





































### STATISTICAL DESCRIPTION OF DATA



## Introduction of Statistics

- The word statistics has been derived from the word.
- Latin Status
- Italian Statista
- German Statistik
- French Statistique



#### **HISTORY OF STATISTICS**

- Kautilya
- famous book 'Arthashastra'
- during Chandragupta's reign
- in the fourth century B.C
- Ain-i-Akbari written by Abu Fazl
- . Referring to Egypt,
- the first census was conducted by the Pharaoh during 300 B.C. to 2000 B.C.



Kautilya's Arthashastra



# Meaning (Definition) of statistics

- i) Singular Sense : (Method)
- ii) Plural Sense: (Data)

 (i) Singular Sense:Scientific method that is employed for collecting; analysing and presenting data, Leading finally to drawing statistical inferences

(ii) Plural Sense:- Data qualitative as well as quantitative.

# Use of statistics

- public services,
- defence
- banking,
- insurance sector
- tourism and hospitality,
- police and military etc



# **Application of Statistics:-**

- (i) Economics
- (ii) Business Managements
- (iii)Industry
- (iv) Commerce







## ECONOMICS

- Time Series Analysis,
- Index Numbers,
- Demand Analysis



- Conducting socio-economic surveys and analysing the data derived from it
- Regression analysis,
- for making future projection of demand of goods, sales, prices, quantities etc.

# **BUSINESS MANAGEMENT**





#### Statistics in Commerce and Industry

- Data on previous sales
- raw materials,
- wages and salaries,



- products of identical nature of other factories etc
- collected, analysed and experts are consulted in order to maximise profits

### **Limitations of Statistics**

- 1. Do not study qualitative phenomenon such as Beauty, Honesty, poverty etc.
- 2.It deals with groups and not with individuals.
- 3. Statistical laws are not exact. Statistical results are true only on average.
- 4.It can be missued.

### VARIABLE

#### DISCRETE

 a variable assumes a finite or a countably infinite number of isolated values, it is known as a discrete variable.

#### Examples

- number of petals in a flower,
- the number of misprints a book contains
- the number of road accidents in a particular locality and so on.

#### **CONTINUOUS**

- it can assume any value from a given interval.
  Examples
- height,
- weight,
- sale
- , profit .

# COLLECTION OF DATA

• (a) Primary;

The data which are collected for the first time by an investigator or agency are known as primary data

• (b) Secondary

if the data, as being already collected, are used by a different person or agency

# **Collection of Primary Data**

- (i) Interview method;
- (ii) Mailed questionnaire method;
- (iii) Observation method;
- (iv) Questionnaries filled and sent by enumerators

#### Interview method

- (a) Personal Interview method,
- (b) Indirect Interview method
- (c) Telephone Interview method

#### personal interview method

- the investigator meets the respondents directly and collects the required information.
- In case of a natural calamity like a super cyclone
- or an earthquake or an epidemic like plague,



#### Indirect Interview method

- If there are some practical problems in reaching the respondents directly,
- as in the case of a rail accident
- investigator collects the necessary information from the persons associated with the problems



#### **Telephone interview method**

the relevant information can be gathered by the researcher himself by contacting the interviewee over the phone.


#### Mailed questionnaire method

 framing a well-drafted and soundly-sequenced questionnaire covering all the important aspects of the problem under consideration and sending

1. Have you he	eard of this test (through a	family member or friend?)
a) Yes		
b) No		
2. Have you do	one this test before?	
a) Yes		
b) No		
3. What is your	biggest fear regarding thi	s test?
a)	Shortness of breath	
ь)	Asphyxia	
c)	Death	
d)	Results of the test	

Rev Port Pneumol. 2012;18:226-32

#### observation method

 data are collected, as in the case of obtaining the data on the height and weight of a group of students, by direct observation or using instrument.



# Questionnaries filled and sent by enumerators.

- Enumerators collects information directly
- by interviewing the persons having information
- Question are explained and hence data is collected.

