Last 7 attempts Weightage of Theory in Statistics

Attempt	Type	Chp13	Chp14	Chp15	Chp16	Chp17	Chp18	Total
Dec 22	**Theory**	**4**	**3**	**1**	**2**	**1**	**3**	**14**
	Practical	0	13	5	2	4	2	26
	Total	**4**	**16**	**6**	**4**	**5**	**5**	**40**
Jun 22	**Theory**	**9**	**3**	**0**	**1**	**4**	**6**	**23**
	Practical	0	6	7	2	2	0	17
	Total	**9**	**9**	**7**	**3**	**6**	**6**	**40**
Dec 21	**Theory**	**3**	**5**	**0**	**0**	**2**	**4**	**14**
	Practical	4	6	7	6	2	1	26
	Total	**7**	**11**	**7**	**6**	**4**	**5**	**40**
Jul 21	**Theory**	**6**	**1**	**0**	**0**	**1**	**0**	**8**
	Practical	2	10	7	4	4	5	32
	Total	**8**	**11**	**7**	**4**	**5**	**5**	**40**
Jan 21	**Theory**	**10**	**5**	**1**	**2**	**2**	**4**	**24**
	Practical	0	4	5	2	3	2	16
	Total	**10**	**9**	**6**	**4**	**5**	**6**	**40**
Dec 20	**Theory**	**8**	**5**	**0**	**4**	**3**	**6**	**26**
	Practical	1	4	4	3	1	1	14
	Total	**9**	**9**	**4**	**7**	**4**	**7**	**40**
Nov 19	**Theory**	**1**	**7**	**0**	**2**	**2**	**5**	**17**
	Practical	0	10	5	4	3	1	23
	Total	**1**	**17**	**5**	**6**	**5**	**6**	**40**

THEORY SUMMARY

Statistical Description of Data – Basics of Statistics

Definition of Statistics

- Plural Sense: Any data – quantitative or qualitative used for statistical analysis.
- Singular Sense: Scientific method of collecting, analyzing, and presenting data to draw statistical inferences. It is also called as Science of Averages or Science of Counting

Origin of Word

Language	Actual Word	Memorize by
Latin	Status	Latus
Italian	Statista	Pasta
German	Statistic	Breadstick
French	Statistique	Barbeque

<table>
<tr><td>Publication</td><td>
<table>
<tr><td>Koutilya's Arthashastra</td><td>• Record of Birth and Deaths
• Chandragupta's reign
• 4th Century B.C</td></tr>
<tr><td>Abu Fezal's Ain-i-Akbari</td><td>• Record on Agriculture
• Akbar Reign
• 16th Century A.D.</td></tr>
<tr><td>First Census</td><td>• Egypt 300 BC to 2000 BC
• By Pharaoh</td></tr>
</table>
</td></tr>
<tr><td>Application of Statistics</td><td>• Economics: Demand Analysis, Future Projection etc.
• Business Management: Decision making using quantitative techniques not intuition
• Industry and Commerce: Profit maximization using business data – sales, purchase, market etc. by consulting experts</td></tr>
<tr><td>Limitation of Statistics</td><td>• It deals with aggregate data and not individual data
• Quantitative data can only be used, however for qualitative – it needs to be converted into quantitative
• Projections are based on conditions/ assumptions and any change in that will change the projection. Example: Future projections of sales
• Sampling based conclusions are used, improper sampling leads to improper results. Random Sampling is must.</td></tr>
<tr><td>Data</td><td>• Quantitative Information shown as number
• Primary: first time collected by agency/ investigator
• Secondary: collected data used by different person/ agency</td></tr>
<tr><td>Variable</td><td>• Measurable Data – Value can vary
<table>
<tr><td>Discrete Variable</td><td>• 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</td></tr>
<tr><td>Continuous Variable</td><td>• When a variable assumes any value from the given interval (can also be in decimals, fractions).
• Example: height, weight, sale, money</td></tr>
</table>
</td></tr>
<tr><td>Attribute</td><td>• Qualitative Characteristics. Example: gender of a baby, the nationality of a person, the colour of a flower etc.</td></tr>
<tr><td>Collection of Primary Data – Interview Method</td><td>
<table>
<tr><th>Method</th><th>Details</th></tr>
<tr><td>Personal Interview</td><td>• Where data is collected directly from respondents.
• Highly Accurate – Low Coverage
• Example: Natural Calamity, Door to Door Survey</td></tr>
<tr><td>Indirect Interview</td><td>• When reaching respondent is difficult, data is collected by contacting associated persons.
• Highly Accurate – Low Coverage
• Example: Rail accident</td></tr>
<tr><td>Telephone Interview</td><td>• Data is collected over phone
• Quick and non-expensive method</td></tr>
</table>
</td></tr>
</table>

