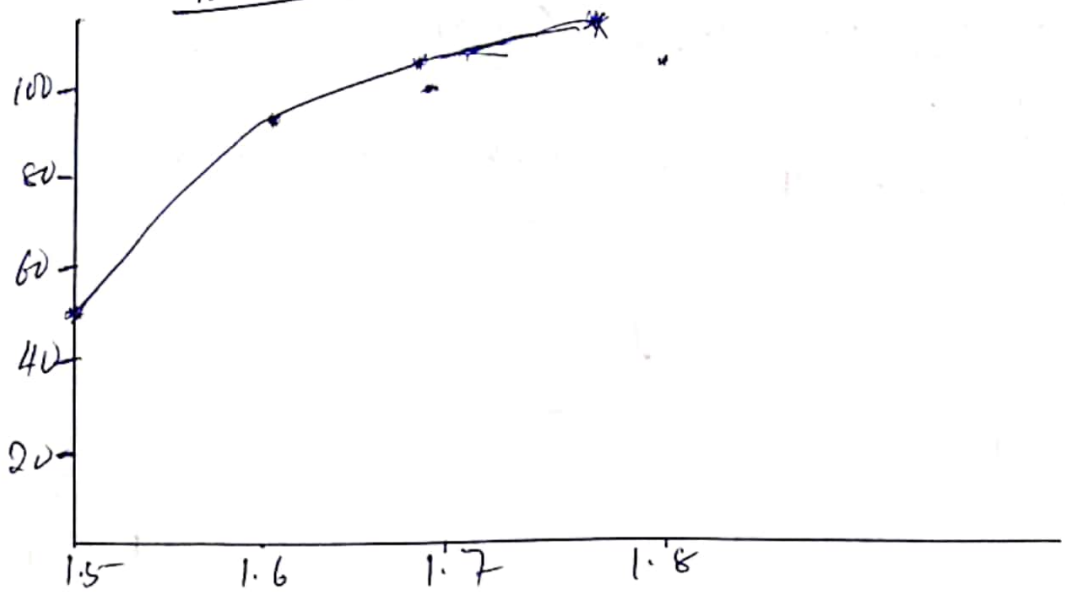
Height	No of Std (Freq.)	Cumulative Freq.
1.50 — 1.59	200	200
1.60 — 1.69	140	340
1.70 — 1.79	30	370
1.80 — 1.89	5	375
	375	

Percentage (%) Cumulative Frequency

$\frac{200}{375} \times 100 = 53.3$ / $\frac{340}{375} \times 100 = 90.7$

$\frac{370}{375} \times 100 = 98.7$ & $\frac{375}{375} \times 100 = 100$.

% C.F Graph



PRESENTATION OF DATA DIAGRAMATICALLY

Statistical data can often be presented by means of a diagram, chart or graphs. This enables relationship, trend and comparison to be grasped more readily.

BAR CHARI

~~Statistical data can often be presented by bar~~

The bar chart is a means of graphical presentation which enables quantities to be compared visually.

Bars are drawn whose lengths are proportional to the magnitude to be presented. When using bar charts, a suitable scale must be chosen and this must be indicated. We have the following bar charts.

