A. INTRODUCTION TO STATS

Definition

Singular Sense:

- Scientific method that is used for collecting, analyzing and presenting data
- Used to draw statistical inferences
- Inferences means conclusion reached on the basis of evidence and reasoning

Example:

After applying statistical methods we have arrived at a conclusion that in last 5 years crime rate is reduced.

Plural Sense:

- Data qualitative or quantitative collected to do statistical analysis

Example: Based on Cricket Match statistic of this stadium, chasing team wins mostly

History of Stats

- Word Origin
  - ✓ Latin word – Status
  - ✓ Italian word – Statista
  - ✓ German word – statistic
  - ✓ French word – statistique
- Publication:
  - ✓ Koutilya's book Arthashastra
  - ✓ Stat records on Agriculture found in Ain-i-Akbari (author Abu Fezal)
- Census: First ever census done in Egypt (300 years BC to 2000 BC)

Application of Stats

There are various but we will confine to below:

1. Economics: Time Series analysis, index, demand analysis, econometrics, regression analysis
2. Business Management: business decisions rely upon QT
3. Commerce/ Industry: Sales, Purchase, RM, Salary Wages etc. data are analyze for business decisions and policy making

Limitation of Stats:

1. Relevant for aggregate data and not individual data
2. Quantitative data can only be used, however for qualitative – it needs to be converted into quantitative
3. Projections are based on conditions/ assumptions and any change in that will change the projection
4. Sampling based conclusions are used, improper sampling leads to improper results

B. COLLECTION OF DATA

Data and Variable

- Variable = measurable quantity
  - Discrete variable: when a variable assumes a finite or count ably infinite isolated values. Example: no. of petals in a flower, no. of road accident in locality
  - Continuous variable: when a variable assumes any value from the given interval (can also be in decimals, fractions). Example: height, weight, sale, profit
  - Attribute: qualitative characteristics. Example: Gender of a baby, nationality of a person
- Data = quantitative information shown as number. These are of two types:
  - Primary : first time collected by agency/ investigator
  - Secondary: collected data used by different person/ agency

How to collect Primary Data?

1. **Interview Method**:
   a. Personal Interview: directly from respondents. Example: Natural Calamity, Door to Door Survey
   b. Indirect Interview: when reaching to person difficult, contact associated persons. Example: Rail accident
   c. Telephone Interview: over phone, quick and non-responsive

Type of Interview/ Parameters	Personal	Indirect	Telephone
Accuracy	High	Low	Low
Coverage	Low	Low	High
Non Response	Low	Low	High

2. **Mailed Questionnaire Method:**
   a. Mailed means by Post or Email
   b. Well drafted + properly sequenced + with guidelines
   c. Non Response is Maximum

3. **Observation Method:**
   a. Data collected by direct observation or using instrument
   b. Example: Height check, Weight check,
   c. Although more accurate but it is time consuming, low coverage and laborious

4. **Questionnaire filled and sent by Enumerators**
   a. Enumerator: Person who directly interact with respondent and fill the questionnaire
   b. Generally used in Surveys

Sources of Secondary Data

1. International sources like World Health Organization (WHO), International Monetary Fund (IMF), International Labor Organization (ILO), World Bank
2. Government Sources – In India – Central Statistics Office (CSO), National Sample Survey Office- NSSO, Regulators – RBI, SEBI, RERA, IRDA
3. Private or Quasi-government sources like Indian Statistical Institute (ISI), Indian Council of Agriculture, NCERT
4. Research Papers and other unpublished sources

Scrutiny of Data

1. Scrutiny – checking accuracy and consistency of data
2. Finding of errors by enumerators while filling or receiving questionnaire
3. Internal consistency check: when two or more series of related data are given check each other
4. Consider enumerators' bias while using data