<table>
<tr><td>Collection of Primary Data – Mailed Questionnaire Method</td><td>• In this method well drafted and soundly sequenced questionnaire,
• covering all the important aspects of the data requirement is sent to respondent for filling.
• Here coverage is wide but amount of non-responses will be maximum</td></tr>
<tr><td>Collection of Primary Data – Observation Method</td><td>• In this method data is collected by direct observation or using instrument.
• For example: data on height and weight for a group of students.
• Although more accurate but it is time consuming, low coverage and laborious method.</td></tr>
<tr><td>Collection of Primary Data – Questionnaire Filled and sent by Enumerators</td><td>• Mix of Interview and Mailed Questionnaire
• Enumerator means a Person who directly interacts with respondent and fills the questionnaire.
• It is generally used in case of Surveys and Census.</td></tr>
<tr><td>Sources of Secondary Data</td><td><table>
<tr><td>International Sources</td><td>World Health Organization (WHO), International Monetary Fund (IMF), International Labor Organization (ILO), World Bank</td></tr>
<tr><td>Government Sources</td><td>In India – Central Statistics Office (CSO), Indian Agricultural Statistics by the Ministry of Food and Agri, National Sample Survey Office- NSSO, Regulators – RBI, SEBI, RERA, IRDA</td></tr>
<tr><td>Private or Quasi-govt. sources</td><td>Indian Statistical Institute (ISI), Indian Council of Agriculture, NCERT</td></tr>
</table></td></tr>
<tr><td>Scrutiny of Data</td><td>• checking accuracy and consistency of data
• There is no rule for it, one must apply his intelligence, patience and experience while scrutinizing the given information.
• Internal Consistency: When two or more series of related data are given, we should check consistency among them.</td></tr>
<tr><td>Presentation of Data – Classification / Organization of Data</td><td>Classification or Organisation: putting data in a neat, precise, and condensed form, making it comparable, suitable for analysis, more understandable.<table>
<tr><td>Chronological/ Temporal/ Time Series Data</td><td>• Data arranged based on Time
• Example: Revenues YoY i.e year on year</td></tr>
<tr><td>Geographical or Spatial Series Data</td><td>• Arrangement based on regions
• Example: Country wise Revenue of a global company</td></tr>
<tr><td>Qualitative or Ordinal Data</td><td>• Based on some attribute
• Nationality Wise Medal Winners in Olympics</td></tr>
<tr><td>Quantitative or Cardinal Data</td><td>• Based on some variable
• Example: Frequency Distribution of a Data</td></tr>
</table></td></tr>
<tr><td>Mode of Presentation of Data –</td><td>• This method comprises presenting data with the help of a paragraph or several paragraphs.
• This is not a suitable mode of presentation as it is dull, monotonous and</td></tr>
</table>

<table>
<tr><td>Mode of Presentation of Data – Tabular Form</td><td>

When data is shown in the form of Table.
Useful in easy comparison
Complicated data can be presented
Table is must to create a diagram
No analysis possible without table
Components of Table

</td></tr>
<tr><td>Components of Table</td><td>
<table>
<tr><th>Description</th><th>Name of Component of Table</th></tr>
<tr><td>Entire Upper Part</td><td>Box Head</td></tr>
<tr><td>Upper Part describing columns and sub-columns</td><td>Caption</td></tr>
<tr><td>Left part of the table describing rows</td><td>Stub</td></tr>
<tr><td>Main Data of Table</td><td>Body</td></tr>
<tr><td>Source of Data at the bottom of Table</td><td>Footnote</td></tr>
</table>
</td></tr>
<tr><td>Mode of Presentation of Data – Diagrams</td><td>

Can be used by educated and uneducated section of society
Hidden trend can be traced
If priority is accuracy, then tabulation is better

</td></tr>
<tr><td>Line Diagram</td><td>

Time Series is generally in x axis
For wide fluctuation – log chart or ratio chart is used
Two or more series of same unit – Multiple Line Chart
Two or more series of different unit – Multiple Axis Chart

</td></tr>
<tr><td>Bar Diagram</td><td>

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
Vertical Bar Diagram: Useful for Data varying over Time and Quantitative Data
Horizontal Bar Diagram: Useful for Data varying over Space and Qualitative Data

</td></tr>
<tr><td>Pie Chart</td><td>

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 = $\frac{\textbf{(segment value x 360°)}}{\textbf{(total value)}}$

</td></tr>
</table>

Statistical Description of Data – Frequency Distribution

Frequency and	• Frequency means number of times a particular observation is repeated.