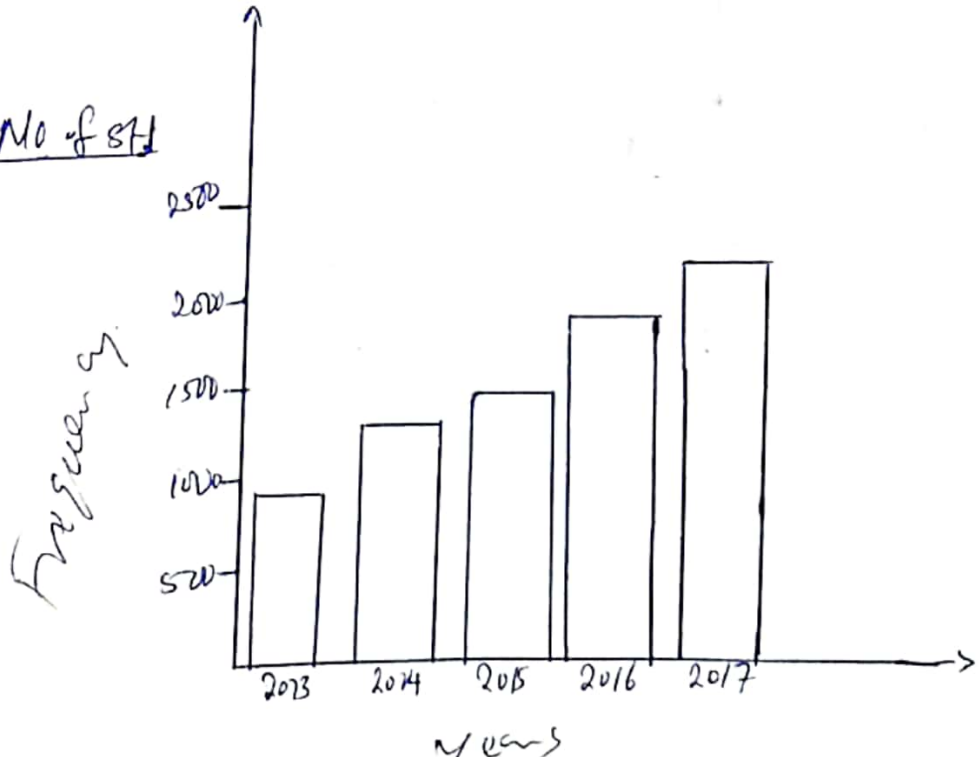
(1) A simple bar chart
(2) A component bar chart
(3) A percentage component bar chart
(4) A Multiple bar chart.

A Simple Bar chart

Example:

A bar chart indicating number of students admitted to YSU.

Year	No of St
2013	950
2014	1200
2015	1450
2016	1800
2017	2100

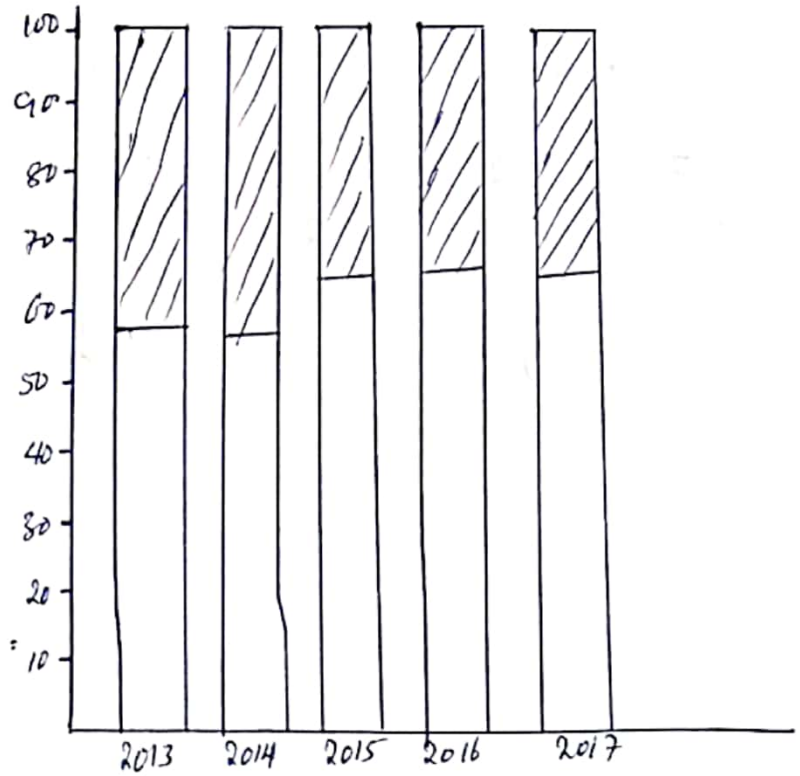


The percentage component Bar - Chart.

In this case the bars are divided in proportion to the percentages that the parts bears to the whole. The scale will be a percentage scale, and all charts will be the same length.