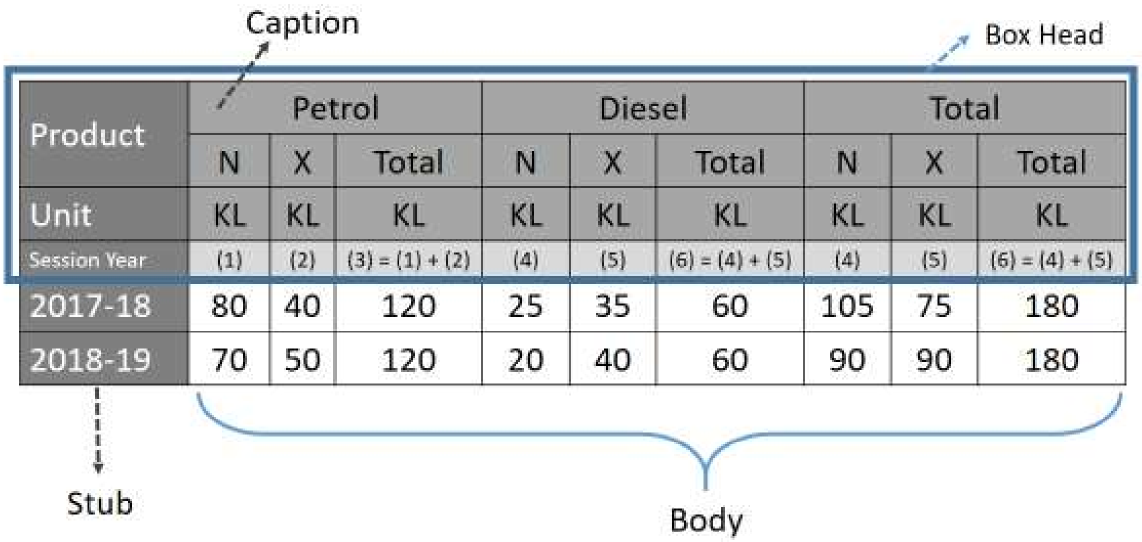
C. PRESENTATION OF DATA

Classification and organization of Data:

- means process of arranging data based on some logic
- there are four types of classification of data
    a. Chronological/ Temporal/ Time Series Data (ex. Profit YoYi.e year on year)
    b. Geographical or Spatial Series Data (ex. Weather in North India and South India)
    c. Qualitative or Ordinal Data (ex. Rating Top 20 songs by Radio Mirchi)
    d. Quantitative or Cardinal Data (no. of left handed batsmen in cricket teams playing CWC19)

Mode of Presentation

1. **Textual:** where text is used in the form of para or sentence. Example: Height of A,B and C is 160cm, 165cm, 175cm respectively
2. **Tabular/ Tabulation:**
    - Data shown in the form of table
    - Some important terms about Table (we will understand by example - next page figure)
    - It is preferred over textual form because
        - Useful in easy comparison
        - Complicated data can be presented
        - Table is must to create a diagram
        - No analysis possible without diagram



Product	Petrol			Diesel			Total		
	N	X	Total	N	X	Total	N	X	Total
Unit	KL	KL	KL	KL	KL	KL	KL	KL	KL
Session Year	(1)	(2)	(3) = (1) + (2)	(4)	(5)	(6) = (4) + (5)	(4)	(5)	(6) = (4) + (5)
2017-18	80	40	120	25	35	60	105	75	180
2018-19	70	50	120	20	40	60	90	90	180

Caption
Box Head
Stub
Body

3. Diagrammatic representation of data

- Can be helpful for layman (without having much knowledge of numbers)
- Hidden trend can be traced
- Table is more accurate than diagrams
- Types of Diagram below:

Line Diagram/ Histogram:

- plotting points in graph and join them to make a line
- used generally for time series (variable y is plotted against time t)
- for wide fluctuation, log chart or ratio chart is used (log y is plotted against t)
- for two or more series of same unit – multiple line chart is used
- for two or more series of distinct unit – multiple axes chart is used
- Refer Material for Diagram

Bar Diagram

- Bar means rectangle of same width and of varying length drawn horizontally or vertically
- For comparable series – multiple or grouped bar diagrams can be used
- For data divided into multiple components – subdivided or component bar diagrams
- For relative comparison to whole, percentage bar diagrams or divided bar diagrams

Pie Chart

- Used for circular presentation of relative data (% of whole)
- Summation of values of all components/segments are equated to 360 Degree (total angle of circle)
- Segment angle = $\frac{\text{segment value x } 360^\circ}{\text{total value}}$

D. FREQUENCY DISTRIBUTION

What is Frequency Distribution?

Frequency means number of times a particular observation is repeated. This applies to both variable and attribute. It is shown in tabular form with class interval or the observation in one column and its frequency in the other.

These are of two types

- Ungrouped/ Simple Frequency Distribution
- Grouped Frequency Distribution

How to construct a frequency distribution

1. Find Range = Largest observation – Smallest Observation
2. Form no. of classes
    - i. In case of discrete variable – number of isolated values
    - ii. In case of continuous variable - no. of class interval = $\frac{\text{Range}}{\text{Class length required}}$
    - iii. Present classes or class interval in a table known as frequency distribution table
    - iv. Apply tally mark/ stroke against occurrence of particular value in a class / class interval
    - v. Count strokes to obtain frequency of each class or class interval

Important Terms

1. **Mutually exclusive classification or Overlapping Classification**: This is usually applicable for continuous variable. An observation as UCL is excluded from the class interval and taken in the class where it is LCL.