### COMPARISON

Types of interview/ parameters	Personal interview	Indirect interview	Telephonic interview
accuracy	high	medium	low
coverage	low	low	maximum
Non response	low	low	maximum

Best method Personal Interview

#### Sources of Secondary Data

• (a) International sources like

WHO, ILO, IMF, World Bank etc.

• (b) Government sources like

Statistical Abstract by CSO,

Indian Agricultural Statistics by the Ministry of Food and Agriculture .

- (c) Private and quasi-government sources like ISI, ICAR, NCERT etc.
- (d) Unpublished sources of various research institutes, researchers etc.

#### Scrutiny of Data

- check whether the data under consideration are accurate as well as consistence.
- Errors in data may creep in while writing or copying the answer on the part of the enumerator. A keen observer can easily detect that type of error.
- Again, there may be two or more series of figures which are in some way or other related to each other.
- If the data for all the series are provided, they may be checked for internal consistency

#### Check accuracy



- Consistency
- Date of birth :-12/3/2005
- Age :- 50
- Accuracy
- Weight :- 120
- But very slim

#### **Classification or Organisation of Data**

- (i) Chronological or Temporal or Time Series Data;
- (ii) Geographical or Spatial Series Data;
- (iii) Qualitative or Ordinal Data;
- (iv) Quantitative or Cardinal Data.



#### Chronological or Temporal or Time Series Data

- When the data are classified in respect of successive time points or intervals, they are known as time series data.
- The number of students appeared for CA final for the last twenty years,
- the production of a factory per month from 2000 to 2015 etc. are examples of time series data.

### **Geographical or Spatial Series Data**



BY: SAKO.COM

- Data arranged region wise are known as geographical data.
- If we arrange the students appeared for CA final in the year 2015 in accordance with different states, then we come across Geographical Data.

#### Qualitative or Ordinal Data

- Data classified in respect of an attribute are referred to as qualitative data
- Examples
- Data on nationality,
- gender,
- smoking habit of a group of individuals

#### Quantitative or Cardinal Data

- When the data are classified in respect of a variable,say
- height,
- weight,
- profits,
- salaries etc.

#### DATA

#### frequency data

#### qualitative

quantitative data

#### non-frequency data.

- time series data
- geographical data

#### Mode of Presentation of Data

- (a) Textual presentation;
- (b) Tabular presentation or Tabulation;
- (c) Diagrammatic representation.

#### **Textual presentation**

- This method comprises presenting data with the help of a paragraph or a number of paragraphs.
- The official report of an enquiry commission is usually made by textual presentation.
- 'In 2009, out of a total of five thousand workers of Roy Enamel Factory, four thousand and two hundred were members of a Trade Union. The number of female workers was twenty per cent of the total workers out of which thirty per cent were members of the Trade Union.
- In 2010, the number of workers belonging to the trade union was increased by twenty per cent as compared to 2009 of which four thousand and two hundred were male. The number of workers not belonging to trade union was nine hundred and fifty of which four hundred and fifty were females.'

#### Tabular presentation or Tabulation

Main parts of a table - (Five Parts):

- (a) Caption:- upper part of the table, describing the columns and sub-columns, if any.
- (b) Box-head:- The entire upper part of the table which includes columns and sub-column numbers, unit of measurement along with caption.
- (c) Stub:- The left part of the table providing the description of the rows.
- (d) Body:- Main part of the table that contains the numerical figures.
- (e) Footnotes & Sources: Should be shown at the bottom part of the table.

### Why tabulation

- (i) It facilitates comparison between rows and columns.
- (ii) Complicated data can also be represented using tabulation.
- (iii) It is a must for diagrammatic representation.
- (iv) Without tabulation, statistical analysis of data is not possible.

#### Diagrammatic representation of data

- Line diagram or Historiagram;
- Bar diagram
- Pie chart

#### Line diagram or Historiagram



### logarithmic or ratio chart



Years	Sales
2018	5
2019	500
2020	5000
2021	500000

## Multiple line chart



years	Sales (in crores)	Profit (in crores)
2018	500	56
2019	567	76
2020	987	95
2021	1098	98

## Multiple axis chart

years	Sales (in crores)	No.of employes
2018	500	3000
2019	678	4567
2020	908	8765
2021	1098	9087

#### Bar diagram

 rectangles of equal width and usually of varying lengths are drawn either horizontally or vertically.