Example:

Year	Male	Female	Total	% Male	% Female	Total %
2013	20	15	35	57	43	100
2014	25	20	45	56	44	
2015	32	18	50	64	36	
2016	45	25	70	64	36	
2017	50	30	80	63	37	
Total						



Key

Female

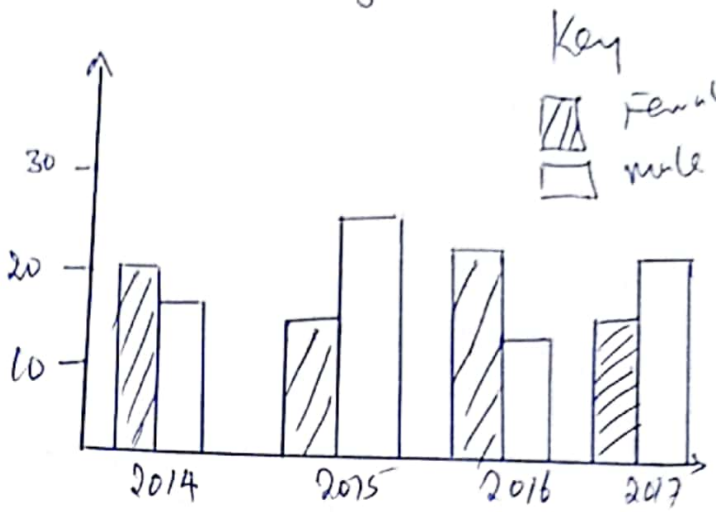
Male

Multiple Bar - Chart.

This chart groups two or bar charts together

Example

Year	Female	Male
2014	20	18
2015	16	25
2016	22	14
2017	15	20

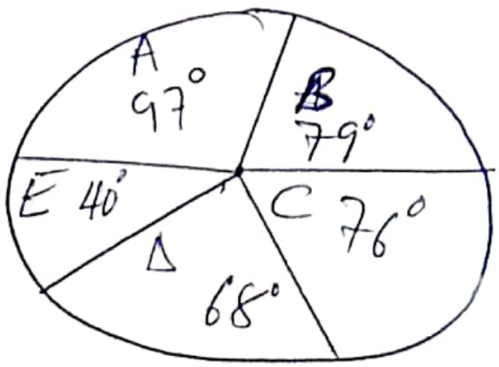


THE PIE CHART

The pie chart is circular presentation of data which is similar to the bar chart. It represent data in degrees.

Example.

School	No of students	Relative Frequency	Angles in degrees
A	35	35/130 = 0.27	0.27 x 360 = 97.2 ≃ 97
B	29	29/130 = 0.22	0.22 x 360 = 79.2 ≃ 79
C	27	27/130 = 0.21	0.21 x 360 = 75.6 ≃ 76
D	25	25/130 = 0.19	0.19 x 360 = 68.4 ≃ 68
E	14	14/130 = 0.11	0.11 x 360 = 39.6 ≃ 40
Total	130		360.

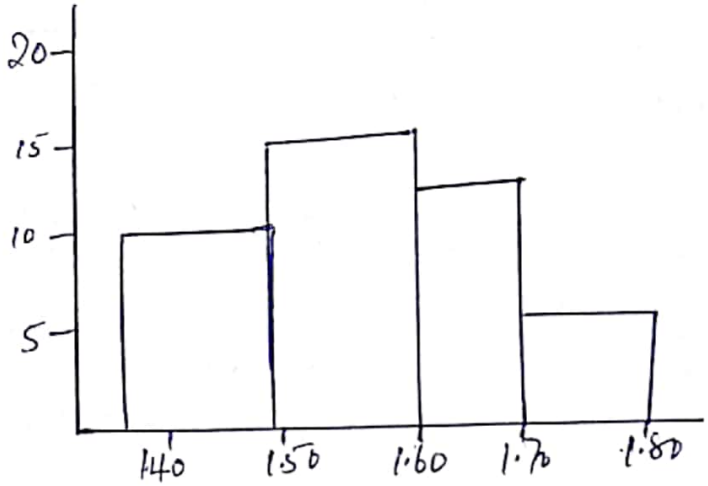


HISTOGRAM

This is a diagramatical representation of data where a series of rectangles having a base measured interval and an area proportional to the frequency

