UNIT II Exploratory Analysis

Exploratory data analysis is an approach to analyze the data using visual techniques. It is used to discover trends, patterns or to check assumptions with the help of statistical summary and graphical representations

Descriptive statistics: -

Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that for eg. pattern might emerge from the data.

Descriptive statistics do not, however, allow us to make conclusion beyond the data we have analysed or reach conclusions regarding any hypotheses, we might have made. They are simply a way to describe our data.

Descriptive statistics are very important because if we simply presented our raw data it would be hard to visualize what the data was showing. especially if there was a lot of it. Descriptive statistics therefore enables us to present the data in a more meaningful way, which allows simpler interpretations of the data.

for eg.

If we had the results of 100 pieces of students coursework, we may be interested in the overall performance of those students. We would also be interested in the distribution or spread of the marks. Descriptive statistics allow us to do this.

There are 2 types of statistic that are used to describe data.

1 Measures of central tendency –

These are ways of describing the central position of frequency distribution for a group of data. In this case, the frequency distribution is simply the distribution and pattern of marks scored by the 100 students from the lowest to the highest. We can describe this central position using a number of statistics, including the mode, median, and mean.

2. Measures of spread

These are ways of summarizing a group of data by describing how spread out the scores are. for eg.

The mean score of our 100 students may be 65 out of 100. However, not all students will have scored 65 marks. Rather, their scores will be spread out. Some will be be lower and others higher. measures of spread help us to summarize how spread out these scores are. To describe this spread, a number of statistics are available to us, including the range, quartiles, absolute deviation, variance and standard deviation.

When we use descriptive statistics it is useful to summarize our group of data using a combination of tabulated description, graphical description and statistical commentary.

Central Tendency

1 Mean –

The mean or average is probably the most commonly used method of describing central tendency. To compute the mean all you do is add up all the values and divide by the number of values.

eg. 15, 10, 21, 20, 36, 15, 25, 15

The sum of the values is = 167.

$$\text{Mean} = \frac{167}{8} = 20.875.$$

2 Median – The median is the score found at the exact middle of the set of values. One way to compute the median to list all scores in numerical order and then locate the score in the center of the sample.

eg. 15, 15, 15, 20, 20, 21, 25, 36

Median is 20

3 Mode –

The mode is the most frequently occurring value in the set of scores. To determine the mode, you might again order the scores and count each one. The most frequently occurring value is the mode.

for eg. 15 is occurs 3 times.

4 Standard Deviation –

The standard deviation is a more accurate and detailed estimate of dispersion

because an outlier can greatly exaggerate the range. The Standard deviation shows the relation that set of scores has to the mean of the sample.

eg. 15, 20, 21, 20, 36, 15, 15, 15

To compute the standard deviation, we first find the distance between each value and the mean.

Mean = 20.875

The differences from mean are.

15 − 20.875 = − 5.875
20 − 20.875 = −0.875
21 − 20.875 = +0.125
20 − 20.875 = −0.875
36 − 20.875 = 15.125
15 − 20.875 = −5.875
15 − 20.875 = +4.125
15 − 20.875 = −5.875

Values that are below the mean have negative discrepancies and values above it have positive one.

We square each discrepancy.

−5.875 × −5.875 = 34.51
−0.875 × −0.875 = 0.765
+0.875 × +0.875 = 0.0156
−0.875 × +0.875 = 0.7656
15.125 × 15.125 = 228.75
−5.875 × −5.875 = 34.51
+4.125 × +4.125 = 17.0156
−5.875 × −5.875 = 34.515

We take these squares and sum them to get the sum of squares values.

Sum = 350.875

We divide this sum by the number of scores minus 1.

$\frac{350.875}{7} = 50.125$ The value is known as the variance.

To get standard deviation, we take the square root of the variance.

SQRT (50.125) = 7.0799.

formula for the standard deviation.

$\Sigma (x - \bar{x})^2 / n-1$ \Sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}

$\sqrt{n-1} / n-1 \Sigma (x - \bar{x})^2$

Where

x is each score

$\bar{x}$ is the mean or average

n is the no. of values

Σ means we sum across the values.

