Q1. What is Statistics?

→ Statistics is the branch of mathematics for collecting, analysing and interpreting data. Statistics can be used to predict the future, determine the probability that a specific event will happen, or help answer questions about a survey.

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

→ Statistics is mainly divided into the following two categories.

1. Descriptive Statistics
2. Inferential Statistics

Descriptive Statistics

In the descriptive Statistics, the Data is described in a summarized way. The summarization is done from the sample of the population using different parameters like Mean or standard deviation. Descriptive Statistics are a way of using charts, graphs, and summary measures to organize, represent, and explain a set of Data.

- Data is typically arranged and displayed in tables or graphs summarizing details such as histograms, pie charts, bars or scatter plots.
- Descriptive Statistics are just descriptive and thus do not require normalization beyond the Data collected.

Inferential Statistics

In the Inferential Statistics, we try to interpret the Meaning of descriptive Statistics. After the Data has been collected, analyzed, and summarised we use Inferential Statistics to describe the Meaning of the collected Data.

- Inferential Statistics use the probability principle to assess whether trends contained in the research sample can be generalized to the larger population from which the sample originally comes.

- Inferential Statistics are intended to test hypotheses and investigate relationships between variables and can be used to make population predictions.
- Inferential Statistics are used to draw conclusions and inferences, i.e., to make valid generalizations from samples.

Example

In a class, the Data is the set of marks obtained by 50 students. Now when we take out the Data average, the result is the average of 50 students' marks. If the average marks obtained by 50 students are 88 out of 100, on the basis of the outcome, we will draw a conclusion.

Q3. What are the different types of data and how do they differ from each other? Provide an example of each type of data.

→ Qualitative or Categorical Data

Qualitative or Categorical Data is data that can't be measured or counted in the form of numbers. These types of data are sorted by category, not by number. That's why it is also known as Categorical Data. These data consist of audio, images, symbols, or text. The gender of a person, i.e., male, female, or others, is qualitative data.

Qualitative data tells about the perception of people. This data helps market researchers understand the customers' tastes and then design their ideas and strategies accordingly.

The other examples of qualitative data are :

- What language do you speak
- Favorite holiday destination
- Opinion on something (agree, disagree, or neutral)
- Colors

The Qualitative data are further classified into two parts :

Nominal Data

Nominal Data is used to label variables without any order or quantitative value. The color of hair can be considered nominal data, as one color can't be compared with another color.

The name "nominal" comes from the Latin name "nomen," which means "name." With the help of nominal data, we can't do any numerical tasks or can't give any order to sort the data. These data don't have any meaningful order; their values are distributed into distinct categories.

Examples of Nominal Data :

- Colour of hair (Blonde, red, Brown, Black, etc.)
- Marital status (Single, Widowed, Married)
- Nationality (Indian, German, American)
- Gender (Male, Female, Others)
- Eye Color (Black, Brown, etc.)

Ordinal Data

Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can't do any arithmetical tasks on them.

Ordinal data is qualitative data for which their values have some kind of relative position. These kinds of data can be considered "in-between" qualitative and quantitative data. The ordinal data only shows the sequences and cannot use for statistical analysis. Compared to nominal data, ordinal data have some kind of order that is not present in nominal data.

Examples of Ordinal Data :

- When companies ask for feedback, experience, or satisfaction on a scale of 1 to 10
- Letter grades in the exam (A, B, C, D, etc.)
- Ranking of people in a competition (First, Second, Third, etc.)
- Economic Status (High, Medium, and Low)
- Education Level (Higher, Secondary, Primary)

Difference between Nominal and Ordinal Data

Nominal Data	Ordinal Data
Nominal data can't be quantified, neither they have any intrinsic ordering	Ordinal data gives some kind of sequential order by their position on the scale
Nominal data is qualitative data or categorical data	Ordinal data is said to be "in-between" qualitative data and quantitative data
They don't provide any quantitative value, neither can we perform any arithmetical operation	They provide sequence and can assign numbers to ordinal data but cannot perform the arithmetical operation
Nominal data cannot be used to compare with one another	Ordinal data can help to compare one item with another by ranking or ordering
Examples: Eye color, housing style, gender, hair color, religion, marital status, ethnicity, etc	**Examples:** Economic status, customer satisfaction, education level, letter grades, etc

