Statistics Basic-1

Q1. What is Statistics?

ANS:

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. In other words, it is a mathematical discipline to collect, summarize data.

In data science, statistics is used to summarize and describe data through measures such as mean, median, mode, and standard deviation. It is also used to test hypotheses and make inferences about populations based on samples of data. Additionally, statistical models are used to make predictions and identify relationships between variables.

Basically, there are two types of statistics.

1. Descriptive Statistics
2. Inferential Statistics

Q2. Define the different types of statistics and give an example of when each type might be used.

ANS:

Basically, there are two types of statistics.

1. Descriptive Statistics
2. Inferential Statistics

1. Descriptive Statistics:

- Descriptive statistics is a term given to the analysis of data that helps to describe, show and summarize data in a meaningful way.
- It is a simple way to describe our data.
- Descriptive statistics is very important to present our raw data ineffective/meaningful way using numerical calculations or graphs or tables.
- This type of statistics is applied to already known data.

Types of Descriptive Statistics:

1. Measure of Central Tendency
    - mean
    - mode
    - median
2. Measure of Variability
    - range
    - varience
    - standard deviation

Example of descriptive statistics:

A researcher collects data on the heights of 100 students in a school. The researcher uses descriptive statistics to summarize the data by calculating the mean height, standard deviation, and creating a histogram to visualize the distribution of heights.

2. Inferential Statistics:

- It consist of using data you have to measure conclusion.
- In inferential statistics, predictions are made by taking any group of data in which you are interested.
- It can be defined as a random sample of data taken from a population to describe and make inferences about the population.
- Any group of data that includes all the data you are interested in is known as population.
- It basically allows you to make predictions by taking a small sample instead of working on the whole population.

Example of inferential statistics:

A company wants to determine whether a new product will be successful among its target market. The company collects a sample of data from a group of customers who have purchased the product and uses inferential statistics to determine whether the data supports the conclusion that the product will be successful among the entire target market. The company may use techniques such as hypothesis testing, confidence intervals, or regression analysis to draw conclusions about the population.

Q4. Categorise the following datasets with respect to quantitative and qualitative data types:

1. Grading in exam: A+, A, B+, B, C+, C, D, E
2. Colour of mangoes: yellow, green, orange, red
3. Height data of a class: [178.9, 179, 179.5, 176, 177.2, 178.3, 175.8,...]
4. Number of mangoes exported by a farm: [500, 600, 478, 672, ...]

ANS:

1. Grading in exam: A+, A, B+, B, C+, C, D, E

- Qualitative (Categorical) Data Type

2. Colour of mangoes: yellow, green, orange, red

- Qualitative (Categorical) Data Type

3. Height data of a class: [178.9, 179, 179.5, 176, 177.2, 178.3, 175.8,...]

- Quantitative (Continuous) Data Type

4. Number of mangoes exported by a farm: [500, 600, 478, 672, ...]

- Quantitative (Discrete) Data Type

Q5. Explain the concept of levels of measurement and give an example of a variable for each level.

ANS: The concept of levels of measurement, also known as scales of measurement.

1. Nominal Level: You can categorize your data by labelling them in mutually exclusive groups, but there is no order between the categories. Examples: City of birth, Gender, Ethnicity, Car brands, Marital status
2. Ordinal Level: You can categorize and rank your data in an order, but you cannot say anything about the intervals between the rankings.Although you can rank the top 5 Olympic medallists, this scale does not tell you how close or far apart they are in number of wins. Examples: Top 5 Olympic medallists, Language ability (e.g., beginner, intermediate, fluent)
3. Interval Level: You can categorize, rank, and infer equal intervals between neighboring data points, but there is no true zero point. For example, the difference between any two adjacent temperatures is the same: one degree. But zero degrees is defined differently depending on the scale – it doesn’t mean an absolute absence of temperature. Other Examples: Test scores (e.g., IQ or exams), Personality inventories, Temperature in Fahrenheit or Celsius
4. Ratio Level: You can categorize, rank, and infer equal intervals between neighboring data points, and there is a true zero point.A true zero means there is an absence of the variable of interest. In ratio scales, zero does mean an absolute lack of the variable. For example, in the Kelvin temperature scale, there are no negative degrees of temperature – zero means an absolute lack of thermal energy. Other

Q6. Why is it important to understand the level of measurement when analyzing data? Provide an example to illustrate your answer.

ANS:

Understanding the level of measurement is important when analyzing data because it determines the type of statistical analyses that can be performed on the data, as well as the appropriate way to interpret the results. Different types of statistical analysis are suited for different levels of measurement, and using an inappropriate analysis can result in inaccurate conclusions.

If a variable is nominal, it is inappropriate to perform calculations such as mean or standard deviation, as these statistics are not meaningful for categorical data. Instead, appropriate measures for nominal data include frequency counts and percentages. If a variable is ordinal, it is appropriate to use measures such as median or mode, but not mean, as the ranking of the values does not necessarily correspond to equal intervals. If a variable is interval or ratio, all three measures (mean, median, and mode) can be used, as well as measures such as standard deviation and correlation.