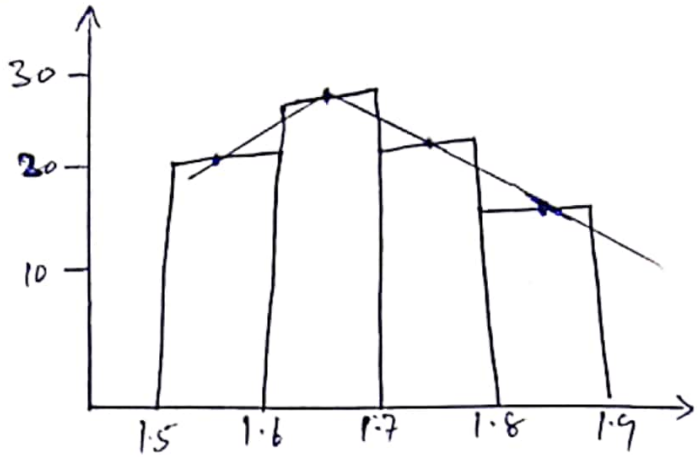
Example

Height	Number of Students
1.40 — 1.50	10
1.50 — 1.60	15
1.60 — 1.70	12
1.70 — 1.80	6



FREQUENCY POLYGON

Where the class intervals of a frequency distribution are equal. It can also be represented by means of a frequency polygon. This is obtained by joining the midpoints of the tops of the rectangle of the histogram.



PICTOGRAM

In this case small symbols or simplified pictures are used to represent the data. There are important rules to follow if pictorial charts are to fulfil their functions. The rules are:

1. The symbols must be simple and clear.
2. The quantity each symbol represents should be given.
3. Large quantities are shown by a great number of symbols and not by larger symbols.

MEASURES OF LOCATION AND DISPERSION

A measure of location of a distribution is a value which is typical or representative of the data. The measure of location will give the central value.

The following are measures of location.

1. Mean:
   - (a) Arithmetic mean.
   - (b) Geometric mean and
   - (c) Harmonic mean.
2. Median
3. Mode.

MEAN

=> (a) Arithmetic mean.

If the data are: $x_1, x_2, x_3 \cdots x_n$

then arithmetic mean is defined as

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad i \text{ is any integer}$$

$$= \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}.$$

Example: Calculate the arithmetic mean of the set of numbers given as: 2, 7, 8, 4, 3, 9, 5, 1, 5, 6.

(2) If the different number $x_1, x_2 \cdots x_n$ have frequencies $f_1, f_2, \cdots f_n$ respectively, then their arithmetic mean is given by:

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

Relationship Between The Three.

Mean − Mode = 3(mean − median).

(3) Given:

x	f	fx
1	5	5
2	3	6
3	2	6
4	4	16
5	2	10
Total	16	43

$\therefore \bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{43}{16}$

$= 2.7.$

The arithmetic mean of grouped data are obtained by first determining the class marks or midpoints of each class.

(4) Given

Mark (x)	Frequency
10 — 19	1
20 — 29	7
30 — 39	8
40 — 49	15

Marks	Class Marks (x)	Frequency f	fx
10 – 19	14.5	1	14.5
20 – 29	24.5	7	171.5
30 – 39	34.5	8	276.5
40 – 47	44.5	15	667.5

$$\therefore \bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{1,130}{31} = 36.5$$

(b) Geometric Mean.

If the data is $x_1, x_2, x_3 \ldots x_n$

the geometric mean

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n}$$

= n^{th} root of the product of data or observation

The geometric mean of number 2, 4 and 8 is $\bar{x}_G = \sqrt[3]{2 \times 4 \times 8} = \sqrt[3]{64} = 4$.

(c) Harmonic Mean.

The harmonic mean $\bar{x}_H$ of N numbers $x_1, x_2 \ldots x_n$ is defined as

$$\bar{x}_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

Example: The harmonic mean of the number 2, 4 and 8 is

$$\bar{x}_H = \frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{8}} = \frac{3}{7/8} = 3.43.$$

MEDIAN

The median is that ~~data~~ value ~~of~~ ~~the~~ variables which divides the group in two equal parts. The ~~part~~ first part are less than the value while the ~~Second~~ second part is greater than the value.

Example:

If the data are 5, 2, 6, 3, & arrange the data in order of size.

2, 3, (5), 6, 8.

median = 5

If the number of values is even add the two central values and divide by two.

example:

If the values are 2, 3, 7, 9, 10, 11. the

median is $\frac{7+9}{2} = \frac{16}{2} = 8$.