   Example: in the below class interval where will the observation 20 fall?

Class	Class where 20 will fall
10-20	No – excluded
20-30	Yes
30-40	No

2. **Mutually inclusive classification or Non Overlapping Classification**: This is usually applicable to discrete variable. All observation including UCL and LCL will be taken in the same class interval as there is no confusion.

   Example:

Class	Class where 20 will fall
10-19	No
20-29	Yes
30-39	No

3. **Class Limit:** for a class interval CL is the minimum and maximum value the class interval may contain. Minimum = Lower Class Interval (LCL) and Maximum = Upper Class Interval (UCL)
   Example:

Class	Type	LCL	UCL	Class	Type	LCL	UCL
10-19	Mutually Exclusive	10	19	10-20	Mutually Inclusive	10	20
20-29	Mutually Exclusive	20	29	20-30	Mutually Inclusive	20	30
30-39	Mutually Exclusive	30	39	30-40	Mutually Inclusive	30	40

4. **Class Boundary:** These are actual class limits of a class interval
   a. **For Mutually Exclusive / Overlapping :** Class Boundary = Class Limit
      LCL = LCB, UCL = UCB
   b. **For Mutually Inclusive / Non Overlapping:** Mid of the two class limits
      LCB = LCL – D/2, UCB = UCL + D/2
      Example:

Class	Type	LCL	UCL	LCB	UCB	Class	Type	LCL	UCL	LCB	UCB
10-19	Mutually Exclusive	10	19	9.5	19.5	10-20	Mutually Inclusive	10	20	10	20
20-29	Mutually Exclusive	20	29	19.5	29.5	20-30	Mutually Inclusive	20	30	20	30
30-39	Mutually Exclusive	30	39	29.5	39.5	30-40	Mutually Inclusive	30	40	30	40

5. **Mid Point/ Mid Value of Class / Class Mark**

$$\frac{\mathbf{LCL+UCL}}{\mathbf{2}} \text{ or } \frac{\mathbf{LCB+UCB}}{\mathbf{2}}$$

6. **Width / Size of Class Interval**
   UCB – LCB

7. **Cumulative Frequency**

Class	Frequency	Less than type CF	More than type CF
10-20	5	5	18
20-30	2	7	13
30-40	8	15	11
40-50	3	18	3
Total	18		

8. **Frequency Density**

$$\frac{\textbf{Frequency of class}}{\textbf{Class length of that class}}$$

9. **Relative Frequency or % Frequency**

$$\frac{\textbf{Frequency of class}}{\textbf{Total Frequency of table}}$$

Class	Frequency	Class Length	Frequency Density	Relative Frequency	Percent Frequency
10-20	5	10	0.5	5/18	27.7%
20-30	2	10	0.2	2/18	11.11%
30-40	8	10	0.8	8/18	44.44%
40-50	3	10	0.3	3/18	16.67%
Total	18				

Graphical Presentation of Frequency Distribution

1. **Histogram/ Area Diagram** [refer study material page 14.20 for diagram]
   a. It is a convenient way to represent FD
   b. Comparison between frequency of two different classes possible
   c. It is useful to calculate mode also
   d. Steps to create
      - Covert CL into CB and plot in x axis
      - Form rectangles taking class interval as base (x axis)
      - And frequency as length (y axis) | Use frequency density in case of uneven length

2. **Frequency Polygon**
   a. Usually preferable for ungrouped frequency distribution
   b. Can be used for grouped also but only if class lengths are even
   c. Steps to create
      - Plot (x_i, f_i) where x_i = class value (in case of ungrouped), mid value (in case of grouped) and f_i = frequency
      - Join all plotted points to make line segments which eventually will become a polygon (a shape with multiple number of line segments)

3. **Ogives/ Cumulative Frequency Graph**
   a. Create a table where cumulative frequency is mapped against each CB (Class Boundary) and make a curve by plotting and joining points by line segments. (curve is called Ogive)
   b. This graph can be made by both type of Cumulative Frequency and called as Less than Ogive or More than Ogive

c. It can be used for calculating quartiles also
d. If we plot both ogives in same graph, perpendicular line drawn from their intersection towards x axis is cutting axis at Median

4. Frequency Curve

a. It is a limiting form of Area Diagram (Histogram) or frequency polygon
b. It is obtained by drawing smooth and free hand curve though the mid points
c. These are of below four types:
   - Bell Shaped
   - U-Shaped
   - J-Shaped
   - Combination of Curves as Mixed Curve