#### Types of bar diagram

#### **Horizontal Bar diagram**

 used for qualitative data or data varying over space



#### **Vertical Bar diagram**

 is associated with quantitative data or time series data



## Multiple or Grouped Bar Diagrams:



## **Component or Sub-divided Bar Diagrams**



#### Divided Bar Charts or Percentage Bar Diagrams

year	total	pass	fail
2018	1000	790	210
2019	1587	1200	387
2020	1879	1500	379
2021	2089	1900	189

#### **Pie Chart**



#### **FREQUENCY DISTRIBUTION**

- A frequency distribution may be defined as
- a tabular representation of statistical data,
- usually in an ascending order,
- relating to a measurable characteristic
- according to individual
- or a group of values of the characteristic under study

## frequency distribution

#### simple Frequency Distribution

- When tabulation is done in respect of a discrete random variable, it is known as Discrete or Ungrouped or simple Frequency Distribution
- Example
- The distribution of the number of car accidents in Delhi during 12 months of the year 2005

#### **Grouped Frequency Distribution**

- in case the characteristic under consideration is a continuous variable, such a classification is termed as Grouped Frequency Distribution.
- Example
- distribution of heights of the students of St. Xavier's College for the year 2004

#### Example

Following are the records of babies born in a nursing home in Bangalore during a week(B denoting Boy and G for Girl) :

## B G G B G G B G G B B B G B B G B B B G B B G B B G G G G G B B G

Construct a frequency distribution according to gender.

#### Frequency Distribution of a Variable

- Find the largest and smallest observations and obtain the difference between them, known as Range, in case of a continuous variable.
- Form a number of classes depending on the number of isolated values assumed by a discrete variable. In case of a continuous variable, find the number of class intervals using the relation.
- No. of class Interval X class length = Range.
- Present the class or class interval in a table known as frequency distribution table.
- Apply 'tally mark' i.e. a stroke against the occurrence of a particulars value in a class or class interval.
- Count the tally marks and present these numbers in the next column, known as frequency column, and finally check whether the total of all these class frequencies tally with the totalnumber of observations.

#### Example

- A review of the first 30 pages of a statistics book reveals the following printing mistakes:
- 042113312325602
- 322334056101044
- Make a frequency distribution of printing mistakes.

# Difference between class limit and class boundary

#### **Class Limit (CL)**

- Corresponding to a class interval, the class limits may be defined as the minimum value and the maximum value the class interval may contain. The minimum value is known as the lower
- class limit (LCL) and the maximum value is known as the upper class limit (UCL).

#### **Class Boundary (CB)**

- Class boundaries may be defined as the actual class limit of a class interval.
- For mutually inclusive classification
- For mutually exclusive classification

#### Mid-point or Mid-value or class mark

The central value of the class interval is called the mid point or mid-value or class mark. Mid Point or class mark = LCL+UCL/ 2 or LCB+UCB/2

#### Width or size of a class interval

 The width of a class interval may be defined as the difference between the UCB and the LCB of that class interval.

#### **Cumulative Frequency**

- The cumulative frequency corresponding to a value for a discrete variable and corresponding to a class boundary for a continuous variable may be defined as
- the number of observations less than the value or less than or equal to the class boundary.
### Frequency density of a class interval

 It may be defined as the ratio of the frequency of that class interval to the corresponding class length Relative frequency and percentage frequency of a class interval

- Relative frequency of a class interval may be defined as the ratio of the class frequency to the total frequency.
- Percentage frequency of a class interval may be defined as the ratio of class frequency to the total frequency, expressed as a percentage

## GRAPHICAL REPRESENTATION OF A FREQUENCY DISTRIBUTION

- (i) Histogram or Area diagram;
- (ii) Frequency Polygon;
- (iii) Ogives or cumulative Frequency graphs.