For grouped data, the median can be obtained by using the formula.

$$\text{median} = L_1 + \left(\frac{\frac{N}{2} - (\Sigma f)_1}{f_{\text{median}}}\right) C$$

where

L_1 = Lower class boundary of the median class

N = Number of items in the data (total frq.)

$(\Sigma f)_1$ = Sum of frequency of all values lower than the median class

f_{median} = frequency of the median class.

C = size of the median class interval

$\frac{N}{2}$ = median class.

THE MODE

The mode of a set of number is that value which occurs with the greatest frequency - i.e the value with the highest frequency

Example

x	2	4	5	7
frequency	3	6	2	4

∴ 4 is the mode of the given data.

$$\text{Mode} = L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)C$$

QUARTILE, DECILES AND PERCENTILE

Quartiles: The values are divided into four equal parts. These values are denoted by Q_1, Q_2, and Q_3 are called first, second and third quartiles respectively.

Decile: The values are divided into ten (10) equal part and are denoted by $D_1, D_2, \cdots D_9$.

Percentiles: Divide the data into 100 equal part and are denoted by $P_1, P_2, \cdots P_{99}$.

Collectory quartiles, deciles and percentile and other values obtained by equal sub-division of the data are called quantiles.

Quantiles: Arrange the data in order of size then the $\frac{n^{th}}{4}$, $\frac{n^{th}}{2}$ and $\frac{3n}{4}$ data values are the quartiles.

Deciles: The first, second, third, ... Ninth deciles are obtained by counting $\frac{N}{10}$, $\frac{2N}{10}$, ... $\frac{9N}{10}$ of the distribution respectively

percentiles: $\frac{\Sigma f_i + 1}{100}$, $\frac{2(\Sigma f_i + 1)}{100}$ $\frac{3(\Sigma f_i + 1)}{100}$ ... $\frac{99(\Sigma f_i + 1)}{100}$

MEASURES OF DISPERSION

Measures of dispersion include: Range, variance, standard deviation, interquartile range, Semi-interquartile range and the mean deviation.

RANGE: The range of a set of numbers is the difference between the greatest and smallest value. Example the range of the numbers 2, 5, 7, 3 4 is $7-2=5$. and for 5, 3, 4, 9, -8, 10, 6, is $10-(-8)=10+8=18$

VARIANCE: If the values are $x_1, x_2, \cdots x_n$ then the variance is

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$ and the standard deviation is equal to the square root of the variance

Therefore

STANDARD DEVIATION: $$\sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} = \sigma$$

Example:

Find the standard deviation of the set

2, 3, 6, 8, 11

Solution

The arithmetic mean $= \bar{x} = \frac{2+3+6+8+11}{5} = \frac{30}{5} = 6.$

$\therefore$ Variance $= \sigma^2 = \frac{(2-6)^2+(3-6)^2+(6-6)^2+(8-6)^2+(11-6)^2}{5}$

$= \frac{(-4)^2+(-3)^2+(0)^2+2^2+5^2}{5}$

$= \frac{16+9+0+4+25}{5} = \frac{54}{5}$

$= 10.8$

$S.D = \sqrt{\text{Variance}} = \sqrt{10.8} = 3.286.$

From frequency table.

x	f	fx	x^2	fx^2
2	5	10	4	20
4	3	12	16	48
5	2	10	25	50
7	4	28	49	196
	14	60		314

Arithmetic Mean $= \bar{x} = \frac{\sum_{i=1}^{n} f x_i}{\sum_{i=1}^{n} f_i} = \frac{60}{14} = 4.29$

Variance $= \sigma^2 = \frac{\sum_{i=1}^{n} f_i x_i^2}{\sum_{i=1}^{n} f_i} - \bar{x}^2 = \frac{314}{14} - (4.29)^2$

$= 22.43 - 18.40$

$= 4.03$

∴ Standard deviation = $\sqrt{\text{Variance}} = \sqrt{4.03} = 2$

From group data

Marks	F	Mid point x	fx	x^2	fx^2
1 — 10	5	5.5	27.5	30.25	151.25
11 — 20	3	15.5	46.5	240.25	720.75
21 — 30	2	25.5	51.0	650.25	1300.5
31 — 40	1	35.5	35.5	1260.25	1260.25
	11		160.5		3432.75

Arithmetic Mean = $\frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{160.5}{11} = 14.6$

∴ Variance = $\frac{\sum_{i=1}^{n} f x_i^2}{\Sigma f} - \bar{x}^2 = \frac{3432.75}{11} - (14.6)^2$

$= 312.07 - 213.16$

$= 98.91$

Standard deviation = $\sqrt{\text{Variance}} = \sqrt{98.91} = 9.95$

MEAN DEVIATION

Is the define as = $\sum_{i=1}^{n} \frac{|x_i - \bar{x}|}{n}$

Example.