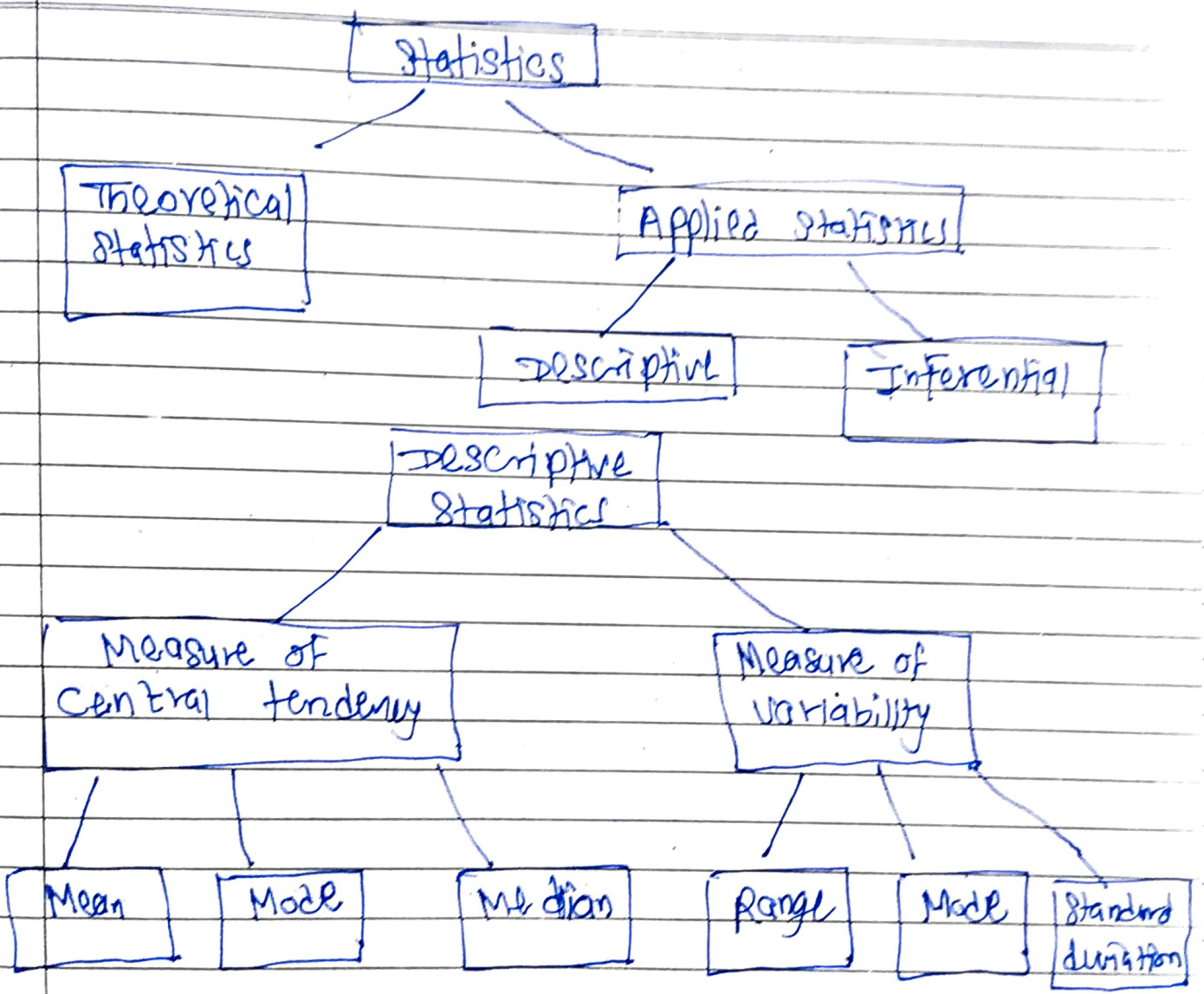
Types of descriptive statistics - 4

1 Measures of frequency
Count, Percent, frequency
Shows how often something occurs.
Use this when you want to show how often a response is given.

2 Measures of central Tendency
mean, median, mode
locates the distribution by various points
Use this when you want to show how an average or most commonly indicated response.

3 Measures of dispersion or variation
Range, variance, Standard deviation
identifies the spread of scores by stating interval.
Range = High / Low points.
Variance or Standard deviation = difference between obtained score and mean.
Use this when you want to show how "spread out" the data are. It is helpful to know when your data are so spread out that it affects the mean.

4 Measures of position
- percentile Ranks, Quartile Ranks
- Describes how scores, fall in relation to one another. Relies on standardized scores.
- Use this when you need to compare scores to a normalized score.



Statistics — the practice or science of collecting and analysing numerical data in large quantities.

Descriptive statistics	Inferential statistics
1. It gives information about raw data which describes the data in some manner.	① It makes inference about population using data drawn from the population.
2. It helps in organizing, analyzing and to present data in a meaningful manner.	② It allows to compare data, make hypothesis and predictions.
3. It is used to describe a situation	③ It is used to explain the chance of occurance of an event.
4. It explain already known data and limited to a sample or population having small size.	④ It attempts to reach the conclusion about the population.
5. It can be achieved with the help of charts, graphs, tables	⑤ It can be achieved by probability.

Comparitive Analysis

Comparative analysis refers to the comparision of two or more processes, documents, data sets or other objects. Pattern analysis, filtering and decision tree analytics are forms of comparative analysis.