For example let's consider a study that examines the relationship between income and job satisfaction among employees in a company. The income variable is a ratio-level variable because it has a true zero point (no income) and equal intervals between values. In contrast, job satisfaction is an ordinal-level variable because it has a specific order or rank, but unequal or non-meaningful differences between categories. If the researcher were to use a correlation analysis to examine the relationship between these two variables, the result would be incorrect because correlation is only appropriate for ratio-level variables. Instead, a more appropriate statistical test for analyzing the relationship between a ratio-level and ordinal-level variable would be a chi-squared test or a t-test. In summary, understanding the level of measurement of variables is crucial for selecting appropriate statistical methods and interpreting results accurately in data analysis.

Q7. How nominal data type is different from ordinal data type.

ANS:

Nominal and ordinal data are both types of categorical data, but they differ in the level of measurement and the type of information they provide.

Nominal data is the lowest level of measurement and refers to data that is categorized without any numerical or quantitative value. In other words, nominal data consists of categories that are mutually exclusive and do not have any inherent order or ranking. Examples of nominal data include gender (male/female), hair color (blonde/brunette/red), and marital status (single/married/divorced).

Ordinal data, on the other hand, is the second level of measurement and refers to data that has an inherent order or ranking, but the differences between categories may not be equal or meaningful. Examples of ordinal data include educational level (elementary school, high school, college), level of agreement (strongly disagree, disagree, neutral, agree, strongly agree), and rating scales (poor, fair, good, excellent).

Q8. Which type of plot can be used to display data in terms of range?

ANS:

- A type of plot that can be used to display data in terms of range is a box plot .
- it is also known as a box-and-whisker plot.
- A box plot is a graphical representation of the distribution of a set of continuous data, using five summary statistics: the minimum value, the first quartile (Q1), the median (Q2), the third quartile (Q3), and the maximum value.
- Box plots are particularly useful for displaying data that has a non-normal distribution, outliers, or multiple groups of data that need to be compared.
- A box plot provides a visual representation of the minimum, maximum, median, and quartiles of a dataset. The box in the plot represents the interquartile range (IQR), which is the range of values between the first quartile (25th percentile) and the third quartile (75th percentile) of the dataset. The median is represented by a line within the box, while the minimum and maximum values are represented by the whiskers (vertical lines) that extend from the box.

For example A box plot could be used to display the range of temperatures recorded in a city over a given period of time. The minimum and maximum values would be represented by the whiskers, and the box would represent the interquartile range (the middle 50% of the data). The median temperature would be represented by a line inside the box. This would allow for easy comparison of temperature ranges across different months or years, and also identify any unusually high or low temperatures as outliers.

- Histogram is also one of the plot that can be shown in terms of ranges
- A histogram is particularly useful for displaying continuous data, as it allows you to see the shape of the distribution, including any peaks or clusters, as well as the range and spread of the data. It can also help identify outliers or unusual patterns in the data.

Q9. Describe the difference between descriptive and inferential statistics. Give an example of each type of statistics and explain how they are used.

ANS:

Descriptive statistics involves summarizing and describing the main features of a data set, such as measures of central tendency (mean, median, mode) and measures of dispersion (range, standard deviation, variance). Descriptive statistics is used to provide a snapshot or summary of the data, in order to gain an understanding of the characteristics and patterns of the data set.

Examples of descriptive statistics include calculating the mean height of a group of people or the standard deviation of their weights.

Inferential statistics , on the other hand, involves using statistical methods to make inferences or draw conclusions about a population based on a sample of data. Inferential statistics is used to test hypotheses or answer research questions, by making predictions or generalizations about a larger group based on a smaller subset of data.

Q10. What are some common measures of central tendency and variability used in statistics? Explain how each measure can be used to describe a dataset.

ANS:

- Some of the common Measures of Central Tendency used in Statistics are:

```
1. Mean:
The arithmetic mean, or average, is calculated by summing all the values in
a dataset and dividing by the number of values. The mean is a useful measure
of central tendency when the data is normally distributed and there are no e
xtreme values or outliers.

2. Median:
The median is the middle value in a dataset when the values are arranged in
order. The median is useful when there are extreme values or outliers in the
dataset, since it is not affected by these values in the same way as the mea
n.

3. Mode:
The mode is the most common value in a dataset. The mode is useful when deal
ing with categorical data or discrete variables, where there may be several
values that occur with the same frequency.
```

- Some of the common Measures of Variability used in Statistics are:

```
1. Range:
The range is the difference between the largest and smallest values in a dat
aset. The range provides a simple measure of variability, but it can be sens
itive to extreme values or outliers.

2. Variance:
The variance measures the average squared deviation from the mean in a datas
et. The variance is useful when you want to quantify the spread of a datase
t, and it is used in many statistical tests and models.

3. Standard deviation:
The standard deviation is the square root of the variance. The standard devi
ation provides a more intuitive measure of variability, since it is in the s
ame units as the original data.
```