Find the mean deviation of 1, 7, 5, 3, 4

$\bar{x} = \frac{\Sigma x}{n} = \frac{1+7+5+3+4}{5} = \frac{20}{5} = 4$

x	$x-4$	$\lvert x-4 \rvert$
1	-3	3
7	3	3
5	1	1
3	-1	1
4	0	0
Total		8

$\therefore$ Mean deviation $= M.D = \frac{8}{5} = 1.6.$

Inter-Quartile Range $= 3^{rd}$ quartile $- 1^{st}$ quartile

Quartile deviation is some time called semi-inter quartile range.

Quartile Deviation $= \frac{1}{2}(3^{rd}$ quartile $- 1^{st}$ quartile$)$

$= \frac{1}{2}(Q_3 - Q_1)$ or

Semi Interquartile Range $= \frac{1}{2}(Q_3 - Q_1)$.

Coefficient of Variation $= \frac{\text{Standard deviation}}{\text{Mean}}$.

PROBABILITY

If an experiment is Conducted, the set of all possible outcomes is Called the sample space denoted by the Symbol S. ie the Universal set is Called the sample Space S

Any element of the sample space is Called an Event Any Subset of the sample space is also Called an event. ie is represented by the Symbol A. ie Simply an occurance

The probability of an event A if the Sample space is S is defined as

$$P(A) = \frac{n(A)}{n(S)}$$

Example

If a Coin is tossed once, the set of all possible outcome = Sample space equal to (Head up, Tail up) ie (H, T).

$$P(\text{Head up}) = P(H) = \frac{n(H)}{n(S)} = \frac{1}{2}$$

$$P(\text{Tail up}) = P(T) = \frac{n(T)}{n(S)} = \frac{1}{2}$$

PROPERTIES OF PROBABILITY

(1) $P(A)$ is all possitive or zero
ie $P(A) \geq 0$ for all A Subset of S

(2) $P(S) = \frac{n(S)}{n(S)} = 1$
ie Probability of the Sample Space equal to 1 (Longest value)

(3) $P[\,] = 0$
i.e probability of the empty set is zero

(4) If $A = B$
Then $P(A) = P(B)$
i.e if two set are identical then, their probabilities are equal

(5) If two events are independent, then
$P(A \cap B) = P(A) \times P(B)$

(6) If two events are Mutually exclusive then
$P(A \cup B) = P(A) + P(B)$.

MUTUALLY EXCLUSIVE EVENTS (Addition of Prob.)

Two events are said to be mutually exclusive, if the one occur the other cannot. ie the two events cannot occur together. The word such as "or", "either", "neither" are used

EQUALLY LIKELY EVENTS

These are events which have the same chance or probability of occuring. e.g event of tossing a head or a tail in a toss of a fair coin.

INDEPENDENT EVENTS (Multiplication of Prob.)

These are events that are in no way affected by the occurance of each other. ie both of them can occur. Example if a die is thrown and a coin is tossed, the probability of a head being tossed is unaffected by the outcome of the die throw, and vice versa. Word "and", "both", "all" are used.

If two coins are tossed one each, the sample space is the set of all possible out come.
$= HH, HT, TH, TT,$ T – tail up.
H – Head up

BINOMIAL DISTRIBUTION

If the random variable has two values (i.e Success and failure) S & F with $P(S) = P$ and $P(F) = 1 - P = Q$ on run the random experiment n times & count the number of successes $x = K$ then:

$$P(x = K) = {}^{n}C_{K}\, P^{K}\, Q^{n-K}$$

$$K = 1, 2, - - - n.$$

$$P(x = K) = \frac{n!}{(n-K)!\,K!} \cdot P^{K} \cdot Q^{n-K}$$

Example

(1) Toss a balanced Coin two (2) time ie $n = 2$
find the probability of getting
(i) $P(x = 0)$ (ii) $P(x = 1)$ (iii) $P(x = 2)$
where $x = K$ = number of heads.