<table>
<tr><td></td><td>
- tabular representation of statistical data, usually in an ascending order,
- relating to a measurable characteristic
- according to individual value or a group of values of the characteristic under study.
</td></tr>
<tr><td>Types of Frequency Distribution</td><td>
<table>
<tr><td>Ungrouped/ Simple Frequency Distribution</td><td>• When there are limited number of distinct observations, frequency can be assigned to each one of them.
• This distribution is simple</td></tr>
<tr><td>Grouped Frequency Distribution</td><td>• When there are large no. of observations, grouping is done among them (generally in ascending order).
• Each group is called as class interval and frequency is assigned to group and not individual values,
• this is called Grouped Frequency Distribution</td></tr>
</table>
</td></tr>
<tr><td>Class Limit</td><td>
• For a class interval CL is the minimum and maximum value the class interval may contain
• Minimum Value – Lower Class Limit
• Maximum Value – Upper Class Limit
<table>
<tr><th>Class Interval</th><th>Frequency</th><th>LCL</th><th>UCL</th></tr>
<tr><td>10-19</td><td>10</td><td>10</td><td>19</td></tr>
<tr><td>20-29</td><td>5</td><td>20</td><td>29</td></tr>
<tr><td>30-39</td><td>8</td><td>30</td><td>39</td></tr>
</table>
</td></tr>
<tr><td>Classification of Grouped of Frequency Distribution</td><td>
<table>
<tr><td>Mutually Exclusive / Overlapping Classification
<table>
<tr><th>Class</th><th>LCL</th><th>UCL</th></tr>
<tr><td>10-20</td><td>10</td><td>20</td></tr>
<tr><td>20-30</td><td>20</td><td>30</td></tr>
<tr><td>30-40</td><td>30</td><td>40</td></tr>
</table>
</td><td>• Here UCL an interval and LCL of next interval are same
• This is usually applicable for continuous variable.
• An observation which is equivalent to common class limit is excluded from the class interval where it is UCL and taken in the class where it is LCL.</td></tr>
<tr><td>Mutually Inclusive / Non-Overlapping Classification
<table>
<tr><th>Class</th><th>LCL</th><th>UCL</th></tr>
<tr><td>10-19</td><td>10</td><td>20</td></tr>
<tr><td>20-19</td><td>20</td><td>30</td></tr>
<tr><td>30-39</td><td>30</td><td>40</td></tr>
</table>
</td><td>• There is no common class limits between two intervals.
• 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.</td></tr>
</table>
</td></tr>
</table>

Class Boundary

In case of Exclusive / Overlapping Classification

Class Boundary = Class Limit

Class	LCL	UCL	LCB	UCB
10-20	10	20	10	20
20-30	20	30	20	30
30-40	30	40	30	40

In case of Inclusive / Overlapping Classification

Lower Class Boundary

LCB = LCL – 0.5

UCB = UCL + 0.5

Class	LCL	UCL	LCB	UCB
10-19	10	19	9.5	19.5
20-29	20	29	19.5	29.5
30-39	30	39	29.5	39.5

Mid-Point / Class Mark / Mid Value of Class Interval

$$\frac{LCL+UCL}{2} \qquad \frac{LCB+UCB}{2}$$

- Useful in calculation of AM, GM, HM, SD in case of grouped frequency distribution

Class Length/ Width or Size

UCB – LCB only

Cumulative Frequency

- Less than type: It shows no. of observations less than UCB
- More than type: It shows no. of observations more than UCB

Class Interval	Freq.	UCB	Less than type CF	More than type CF	Total of both CF
44-48	3	48.5	3	33	36
49-53	4	53.5	7	29	36
54-58	5	58.5	12	24	36
59-63	7	63.5	19	17	36
64-68	9	68.5	28	8	36
69-73	8	73.5	36	0	36
Total	36				

Frequency Density

$$\frac{\text{Class Frequency}}{\text{Class Length of class}}$$

Relative Frequency

$$\frac{\text{Class frequency}}{\text{Total Frequency}}$$

Its can have values between 0 and 1

Percentage Frequency

$$\frac{\text{Class frequency}}{\text{Total Frequency}} \times 100$$

Frequency Dist. Diagram – Histogram

- It is a convenient way to represent FD
- Comparison between frequency of two different classes possible
- It is useful to calculate mode also

Frequency Polygon

- Usually preferable for ungrouped frequency distribution
- Can be used for grouped also but only if class lengths are even

Ogives/ Cumulative

- This graph can be made by both type of Cumulative Frequency and called as Less than Ogive or More than Ogive