Quantitative Data

Quantitative data can be expressed in numerical values, making it countable and including statistical data analysis. These kinds of data are also known as Numerical data. It answers the questions like "how much," "how many," and "how often." For example, the price of a phone, the computer's ram, the height or weight of a person, etc., falls under quantitative data.

Quantitative data can be used for statistical manipulation. These data can be represented on a wide variety of graphs and charts, such as bar graphs, histograms, scatter plots, boxplots, pie charts, line graphs, etc.

Examples of Quantitative Data :

- Height or weight of a person or object
- Room Temperature
- Scores and Marks (Ex: 59, 80, 60, etc.)
- Time

The Quantitative data are further classified into two parts :

Discrete Data

The term discrete means distinct or separate. The discrete data contain the values that fall under integers or whole numbers. The total number of students in a class is an example of discrete data. These data can't be broken into decimal or fraction values.

The discrete data are countable and have finite values; their subdivision is not possible. These data are represented mainly by a bar graph, number line, or frequency table.

Examples of Discrete Data :

- Total numbers of students present in a class
- Cost of a cell phone
- Numbers of employees in a company
- The total number of players who participated in a competition
- Days in a week

Continuous Data

Continuous data are in the form of fractional numbers. It can be the version of an android phone, the height of a person, the length of an object, etc. Continuous data represents information that can be divided into smaller levels. The continuous variable can take any value within a range.

The key difference between discrete and continuous data is that discrete data contains the integer or whole number. Still, continuous data stores the fractional numbers to record different types of data such as temperature, height, width, time, speed, etc.

Examples of Continuous Data :

- Height of a person
- Speed of a vehicle

- “Time-taken” to finish the work
- Wi-Fi Frequency
- Market share price

Difference between Discrete and Continuous Data

Discrete Data	Continuous Data
Discrete data are countable and finite; they are whole numbers or integers	Continuous data are measurable; they are in the form of fractions or decimal
Discrete data are represented mainly by bar graphs	Continuous data are represented in the form of a histogram
The values cannot be divided into subdivisions into smaller pieces	The values can be divided into subdivisions into smaller pieces
Discrete data have spaces between the values	Continuous data are in the form of a continuous sequence
Examples: Total students in a class, number of days in a week, size of a shoe, etc	**Example:** Temperature of room, the weight of a person, length of an object, etc

Q4. Categorise the following datasets with respect to quantitative and qualitative data types:
(i) Grading in exam: A+, A, B+, B, C+, C, D, E
(ii) Colour of mangoes: yellow, green, orange, red
(iii) Height data of a class: [178.9, 179, 179.5, 176, 177.2, 178.3, 175.8,...]
(iv) Number of mangoes exported by a farm: [500, 600, 478, 672, ...]

→ (i) qualitative
(ii) qualitative
(iii) quantitative
(iv) quantitative

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

→ First, knowing the level of measurement helps you decide how to interpret the data from that variable. When you know that a measure is nominal (like the one just described), then you know that the numerical values are just short codes for the longer names. Second, knowing the level of measurement helps you decide what statistical analysis is appropriate on the values that were assigned. If a measure is nominal, then you know that you would never average the data values or do a t-test on the data.
There are typically four levels of measurement that are defined:

- **Nominal**
- **Ordinal**
- **Interval**
- **Ratio**

In nominal measurement the numerical values just "name" the attribute uniquely. No ordering of the cases is implied. For example, jersey numbers in basketball are measures at the nominal level. A player with number `30` is not more of anything than a player with number `15`, and is certainly not twice whatever number `15` is.