(2) Sample space: {TTT, TTH, THT, HTT, HHT, HTH, THH, HHH}
Find $P(HTT, THT, TTH)$

(2) Toss a balanced Coin 3 times. In each toss
$H = S$ & $T = F$
$P(H) = P(S) = \frac{1}{2} = P$
$P(T) = P(F) = \frac{1}{2} = 1 - P$
Find (1) $P(TTT)$
(2) $P(HTT, THT, TTH)$
(3) $P(HHT, HTH, TTH)$

(3) A die is rolled five times. Determine the probability of obtaining three Sixes.

Here $n = 5$, $x = 3$, $p(6) = 1/6$ & $q = 1 - \frac{1}{6} = 5/6$

(4) Twenty percent of items produced on a machine are outside stated tolerances. Determine the probability distribution of the number of defectives in a pack of five items

∴ The probability distribution is the set of probabilities for $x = 0, 1, 2, 3, 4, 5$ successes. ie defective in this case. we have $p = 0.2$ $q = 0.8$.

Bernoulli Trial

When $n = 1$, e.g Tose a Coin once, B, G S, F

If we have a bernoulli trial then

$p(S) = p$, $p(F) = 1 - p = q$

$p(x = k) = {}^{1}C_k \, p^k q^{1-k}$ $k = 0, 1$

(5) A balance Coin tossed 5 times what is the probability of getting:

(i) 2 head (ii) 3 head (3) 5 head.

(6) In a class of 100 std 40 of which one girl. A random Sample of 4 is taken. what is the probability that in the Sample Their are (0 girl), two girl 1 girl, 3 girl and 4 girl

MEAN & STANDARD DEVIATION OF BINOMIAL DIST

$\text{Mean} = np$

$\text{S.D} = \sqrt{npq}$.

Two fair die are tossed, the sample space is

	1	2	3	4	5	6
1	1,1	1,2	1,3	1,4	1,5	1,6
2	2,1	2,2	2,3	2,4	2,5	2,6
3	3,1	3,2	3,3	3,4	3,5	3,6
4	4,1	4,2	4,3	4,4	4,5	4,6
5	5,1	5,2	5,3	5,4	5,5	5,6
6	6,1	6,2	6,3	6,4	6,5	6,6

first die x_1 and second die x_2

Binomial Distribution

It has the following assumption

1. There are only two possible outcome for each trials arbitrarily called success and failure
2. The probability of success is the same for each trial
3. There are n trials where n is constant.
4. The n-trial are independent.

CORRELATION ANALYSIS

The Correlation is a Statistical tool which studies the relationship between two variables and Correlation analysis includes various methods and techniques used for studying and measuring the extend of the relationship between the two variables.

TYPES OF CORRELATION

(1) positive Correlation
(2) Negative "

If the values of the two variables deviate in the same direction i.e, if the increase in the values of one variable results, on an average, in a corresponding increase in the values of the other variable or if a decrease in the values of one variable results, on an average, in a corresponding decrease in the values of the other variable, then the Correlation is said to be positive or direct.

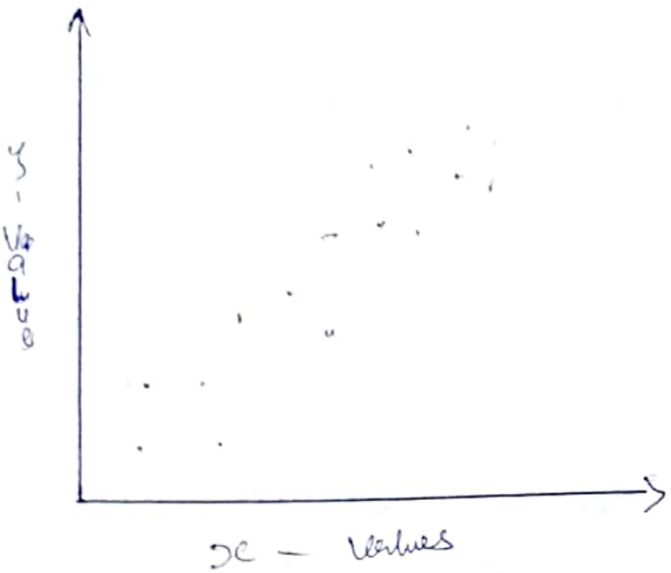
On the other hand, Correlation is said to be negative or inverse if the variables deviate in the opposite direction i.e, if the increase (decrease) in the values of one variable result on the average, in a corresponding decrease (increase) in the values of the other variable.

