Assignment: [18-02-26]

1. What is statistics?
   Statistics is the branch of mathematics that deals with collecting, organizing, analyzing, interpreting, and presenting data. It helps in making decisions and drawing conclusions from data.

2. Types of statistics
   There are two main types: Descriptive statistics and Inferential statistics. Descriptive statistics summarize and describe data using measures like mean, median, and standard deviation. Inferential statistics use sample data to make predictions or conclusions about a population using tests like t-test, z-test, and ANOVA.

3. What is central tendency? Explain with hands-on code
   Central tendency refers to the measure that identifies the center or typical value of a dataset. The main measures are mean, median, and mode.

   Example code:

```
import pandas as pd
data = pd.Series([10, 20, 30, 40, 50])
print("Mean:", data.mean())
print("Median:", data.median())
print("Mode:", data.mode()[0])
```

4. What is measures of dispersion? Explain range, variance, standard deviation, IQR
   Measures of dispersion show how spread out the data is.
   Range is the difference between maximum and minimum value.
   Variance measures how far each value is from the mean.
   Standard deviation is the square root of variance and shows spread in the same unit as data.
   IQR (Interquartile Range) is Q3 - Q1 and measures the spread of the middle 50% of data.

5. What is t-test?
   A t-test is a statistical test used to compare the means of two groups when the sample size is small and population standard deviation is unknown.

6. What is z-test?
   A z-test is used to compare means when the sample size is large and population standard deviation is known.

7. What is ANOVA? What are its assumptions?
   ANOVA (Analysis of Variance) is used to compare means of more than two groups.
   Assumptions:
   Data should be normally distributed.
   Groups should have equal variances (homogeneity).
   Observations should be independent.

8. What is p-value?
   P-value is the probability of obtaining results at least as extreme as the observed results assuming the null hypothesis is true. A small p-value indicates strong evidence against the null hypothesis.

9. What is hypothesis testing?
   Hypothesis testing is a statistical method used to make decisions about a population parameter using sample data. It involves null hypothesis (H0) and alternative hypothesis (H1).

10. What is z-score?
    Z-score measures how many standard deviations a data point is away from the mean.

11. Why significance level is important?
    Significance level (alpha) is the threshold used to decide whether to reject the null hypothesis. Common value is 0.05. It controls the probability of making a Type I error.

12. What is chi-square?
    Chi-square test is used to determine whether there is a significant association between categorical variables.

13. How the data is distributed?
    Data can be normally distributed (bell-shaped), positively skewed (right skewed), negatively skewed (left skewed), or uniform distribution. Example: heights follow normal distribution, income often follows right-skewed distribution.

14. What is inferential statistics?
    Inferential statistics uses sample data to make generalizations or predictions about a larger population.

15. What is correlation?
    Correlation measures the strength and direction of relationship between two variables. It ranges from -1 to +1.

16. What is regression analysis?
    Regression analysis is a statistical method used to model the relationship between dependent and independent variables.

17. What are independent and dependent variables?
    Independent variable is the input or predictor variable. Dependent variable is the output or target variable that depends on the independent variable.

18. How statistics impact ML models?
    Statistics helps in understanding data distribution, feature selection, handling outliers, hypothesis testing, and model evaluation. Many ML algorithms are based on statistical concepts.

19. What are data attributes?
    Data attributes are characteristics or features of data such as name, age, salary, category, etc.

20. What is qualitative and quantitative?
    Qualitative data is non-numerical data like gender or color. Quantitative data is numerical data like age or salary.

21. Difference between continuous and categorical data
    Continuous data can take any value within a range (height, weight). Categorical data represents categories or groups (gender, city).

22. What is data?
    Data is raw facts, numbers, or information collected for analysis.

23. Difference between structured and unstructured data
    Structured data is organized in tabular form like databases. Unstructured data includes text, images, videos, and audio.

24. What are outliers? Why are they important?
    Outliers are data points that are significantly different from other observations. They are important because they can affect statistical results and model performance.

25. How to find outliers in dataset?
    Using IQR method:
    Lower limit = Q1 - 1.5 × IQR
    Upper limit = Q3 + 1.5 × IQR
    Values outside this range are outliers.

    Example code:

```
import pandas as pd
data = pd.Series([10, 12, 14, 15, 100])
Q1 = data.quantile(0.25)
Q3 = data.quantile(0.75)
IQR = Q3 - Q1
lower = Q1 - 1.5 * IQR
upper = Q3 + 1.5 * IQR
outliers = data[(data < lower) | (data > upper)]
print("Outliers:", outliers)
```