In ordinal measurement the attributes can be rank-ordered. Here, distances between attributes do not have any meaning. For example, on a survey you might code Educational Attainment as 0=less than high school; 1=some high school.; 2=high school degree; 3=some college; 4=college degree; 5=post college. In this measure, higher numbers mean *more* education. But is distance from 0 to 1 same as 3 to 4? Of course not. The interval between values is not interpretable in an ordinal measure.

In interval measurement the distance between attributes *does* have meaning. For example, when we measure temperature (in Fahrenheit), the distance from 30-40 is same as distance from 70-80. The interval between values is interpretable. Because of this, it makes sense to compute an average of an interval variable, where it doesn't make sense to do so for ordinal scales. But note that in interval measurement ratios don't make any sense - 80 degrees is not twice as hot as 40 degrees (although the attribute value is twice as large).

Finally, in ratio measurement there is always an absolute zero that is meaningful. This means that you can construct a meaningful fraction (or ratio) with a ratio variable. Weight is a ratio variable. In applied social research most "count" variables are ratio, for example, the number of clients in past six months. Why? Because you can have zero clients and because it is meaningful to say that "…we had twice as many clients in the past six months as we did in the previous six months."

It's important to recognize that there is a hierarchy implied in the level of measurement idea. At lower levels of measurement, assumptions tend to be less restrictive and data analyses tend to be less sensitive. At each level up the hierarchy, the current level includes all of the qualities of the one below it and adds something new. In general, it is desirable to have a higher level of measurement (e.g., interval or ratio) rather than a lower one (nominal or ordinal).

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

→ Difference between Nominal and Ordinal

Parameter	Nominal	Ordinal
Definition	Nominal data is defined as the data used for naming or labeling variables, without any quantitative value.	Similar to nominal data, ordinal data is categorical data with an order.
Uses	Used to collect data of people, places, or things.	Used to collect feedback, reviews, or ratings.
Degree of Quantitative Value	There is no quantitative value associated with variables. It is a qualitative measure.	Quantitative values are attached to ordinal variables, but you can’t conduct arithmetic calculations.
Mathematical Operation	Equality	Equality and Comparison

Measures of Central Tendency	**Mode**	**Mode and Median**
Measure of Variability	**None**	**Range and Interquartile Range**
Example	**Sex (male and female)**	**Customer Satisfaction Rating (on a scale 0 – 5)**

What is Nominal Data?

The term ‘nominal’ comes from the Latin word ‘nomen’ or ‘nominalis’, which means name. It is defined as the data type that is used for naming or labeling variables.

- **Classifies the data into named groups without any quantitative meaning.**
    - **i.e., it divides variables into mutually exclusive and labeled categories.**

- **Statistical tests like chi-square goodness of fit and chi-square test of independence are used to analyze the nominal data.**
- **Frequency distribution and mode are used to analyze nominal data.**
- **Nominal variables can be coded with numerical values (arbitrary order), but they can't be used for arithmetic operations.**
- **Example:**
    - **Blood Type: A, B, O, AB**
    - **Sex: Male, Female**
    - **Marital Status: Single, Married, Divorced**

What is Ordinal Data?

Ordinal Data classifies variables that have a natural order or rank.

- **Ordinal data can be presented in tabular and graphical format**
- **Frequency distribution, mode, median, and range are used to analyze Ordinal Data.**
- **Statistical tests like Spearman's rho correlation test, mood's median test, etc., are used to analyze the data.**
- **Example:**
    - **Grades: A, B, C, D, F**
    - **Rating: Worst, Poor, Neutral, Good, Excellent**
    - **Income Group: High, Medium, Low**
    - **Education: 12th, Graduate, Post-Graduate, PhD**

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

→ **The best graph to illustrate ranges in a data series is the box plot, also known as a box-and-whisker plot. This type of graph displays the minimum, first quartile, median, third quartile, and maximum values of a data set, as well as any outliers that may exist. It is a useful tool for quickly understanding the distribution of a data set, and can be particularly helpful for identifying patterns or anomalies in large data sets.**

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

→ Difference Between Descriptive and Inferential statistics

As you can see, Descriptive statistics summarize the features or characteristics of a data set, while Inferential statistics enables the user to test a hypothesis to check if the data is generalizable to the wider population. Now, how can we go from descriptive to inferential statistics? The difference lies in finding the answer to "What is?" vs. "What else it might be?".