METHODS OF STUDYING CORRELATION

The commonly used methods for studying the Correlation between two variables are:

(1) Scatter diagram method
(2) Karl pearson's Coefficient of Correlation
(3) Rank Correlation Method
(4) Concurrent deviations Method
(5) Two-way frequency table

SCATTER DIAGRAM



=> Scatter diagram enables us to obtain an approximate estimating line or line of best fit.

=> Scatter diagram only tell us about the nature of the relationship between the two variables.

KARL PEARSON's COEFFICIENT OF CORRELATION

This is a measurement of the degree of relationship b/w the variable. It will vary between +1 and -1. There are different measures of correlation but the most generally use is one called pearson product moment correlation coefficient. The symbol "r" is used as the pearson coefficient of correlation. and is defined as:

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

From Example

From the following table calculate the r.

(1)

X	6	2	10	4	8
Y	9	11	5	8	7

(2)

X	Y
2	60
5	100
4	70
6	90
3	80

2

LINEAR REGRESSION ANALYSIS

Regression analysis, in the general sense, means the estimation or prediction of the unknown value of one variable from the the known value of the other variable. It is specially used in business and economics to study the relationship between two or more variables that are related.

It is also defined as the mathematical measure of the average relationship between two or more variables interms of the original units of the data.

In the regression analysis there are two types of variables.

— Dependent variable: the variable whose value is to be predicted.

— Independent variable: Variable which influences the values or is used for prediction.

In regression analysis independent variable is also known as regressor or predictor or explanator while the dependent variable is also known as regressed or explained variable.

LINES OF REGRESSION

Line of regression is the line which gives the best estimate of one variable for any given value of the other variable ~~In case of two variable x and y~~

The term best fit is interpreted in accordance with the principle of least squares method.

The equation of the line of regression of y on x is:

$$y = a + bx \quad -- \quad (1)$$

Without proof the value of the "a" and "b" in the (1) will be calculated by solving the following equations simultaneously:

(1)

3 $\Sigma y = na + b\Sigma x$ - - (1)

$\Sigma xy = a\Sigma x + b\Sigma x^2$ - - (ii)

$\Rightarrow \bar{y} = a + b\bar{x}$

$a = \bar{y} - b\bar{x}$ - - (2)

$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} \quad - - (3)$$

Examples

(1) From the following table obtain the regression equation

Sale (x)	91	97	108	121	67	124	51	73	111	57
purchases (y)	71	75	65	97	70	91	39	61	80	47

(2) The adjoining table shows the number of motor registrations in a certain territory for a term of 5 years and the sale of motor tyres by a firm in that territory for the same period.

Find the regression equation to estimate the sale of tyres when motor registration is 850.

year	Motor Regn (x)	No. of Tyres sold (y)
1	600	1250
2	630	1100
3	720	1300
4	750	1350
5	800	1500

RANK CORRELATION

Sometimes we come across statistical series in which the variables under consideration are not capable of quantitative measurement but can be arranged in serial order. This happen when we are dealing with qualitative characteristics such as honesty, beauty, character, morality etc. which cannot be measured quantitatively but can be arranged serially. In such situation Karl pearson's coefficient of correlation cannot be used.

Charles Edward spearman developed a formula called Spearman's rank correlation coefficient. usually denoted by ρ (Rho) and is defined as

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

where d is the difference between the pair of ranks of the same individual in the two characteristics and n is the number of pairs.

Example: From the following data Calculate the rank Correlation Coefficient after making adjustment for tied ranks

x	48	33	40	9	16	16	65	24	16	57
y	13	13	24	6	15	4	20	9	6	19

(2)

Std	A	B	C	D	E
Marks	70	80	40	45	5
STA	60	70	80	65	45