Index Numbers

Practical Examples of Index Numbers	• Index numbers are convenient devices for **measuring relative changes (generally in %)** of differences from **time to time** or from **place to place** • Series of numerical figures which show relative position • Index Numbers show percentage changes rather than absolute amounts of change
Data Selection	• It **depends on the purpose** for which the index is used. • Index numbers are often constructed from the **sample**. **Random sampling,** and if need be, a **stratified random sampling** can be used to ensure that sample is representative. • Data should be **comparable** by ensuring consistency in selection method.
Base Period	• It is a **point of reference** in comparing various data. • Standard point of comparison. • The period should be **normal**. • It should be **relatively recent** • Choice of suitable base period is a temporary solution
Use of Averages	• The **geometric mean is better** in averaging relatives, • But for most of the index's **arithmetic mean is used because of its simplicity**
Price/ Quantity/ Value Relative	For Individual Commodity, $\frac{\textbf{Current Period Price/ Quantity/ Value}}{\textbf{Base Period Price/ Quantity/ Value}}$
Link Relative	$\frac{P_1}{P_0}, \frac{P_2}{P_1}, \frac{P_3}{P_2}, \ldots, \frac{P_n}{P_{n-1}}$ Same can be created for quantities also
Chain relatives	When the above relatives are in respect to a fixed base period these are also called the chain relatives $\frac{P_1}{P_0}, \frac{P_2}{P_0}, \frac{P_3}{P_0}, \ldots, \frac{P_n}{P_0}$
Formula for Chain Index (when direct data is not available)	$\frac{\textbf{Link relative of current year} \times \textbf{Chain Index of previous year}}{\textbf{100}}$ The chain index is an unnecessary complication unless of course where data for the whole period are not available or where commodity basket or the weights have to be changed.
Limitations of Index Numbers	• Chances of errors due to Sampling • It gives broad trend not real picture • Due to many methods, at times it creates confusion
Usefulness of Index Numbers	• Index numbers are very useful in deflating (eg. Nominal wages into real) • Framing suitable policies in economics and business

	• They reveal trends and tendencies in making important conclusions • They are used in time series analysis to study long-term trend, seasonal variations and cyclical developments
Formula for Deflated Value	**Deflated Value** $= \frac{\text{Current Value}}{\text{Price Index of the current year}}$
Shifted Price Index	$\frac{\text{Original Price Index}}{\text{Price Index of the year on which it has to be shifted}} \times 100$
Unit Test	• This test requires that the formula should be independent of the unit in which or for which prices and quantities are quoted. • Except for the simple (unweighted) aggregative index all other formulae satisfy this test.
Time Reversal Test	• It is a test to determine whether a given method will work both ways in time, forward and backward. • $P_{01} \times P_{10} = 1$ • Laspeyres' method and Paasche's method do not satisfy this test, but Fisher's Ideal Formula does.
Factor Reversal Test	• This holds when the product of price index and the quantity index should be equal to the corresponding value index. • Symbolically $P_{01} \times Q_{01} = V_{01}$ • Fisher's Index Number is ideal as it satisfies Unit, Time Reversal and Factor Reversal Test
Circular Test	• This property therefore enables us to adjust the index values from period to period without referring each time to the original base. • It is an extension of time reversal test • The test of this **shiftability of base** is called the circular test. • This test is not met by Laspeyres, or Paasche's or the Fisher's ideal index. • The weighted GM of relative, **simple geometric mean of price relatives** and the **weighted aggregative with fixed weights meet this test.** (These methods are not in syllabus)
Cost of Living Index (also called General Index)	• CLI is defined as the **weighted AM of index numbers** of few groups of basic necessities. • AM of group indices gives the General Index • Generally, for calculating CLI; food, clothing, house rent, fuel & lightning and miscellaneous groups are taken into consideration. • Examples of CLI: WPI, CPI, etc.
Symbol	• P_{01} is the index for time 1 on 0 • P_{10} is the index for time 0 on 1

Arithmetic Mean

Property 1	If all the observations are constant, AM is also constant
Property 2	the algebraic sum of deviations of a set of observations from their AM is zero
Property 3	AM is affected both due to change of origin and scale If $y = a + bx$ then $\bar{y} = a + b\bar{x}$
Property 4	Combined AM $\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$
General Review	• AM is best measure of central tendency • AM is based on all observations • AM is affected by sampling fluctuations • AM is amenable to mathematical property • AM cannot be used in case of open end classification

Median

Property 1	For a set of observations, the sum of absolute deviations is minimum, when the deviations are taken from the median. $\sum\lvert x_i - Me\rvert$
Property 2	Median is also affected by both change of origin and scale.
General Review	• Median is also called as positional average • Median is not based on all observations • Median is not affected by sampling fluctuations • Median is best measure of central tendency in case of open end classification

Partition Values

Meaning

- These may be defined as **values dividing** a given **set of observations** into number of **equal parts**
- When we want to divide the given set of observations into two equal parts, we consider median, similarly there are quartiles, deciles, percentiles

Name of PV	No. of equal parts	No. of PVs	Symbol
Median	2	1	**Me**
Quartile	4	3	Q_1, Q_2, Q_3
Decile	10	9	$D_1, D_2, \ldots, D_9$
Percentile	100	99	$P_1, P_2, \ldots, P_{99}$