## Histogram or Area diagram

- A comparison among the frequencies for different class intervals is possible in this mode of diagrammatic representation.
- In order to draw a histogram,
- the class limits are first converted to the corresponding class boundaries
- and a series of adjacent rectangles, one against each class interval, with the class interval as base or breadth
- and the frequency or frequency density usually when the class intervals are not uniform as length or altitude, is erected.

# histogram



# Histogram (Area diagram)

- A Histogram is a graph containing a set of rectangles,
- each being constructed to represent the size of the class interval by its width
- and the frequency in each class-interval by its height.
- The area of each rectangle is proportional to the frequency in the respective class-interval and the total area of the histogram is proportional to the total frequency.

## Important points

- When the class-intervals are unequal the heights of rectangles are made proportional not to the class frequencies, but to the frequency densities.
- In construction of histogram the class intervals should be in exclusive form.
- We can find mode graphically by histogram.

# Frequency Polygon

 If we mark the mid-points of the top horizontal sides of the rectangles in a histogram



- and join them by straight lines,
- the figure so formed is called a frequency polygon.
- A frequency polygon is useful in comparing two or more frequency distribution.

# Frequency polygon can be drawn in two ways:

- (a) By preparing histogram first.
- (b) Direct method.
- Usually frequency polygon is meant for single frequency distribution.
- However, we also apply it for grouped frequency distribution provided the width of the class intervals remains the same.

## Ogives or Cumulative Frequency Graph

- A graph which represents the data of a cumulative frequency distribution is called ogive curve.
- Ogive is a line diagram



# **TYPES OF OGIVES**

#### Less than ogive

#### More than ogive

- If the cumulative frequencies are plotted at the upper limit of the class interval, it is a less than ogive.
- Cumulative frequencies are plotted against the lower class boundaries of the respective class, intervals.

## Frequency Curve

- A frequency curve is drawn by smoothing the frequency polygon.
- It is smoothed in such a way that the sharp turns are avoided.
- A frequency curve can be regard as a limiting form of frequency polygon or histogram.

# types of frequency curves

- (a) Bell-shaped curve;
- (b) U-shaped curve;
- (c) J-shaped curve;
- (d) Mixed curve.



## **Bell Shaped Curves**



- Most of the commonly used distributions are bell shaped curves.
- The distribution of height, marks, profit etc. belongs to this category.
- On a bell shaped curves, the frequency starting from a rather low value gradually reaches the maximum value
- and then gradually decreases to reach its lowest
- value at the other extremity frequency is maximum at central part.

## **U-Shaped Curve**



- Frequency is minimum near the central part and the frequency reaches maximum at the two extremities.
- The distribution of Kolkata bound commuters belongs to this type of curve
- as there are maximum number of commuters during the peak hours in the morning and in the evening.

## J-shaped curve



- J-shaped curve starts with a minimum frequency and then gradually reaches its maximum frequency at the other
  extremity.
- The distribution of commuters coming to Kolkata from the early morning hour to peak morning hour follows such a distribution

## Mixed Curve



## False Base Line

- The false base line graph technique is useful from two point of views:
- (i) To magnify the minor fluctuation in time series data.
- (ii) To economic the space.









Some Useful formular P(AUB) = P(A) + P(B) - P(ADB) P(A-B)= P(A) - P(ADB) P(A-B) = P(AnB')P(B-A) = P(B) - P(ADB) P(B-A)= P(BnA') PLAUBIC = 1- PLAUB) PLANB) = PLAUBIC PLANBI - 1- PLANB) P(AUB') = PLANBIC Important Points Nothing mention - Cousider without suplacement. with suplacement -> Independent without ruplacement -> Dependent (If order not given without suplacement -> Combination

Independent <=>> Mutually Exclusive P(A) = 1 Sume event P(A) = 0 Impossible event.  $0 \leq P(A) \leq 1$ ; P(A) + P(A') = 1Properties of Expected Values: E(x+y) = E(x) + E(y) <E(Rn) = RE(n)E(9x+b)= 9E(x)+b E(x.y) = E(x) . E(y) E(n-3) = E(x) - E(3)E(A) = A. whenever nandy are Independent