The differences between descriptive statistics vs inferential statistics lie as much in the process as in the statistics reported. Given below are the key points of difference in descriptive vs inferential statistics.

- **Descriptive Statistics gives information about raw data regarding its description or features. Inferential statistics, on the other hand, draw inferences about the population by using data extracted from the population.**
- **We use descriptive statistics to describe a situation, while we use inferential statistics to explain the probability of occurrence of an event.**

- **As for descriptive statistics, it helps to organize, analyze and present data in a meaningful manner. Inferential statistics helps to compare data, make hypotheses and predictions.**
- **Descriptive statistics explains already known data related to a particular sample or population of a small size. Inferential statistics, however, aims to draw inferences or conclusions about a whole population.**
- **We use charts, graphs, and tables to represent descriptive statistics, while we use probability methods for inferential statistics.**
- **It is simpler to perform a study using descriptive statistics rather than inferential statistics, where you need to establish a relationship between variables in an entire population.**

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.

⟶ Central tendency and variation are two measures used in statistics to summarize data. Measure of central tendency shows where the center or middle of the data set is located, whereas measure of variation shows the dispersion among data values.

Mean

The mean is the arithmetic average, and it is probably the measure of central tendency that you are most familiar. Calculating the mean is very simple. You just add up all of the values and divide by the number of observations in your dataset.

Median

The median is the middle value. It is the value that splits the dataset in half, making it a natural measure of central tendency.

To find the median, order your data from smallest to largest, and then find the data point that has an equal number of values above it and below it. The method for locating the median varies slightly depending on whether your dataset has an even or odd number of values. I'll show you how to find the median for both cases.

Mode

The mode is the value that occurs the most frequently in your data set, making it a different type of measure of central tendency than the mean or median.

To find the mode, sort the values in your dataset by numeric values or by categories. Then identify the value that occurs most often.

On a bar chart, the mode is the highest bar. If the data have multiple values that are tied for occurring the most frequently, you have a multimodal distribution. If no value repeats, the data do not have a mode.

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

→ **There are 4 levels of measurement:**

- **Nominal: the data can only be categorized**

- **Ordinal**: the data can be categorized and ranked
- **Interval**: the data can be categorized, ranked, and evenly spaced
- **Ratio**: the data can be categorized, ranked, evenly spaced, and has a natural zero.

Nominal, ordinal, interval, and ratio data

Going from lowest to highest, the 4 levels of measurement are cumulative. This means that they each take on the properties of lower levels and add new properties.

Nominal level	Examples of nominal scales
You can categorize your data by labelling them in mutually exclusive groups, but there is no order between the categories.	• City of birth • Gender • Ethnicity • Car brands • Marital status

Ordinal level	Examples of ordinal scales
You can categorize and rank your data in an order, but you cannot say anything about the intervals between the rankings.	• Top 5 Olympic medallists • Language ability (e.g., beginner, intermediate, fluent)

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.	• Likert-type questions (e.g., very dissatisfied to very satisfied)
Interval level	**Examples of interval scales**
You can categorize, rank, and infer equal intervals between neighboring data points, but there is no true zero point. 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. The same is true for test scores and personality inventories. A zero on a test is arbitrary; it does not mean that the test-taker has an absolute lack of the trait being measured.	• Test scores (e.g., IQ or exams) • Personality inventories • Temperature in Fahrenheit or Celsius
Ratio level	**Examples of ratio scales**

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.**	• **Height** • **Age** • **Weight** • **Temperature in Kelvin**