Day -13 : Assignment Answers

1. What is Statistics?

Statistics is the branch of mathematics that deals with collecting, organizing, analyzing, interpreting, and presenting data.
It helps us understand patterns and make decisions based on data. In data science, statistics is the foundation for analysis and machine learning models.

2. Types of Statistics

There are two main types:

Descriptive Statistics – It summarizes data using measures like mean, median, mode, range, and standard deviation.

Inferential Statistics – It makes predictions or conclusions about a population based on sample data. It includes hypothesis testing, t-test, z-test, and ANOVA.

3. What is Central Tendency?

Central tendency represents the center value of a dataset. The main measures are mean, median, and mode.

Hands-on Code Example:

```
import numpy as np

data = [10, 20, 30, 40, 50]

mean = np.mean(data)
median = np.median(data)
mode = max(set(data), key=data.count)

print("Mean:", mean)
print("Median:", median)
print("Mode:", mode)
```

4. Measures of Dispersion

Measures of dispersion show how spread out the data is.

Range – Maximum minus minimum value.
Variance – Average of squared differences from the mean.
Standard Deviation – Square root of variance.
IQR – Difference between third quartile (Q3) and first quartile (Q1).

5. What is t-test?

A t-test compares the means of two groups when the sample size is small and population variance is unknown.

6. What is z-test?

A z-test compares sample mean with population mean when sample size is large and population variance is known.

7. What is ANOVA?

ANOVA (Analysis of Variance) compares means of more than two groups.

Assumptions:
- Data should be normally distributed.
- Variances should be equal.
- Observations should be independent.

8. What is p-value?

P-value shows the probability of obtaining results if the null hypothesis is true. If p-value is less than 0.05, we reject the null hypothesis.

9. What is Hypothesis Testing?

Hypothesis testing is a statistical method to test assumptions about population data using sample data. It includes null hypothesis (H0) and alternative hypothesis (H1).

10. What is Z-score?

Z-score tells how many standard deviations a data point is away from the mean.

11. Why Significance Level is Important?

Significance level (usually 0.05) sets the threshold to reject the null hypothesis and controls Type I error.

12. What is Chi-Square?

Chi-square test checks the relationship between categorical variables by comparing observed and expected values.

13. How Data is Distributed?

Data can follow different distributions such as:
- Normal distribution
- Uniform distribution
- Binomial distribution
- Poisson distribution

14. What is Inferential Statistics?

Inferential statistics uses sample data to make conclusions about the entire population.

15. What is Correlation?

Correlation measures the strength and direction of relationship between two variables.

16. What is Regression Analysis?

Regression analysis predicts dependent variable based on independent variable(s).

17. Independent and Dependent Variables

Independent variable – Input variable.
Dependent variable – Output variable.

18. How Statistics Impacts ML Models?

Statistics helps in data preprocessing, feature selection, model evaluation, and improving accuracy.

19. What are Data Attributes?

Data attributes are features or characteristics of data like age, salary, gender.

20. Qualitative and Quantitative Data

Qualitative – Non-numerical data.

Quantitative – Numerical data.

21. Continuous vs Categorical Data

Continuous – Can take any value within a range.
Categorical – Represents categories or groups.

22. What is Data?

Data are raw facts collected for analysis.

23. Structured vs Unstructured Data

Structured data is organized in tables. Unstructured data includes text, images, and videos.

24. What are Outliers?

Outliers are extreme values that differ from other observations. They may affect model accuracy.

25. How to Find Outliers?

Outliers can be found using:
- Z-score method
- IQR method
- Boxplot visualization